Some Investigations on Protein Structure, Function and Dynamics in Disordered states and Non-ideal conditions

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

M. VENKATA SATISH KUMAR

ROLL NO. 06610609

DEPARTMENT OF BIOTECHNOLOGY

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI, INDIA

FEBRUARY 2010

Some Investigations on Protein Structure, Function and Dynamics in Disordered states and Non-ideal conditions

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

M. VENKATA SATISH KUMAR

ROLL NO. 06610609

DEPARTMENT OF BIOTECHNOLOGY

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI, INDIA

FEBRUARY 2010

Dedicated
to
My Parents ,
my wife , and
my son Shriram

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

DEPARTMENT OF BIOTECHNOLOGY

STATEMENT

I do hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the Department of Biotechnology, Indian Institute of Technology Guwahati, India under the guidance of Prof. Rajaram Swaminathan.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work described is based on the findings of other investigators.

Date: 17th February, 2010 **M. VENKATA SATISH KUMAR**

IIT Guwahati

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

DEPARTMENT OF BIOTECHNOLOGY

Prof. Rajaram Swaminathan
Department of Biotechnology,
IIT Guwahati, Assam-781039, INDIA

Phone: +91-361-258-2248
Fax: +91-361-258-2249
Email: rsw@iitg.ernet.in

CERTIFICATE

It is certified that the work described in this thesis entitled "***Some investigations on protein structure, function and dynamics in disordered states and non-ideal conditions***" done by Mr. M. Venkata Satish Kumar for the award of degree of Doctor of Philosophy is an authentic record of the results obtained from the research work carried out under my supervision in the Department of Biotechnology, Indian Institute of Technology Guwahati, India, and this work has not been submitted elsewhere for a degree.

Date: 17th February, 2010 **Prof. Rajaram Swaminathan**

IIT Guwahati **(Supervisor)**

INDIAN INSTITUTE OF TECHNOLOGY, GUWAHATI

Department of Biotechnology

CERTIFICATE OF COURSE WORK

This is to certify that M. Venkata Satish Kumar has satisfactorily completed all the courses required for the Ph.D degree program. These courses include

BT 601:	Analytical Biotechnology
BT 602:	Basic Biotechnology
BT 604:	Enzymology
BT 610:	Frontiers in Biomolecular simulations

M. Venkata Satish Kumar has successfully completed his Ph.D qualifying examination in May 2007.

Professor Arun Goyal
Head, Department of Biotechnology
I. I. T. Guwahati

Dr. Utpal Bora
Secretary, DPPC
I. I. T. Guwahati

Acknowledgements

I would like to express my sincere gratitude toward toward my adviser, Prof. Rajaram Swaminathan, for his insightful guidance, invaluable encouragement and countless support. I did not have much research experience when I started my Ph.D. studies. It was under the direction of Prof. Rajaram Swaminathan that I finished the first research project in my life. Challenging and smart questions raised by my adviser taught me how to check the results of research from a variety of ways, and how to ask the right questions to obtain the most interesting results. Prof. Rajaram Swaminathan is an exceptional expert on explaining complicated phenomenon by simple models. His teaching and principles are invaluable to my future work. Prof. Rajaram Swaminathan gave me many suggestions on this thesis.

I would like to acknowledge my sincere gratitude to all my doctoral committee members, Dr. Subhradip Ghosh, Dr. Latha Rangan and Dr. Biplab Bose for their insightful advices and valuable suggestions.

I am much obligated to Dr. Pradipta Bandyopadhyay for introducing me the subject, Computational Biology.

I am much obliged to all other faculty members in the Department of Biotechnology for their help and encouragement and the non-teaching staffs of the Department for their technical support. I would like to take this opportunity to thank Department of Biotechnology for proving me the necessary computational facilities during the entire duration of my research tenure. I am thankful to the Institute, Indian Institute of Technology Guwahati for providing me with the state of the art infrastructure and facilities for advanced research. The financial support from All India Council for Technical Education (AICTE), New Delhi in the form of QIP is duly acknowledged.

I would like to thank my research group members Satish, Ravi, Meher, Nividh Chandra, Akshay, Narendra Reddy, for their timely help, constant support, and for the wonderful time we shared during this period.

I would like to extend my gratitude to my departmental friends Kausik, Achleesh, Naresh, Preeti, Sanjay, Vinod, Shadab, Seema, Priyanka, Uzma, Souvika, Mithilesh, Seraj, Kasturi, Vigya, Atul, Chockalingam, Asim, Sahil, Saravanan, Santhosh, Leela Krishna, Shashank, Amit, Alpana, Prarthana, Sharan, Nurul ,Anitha, Rashmi, Niranjan, Krishna Das, for their timely support and steady help.

At last I would like to extend my thanks to my father, my wife and my son. I would not have been able to finish this challenging journey of Ph.D. study without their support and understanding.

Date: 17th February, 2010. **M. Venkata Satish Kumar**

Abstract:

The proteins we observe in nature have evolved through selective pressure to do certain specific functions such as catalyzing and regulating biochemical reactions, transporting molecules that permit cells to grow and reproduce. The three dimensional structure of protein is fundamentally related and tied to its biological function. So understanding the structure of protein is important to obtain insights on its function.

The details of ordered secondary structures like helices and sheets in proteins are extensively discussed in the literature, but there are few reports on the remaining secondary structures comprising of turns, bends, bulges and other irregular structures that are commonly grouped and referred to as loops. Loops are essentially regions of non-repetitive conformation connecting regular secondary structures. These regions are important parts of protein structures as they have been shown in some cases to be the nucleation sites for protein folding, the sites of catalytic activity in enzymes or interfacial regions in the interaction between other proteins, ligands and nucleic acids. Thus loops play an important role in protein function.

The structural-genomics initiative is expected to boost the population of high resolution protein 3D structures. Consequently, it has been argued that manual analysis of protein structure must be replaced by automated approaches of protein structure analysis. Loops play a major role in determining protein fold, hence methods that can facilitate automated analysis of loops in proteins are desirable. We have devised a method to identify and investigate functionally active loops and unstructured regions in protein structures using the MSRP parameter. This method (a) provides a unique classification tool for loops and folds among proteins, (b) permits automated identification of functional loops in protein structures, (c) provides clues on the diversity of conformations sampled by a disordered region during a molecular dynamics simulation.

Traditionally acknowledged concept of protein function was the structure-function paradigm, represented as Amino acid sequence-> 3D structure->Function. The 3D structure of protein in the folded state is related to its function and thus the native protein structure is the ordered 3D structure. However, recently it has emerged that the actual functional state for many proteins and protein domains are intrinsically unstructured. These unstructured proteins are also called intrinsically disordered proteins or natively unfolded proteins. They comprise a large fraction of eukaryotic

proteins with 33% that are either completely or partially disordered. These proteins fulfill essential biological functions like regulation of transcription and translation, cellular signal transduction, protein phosphorylation, storage of small molecules and the regulation of the self-assembly of large multiprotein complexes where their structural plasticity plays a crucial role. They are characterized by the lack of stable secondary and tertiary structure under physiological conditions and in the absence of a binding partner or ligand. The unstructured or disordered regions in these proteins are mainly in the form of short and long irregularly shaped loops. Despite the large abundance and importance of disordered proteins, the disordered regions in these proteins are still poorly understood. Several experimental methods such as nuclear magnetic resonance, circular dichroism spectroscopy and proteolysis are used for the characterization of disordered state. But these methods cannot directly sample protein dynamics in the time scale (ns) relevant to conformational changes in such disordered regions and therefore provide only indirect information on the disordered state. Thus in disordered proteins, the structural heterogeneity and rapid inter-conversion among conformers lead to problems in characterization and present practical challenges. Nowadays, MD simulations have been widely used as a powerful tool to understand the conformational dynamics of proteins at atomic level. This method has been less explored in the field of disordered proteins. It would be worthwhile to investigate the dynamics of disordered proteins using MD simulations. To probe the unique dynamical features of disordered regions in disordered proteins, we have studied and compared well ordered proteins with disordered proteins and also ordered regions and disordered regions within disordered proteins. We have characterized intrinsically disordered proteins with MD simulation and identified features unique to disordered regions and disordered proteins.

Generally the information regarding the kinetic parameters (k_{cat} and K_m), equilibria, and mechanism of biochemical reactions are obtained through experiments conducted on solutions containing low concentrations (less than about 1 mg/ml) of total protein, nucleic acid, and/or polysaccharides together with buffer salts, low molecular weight substrates, and cofactors as required. In contrast, biochemical reactions in eukaryotic cells take place in an environment containing substantially greater total concentrations (50-400 mg/ml) of macromolecules such as cytoskeleton filaments and microtubules, various organelles and a variety of other macromolecular species like proteins, nucleic acids etc. In general, no single macromolecular species

exists at high concentration but when taken together, macromolecules occupy a significant portion (typically 20-30%) of the total volume. This volume is unavailable to other molecules. Such environments are called crowded. Thus the milieu inside the cell is crowded and thermodynamically non ideal. The consequences of such crowding on reaction equilibria and diffusion phenomena occurring in such media with a high volume fraction of obstacles needs to be understood.

But most of the studies used crowding agents of fixed size only. But it is known that the macromolecules present inside the cell, exist in various sizes and shapes. Thus, it is desirable to investigate the effect of different sizes of crowding agents on enzyme activity. Accordingly, we investigated how the kinetics of an enzymatic reaction is dependent on size and concentration of crowding species. In addition, we have also studied the effect of different substrates on enzyme activity under crowding conditions. We have used a newer method to ensure efficient mixing of large molecular weight dextrans with enzyme. Our findings depicts that size and concentration of macromolecule play a crucial role in influencing the rate of an enzymatic reaction. The effect of crowding by smaller dextrans (40 kDa) showed minor decrease in rate in comparison with larger dextrans (500 and 2000 kDa) in the case of AP vs. PNPP. We observed the effect of crowding by dextrans to have opposite effects on two different substrates IA and NA with acetyl cholinesterase. The effect of crowding by 200 kDa dextran size clearly stimulates the reaction at low concentrations with NA, while it inhibits it appreciably with IA unlike other dextrans of larger size. This finding shows that the increase in activity of $ES^{\neq}$ complex is selective on substrate. Thus our results reveal that the effect of crowding on enzymatic reactions is not simple as it appears and depends on crowder size and substrate nature.

Another outcome of macromolecular crowding is protein aggregation. The question of the overall effect of macromolecular crowding on protein aggregation and fibrillation is not as simple as it appears. There are evidences to show enhancement of the undesirable aggregation of partially unfolded proteins by macromolecular crowding when the intrinsic folding rate of the protein is relatively slow. Irreversible unfolding due to aggregation of unfolded states in the presence of crowding agents has been observed for dihydrofolate reductase, enolase and green fluorescent protein. Overall, these studies infer that the macromolecular crowding might dramatically enhance the competition between protein folding and aggregation, favoring the latter

when folding is relatively slow. Currently, the protein aggregation represents an important problem in biomedicine and biotechnology. In the recent past, there are studies highlighting recognition of the protein deposition or conformational disorders which include Parkinson's diseases, Down's syndrome, Alzheimer's disease, Huntington disease and so on.

The structural details of species formed early during the aggregation process and their features which trigger amyloidosis seems to be important. The interest in amyloid forming proteins has led to the investigation of well characterized protein such as Human Lysozyme, as our model system to examine structural and mechanistic principles that may generally applicable to all amyloid fibrils. There are two known natural mutations of the human lysozyme: D67H and I56T (Pepys *et al*., 1993). They are shown to cause autosomal dominant hereditary nonneuropathic systemic amyloidosis (this is a condition whereby there is amyloid deposition in the viscera and other body cavities). Amyloids are formed because of diverse sequence, fold and function. The mechanism of fibrilogenesis is not clear but appears related to changes in stability and tendency to aggregate due to mutations.

. So it will be worthwhile to investigate the structural feature of the partially folded structures of wild type and the mutants of human lysozyme that may trigger amyloid formation. MD simulations can be employed to understand the conformational dynamics of the proteins at atomistic level. We have used the MD simulations to analyze and compare the conformational dynamics originating from wild type and mutants of human lysozyme. The analyses of backbone RMSD, B factor values, SASA, end to end chain distance, secondary structure content, distance matrix, S^2 order values, conformational entropy, water movement around residues in core domain, and hydrophobic contacts all show the appreciable structural destabilization in mutants compared to wild type. The higher content of beta sheet secondary structure, increased flexibility and disruption in hydrophobic contacts near the alpha/beta domain interface and in beta domain in the case of mutants I56T, D67H perhaps leads to amyloidogenicity. Our results also to some extent confirms and indicates the residue involved in hydrophobic core Y38 (near the alpha/beta domain interface) to be the seed for fibril formation in mutants.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

μs	microsecond
μM	micro molar
Å	Angstroms
AchE	Acetyl Cholinesterase
ADP	Adenosine 5-diphosphate
AMBER	Assisted Model Building with Energy Refinement
AP	Alkaline Phosphatase
ATP	Adenosine triphosphate
BPTI	Bovine pancreatic trypsin inhibitor
Da	Dalton
3DFFT	Three dimensional fast Fourier transform
DNA	Deoxyribo Nucleic Acid
DSSP	Dictionary of Secondary Structure for Proteins
ER	Endoplasmic reticulum
FFT	Fast Fourier Transforms
IA	3-Indoxyl Acetate
IDP	Intrinsically Disordered Protein
K	Kelvin
LEaP	Link Edit and Parm
MD	Molecular Dynamics
mg	milli grams
ml	milli litres
mM	milli molar
MSRP	Mean Separation of Points in Ramachandran Plot
NA	2-Naphthyl Acetate
NMR	Nuclear Magnetic Resonance
ns	nanosecond
PBC	Periodic Boundary Condition

PDB	Protein Data Bank
PME	Particle-Mesh Ewald
PNPP	Para-nitrophenyl phosphate
ps	picosecond
R_g	Radius of Gyration
RMSD	Root Mean Square Deviation
RNA	Ribo Nucleic Acid
SASA	Solvent Accessible Surface Area
SCOP	Structural Classification of Proteins
SPDBV	Swiss PDB viewer
TIP3P	Transferable Intermolecular Potential 3-Point
VMD	Visual Molecular Dynamics
WT	Wild Type

CHAPTER 1
Introduction

1. Introduction:

The structure, function and dynamics of few proteins in disordered and non ideal conditions were investigated in this thesis work. The structurally disordered loop regions in proteins have important biological functions. Although they are part of the protein, their structures are least accurately predicted in comparison to helices or sheets. This led to the extensive studies on the structural details and identification of loop regions in proteins.

Despite the fact that which they fail to form fixed 3D structure under physiological conditions, intrinsically disordered proteins exist as ensembles of conformations, carry out critically important biological functions. Several recent studies provide evidence to the growing interest in these proteins.

The phenomenon of macromolecular crowding that is ever-present in cells is generally under appreciated. Actually crowding results in non ideal environment inside the cell and thereby affects the rates and the equilibria of interactions involving macromolecules, but such interactions are commonly studied outside the cell in uncrowded buffers. However, in recent years a number of reports have appeared in literature, indicating that the phenomenon is getting attention. The association of protein misfolding, aggregation processes with the pathological conditions that include Alzheimer's and Parkinson's diseases, systemic amyloidoses and type II diabetes has led to extensive studies on this disordered condition. A brief review of these areas is presented below, highlighting their respective importance and applications. The current status of research in these areas is also reviewed briefly along with the problems that we have attempted to address.

1.1. Protein structure, function and dynamics:

When we view a cell under microscope or analyze its activity, we are in fact observing proteins. This is because proteins constitute most of the cell's dry mass and serve as tiny molecular machines to promote cell's functions. The proteins have evolved through selective pressure to do certain specific functions. Proteins play crucial, life-sustaining biological roles, both as constituent molecules and as triggers of physiological processes for all living things. For example, proteins provide architectural support in muscle tissues, ligaments, tendons, bones, skin, hair, organs, and glands. Their

environment-tailored structures make possible the coordinated function (motion, regulation, etc.) in some of these assemblies.

Proteins also provide the essential services of transport and storage, such as oxygen and iron in muscle and blood cells. The foremost pair of solved protein structures hemoglobin and myoglobin, serve as the crucial oxygen carriers in vertebrates. Hemoglobin is found in red blood cells and is the chief oxygen carrier in the blood (it also transports carbon dioxide and hydrogen ions). Myoglobin is found in muscle cells, where it stores oxygen and facilitates oxygen movement in muscle tissue. The sperm whale depends on myoglobin in its muscle cells for large amounts of oxygen supplies during long underwater journeys.

Proteins further play critical regulatory roles in many basic processes fundamental to life, such as reaction catalysis (e.g., digestion), immunological and hormonal functions, and the coordination of neuronal activity, cell and bone growth, and cell differentiation. Thus they are involved in the house-keeping activities taking place inside the cell. The three dimensional structure of protein is fundamentally related and tied to its biological function. Proteins that perform similar functions tend to show a considerable degree of structural homology (**Chan and Dill, 1993**; **Voet and Voet, 1990**). So understanding the structure of protein is important to obtain insights on its function.

From a chemical point of view, proteins are by far the most structurally complex and functionally sophisticated molecules known. Proteins consist of a linear chain of a particular sequence of varying length, containing anywhere from tens to thousands of the twenty naturally occurring amino acids each linked to its neighbor through a covalent peptide bond. Proteins are therefore also known as polypeptides. The amino acid sequence of protein chain is called its primary structure. Each type of protein has a unique sequence of amino acids, and there are several thousands of different proteins, each with its own particular amino acid sequence. Different regions of the protein chain form local regular secondary structures such as α-helices, β-sheets and random coils. The ϕ and ψ angles between adjacent amino acids in the protein chain are called the torsion angles (see **Figure 1.1**), determine the twists and turns in the sequence which result in these secondary structures. The tertiary structure is formed by packing such

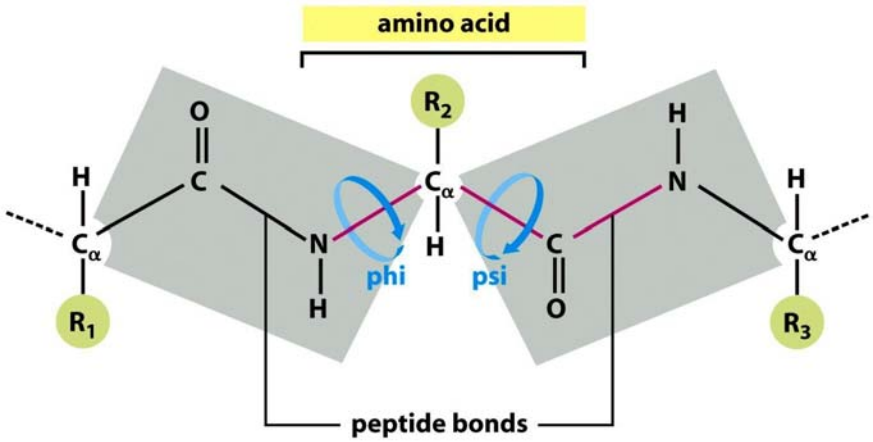

Figure 1.1: *Each amino acid contributes three bonds (red) to the backbone of the chain. The peptide bond is planar (gray shading) and does not permit rotation. By contrast, the rotation can occur about the C_α-C bond, whose angle of rotation is called psi (ϕ), and about the N- C_α bond, whose angle of rotation is called phi (ψ) (adapted from Molecular Biology of THE CELL, 5/e).*

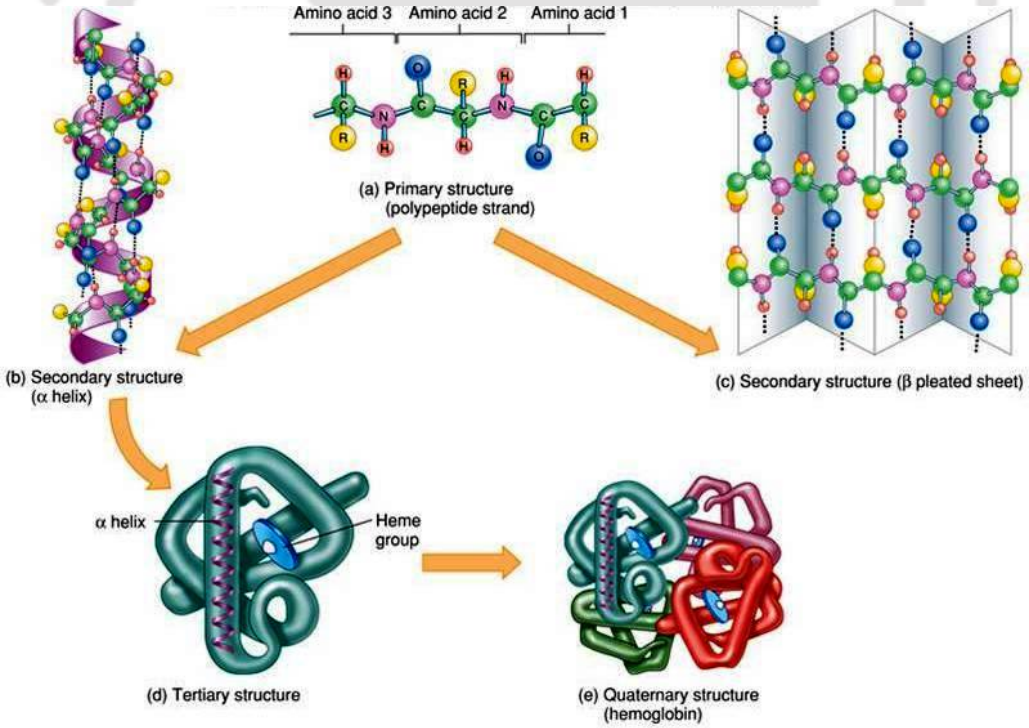

Figure 1.2: *The Protein structure hierarchy, from primary to quaternary structure (adapted from stevebambas.com/images).*

structural elements into one or several compact globular units called domains. The final protein may contain several polypeptide chains arranged in a quaternary structure. The details about different structures of protein in nutshell are shown in **Figure 1.2.** By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region or an active site. Thus protein structures are built up from combinations of secondary structure elements like alpha helices, beta strands and random coils.

"Biology is all about wiggling and jiggling of atoms" said Richard Feynman several decades ago. The functioning of all the biological processes involves simple elementary principles but with complex structures and their associated dynamics. Thus, a complete understanding of biological processes requires information on dynamics apart from the knowledge of their high resolution structure. The biological processes such as Protein folding, binding, catalytic activity and molecular recognition all involve molecular movements of different extents. The molecular movements are brought upon via flexible regions. Stemming from sequence, modifications of electrostatic and hydrophobic properties of the protein fold determine flexible and rigid regions. Flexible regions are central to both protein folding and function.

$$Sequence \xrightarrow{Flexible\ regions} Structure(fold) \xrightarrow{Flexible\ regions} Function \quad \text{-- (1.1)}$$

Protein flexible regions allow the exact movement in thousands of atomic coordinates to perform function. Several examples are available in the literature that accounts for the protein movement. This includes flap movements in retroviral HIV-1 protease, domain movements in T-4 lysozyme, calmodulin and adenylate kinase and fragment movements in lactate dehydrogenease. The relations between their movements and function have been established by X-ray crystallography and other experimental observations. For example, flap movements in retroviral protease are conserved throughout the family of aspartyl protease. Consisting of a β-hairpin, the flaps at the ceiling of the binding pocket move about 7 Å between its 'closed 'and 'open' conformations (**Wlodawer *et al.*, 1989**; **Lapatto *et al.*, 1989**; **Spinelli *et al.*, 1991**). The speedy flap movements have been shown by NMR and fluorescence changes. In the case of adenylate kinase a four stranded anti-parallel β-sheet is shown to undergo considerable displacement upon substrate binding (**Vonrhein *et al.*, 1995**; **Anderson *et al.*, 1979**).

Calmodulin, T-4 lysozyme, troponin C, lactoferrin and glutamate dehydrogenase are other such authentic examples, where the movements have been crystallized, indicating clear movements (**Gerstein and Krebs, 1998**). The extent of movement generally varies and depends on their functional requirements. For example DNA polymerase- β undergoes a very large movement (about 11 Å to accommodate DNA) as compared to glutamate dehydrogenase (about 0.5 Å) (**Sinha *et al.*, 2001a**). "Domain-swapping" (**Schlunegger *et al.*, 1997**), the binding to multiple substrates under different conditions, allosteric-regulation, operation of molecular motors and binding cascades all are also due to conformational adaptabilities (**Sinha *et al.*, 2001a**; **Sundberg and Mariuzza, 2000**; **Sinha *et al.*, 2001b**; **Ramakrishnan *et al.*, 2002**), arising from flexible regions. Also, the Flexibility/rigidity compensations determine protein thermostability. The binding site of a secondary antibody, of high affinity towards its antigen, would consist of flexible and rigid regions, "preselected" for their respective roles. Similarly, in hinge-bending type of movement, as observed in adenylate kinase or calmodulin, the flexible hinge-points are selected to allow the motion. "Lock and key" or "induced fit" type of binding are also selected, rather than just an outcome of the structural details. It is apparent that the protein movements, whether involving subunits, domains or any secondary structural elements, are uniquely selected for the respective function. Since fold also relates to function, a particular sequence is evolutionally selected for both structure and function.

1.2. Loops and their role in protein function:

The details of ordered secondary structures like helices and sheets in proteins are extensively discussed in the literature (**Chothia *et al.*, 1977**; **Chothia, 1984**; **Efimov, 1979**; **Chou and Fasman, 1977**), but there are few reports (**Tramontano, 1996; Kanagasabai *et al.*, 2007**) on the remaining secondary structures comprising of turns, bends, bulges and other irregular structures that are commonly grouped and referred to as loops. Loops are essentially regions of non-repetitive conformation connecting regular secondary structures. A combination of secondary structure elements forms the stable hydrophobic core of the molecules. The loop regions are at the surface of the molecule. The main chain C=O and NH groups of these loop regions, which in general do not form hydrogen bonds to each other, are exposed to the solvent and can form hydrogen bonds to water molecules. These regions are important parts of protein structures as they have

been shown in some cases to be the nucleation sites for protein folding, the sites of catalytic activity in enzymes or interfacial regions in the interaction between other proteins, ligands and nucleic acids. Thus loops play an important role in protein function. There are many examples in the literature that relate loops to protein function: (a) recognition sites , such as Complementarity Determining Regions, (**Kim *et al.*, 1999**); (b) protein-protein interactions, such as signaling cascades (**Bernstein *et al.*, 2004**; **Zomot and Kanner, 2003**), dimerisation (**Feng *et al.*, 2003**; **Fritz-Wolf *et al.*, 1996**) and protease inhibitors (**Jackson and Rusell, 2000**); (c) ligand binding, such as the P-loop (**Saraste *et al.*, 1990**), EF-hand (**Kawasaki and Kretsinger, 1995**), NAD(P)-binding loops (**Wierenga *et al.*, 1986**) and glycine-rich-loop (**Schenk and Snaar-Jagalska, 1999**); (d) DNA-binding (**Tainer *et al.*, 1995**); (e) forming enzyme active sites, such as Ser-Thr kinases (**Johnson *et al.*, 1998**) and serine proteases (**Wlodawer *et al.*, 1989**); (f)'triggering' loops whose conformational change is required for the catalytic process of enzymes such as β 1,4-Galactosyltransferase (**Gunasekaran and Nussinov, 2004**); (g) driving membrane insertion of pore forming bacterial proteins, such as the anthrax protective antigen (**Benson *et al.*, 1998**) and aerolysin (**Iacovache *et al.*, 2006**).

In most of the enzyme catalyzed reactions, loops play important role and are involved in interaction with the substrate and other ligands like metal ions, metabolites and so on. Such interactions between ligand and protein have been studied by measuring spatial displacement of the concerned chain or atoms subsequent to ligand binding. To get the finer details on such interactions, it would be important to know the regions that undergo structural changes subsequent to binding. This information may be necessary for designing better inhibitors, drugs and so on.

Loops occur ubiquitously among proteins and long escaped from structural classification because of their flexibility and non-periodic nature. But there have been attempts to classify and categorize them in the past based on their conformations (**Venkatachalam, 1968**; **Srinivasan *et al.*, 1991**; **Donate *et al.*, 1996**). The dihedral angles of these irregular regions has been shown to be scattered over several regions in the Ramachandran map, while their structure depends on their sequence, their length and also on the elements and molecular packing of the regular secondary structure, like α-helices or β-strands, to which they connect (**Efimov, 1993**). During attempts to model

protein structures, for example using homology modeling, loops have always posed a major challenge owing to their large variability among different protein families. In fact prediction of loops from the protein sequence is a major issue (Burke **and Deane, 2001**; **Dovidchenko *et al.*, 2008**) in contrast to the fairly accurate predictions that are possible with regular secondary structure (**Raghava, 2002; Guharoy and Chakrabarti, 2007**).

The structural-genomics initiative is expected to boost the population of high resolution protein 3D structures. Consequently, it has been argued that manual analysis of protein structure must be replaced by automated approaches of protein structure analysis (**Cootes *et al.*, 2003**). Loops play a major role in determining protein fold, hence methods that can facilitate automated analysis of loops in proteins are desirable (**Espadaler *et al.*, 2004**).

Scope of my work

In the work described in **Chapter 2**, we analyze the Ramachandran maps of loop regions and other secondary structure elements with the objective of obtaining some features unique to loop regions. The tightly clustered dihedral points at confined regions of the Ramachandran map strongly correlate with the presence of regular secondary structure in the protein. An irregular structure like loop in contrast has no dihedral angle constraints or preferences yielding a scattered distribution of points in the Ramachandran space. This extent of scattering is quantitatively measured to arrive at a parameter (MSRP) that has unique value for loop regions in the protein structure. Subsequently changes in MSRP upon ligand binding or during molecular dynamics was also investigated.

1.3. Intrinsically disordered proteins and their importance:

Traditionally acknowledged concept of protein function was the structure-function paradigm, represented as

$$\text{Amino acid sequence} \rightarrow \text{3D structure} \rightarrow \text{Function} \qquad \text{----- (1.2)}$$

The 3D structure of protein in the folded state is related to its function (e.g. catalysis) and thus the native protein structure is the ordered 3D structure. Large number of experimental evidence has been reported since the 1890s to support this observation. Some important among them include theoretical models postulated by Pauling (**Pauling *et al.*, 1951**), Fischer's lock and key hypothesis (**Fischer, 1894**), the first crystal

structures of globular proteins (**Kendrew *et al.*, 1958, Kendrew *et al.*, 1960**) and of enzymes (**Blake *et al.*, 1965**), and the studies that supported the refoldability proteins into their functional states (**Anson and Mirsky, 1925**; **Anfinsen, 1973**), in which a protein was shown to regain its function if the necessary environmental conditions were restored after the intial perturbation. Infrequent counter examples to the general view presented above have been observed over many years, but these were mostly under appreciated and largely overshadowed by the success of the studies of proteins with specific 3D structures, or what we call ordered proteins. However, recently it has emerged that the actual functional state for many proteins and protein domains are intrinsically unstructured. These unstructured proteins are also called intrinsically disordered proteins (IDPs) (**Dunker *et al.*, 2001**), natively disordered proteins (**Daughdrill *et al.*, 2005**), natively unfolded proteins (**Weinreb *et al.*, 1996**), and intrinsically unstructured proteins (**Wright and Dyson, 1999**). They comprise a large fraction of eukaryotic proteins with 33% that are either completely or partially disordered (**Ward *et al.*, 2004**). These proteins fulfill essential biological functions like regulation of transcription and translation (**Uversky *et al.*, 2005**), cellular signal transduction (**Iakoucheva *et al.*, 2002**), protein phosphorylation, storage of small molecules and the regulation of the self-assembly of large multiprotein complexes (**Romero *et al.*, 2004**; **Dyson and Wright, 2005**) where their structural plasticity plays a crucial role and thus complement the functional repertoire of ordered regions, evolved mainly to carry out efficient catalysis. The IDPs are characterized by the lack of stable secondary and tertiary structure under physiological conditions and in the absence of a binding partner or ligand. The lack of stable secondary and tertiary structure is thought to provide several advantages, such as (i) an increased interaction surface area, (ii) conformational flexibility to interact with several targets, (iii) the presence of molecular recognition elements that fold upon binding, (iv) accessible post-transitional modification sites, and (v) the availability of short linear interaction motifs (**Wright and Dyson, 1999**; **Kriwacki *et al.*, 1996; Tompa, 2005; Oldfield *et al.*, 2005a**). In most of the IDPs, the unstructured or disordered regions exist mainly in the form of short and long irregularly shaped loops. Despite the large abundance, unusual structural and functional importance of disordered proteins, the disordered regions in these proteins are still poorly understood. It has been observed that

altered abundance of IDPs in cell is associated with several disease conditions. For example, over expression of thyroid cancer 1 (TC-1) (**Sunde *et al.*, 2004**) or under expression of adenosine 5-diphosphate (ADP) ribosylation factor (Arf) (**Sherr, 2006**) and p27 (**Grimmler *et al.*, 2007**) has been linked with various types of cancer. Similarly, over expression of α-synuclein and tau proteins increases the risk of aggregate formation and has been linked to Parkinson's disease and Alzheimer's disease (**Chiti and Dobson, 2006; Goedert, 2001**).

The disorder in IDPs has been characterized by several physico-chemical methods such as nuclear magnetic resonance (**Dyson and Wright, 2002, 2004, 2005**; **Bracken *et al.*, 2004**), near ultra violet circular dichroism (CD) (**Fasman, 1996**), far-ultraviolet CD (**Adler *et al.*, 1973**; **Provencher and Glockner, 1981**; **Woody, 1995**; **Uversky *et al.*, 2000**), ORD (**Adler *et al.*, 1973**; **Uversky *et al.*, 2000**), Fourier transform infrared (**Uversky *et al.*, 2000**), Raman spectroscopy and Raman optical activity (**Smyth *et al.*, 2001**), different fluorescence techniques (**Uversky, 1999**; **Receveur-Brechot *et al.*, 2006**), numerous hydrodynamic techniques (including gel-filtration, viscometry, small angle x-ray scattering (SAXS), small angle neutron scattering (SANS), sedimentation, and dynamic and static light scattering) (**Uversky, 1999; Receveur-Brechot *et al.*, 2006**), rate of proteolytic degradation (**Markus, 1965**; **Mikhalyi, 1978**; **Hubbard *et al.*, 1994; Fontana *et al.*, 1997, 2004**), aberrant mobility in SDS-gel electrophoresis (**Iakoucheva *et al.*, 2001; Tompa, 2002**), low conformational stability (**Uversky, 1999**; **Privalov, 1979; Ptitsyn, 1995; Ptitsyn and Uversky, 1994; Uversky and Ptitsyn, 1996**), H/D exchange (**Receveur-Brechot *et al.*, 2006**), immunochemical methods (**Westhof *et al.*, 1984; Berzofsky, 1985**), interaction with molecular chaperones (**Uversky, 1999**), electron microscopy or atomic force microscopy (**Uversky, 1999**; **Receveur-Brechot *et al.*, 2006**), and the charge state analysis of electrospray ionization mass-spectrometry (**Kaltashov and Mohimen, 2005**). But most of these methods cannot directly sample protein dynamics in the nanosecond time scale relevant to conformational changes in such disordered regions and therefore provide only indirect information on the disordered state. Thus in disordered proteins, the structural heterogeneity and rapid inter-conversion among conformers lead to problems in characterization and present practical challenges. Nowadays, MD simulations have been widely used as a powerful tool to

understand the conformational dynamics of proteins at atomic level. This method has been less explored in the field of disordered proteins. It would be worthwhile to investigate the dynamics of disordered proteins using MD simulations.

Scope of my work

In the work described in **Chapter 3**, we have used multiple MD simulations to compare the conformational dynamics originating from two ordered proteins namely, STAT-4 N-domain (1BGF) and Catalytic domain of MutY from *E coli* (1MUN), a partially ordered protein Brak/CXCL 14 (2HDL) and four disordered proteins, Sub domain of staphylococcal nuclease (2SOB), Apoptosis regulator Bcl-X_L (1LXL), F6 subunit of ATP synthase (1VZS) and Tyrosyl-tRNA synthetase (1JH3). All the seven simulations (10 ns each) were performed using ff99SB Amber force field (**Hornak *et al.*, 2006; Wickstrom *et al.*, 2009**). We have analyzed the trajectories arising from these simulations to compare parameters like RMSD, SASA, radius of gyration, conformational entropy between ordered and disordered proteins.

1.4. Molecular crowding and its consequences:

Generally the information regarding the kinetic parameters (k_{cat} and K_m), equilibria, and mechanism of biochemical reactions are obtained through experiments conducted in solutions containing dilute solutions of total protein, nucleic acid, and/or polysaccharides together with buffer salts, low molecular weight substrates, and cofactors as required. In contrast, biochemical reactions in intracellular environments occur in presence of high concentrations of macromolecules such as cytoskeletal filaments and microtubules, various organelles and a variety of other macromolecular species like proteins, nucleic acids etc. This is referred to as macromolecular crowding (**Hall and Minton, 2003**). The media is crowded (as shown in **Figure 1.3.**) rather than concentrated because no single macromolecule is at high concentration, but taken together, the total macromolecular concentration is 50-400mg/ml (**Fulton, 1982**; **Lanni *et al.*, 1985**; **Cayley *et al.*, 1991**), implying that between 5% and 40% of the total volume is physically occupied by these molecules (**Fulton, 1982**; **Gershon *et al.*, 1985**). This total volume occupancy is significant when compared with values for in vitro conditions.

Macromolecular crowding has significant consequences on the thermodynamics of the cell (**Minton, 1993**, **1998**) and strongly affects diffusion processes (**Luby-Phelps *et al.*, 1987**). The thermodynamics of the intracellular environment is liable to get affected by the excluded volume effect arising from a mutual impenetrability of solute molecules and strongly affects substances with large molecular weights (**Zimmerman and Trach, 1991**). The cytoplasm can also undergo phase separation due to macromolecular crowding (**Johansson *et al.*, 2000**; **Brooks, 2000**; **Walter, 2000**). Macromolecular crowding enhances protein association (**Rivas *et al.*, 1999**) and self association of monomers (**Rivas *et al.*, 2001**). It also increases the rate of folding and refolding (**vanden Berg *et al.*, 1999, 2000**). The stability of proteins in crowded conditions is mainly due to decrease in the entropy of denaturation of proteins with no associated change in enthalpy (**Sasahara *et al.*, 2003**). Diffusion of macromolecules in the cytoplasm can be 5-20 times slower than in saline solutions (**Elowitz *et al.*, 1999**; **Verkman, 2002**). Furthermore, many reactions occur on two dimensional membranes or one dimensional channels (**Clegg, 1984**; **Srere *et al.*, 1989**). The complexity of the cytoplasm structure and thermodynamics indicates that we must be careful when considering our description of reaction kinetics in intracellular conditions.

Macromolecular crowding and its influence on both thermodynamics and reaction rates in cellular media have been known since the 1960s by the pioneering investigations of A. G. Ogston and T. C. Laurent providing the initial insight (**Ogston and Phelps, 1960**; **Laurent, 1963**, **1971**; **Laurent and Ogston, 1963**; **Edmond and Ogston, 1968**). The physico-chemical properties of the cytoplasm are considered to be important for biochemists and physiologists in order to understand the dynamics from *in vitro* experimental assays. While the above points have only begun to be valued within the last 15 years, the majority of biochemistry and physiology textbooks are still ignoring them (**Ralston, 1990**). This is not surprising, because in the majority of the literature scientists discuss the physiological or biochemical relevance of their *in vitro* experiments in dilute solutions (**Ellis, 2001a**). In response to this circumstances, a number of reviews on qualitative and quantitative thermodynamic aspects of macromolecular crowding (**Minton, 2000, 2001**; **Ellis, 2001a, b**), its role in the regulation of cell volume (**Al-Habori, 2001**) and the physical properties of the cytoplasm (**Luby-Phelps, 2000**) have

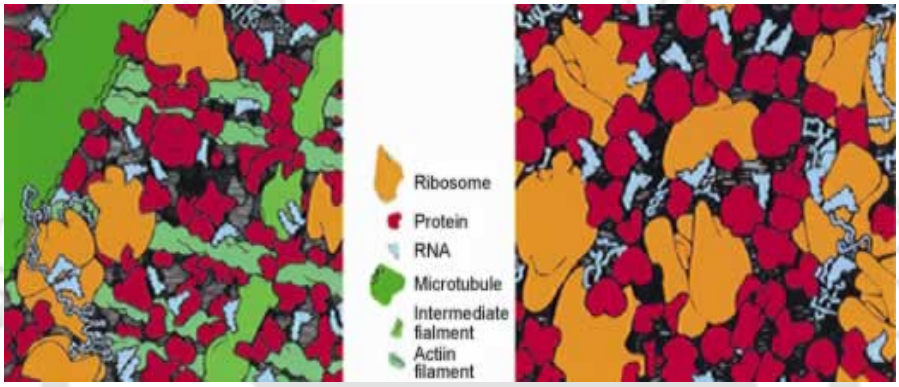

Figure 1.3 : *The crowded state of the cytoplasm in (left) eukaryotic and (right) E. coli cell. Each square illustrates the face of a cube with an edge 100 nm in length. (Picture Courtesy: David S. Goodsell)*

been published over the last few years. However, the literature remains not sufficiently understood and contradictory for the study of biochemical reaction kinetics in intracellular environments. For example, there are investigators formulating the explicit assumption that reactions occur in a region of the cytosol where the law of mass action is valid (see, for example, **Hunding and Kaern, 1998**) whilst others assume they are dealing with large numbers of molecules (**Heinrich and Schuster, 1996**). Moreover assuming the law of mass action, **Minton (1981)** developed a general theory of the effects of volume exclusion on the thermodynamics of globular macromolecules and macromolecular complexes in solution using the hard-particle approximation.

1.4.1. Effects of macromolecular crowding on classical kinetics:

Biochemical reactions are usually carried out under relatively idealized conditions to minimize the effects of non-specific interactions. The experimental practice to mimic *in vivo* environment over the last 40 years has been to add to the solution high concentrations of macromolecules, such as polyethylene glycol, polyvinyl alcohol, Ficolls, dextrans, ovalbumin, serum albumin and hemoglobin. These experiments have helped us to appreciate the effects of macromolecular crowding in cell biochemistry and physiology. A. G. Ogston, T. C. Laurent and A. P. Minton who pioneered the physico-chemical study of proteins in environments with macromolecular crowding developed the thermodynamic principles, based on the law of mass action, to explain the effects observed experimentally in such conditions.

1.4.2. The volume exclusion effect on the equilibrium constants and activity coefficients:

The work of A. G. Ogston and T. C. Laurent mainly focused on the study of the equilibrium constants and the activity coefficients of the components in media with macromolecular crowding (**Ogston, 1970**; **Laurent, 1995**). They observed that the chemical potential of proteins often increases when a macromolecular crowding agent is added to the solution (**Ogston, 1962**; **Laurent and Ogston, 1963**; **Edmond and Ogston, 1968**). This effect has been explained to be the result of mutual exclusion of proteins from the media due to the macromolecular crowding and has been discussed in terms of the activity coefficients (**Ogston, 1958, 1962**; **Ogston and Phelps, 1960**). According to **Minton (1990)**, who has extended the volume exclusion theory of A. G. Ogston, the most

important consequence of crowding on the reaction media in cells is the mutual impenetrability of the reactants causing an excluded volume effect in the reaction media. Reactants are spatially constrained on the microscopic level by force fields, such as steric repulsion and attractive interactions which occur between molecules. These forces can be either specific if they depend on the structure of the interacting molecules or non-specific if they depend on the global properties of the solvent or reaction medium. The availability of reaction volume for a given molecule depends upon the numbers, sizes and shapes of the other molecules present in the reaction compartment (**Figure 1.4**). The effect of steric repulsive forces on the volume available to a given molecule depends on the centre of mass of the molecule and the molecules already present in the solution or "background molecules". If the molecule to be introduced in the reaction is much smaller than the background molecules, the available reaction volume is large, as the small molecules can diffuse between the large molecules. However, if the molecule introduced in the reaction has a similar size to the background molecules, then the available volume is substantially smaller, as the centre of a molecule can approach the centre of another only to the distance at which the surface of the molecules contact each other (**Minton, 2001**).

The thermodynamic effects of volume exclusion on the equilibrium constants were discussed by **Giddings (1970)** to explain certain physico-chemical properties of porous networks and membranes. Consider the reaction

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \qquad \text{----- (1.3)}$$

Applying the law of mass action, the equilibrium constant for the association of C in reaction (1.3) is given by

$$K_{eq} = \frac{\gamma_C [C]}{\gamma_A [A]\, \gamma_B [B]} \qquad \text{----- (1.4)}$$

where the square brackets represent the concentration of the chemical species and γ_i their activity coefficients. The activity is a measure of non-ideal behavior arising from the interaction between the reactants and the macromolecules in the reaction media (**Giddings, 1970**). The activity coefficients are functions of the concentrations of all species present during the reaction. If the reaction occurs in an ideal solution, that is,

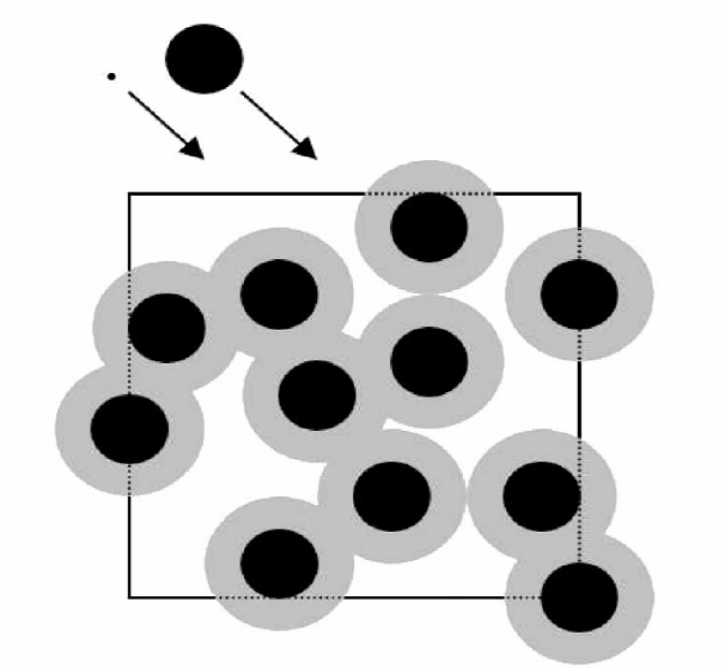

Figure 1.4: *The volume exclusion effect. A small molecule is free to occupy the entire volume in the box between the black obstacles (i.e. the white and gray areas). A molecule of comparable size to the obstacles is much more limited in the volume it can occupy, the centre of such a molecule is limited to the white areas only within the enclosed box.*

the reactants do not interact with reaction media and background molecules, the activity coefficients are equal to 1. In crowded environments, the activity coefficients will vary depending upon the sizes of molecules A, B and C relative to the size of the crowding agents (**Minton, 1981**). To facilitate the experimental determination of the effects of macromolecular crowding, Minton (1981) defined the apparent equilibrium constant

$$\widetilde{K_{eq}} = \Gamma\ K_{eq} \qquad \text{----- (1.5)}$$

where K_{eq} is the equilibrium constant measured in an ideal solution and $\Gamma = (\gamma_C / \gamma_A\ \gamma_B)$ is a non-ideal correction factor, which takes constant values. The value of Γ calculated using the hard spherical particle approximation increases as the sizes of A and B increase relative to the macromolecular agent.

Using the thermodynamic theory developed by **Giddings (1970)**, **Laurent (1971)** was the first biochemist to study the effects of macromolecular crowding in enzyme catalyzed reactions. Considering the Michaelis-Menten reaction (**Michaelis and Menten, 1913**) as a protype, that is the reversible reaction between an enzyme E and a

substrate S; giving the enzyme-substrate complex C; which irreversibly yields product P, we have

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_2} E + P \qquad \text{----- (1.6)}$$

Where k_1, k_{-1} and k_2 are rate constants. Under [S] > > [E], the behavior of this enzymatic reaction can be studied by measuring the initial rate of product formation v_0 and applying the Michaelis-Menten equation (**Boyde, 1980**)

$$v_0 = \frac{V_{max}\,[S_0]}{[S_0] + K_m} \qquad \text{----- (1.7)}$$

Where V_{max} is the maximum velocity and $K_m = (k_{-1} + k_2) / k_1$, is the Michaelis-Menten constant. In order to understand the effects of macromolecular crowding, **Laurent (1971)** considers the Michaelis-Menten constant and the equilibrium dissociation constant of enzyme-substrate complex to be equivalent. He thus rewrites the Michaelis-Menten equation as

$$v_0 = \frac{V_{max}\,[S_0]}{[S_0] + \widetilde{K_m}} \qquad \text{----- (1.8)}$$

Where $\widetilde{K_m} = \Gamma\, K_m$ is the apparent Michaelis-Menten constant, K_m is the Michaelis-Menten constant in an ideal solution and $\Gamma = \gamma_C / (\gamma_S\, \gamma_E)$ is the non-ideal correction factor. Note that the conditions in which the Michaelis-Menten constant equals the equilibrium dissociation constant of the enzyme-substrate complex have established by **Schnell and Maini (2000)**.

The most significant discovery by **Laurent (1971)** is that the apparent value of the Michaelis-Menten constant decreases in the presence of a macromolecular agent in the single enzyme-substrate reaction and in other enzymatic reactions such as two-substrate reactions and competitive inhibitions. **Laurent *et al.* (1974)** also studied the stability of the DNA double helix in the presence of crowding agents and found that macromolecular crowding raised the melting temperature of the DNA dramatically. This result together with observations of the packed nature of DNA in *in vivo* conditions led to the study of macromolecular crowding in several enzymes and proteins involving nucleic acids. **Fuller *et al.* (1981)** and **Zimmerman and Harrison (1987)** found that the excluded volume effect increases the binding and activity of polymerase to DNA. Similar results have also been noticed for DNA ligases (**Zimmerman and Pheiffer,**

1983). **Zimmerman and Minton (1993)**, **Minton (2001)** and **Hall and Minton (2003)** have tabulated more than 30 biochemical systems in which protein-protein, protein-nucleic acid and nucleic acid-nucleic acid interactions increase their activity in solutions when macromolecular agents are added. However, the majority of the data on these systems is focused on the enzyme activity and not on the reaction rates. **Zimmerman (1993)** has also discussed macromolecular crowding effects on genome structure and function. It is significant to note that macromolecular crowding can also decrease enzyme activity, depending on the reaction kinetics and the state of the enzyme **(Minton, 1990)**. For example, the activity of d-glyceraldehyde 3-phosphate dehydrogenase decreases as macromolecular crowding increases. This is due to the increasing association of the active monomeric form of the enzyme into its inactive tetrameric form (**Ovadi *et al.*, 1979**; **Minton, 1983**). In order to explain these experimental observations, **Minton (1981)** developed a general theory describing the effects of volume exclusion on the thermodynamics of globular macromolecules and macromolecular complexes in solution. According to the theory developed by Minton (**Minton, 1981**):

(1) The conformation of the enzyme may be altered, if the enzyme exists in two or more conformations with different enzymatic activities. The presence of macromolecules shifts the equilibrium between the different conformations, favouring the more compact conformation and thereby alters the catalytic activity of the enzyme.

(2) The enzymes might self-associate in volume occupied media, leading to a change in catalytic activity. This change will depend upon the extent of self-association (Minton, 1981).

(3) If the reaction is encounter rate controlled, the presence of macromolecules will lead to a decrease in the rate of the reaction.

This formalism is an extension of the theory of the thermodynamic state of macromolecules reacting in membranes and other porous media (**Giddings, 1970**), which is itself based on the law of mass action whereby the rate of reaction is proportional to the product of concentrations of the reactants. By applying the law of mass action to the reaction (1.6), the governing equations for the species are,

$d[S]/dt = -k_1[S][E] + k_{-1}[C]$ ----- (1.9)

$d[E]/dt = -k_1[S][E] + (k_{-1} + k_2)[C]$ ----- (1.10)

$d[C]/dt = k_1[S][E] - (k_{-1} + k_2)[C]$ ----- (1.11)

$d[P]/dt = k_2[C]$ ----- (1.12)

with initial conditions at t = 0

$([S], [E], [C], [P]) = ([S_0], [E_0], 0, 0)$, ----- (1.13)

where the subscript 0 implies initial concentration. In this system, following Minton (1981), the rate constants k_i are written as products of the ideal constants k_i^0 and correction factors Γ_i

$k_1 = k_1^0 \Gamma_1$, ----- (1.14)

$k_{-1} = k_{-1}^0 \Gamma_{-1}$, ----- (1.15)

$k_2 = k_2^0 \Gamma_2$ ----- (1.16)

According to Minton's theory, if the enzymic reaction occurs in an ideal and homogenous environment, $\Gamma_i = 1$ and no correction factor is required. However, if S is a macromolecule and the reaction occurs in a crowded environment, k_i will increase with the volume exclusion. One of the important issues with this formalism for enzyme action is that it is based on the law of mass action or transition-state theory of reaction rates. By considering that the reaction rate of S and E is proportional to the sum of their diffusion coefficients, this law relies on the classical chemical kinetics assumption that reactants are well diluted and perfectly mixed, and reactions occur in a three-dimensional environments with homogenous medium (**Minton, 1981**). These conditions are difficult to meet in macromolecular crowding, and as such the adequacy of the law of mass action has been questioned for describing intracellular reactions (**Clegg, 1984**; **Halling, 1989**; **Kuthan, 2001**).

A number of authors (see, for example, **Gillespie, 1992**; **Kepler and Elston, 2001**) consider that stochastic modeling as more realistic alternative for modeling intracellular reactions because it studies the individual behavior of small molecules.

Another approach is the power-law approximation, a phenomenological approach developed by **Savageau (1969, 1976)** to describe reactions following non-ideal kinetics in crowded environments (**Savageau, 1992**). On the other hand, **Kopelman (1986)** has discovered that chemical reactions occurring in heterogeneous media, where reactants are spatially constrained on the microscopic level, follow a fractal-like kinetics. However, this fractal-like kinetics has been rarely applied to the study of biochemical reactions.

However, though a number of theoretical approaches exist for describing the effects of crowding on the kinetics of biochemical reactions, few studies have explored the consequences of crowding on enzyme catalysis *in vitro*. The effect of different concentrations of Ficoll 70 on the rate of EcoRV-catalyzed cleavage of pBR 322 was studied by **Wenner and Bloomfield (1999)**. They observed that Ficoll 70 had little effect on the overall reaction velocity of EcoRV in the concentration range 0-20% g/dL owing to offsetting increases in V_{max}, K_m, and stronger non-specific binding between enzyme and substrate/product. The effect of PEG 6K on enzyme activity of *Escherichia coli* AspP was investigated recently (**Moran-Zorzano *et al.*, 2007**). They reported that 50 g/L PEG decrease K_m fourfold and increases V_{max} six fold. But most of the studies used crowding agents of fixed size only. But it is known that the macromolecules present inside the cell, exist in various sizes and shapes. Thus, it is desirable to investigate the effect of different sizes of crowding agents on enzyme activity.

Scope of my work

In the **Chapter 4**, we describe investigations on the effect of macromolecular crowding on the kinetics of a) alkaline phosphatase (AP) catalyzed hydrolysis of p-nitro phenyl phosphate (PNPP) and b) acetyl cholinesterase (AchE) catalyzed hydrolysis of 2-naphthyl acetate (NA) and 3-indoxyl acetate (IA). The crowding agent, dextran, of different sizes (15-20, 40, 70, 200, 500 & 2000 kDa) were added to mimic the cellular conditions where numerous such enzyme catalyzed reactions take place. In addition, we also studied the effect of different substrates on enzyme activity under the crowded milieu. The products of above mentioned enzyme catalyzed hydrolysis reactions being colored; the reaction kinetics could be monitored conveniently using UV-visible spectroscopy. We monitored the change in enzyme activity during hydrolysis using multiple concentrations and different sizes of crowding species.

1.5. Protein Aggregation and its significance:

An important outcome of macromolecular crowding is protein aggregation. The question of the overall effect of macromolecular crowding on protein aggregation and fibrillation is not as simple as it appears. There are evidences to show enhancement of the undesirable aggregation of partially unfolded proteins (**Kinjo and Takada, 2003**) by macromolecular crowding when the intrinsic folding rate of the protein is relatively slow. Irreversible unfolding due to aggregation of unfolded states in the presence of crowding agents has been observed for dihydrofolate reductase, enolase and green fluorescent protein (**Martin, 2002**). Overall, these studies infer that the macromolecular crowding might dramatically enhance the competition between protein folding and aggregation, favoring the latter when folding is relatively slow. Currently, protein aggregation represents an important problem in biomedicine and biotechnology.

Protein aggregation is essentially a self-association process in which many identical protein molecules form higher order conglomerates of low solubility that eventually precipitate. On the basis of their macroscopic morphology, they are generally classified as either ordered or disordered aggregates (**Dobson, 2004**; **Lopez de la Paz and Serrano, 2004**).

The activities inside the cell are generated and abolished by means of folding and unfolding of proteins. Some times the unfolding is also the precursor to the degradation of proteins (**Matouschek, 2001**). It is evident that, some events in the cell, such as translocation across membranes, require proteins to be in unfolded or partially folded states. Other miscellaneous biological processes such as trafficking, secretion, immune response and the regulation of the cell cycle, are in fact known to be directly dependent on folding and unfolding of proteins (**Radford and Dobson, 1999**). It is therefore quite obvious that failure to fold correctly, or to remain correctly folded, will give rise to the malfunctioning of living systems and therefore lead to disease. In fact it is becoming clear that a wide range of human diseases are associated with aberrations in the folding process (see **Table 1.1**) (**Thomas *et al.*, 1995**; **Dobson, 2001**).

Table 1.1: *Representative protein folding diseases*

Disease	Protein	Site of folding
Hypercholesterolaemia	Low-density lipoprotein receptor	ER
Cystic fibrosis	Cystic fibrosis trans-membrane regulator	ER
Phenylketonuria	Phenylalanine hydroxylase	Cytosol
Huntington's disease	Huntingtin	Cytosol
Marfan syndrome	Fibrillin	ER
Osteogenesis imperfecta	Procollagen	ER
Sickel cell anaemia	Haemoglobin	Cytosol
α1-Antitrypsin deficiency	α1-Antitrypsin	ER
Tay-Sachs disease	β-Hexosaminidase	ER
Scurvy	Collagen	ER
Alzheimer's disease	Amyloid β-peptide/tau	ER
Parkinson's disease	α-Synuclein	Cytosol
Scrapie/Creutzfeldt-Jakob disease	Prion disease	ER
Familial amyloidoses	Transthyretin/lysozyme	ER
Retinitis pigmentosa	Rhodopsin	ER
Cataracts	Crystallins	Cytosol
Cancer	p53	Cytosol

It has been observed that some of the diseases for example, cystic fibrosis result from the simple reason that if proteins do not fold correctly they will not be able to exercise their proper functions. In other cases, the misfolded proteins form intractable aggregates within cells or in the extracellular space. An increasing number of pathologies, including Alzheimer's and Parkinson's diseases, are identified to be directly associated with the deposition of such aggregates in tissue (**Thomas *et al.*, 1995; Dobson, 2001; Tan and Pepys, 1994**).

1.5.1. Aggregation and Amyloid formation:

Deposits of proteins in the form of amyloid fibrils and plagues in brain, in vital organs such as the liver and spleen, or in skeletal tissue, depending on the type of disease (**Thomas *et al.*, 1995; Dobson, 2001; Tan and Pepys, 1994**) is one of the most characteristic features of many of the aggregation diseases. In the case of neurodegenerative diseases, the amount of aggregates deposited is very less and some times can be almost untraceable, while in systemic diseases, kilograms of protein deposits can be noticed. Every type of amyloid disease involves the aggregation of a definite protein although a range of other components, including other proteins and

carbohydrates, is also incorporated into the deposits when they form *in vivo*. The characteristics of the soluble forms of the 20 or so human proteins, involved in the formation of well defined amyloids are varied – they range from intact globular proteins to largely unstructured peptide molecules-but the aggregated forms have many similar characteristics (**Sunde and Blake, 1997**). All the amyloid deposits on binding to certain dye molecules, notably Congo red shows unique optical properties (such as birefringence). These properties have been used over a century for the diagnosis. The fibrillar structures of many aggregates that are considered to be one of the characteristic feature, have very related morphologies (long, unbranched and often twisted structures a few nm in diameter) and a characteristic "cross-beta" X-ray fibre diffraction pattern (**Sunde and Blake, 1997**). The latter reveals that the organized core structure is composed of β-sheets having strands running perpendicular to the fibril axis. Earlier it was assumed that the ability to form amyloid fibrils with the above characteristics was limited to a relatively few number of proteins, mainly those seen in disease states, and that these proteins possess unique sequence motifs encoding the amyloid core structure. But from the recent studies, it has been suggested that the ability of polypeptide chains to form such structures is common and in fact can be considered a generic feature of polypeptide chains (**Chiti *et al.*, 1999**; **Dobson, 1999**). The latter statement is evident from the fact that fibrils can be formed by many different proteins that are not associated with disease, this includes the well known proteins myoglobin (**Fandrich *et al.*, 2001**), homopolymers such as polythreonine or polylysine (**Fandrich and Dobson, 2002**). Thus one can regard amyloid fibrils as highly organized structures. The characteristic features of such structures can be known from the physico-chemical properties of the polymer chain. Information gathered from the fibrils formed from disease associated proteins and also from the other proteins has enabled many of the features of these structures to be defined (**Petkova *et al.*, 2002**), although no complete structure has yet been determined in atomic detail. It is apparent that the interactions, particularly hydrogen bonds, involving the polypeptide main chain essentially stabilizes the core structure of the fibrils. As the main chain is common to all peptides, this observation explains why fibrils formed from polypeptides of very different amino acid sequence are similar in appearance. The side chains are likely to be incorporated in whatever manner is most favorable for a given

sequence within the amyloid structures; they affect the details of the fibrillar assembly but not their general structure (**Chamberlain *et al.*, 2000**). The tendency to form amyloid fibrils appears to be generic but the propensity to do so vary dramatically between different sequences (**Du Bay, 2004**). It has also been observed that there is a considerable change (by an order of magnitude or more) in the rate of aggregation of unfolded polypeptide by single changes of amino acid in protein sequences. This in turn made possible to correlate the changes in aggregation rates caused by such mutations with changes in simple properties that arise from such substitutions, such as charge, secondary structure propensities and hydrophobicity (**Chiti *et al.*, 2003**). As this correlation has been found to hold for a wide range of different sequences, it strongly endorses the concept of the generic nature of amyloid formation.

The formation of fibrils by certain proteins that display a globular structure in native state, such as lysozyme or transthyretin and β2-microglobulin, still remains a challenging issue in protein biochemistry. The common feature of all these proteins is their relative structural instability and the capability to change conformation adopting the common beta sheet fibrillar structure (**Merlini and Bellotti, 2003**). Several studies are reported to support the conformational change hypothesis (**Lai *et al.*, 1996; Wetzel, 1996; Funahashi *et al.*, 1996**) and are now believed to be the main cause leading to fibril formation in conditions of protein denaturation and partial unfolding (**Forloni *et al.*, 1993**).

The discovery that lysozyme can cause systemic amyloidoses offered a unique opportunity to explore in details the relationship between structure and folding in amyloid proteins exploiting the available wealth of information on structure and folding of lysozyme. **Pepys *et al.*, (1993)** identified that single point mutations in human lysozyme gene associated with hereditary systemic amyloidosis. This has led to extensive study on lysozyme amyloid fibrils and there are reports that have shown several proteins in lysozyme family to be capable of amyloid fibril formation *in vitro* including several lysozymes (human, hen, turkey, equine) and two α-lactalbumin (bovine, human) proteins (**Trexler and Nilsson, 2008**). All of these proteins share significant sequence identity and are structurally similar (**Boeckmann *et al.*, 2003**; **Huang and Miller, 1991**). There are two known natural mutations of the human lysozyme: D67H and I56T (**Pepys *et al.*,**

1993). They are shown to cause autosomal dominant hereditary nonneuropathic systemic amyloidosis (this is a condition whereby there is amyloid deposition in the viscera and other body cavities). Amyloids are formed because of diverse sequence, fold and function. The mechanism of fibrilogenesis is not clear but appears related to changes in stability and tendency to aggregate due to mutations. **Booth *et al.* (1997)** proposed that partly folded forms of amyloidogenic proteins which lack global co-operativity undergoes a helix to sheet transition and form the initial seed for the generation of amyloids. **Morozova-Roche *et al.* (2000)** showed the existence of template seed for the wild type and the two natural mutants of human lysozyme that assist the formation of fibrils. The structural details of species formed early during the aggregation process and their features which trigger amyloidosis seems to be important. So it will be worthwhile to investigate the structural feature of the partially folded structures of wild type and the mutants of human lysozyme that may trigger amyloid formation. MD simulations can be employed to understand the conformational dynamics of the proteins at atomistic level.

Scope of my work

In **Chapter 5**, we have compared the conformational dynamics of wild type and mutants of Human lysozyme that are reported to cause amyloidosis using MD simulations. The crystal structures are available for wild type and three of the human lysozyme mutants: D67H, I56T, and T70N. The mutants D67H and I56T cause amyloidosis independently whereas T70N results in amyloidosis if present as double mutant T70N/W112R (**Rocken *et al.*, 2006**). T70N is the only known naturally occurring destabilized mutant of human lysozyme that has not been observed in amyloid deposits in human patients. But for understanding the determinants of amyloid disease, it is important to study and compare the properties of T70N mutant with those of amyloidogenic mutants of human lysozyme. In general these structures are very much similar to the wild type (WT) protein. All the four simulations (20 ns each) were performed using ff99SB Amber force field to investigate the conformational dynamics. We have analyzed the trajectories arising from these simulations to obtain insights on conformational features triggering amyloidosis.

1.6. Techniques:

To monitor the structure, function and dynamics of proteins in disordered states and non-ideal conditions, we have used two important techniques, Molecular dynamics simulation and UV-visible spectroscopy. The principles and theory involved in these techniques are discussed here.

1.6.1. Molecular Dynamics Simulation:

1.6.1.1. Why Molecular Dynamics?

Molecular dynamics (MD) simulations represent the in silico approach to statistical mechanics. As a counterpart to experiment, MD simulations are used to estimate equilibrium and dynamic properties of complex systems that cannot be calculated analytically. Representing the exciting interface between theory and experiment, MD simulations occupy a venerable position at the crossroads of mathematics, biology, chemistry, physics, and computer science. The static view of a biomolecule, as obtained from X-ray crystallography, for example - while extremely valuable - is still insufficient for understanding a wide range of biological activity. It only provides an average, frozen view of a complex system. Certainly, molecules are live entities, with their constituent atoms continuously interacting among themselves and with their environment. Their dynamic motions can explain the wide range of thermally-accessible states of a system and thereby connect structure and function. By following the dynamics of a molecular system in space and time, we can obtain a rich amount of information concerning structural and dynamic properties. Such information includes molecular geometries and energies; mean atomic fluctuations; local fluctuations (like formation/breakage of hydrogen bonds, water/solute/ion interaction patterns, or ring flips, nucleic acid sugar repuckering, diffusion), enzyme/substrate binding; free energies; and the nature of various types of concerted motions. Ultimately perhaps, large scale deformations of macromolecules such as protein folding might be simulated. This formidable aspect, however, is more likely to be an outgrowth of hand-in-hand advances in both experiment and theory. Though the MD approach remains popular because of its essential simplicity and physical appeal, it complements many other computational tools for exploring molecular structures and properties, such as Monte Carlo simulations, Poisson-Boltzmann analyses, energy minimization and Brownian dynamics.

1.6.1.2. Background of Molecular Dynamics Simulations:

The molecular dynamics simulation method was first introduced by Alder and Wainwright in the late 1950's (**Alder and Wainwright, 1957, 1959**) to study the interactions of hard spheres. Many important insights regarding the behavior of simple liquids emerged from their studies. The next major advance was in 1964, when Rahman carried out the first simulation using a realistic potential for liquid argon (**Rahman, 1964**). The first molecular dynamics simulation of a realistic system was done by Rahman and Stillinger in their simulation of liquid water in 1974 (**Stillinger and Rahman, 1974**). The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) (**McCammon, *et al*, 1977**). Today in the literature, one regularly finds molecular dynamics simulations of solvated proteins, protein-DNA complexes as well as lipid systems addressing a variety of issues including the thermodynamics of ligand binding and the folding of small proteins. The number of simulation techniques has greatly expanded; there exist now many specialized techniques for particular problems, including mixed quantum mechanical - classical simulations that are being employed to study enzymatic reactions in the context of the full protein. Molecular dynamics simulation techniques are widely used in experimental procedures such as X-ray crystallography and NMR structure determination.

1.6.1.3. Theory of Molecular Dynamics Simulations:

In molecular dynamics, successive configurations of the system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time. From this trajectory, the average values of properties can be determined. The method is deterministic; once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past. The trajectory is obtained by solving the differential equations embodied in Newton's second law

$$F_i = m_i a_i \qquad \text{----- (1.17)}$$

Where F_i is the force exerted on particle I, m_i is the mass of particle i and a_i is the acceleration of particle i. The force can also be expressed as the gradient of the potential energy,

$$F_i = -\nabla i\ V \qquad \text{----- (1.18)}$$

Combining these two equations yields

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \qquad \text{----- (1.19)}$$

Where V is the potential energy of the system. Newton's equation of motion can then relate the derivative of the potential energy to the changes in position as a function of time.

Molecular dynamics simulations can be time consuming and computationally expensive. However, computers are getting faster and cheaper. Simulations of solvated proteins are calculated up to the nanosecond time scale; however, simulations into the millisecond regime have been reported.

In this way an MD simulation generates a trajectory that describes how the system moves through phase space as a function of time. Several computationally efficient algorithms exist for integrating the equations of motion. All of them use Taylor series expansions of the positions and dynamic properties:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \frac{1}{6}b(t)\delta t^3 + \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.20)$$

where v is the velocity (the first derivative of the positions with respect to the time), a is the acceleration (the second derivative), b is the third derivative, and so on. One of the most commonly used algorithms is the Verlet algorithm (**Verlet, 1967**). It uses the positions and accelerations at time t, and the positions from the previous step to calculate the new positions at time $t+\delta t$.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.21)$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 - \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.22)$$

Adding these two equations gives:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.23)$$

The Verlet algorithm uses no explicit velocities. The advantages of the Verlet algorithm are *i)* it is straightforward, and *ii)* the storage requirements are modest. The disadvantage is that the algorithm is of moderate precision.

A variation on the Verlet algorithm is the leap-frog algorithm (**Hockney, 1970**). In this algorithm, the velocities are first calculated at time $t+\frac{1}{2}\delta t$; these are used to calculate the positions, *r*, at time $t+\delta t$. In this way, the velocities *leap* over the positions, then the positions *leap* over the velocities. The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions. It uses the following relationships:

$$r(t+\delta t) = r(t) + v(t+\frac{1}{2}\delta t)\delta t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.24)$$

$$v(t+\frac{1}{2}\delta t) = v(t-\frac{1}{2}\delta t) + a(t)\delta t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.25)$$

The velocities at time *t* can be approximated by the relationship:

$$v(t) = \frac{1}{2}\left[v(t+\frac{1}{2}\delta t) + v(t-\frac{1}{2}\delta t) \right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.26)$$

1.6.1.4. Potential Energy function:

MD simulations start with a knowledge of the energy of the system as a function of the atomic coordinates. The potential energy surface establishes the relative stabilities of the different possible stable or metastable structures. The forces acting on the atoms of the system, which are related to the first derivatives of the potential with respect to the atom positions, can be used to compute the dynamic behavior of the system by solving Newton's equations of motion for the atoms as a function of time as shown in equation.

Although quantum mechanical calculations can yield potential surfaces for small molecules, only empirical energy functions can make available this information for proteins and their environment. As at ordinary temperatures most motions leave the bond lengths and angles of polypeptide chains near their equilibrium values, the exactness of the energy function representation of bonding is comparable to that of vibrational analyses of small molecules. In globular proteins, contacts between non-bonded atoms also make a large contribution to the potential energy of the folded or native structure. The success of hard sphere non-bonded radii in conformational studies suggests that relatively simple functions may adequately describe them. The energy functions used for

proteins are generally composed of bonding terms representing bond lengths, bond angles, and torsional angles, and non-bonding terms consisting of van der Waals interaction and electrostatic contributions. One widely used expression (**Brooks *et al.*, 1983**) is

$$E(R) = \sum_{bonds} K_l (l - l_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \frac{1}{2} \sum_{torsional} K_\phi (1 + \cos[n\phi - \phi_0]) + \sum_{i<j}^{atoms} \left(\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr} \right)$$

..... (1.27)

The energy, E, is a function of the Cartesian coordinate set, *R*, specifying the positions of all the atoms, from which are calculated the internal coordinates for bond lengths (l), bond angles (θ), dihedral angles (ϕ) and interparticle distances (r). The first term in equation (1.27) represents instantaneous displacements from the equilibrium bond length, l_{eq}, by a Hooke's law (harmonic) potential. This is shown in **Figure 1.5**.

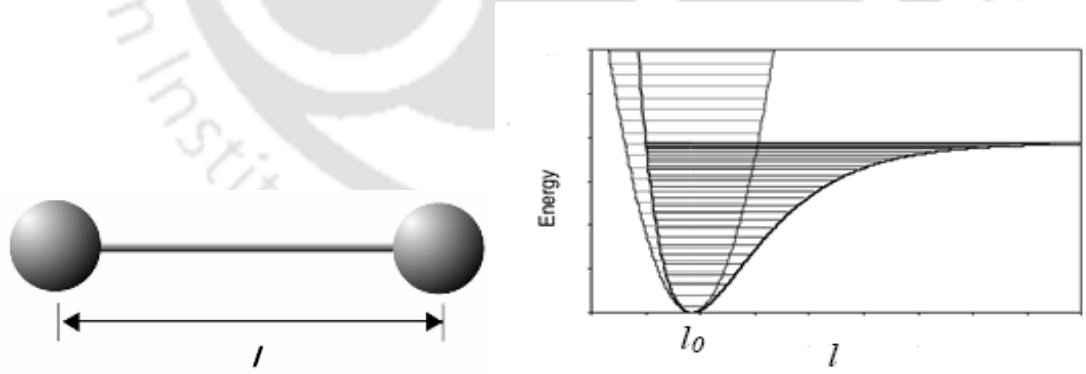

Figure 1.5: *(a) Bonds between atoms separated by a distance "l". (b) Near to the equilibrium value "l_0 "the harmonic potential is a good estimate for the more accurate Morse curve.*

Such a harmonic potential is the first approximation to the energy of a bond as a function of its length. The bond force constant K_l determines the flexibility of the bond and can

be evaluated from infrared stretching frequencies or quantum mechanical calculations. Equilibrium bond lengths can be inferred from high resolution, low temperature crystal structures or microwave spectroscopy data. The energy connected with alteration of bond angles given by the second term in equation (1.27) is also represented by a harmonic potential. The contribution of each angle is characterized by a force constant K_θ and an equilibrium value θ_0. Vibrational motions involving angle bending normally occur at lower frequencies than those of typical bond vibrations, less energy is required to distort an angle from its equilibrium value than to stretch a bond. This fact is reflected in the smaller force constants used for angle terms compared to those of bond terms in most force field implementations. The definition of angle bending terms is illustrated in **Figure 1.6.**

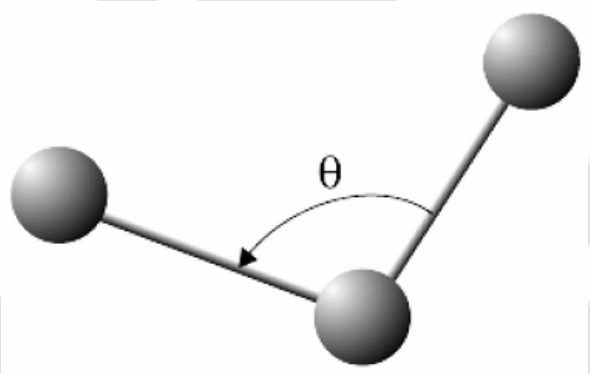

Figure 1.6: *Angle bending terms are three atom terms characterized by a harmonic potential dependent on the angle between the three atoms.*

For rotations about bonds, torsion angle potential functions given by the third term in equation (1.27) are used. This potential is assumed to be periodic and modeled by a cosine or sum over cosine functions. K_ϕ represents the barrier height, ϕ is the torsion angle between the 1,4-pair, ϕ_0 is an offset which defines the angular position of the first minimum in the potential, and n the multiplicity which gives the number of minima in the function as the bond is rotated through 360°. In most force fields two or more torsion terms may be assigned to the same 1, 4-pair with different values of K, n, and ϕ, in the energy evaluation these terms are summed together. This allows the reproduction of very complex shapes for the rotational energy barrier. An example of this is illustrated in **Figure 1.7** along with the definition of the torsion angle ϕ.

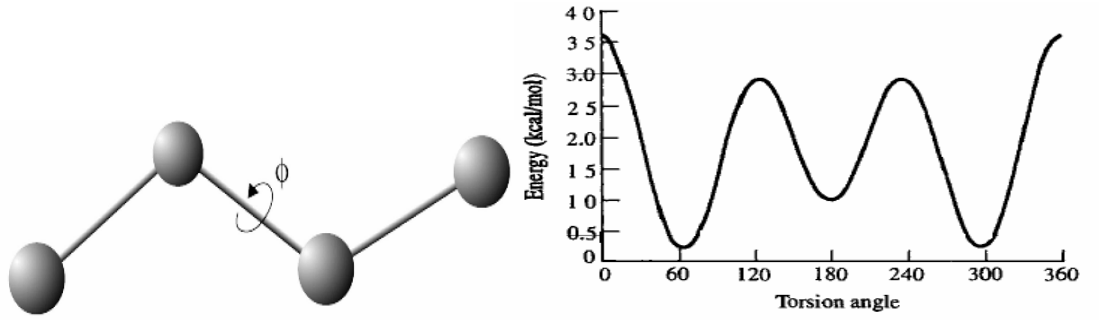

Figure 1.7: *(a) Definition of the torsion angle ϕ. (b) Energy curve for the torsional terms.*

The final term in equation (1.27) represents the contribution of non-bonded interactions and has three parts: a repulsive term preventing atoms from interpenetrating at very short distances; an attractive term accounting for the London dispersion forces between atoms; and an electrostatic term that is attractive or repulsive depending on whether the charges q_1 and q_2 are of opposite or the same sign. The first two non-bonded terms combine to give the familiar Lennard-Jones 6-12 potential, which has a minimum at an interatomic separation equal to the sum of the van der Waals radii of the atoms; parameters A and B depend on the atoms involved and have been determined by a variety of methods, including non-bonding distances in crystals and gas-phase scattering measurements.

Electrostatic interactions between pairs of atoms are represented by a Coulomb potential with D the effective dielectric function for the medium and r the distance between the two charges. Use of atomic partial charges avoids the need for a separate term to represent the hydrogen bond interaction; that is, when the positive hydrogen attached to an electronegative atom comes within van der Waals distance of a negative acceptor atom, the Coulomb attraction adds to the Lennard-Jones potential and results in a hydrogen bond.

The application of empirical energy function depends on the extent to which the parameters determined for equation (1.27) by the study of model systems, such as amino acids, can be employed for macromolecules, such as proteins. Evidence from a number of comparisons suggests that this transferability condition is satisfied in many applications.

The standard simulation package AMBER 8 (**Case *et al.* 2002 & 2004**) is used in the present work. *Sander*, one of the AMBER modules carries out the energy

minimization, molecular dynamics and NMR refinements. It provides direct support for several force fields (non-polarizable and polarizable) for proteins and nucleic acids, and for several water models and other organic solvents. The basic force field implemented here has the form shown in equation (1.27).

1.6.1.5. Treatment of Solvent in MD simulations:

Solvent, generally water, has a primary influence on the structure, dynamics and thermodynamics of biological molecules, both locally and globally. One of the most important effects of the solvent is the screening of electrostatic interactions. The electrostatic interaction between two charges is given by Coulomb's law,

$$V_{elec} = \frac{q_i q_j}{\varepsilon_{eff} r_{ij}} \qquad \text{-----(1.28)}$$

where q_i, q_j are the partial atomic charges, ε_{eff} is the effective dielectric constant and r_{ij} is the relative distance between the two particles. It is important to include solvent effects in an MD simulation. This can be done at several levels. The simplest treatment is to simply include a dielectric screening constant in the electrostatic term of the potential energy function. In this **implicit** treatment of the solvent, water molecules are not included in the simulation but an effective dielectric constant is used. Often the effective dielectric constant is taken to be distance dependent, $e_{eff} = r_{ij}\ \varepsilon)$, where ε ranges from 4 to 20. Although this is a crude approximation, it is still much better than using unscreened partial charges. Other implicit solvent models have been developed that range from the relatively simple distance-dependent dielectric constants to models that base the screening on the solvent exposed surface area of the protein. The distance-dependent dielectric coefficient is the simplest way to include solvent screening without including explicit water molecules and it is available in most simulation programs. Recently, several implicit solvent models based on continuum electrostatic theory have been developed. If water molecules are **explicitly** included in the simulation, then ε = 1 in the energy function; the explicit water molecules provide the electrostatic shielding. In this more detailed treatment of the solvent, boundary conditions must be imposed, first, to prevent the water molecules from diffusing away from the protein during the simulation, and second to allow simulation and calculation of macroscopic properties using a limited

number of solvent molecules. Several different treatments of the boundary exist, the use of one over another depends strongly on the type of problem the simulation is to address.

1.6.1.6. Water Models:

For the simulation of water clusters, liquid water and aqueous solutions with explicit solvent, water models are used. These models use the approximations of molecular mechanics. Many different models have been proposed; they can be classified by the number of points used to define the model, whether the structure is rigid or flexible, and whether the model includes polarization effects. An alternative to the explicit water models is to use an implicit solvation model, also known as a continuum model.

The simplest water models treat the water molecule as rigid and rely only on non-bonded interactions. The electrostatic interaction is modeled using Coulomb's law and the dispersion and repulsive forces using the Lennard-Jones potential. The potential for models such as TIP3P and TIP4P is represented by

$$E_{ab} = \sum_{i}^{on\ a} \sum_{j}^{on\ b} \frac{K_c q_i q_j}{r_{ij}} + \frac{A}{r_{oo}^{12}} - \frac{B}{r_{oo}^{6}} \quad \text{------(1.29)}$$

where k_C is the electrostatic constant, q_i and q_j are the partial charges relative to the charge of the electron; r_{ij} is the distance between two atoms or charged sites; and A and B are the Lennard-Jones parameters. The charged sites may be on the atoms or on dummy sites (such as lone pairs). In most water models, the Lennard-Jones term applies only to the interaction between the oxygen atoms. The TIP3P model has three interaction sites which correspond to the three atoms of the water molecule. A point charge is assigned to each atom; also the oxygen atom gets the Lennard-Jones parameters. This model is widely used for molecular dynamics simulations because of its simplicity, reasonable structural and thermodynamic descriptions and computational efficiency. The model is fit to explain the properties of liquid water at other thermodynamic points, its complicated phase diagram, ionic, and other aqueous solutions, and confined and biological water. The model is advanced over the two-site models in both static and dynamic properties. The H-O-H angle for the water molecule is of 104.5° and the O-H distance is 0.957 Å. The simple models for TIP3P water is shown in **Figure 1.8.**

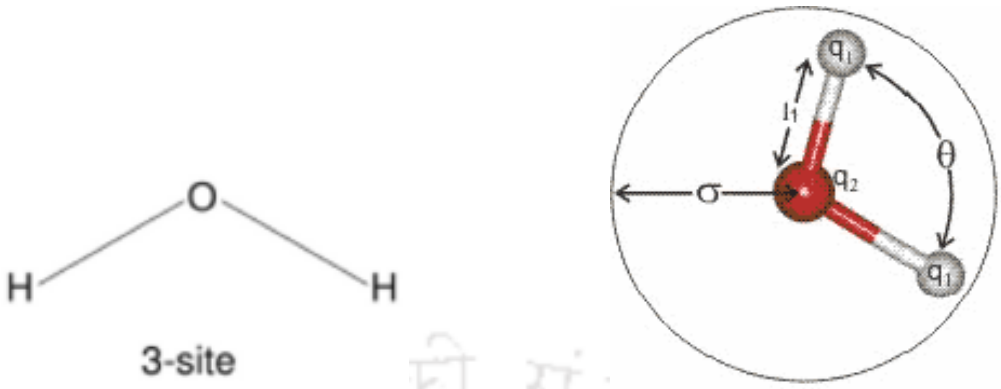

Figure 1.8: *Schematic representation of the TIP3P water models.*

1.6.1.7. Periodic Boundary Conditions:

Periodic boundary conditions make possible a simulation to be performed using a relatively small number of particles in such a way that the particles experience forces as though they were in a bulk solution. See, for example, the two dimensional box as shown in **Figure 1.9.** The central box is surrounded by eight neighbors. The coordinates of the image particles, those found in the surrounding box are related to those in the primary box by simple translations. The simplest box is the cubic box. Forces on the primary particles are calculated from particles within the same box as well as in the image box. The cutoff is chosen such that a particle in the primary box does not see its image in the surrounding boxes.

1.6.1.8. Particle Mesh Ewald:

The PME is, as the name suggests, a modified form of Ewald summation that is inspired by and closely related to the original Hockney-Eastwood PPPM method (**Hockeny & Eastwood, 1981**). Ewald summation is a method to efficiently calculate the infinite range Coulomb interaction under periodic boundary conditions (PBC), and PME is a modification to accelerate the Ewald reciprocal sum to near linear scaling, using the three dimensional fast Fourier transform (3DFFT). Because the Coulomb interaction has infinite range, under PBC particle i within the unit cell interacts electrostatically with all other particles j within the cell, as well as with all the periodic images of j. It also interacts with all of its own periodic images. The electrostatic energy of the unit cell, and related quantities such as forces on individual particles, are found by summing the

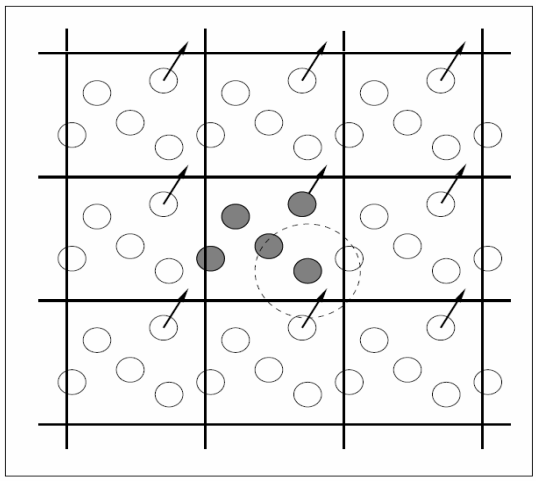

Figure 1.9: *Periodic boundary conditions. As a particle move out of the simulation box, an image particle moves in to replace it. In calculating particle interactions within the cut-off range both real and image neighbors are included.*

resulting infinite series. This latter converges (slowly) to a finite limit only if the unit cell is electrically neutral, and furthermore, the limit is found to depend on the order of summation (conditional limit rather than absolute convergence). Ewald (**Ewald, 1921**) applied a Jacobi theta transform to convert this slowly, conditionally convergent series to a pair of rapidly, absolutely convergent series, called the Ewald direct and reciprocal sums. The conditional convergence of the original series is expressed in a third term (**De Leeuw *et al.*, 1986**) as a quadratic function of the dipole moment of the unit cell, whose form depends on the order of summation. It is standard to assume that the whole assembly of unit cells is immersed in an external dielectric. Most commonly this dielectric is assumed to be fully conducting, in which case the third term vanishes and the order of summation becomes irrevelant. A more elementary derivation of the Ewald sum, using compensating Gaussian charge densities together with Poisson's equation under PBC in place of the Jacobi theta transform, can be found in the appendix to Kittel (**Kittel, 1986**). The Ewald direct sum resembles the standard Coulomb interaction, but with the term $(q_i q_j / r_{ij})$, representing the Coulomb energy of interaction between particles I and j, replaced by $q_i q_j \mathrm{erfc}(\beta r_{ij})/r_{ij,}$ where β is the so called Ewald convergence parameter. This latter term involving erfc converges rapidly to zero as a function of the interparticle distance r_{ij}, allowing the use of a finite cutoff. In the sander and pmemd programs the Ewald direct sum is calculated together with the van der Waals interactions. The default

value of the direct sum cutoff is 8 Å, independent of system size. Accordingly, β is chosen to be ~0.35 $Å^{-1}$, leading to a relative RMS force error due to truncation below 5 x 10^{-4}. The Ewald reciprocal sum is the sum, over all reciprocal lattice vectors m, ($m \neq 0$), of a Gaussian-like weight factor $\exp(-\pi^2 m^2/\beta^2)/(2\pi m^2)$ multiplied by $|S(m)|^2$, where the so called structure factor S(m) is given by the sum of $q_j \exp(2\pi i m r_j)$ over all particles j in the unit cell (r_j is the Cartesian coordinate vector of particle j). A cutoff can also be applied to the reciprocal sum. With the above choice of β, the number of reciprocal vectors needed so that the relative RMS force error due to truncation is below 5 x 10^{-4} is typically several times the number of particles in the unit cell. Because of the computational cost of calculating S(m) is of order N for each such vector m, the cost of the reciprocal sum is thus of order N^2. Unfortunately, this cost becomes prohibitive for systems containing tens of thousands of particles as is typical today.

What the PME algorithm does is to accurately approximate the structure factors S(m) using the 3DFFT. The essential idea is to first note that $\exp(2\pi i m r_j)$ can be factored into three one dimensional trigonometric terms (even in triclinic unit cells). One can then simply apply table lookup to these terms, approximating the trigonometric functions (evaluated at the crystallographic fractional coordinates of particle j) in terms of their values at nearby grid points. By this means the structure factors are approximated as sums over regular grid points, that is as a discrete Fourier transform that can be rapidly calculated using the 3DFFT, delivering all the needed structure factors at order N log(N) computational cost.

1.6.1.9. Energy minimization methods:

The potential energy calculated by adding the energies of various interactions is a numerical value for a single configuration. This number can be used to assess a particular configuration, but it may not be a useful measure of a configuration because it can be dominated by a few bad interactions. For instance, a large molecule with an excellent configuration for nearly all atoms can have a large overall energy because of a single bad interaction, for instance two atoms too near each other in space and having a huge van der Waals repulsion energy. It is often preferable to carry out energy minimization on a configuration to find the best nearby configuration.

Energy minimization is a numerical method for finding a minimum on the potential energy surface starting from a higher energy initial structure, labeled "1" as illustrated in **Figure 1.10.** During energy minimization, the geometry is changed in a stepwise manner so that the energy of the molecule is reduced, from steps 2 to 3 to 4 as shown in Figure 1.9. After a number of steps, a local or global minimum on the potential energy surface is reached.

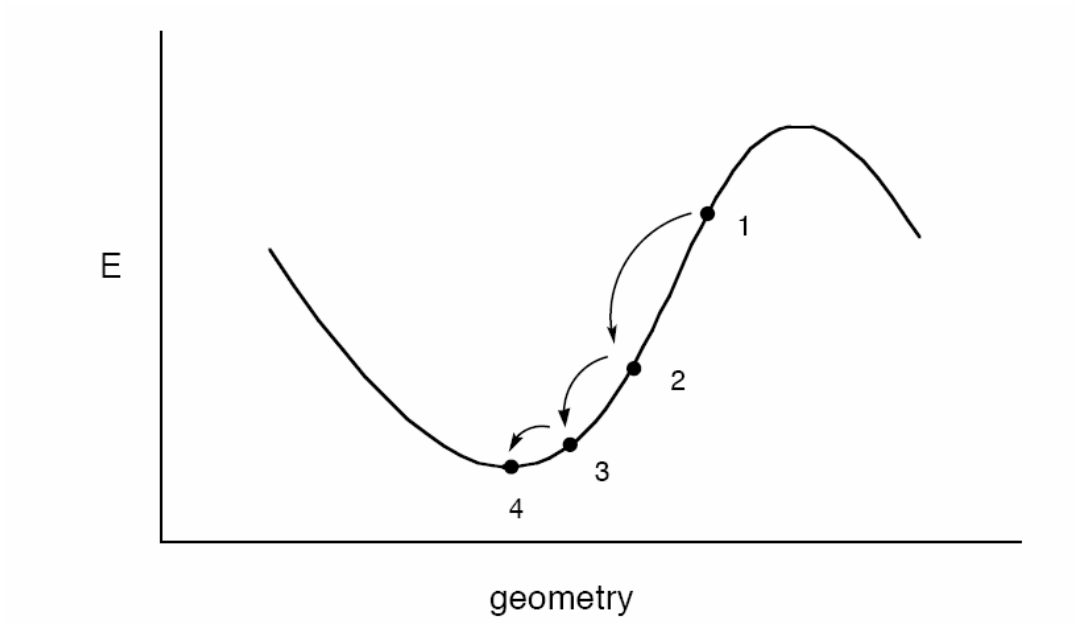

Figure 1.10: *The process of energy minimization changes the geometry of the molecule in a step-wise fashion until a minimum is reached.*

Most of the energy minimization methods proceed by determining the energy and the slope of the function at point 1. If the slope is positive, it is an indication that the coordinate is too large (as for point 1). If the slope is negative, then the coordinate is too small. The numerical minimization technique then adjusts the coordinate; if the slope is positive, the value of the coordinate is reduced as shown by point 2. The energy and the slope are again calculated for point 2. If the slope is zero, a minimum has been reached. If the slope is still positive, then the coordinate is reduced further, as shown for point 3, until a minimum is obtained. There are numerous methods for actually varying the geometry to find the minimum. Many of the methods used to find a minimum on the potential energy surface of a molecule use an iterative formula to work in a step-wise fashion. These are all based on formulas of the type:

$$X_{new} = X_{old} + correction \quad \text{-----(1.30)}$$

In the equation, X_{new} refers to the value of the geometry at the next step (for example, moving from step 1 to 2 in the figure), X_{old} refers to the geometry at the current step, and the correction is some adjustment made to the geometry. In all these methods, a numerical test is applied to the new geometry (X_{new}) to decide if a minimum is reached. For example, the slope may be tested to see if it is zero within some numerical tolerance. If the criterion is not met, then the formula is applied again to make another change in the geometry. There are three main energy minimization methods.

(a) Newton-Raphson method:

The Newton-Raphson method is the most computationally expensive per step of all the methods utilized to perform energy minimization. It is based on Taylor series expansion of the potential energy surface at the current geometry. The equation for updating the geometry is

$$X_{new} = X_{old} - \frac{E'(X_{old})}{E''(X_{old})} \quad \text{-----(1.31)}$$

Notice that the correction term depends on both the first derivative (also called the slope or gradient) of the potential energy surface at the current geometry and also on the second derivative (also called the curvature). It is the necessity of calculating these derivatives at each step that makes the method very expensive per step (especially for a multidimensional potential energy surface where there are many directions in which to calculate the gradients and curvatures). However, the Newton-Raphson method usually requires the fewest steps to reach the minimum.

(b) Steepest Descent Method:

Rather than requiring the calculation of numerous second derivatives, the steepest descent method relies on an approximation. In this method, the second derivative is assumed to be a constant. Therefore, the equation to update the geometry becomes

$$X_{new} = X_{old} - \gamma E'(X_{old}), \quad \text{-----(1.32)}$$

Where γ is a constant. In this method, the gradients at each point still must be calculated, but by not requiring second derivatives to be calculated, the method is much faster per step than the Newton-Raphson method. However, because of the

approximation, it is not as efficient and so more steps are generally required to find the minimum.

(c) Conjugate Gradient Method:

In this method, the gradients of the current geometry are first computed. The direction of the largest gradient is determined. The geometry is minimized along this one direction (this is called a line search). Then a direction orthogonal to the first one is selected (a "conjugate" direction). The geometry is minimized along this direction. This continues until the geometry is optimized in all the directions.

1.6.1.10. SHAKE Algorithm:

In the MD simulations the size of time step should be smaller than the motions characterized by the highest frequencies. In classical MD simulations these are typically the bond stretching motions involving hydrogen atoms. In order to use a larger time step all the bonds involving hydrogen atoms were constrained to their equilibrium position with the SHAKE algorithm (**Armstrong, 1998**; **Ryckaert *et al.* 1977**). In this algorithm at each step a correction $\vec{g}_a^{(r)}$ directed along the bond, is introduced in the forces such as to guarantee that the constraint is satisfied.

$$\vec{r}(t+\delta t)=\vec{r}(t+\delta t)+\frac{\delta t^2}{m_a}\vec{g}_a^{(r)} \qquad \text{.......................................} \quad (1.33)$$

where $r\ (t+\delta t)$ is the position that the system would have reached in the absence of the constraint. The procedure is then applied on the next constraints. In a chain of atoms the correction applied on constraint $i+1$ would partially disrupt constraint i. For this reason the correction is applied cyclically on each constraint in the system, until the desired convergence is reached for all the constraints.

1.6.1.11. The Berendsen thermostat:

Constant temperature MD simulations could be obtained by coupling to a Berendsen thermal bath (**Berendsen *et al.*, 1984**). At each step the velocities are scaled by a factor

$$\chi=\left(1+\frac{\delta t}{\tau_T}\left(\frac{T}{T_0}-1\right)\right) \qquad \text{.......................................} \quad (1.34)$$

where T_0 is the reference temperature and τ_T is a time constant that determines the strength of the coupling between the system and the thermal bath. A similar algorithm can be used to obtain constant pressure by periodic scaling of the simulation cell size and atomic positions.

1.6.1.12. Setting up and running a MD simulation:

The various steps involved in setting up and running a MD simulation is shown in detail in the form of flowchart (**Figure 1.11**).

(a) Initialization stage:

To set up a molecular dynamics simulation, we have to first select an initial configuration of the system, a starting point, or t = 0. Very often, in simulations of biomolecules, an X-ray crystal structure or an NMR structure is obtained from the Brookhaven Protein Databank and used as the initial structure. It is also possible to use a theoretical structure developed by homology modeling. The selection of the initial configuration must be done carefully as this can influence the quality of the simulation. It is often good to choose a configuration close to the state that one wish to simulate. Prior to starting a molecular dynamics simulation, it is sensible to do an energy minimization of the structure. This removes any strong van der Waals interactions that may exist, which might otherwise lead to local structural distortion and result in an unstable simulation. At this point, explicit water molecules are added to solvate the protein. If the starting structure is an X-ray crystal structure, then it is likely that some water molecules are already present, but the amount is usually insufficient for solvation. The solvating water molecules are usually obtained from a suitable large box of water that has been previous equilibrated. The entire box of water is overlayed onto the protein and those water molecules that overlap the protein are removed. At this point, another energy minimization should be done with the protein fixed in its energy minimized position. This allows the water molecules to readjust to the protein molecule.

(b) Heating the system:

Initial velocities at a low temperature are assigned to each atom of the system and Newton's equations of motion are integrated to propagate the system in time. If you are running an explicit solvent simulation, first fix the protein positions and let the waters

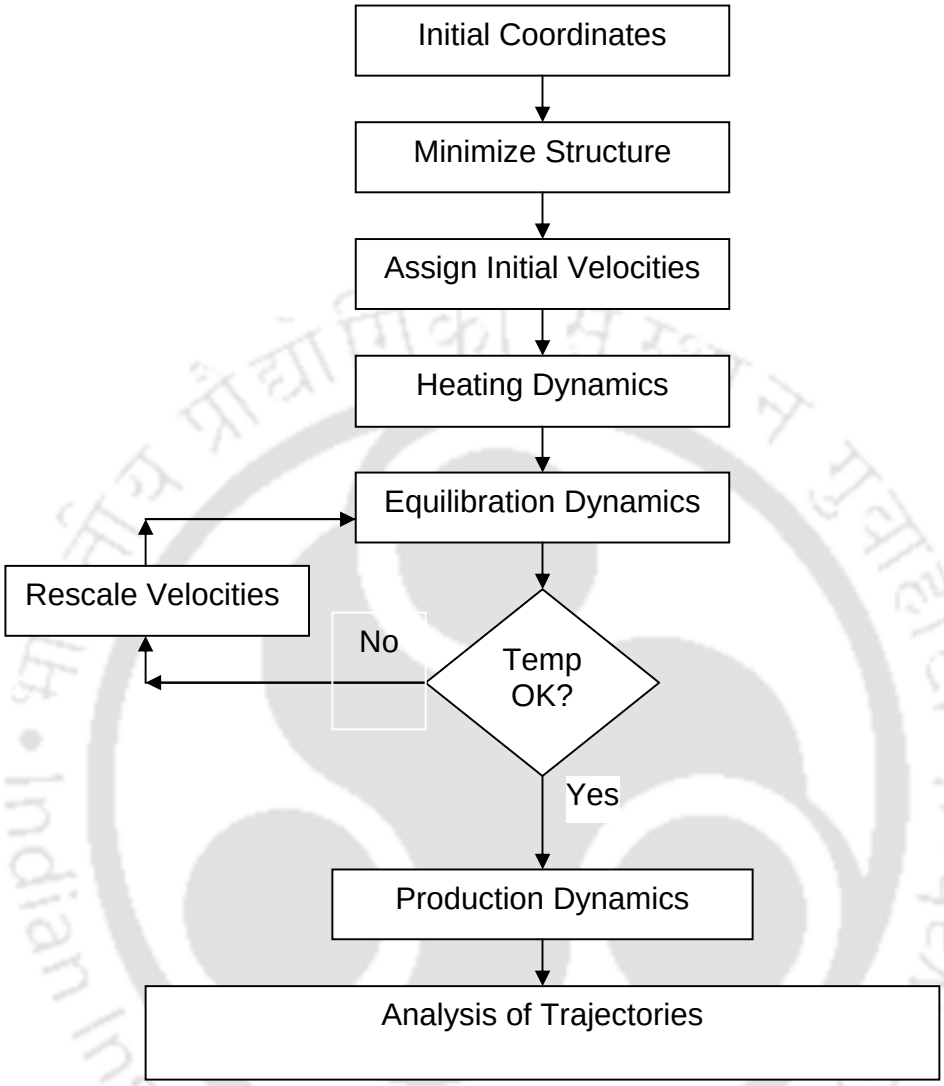

Figure 1.11: *Flowchart representation of various steps involved in running general MD simulation*

move to adjust to the present of the protein. Once the waters are equilibrated, the constraints on the protein can be removed and the whole system (protein+water) can evolve in time. During the heating phase, initial velocities are assigned at a low temperature and the simulation is started. Periodically, new velocities are assigned at a slightly higher temperature and the simulation is allowed to continue. This is repeated until the desired temperature is reached.

(c) Equilibration stage:

Once the preferred temperature is reached, the simulation of protein/water system continues and during this stage several properties are monitored; in particular, the structure, the pressure, the temperature and the energy. The point of the equilibration phase is to run the simulation until these properties become stable with respect to time. If the temperature increases or decreases significantly, the velocities can be scaled such that the temperature returns to near its desired value.

(d) Production phase:

The last step of the simulation is to run the simulation in "production" phase for the time length desired. This can be from some hundred ps to ns or more. It is during the production phase that thermodynamic parameters can be calculated so the simulation must conform to one of the ensembles described earlier.

(e) Analysis of Trajectory:

In the MD simulation, coordinates and velocities of the system are saved for supplementary analysis. Many time dependent properties can be displayed graphically, where one of the axes corresponds to time and the other to the quantity of interest, such as energy, RMSD, RMSF, B-factors, time dependent properties such as correlation functions etc. Average structures can be calculated and compared to experimental structures. The quantities that are usually calculated from an MD simulation include:

(a) **Mean Energy:** $$\langle E \rangle = \frac{1}{N}\sum_{i=1}^{N} E_i \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.35)$$

(b) **RMSD:** $$\left\langle \left(r_i^{\alpha} - r_i^{\beta} \right) \right\rangle^{1/2} = \sqrt{\frac{1}{N}\sum_i \left(r_i^{\alpha} - r_i^{\beta} \right)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.36)$$

(c) **RMSF:** $$\sqrt{\frac{1}{N_f}\sum_f \left(r_i^{f} - r_i^{ave} \right)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.37)$$

(d) **B-factors** (Temperature factors): $$B_i = \frac{8}{3}\pi^2 \sqrt{\frac{1}{N_f}\sum_f \left(r_i^{f} - r_i^{ave} \right)^2} \qquad \ldots. \quad (1.38)$$

(e) **Radius of Gyration** (R_g): $\sqrt{\frac{1}{N_i}\sum_i \left(r_i - r_{cm}\right)^2}$ ..... (1.39)

where r_i - r_{cm} is the distance between atom *i* and the center of mass of the molecule.

(f) **Time-correlation function**: $C(t) = P2(t) / R(t)$ ..... (1.40)

where *P2 (t)* is *<P2 (v (0).v (t))>* with *P2* is the legendre polynomial of order 2 and "*v*" is the vector chosen and *R (t)* is *<b^3(0) b^3(t)>* with b as the length of the vector chosen.

(g) **S^2 order parameter**: It is the measure of the fluctuations in the orientation of an N-H bond internuclear vector. It is a time dependent property such as correlation functions and is described by

$$S^2 = \lim_{t\to\infty} C(t) \qquad \text{..... (1.41)}$$

1.6.1.13. Molecular Modeling and Visualization:

Molecular graphics programs along with the MD simulation can be used to visualize the structure of biomolecules to get the better understanding about the conformational changes at an atomic level. Programs like XGrace can display the structural parameters of interest in a time dependent way and the programs like CHIMERA, RasMol, Sybyl, VMD, PyMol, SPDBV are useful tools for protein visualization and modeling.

1.6.2. Monitoring enzyme kinetics:

1.6.2.1. What are enzymes?

Enzymes are biological catalysts. They increase the rate of chemical reactions taking place within living cells without themselves suffering any overall change. The reactants of enzyme catalyzed reactions are termed substrates and each enzyme is quite specific in character, acting on a particular substrate or substrates to produce a particular product or products. However, without the presence of a non-protein component called a cofactor, many enzyme proteins lack catalytic activity. When this is the case, the inactive protein component of an enzyme is termed the apoenzyme, and the active enzyme, including cofactor, the holoenzyme. The cofactor may be an organic molecule, when it is known as a coenzyme, or it may be a metal ion. Some enzymes bind cofactors more

tightly than others. When a cofactor is bound so tightly that it is difficult to remove without damaging the enzyme it is sometimes called prosthetic group.

1.6.2.2. Enzyme Kinetics:

The sole function of an enzyme is to catalyze a reaction. Enzyme kinetics is that branch of enzymology that deals with the factors affecting the rates of enzyme-catalyzed reactions. The most important factors are: enzyme concentration, ligand concentrations (substrates, products, inhibitors, and activators), pH, ionic strength, and temperature. When all these factors are analyzed properly, it is possible to learn a great deal about the nature of the enzyme. For example, by varying the substrate and product concentrations, it is possible to deduce the kinetic mechanism of the reaction, that is, the order in which substrates add and products leave the enzyme. Such studies establish the kinds of enzyme-substrate and enzyme-product complexes that can form and thereby tell us something about the architecture of the active site. In some cases the kinetics of a reaction provides evidence for stable, covalently-bound intermediates that are undetectable by ordinary chemical analyses. Certain kinetic constants can be determined and from these we can make an educated guess concerning the usual intracellular concentrations of substrates and products and the physiological direction of the reaction. The kinetics of a reaction may indicate the way in which the activity of the enzyme is regulated *in vivo*. A study of the effect of varying pH and temperature on the kinetics constants can provide information concerning the identities of the amino acid R-groups of the active site. A kinetic analysis can lead to a model for an enzyme-catalyzed reaction and, conversely, the principles of enzyme kinetics can be used to write the kinetic equation for an attractive model. The kinetic equation tells us exactly how all the ligands of a system interact to affect the velocity of the reaction. Consequently, once we have possible equations, the model can be tested experimentally. For many biologists, a thorough understanding of enzyme kinetics is indispensable to their research.

1.6.2.3. Michealis-Menten Equation:

Named in honour of L. Michaelis and M. L. Menten (**Michaelis and Menten, 1913**), this equation was derived in 1902 by A. Brown (**Brown, 1902**) and V. Henri (**Henri, 1902**) and is based in its final version on the steady state assumption of G. E. Briggs and J. B. S. Haldane (**Briggs and Haldane, 1925**), derived in 1925. The equation

states that the simplest type of reaction catalyzed by an enzyme (E) with a single substrate (S), assuming irreversible conversion to product (P) is:

$$E + S \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} ES \xrightarrow{k_2} E + P \qquad \text{----- (1.42)}$$

To describe this reaction, a differential equation can be formulated for each component:

$$d[S]/dt = -k_1 [S][E] + k_{-1}[ES] \qquad \text{----- (1.43)}$$

$$d[E]/dt = -k_1 [S][E] + (k_{-1} + k_2)[ES] \qquad \text{----- (1.44)}$$

$$d[ES]/dt = k_1 [S][E] - (k_{-1} + k_2)[ES] \qquad \text{----- (1.45)}$$

$$d[P]/dt = k_2 [ES] = v_0 \qquad \text{----- (1.46)}$$

It is not possible to derive a simple rate equation from these relationships and therefore simplifications were sought. They concentrate on the plausible assumption that the binding equilibrium set up in the first part of the reaction should be fast compared with the following catalytic step, so that the rate constants would be related according to $k_1 \sim k_{-1} > k_2$. In the original assumption the catalytic constant k_2 should be so low that it does not influence the binding equilibrium. However, since the reaction velocity depends directly on the amount of ES ($v_0 = k_2$ [ES]), v_0 is a direct measure of ES and therefore can be used to determine the dissociation constant according to the law of mass action:

$$K_d = \frac{k_{-1}}{k_1} = \frac{[S][E]}{[ES]} \qquad \text{----- (1.47)}$$

In their more differentiated view, Briggs and Haldane recognized that k_2 cannot be completely neglected; rather, formation (due to k_1) and decay (due to k_{-1} and k_2) of the enzyme-substrate complex ES compensate for one another, so that [ES] remains constant: d [ES] / dt = 0. This state, however, is maintained for only a limited time and was designated 'steady state' by the authors, to differentiate it from real equilibrium. For this time period the rate equation for the overall reaction can be simplified. The differential equation for the enzyme-substrate complex (1.42) can thus be simplified:

$$d[ES]/dt = k_1[S][E] - (k_{-1} + k_2)[ES] = 0 \qquad \text{----- (1.48)}$$

Taking into account that the total amount of enzyme used in the assay consists of free and complex-bound enzyme, $[E]_0 = [E] + [ES]$, this equation can be rearranged:

$$[ES] = \frac{k_1[S][E]_0}{k_1[S] + (k_{-1} + k_2)} \qquad \text{----- (1.49)}$$

and inserted into the expression for the velocity, dividing the nominator and denominator by k_1:

$$v_0 = \frac{d\ [P]}{dt} = k_2[ES] = \frac{k_2[S][E]_0}{[S] + ((k_{-1} + k_2)/k_1)} \qquad \text{---- (1.50)}$$

The constant term of the three rate constants in the denominator is replaced by one kinetic constant, the Michaelis constant: $K_m = (k_{-1} + k_2) / k_1$. Since during the experiment the total amount of enzyme should remain constant, it is combined with the catalytic rate constant to give the maximum velocity $V_{max} = k_2\ [E]_0$, which is reached when all enzyme molecules present in the assay are taking part in catalysis.

$$v_0 = \frac{V_{max}\ [S]}{[S] + K_m} \qquad \text{----- (1.51)}$$

This is the final form of the Michaelis-Menten equation. Although the catalytic constant k_2 (generally designated k_{cat}) is a characteristic constant for a given enzyme, the maximum velocity V_{max} depends on both the immediate amount and activity of enzyme present in the reaction mixture and cannot easily be compared between different assay conditions (especially when carried out in different laboratories). For its exact determination, both the molarity and the specific activity of the enzyme must be known. In contrast to this, the Michaelis constant is independent of enzyme amount and activity, and corresponding values should be obtained under similar test conditions.

In comparison to the dissociation constant K_d, the Michaelis constant K_m is extended by the catalytic constant k_2. Consequently, it is similar to K_d only if k_2 becomes very small compared with k_1 and k_{-1}, but it differs considerably from the dissociation constant the closer k_2 becomes to k_1 and k_{-1}. But since in most cases k_2 is rather small, the contribution of the binding constant to the Michaelis constant dominates the contribution of the catalytic constant and in these cases the Michaelis constant can be regarded essentially as an indication of affinity. Like the dissociation constant, K_m has the dimension of concentration (M), but the catalytic constant k_2 is a first-order rate constant and has the dimension of s^{-1}, and the maximum velocity V_{max} is in units of concentration per time ($M\ s^{-1}$). To check the validity of Michaelis-Menten equation for a special

enzyme reaction and to determine the Michaelis constant and the maximum velocity, the reaction rate must be determined with different amounts of substrate. The actual enzyme velocity is obtained from the slope of the initial, linear part of the progress curve, as d [P] / dt or –d [S] / dt (M s^{-1} or µmol min^{-1}). Plotting these values against the initial substrate concentration should yield a hyperbolic saturation curve as shown in **Figure 1.12.**

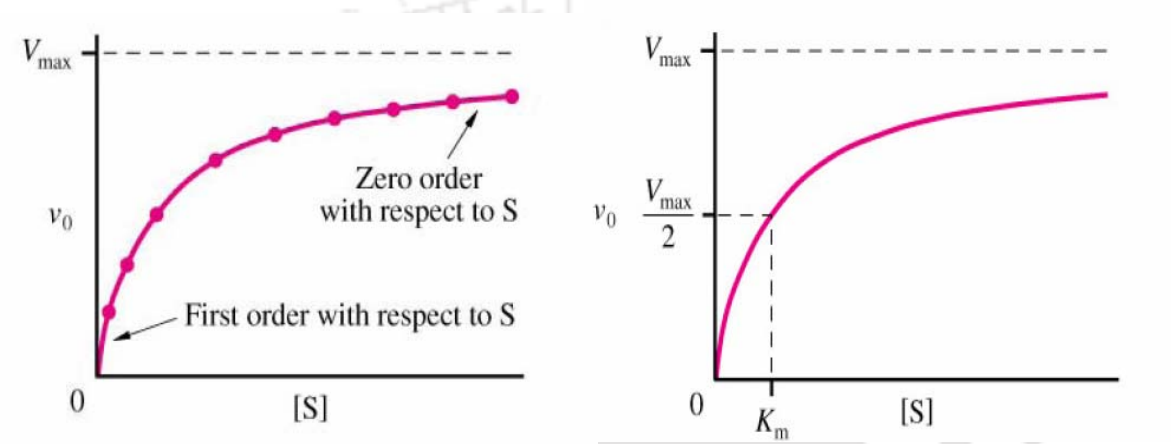

Figure 1.12: *(a) Saturation curve according to the Michaelis-Menten equation (b) Determination of maximum velocity V_{max} and the Michaelis constant K_m are shown.*

The velocity approaches a plateau at infinite substrate concentration, since for [S] >> K_m, the latter can be neglected and the Michaelis-Menten equation reduces to $v_0 = V_{max}$. For [S] = K_m, the equation becomes $v_0 = V_{max}/2$, and the substrate at half-maximum velocity has the value of K_m. Because of this relationship the substrate concentrations for studying the Michaelis-Menten kinetics should be chosen within the range of the Michaelis constant, preferentially from one order of magnitude below to one order of magnitude above K_m.

If we examine the v_0 versus [S] curve, we find three distinct regions where the velocity responds in a characteristic way to increasing [S] (**Figure 1.12**). At very low substrate concentrations (eg. [S] < 0.01 K_m), the v_0 versus [S] curve is essentially linear; that is, the velocity (for all practical purposes) is directly proportional to the substrate concentration. This is the region of first-order kinetics. At very high substrate concentration (eg. [S] > 100 K_m), the velocity is essentially independent of the substrate

concentration. This is the region of zero-order kinetics. At intermediate substrate concentrations, the relationship between v_0 and [S] follows neither first-order nor zero-order kinetics.

1.6.2.4. Principles of UV/Visible Photometry:

Photometric measurements are the most frequent methods employed in enzyme analysis, mainly because of the accuracy of the method, the relative ease of use, the moderate costs, and multiple usability of the instruments. To understand the principle of photometry, one must keep in mind that the instrument, although called an absorption photometer, actually measures light transmission, that is, the decrease in the intensity of incident light I_0 by the solution in the cuvette to become I:

$$I / I_o = 10^{-\varepsilon. c. d} \qquad \text{----- (1.52)}$$

Where d is the path length of the cuvette, c the concentration of the substance being measured, and ε its absorption coefficient (dimension: mol^{-1} cm^{-1}), a material constant, of the absorbing substance. I_0 is defined as the total light intensity (100%) under the prevailing conditions, that is the design of the photometer, the relative portion of light supplied by the lamp, the light path, monochromator arrangement, band width, photomultiplier sensitivity, selected wavelength, and the blank solution in the cuvette, and the completely closed light path is taken as 0% intensity. The quotient I/I_0 allows absolute quantitative determinations independent of specific conditions and photometer types. The inconvenient exponential dependence on the substance concentration is transformed to a linear one by the well known Beer-Lambert law, in which the negative logarithm of the transmission is defined as absorption A:

$$A = - \log (I / I_0) = \varepsilon\, c\, d \qquad \text{----- (1.53)}$$

Most photometers allow the measurement of both transmission and absorption.

1.7. Specific objectives of the thesis:

1. a) To develop a method to isolate and study loop regions in protein structures using their Ramachandran maps.
   b) To attempt an approach to quantify disorder in unstructured proteins/regions.
2. To characterize the dynamics of disordered regions and disordered proteins using molecular dynamics simulations by analyzing structural parameters such as radius of gyration, solvent accessible surface area, end to end chain distance, secondary structural analysis and so on.
3. To study the effect of molecular crowding by dextrans of various sizes such as 15-20, 40, 70, 200, 500 & 2000 kDa on the
   a) Rate of Alkaline Phosphatase-catalyzed hydrolysis of p-nitro phenyl phosphate (PNPP).
   b) Rate of Acetyl Cholinesterase-catalyzed hydrolysis of 2-naphthyl acetate and 3-indoxyl acetate.

4. To understand the Human Lysozyme amyloidosis: Insights from the conformational dynamics of wild type and mutants (I56T), (D67H), and (T70N) using molecular dynamics simulations.

CHAPTER 2
A novel approach to segregate and identify functional loop regions in Protein Structures using their Ramachandran Maps

A novel approach to segregate and identify functional loop regions in Protein Structures using their Ramachandran Maps

2.1. Introduction:

Loops occur ubiquitously among proteins. These are the flexible short segments that connect the two stable secondary structural units and plays important role in protein function such as ligand binding, catalysis, recognition, and allosteric regulation. The most challenging task in protein structure determination and modeling is the prediction of loops. The favored conformation of loops often remains unclear even when the rest of the protein is resolved at high resolution. This is mainly due to the high flexibility of loops that is often tied to their function. The dihedral angles of these irregular regions has been shown to be scattered over several regions in the Ramachandran map, while their structure depends on their sequence, their length and also on the elements and molecular packing of the regular secondary structure, like α-helices or β-strands, to which they connect (**Efimov, 1993**). The sequence inconsistency in loops also presents a particular challenge for homology modeling methods. So, for understanding the function of proteins and the biological processes the loops mediate, the characterization of loop structures and their motions is thought to be important.

In this work we analyze the Ramachandran maps of loop regions and other regular secondary structure elements with the objective of obtaining some features unique to loop regions. The tightly clustered dihedral points at confined regions of the Ramachandran map strongly correlate with the presence of regular secondary structure in the protein. An irregular structure like loop in contrast has no dihedral angle constraints or preferences yielding a scattered distribution of points in the Ramachandran space. This extent of scattering is quantitatively measured to arrive at a parameter (**MSRP**) that has unique value for loop regions in the protein structure. We also measure the **MSRP** of loop regions in protein complexes, both in their free and ligand bound states, and show how MSRP changes can correlate with structural perturbations in loop regions subsequent to binding of ligand. Finally, using MD simulations, we demonstrate that time-dependent changes in MSRP along the trajectory are more for unstructured proteins in comparison to folded globular proteins.

2.2. Methods:

2.2.1. Selection of Structured Proteins:

In this work, four sets of structured proteins were used. A total of 150 protein domains were selected from the **SCOP** (Structural Classification of Proteins) (**Murzin *et al.*, 1995**) database, release 1.65, belonging to class a, b, c and d. Out of 150 protein domains, 50 were taken from class a (all-α proteins- proteins which form compact structure by packing mainly α-helices, often in a symmetric arrangement around a central hydrophobic core), 50 were taken from class b (all-β proteins- proteins which pack together mainly β-sheets, with adjacent strands linked by turns and loops and various hydrogen-bonding networks formed among the individual strands, often resulting in layered or barrel structures), 25 were taken from class c (α/β proteins- proteins that are folded with alternating α-helices and β-strands, often forming layered or barrel-like structures) and 25 were taken from class d (α+β proteins- proteins that combine largely-separated that is non-alternating helical and strand regions, often by hairpins). All these structured proteins were selected based on quality, where no structure in the set has a resolution poorer than 2.5 Å, with less than 25% sequence identity. Furthermore, they are globular domains with important biological function. See **Tables 2.1- 2.4** for details.

Table 2.1: *Class a, All α Proteins**

S.No.	PDB	Name	Residue Range
1	1dlwA	Protozoan/bacterial hemoglobin {Ciliate (Paramecium caudatum)}	
2	1kr7A	Nerve tissue mini-hemoglobin (neural globin) {Milky ribbon-worm (Cerebratulus lacteus)}	
3	1jj2U	Ribosomal protein L29 (L29p) {Archaeon Haloarcula marismortui}	
4	1fpoA	HSC20 (HSCB), N-terminal (J) domain {Escherichia coli}	001-76
5	1aqt_	Epsilon subunit of F1F0-ATP synthase C-terminal domain {Escherichia coli}	87-136
6	1kcfA	Mitochondrial resolvase ydc2 N-terminal domain {Fission yeast (Schizosaccharomyces pombe)}	003-38
7	1c75A	Cytochrome c6 (synonym: cytochrome c553) {Bacillus pasteurii}	
8	1qksA	N-terminal (heme c) domain of cytochrome cd1-nitrite reductase {Paracoccus denitrificans}	009-135
9	1kb0A	Quinoprotein alcohol dehydrogenase, C-terminal domain {Comamonas testosteroni}	579-675
10	1mh4A	Mating type protein A1 Homeodomain {Baker's yeast (Saccharomyces cerevisiae)}	472-523
11	1gvdA	c-Myb, DNA-binding domain {Mouse (Mus musculus)}	
12	1bl0A	MarA {Escherichia coli}	63-124
13	2tct1	Tetracyclin repressor (Tet-repressor, TetR) {Escherichia coli}	002-67

14	1mgtA	O6-alkylguanine-DNA alkyltransferase {Archaeon Pyrococcus kodakaraensis}	89-169
15	1mkmA	Transcriptional regulator IclR, N-terminal domain {Thermotoga maritima}	001-75
16	1hw1A	Fatty acid responsive transcription factor FadR, N-terminal domain {Escherichia coli}	005-78
17	1oaiA	FG-binding, C-terminal domain of TAP {Human (Homo sapiens)}	
18	2erl_	ER-1 {Euplotes raikovi}	
19	1ail_	N-terminal, RNA-binding domain of nonstructural protein NS1 {Influenza A virus}	
20	1a32_	Ribosomal protein S15 {Bacillus stearothermophilus}	
21	1b67A	Archaeal histone {Archaeon Methanothermus fervidus, histone A}	
22	1n1jB	Nuclear transcription factor Y subunit gamma (Nf-Yc2) {Human (Homo sapiens)}	
23	1ls1A	Signal sequence recognition protein Ffh {Thermus aquaticus}	001-88
24	1h99A	Transcriptional antiterminator LicT {Bacillus subtilis}	54-168
25	1o9rA	Dodecameric ferritin homolog {Agrobacterium tumefaciens, Dps}	
26	1fr2A	ImmE9 protein (Im9) {Escherichia coli}	
27	1e3oC	Oct-1 {Human (Homo sapiens)}	001-75
28	1ji7A	Etv6 transcription factor pointed domain (Tel SAM) {Human (Homo sapiens)}	
29	1ecmA	Chorismate mutase domain of P-protein {Escherichia coli}	
30	1kncA	Antioxidant defence protein AhpD {Mycobacterium tuberculosis}	
31	1pbwA	p85 alpha subunit RhoGAP domain {Human (Homo sapiens)}	
32	1ks9A	Ketopantoate reductase PanE {Escherichia coli}	168-291
33	1a6m_	Myoglobin {Sperm whale (Physeter catodon)}	
34	1k6kA	N-terminal, ClpS-binding domain of ClpA, an Hsp100 chaperone {Escherichia coli}	
35	1g8qA	CD81 extracellular domain {Human (Homo sapiens)}	
36	1g72B	Methanol dehydrogenase, light chain {Methylophilus methylotrophus, w3a1}	
37	1elwA	Hop {Human (Homo sapiens)}	
38	1n8vA	Chemosensory protein Csp2 {Cabbage moth (Mamestra brassicae)}	
39	1or7A	SigmaE factor (RpoE) {Escherichia coli}	001-111
40	1h3lA	Sigma factor SigR {Streptomyces coelicolor a3(2)}	
41	1g8eA	Flagellar transcriptional activator FlhD {Escherichia coli}	
42	1dd3A	Ribosomal protein L7/12, oligomerisation (N-terminal) domain {Thermotoga maritima}	001-57
43	1n1eA	Glycerol-3-phosphate dehydrogenase {Trypanosome (Leishmania mexicana)}	198-357
44	1mv8A	GDP-mannose 6-dehydrogenase, middle domain {Pseudomonas aeruginosa}	203-300
45	1f0yA	Short chain L-3-hydroxyacyl CoA dehydrogenase {Human (Homo sapiens)}	204-302
46	1ko9A	8-oxoguanine glycosylase {Human (Homo sapiens)}	136-323
47	1b0nB	SinR repressor, DNA-binding domain {Bacillus subtilis}	
48	1b4fA	EphB2 receptor {Human (Homo sapiens)}	

49	1on2A	Manganese transport regulator MntR {Bacillus subtilis}	63-136
50	1hlvA	DNA-binding domain of centromere binding protein B (CENP-B) {Human (Homo sapiens)}	001-66

Table 2.2 : *Class b, All β Proteins**

S.No.	PDB	Name	Residue Range
1	1mqkL	Immunoglobulin light chain kappa variable domain, VL-kappa {Mouse (Mus musculus), cluster 4}	
2	1ogaE	T-cell antigen receptor {Human (Homo sapiens), beta-chain}	005-118
3	1cdy_	N-terminal domain of CD4 {Human (Homo sapiens)}	001-97
4	1ncwL	Immunoglobulin light chain kappa constant domain, CL-kappa {Mouse (Mus musculus)}	108-214
5	1ogaD	T-cell antigen receptor {Human (Homo sapiens), beta-chain}	118-202
6	1k3iA	Galactose oxidase, C-terminal domain {Fungi (Fusarium spp)}	538-639
7	1qhoA	Five domain "maltogenic" alpha-amylase (glucan 1,4-alpha-maltohydrolase), domain D {Bacillus stearothermophilus}	496-576
8	1e5bA	Endo-1,4-beta xylanase D, xylan binding domain, XBD {Cellulomonas fimi}	
9	1qhoA	Five domain "maltogenic" alpha-amylase (glucan 1,4-alpha-maltohydrolase), domain D {Bacillus stearothermophilus}	577-686
10	1js8A	C-terminal domain of mollusc hemocyanin {Giant octopus (Octopus dofleini)}	2792-2892
11	1kzqA	Major surface antigen p30, SAG1 {Toxoplasma gondii}	003-131
12	1qpxA	PapD {Escherichia coli}	125-215
13	1jk4A	Neurophysin II {Cow (Bos taurus)}	
14	1lox_	15-Lipoxygenase {Rabbit (Oryctolagus cuniculus)}	002-72
15	1bu8A	Pancreatic lipase, C-terminal domain {Rat (Rattus norvegicus)}	337-449
16	1k3iA	Galactose oxidase, N-terminal domain {Fungi (Fusarium spp)}	-12-150
17	1of4A	Beta-mannosidase, C-terminal domain {Thermotoga maritima}	
18	1n7oA	Hyaluronate lyase {Streptococcus pneumoniae}	815-890
19	1d2sA	Sex hormone-binding globulin {Human (Homo sapiens)}	
20	1h2cA	EV matrix protein {Ebola virus}	
21	1fx7A	Iron-dependent regulator IdeR {Mycobacterium tuberculosis}	145-230
22	1h9rA	C-terminal domain of molybdate-dependent transcriptional regulator ModE {Escherichia coli}	123-199
23	1a8p_	Ferredoxin reductase (flavodoxin reductase) N-terminal domain {Azotobacter vinelandii}	002-100
24	1exmA	Elongation factor Tu (EF-Tu) {Thermus thermophilus}	313-405
25	2hrvA	2A cysteine proteinase {Human rhinovirus 2}	
26	1bco_	mu transposase, C-terminal domain {Bacteriophage mu}	481-560
27	1e79A	F1 ATP synthase alpha subunit, domain 1 {Cow (Bos taurus)}	19-94
28	1kzkA	Human immunodeficiency virus type 1 protease {Human immunodeficiency virus type 1}	
29	1qqgA	Insulin receptor substrate 1, IRS-1 {Human (Homo sapiens)}	159-262
30	1bebA	beta-Lactoglobulin {Cow (Bos taurus)}	

31	1mxgA	Bacterial alpha-Amylase {Archaeon Pyrococcus woesei}	362-435
32	2arcA	Regulatory protein AraC {Escherichia coli}	
33	1qiuA	Adenovirus {Human adenovirus type 2}	319-395
34	1g8lA	Molybdenum cofactor biosynthesis protein MoeA, C-terminal domain {Escherichia coli}	327-409
35	1k3xA	Endonuclease VIII {Escherichia coli}	001-124
36	1iv8A	Maltooligosyl trehalose synthase {Archaeon Sulfolobus acidocaldarius}	654-720
37	1c39A	Cation-dependent mannose 6-phosphate receptor, extracytoplasmic domain {Cow (Bos taurus)}	
38	1swuA	Streptavidin {Streptomyces avidinii}	
39	1e0tA	Pyruvate kinase (PK) {Escherichia coli}	70-167
40	1fmtA	Methionyl-tRNAfmet formyltransferase, C-terminal domain {Escherichia coli}	207-314
41	1arb_	Achromobacter protease {Achromobacter lyticus, strain m497-1}	
42	1ep3B	Dihydroorotate dehydrogenase B, PyrK subunit {Lactococcus lactis, isozyme B}	002-102
43	1jb9A	Ferredoxin reductase (flavodoxin reductase) N-terminal domain {Maize (Zea mays), root isoform}	006-162
44	1m4vA	Superantigen-like protein SET3 {Staphylococcus aureus}	005-100
45	1dj7B	Ferredoxin thioredoxin reductase (FTR), alpha (variable) chain {Synechocystis sp.}	
46	1guiA	Carbohydrate binding module from laminarinase 16A {Thermotoga maritima}	
47	1oe1A	Nitrite reductase, NIR {Alcaligenes xylosoxidans}	160-336
48	1igqA	Transcriptional repressor protein KorB {Escherichia coli}	
49	1nteA	Syntenin 1 {Human (Homo sapiens)}	
50	1k0rA	S1 domain of NusA {Mycobacterium tuberculosis}	108-183

Table 2.3: *Class c, All α / β Proteins**

S.No.	PDB	Name	Residue Range
1	1n55A	Triosephosphate isomerase {Leishmania mexicana}	
2	1li4A	S-adenosylhomocystein hydrolase {Human (Homo sapiens)}	190-352
3	1b8pA	Malate dehydrogenase {Aquaspirillum arcticum}	003-158
4	1ceqA	Lactate dehydrogenase {Malaria parasite (Plasmodium falciparum)}	19-352
5	1ks9A	Ketopantoate reductase PanE {Escherichia coli}	001-167
6	1fl2A	Alkyl hydroperoxide reductase subunit F (AhpF), C-terminal domains {Escherichia coli}	326-451
7	1d7yA	NADH-dependent ferredoxin reductase, BphA4 {Pseudomonas sp., KKS102}	116-236
8	1a9xB	Carbamoyl phosphate synthetase, small subunit N-terminal domain {Escherichia coli}	1502-1652
9	1ay7B	Barstar (barnase inhibitor) {Bacillus amyloliquefaciens}	
10	1knxA	HPr kinase/phoshatase HprK N-terminal domain {Mycoplasma pneumoniae}	001-132
11	1jj2K	Ribosomal protein L15 (L15p) {Archaeon Haloarcula marismortui}	
12	1l0bA	Breast cancer associated protein, BRCA1 {Rat (Rattus norvegicus)}	1591-1702

13	1npyA	Shikimate 5-dehydrogenase-like protein HI0607 {Haemophilus influenzae}	001-102
14	1jbeA	CheY protein {Escherichia coli}	
15	1dbwA	Transcriptional regulatory protein FixJ, receiver domain {Rhizobium meliloti}	
16	1eucB	Succinyl-CoA synthetase, beta-chain, C-terminal domain {Pig (Sus scrofa)}	246-393
17	1f8yA	Nucleoside 2-deoxyribosyltransferase {Lactobacillus leichmannii}	
18	1ccwA	Glutamate mutase, small subunit {Clostridium cochlearium}	
19	1gh2A	Thioredoxin-like protein, N-terminal domain {Human (Homo sapiens)}	
20	1e6bA	Class zeta GST {Mouse-ear cress (Arabidopsis thaliana)}	008-87
21	1qmhA	RNA 3'-terminal phosphate cyclase, RPTC, insert domain {Escherichia coli}	185-279
22	1dzfA	Eukaryotic RPB5 N-terminal domain {Baker's yeast (Saccharomyces cerevisiae)}	005-143
23	1sfe_	Ada DNA repair protein {Escherichia coli}	012-92
24	1mgtA	O6-alkylguanine-DNA alkyltransferase {Archaeon Pyrococcus kodakaraensis}	001-88
25	1b74A	Glutamate racemase {Aquifex pyrophilus}	001-105

Table 2.4: *Class d, All α + β Proteins**

S.No.	PDB	Name	Residue Range
1	1lniA	RNase Sa {Streptomyces aureofaciens}	
2	1i0vA	RNase T1 {Aspergillus oryzae}	
3	3lztC	Lysozyme {Chicken (Gallus gallus)}	
4	1i4mA	Prion protein domain {Human (Homo sapiens)}	
5	1qmeA	Penicillin-binding protein 2x (pbp-2x), C-terminal domain {Streptococcus pneumoniae}	632-692
6	1ogwA	Ubiquitin {Human (Homo sapiens)}	
7	1hz6A	Immunoglobulin light chain-binding domain of protein L {Peptostreptococcus magnus}	
8	1i8tA	UDP-galactopyranose mutases {Escherichia coli}	245-313
9	1pcfA	Transcriptional coactivator PC4 C-terminal domain {Human (Homo sapiens)}	
10	1itxA	Chitinase A1 {Bacillus circulans}	338-409
11	1jj2W	Ribosomal protein L31e {Archaeon Haloarcula marismortui}	
12	1psuB	Phenylacetic acid degradation protein PaaI {Escherichia coli}	
13	3proC	Alpha-lytic protease prodomain {Lysobacter enzymogenes}	006-85
14	1egaA	GTPase Era C-terminal domain {Escherichia coli}	183-295
15	1kskA	Ribosomal small subunit pseudouridine 516 synthase RsuA {Escherichia coli}	125-231
16	1iqoA	Hypothetical protein MTH1880 {Archaeon Methanothermobacter thermautotrophicus}	001-96
17	1ufyA	Chorismate mutase {Thermus thermophilus}	
18	1kbiA	Flavocytochrome b2, N-terminal domain {Baker's yeast (Saccharomyces cerevisiae)}	001-97
19	1seiA	Ribosomal protein S8 {Bacillus stearothermophilus}	

20	1b5qA	Polyamine oxidase {Maize (Zea mays)}	294-405
21	1gy7A	Nuclear transport factor-2 (NTF2) {Baker's yeast (Saccharomyces cerevisiae)}	
22	1pinA	Mitotic rotamase PIN1, domain 2 {Human (Homo sapiens)}	45-163
23	1di2A	Double-stranded RNA-binding protein A, second dsRBD {Xenopus laevis}	
24	1pda_	Porphobilinogen deaminase (hydroxymethylbilane synthase), C-terminal domain {Escherichia coli}	220-307
25	1gmuA	Urease metallochaperone UreE, C-terminal domain {Klebsiella aerogenes}	71-138

** each entry is provided with the four-digit PDB codes and the protein chain identifiers, the residue range of the domain if not the whole chain.*

2.2.2. Selection of Unstructured Proteins:

A total of 20 unstructured proteins were identified from DisProt database (**Sickmeier *et al.*, 2007**), v3.8, while their structures were procured from the PDB. See **Table 2.5** for details.

Table 2.5: *Unstructured Proteins*

S.No.	PDB	Name
1	2fft	Thylakoid soluble phosphoprotein
2	2eze_A	High mobility group protein HMG-I/HMG-Y
3	1hn3	p19 ARF protein
4	2ju4	Retinal rod rhodopsin-sensitive GMP-PDE gamma-subunit
5	1vzs	F6 subunit of ATP synthase, mitochondrial precursor
6	1uss	Hho1p
7	1r8u_A	CIT-ED2
8	1jh3	Tyrosyl-tRNA synthetase
9	2rn9	Cytochrome c oxidase copper chaperone
10	1hcp	Estrogen receptor alpha
11	1zza	Stannin
12	2c55	HIV Type 1 p6 Protein
13	1qk9	Methyl-Cpg-binding protein 2
14	1ueo	Penaeidin-3a
15	1tv0	Cryptdin-4
16	2sob	SN-OB, OB-fold sub-domain of Staphylococcal nuclease
17	2ddn	CP12
18	1bnb	Beta-defensin 12
19	1xq8	Alpha-synuclein
20	1lxl	Apoptosis regulator Bcl-x_L

Out of 20 unstructured proteins, except for one which is theoretical model structure, all were NMR structures. All these proteins contain unstructured region to an extent of 60% or more.

2.2.3. Selection of APO/HOLO sets:

(a) DNA binding Proteins:

A total of 43 pairs of DNA-binding proteins having their structures determined both in the DNA-bound (HOLO) and unbound (APO) forms were selected from a previous study (**Gao and Skolnick, 2009**). These proteins were selected using the following criteria: (i) the holo- and apo-structures share 90% global sequence identity; (ii) the protein is bound to a specific DNA molecule in the holo-form; (iii) the protein chain length is less than 400 residues; and (iv) the DNA bound to protein has more than 7 and less than 40 base pairs. No two pairs of proteins share a sequence identity greater than 35%. See **Table 2.6** for details

(b) Other Ligand bound Proteins:

Another set containing a total of 98 pairs of ligand bound (HOLO) and free form (APO) of protein structures determined by X-ray crystallography at a resolution of 2.5 Å or better was used in this work. These proteins were selected from a previous study (**Gunasekaran and Nussinov, 2007**). This protein dataset contains 3 classes of proteins. A total of 41 proteins in class I where no conformational change (C^{α} displacement with less than 0.5 Å) occurs upon ligand binding, 35 proteins in class II where moderate conformational change (greater than or equal to 0.5 Å but less than or equal to 2.0 Å) occurs and 22 proteins in class III where a large conformational change (greater than 2.0 Å) occurs upon ligand binding. See **Table 2.6** for details.

Table 2.6: *APO / HOLO Proteins**

DNA binding Proteins				
S.No.	**APO**	**HOLO**	**Residue Range**	**Protein Description**
1	1a41_	2h7gX		Type IB Topoisomerase
2	1af5_	1n3fA		Endonuclease I-CreI
3	1ajyA	1zmeC		PUT3
4	1arrA	1parA		Arc repressor
5	1aw6_	1d66A		GAL4

6	1es8A	1dfmA		Endonuclease BglII
7	1ev7A	1iawA	177-309	Endonuclease NaeI
8	1evxA	1cz0A		Endonuclease I-PpoI
9	1f9fA	1jj4A		Papillomavirus E2
10	1fc3A	1lq1A		Spo0A
11	1g6nA	1zreA	138-206	Catabolite gene activator
12	1gv2A	1h8aC	89-143	c-Myb
13	1gvjA	1k7aA	333-436	ETS-1
14	1gxqA	1gxpA		PhoB
15	1h56A	1f0oA		Endonuclease PvuII
16	1hom_	9antA	005-060	Antennapedia homeodomain
17	1iknA	1vkxA	19-191	NF-kappa B
18	1irqA	2bnwA		Repressor omega
19	1j0rA	1f4kA		Replication terminator protein
20	1jbgA	1r8dA		MerR
21	1jtxB	1jt0A	002-072	QacR
22	1mjkA	1mjmA		MetJ
23	1mn4A	2etwA		Ndt80
24	1okrA	1saxA		Methicillin repressor MecI
25	1or7A	2h27A	123-187	Group IV sigma factor
26	1pra_	1rpeL	001-063	bacteriophage 434 repressor
27	1pyc_	1hwtC	60-97	Hap1
28	1r05A	1nlwB		Max
29	1rveA	1eopA		Endonuclease EcoRV
30	1rxr_	1by4A		Retinoid X receptor
31	1sdoA	1vrrA	1-203	Endonuclease BstYI
32	1tfb_	1c9bA	111-207	Transcription factor IIB
33	1vf9A	1w0uA	446-500	Telomeric protein TRF2
34	1vhiA	1b3tA		Epstein-Barr nuclear antigen 1
35	1vokA	1qnbA	16-115	TATA box-binding protein
36	1wpkA	1u8bA	009-076	Ada
37	1wtdA	1wteA		Endonuclease Eco0109I
38	1xwrA	1zs4A		Lambda cII
39	1z91A	1z9cA		OhrR
40	2audA	1tx3A		Endonuclease HincII

41	2cpgA	1b01A		CopG
42	2jcgA	1rzrG	003-060	Catabolite Control Protein A
43	2tdx_	1ddnA		Diphtheria toxin repressor

Class I: no conformational change (C^{α} displacement less than 0.5 Å)

S.No.	APO	HOLO	Protein Description
1	153l	154l	Goose lysozyme (trisaccharide)
2	1a0s P	1a0t P	Sucrose-specific porin (sucrose)
3	1a7u A	1a8u A	Chloroperoxidase T (benzoate)
4	1afd	1afa 1	C-type animal lectins (galactose)
5	1fgk A	1agw A	Tyrosine kinase domain of fibroblast growth factor receptor 1 (su4984 inhibitor)
6	1ahc	1aha	α-Momorcharin (adenine bitter melon)
7	1aqh	1aqm	α-Amylase (TRIS)
8	1yme	1arm	Carboxypeptidase (tromethamine)
9	2hhm A	1awb A	Human myo-inositol monophosphatase (D-inositol-1-phosphate and calcium)
10	1b49 A	1b5d A	DCMP hydroxymethylase from T4 (DCM 2′-deoxycytidine-5′-monophosphate)
11	1bjz	1bj0	Tetracycline chelated (CTC 7-chlorotetracycline)
12	1psj	1bk9	Phospholipase A2 (P-bromo-phenacyl-bromide, 1,4-butanediol)
13	1bmz A	1bm7 A	Transthyretin (flufenamic acid)
14	1rtc A	1br6 A	Ricin A chain (pteroic acid)
15	2blg A	1bso A	Bovine β-lactoglobulin (12-bromododecanoic acid)
16	3app A	1bxq A	Acid proteinase (phosphonate inhibitor)
17	1yer A	1byq A	Hsp90 N-terminal (ADP-Mg)
18	2chs A	1com A	Monofunctional chorismate mutase (prephenate)
19	1dco A	1dcp A	Bifunctional protein-binding transcriptional coactivator DCOH (biopterin)
20	1xla A	1did A	D-Xylose isomerase (2,5-dideoxy-2,5-imino-D-*glucitol)
21	2sil	1dil	Sialidase from salmonella typhimurium (apana and epana inhibitors)
22	1dmx	1dmy	Murine mitochondrial carbonic anyhdrase V (acetazolamide)
23	1dun	1duc	EIAV dutpase (DUDP strontium)
24	1dup	1dud	Deoxyuridine 5′-triphosphate nucleotide hydrolase (deoxyuridine 5′- diphosphate)
25	1eur	1eus	Bacterial sialidase (2-deoxy-2,3-dehydro-N- acetylneuraminic acid)
26	1gmq A	1gmp A	Ribonuclease (2′-GMP)
27	1dea A	1hor A	Glucosamine 6-phosphate deaminase (2-deoxy-2-amino glucitol-6-phosphate)
28	2hvm	1hvq	Hevamine (N-acetyl-D-glucosamine)

29	1ifb	1icm	Rat intestinal fatty acid binding protein (myristate)
30	1ped A	1kev A	NADP-dependent alcohol dehydrogenase (dihydro-nicotinamide-adenine-dinucleotide)
31	1loe A	1log A	Legume, isolectin I (α-D-mannose, N-acetyl-D-glucosamine)
32	1tfa A	1nft A	Ovotransferrin (nitrilotriacetic acid)
33	1png	1pnf	Pngase F (Di-N-acetylchitobiose)
34	1pnk AB	1pnl AB	Penicillin acylase (2-phenylacetic acid)
35	1aqp	1rca	Ribonuclease A (2′-deoxycytidine-2′-deoxyguanosine-3′,5′-monophosphate)
36	2alp	1tal	α-Lytic protease (TRIS (hydroxyethyl) amino methane)
37	1vpn A	1vps A	Polyomavirus vp1 pentamer (disialylated hexasaccharide)
38	1xza	1xzb	Fusarium solani cutinase (mercury acetate)
39	1ena	2enb	Staphylococcal nuclease (thymidine 3′,5′-diphosphate)
40	1ag1 B	4tim B	Trypanosomal triosephosphate isomerase (2-phosphoglycerate)
41	3enl	5enl	Enolase (2-phospho-D-glyceric acid and calcium)
Class II: moderate conformational change (greater than or equal to 0.5 Å but less than or equal to 2.0 Å)			
S.No.	**APO**	**HOLO**	**Protein Description**
1	3icd	1ai2	Isocitrate dehydrogenase (isocitrate, NADP+ and calcium)
2	1fus	1fut	Ribonuclease F1 of Fusarium moniliforme (guanosine-2′-monophosphate)
3	2gd1 O	1gd1 O	Glyceraldehyde-3-phosphate dehydrogenase (nicotinamide adenine dinucleotide)
4	1amp	1igb	Aeromonas proteolytica aminopeptidase (para-iodo-D-phenylalanine hydroxamate)
5	1qpo A	1qpq A	Quinolinic acid phosphoribosyltransferase (quinolinic acid)
6	5dfr	1ra1	Dihydrofolate reductase (nicotinamide adenine dinucleotide phosphate)
7	1bhj A	1xva A	Methyltransferase (S-adenosylmethionine)
8	2paw	1a26	Catalytic fragment of poly(adp-ribose) polymerase (carba-nicotinamide-adenine-dinucleotide)
9	1alb	1adl	Adipocyte lipid-binding protein (arachidonic acid)
10	1ajz	1ajo	E. coli dihydropteroate synthase (dihydropterine, sulfanilamide)
11	1alv A	1alw A	Domain VI of porcine calpain (3-(4-iodophenyl)-2-mercapto-(z)-2-propenoic acid)
12	2ptn	1aq7	Trypsin (inhibitor aeruginosin 98-B)
13	1bd9 A	1beh A	Human phosphatidylethanolamine binding protein (cacodylate)
14	1trz C	1ben C	Insulin (4-hydroxybenzamide)
15	1sgk	1ddt	Diphtheria toxin (adenylyl 3′-5′ uridine 3′ monophosphate)
16	1epa A	1epb A	Epididymal retinoic acid binding protein (retinoic acid)
17	2fgf	1fga	Human basic fibroblast growth factor (β-mercaptoethanol, selenate)
18	1fkk	1fkl	FKBP 12-rapaymycin (rapamycin immunosuppressant drug)
19	1gcg	1gca	Periplasmic glucose/galactose receptor (galactose)

20	1jda	1jdc	Maltotetraose-forming exo-amylase (maltotetraose)
21	135l	1jef	Turkey lysozyme (GlcNac-3)
22	1igs	1ju1	Indole-3-glycerolphosphate synthase (N-(ethyl sulfite)morpholine)
23	1kem LH	1kel LH	Fab fragment (hapten 1-[n-4′-nitrobenzyl-n-4′-carboxybutylaminomethylphosphonic acid]
24	1lts D	1ltt D	Heat-labile enterotoxin (lactose)
25	1mjk A	1mjl A	Methionine repressor (S-adenosyl methionine)
26	1mzl	1mzm	Maize nonspecific lipid transfer protein (palmitate)
27	2nck R	1nhk R	Nucleoside diphosphate kinase (5′-cyclic adenosine monophosphate)
28	2pgd	1pgn	6-Phosphogluconate dehydrogenase(nicotinamide 8-bromo-adenine dinucleotide phosphate)
29	1ptq	1ptr	Protein kinase C δ cys2 domain (phorbol-13-acetate)
30	1sry A	1ses A	Seryl-tRNA synthetase (seryl adenylate)
31	1ubv	1ubw	Farnesyl pyrophosphate synthetase (geranyl diphosphate)
32	2izd B	2izf B	Streptavidin-biotin (biotin)
33	2pcd	3pca A	Protocatechuate 3,4-dioxygenase (3,4-dihydroxybenzoate)
34	4cha A	6cha A	α-Chymotrypsin(phenylethane boronic acid)
35	1bkz A	2gal A	Human galectin-7 (galactose)
Class III: large conformational change (greater than 2.0 Å)			
S.No.	**APO**	**HOLO**	**Protein Description**
1	1a9o	1a9p	Bovine purine nucleoside phosphorylase (9-deazainosine and phosphate)
2	2rcs LH	1Aj7 LH	Immunoglobulin 48g7 germline fab (hapten 5-(para-nitrophenyl phosphonate)-pentanoic acid)
3	4ake A	1ake A	Adenylate kinase (PA,PE-bis(adenosine-5′-)pentaphosphate)
4	1omp	1anf	Maltodextrin binding protein (maltose)
5	1brq	1brp	Human plasma retinol-binding protein (retinol)
6	1bya	1byb	Soybean β-amylase (β-maltose and maltal)
7	1ceo	1cen	Cellulase (cellohexaose)
8	1ebh A	1ebg A	Enolase (phosphonoacetohydroxamate)
9	1xaa	1hex	3-Isopropylmalate dehydrogenase (β-nicotinamide adenine dinucleotide)
10	1hsi	1hii	HIV-2 proteases (CGP 53820)
11	1lz4	1hnl	Glutathionylated human lysozyme (glutathione)
12	4tms	1lca	Thymidylate synthase (DUMP and cb3717)
13	3ins AB	1mpj AB	Insulins (phenol)
14	1fgx A	1o0r A	β1,4-Galactosyltransferase (UDP-galactose)
15	1swa A	1swd A	Streptavidin (biotin)
16	1rtj	1vrt A	HIV-1 RT (dipyridodiazepinone nevirapine)
17	2bgt	2bgu	β-Glucosyltransferase (substrate uridine diphosphoglucose)

18	2chs A	2cht A	Monofunctional chorismate mutase (endo-oxabicyclic inhibitor)
19	1bkz A	3gal A	Human galectin-7 (galactosamine)
20	1fua	4fua	L-Fuculose-1-phosphate aldolase (phosphoglycolohydroxamic acid)
21	1enq	5cna A	Concanavalin-A (α-methyl-D-mannopyranoside)
22	1ypi A	7tim A	Yeast triosephosphate isomerase (phosphoglycolohydroxamate)

** each entry is provided with the four-digit PDB codes and the protein chain identifiers, the residue range of the domain if not the whole chain and the description of protein with ligand name in parenthesis.*

2.2.4. Assignment of secondary structure:

For each amino acid in a protein in the above data sets, the secondary structure was assigned using **DSSP** software (**Kabsch and Sander, 1983**). We have considered all the seven-state assignments of **DSSP** (G, H, I, B, E, T and S) with minor modification. Loops (Z) were assigned to regions not assigned among these seven. T was retained only if it occurred as a sole structure between helices/strands, otherwise it was included as part of loop region. Bends (S) were grouped under loops (Z). Furthermore, we classified loops into two classes: terminal loops (Z_T), that end at N or C terminal of the polypeptide chain and loops that occur between secondary structures (Z_B).

2.2.5. MSRP calculation from Ramachandran Map:

For all the above sets of proteins the Ramachandran map (**Ramachandran *et al.*, 1963**) was plotted using the **Swiss-PDB viewer** (v.3.7). The clustering of points (dihedral angle pairs) among consecutive residues in the Ramachandran plot was quantified by MSRP parameter for the segregation of loops in different classes of proteins. **MSRP** for a given region was measured from the **M**ean **S**eparation between points of consecutive residues in the **R**amachandran **P**lot (**MSRP**) (see **Figure 2.1**). In each protein set, for each protein, the mean distance (**MSRP**) between points of consecutive residues in each secondary structure and its associated standard deviation (σ) for that mean amongst all pairs of consecutive points in that structure was computed from the Ramachandran map. **Figures 2.2** A & B illustrate the **MSRP** and its associated σ for two proteins of identical polypeptide chain length but different Ramachandran map profiles. In case of proteins with NMR structures, **MSRP** was calculated for all the ensembles and its average value was displayed with an error bar (standard deviation) highlighting the variation of **MSRP** among different structures in the ensemble. Similarly an average value for σ was also

calculated from the individual σ of different NMR structures. For APO/HOLO protein pairs, MSRP was calculated for the loop regions of APO proteins first. Later, MSRP of identical regions in the HOLO proteins were calculated for comparison with APO proteins.

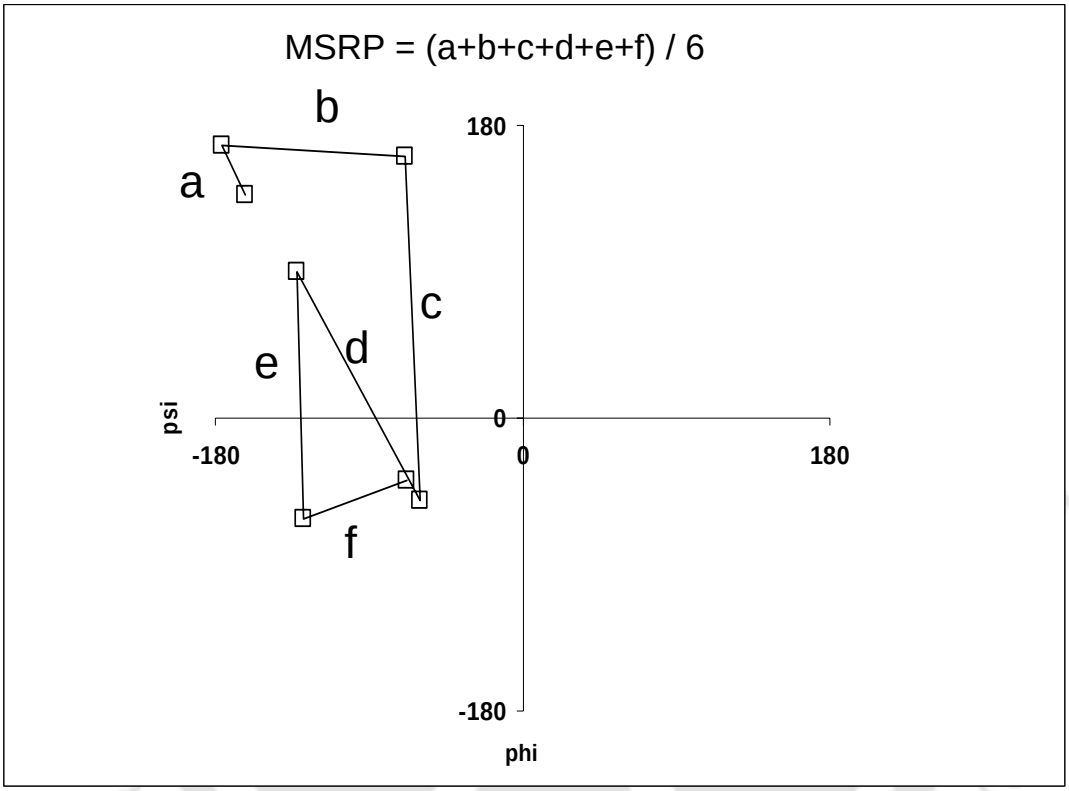

Figure 2.1: *Calculation of MSRP from Ramachandran Map*

2.2.6. Molecular Dynamics Simulations:

The initial structure for the molecular dynamics simulation of proteins (see **Table 2.8**) with PDB codes 1BGF, 1MUN, 2HDL (NMR, Model 1), 2SOB (NMR Model 10), 1LXL (NMR minimized average structure), 1VZS (NMR Model 1) were downloaded from PDB. The LEaP module of the AMBER program package (**Pearlman *et al.*, 1995; Case *et al.*, 2005**) was used to prepare the system for simulation. Each protein was solvated with TIP3P (**Mahoney and Jorgensen, 2000**) waters and neutralized with the counter ions using the LEaP module. Energy minimization and MD simulations were carried out using the SANDER module of AMBER 8. The Amber force field ff99SB (**Hornak *et al.*, 2006, Wickstrom *et al.*, 2009**) is used to describe the atomic interactions.

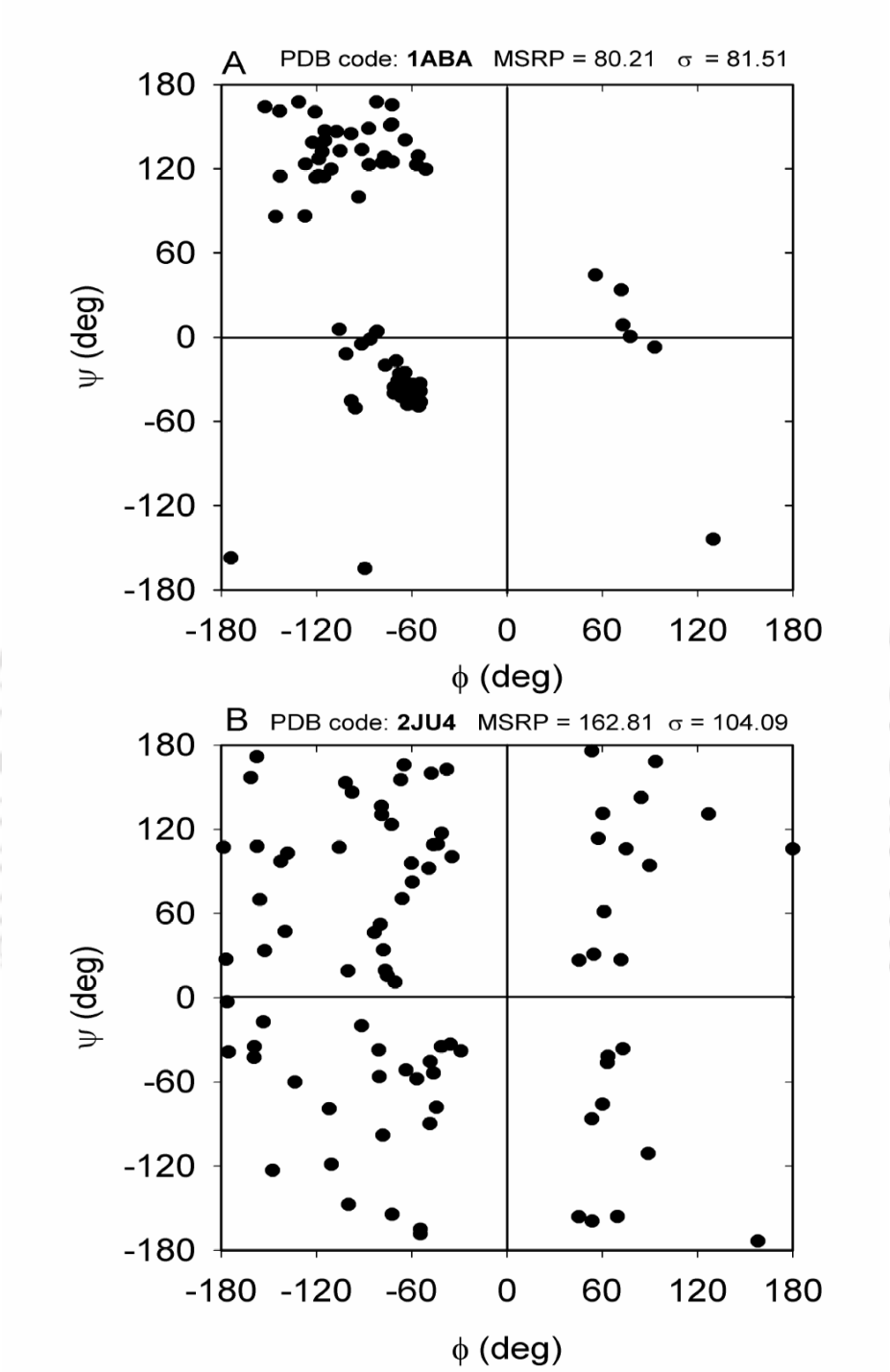

Figure 2.2: *Ramachandran Plot for (A), oxidized bacteriophage T4 glutaredoxin (PDB code: 1ABA) and (B), rhodopsin-sensitive GMP-PDE gamma-subunit (PDB code: 2JU4 model 94/100 of NMR ensemble) is shown. The MSRP values for the proteins are 80.2 ± 81.5 and 162.8 ± 104.1, respectively, while the polypeptide chain lengths for both are 87 residues.*

For the correct treatment of long range electrostatics, we make use of the Particle Mesh Ewald (PME) method (**Essmann *et al.*, 1995**). Constant temperature and pressure conditions in the simulations were achieved by coupling the system to a Berendsen's thermostat and barostat (**Berendsen *et al.*, 1984**). Bonds involving the hydrogen atoms were constrained to their equilibrium position with the SHAKE algorithm. For these simulations, we used an HP Proliant server with 8 processors. The system was minimized in two phases to avoid bad contacts. In the first phase, the system was minimized giving restraints (30 kcal/mol/Å^2) to protein and crystallographic waters for 500 steps with subsequent second phase minimization of the whole system. Then the system was heated to 300 K over 50 ps with a 1 fs time step. The protein atoms were restrained with force constant of 30 kcal/mol/Å^2 at the NVT ensemble. After that the force constant was reduced by 10 kcal/mol/ Å^2 in each step to reach the unrestrained structure in three steps of 10 ps each. The system was then switched over to the NPT ensemble and equilibrated without any restraints for 180 ps. The system was equilibrated in total of 260 ps. The time step for MD simulation for the production run was 2 fs. All the six trajectories were each run for 10 ns and were performed with an 8.0 Å cutoff on real-space interactions. Analysis of parameters of the trajectories was carried out using the ptraj modules of AMBER 8. Graphic visualization of protein structures was done using Chimera. For MSRP calculations shown in **Figure 2.7** and **Table 2.8**, snapshots at every 100 ps interval were taken.

2.3. Results:

We calculated the MSRP values for the sets of ordered proteins/regions listed in **Tables 2.1-2.4** using their Ramachandran maps. The MSRP values are plotted in **Figure 2.3** A, B, C, and D for classes a, b, c, and d, respectively. A detailed statistical summary of **Figure 2.3** is tabulated in the **Table 2.7**. Among all the classes of proteins, it is observed that the α-helices are tightly clustered in a narrow range of MSRP values, averaging ~13-17. The 3_{10} helices show a slightly higher MSRP value among all classes in **Figure 2.3**, averaging ~30-41, but statistically their population is considerably small in comparison to α-helices. In contrast to α-helices, the extended strand has a higher MSRP value and is more scattered with a large spectrum of MSRP values, averaging ~47-63

among all classes. The isolated β-bridges are few in number, averaging ~117-135.

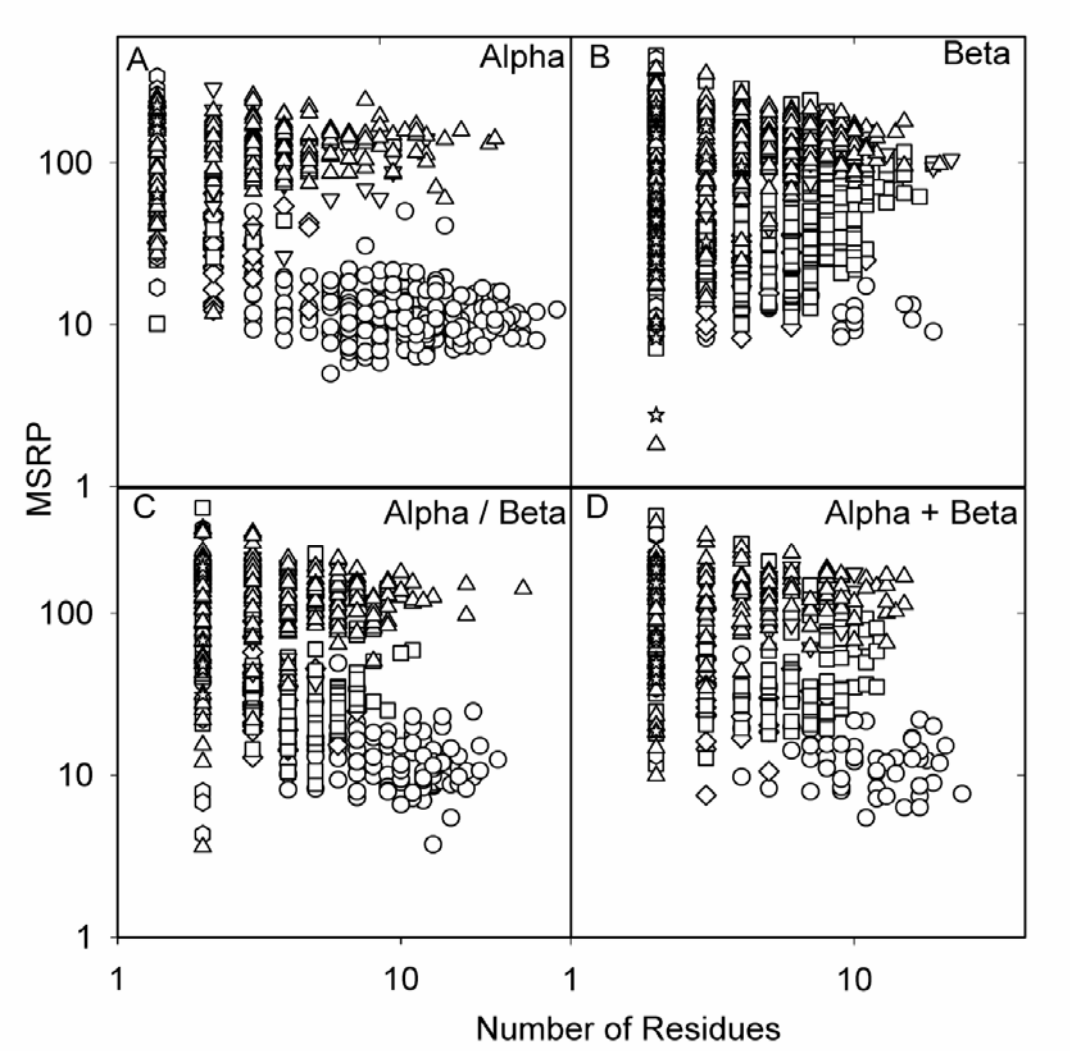

Figure 2.3: *Plots of MSRP against the number of residues in different secondary structure regions. Plot A, B, C and D correspond to datasets from all-α proteins, all-β proteins, α / β and α + β proteins respectively. See* ***Tables 2.1-2.4*** *for details. The X axis has a maximum value of 40. The symbols are as follows: circle, H (α-helix); diamond, G (3_{10}-helix); square, E (Extended strand); hexagon, B (Residue in isolated β–bridge); inverted triangle, Z_T (Terminal Loop); triangle, Z_B (Loop between secondary structures) and star, T (Turn).*

Loops that are in between secondary structures predominantly possess a high MSRP value, averaging ~132-142. But it is noticeable that these loops show a large variation in MSRP values. On the other hand, the loop regions in the terminal positions display MSRP values a shade lower than loops between secondary structures, averaging ~113-

137. Finally, β-turns which connect secondary structures show largest variations amongst different classes, displaying the lowest MSRP value for class b proteins.

It is now clear that loop regions possess a higher MSRP values in comparison to regular secondary structures like helices and strands and thus can be segregated from them. It can be seen that there is large distribution among MSRP values in the case of beta bridges, extended strands and among loops in contrast to helices. This perhaps suggests the structural heterogeneity in these secondary structures other than helices. Among different classes of proteins, the difference in MSRP for a given secondary structure appears small with exception of β-turn. It is also observed that the MSRP values do not differ appreciably for loops in between secondary structures and terminal loops.

Table 2.7: *Mean MSRP values with standard deviation (number of samples) for different secondary structures of different classes of proteins*

Type of Sec. structure	Class a, All Alpha Proteins	Class b, All Beta Proteins	Class c, All Alpha/Beta Proteins	Class d, All Alpha + Beta Proteins	Unstructured Proteins
H (Alpha helix)	13 ± 10 (234)	17 ± 6 (41)	14 ± 7 (112)	16 ± 9 (57)	25 ± 21 (35)
G (3-helix)	30 ± 13 (35)	36 ± 24 (68)	33 ± 16 (40)	41 ± 29 (32)	48 ± 26 (10)
E (Extended strand)	48 ± 27 (14)	63 ± 57 (418)	47 ± 50 (141)	57 ± 53 (115)	94 ± 68 (22)
B (Residue in beta bridge)	135 ± 102 (21)	125 ± 112 (75)	117 ± 90 (31)	119 ± 85 (28)	121 ± 66 (6)
Z_T (Terminal Loop)	113 ± 51 (42)	130 ± 55 (43)	137 ± 68 (20)	133 ± 51 (17)	144 ± 37 (34)
Z_B (Loop between sec. str.)	132 ± 48 (231)	142 ± 63 (448)	142 ± 56 (267)	133 ± 58 (162)	145 ± 47 (51)
T (Hydrogen bonded turn)	157 ± 68 (4)	84 ± 69 (52)	111 ± 77 (9)	123 ± 72 (18)	51 ± 0 (1)

In **Figure 2.4** A, B, C, D, the calculated MSRP is plotted against its own standard deviation for classes a, b, c, and d respectively. The standard deviation (σ) of the MSRP would point out the irregularity in the distance between points in the Ramachandran map or the uncertainty in MSRP. For example, a small MSRP with a low σ would indicate minimal variations in dihedral angles. We observe this mostly with regular secondary structure regions like helices/strands (**Figure 2.4**), where the MSRP is tightly proportional to σ. In general MSRP is expected to be fairly proportional to σ for a regular structure as seen for strands and a minor fraction of loops. Among several other loops in the bottom right side of the plots, this correlation is considerably poor due to perhaps

largely irregular structure. It is apparent that σ does not increase in proportion to the MSRP here, suggesting that the jumps in the Ramachandran map are large and fairly uniform in distance. Obviously, in this case, the points in the Ramachandran map are farther apart (high MSRP) and more uniformly separated (low σ). Thus σ << MSRP may bear the telltale signature of irregularly structured region.

Another feature that is eye-catching in **Figure 2.4** A, B, C, D, is a clear separation between the tightly populated clusters in the bottom left corner and the hugely scattered population further up diagonally on the other side of the divide, as indicated by a dashed line in **Figure 2.4** A, B, C, D**.** This separation is evident for all classes of proteins at more or less similar positions. Perhaps, this separation suggests that loops and a few strands always possess a lot of inconsistency (higher σ) in their dihedral angles. Thus loops are characterized by not only a higher MSRP as revealed in **Figure 2.3**, but also both higher and lower standard deviation as displayed in **Figure 2.4**. It would be interesting to investigate and analyze the structural characteristics of loops in relation with their MSRP and σ.

So far, we have concentrated on ordered protein structures derived from protein crystals. Next we look at unstructured proteins (**Table 2.5**), whose structures have been determined using NMR. **Figure 2.5** A shows the profile observed for several secondary structures, while their statistical summary is shown in the **Table 2.7**. Among unstructured proteins, it is observed that MSRP values for the individual secondary structures is marginally higher, more so with strands. The trend observed in MSRP among different secondary structures is some what similar to proteins in **Figure 2.3**.

Figure 2.5 B shows that similar to **Figure 2.4** ABCD, the loops and several strands exhibit high MSRP and σ, with a fair correlation, while the helices appear confined to the bottom left corner. In contrast to **Figure 2.4**, we observe a substantial population of points in the middle right region of the plot. These are loops that possess high MSRP but not so high σ, presumably because they are unstructured proteins. The error bars in MSRP clearly indicate a significant degree of flexibility in these structures. Thus apparently secondary structures in unstructured proteins appear a shade more irregular in comparison to folded globular proteins (**Figure 2.3**).

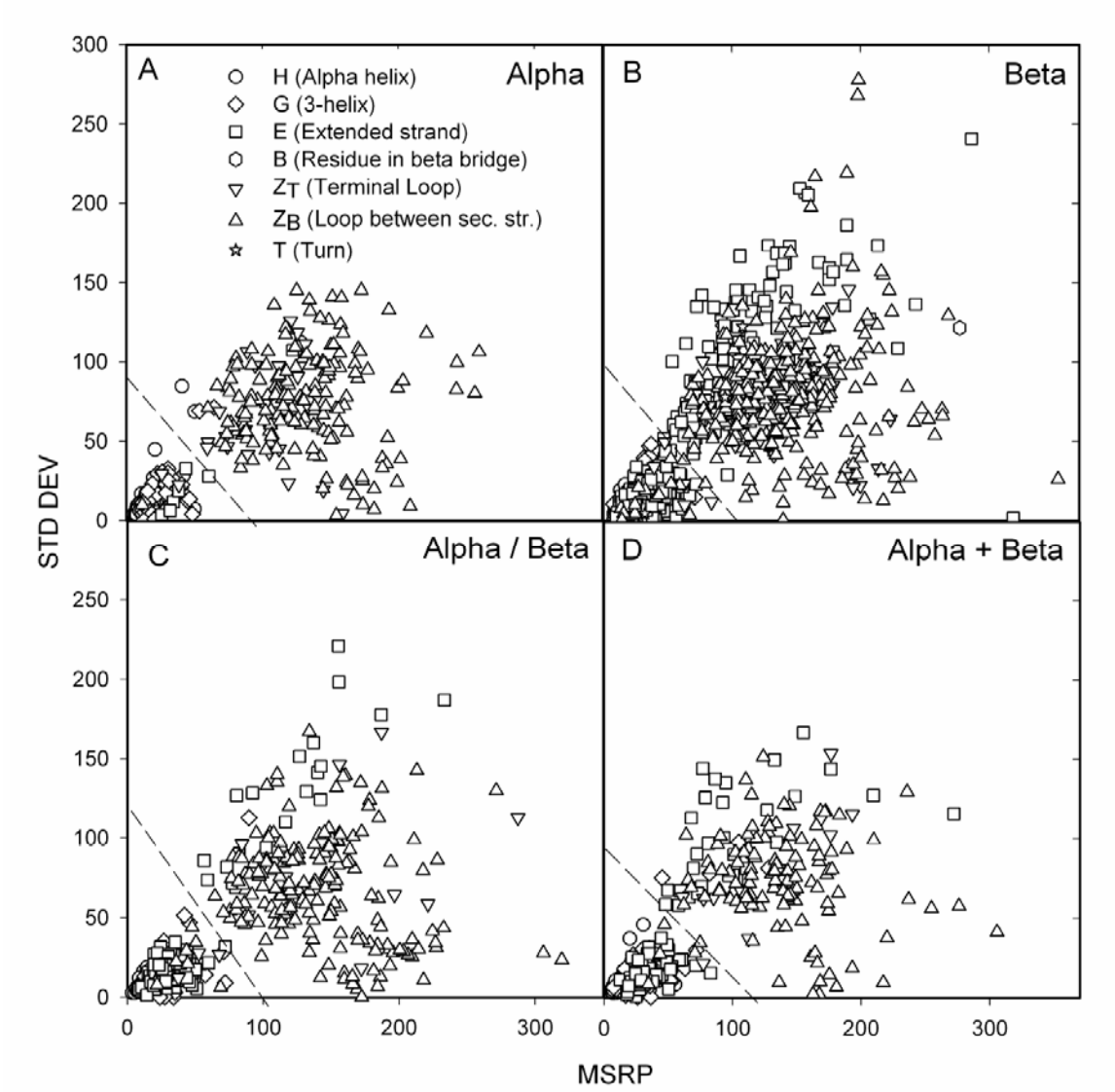

Figure 2.4: *Plots of MSRP against its standard deviation (σ) is shown for different secondary structure regions of different classes of proteins. Plot A, B, C and D correspond to All-α proteins, All-β proteins, α / β and α + β proteins respectively. See* ***Tables 2.1-2.4*** *for details.*

Loops play very important role in protein function (**Kempner, 1993**) and these regions undergo significant changes in their structure while binding to ligands or DNA in order to accommodate the bound molecule inside the protein in an energetically favorable way. It would be valuable and interesting to investigate how the MSRP of the loop changes after it binds to the ligand. The MSRP values of loops in APO and HOLO forms for several DNA binding proteins (**Table 2.6**) are shown in **Figure 2.6** A. We observe that a major population of points lie along the diagonal. This indicates that in these loops

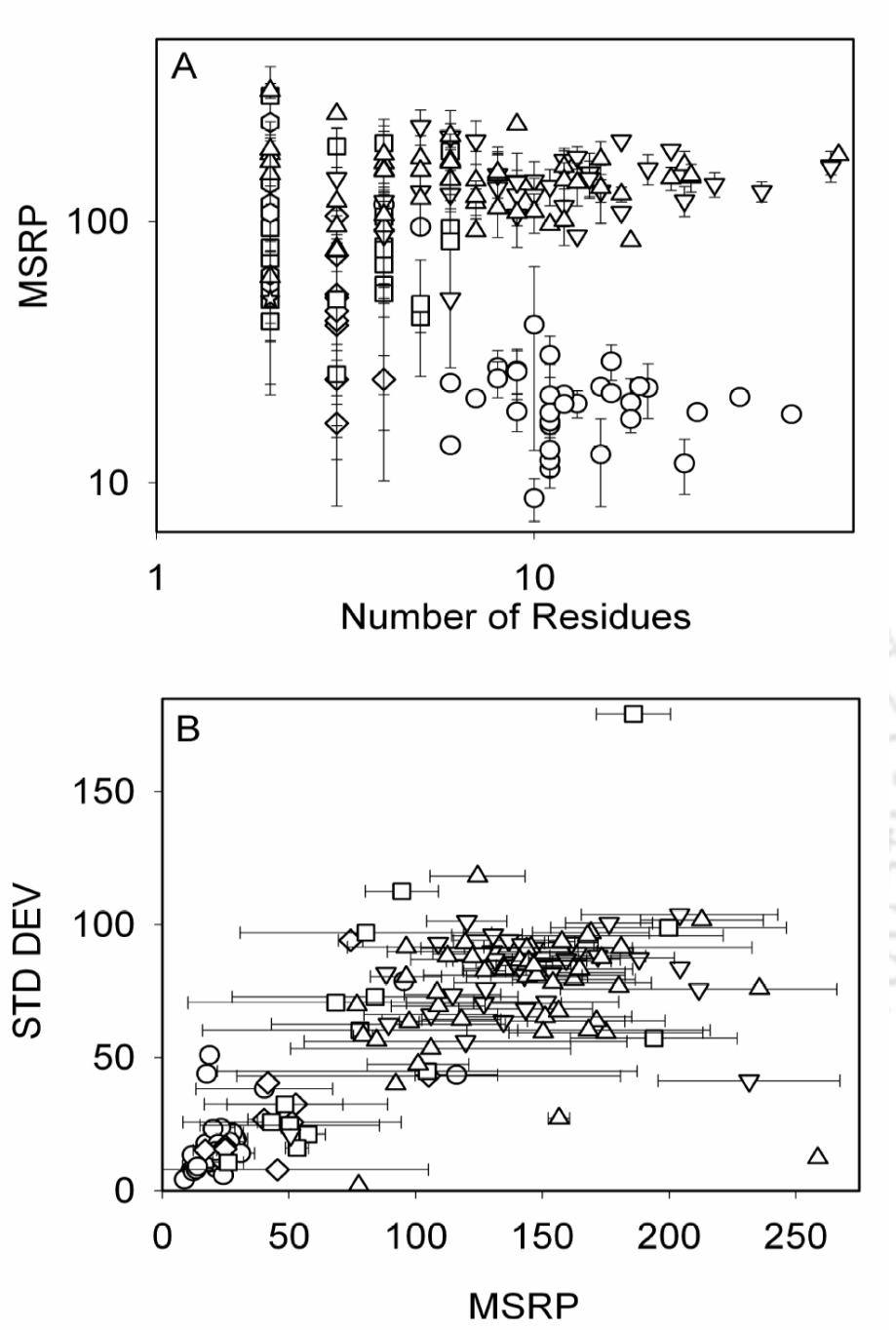

Figure 2.5: *Plots of MSRP against the number of residues in the secondary structure regions and its standard deviation (σ) is shown for unstructured proteins in A and B, respectively. See* ***Table 2.5*** *for details. The error bars in A & B show the range of values observed in MSRP among the NMR ensembles of the same protein. See legend for* ***Figure 2.3*** *for information on symbols.*

the MSRP values have not changed significantly. It is also clear that among a major fraction of these points, the MSRP values lie in the range of ~100-200, consistent with data in **Figure 2.3 & 2.5**.

Thus these loops are not much affected in structure and dynamics subsequent to binding. Points above the diagonal represent loops that had significantly higher MSRP before binding of DNA, but which dipped markedly after binding. We do observe few points that undergo a major dip in their MSRP subsequent to binding, in the upper left corner of the plot. Similarly, one can also observe points to the right of the diagonal, which identify loops that possessed lower MSRP prior to binding DNA, but which has undergone an increase in MSRP after binding.

Statistically, it was determined that 11% of points were to the right of the diagonal (i.e. APO: HOLO < 0.8), while 28% of points were above the diagonal (APO: HOLO > 1.2). It is interesting to observe that relatively less number points to the right of the diagonal in comparison to points above the diagonal. Then we looked into the variation in MSRP among three classes of proteins in their APO/HOLO forms (**Table 2.6**). This is depicted in **Figure 2.6** B. In the case of class I, where C^{α} displacement was < 0.5 Å, we observe that a majority of points lie along the diagonal, as expected.

It was observed that 4% of total points exist above the diagonal (APO: HOLO > 1.2), while 3% of points exist to the right of the diagonal (APO: HOLO < 0.8). In class II, where C^{α} displacement was 0.5-2.0 Å, we observe 7% of points above the diagonal and 5% to its right, while in class III, where C^{α} displacement was > 2.0 Å, we observe 12% of points above diagonal and 9% to the right of diagonal. In all the classes, we have a significant fraction of loops that remain unchanged in their MSRP values subsequent to binding. But it is interesting to observe that the increase in fraction of loops that deviate from the diagonal roughly follows the order of class III > class II > class I, in close correlation with the C^{α} displacements. The above results explicitly establish that deviations from the diagonal in **Figure 2.6** are clearly correlated with structural changes happening in the loops of APO proteins subsequent to DNA/ligand binding. In comparison to ligand binding loops, it appears that DNA binding protein loops undergo far more structural perturbations, judging by the statistical values above.

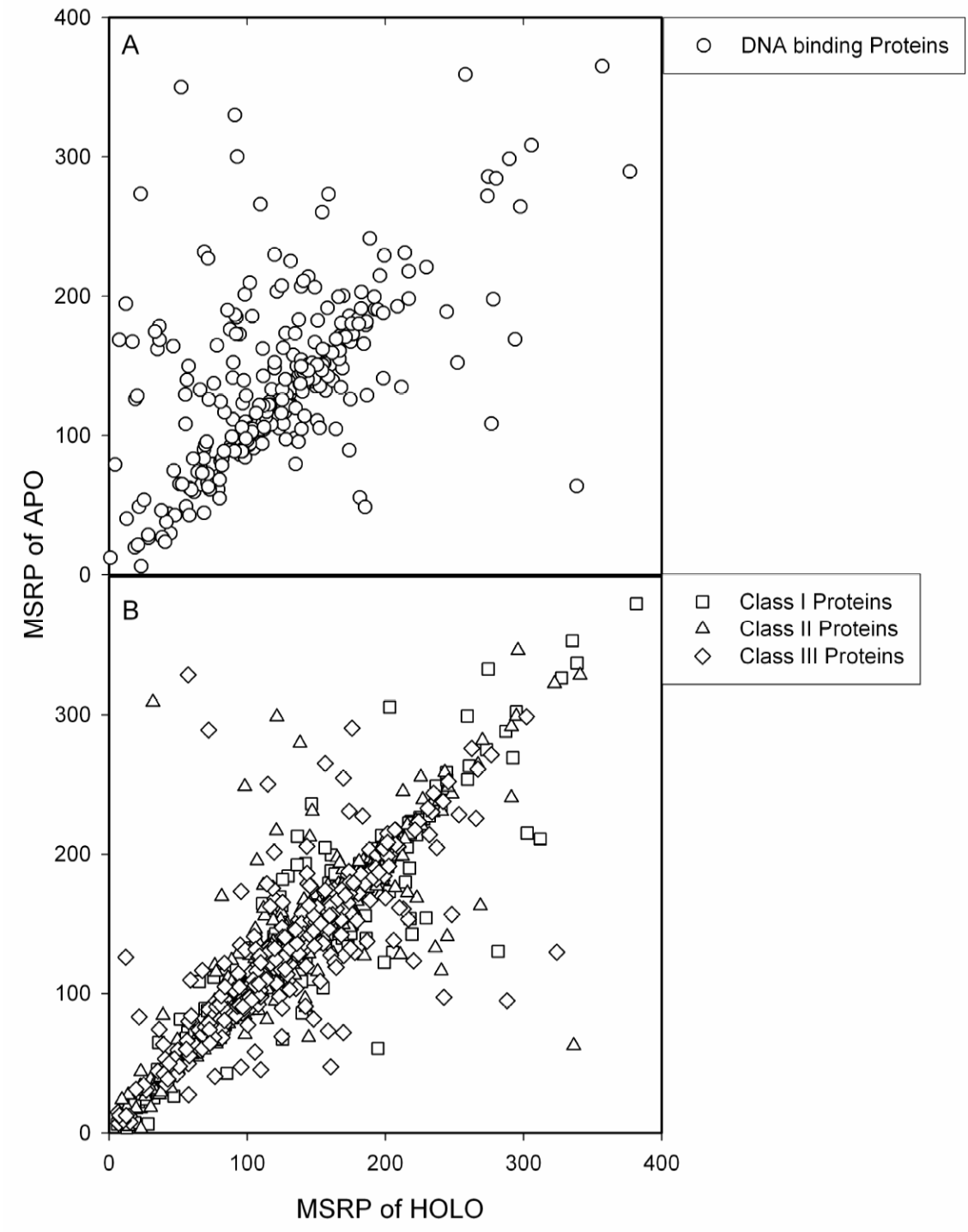

Figure 2.6: *MSRP of APO protein is plotted against HOLO protein for different protein datasets. Plot A shows loops in DNA binding proteins as circles. Plot B reveals loops in: class I proteins (C^{α} displacement < 0.5 Å) as squares, class II proteins (C^{α} displacement 0.5—2.0 Å) as triangles and class III proteins (C^{α} displacement > 2.0 Å) as diamonds. See* ***Table 2.6*** *for details on proteins.*

Finally, it would be worthwhile to investigate how sensitive the MSRP parameter is to the molecular dynamics of the protein. It has been suggested that unstructured regions are dynamic ensembles where atomic coordinates and dihedral angles oscillate widely and randomly with time, having no equilibrium values (**Radivojac *et al.*, 2007**). It would be interesting to compare changes in MSRP during a molecular dynamics simulation for normally folded and unstructured proteins. Specifically, the objective for such an endeavor is twofold: A) to monitor MSRP fluctuations with time among different secondary structure regions in the same protein and subsequently compare them between a regularly folded protein and an unstructured protein and B) To compare MSRP fluctuations across the entire folded protein against entire unstructured protein and unstructured region alone. These investigations are necessary to confirm if temporal oscillations in MSRP can be correlated with structurally disordered regions in proteins.

Previous work has shown that proteins with PDB codes *1BGF* (**Vinkemeier *et al.*, 1998**) and *1MUN* (**Guan *et al.*, 1998**) are ordered, while *2HDL* (**Peterson *et al.*, 2006**), *2SOB* (**Alexandrescu *et al.*, 1995**), *1LXL* (**Muchmore *et al.*, 1996**) & *1VZS* (**Carbajo *et al.*, 2004**) are predominantly disordered proteins. We first observed changes in MSRP for the whole protein. The average MSRP and its associated standard deviation (σ_{MD}) over a 10 ns period are tabulated in **Table 2.8**, while its trajectory is shown in **Figure 2.7**. For the ordered proteins *1BGF* and *1MUN*, the MSRP maintains a fairly steady profile with time. This is reflected in their low σ_{MD} values too. The protein *2HDL*, which is partially disordered, shows large amplitude fluctuations in MSRP compared to ordered proteins above. This protein also displays the highest σ_{MD} among the lot. The partially disordered protein *2SOB* also displays large amplitude oscillations in MSRP, while the other IDP, *1VZS* show only moderate fluctuations in MSRP with time in the MD simulation trajectories. *1LXL* is a protein containing a disordered region between residues 28 & 82. The whole protein 1LXL shows low MSRP fluctuations in the MD simulation. However, the 55 residue disordered region taken separately (*1LXL**) clearly shows a significantly higher MSRP with large σ_{MD} in comparison to whole protein. This implies that MSRP fluctuations occur predominantly in the disordered regions.

Table 2.8: *Ordered and Unstructured proteins considered for molecular dynamics simulation*

PDB ID	Name of the Protein	Average MSRP	σ_{MD}*	Chain length
1BGF	STAT-4 N-Domain (ordered)	49	2.4	124
1MUN	Catalytic domain of MutY from *E coli* (ordered)	71	2.2	225
2HDL	Brak/CXCL14 (unstructured)	95	6.3	78
2SOB	Sub-domain of staphylococcal nuclease (unstructured)	108	5.7	103
1LXL	Apoptosis regulator Bcl-x_L (unstructured)	84	2.9	221
1LXL*	Apoptosis regulator Bcl-x_L (unstructured region 28-82 alone)	140	6.4	55
1VZS	F6 subunit of ATP synthase (unstructured)	87	4.1	76

** refers to statistics over 100 snapshots each separated by 100 ps in a 10 ns MD simulation trajectory*

Next we compare the σ_{MD} among different secondary structure regions for proteins with PDB codes *1BGF, 1MUN, 2SOB and 1LXL*. These results are summarized in **Table 2.9**. It is observed that higher MSRP values do not always possess higher σ_{MD} values. Thus the magnitude of MSRP may not always reflect the structural disorder in the region. When comparing σ_{MD} for two regions, it is important to ensure that number of residues in those regions is nearly same to rule out averaging effects. It is conspicuous that irrespective of whether it is a folded or unstructured protein, the loops in general and terminal loops in particular reveal higher σ_{MD} values in contrast to α-helices which consistently possess the lowest σ_{MD}. Other secondary structures like 3_{10} helix, extended strand, isolated β-bridge and turns display moderately high σ_{MD}. However, these data are of limited utility owing to their small sample number.

2.4. Discussion:

The conformational preferences of each residue in the protein backbone, as defined by the ϕ and ψ dihedral angles in the Ramachandran map is a well established method for characterizing protein structure (**Ramakrishnan and Ramachandran, 1965; Ramachandran *et al.*, 1966**). In the recent past, improved versions of Ramachandran maps have been proposed and adopted for structural validation (**Lovell *et al.*, 2003**). The spatial proximity between any two points of consecutive residues in the Ramachandran Map for an irregularly structured region is significantly more when compared with a regular secondary structure like a α-helix. In the case of former, points of consecutive residues (dihedral angles) are far apart in the Ramachandran map. Often successive points may lie in different quadrants in the plot.

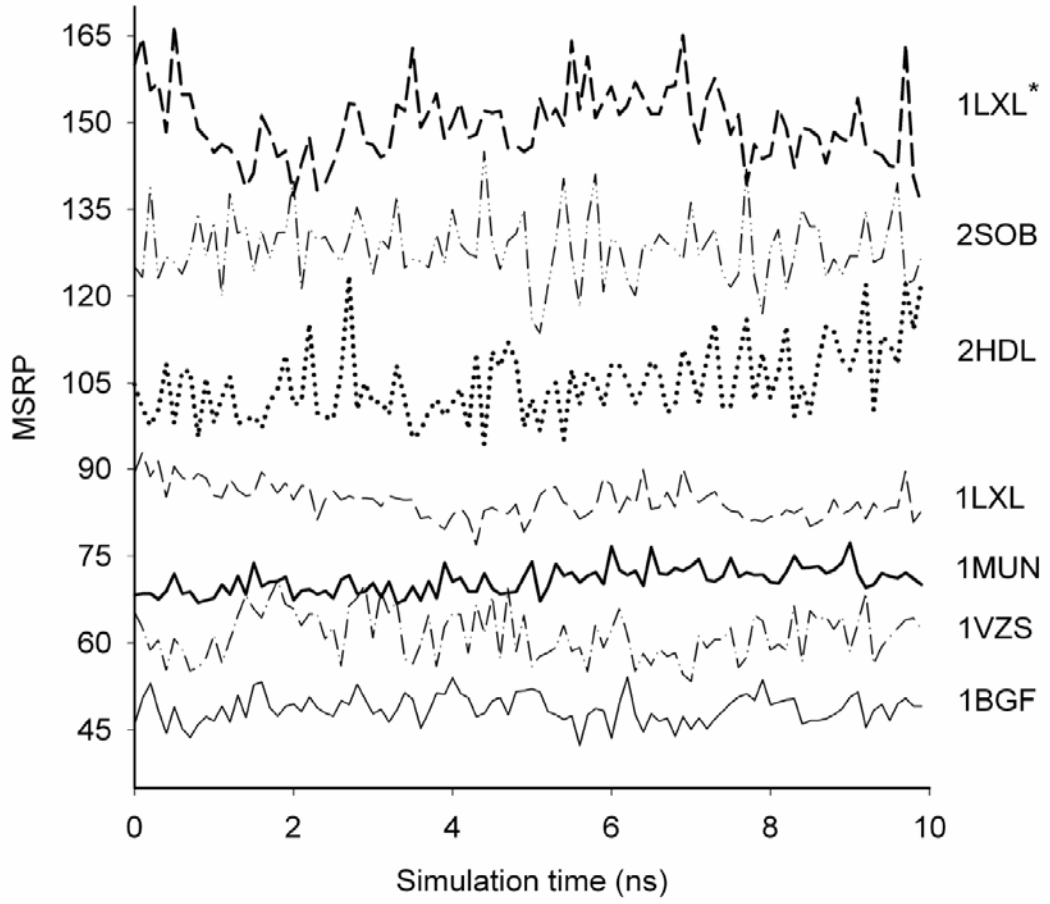

Figure 2.7: *Variation of MSRP along a 10 ns MD simulation trajectory is shown for ordered and unstructured proteins (see* ***Table 2.8*** *for details). The protein PDB codes are as follows: solid line, 1BGF; bold solid line, 1MUN; dotted line (shifted up by 10 MSRP units) 2HDL; dash dot dot line (shifted up by 20 MSRP units), 2SOB; dashed line, 1LXL; bold dashed line (shifted up by 10 MSRP units), 1LXL* disordered region alone and dash dot line (shifted down by 25 MSRP units), 1VZS. A few traces were shifted vertically to avoid overlap and enhance clarity.*

Table 2.9: *Mean MSRP* values among different secondary structures of ordered and disordered proteins from a 10 ns Molecular Dynamics simulation.*

Sec. Str.	Range		# Residues	Mean MSRP	σ_{MD}
1BGF (Ordered Protein)					
H	3	9	7	23	6.1
	15	17	3	33	10.3
	28	33	6	26	7.1
	35	40	6	39	10.8
	43	46	4	27	10.3
	50	74	25	20	2.6
	77	95	19	22	4.1

	98	118	21	23	4
G	12	14	3	22	9.1
	18	21	4	48	11
B	23	24	2	128	65.7
	26	27	2	104	26.2
Z_B	10	11	2	51	69.6
	22	23	2	92	60.2
	27	28	2	187	14.9
	41	42	2	97	27.4
	47	49	3	93	14.8
	75	76	2	148	14.4
	96	97	2	227	42.3
Z_T	0	2	3	229	132.1
	119	123	5	61	34.2
T	24	25	2	54	29.2
1MUN (Ordered Protein)					
H	3	17	15	24	3.7
	30	40	11	20	4.2
	45	68	24	27	6.6
	62	67	6	21	6.4
	70	77	8	22	5.1
	84	99	16	20	4.1
	108	112	5	30	9
	119	130	12	25	4.4
	139	148	10	20	4.6
	158	171	14	20	4.1
	177	190	14	27	5
	209	213	5	43	10.9
G	23	25	3	33	10.8
	199	201	3	34	10.6
	216	218	3	24	11.2
Z_B	18	22	5	92	12.3
	26	29	4	135	11.9
	41	44	4	97	8.5
	59	61	3	153	14.7
	68	69	2	21	15.6
	78	83	6	177	10.7
	100	107	8	118	9.9
	113	118	6	153	26.8
	131	138	8	105	18.5
	149	157	9	128	17.4
	172	176	5	163	16.4
	191	198	8	159	9.1
	202	208	7	119	36.6
	214	215	2	141	14.8
Z_T	219	225	7	57	43.7
2SOB (Disordered Protein)					
H	58	67	10	37	6.7
E	14	16	3	166	29.7
	23	26	4	31	7.8
	31	34	4	57	34.2

B	75	76	2	137	23.6
	91	92	2	25	13.3
Z_B	17	22	6	190	16.8
	27	30	4	205	46.5
	35	57	23	134	14.7
	68	74	7	110	26.7
	76	90	15	120	17.1
Z_T	1	13	13	86	27.6
	92	103	12	121	16.8
1LXL (Disordered Protein)					
H	6	20	15	20	3.7
	85	95	11	20	4.1
	105	110	6	21	5.8
	120	131	12	28	6.2
	137	155	19	18	2.9
	160	175	16	26	4.9
	179	184	6	23	7.7
	188	194	7	34	10
Z_B	21	84	64	134	5.4
	96	104	9	87	7.6
	111	119	9	111	13.5
	132	136	5	173	37.5
	156	159	4	132	11.1
	176	178	3	35	12.3
	185	187	3	184	16.3
Z_T	-3	5	9	72	21.8
	195	217	23	144	18

***** *refers to statistics over 100 snapshots each separated by 100 ps in a 10 ns MD simulation trajectory.*

This leads to higher Mean Separation of points in Ramachandran Plot (MSRP) for such regions. For helices/sheets, the points for consecutive residues often lie closer to each other in the Ramachandran map, while successive points jump from one quadrant to another only occasionally when there is a change in secondary structure, occurrence of glycine residue or residues with disallowed Ramachandran conformations. Thus, the average proximity between points of two consecutive residues is likely to remain quite close (low MSRP) in the case of regions with regular structure, while any irregularity in the structure shall increase the MSRP.

Figure 2.8 summarizes the averages and standard deviations of results displayed in **Figure 2.3** and **2.5** for all datasets employed. The α-helices not only have low MSRP, but also low σ. The 3_{10} helices appear to possess a shade higher range of values in comparison to α-helices, followed by strands. The strands display large error bars, emphasizing their higher variability in comparison to helices. The β-bridge too shows

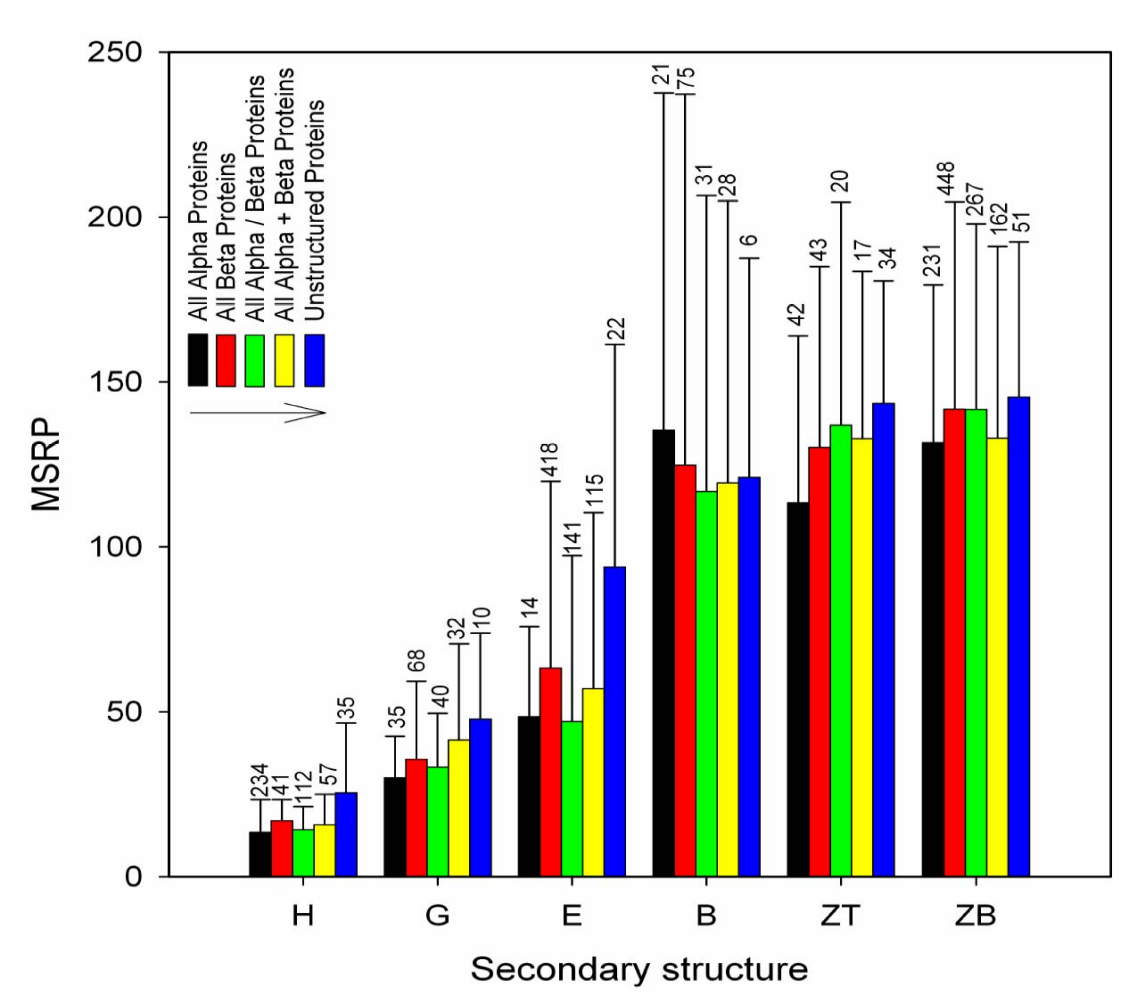

Figure 2.8: *The mean MSRP with its associated standard deviation is shown for each secondary structure among different protein datasets employed. The number above the error bars refers to number of samples used to obtain the statistic. For each secondary structure, the vertical bars represent from left to right: all α proteins, all β proteins, all α/β proteins, all α + β proteins and unstructured proteins. See* ***Figure 2.3 & 2.5*** *for the individual plots and* ***Table 2.7*** *for complete statistical details.*

high MSRP and high variability. Both the loops Z_T and Z_B look fairly similar amongst different protein datasets. It would be worthwhile to a) analyze the sequences that are specific to high or low MSRP loops b) define MSRP ranges for rigid and flexible loops.

The consistently higher MSRP observed for all secondary structures among unstructured proteins may also be accounted by the fact that majority (19/20) of these proteins have structures determined by NMR unlike proteins in **Figure 2.3** which are from X-ray diffraction. NMR structures that do not have residual dipolar coupling constraints do not have directly constrained ϕ and ψ dihedral angles.

It can be argued that the ϕ / ψ space is periodic and measuring the Euclidean distance can introduce higher MSRP values if two close points wrap around to different quadrants in the plot. We observed on an average ~2-3 points per protein in the boundaries (+175 to +180 and -175 to -180) of phi and psi for classes a, b, c and d, while on an average ~4-6 points per protein were seen for unstructured proteins. As we are calculating MSRP for small regions (~10 amino acid sequence), it is unlikely that significant fraction of points are present in the boundary. This error can marginally affect the MSRP calculations when the length of the region gets larger, especially for unstructured proteins. Perhaps, the marginally higher MSRP observed for different secondary structures in unstructured protein may be partly attributed to this error.

The data presented in **Figure 2.6** hint at a new approach to search for functional loops. In addition of to DNA/ligand binding, this approach may also be applied to explore enzyme-substrate, protein-protein interactions. Further analysis of data presented in **Figure 2.6** revealed that among class I/II/III proteins, the majority of loop population which deviate from the diagonal are loops that lie > 4 Å away from the ligand (see **Figure 2.9**). For DNA binding proteins however, majority of loops deviating from the diagonal, lie in the vicinity of the ligand (~3 Å). This is consistent with the fact that compared to ligand binding loops, a greater proportion of DNA binding loops undergo structural perturbations upon binding.

It is remarkable to observe that MSRP values are able to pick the subtle allosteric effects subsequent to ligand/DNA binding. Differences in MSRP between APO & HOLO proteins may reveal structural changes arising from induced fit. More investigations are

required to map the correlations between changes in MSRP and C^{α} displacement in a quantitative way for different regions.

It is apparent from **Table 2.8** that the amplitude of fluctuations in MSRP, as revealed by σ_{MD} is closely connected with the chaotic dynamics of the disordered polypeptide chain. Therefore, deviation in MSRP in the MD trajectory (σ_{MD}) clearly quantifies the disorder in the protein structure specifically when the disordered region alone is considered.

Taking this further to individual secondary structure regions (**Table 2.9**), it is observed that both ordered and unstructured proteins possess loops that show more or less a similar range of σ_{MD} values, suggesting that loops in ordered and unstructured proteins are similar in dynamics. Perhaps loops constitute a common feature between otherwise disparate classes of ordered and unstructured proteins. Unstructured proteins (1LXL & 2 SOB) seem to possess longer loops in comparison to ordered proteins (1BGF & 1MUN). More comparisons with a larger dataset are necessary to arrive at more definitive conclusions.

Describing the ensemble of conformations sampled by a disordered protein remains a challenge (**Eliezer, 2009**). Owing to the huge computational resources demanded for MD simulations of disordered state ensemble, such applications have been restricted to short peptides or short averaging times (**Makowska *et al.*, 2006; Mittag and Forman-Kay, 2007**). In this context, it is pertinent to note that σ_{MD} obtained from the MD trajectories appears as a useful quantifiable parameter to observe conformational diversity among the multiple conformations sampled by disordered protein. This parameter along with MSRP, appears to be most effective when it is measured for a specified disordered region as demonstrated for *1LXL**. It is hoped that this approach can provide insights into

previously unknown disordered regions in a given protein, while also enabling comparison between two different regions in a protein in terms of their disorder.

Finally it must be mentioned that the entire MSRP calculation lends itself for easy automation. Hence the MSRP method, once coded in software, should smooth the progress of structural analysis of a huge number of protein 3D structures to discover functional and dynamic loop regions.

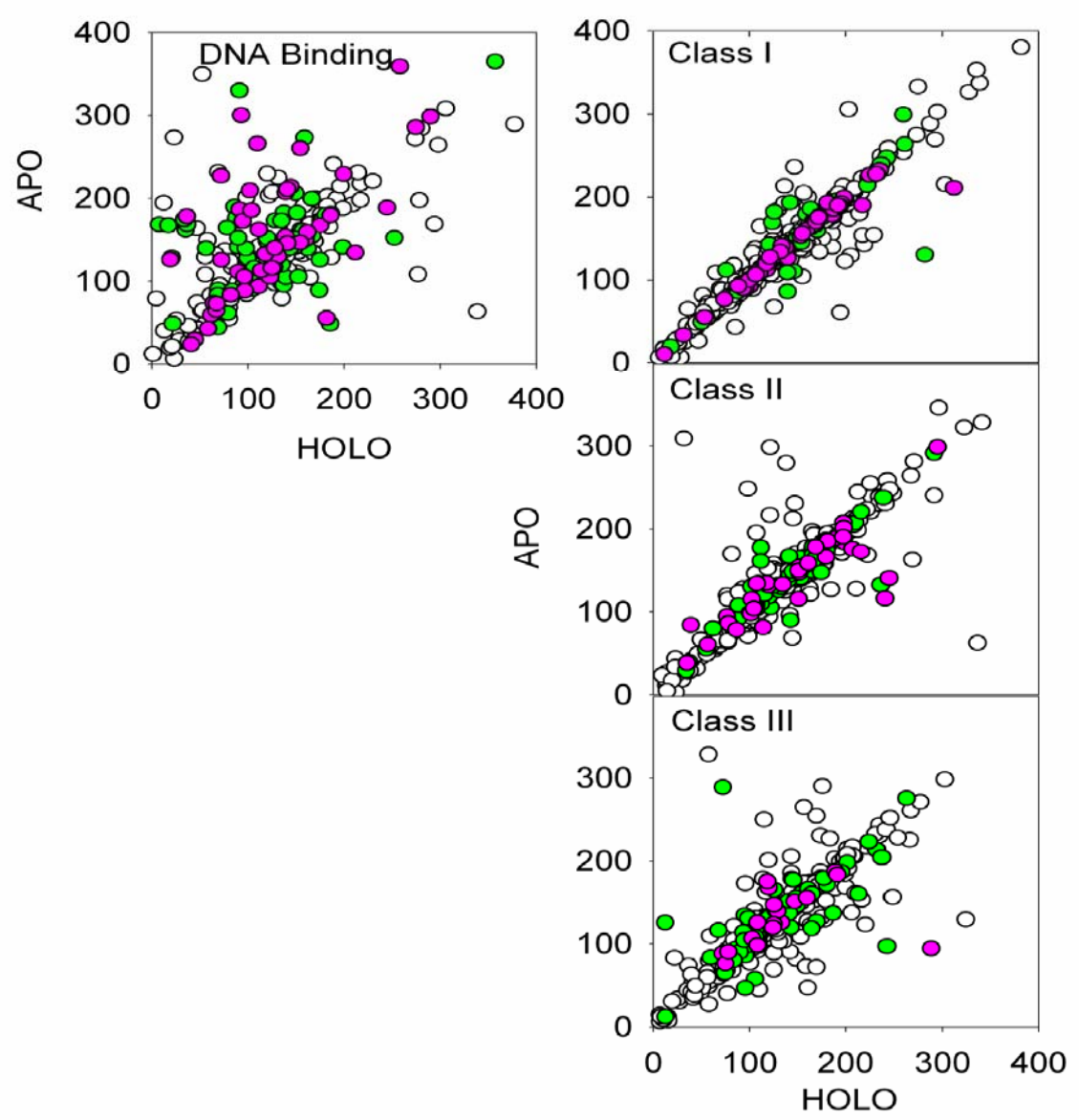

Figure 2.9: *Plots of MSRP of APO against HOLO for DNA binding proteins, Class I, II and III proteins. The symbols are as follows: unfilled circle, loops that are >3 Å away from the center of mass of the ligand; green circle, loops within 3 Å from the ligand and magenta circle, loops within 3 Å from the ligand if three flanking residues on either side are included. All these observations were done using UCSF Chimera a molecular visualization application.*

2.5. Conclusions:

In this work, we proposed a new method to isolate and analyze loop regions in a solved three dimensional structure of a protein using the MSRP parameter. Both the magnitude and the associated standard deviation of this parameter show unique characteristics for loop regions. This method is found to be a useful tool for identifying loop regions that are structurally perturbed after the protein is bound to a ligand. The MSRP parameter is shown to fluctuate more in unstructured regions like loops in comparison to regular regions like α-helices, during molecular dynamics simulations. It is envisaged that this method will a) enable a better categorization of loops and folds among proteins b) permit automated identification of functional loops in protein structures and c) provide clues on the diversity of conformations sampled by a disordered region during a molecular dynamics simulation.

CHAPTER 3
Characterization of Intrinsically Disordered Proteins using Molecular Dynamics Simulations

Characterization of Intrinsically Disordered Proteins using Molecular Dynamics Simulations

3.1. Introduction:

Many proteins lack well defined 3D structure under physiological conditions in the absence of a binding partner. These proteins are also called intrinsically disordered (**Dunker *et al.*, 2001**), natively denatured (**Schweers *et al.*, 1994**), natively unfolded (**Weinreb *et al.*, 1996**), intrinsically unstructured (**Wright and Dyson, 1999**), and natively disordered proteins (**Daughdrill *et al.*, 2005**). They have a broad occurrence in living organisms. It has been reported that bioinformatics analyses of whole genome sequences using disorder predictors (**Oldfield *et al.*, 2005b**) indicated that 6-33% of proteins in bacteria and 35-51% of proteins in eukaryota contain disordered regions of 40 or more consecutive residues (**Oldfield *et al.*, 2000; Dunker *et al.*, 2000**). Disorder has been observed in proteins involved in various cellular functions such as regulation of transcription and translation (**Uversky *et al.*, 2005**), cellular signal transduction (**Iakoucheva *et al.*, 2002**), protein phosphorylation, the storage of small molecules, and the regulation of the self-assembly of large multiprotein complexes (**Romero *et al.*, 2004**; **Dyson and Wright, 2005**). In most of these applications, binding to multiple partners, ability to overcome steric restrictions and high-specificity/low-affinity interactions is highly required. This is provided by IDPs as a result of their high conformational flexibility. The frequency of long contiguous regions of disorder or even whole proteins that are disordered under non denaturing conditions has recently spurred interest in these proteins. The number of experimentally characterized IDPs and ID regions is increasing rapidly and therefore these proteins are gaining much attention from the structural biology community. In the first release of Disprot (in February 2004), which is a database containing experimentally characterized IDPs and ID regions, there were 154 proteins and 190 disordered regions whereas in release 5.0 (February 2010), the database contained 517 proteins and 1183 disordered regions.

It has been observed that altered abundance of IDPs in cell is associated with several disease conditions. For example, over expression of thyroid cancer 1 (TC-1) (**Sunde *et al.*, 2004**) or under expression of adenosine 5-diphosphate (ADP) ribosylation factor (Arf) (**Sherr, 2006**) and p27 (**Grimmler *et al.*, 2007**) has been linked with various types of cancer. Similarly, over expression of α-synuclein and tau proteins increases the risk of aggregate formation and has been

linked to Parkinson's disease and Alzheimer's disease (**Chiti and Dobson, 2006; Goedert, 2001**), respectively.

In a typically ordered protein region, the Ramachandran angles and backbone atoms of each residue undergo non-isotropic small amplitude motions relative to their local neighborhood and are characterized by the equilibrium positions defined by the time averaged values. In contrast to ordered protein regions, IDPs or ID regions exist instead as dynamic ensembles in which atom positions and backbone Ramachandran angles vary chaotically over time with no specific equilibrium values.

Despite the large abundance, unusual structural and functional importance of disordered proteins, the disordered regions in these proteins are still poorly understood. Therefore, characterization of the structure and dynamics of such proteins must be considered as important as the structural studies of well ordered proteins.

The disorder in IDPs has been characterized by several physico-chemical methods such as nuclear magnetic resonance (**Dyson and Wright, 2002, 2004, 2005; Bracken *et al.*, 2004**), near ultra violet circular dichroism (CD) (**Fasman, 1996**), far-ultraviolet CD (**Adler *et al.*, 1973; Provencher and Glockner, 1981; Woody, 1995; Uversky *et al.*, 2000**), ORD (**Adler *et al.*, 1973; Uversky *et al.*, 2000**), Fourier transform infrared (**Uversky *et al.*, 2000**), Raman spectroscopy and Raman optical activity (**Smyth *et al.*, 2001**), different fluorescence techniques (**Uversky, 1999; Receveur-Brechot *et al.*, 2006**), numerous hydrodynamic techniques (including gel-filtration, viscometry, small angle x-ray scattering (SAXS), small angle neutron scattering (SANS), sedimentation, and dynamic and static light scattering) (**Uversky, 1999; Receveur-Brechot *et al.*, 2006**), rate of proteolytic degradation (**Markus, 1965; Mikhalyi, 1978; Hubbard *et al.*, 1994; Fontana *et al.*, 1997, 2004**), aberrant mobility in SDS-gel electrophoresis (**Iakoucheva *et al.*, 2001; Tompa, 2002**), low conformational stability (**Uversky, 1999; Privalov, 1979; Ptitsyn, 1995; Ptitsyn and Uversky, 1994; Uversky and Ptitsyn, 1996**), H/D exchange (**Receveur-Brechot *et al.*, 2006**), immunochemical methods (**Westhof *et al.*, 1984; Berzofsky, 1985**), interaction with molecular chaperones (**Uversky, 1999**), electron microscopy or atomic force microscopy (**Uversky, 1999; Receveur-Brechot *et al.*, 2006**), and the charge state analysis of electrospray ionization mass-spectrometry (**Kaltashov and Mohimen, 2005**). But most of these methods cannot directly sample protein structure on the time scale relevant to conformational changes (typically ns) in such regions and therefore provide only indirect

information on the unstructured state (**Mittag and Forman-Kay, 2007**). Thus in disordered proteins, the structural heterogeneity and rapid inter-conversion among conformers lead to problems in characterization and present practical challenges. It is apparent that the operating principles will be different for disordered protein regions than for folded protein domains, and there is currently very little knowledge of the biophysics of such regions, with many elementary questions unanswered.

In the recent past, MD simulations have been widely used as a powerful tool to understand the dynamic conformational changes of well structured proteins at atomic level. And this method has been less explored in the field of unstructured proteins possibly owing to paucity of high resolution structures among IDPs. There are few studies on disordered proteins using MD simulation. Recently, **Wang *et al.*, (2009)** has studied molecular dynamics simulation of intrinsically disordered proteins in Human diseases. **Chen (2009)** performed explicit-solvent molecular dynamics for both bound and apo phosphorylated KID (pKID) to capture the average properties in the protein folding and unfolding. In another study **Wells *et al.*, (2008)** identified differential flexibility in the N-terminal unfolded domain of tumor suppressor p53 using RDCs, in combination with SAXs and MD simulation. So this motivated us to study the characterization of IDPs with Molecular dynamics simulations. In our study, we attempt to examine and compare the dynamics between three well ordered proteins, one partially ordered and four intrinsically disordered proteins. We have performed 10 ns MD simulation of seven proteins with ff99SB force fields and TIP3P solvent model.

3.2. Materials and Methods:

Three sets of proteins were taken for this study. The first set comprises of ordered proteins, STAT-4 N-domain (pdb code: 1BGF), Catalytic domain of MutY from *E coli* (pdb code: 1MUN). The second set comprises of partially ordered protein, Brak/CXCL 14 (pdb code: 2HDL). The third set comprises of disordered proteins taken from DisProt (**Sickmeier *et al.*, 2007**), Sub domain of staphylococcal nuclease (pdb code: 2SOB), Apoptosis regulator Bcl-X_L (pdb code: 1LXL), F6 subunit of ATP synthase (pdb code: 1VZS) and Tyrosyl-tRNA synthetase (pdb code: 1JH3).

The initial structure for the molecular dynamics simulation of proteins with PDB codes 1BGF, 1MUN, 2HDL (NMR, Model 1), 2SOB (NMR Model 10), 1LXL (NMR minimized average structure), 1VZS (NMR Model 1), 1JH3 (NMR Model 1) were downloaded from PDB.

The LEaP module of the AMBER program package (**Pearlman *et al.*, 1995; Case *et al.*, 2005**) was used to prepare the system for simulation. Each protein was solvated with TIP3P (**Mahoney and Jorgensen, 2000**) waters and neutralized with the counter ions using the LEaP module. Energy minimization and MD simulations were carried out using the SANDER module of AMBER 8. The Amber force field ff99SB (**Hornak *et al.*, 2006; Wickstrom *et al.*, 2009**) is used to describe the atomic interactions. For the correct treatment of long range electrostatics, we make use of the Particle Mesh Ewald (PME) method (**Essmann *et al.*, 1995**). Constant temperature and pressure conditions in the simulations were achieved by coupling the system to a Berendsen's thermostat and barostat (**Berendsen *et al.*, 1984**). Bonds involving the hydrogen atoms were constrained to their equilibrium position with the SHAKE algorithm. For these simulations, we used an HP Proliant server with eight processors.

The system was minimized in two phases to avoid bad contacts. In the first phase, the system was minimized giving restraints (30 kcal/mol/Å^2) to protein and crystallographic waters for 500 steps with subsequent second phase minimization of the whole system. Then the system was heated to 300K over 50 ps with a 1 fs time step. The protein atoms were restrained with force constant of 30 kcal/mol/Å^2 at the NVT ensemble. After that the force constant was reduced by 10 kcal/mol/ Å^2 in each step to reach the unrestrained structure in three steps of 10 ps each. The system was then switched over to the NPT ensemble and equilibrated without any restraints for 180 ps. The system was equilibrated in total of 260 ps. The time step for MD simulation for the production run was 2 fs. All the seven trajectories were each run for 10 ns and were performed with an 8.0 Å cutoff on real-space interactions. Analysis of parameters of the trajectories was carried out using the ptraj modules of AMBER 8. Graphic visualization of protein structures was done using Chimera.

The RMSD, secondary structure analysis, radius of gyration analysis, End to end chain distance, solvent accessible surface area, and distance matrix analysis were carried out using ptraj action commands. The conformational entropy for ordered and disordered proteins were calculated using gas phase statistical mechanics. This involves principal component analysis (the quasi-harmonic approximation) that provides the first decomposition of correlations in particle motion (**Andricioaei and Karplus, 2001**). Thus the entropy is calculated analytically as a sum of independent quantum harmonic oscillators. Origin 6.1 was used to plot the secondary structure analysis and distance matrix analysis.

3.3. Results and Discussion:

3.3.1. Root Mean Square Deviation (RMSD):

The degree of conformational changes for the ordered and disordered proteins during the simulations is easily monitored by the C-alpha root mean square deviation (RMSD). The backbone RMSD of structures relative to the lowest-energy conformation has been calculated and is represented in **Figure 3.1 A, B, C, D, E, F, and G.** For the ordered proteins, 1BGF (**Figure 3.1 A**) and 1MUN (**Figure 3.1 B**), it can be seen that the corresponding structure and conformational variability of the ensembles obtained over the 10 ns simulation time remains almost constant and RMSD value remain well below 2 Å. In contrast to ordered proteins, the trajectories of the partially ordered protein 2HDL (**Figure 3.1 C**) and the disordered proteins 2SOB (**Figure 3.1 D**), 1LXL (**Figure 3.1 E**), 1VZS (**Figure 3.1 F**) and 1JH3 (**Figure 3.1 G**) exhibit higher structure and conformational variability among the ensembles. For all these partially ordered and disordered proteins, the RMSD fluctuates at higher rate and also settles at higher values. This is perhaps due to the instability in structure of these proteins. The flexible disordered regions present in the disordered proteins leads to population of diverse conformers. Very often these regions undergo transition in secondary structure from one form to other. This is discussed under secondary structure analysis. As a result of this, they exhibit higher RMSD values. One can easily infer the order of stability of all these proteins from the RMSD values and its fluctuations (std. dev.) (see **Table 3.1**). So it is apparent that the stability of these proteins follows the order; 1MUN > 1BGF > 2HDL > 1JH3 > 2SOB > 1VZS > 1LXL.

In 1LXL, if we calculate the RMSD values for the ordered region (Residue 1 to 27 & Residue 81 to 221) and disordered region (Residue 28 to 80) separately, we can clearly observe the variation in time course of RMSD as shown in **Figure 3.1 E**. The thin solid line and medium thick solid line represents the time course of RMSD for the ordered region (Residue 1 to 27 & Residue 81 to 221) and disordered region (Residue 28 to 80) in 1LXL respectively. The ordered region in 1LXL shows fairly stable and low RMSD values throughout the period of simulation and behaves similar to that of ordered regions in well ordered proteins while the disordered region in 1LXL shows more fluctuation in RMSD and settles at higher value. So the presence of disordered region in a protein can be known from the time course of RMSD.

The above results reveal that partially ordered and disordered proteins possess highly flexible regions and fragile structure compared to ordered proteins.

Table 3.1: Summary of analyzed trajectory parameters for ordered and disordered proteins

Type of Protein	PDB code	Number of Residues	RMSD (Å)	End to end distance (Å)	Solvent accessible surface area ($Å^2$)	Radius of gyration (Å)
Ordered	1BGF	124	1.6 ± 0.19	20 ± 3.96	8022 ± 141	15 ± 0.14
	1MUN	225	1.2 ± 0.15	33 ± 1.68	12036 ± 171	18 ± 0.15
Partially ordered	2HDL	78	3.4 ± 0.72	30 ± 5.16	7037 ± 162	14 ± 0.27
Disordered	2SOB	103	4.8 ± 0.86	27 ± 6.99	8711 ± 274	16 ± 0.38
	1LXL	221	11.4 ± 3.03	42 ± 7.53	14967 ± 728	21 ± 1.10
	1VZS	76	5.4 ± 0.81	32 ± 5.74	6723 ± 234	14 ± 0.32
	1JH3	99	3.6 ± 0.86	27 ± 3.43	6759 ± 224	13 ± 0.14

3.3.2. Radius of Gyration (R_g):

To investigate the compactness of protein during the simulation, radius of gyration was analyzed. Information regarding the overall shape of the molecule can be gleaned from R_g. The time course of the radius of gyration (R_g) for the ordered, partially ordered and disordered proteins are shown in the **Figure 3.2 A, B, C, D, E, F, and G.** The degree of packing of amino acid residues in the protein can be known from the radius of gyration. This value depends on the number of amino acids in the protein. This parameter is known to affect both the stability and folding rate of proteins. It is observed that ordered proteins, 1BGF (**Figure 3.2 A**) and 1MUN (**Figure 3.2 B**), shows little fluctuations in their R_g during the course of simulation while the partially ordered protein 2HDL (**Figure 3.2 C**) and the disordered proteins 2SOB (**Figure 3.2 D**), 1LXL (**Figure 3.2 E**), 1VZS (**Figure 3.2 F**) shows much higher fluctuations (See **Table 3.1**) in their R_g values during the simulation. This is mainly due to structural rigidity in the case of ordered proteins. In contrast, partially ordered and disordered proteins lacks rigidity in the structure perhaps due to presence of disordered regions. These disordered regions are irregular in structure, predominantly exist in the form of flexible loops. In the time course of simulation, these disordered regions in the partially ordered and disordered proteins undergo extensive change in structure and cause higher fluctuations in the R_g. Among the disordered proteins, 2SOB (**Figure 3.2 D**), and 1VZS (**Figure 3.2 F**), display wild fluctuations in R_g, while 1LXL (**Figure 3.2 E**) shows a downward trend. Particularly in 1LXL, the hydrophobic core has lost its compressibility and thus showing much variation in the R_g. This is mainly due to the presence of extensive irregular structure in these proteins. It seems that the irregular structural part in these proteins often move randomly in and out from its position. In the disordered protein 1JH3

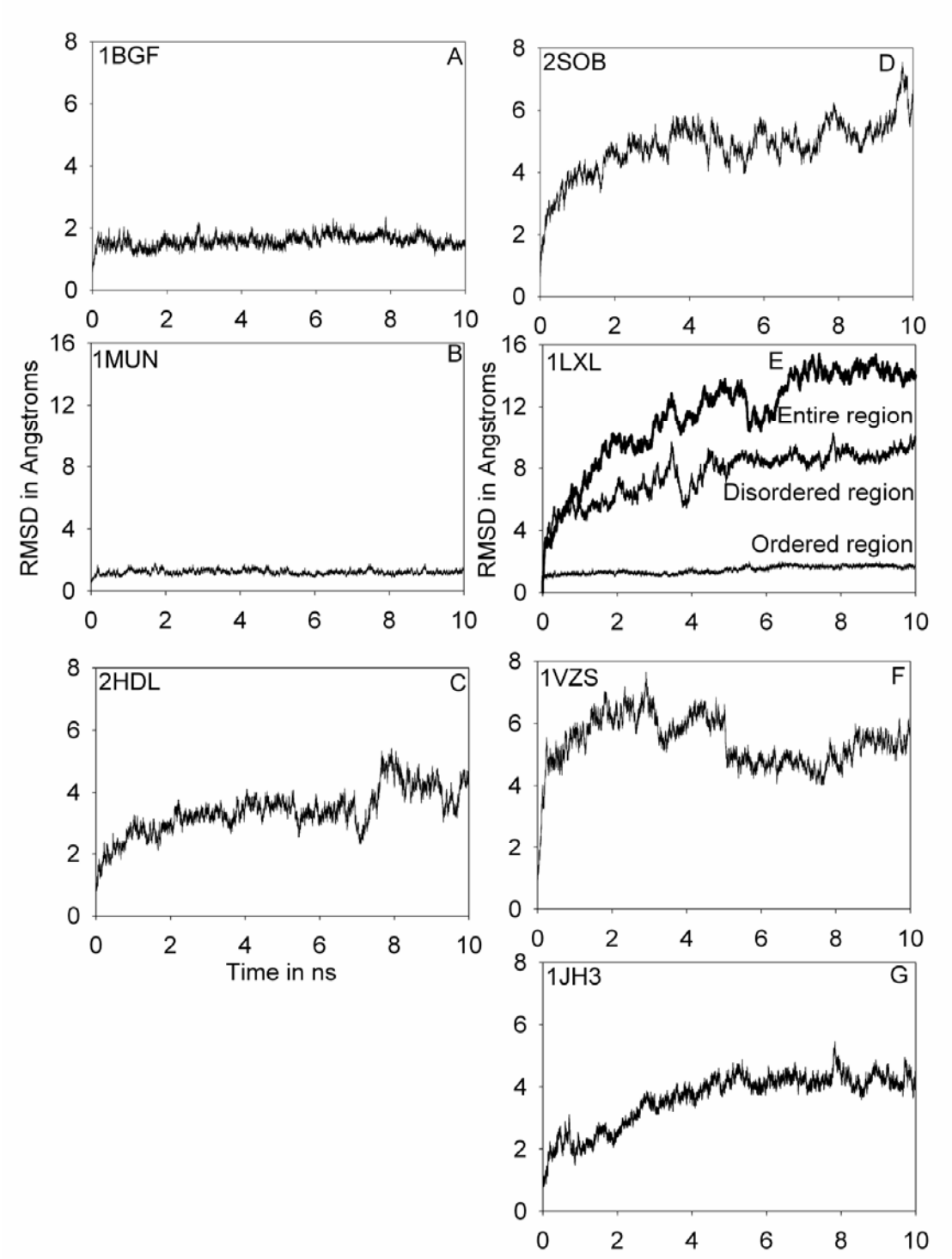

Figure 3.1:`*Root-mean-square deviation (RMSD) to the starting structure as a function of time is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3. Calculations were performed for the backbone atoms of the respective structure versus the backbone atoms of the respective simulation's starting structure.*

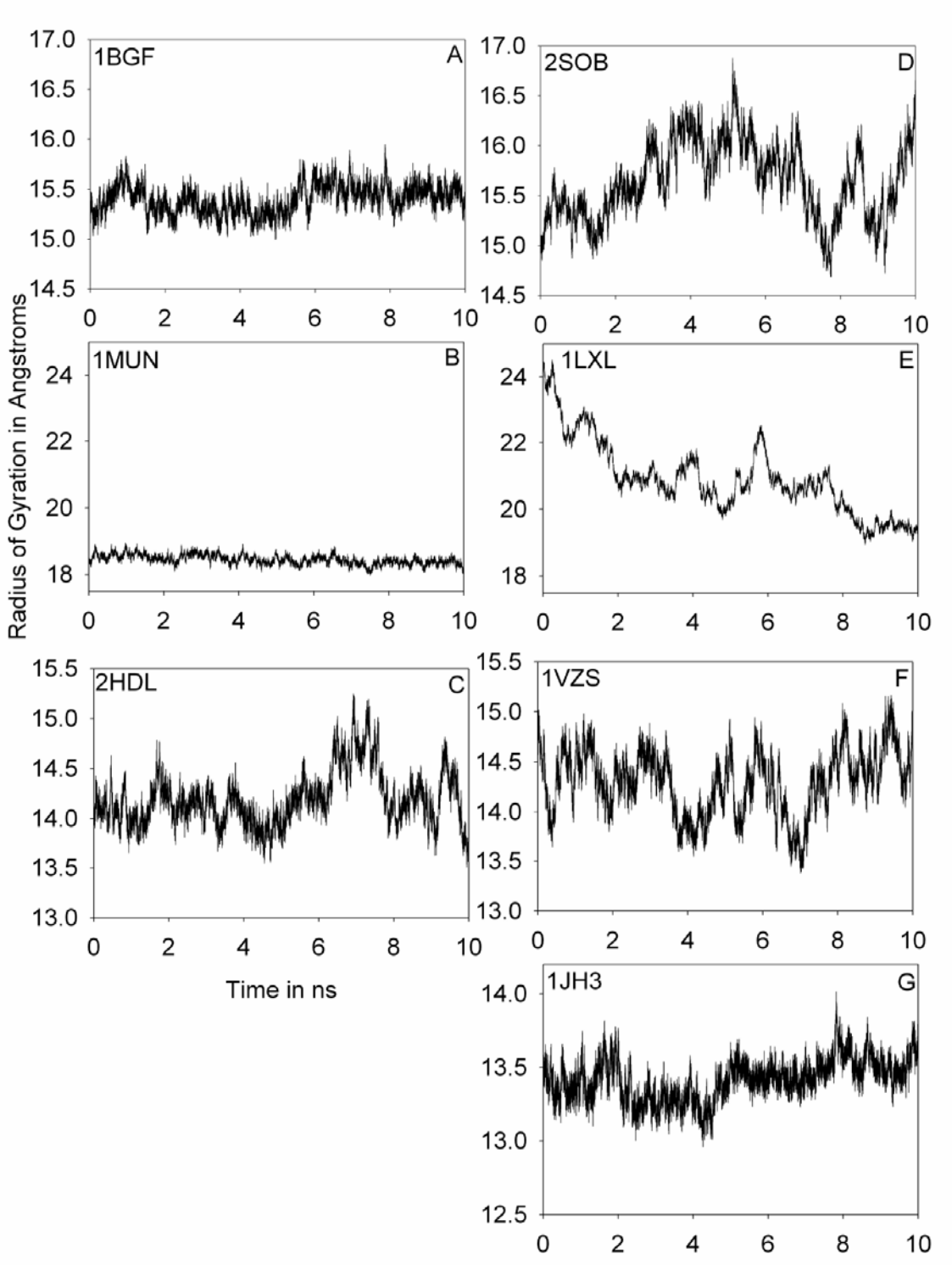

Figure 3.2:`*Radius of gyration of α-carbon atoms as a function of time is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

(**Figure 3.2 G**), we observe fluctuation in R_g similar to that of ordered proteins. This may be due to high degree of packing of amino acids in the regular secondary structure of this protein. Thus the fluctuation in Rg follows the order: 1BGF < 1JH3 < 1MUN < 2HDL < 1VZS < 2SOB < 1LXL. This result is consistent with the backbone RMSD analysis as shown in **Figure 3.1**.

3.3.3. Solvent Accessible Surface Area (SASA):

In order to investigate the finer details about the mobility of flexible regions in the proteins, we calculated the SASA. The results are depicted in **Figure 3.3 A, B, C, D, E, F, and G.** The degree of mobility of the flexible hydrophobic regions in the corresponding protein can be obtained by observing the time course of SASA. In the ordered proteins 1BGF (**Figure 3.3 A**) and 1MUN (**Figure 3.3 B**), SASA fluctuates to a marginal extent. This can be inferred from the standard deviation of SASA (**Table 3.1**). In the case of partially ordered protein, 2HDL (**Figure 3.3 C**), the standard deviation of SASA is slightly higher than the ordered protein 1BGF. But for the disordered proteins, 2SOB (**Figure 3.3 D**), 1LXL (**Figure 3.3 E**), 1VZS (**Figure 3.3 F**), and 1JH3 (**Figure 3.3 G**), SASA fluctuates to a greater extent, which can be inferred from corresponding higher standard deviation value of SASA (See **Table 3.1**). The standard deviation of SASA follows the order 1BGF < 2HDL < 1MUN < 1JH3 < 1VZS < 2SOB < 1LXL. The results of SASA analysis are nearly consistent with the analysis of radius of gyration. The results reflect that disordered proteins contain flexible regions predominantly in the form of loops or other irregular structures and this is often get exposed to the solvent due to high mobility. The unfolded feature in disordered proteins can also be inferred from these results. As a whole we observe that ordered proteins maintains structural integrity during the course of simulation in comparison with disordered proteins.

3.3.4. End to End distance (between first C-alpha atom and last C-alpha atom):

This is another trajectory parameter that gives information about the structural integrity of the protein during the simulation period. The distance between the first and last C^{α} atom was calculated during the time course of simulation for all the proteins under study and the results are depicted in the **Figure 3.4 A, B, C, D, E, F, and G.** The end to end distance in ordered proteins (**Figure 3.4 A & B**) oscillates relatively less in comparison with partially ordered (**Figure 3.4 C**), and disordered proteins (**Figure 3.4 D, E, & F**) which can be inferred from the standard deviation value (See **Table 3.1**). Interestingly 1JH3 shows significantly less fluctuation. The increasing

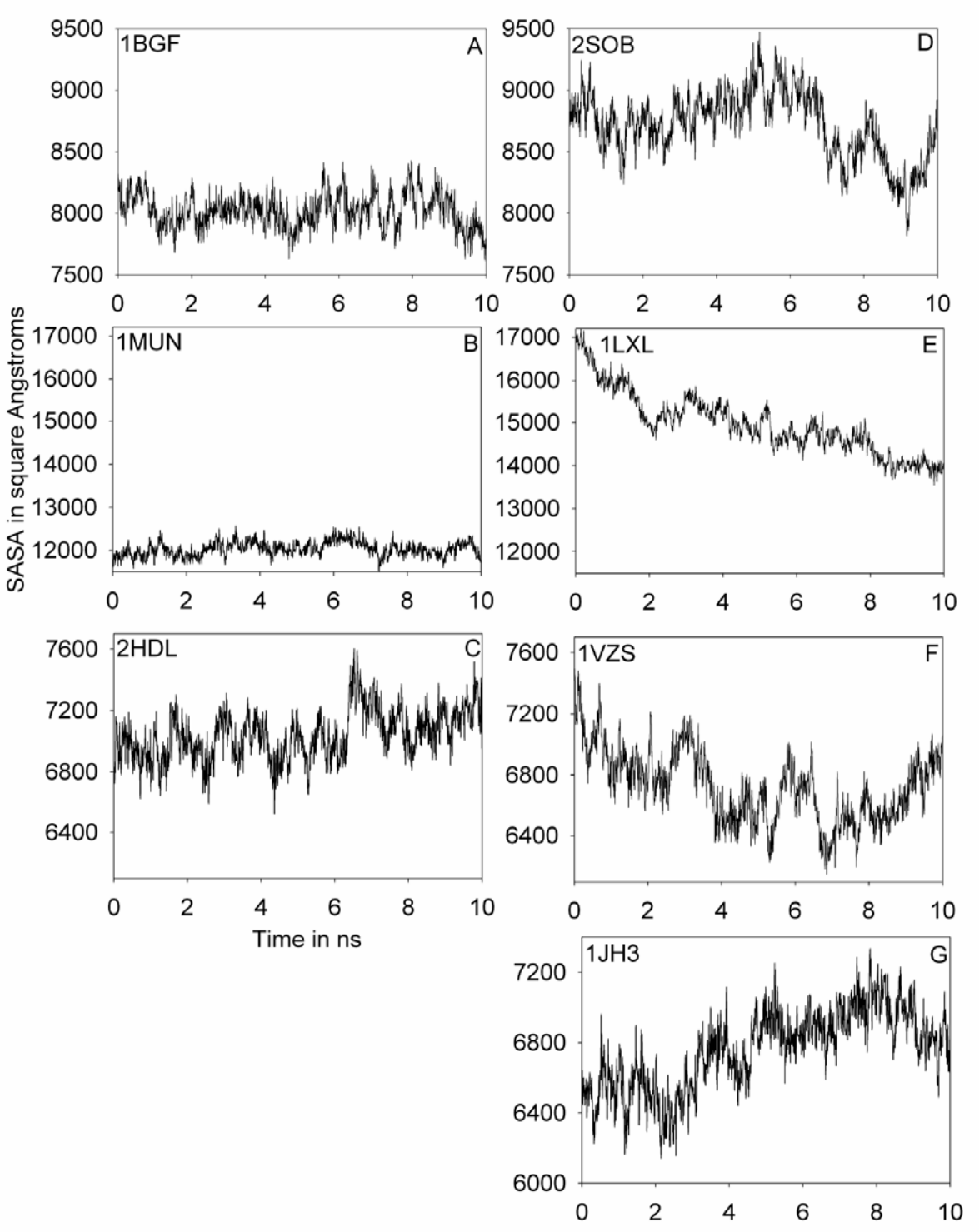

Figure 3.3:`*The solvent accessible surface area (SASA) of entire protein during the time course of simulation is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

order of oscillation of end to end distance value is 1MUN < 1JH3 < 1BGF < 2HDL < 1VZS < 2SOB < 1LXL. In 1BGF (**Figure 3.4 A**), the oscillation in end to end distance is little higher compared with disordered protein 1JH3 (**Figure 3.4 G**) because of the presence of flexible loops at both extremes of the protein chain. The persistence of structural integrity (high degree of packing of amino acids in the regular secondary structure) in 1JH3 actually decreases the fluctuation in end to end distance. The results are nearly consistent with radius of gyration data. Thus our results reveal that ordered proteins are more compact and rigid when compared with disordered proteins. And also, the disordered region in the disordered proteins undergoes rapid conformational dynamics owing to their flexible nature, thus causing higher fluctuations in end to end distance during course of the simulation period.

3.3.5. Analysis of Secondary structure:

The secondary structure analysis was carried out using the Kabsch and Sander algorithm incorporated in their DSSP (Dictionary of secondary structure for proteins) program (**Kabsch and Sander, 1983**). The results are plotted in **Figure 3.5 A, B, C, D, E, F, and G**. The plots show the structural variation of each residue during the time course of simulation. In the case of ordered proteins, that is 1BGF (**Figure 3.5 A**) and 1MUN (**Figure 3.5 B**), most of the regions in the protein remains unchanged in the secondary structure, although some transition in certain regions (Residue 25, 35-40, 47-48 in 1BGF & Residue 43-44, 82-85 in 1MUN) do notably appear that oscillates between turn and alpha helix, turn and irregular structure, and turn and 3_{10} helix .

In the case of partially ordered protein 2HDL (**Figure 3.5 C**), most of the region dominates with irregular structure and remains so during the simulation period. Some regions (Residue No. 6 &7) involves transition from irregular structure to parallel sheet structure and then some regions (Residue No. 51 & 52) from anti parallel sheet structure to parallel sheet structure. Residue No.14, 15, 16 oscillates between turn and irregular structure. Residue No. 66 & 67 oscillates between alpha helix, 3_{10} helix, turn and irregular structure. In the case of disordered proteins, 2SOB (**Figure 3.5 D**), 1LXL (**Figure 3.5 E**), 1VZS (**Figure 3.5 F**), and 1JH3 (**Figure 3.5 G**), most of the time majority of the regions possess irregular structure. Most of regions oscillate between turn and irregular structure. In the case of 1LXL, the experimentally designated disordered region that is between Residue No.28 and 80, involves frequent transitions between turn, irregular structure and 3_{10} helix. Some regions involve transition between alpha helix, turn and 3_{10} helix.

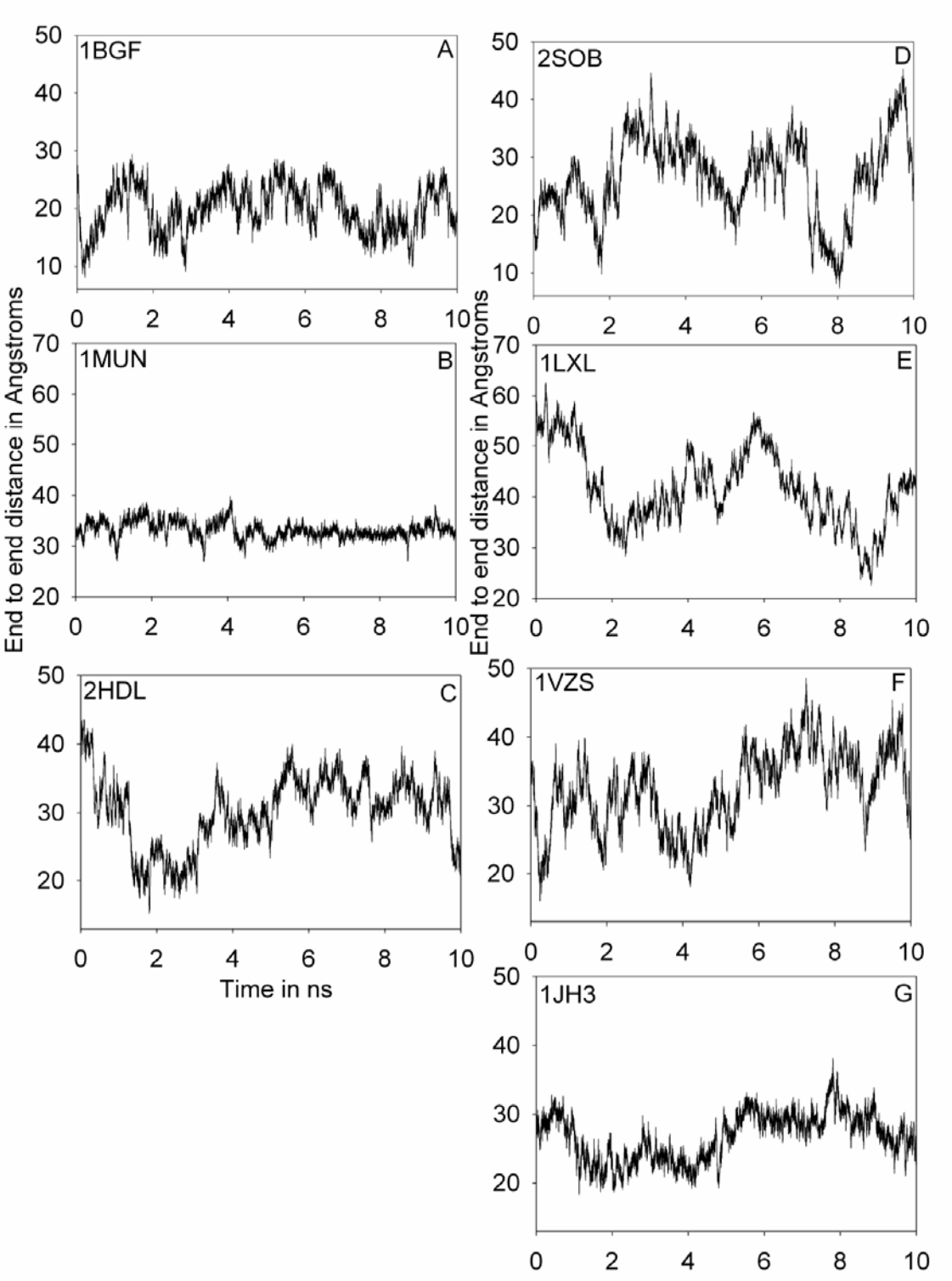

Figure 3.4:`*End to end chain distance of α-carbon atoms as a function of simulation time period is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

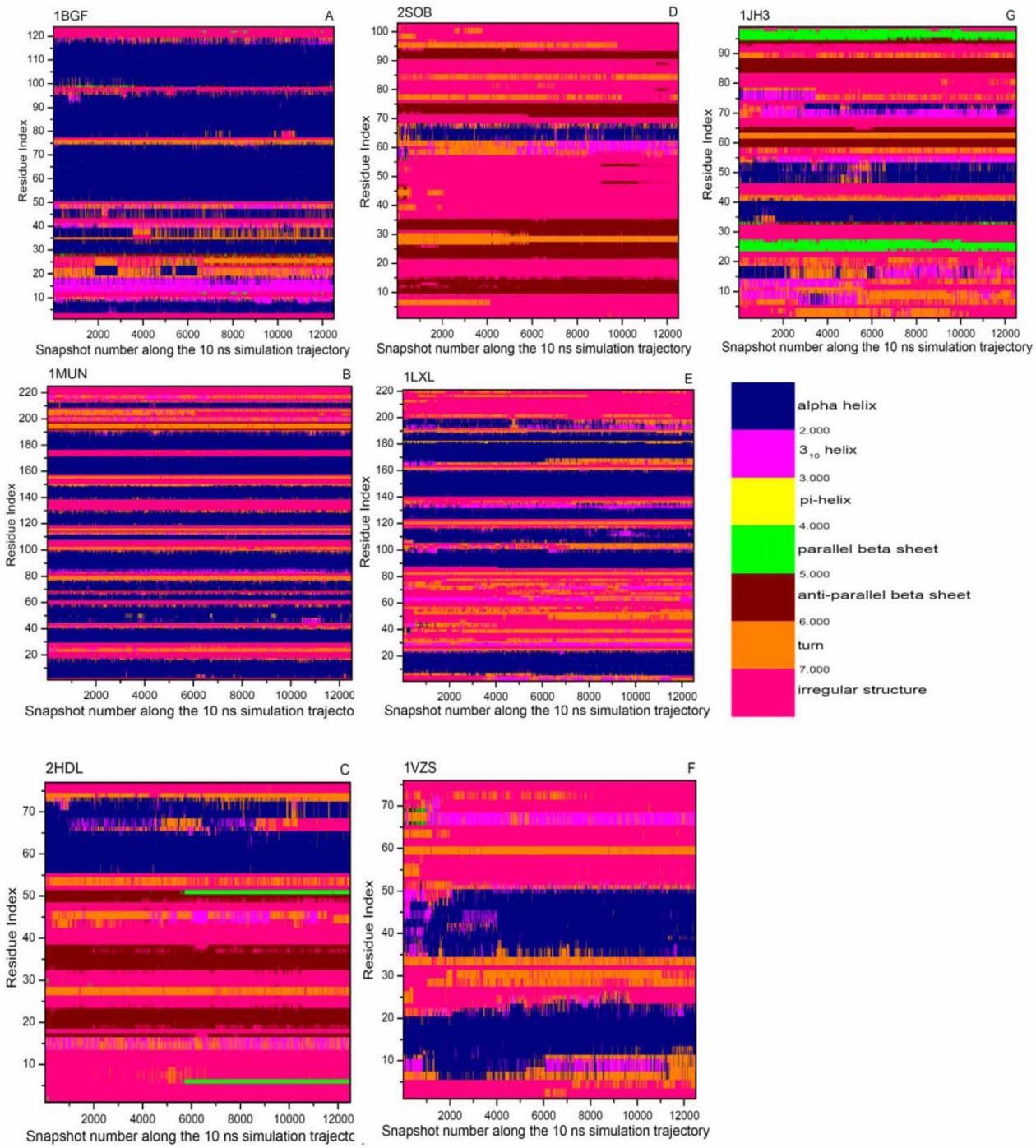

Figure 3.5:`*Detailed secondary structure data for each residue along the complete trajectory is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3. Color code may be interpreted from the legend.*

The above analysis reveal that in ordered proteins, the secondary structure profile remains fairly invariant in vast majority of regions during the time period of simulation while in partially ordered and disordered proteins, there is extensive change in the secondary structure profile in several regions. In disordered proteins, most of the residues exist predominantly in irregular structure; and helical elements are eventually replaced by irregular structure like loops. So the existence and rapid change in structural dynamics of disordered regions in disordered proteins is clearly visible from the secondary structural analysis.

3.3.6. Distance Matrix Analysis:

To investigate the proximity of C^{α}-atoms in the protein structure, we analyze the average minimum distance matrix. The distance between the C^{α}-atoms for all pairs of residues in the corresponding proteins under study is depicted in the **Figure 3.6 A, B, C, D, E, F, and G**. The points near the diagonal represent distances between adjacent residues along the protein backbone. We can infer the details about the regions containing standard secondary structure and irregular structure in the protein. Helical elements are indicated by thick bulges (cylinders) along the diagonal. While the lines that are parallel and perpendicular to the diagonal represents other secondary structures (β-strands) in the protein. In the case of ordered proteins, 1BGF (**Figure 3.6 A**) and 1MUN (**Figure 3.6 B**), we can clearly see the existence of well ordered secondary structure throughout the protein. But on the other hand in partially ordered (**Figure 3.6 C**) and disordered proteins (**Figure 3.6 D, E, F, and G**), we can see the presence of both well ordered secondary structure and irregular structure. The disordered region in 1LXL is experimentally known to be between residue No.28 and 80. From the **Figure 3.6 E**, it can be observed that the region between these two residues lacks well ordered secondary structure. Further it can be seen that in the disordered regions, there is abnormal variation in the distance between C^{α}-atoms in comparison with other ordered regions of the protein.

3.3.7. S^2 order parameter:

The generalized order parameter S^2 calculated from MD simulation trajectory using time dependent correlation motion function generally agrees well with NMR S^2 values (**Chandrasekhar *et al.*, 1992**). This is often useful to validate the conformational dynamics of proteins. The S^2 order parameter is an indicator of protein backbone motions in computationally feasible time scales. We have computed the S^2 values for the ordered and disordered proteins from the corresponding MD simulation trajectory. The results are shown in **Figure 3.7 A, B, C,**

D, E, F, and G. The flexible regions or highly mobile regions in the protein can be inferred from the S^2 values (lower the S^2 value, higher will be flexibility and vice versa). From the plots, it is observed that residues in the disordered regions possess lower S^2 values. In contrast to residues in disordered regions, residues in ordered regions possess fairly higher S^2 values. This is more clearly visible in **Figure 3.7 E**, where the experimentally known disordered region (between residue no. 28 and 80) possesses lower S^2 values in comparison with the rest of the ordered region of the protein. Thus the information regarding the highly flexible disordered regions can be known from the generalized S^2 order parameter.

3.3.8. Conformational Entropy:

To investigate the disorder associated with the structural arrangement of the protein, we analyzed the conformational entropy from the MD simulation trajectory. The results are shown in the **Figure 3.8**. The normalized conformational entropy per residue was observed to be more for the disordered proteins than ordered proteins. The calculated conformational entropy follows the order: 1MUN < 1BGF < 1JH3 < 2HDL < 1LXL < 2SOB < 1VZS. Thus the extent of disorder in the protein can be known from the magnitude of conformational entropy. The results obtained are in consistent with the RMSD trend except for the 1LXL. This is because the normalized conformational entropy per residue depends on the length of disordered region and the length of the protein chain. As the length of the disordered region reaches the length of the protein chain, the conformational entropy value will reach the maximum. But in 1LXL the effect of disordered region on the conformational entropy value is masked because of the larger length of ordered region in the protein.

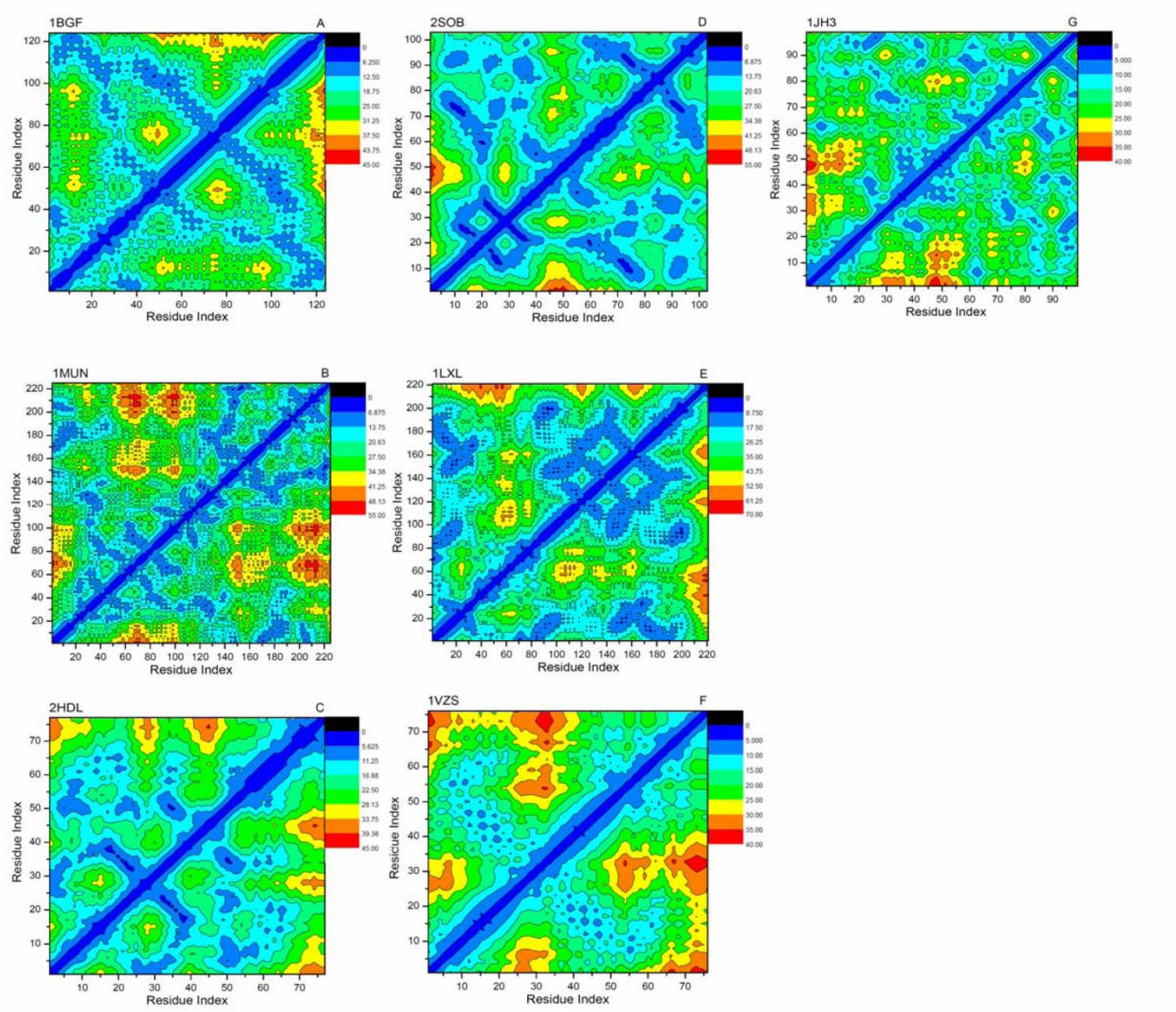

Figure 3.6: *`C^{α} atoms distance matrices are shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

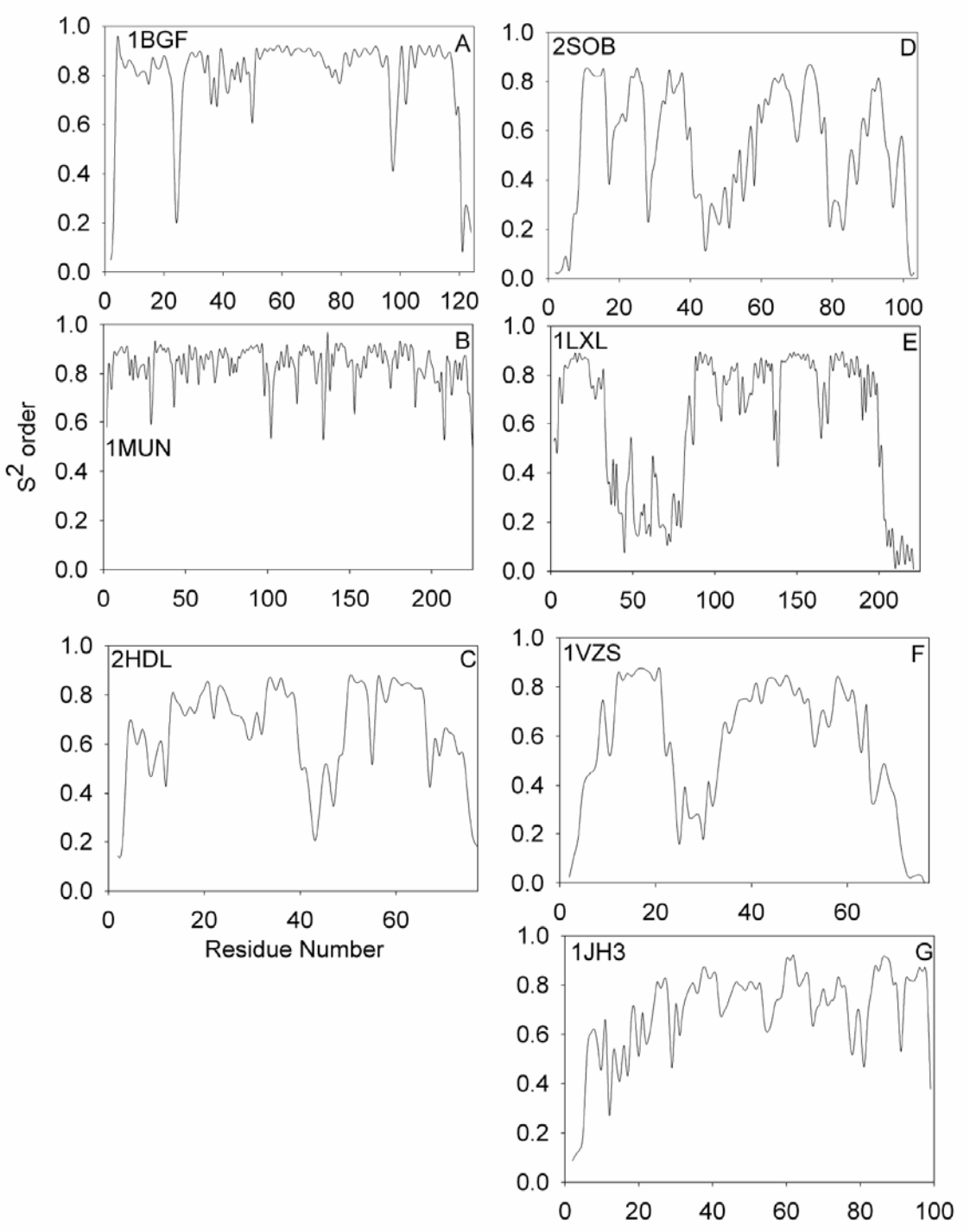

Figure 3.7: *`Calculated Generalized order parameter as a function of residue number is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

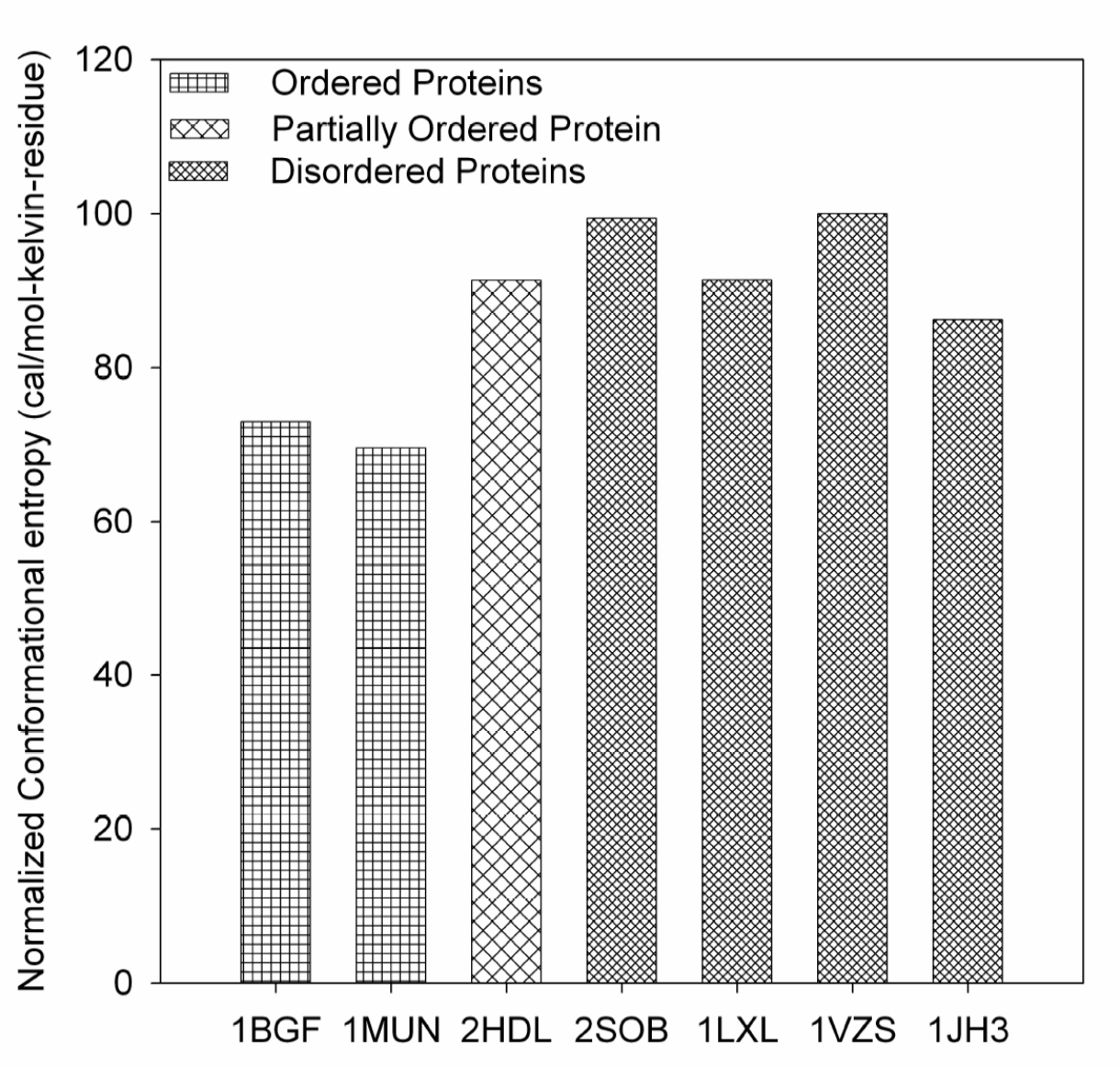

Figure 3.8:`*Normalized conformational entropy is shown. (A) 1BGF. (B) 1MUN. (C) 2HDL. (D) 2SOB. (E) 1LXL. (F) 1VZS. (G) 1JH3.*

3.4. Conclusions:

In this work, we have used the MD simulations to analyze and compare the conformational dynamics originating from two ordered, one partially ordered and four disordered proteins. The observed conformational dynamics features of disordered proteins are unique and can be used as characteristic features to distinguish them from well ordered proteins. Particularly the structure instability feature in disordered protein due to disordered region is seen from high RMSD values, large fluctuations in radius of gyration, end to end C^{α} atom distance, SASA and

RMSD, and higher conformational entropy. The other characteristic feature of disordered proteins is rapid change in conformational dynamics. This is incidental from the secondary structure analysis. Apart from this, the feature of high mobility or flexibility in disordered regions of disordered proteins is evident from lower generalized S^2 order parameter, and SASA analysis. In addition, the structural irregularity in disordered regions is reflected from distance matrix analysis and secondary structure analysis. Thus our results clearly suggest the various approaches employed to analyze MD simulation trajectories clearly bring out the dynamic nature of disordered regions.

CHAPTER 4
Effect of macromolecular crowding on the rate of acetyl cholinesterase reaction: dependence on crowder size and substrate

Effect of macromolecular crowding on the rate of acetyl cholinesterase reaction: dependence on crowder size and substrate

4.1. Introduction:

Inside a living cell living cell, biological processes occur in a medium containing high concentrations of assemblies such as cytoskeletal filaments and microtubules, various sub-cellular organelles and a host of other macromolecular species in an assortment of size and shape like proteins, nucleic acids and so on. Such a milieu is often referred to as macromolecular crowding (**Hall and Minton, 2003**). The media is crowded rather than concentrated because no single macromolecule is at high concentration, but taken together, the total macromolecular concentration is 50-400 mg/ml (**Fulton, 1982; Lanni *et al.*, 1985; Cayley *et al.*, 1991**), implying that between 5% and 40% of the total volume is physically occupied by these molecules (**Fulton, 1982; Gershon *et al.*, 1985**). Therefore, the accessible volume in the cell is reduced. This volume occlusion is significant when compared with *in vitro* conditions. Thus the intra-cellular milieu is crowded and thermodynamically nonideal.

In the cell, the consequences of crowding have been recognized and studied for many years (**Laurent, 1995; Zimmerman and Minton, 1993; Minton, 2001**). Crowding can influence the equilibrium properties of the system such as equilibrium constants by changing the activities of chemical species. Another prominent outcome of crowded conditions inside the cell is that it causes intrinsically disordered proteins to fold into 3D structure containing stable secondary and tertiary structure (**Flaugh *et al.*, 2001**). Crowding can also affect the molecular diffusion rates. This is applicable for both large and small molecules, but the impact is more for large molecules. It is well established that crowding can considerably decrease the diffusion coefficients of macromolecules (**Luby-Phelps *et al.*, 1987; Gershon *et al.*, 1985; Verkman, 2002**), influence diffusion-controlled reaction rates (**Schnell and Turner, 2004**), lead to shifts in chemical equilibria (**Hall and Minton, 2003**), alter protein folding processes and influence protein assembly (**Eggers and Valentine, 2001; van den Berg *et al.*, 1999; Ellis and Hartl, 1999; Zimmerman and Trach, 1991; Zhou, 2004**). In cellular metabolism, the molecular diffusion of enzymes and their substrates play significant role where the encounter of the free substrate with the active site of the freely diffusing enzyme can often be the rate

determining step. Few studies have explored the consequences of crowding on enzyme catalysis *in vitro*. Minton has investigated the effect of macromolecular crowding on the different kinetic steps of enzyme catalysis namely, 1) the enzyme (E), substrate (S) encounter and formation of ES complex, 2) the formation of enzyme-product complex from ES complex, and 3) release of product from the enzyme-product complex. It is predicted that under the condition when the encounter between the substrate and enzyme is rate-limiting, the rate of an enzyme-catalyzed reaction will experience a monotonic decrease with increase in the fractional volume occupancy of the crowding agent (**Minton, 1981**). **Sasaki et.al., (2007)** observed addition of PEG (4K to 20K) enhances activity of DNase I and S1 nuclease, does not significantly affect activity of exonuclease III, and decreases activity of exonuclease I. The effect of different concentrations of Ficoll 70 on the rate of EcoRV-catalyzed cleavage of pBR 322 was studied by **Wenner and Bloomfield (1999)**. They observed that Ficoll 70 had little effect on the overall reaction velocity of EcoRV in the concentration range 0-20% g/dL owing to offsetting increases in V_{max}, K_m, and stronger non-specific binding between enzyme and substrate/product. The effect of PEG 6K on enzyme activity of *Escherichia coli* AspP was investigated recently (**Moran-Zorzano *et al.*, 2007**). They reported that 50 g/L PEG decreases K_m fourfold and increases V_{max} sixfold. **Derham and Harding (2006)** studied the effect of the presence of globular proteins and elongated polymers on enzyme activity. They noticed that enzymatic activity increases and then decreases with increasing concentration of protein crowding agents, but decreases monotonically with increasing concentration of polymeric crowding agents. Often in most of the studies, crowding agents of limited range of sizes were considered. However, it is known that the macromolecules present inside the cell, exist in a wide range of shapes and sizes. Thus, it is imperative to investigate the effect of different sizes of crowding agents on enzyme activity. Theoretical studies have predicted that classical Michaelis-Menten kinetics may not apply to enzyme reactions in crowded media (**Berry, 2002; Savageau, 1995; Schnell and Turner, 2004**). Our group has previously reported the effect of crowding by Ficolls (30 kDa & 400 kDa) and dextrans (15 kDa to 500 kDa) on the activity of alkaline phosphatase. We also demonstrated that Michaelis-Menten kinetics may not be applicable in crowded media (**Homchaudhuri, *et al.*, 2006**). In this work, we investigate

the effect of crowding, using dextran in sizes ranging from 40 to 2000 kDa on the rate of two enzyme catalyzed reactions: a) alkaline phosphatase (AP) catalyzed hydrolysis of p-nitro phenyl phosphate (PNPP) and b) acetyl cholinesterase (AchE) catalyzed hydrolysis of 2-naphthyl acetate (NA) and 3-indoxyl acetate (IA). A noticeable difference in the effects of smaller dextran (40 kDa) in comparison with larger dextrans (500 and 2000 kDa) was observed in the case of AP vs. PNPP. Our findings reveal that smaller dextrans (40 kDa) do not appreciably influence the rate until reaching a concentration of 20 % w/w whereas the larger dextrans shows a steep monotonic decrease in normalized reaction rate. In the case of AchE vs. IA, we observe a similar trend as in AP vs. PNPP. However, in the case of AchE vs. NA, the thermodynamic activity of AchE seems to predominate over the crowding effect at the lower concentrations of dextrans. It is interesting to observe the opposite effects of crowding with two different substrates in AchE. The present work also employs a newer method to ensure efficient mixing of large molecular weight dextrans with enzyme.

4.2. Materials and Methods:

Alkaline phosphatase (from bovine intestinal mucosa), Acetylcholine esterase (from electrophorus electricus, electric eel), were purchased from Sigma-Aldrich Chemicals Pvt. Ltd., India. 2-Naphthyl acetate, 3-indoxyl acetate, Dextran (from Leuconostoc mesenteriodes) of molecular weight 15, 40, 70, 200, 500 and 2000 kDa were purchased from Fluka. The polydispersities of the dextrans were typically less than 2.0 as reported by the manufacturer. Glycine, disodium hydrogen phosphate and glycerol (98% purity) were obtained from Merck. p-nitro phenyl phosphate disodium salt was bought from Sisco Research Laboratories, India. All other chemicals employed were of analytical grade.

(a) Hydrolysis of PNPP by Alkaline phosphatase:

A typical reaction mixture contained alkaline phosphatase (2 μM) and PNPP of desired concentration dissolved in an aqueous solution of 100 mM glycine buffered at pH 9.5. The substrate concentration was kept at 1 mM, which is well above the measured K_m under the reaction conditions employed (0.25 mM). The ratio of molar concentration of substrate to enzyme is higher. This is in agreement with similar ratios inside living cells. However, to rule out situations, such as substrate being trapped/bound in dextrans, a few

experiments were also carried out with 20 mM PNPP. These experiments revealed that the profile observed with 20 mM PNPP against increasing dextran size is similar to that observed with (data not shown) 1mM PNPP. The concentration of dextran in the medium was varied between 0 and 30% (w/w). The total weight of the reaction medium was kept constant at 1.0 g.

The reaction was initiated by forcefully mixing the enzyme (typically ~50 µL in buffered aqueous medium) with an aqueous buffered mixture containing PNPP and crowding agent (typically ~950 µL) in an eppendorf tube using a syringe. This mixture was vigorously vortexed for 30 seconds. Immediately after, the mixture was transferred to a cuvette and the progress of the reaction was conveniently monitored using a spectrophotometer by recording the absorbance of the product p-nitro phenol at 450 nm after a dead time of 30 seconds. The above procedure ensured efficient mixing of substrate and enzyme in the midst of crowding agents.

(b) Hydrolysis of 2-Naphthyl acetate by Acetyl choline esterase:

A typical reaction mixture contained Acetyl choline esterase (61.7 nM) and 2-Naphthyl acetate of desired concentration in an aqueous solution of 10 % methanol and 20 mM phosphate, buffered at pH 7.5. The substrate concentration was kept at 1.5 mM which is lesser than K_m (2.3 mM). This is because we observe formation of precipitate upon mixing high concentrations (> 2 mM) of substrate with enzyme in the presence of crowding agent (dextran). The concentration of dextran in the medium was varied between 0 and 30% (w/w). The total weight of the reaction medium was kept constant at 1.0 g. The reaction was initiated by mixing the enzyme (typically ~50 µL in buffered aqueous medium) with an aqueous buffered mixture containing 2-Naphthyl acetate and crowding agent (typically ~950 µL) in an identical procedure as described for AP reaction above. Subsequently the progress of the reaction was conveniently monitored using a spectrophotometer by recording the absorbance of the product 2-Naphthol at 327 nm.

(c) Hydrolysis of 3-Indoxyl acetate by Acetyl choline esterase:

A typical reaction mixture contained Acetyl choline esterase (61.7 nM) and 3-Indoxyl acetate of desired concentration in an aqueous solution of 10 % methanol and 20 mM phosphate buffered at pH 7.5. The substrate concentration was kept at 5 mM which

is near to the K_m (5.1 mM). This is because we observe formation of precipitate upon mixing high concentrations (> 5.5 mM) of substrate with enzyme in the presence of crowding agent (dextran). The concentration of dextran in the medium was varied between 0 and 30% (w/w). The total weight of the reaction medium was kept constant at 1.0 g. The reaction was initiated by forcefully mixing the enzyme (typically ~50 µL in buffered aqueous medium) with an aqueous buffered mixture containing 3-Indoxyl acetate and crowding agent (typically ~950 µL) as described for AP reaction. The progress of the reaction was conveniently monitored using a spectrophotometer by recording the absorbance of the product 3-hydroxy indole at 385 nm.

In all the reactions above, the completeness of the mixing was ensured and supported by the following observations: 1) the absorbance of product formed initially increased steadily with time from the start, maintaining a linear (monophasic) profile. ii) The initial slope of the absorbance/time plot obtained above was reproducible when the experiment was repeated subsequently multiple times under identical conditions.

(d) Calculation of Normalized rate of reaction:

The initial velocity, V, was obtained by linear regression of the first 20 s of the recorded absorbance/time data, so that inhibition from appreciable build up of the product is negligible. The initial velocity observed under identical conditions, but in the complete absence of the crowding species, was referred to as V_0. The normalized rate, V_{norm}, was calculated from the following equation:

$V_{norm} = 100 (V/V_0)$

The points depicted in the figures are the averages of at least three independent experiments done on different days. Blank solutions containing a mixture of 25% (w/w) of the crowding agent employed and the corresponding substrate showed negligible change in the absorbance in the complete absence of the enzyme under identical conditions, proving that all of the crowding agents employed in the study are indeed chemically inert. All samples were made in deionized water. All the experiments were carried out at 25 ^{0}C.

(e) Calculation of rate of reaction at different viscosity:

The dependence of the rate of the above enzymatic reactions on solution viscosity are studied by using glycerol water mixtures. The method adopted to calculate the rate is

same as above. The relative viscosity of aqueous solutions of glycerol (at 25 ^{0}C) as a function of concentration expressed in weight percent were obtained from literature (**Borchers, 1955**). For some concentrations (15, 25, 35, 45 % w/w) of glycerol, the relative viscosity values are not given in the literature. These values are calculated using interpolation from the plot of known values of relative viscosity against glycerol concentration (0 - 60% w/w). The fitted points yielded an $R^2 = 0.985$.

4.3. Results:

In this work, three different enzymatic reactions were chosen to study the effect of macromolecular crowding by different sizes of dextran on the kinetics of the reaction. The reasons for selecting these three enzymatic reactions were follows: (i) the reactions are accompanied by minimal change in the excluded volume. The substrates, PNPP (p-nitro phenyl phosphate) (M_w ~ 220 Da), NA (2-naphthyl acetate) (M_w ~186 Da), IA (3-indoxyl acetate) (M_w ~175 Da) and the products, p-nitrophenol (M_w ~140 Da), 2-naphthol (M_w ~144 Da), 3-hydroxy indole (M_w ~133 Da) are relatively smaller in size when compared with surrounding macromolecules, dextrans (M_w ~ 15,000 Da or more), such that excluded volume effects from substrate, and product can be safely neglected. (ii) The progress of all these reactions can be easily monitored by UV-visible spectroscopy. (iii) The enzymes, Alkaline phosphatase (M_w ~1,46,300 Da) and Acetyl cholinesterase (M_w ~69,139 Da) diffuses relatively slowly compared with their respective substrates. The concentration of substrates taken in all these reactions is much higher than the concentration of enzyme. Thus we have the medium where in the substrate diffuses freely in the midst of relatively immobile enzyme and background macromolecules.

First we concentrate on the effect of crowding on kinetics of AP vs. PNPP. **Figure 4.1 A, B** and **C** depict the observed initial rise and succeeding plateau in the absorbance due to formation of the yellow product, p-nitrophenol, during the hydrolysis of PNPP in the presence of increasing amounts of 40, 500 and 2000 kDa dextran in the reaction medium, respectively. In **Figure 4.1 A**, we observe only a marginal decrease in the initial velocity of the reaction with increasing amounts of 40 kDa dextran (noticeably 25 & 30 % w/w) as revealed by the initial slope of the absorbance/time data. In contrast, **Figure 4.1 B & C** reveal a significant decline in the initial velocity when increasing

amounts of 500 and 2000 kDa dextran were added to the medium, respectively. The reaction slows down appreciably at 25% w/w for 500 kDa & 20% w/w for 2000 kDa.

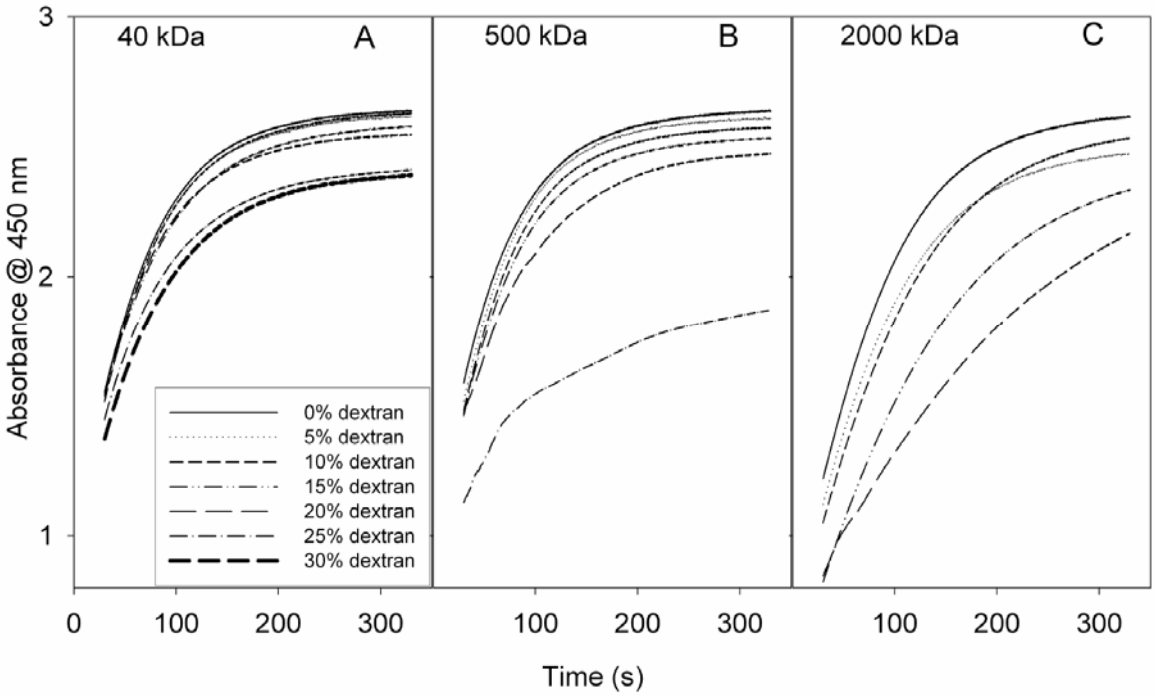

Figure 4.1: *`The reaction profile of an alkaline phosphatase-catalysed hydrolysis of PNPP (at 1 mM) monitored by measuring the absorbance at 450 nm is shown. (A) In presence of 40 kDa dextran. (B) In presence of 500 kDa. (C) Inpresence of 2000 kDa. The different w/w % of dextrans are represented by the symbols as follows: solid line, 0%; dotted, 5%; short dash, 10%; dash dot dot, 15%; long dash, 20%; dash dot, 25%; medium dash, 30%.*

The effect of differing concentrations of 40, 500 and 2000 kDa dextran, taken individually on the normalized reaction rate is shown in **Figure 4.2**. The difference in trend of smaller dextrans (40 kDa) and larger dextrans (500 and 2000 kDa) is clearly visible. While smaller dextrans do not appreciably influence the rate until reaching a concentration of 20% w/w, larger dextrans reveal (especially 2000 kDa) a steep monotonic decrease in normalized rate. The profile observed with 500 (at 20 & 25% w/w) and 2000 kDa dextrans is similar to that expected when the rate of reaction is limited by the diffusional encounter between AP and PNPP. The alkaline phosphatase reaction, in the absence of crowding is not diffusion limited (k_{cat}/K_m = (0.76 ± 0.05) x 10^4 M^{-1} s^{-1}) under our experimental conditions employed. However, significant crowding

by large macromolecules (500 kDa and higher) appears to decrease the probability of enzyme-substrate encounters resulting in the observed profile shown in **Figure 4.2**.

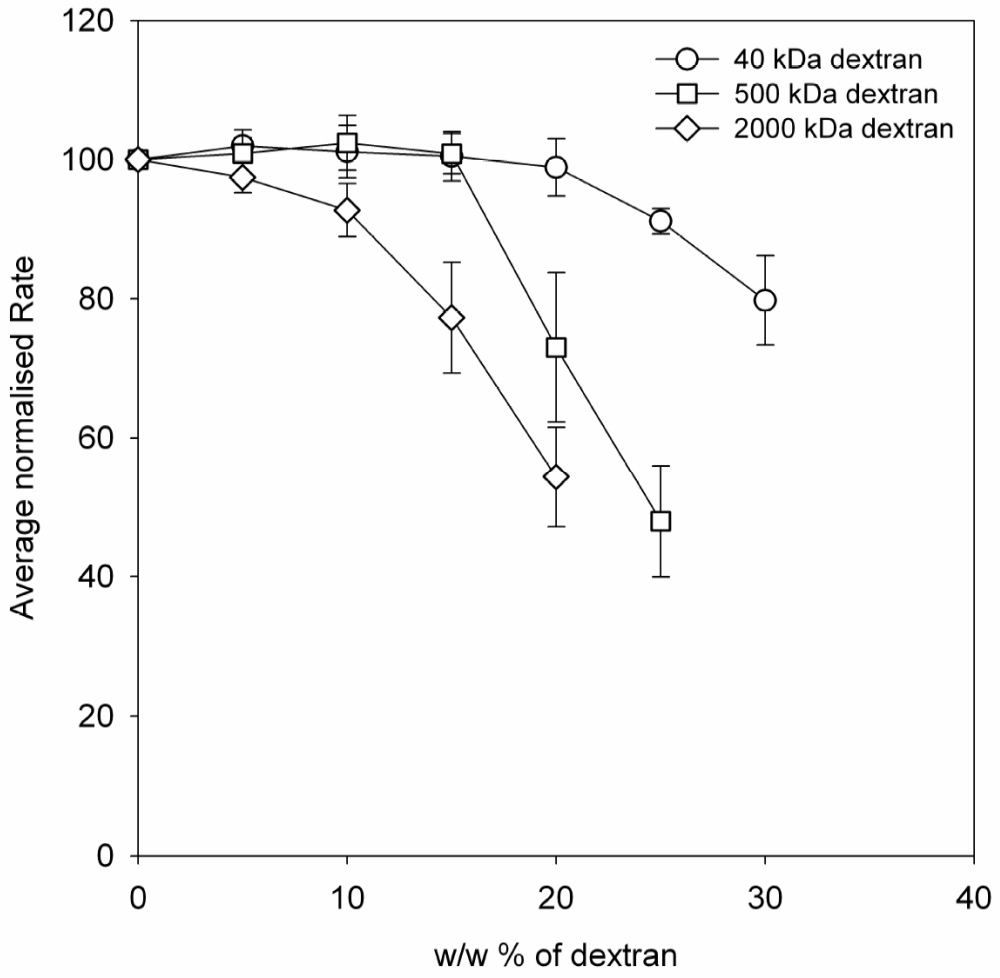

Figure 4.2: *Dependence of enzymatic rate of an alkaline phosphatase-catalysed hydrolysis of PNPP (at 1 mM) on size and concentration of dextrans is shown. The average normalized rate is plotted against the different w/w % of dextrans of different sizes. The dextran sizes are as follows: circle, 40 kDa; square, 500 kDa; diamond, 2000 kDa. The error bars show the range of average normalized rate values obtained from different experiments done on different days.*

Although the effect of crowding by different sizes of dextran (15 – 500 kDa) has been reported by us before (**Homchaudhuri, *et al.*, 2006**), this experiment was repeated for two reasons: a) The approach presented here uses a more efficient method to mix enzyme and substrate and b) The present study uses a much larger dextran size (40 to 2000 kDa) unlike the last attempt. Comparison of results in **Figure 4.2** with our previous

work, clearly emphasizes the importance of efficient mixing of reactants. Importantly, the trend observed with the reaction rates for different sizes of dextran appears same.

Now we investigate the effect of crowding on kinetics of AchE vs. IA and AchE vs NA. **Figure 4.3 A, B, C, and D** depict the typically observed initial rise and subsequent plateau in the absorbance due to formation of the product 3-hydroxy indole,

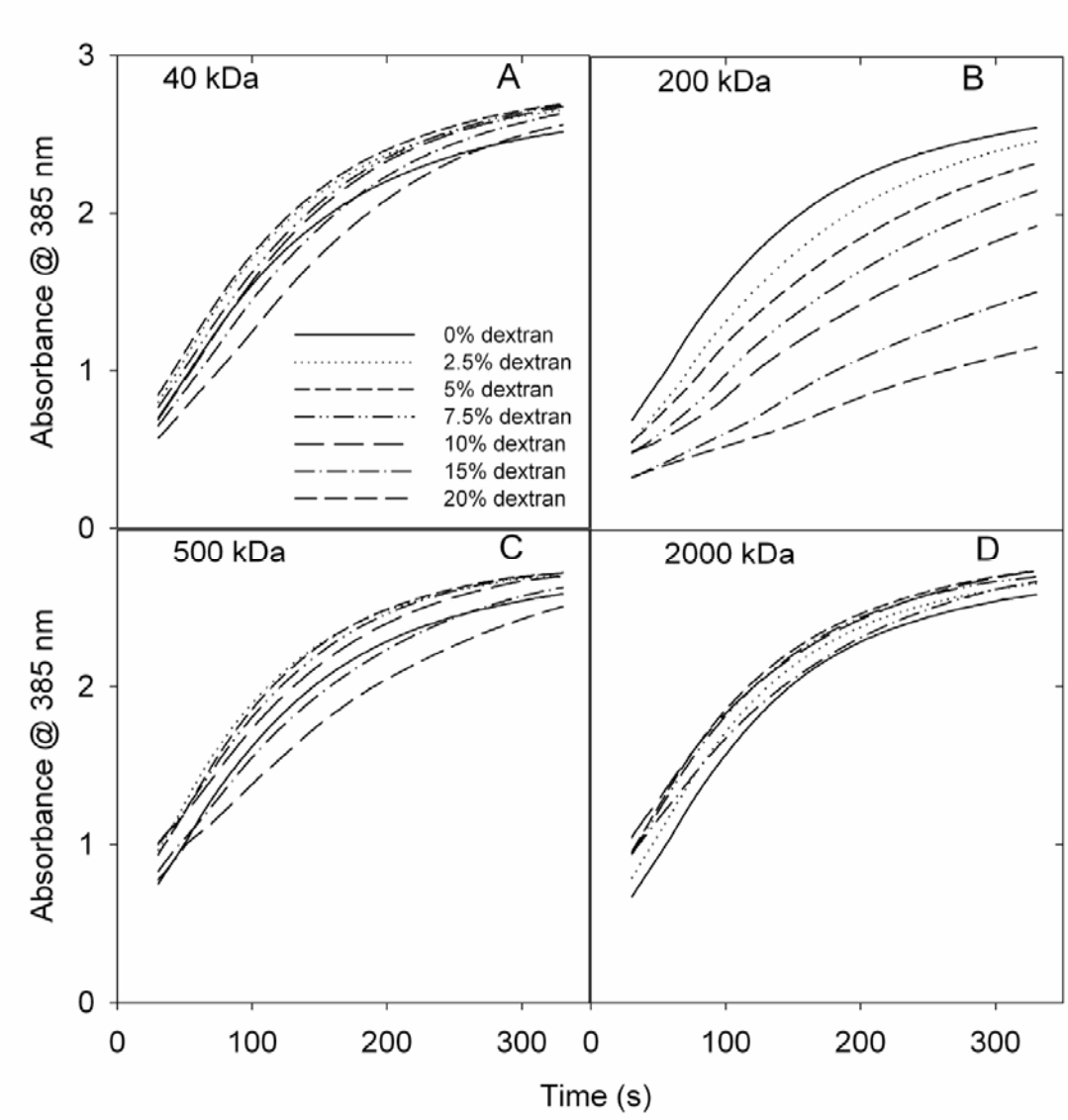

Figure 4.3: *The typical reaction profile of an Acetyl cholineesterase-catalysed hydrolysis of 3-indoxyl acetate (at 5 mM) monitored by measuring the absorbance at 385 nm is shown. (A) In presence of 40 kDa dextran. (B) In presence of 200 kDa. (C) In presence of 500 kDa. (D) In presence of 2000 kDa. The different w/w % of dextrans are represented by the symbols as follows: solid line, 0%; dotted, 2.5%; short dash, 5%; dash dot dot, 7.5%; long dash, 10%; dash dot, 15%; medium dash, 20%.*

during the hydrolysis of IA in the presence of increasing amounts of 40, 200, 500 and 2000 kDa dextran in the reaction medium, respectively. In **Figure 4.3 A**, we observe a marginal increase in the initial velocity of the reaction, followed by decrease with increasing amounts of 40 kDa dextran. **Figure 4.3 C & D** also show a similar trend with 500 & 2000 kDa dextrans respectively. However, there is a major decline in the initial velocity when increasing amounts of 200 kDa dextran was added to the medium. With 40, 500 & 2000 kDa dextran, the reaction begins to slows down from 15% w/w onwards.

Figure 4.4 A, B, C, D, E, and F depict the characteristically observed initial rise and subsequent plateau in the absorbance due to formation of the product 2-naphthol, during the hydrolysis of NA in the presence of increasing amounts of 15-20, 40, 70, 200, 500, and 2000 kDa dextran in the reaction medium, respectively. A rise and subsequent decline in reaction rate is clearly noticed for all dextran sizes with increasing concentrations.

The effect of different concentrations of different sizes of dextran, taken individually on the normalized reaction rate of AchE vs. IA and AchE vs. NA is shown in **Figure 4.5 & 4.6** respectively. In the case of AchE vs. IA, it can be seen from **Figure 4.5** that there is a clear difference in profile between smaller (40 kDa) and larger dextrans (500 and 2000 kDa). It is observed that smaller dextrans (40 kDa) do not show considerable decrease in rate until reaching a concentration of 10 % w/w. While larger dextrans (500 and 2000 kDa) shows a marginal increase at 2.5 & 5 % w/w followed by decrease at 10% w/w and higher concentrations. For the intermediate 200 kDa dextran, we observed a steep decline in rate, quite in contrast to remaining sizes of dextran. The profile observed with 500 and 2000 kDa dextrans is similar to that expected when the rate of reaction is limited by the diffusional encounter between AchE and IA. The AchE vs. IA reaction, in the absence of crowding is not diffusion limited (k_{cat}/K_m = (2.6 ± 0.12) x 10^4 M^{-1} s^{-1}) under our reaction conditions. Here, also we observe crowding by large macromolecules (500 kDa and higher) appears to decrease the enzyme-substrate encounters resulting in the observed profile shown in **Figure 4.5**. Thus it is apparent that smaller dextrans in contrast to larger dextrans do not affect the enzyme-substrate encounters unless they are present in high concentrations.

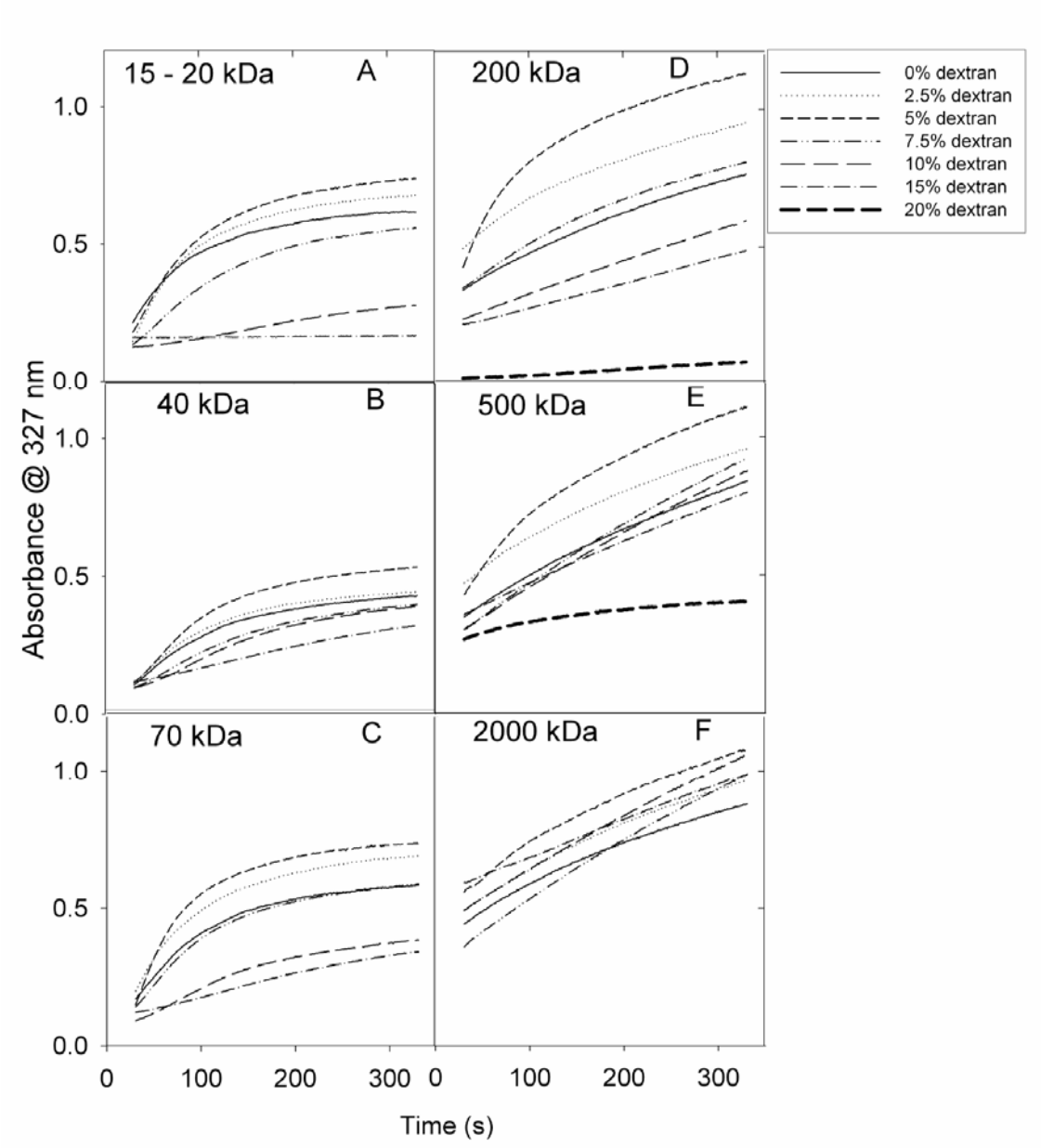

Figure 4.4: *The typical reaction profile of an Acetyl cholinesterase-catalysed hydrolysis of 2-naphthyl acetate (at 1.5 mM) monitored by measuring the absorbance at 327 nm is shown. (A) In presence of 15-20 kDa dextran. (B) In presence of 40 kDa. (C) In presence of 70 kDa. (D) In presence of 200 kDa. (E) In presence of 500 kDa . (F) In presence of 2000 kDa. The different w/w % of dextrans are represented by the symbols as follows: solid line, 0%; dotted, 2.5%; short dash, 5%; dash dot dot, 7.5%; long dash, 10%; dash dot, 15%; medium dash, 20%.*

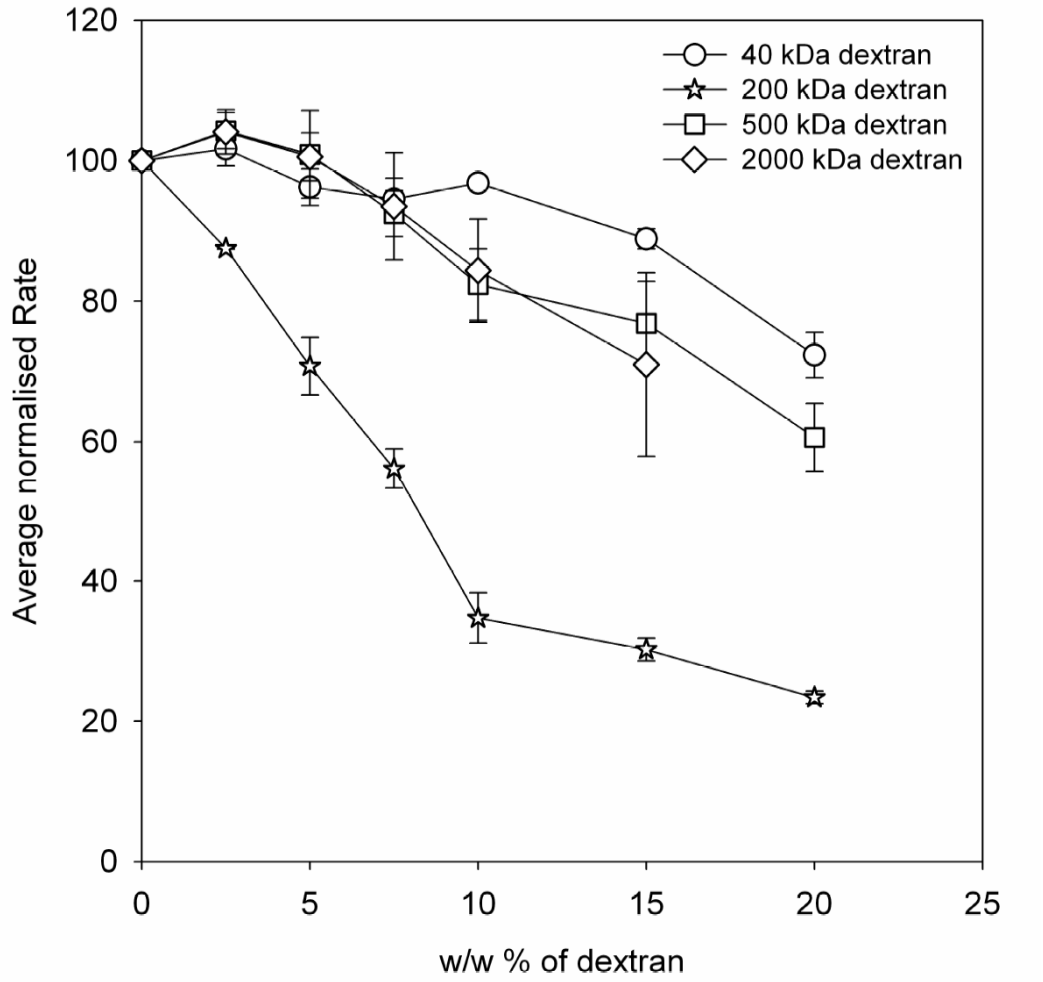

Figure 4.5: *Dependence of enzymatic rate of an Acetyl cholinesterase-catalysed hydrolysis of 3-indoxyl acetate (at 5 mM) on size and concentration of dextrans is shown. The average normalized rate is plotted against the different w/w % of dextrans of different sizes. The dextran sizes are as follows: circle, 40 kDa; star, 200 kDa; square, 500 kDa; diamond, 2000 kDa. The error bars show the range of average normalized rate values obtained from different experiments done on different days.*

In the case of AchE vs. NA (**Figure 4.6**), it can be seen that in smaller (15-20, 40, 70 kDa) as well as in larger dextrans (200, 500 and 2000 kDa), the reaction profile follows the trend that is clearly different from the profile observed in the two enzymatic reactions above. We observe increase in rate of reaction at lower w/w % of smaller as well as larger dextrans. This increase in rate was observed to be maintained till % w/w of dextran reaches 5% in the case of 15-20, 40, 70, 200, and 2000 kDa dextran. But in the

case of 500 kDa the increase in rate was observed to continue till the %w/w reaches 7.5%.

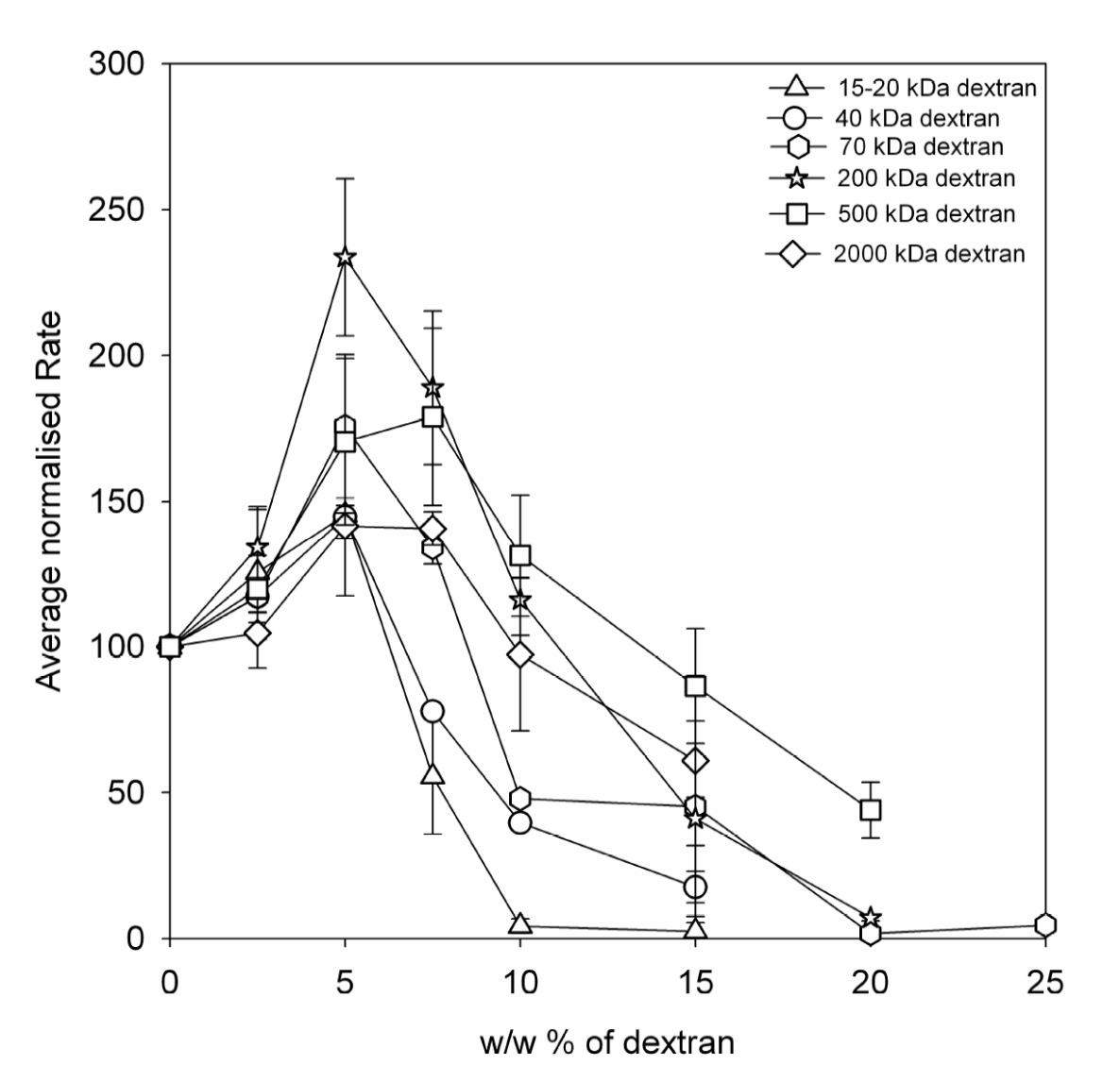

Figure 4.6: *Dependence of enzymatic rate of an Acetyl cholinesterase-catalysed hydrolysis of 2-naphthyl acetate (at 1.5 mM) on size and concentration of dextrans is shown. The average normalized rate is plotted against the different w/w % of dextrans of different sizes. The dextran sizes are as follows: triangle, 15-20 kDa; circle, 40 kDa; hexagon, 70 kDa; star, 200 kDa; square, 500 kDa; diamond, 2000 kDa.*

The 200 kDa dextran shows marked increase in rate at 5% w/w in comparison with other dextrans. This is followed by 70, 500, 40, 15-20, and 2000 kDa dextran. It is likely that, this reaction is transition state limited and thus there is increase in rate due to increase in thermodynamic activity of the enzyme. But with increase in concentration of

dextran, the rate eventually decreases owing to increase in crowding that hinders the enzyme-substrate encounters resulting in the observed profile shown in **Figure 4.6** and thus the reaction becomes diffusion limited at higher w/w % of dextran (> 10%). The occurrence of bell shape profile with increase in crowded concentration has been proposed previously (**Ellis, 2001b**). We have observed that the AchE vs. NA reaction, in the absence of crowding is not diffusion limited (k_{cat}/K_m = (2.18 ± 0.18) x 10^4 M^{-1} s^{-1}) under the conditions employed by us.

We also looked into the effect of solution viscosity on the rate of AchE vs. IA and AchE vs. NA reactions. This is shown in **Figure 4.7** respectively. In both the cases, the dependence of reaction rate on bulk solution viscosity was fairly identical. This was expected as both substrates are fairly similar in size and both encounter the same enzyme.

4.4. Discussion:

Our earlier work using PNPP and alkaline phosphatase had revealed a marginal slowing (< 2 fold) with smaller dextrans (15-70 kDa) and appreciable slowing (> 5 fold) with larger dextrans (200 kDa or more) (**Homchaudhuri, *et al.*, 2006)**. With a more efficient mixing of enzyme, substrate and dextrans before onset of reaction it is observed from the current observations that the extent of slowing in the rate of AP/PNPP reaction is 2-fold far less for all dextrans sizes in the range 40- 2000 kDa. This work presents two interesting results. Firstly the effect of crowding by identical dextrans seems to have opposite effects on two different substrates IA and NA. This was indeed surprising. It implies that increase in activity of $ES^{\neq}$ complex is selective on the substrate. Such an event could imply different reaction mechanisms for the two substrates with acetylcholinesterase.

Acetylcholinesterase from electric eel has been shown to be a tetrameric enzyme (**Chothia and Leuzinger, 1975**) in its active state. Hence it is unlikely that this enhancement in its activity is due to further oligomerization under crowded conditions.

Figure 4.7 shows that effect of viscosity is identical for both substrates, suggesting that it has no role to play on the contrasting kinetic profiles observed with two different substrates.

The second interesting result is the effect of 200 kDa dextran. Crowding by this dextran size clearly stimulates the reaction at low concentrations with NA, while it

inhibits the rate of reaction appreciably with IA unlike other dextrans of larger size. Napthyl acetate and Indoxyl acetate are to a large extent similar in nature, hence it is difficult to correlate these observations with their chemical composition. Increase in rate of enzymatic reaction has been reported in the past (**Ellis, 2001b**).

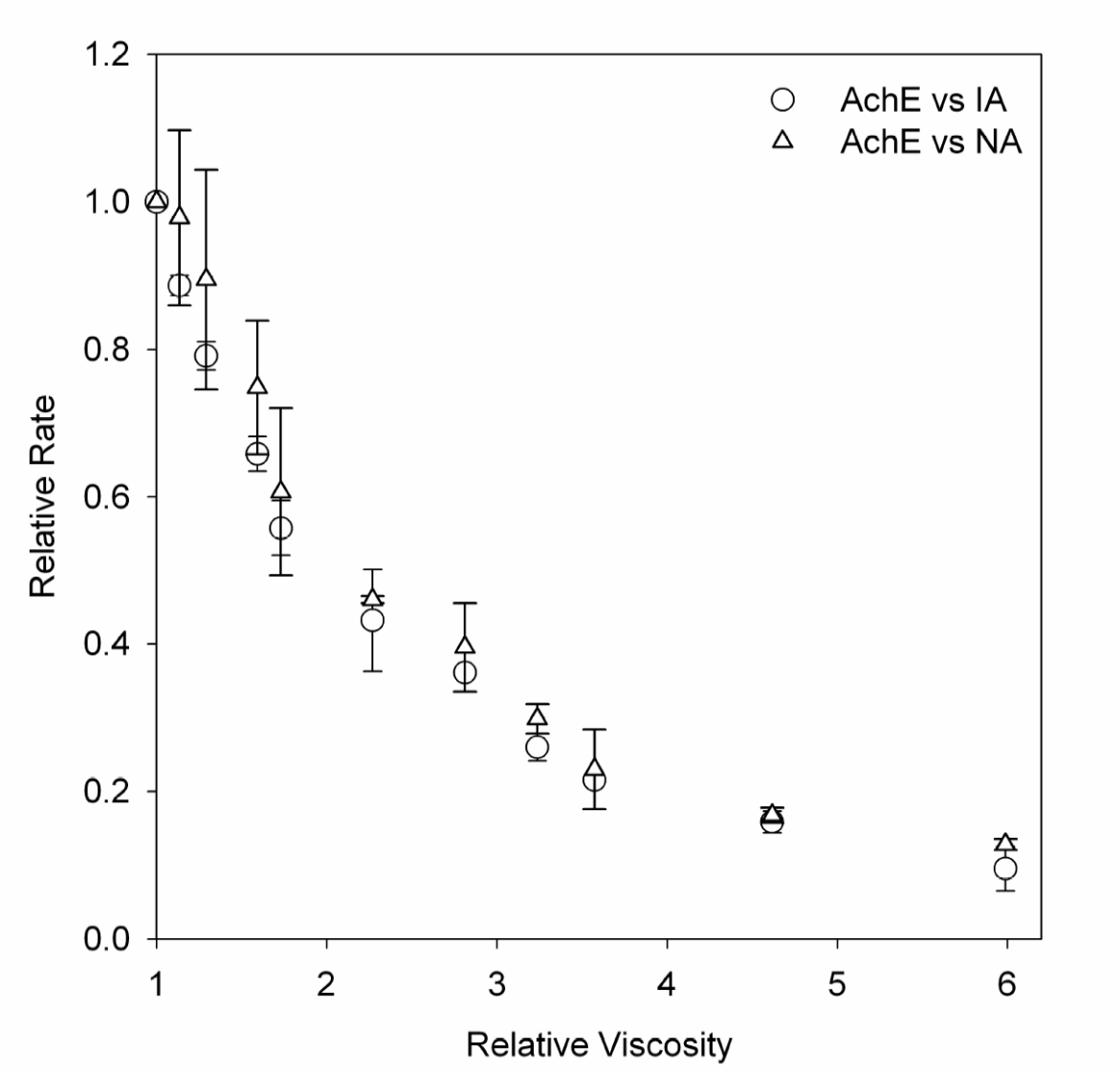

Figure 4.7: *Dependence of the rate on solution viscosity is shown. Circle represents the Acetyl cholinesterase-catalysed hydrolysis of 3-indoxyl acetate (at 5 mM). Triangle represents the Acetyl cholinesterase-catalysed hydrolysis of 2-naphthyl acetate (at 1.5 mM). The relative rate of hydrolysis of reaction is plotted against the relative viscosity in glycerol-water mixtures. Other conditions are similar to those described under experimental. The error bars show the range of relative rate values obtained from different trials of experiment.*

Jiang and Guo (2007) have observed increased k_{cat}/K_m with isochorismate synthase in 30% w/v Ficoll. They attribute the increase to change in structure of enzyme. However such a scenario is unlikely to result in opposite effects with two different substrates observed here. **Derham and Harding (2006)** have shown that decarboxylating enzymes like urease, glutamate decarboxylase and pyruvate decarboxylase can have upto- 10-fold increase in rate in presence of 30 wt-% hemoglobin or lysozyme. However they observe a steady decrease in rate in the midst of crowding by 10 and 120 kDa dextran. They attribute the rate increase to decreased amount of free water and self-association of enzyme. Our results thus show that effect of crowding on enzymatic reactions is still far from being fully understood and predicted.

4.5. Conclusions:

In this work, we have used a newer method to ensure efficient mixing of large molecular weight dextrans with enzyme. Our findings depicts that size and concentration of macromolecule play a crucial role in influencing the rate of an enzymatic reaction. The effect of crowding by smaller dextrans (40 kDa) showed minor decrease in rate in comparison with larger dextrans (500 and 2000 kDa) in the case of AP vs PNPP. We observed the effect of crowding by dextrans to have opposite effects on two different substrates IA and NA with acetyl cholinesterase. The effect of crowding by 200 kDa dextran size clearly stimulates the reaction at low concentrations with NA, while it inhibits it appreciably with IA unlike other dextrans of larger size. This finding shows that the increase in activity of $ES^{\neq}$ complex is selective on substrate. Thus our results reveal that the effect of crowding on enzymatic reactions is not simple as it appears and depends on crowder size and substrate nature.

CHAPTER 5
Human Lysozyme Amyloidosis: Insights from the Conformational Dynamics of Wild type and Mutants (I56T), (D67H), and (T70N) using Molecular Dynamics Simulations.

Human Lysozyme Amyloidosis: Insights from the Conformational Dynamics of Wild type and Mutants (I56T), (D67H), and (T70N) using Molecular Dynamics Simulations.

5.1. Introduction:

Amyloidosis refers to a condition wherein protein aggregates into insoluble ordered structures that share a common fibrillar conformation known as amyloid. Amyloid fibril formation has been linked to disorders that may be categorized as "protein deposition diseases". This includes several pathological conditions such as Alzheimer's disease, Parkinson's disease, the spongiform encephalopathies, systemic amyloidosis and so on. So far, more than 40 human diseases have been attributed to amyloid deposition and each one observed to have a distinct clinical profile (**Chiti & Dobson, 2006**). Among the amyloid forming proteins, there is no sequence similarity and folding analogy, this made Dobson to conclude that the ability to form cross β-structure, wherein hydrogen bonds are formed between polypeptide chains in directions parallel to the fiber axis, is a generic property of polypeptide chains (**Dobson, 1999**).

In most of the amyloidosis, especially extracellular amyloidosis, the formation of amyloid is not because of errors in the normal folding pathway of a protein but due to decreased structural stability, unfolding events or even conformational fluctuations around the native structure ensemble, allowing for alternative stable conformations. Mutations, in particular, seem to play a critical role in the decrease of protein conformational stability and in triggering unfolding events (**Brito *et al.*, 2003**).

The formation of fibrils by certain proteins that display a globular structure in native state, such as lysozyme or transthyretin and β2-microglobulin, still remains a challenging issue in protein biochemistry. The common feature of all these proteins is their relative structural instability and the capability to change conformation adopting the common beta sheet fibrillar structure (**Merlini and Bellotti, 2003**). Several studies are reported to support the conformational change hypothesis (**Lai *et al.*, 1996; Wetzel, 1996; Funahashi *et al.*, 1996**) and are now believed to be the main cause leading to fibril formation in conditions of protein denaturation and partial unfolding (**Forloni *et al.*, 1993**).

The discovery that lysozyme can cause systemic amyloidoses offered a unique opportunity to explore in details the relationship between structure and folding in amyloid proteins exploiting the available wealth of information on structure and folding of lysozyme. **Pepys *et al.*, (1993)** identified that single point mutations in human lysozyme gene associated with hereditary systemic amyloidosis. This has led to extensive study on lysozyme amyloid fibrils and there are reports that have shown several proteins in lysozyme family to be capable of amyloid fibril formation *in vitro* including several lysozymes (human, hen, turkey, equine) and two α-lactalbumin (bovine, human) proteins. All of these proteins share significant sequence identity and are structurally similar (**Boeckmann *et al.*, 2003; Huang and Miller, 1991**).

Human lysozyme is a protein with 130 residues belonging to the c-type class of lysozymes and found in secretions and more generally in leukocytes and kidneys. It is coded by a gene located on chromosome 12 and organized in 4 exons and 3 introns (**Peters *et al.*, 1989**). It is an enzyme that cleaves preferentially β-1,4 glycosidic linkages between the N-acetylmuramic acid and N-acetylglucosamine that occur in the peptidoglycan cell wall structure of certain microorganisms and has an enzyme classification number of 3.2.1.17. It has a bipartite structure composed of two independent domains. The larger alpha domain composed of residues 1-42 and 81-130 with four alpha helices and one 3_{10} helix and the smaller beta domain composed of residues 43-80 with a three-stranded antiparallel beta sheet and an irregular loop. Its active site is located in a cleft that is formed between these two domains. So far, five mutations in human lysozyme gene have been reported that give rise to six variant proteins (I56T, F57I, W64R, D67H, T70N, and the double mutants F57I / T70N, T70N/W112R). Four of these (I56T, F57I, W64R, D67H) have been associated with systemic amyloidosis involving the kidney, liver and spleen (**Pepys *et al.*, 1993; Valleix *et al.*, 2002; Yazaki *et al.*, 2003**) while the variant T70N is not amyloidogenic and is common in the normal British population (**Booth *et al.*, 1997**). But it has been reported that T70N results in amyloidosis if present as double mutant T70N/W112R (**Rocken *et al.*, 2006**) or F57I/T70N (**Yazaki et al., 2003**).

Booth *et al.* (1997) proposed that partly folded forms of amyloidogenic proteins which lack global co-operativity undergo a helix to sheet transition and form the initial

seed for the generation of amyloids. **Morozova-Roche *et al.* (2000)** showed the existence of template seed for the wild type and the two natural mutants of human lysozyme that assist the formation of fibrils. The structural details of species formed early during the aggregation process and their features which trigger amyloidosis seems to be important. So it will be worthwhile to investigate the structural feature of the partially folded structures of wild type and the mutants of human lysozyme that may trigger amyloid formation. Recently, **Chiti and Dobson (2009)** has reported about the amyloid formation by globular proteins under native conditions. They showed that a transition across the major energy barrier for unfolding is not essential and that aggregation may well be initiated from locally unfolded states that become accessible, for example, via thermal fluctuations occurring under native conditions.

In this work, we have compared the conformational dynamics of wild type and mutants of Human lysozyme using MD simulations under native conditions. The crystal structures are available for wild type and three of the human lysozyme mutants: D67H, I56T, and T70N. T70N is the only known naturally occurring destabilized mutant of human lysozyme that has not been observed in amyloid deposits in human patients. But for understanding the determinants of amyloid disease, it is important to study and compare the properties of T70N mutant with those of amyloidogenic mutants of human lysozyme. In general these structures are very much similar to the wild type (WT) protein. All the four simulations (20 ns each) were performed using ff99SB Amber force field to investigate the conformational dynamics. We have analyzed the trajectories arising from these simulations to obtain insights on conformational features triggering amyloidosis.

5.2. Materials and Methods:

The initial structure for the molecular dynamics simulation of proteins with PDB codes 1REX (wild type), 1LOZ (I56T), 1LYY (D67H), and 1W08 (T70N) were downloaded from PDB. The LEaP module of the AMBER program package (**Pearlman *et al.*, 1995; Case *et al.*, 2005**) was used to prepare the system for simulation. Each protein was solvated with TIP3P (**Mahoney and Jorgensen, 2000**) waters and neutralized with the counter ions using the LEaP module. Energy minimization and MD simulations were carried out using the SANDER module of AMBER 8. The Amber force

field ff99SB (**Hornak *et al.*, 2006, Wickstrom *et al.*, 2009**) is used to describe the atomic interactions. For the correct treatment of long range electrostatics, we make use of the Particle Mesh Ewald (PME) method (**Essmann *et al.*, 1995**). Constant temperature and pressure conditions in the simulations were achieved by coupling the system to a Berendsen's thermostat and barostat (**Berendsen *et al.*, 1984**). Bonds involving the hydrogen atoms were constrained to their equilibrium position with the SHAKE algorithm. For these simulations, we used an HP Proliant server with eight processors.

The system was minimized in two phases to avoid bad contacts. In the first phase, the system was minimized giving restraints (30 kcal/mol/$Å^2$) to protein and crystallographic waters for 500 steps with subsequent second phase minimization of the whole system. Then the system was heated to 300K over 50 ps with a 1 fs time step. The protein atoms were restrained with force constant of 30 kcal/mol/$Å^2$ at the NVT ensemble. After that the force constant was reduced by 10 kcal/mol/ $Å^2$ in each step to reach the unrestrained structure in three steps of 10 ps each. The system was then switched over to the NPT ensemble and equilibrated without any restraints for 180 ps. The system was equilibrated in total of 260 ps. The time step for MD simulation for the production run was 2 fs. All the four trajectories were each run for 20 ns and were performed with an 8.0 Å cutoff on real-space interactions. Analysis of parameters of the trajectories was carried out using the ptraj modules of AMBER 8. Graphic visualization of protein structures was done using Chimera.

The RMSD, secondary structure analysis, B factor values, solvent accessible surface area, radius of gyration analysis, End to end chain distance, water movement analysis, distance matrix analysis and hydrophobic contact analysis were carried out using ptraj action commands. The conformational entropy for wild type and mutant human lysozyme were calculated using gas phase statistical mechanics. This involves principal component analysis (the quasi-harmonic approximation) that provides the first decomposition of correlations in particle motion (**Andricioaei and Karplus, 2001**). Thus the entropy is calculated analytically as a sum of independent quantum harmonic oscillators. Origin 6.1 was used to plot the secondary structure analysis, water movement analysis, distance matrix analysis and hydrophobic contact analysis.

5.3. Results:

5.3.1. Root Mean Square Deviation (RMSD):

The degree of conformational changes for the wild type and mutants of human lysozyme during the simulations is simply monitored by the C-α root mean square deviation (RMSD). The backbone RMSD of structures relative to the lowest-energy conformation has been calculated and is represented in **Figure 5.1 A, B, C, and D**. The RMSD values for the wild type (**Figure 5.1 A**) and mutants I56T (**Figure 5.1 B**), D67H (**Figure 5.1 C**), and T70N (**Figure 5.1 D**) remains fairly constant with the time course of simulation and settles well below 2 Å. This implies that wild type and the mutants of human lysozyme possess relatively stable structure. Among the wild type and mutants, in D67H mutant, we observe most of the times the RMSD values are at higher side. This happens in D67H perhaps due to the existence of high population of conformers that are relatively less stable than the initial lowest energy conformation. The individual average and standard deviation of RMSD values for the wild type and mutants are shown in **Table 5.1**. The order of stability from the average RMSD values (see **Table 5.1**) was observed to be D67H < T70N < I56T & wild type. The above results indicate that structural integrity is affected to certain extent in the case of mutants than in wild type. Thus mutants appear comparatively less stable than wild type. We have also calculated the RMSD of mutants with respect to wild type as initial structure. From the values obtained we noticed that among the three mutants, I56T shows least variation in global structure with respect to wild type. The other two mutants (D67H and T70N) showed appreciable differences.

5.3.2. Radius of Gyration (R_g):

To investigate the compactness of protein during the course of simulation, radius of gyration was determined. Information regarding the overall shape of the molecule can be gleaned from R_g. The time course of the radius of gyration (R_g) for the wild type and mutants are shown in the **Figure 5.2 A, B, C, and D**. The degree of packing of amino acid residues in the protein can be known from the radius of gyration. This parameter is known to affect both the stability and folding rate of proteins. It is observed that wild type (**Figure 5.2 A**) and mutants I56T (**Figure 5.2 B**), D67H (**Figure 5.2 C**), and T70N (**Figure 5.2 D**) show more or less similar profile in R_g during the course of simulation.

The order of average R_g values (see **Table 5.1**) for the wild type and mutants are observed to be D67H > T70N > I56T > wild type. This trend is quite similar to that observed above for the RMSD analysis.

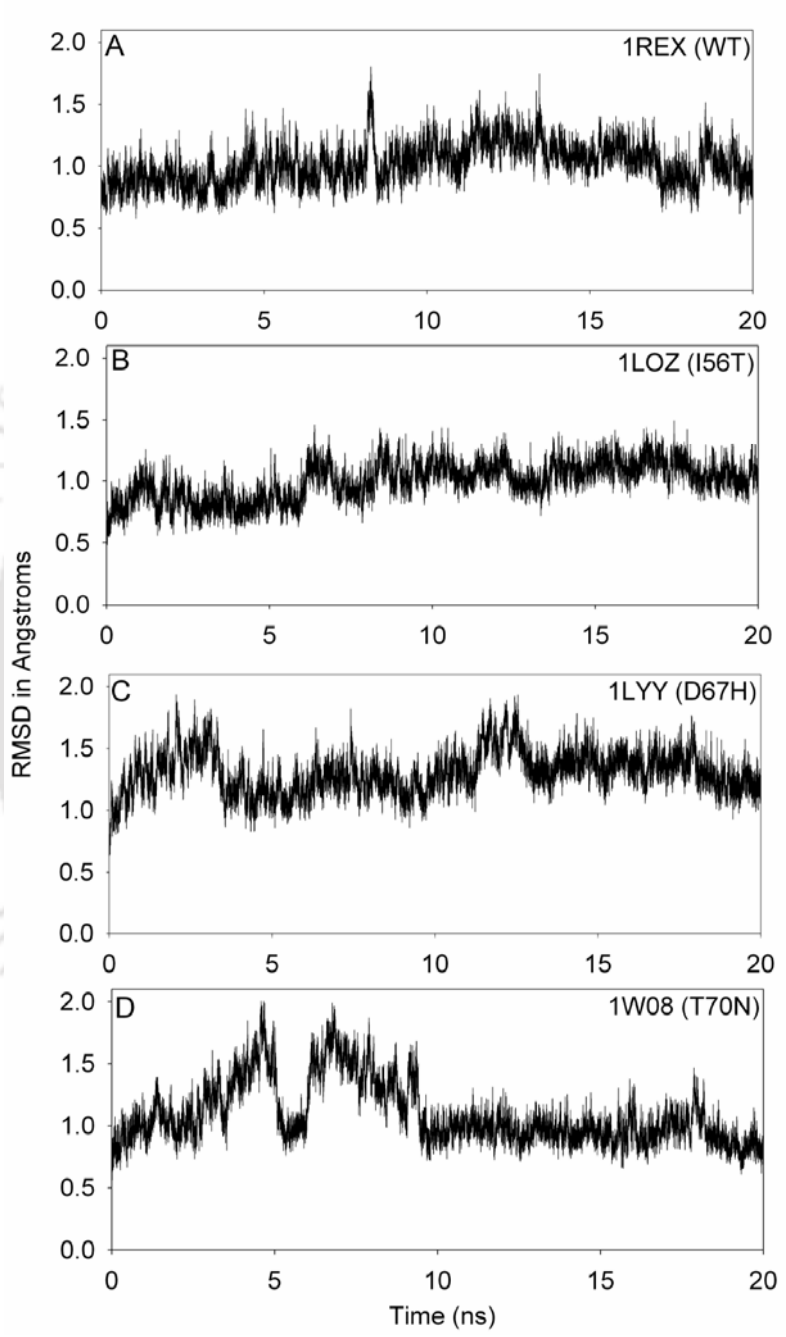

Figure 5.1:`*Root-mean-square deviation (RMSD) to the starting structure as a function of time is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N). Calculations were performed for the backbone atoms of the respective structure versus the backbone atoms of the respective simulation's starting structure.*

Table 5.1: *Summary of analyzed trajectory parameters for wild type and mutants of Human lysozyme*

PDB code	Variant	RMSD (Å)	RMSD (Å) In reference to 1REX	End to end distance (Å)	Solvent accessible surface area ($Å^2$)	Radius of gyration (Å)	Normalized Conformational entropy (cal/mol-Kelvin-residue)
1REX	WT	1.01 ± 0.16	0	21.66 ± 0.71	7232 ± 147	14.38 ± 0.08	93.81
1LOZ	I56T	0.99 ± 0.15	0.83 ± 0.01	22.60 ± 1.09	7283 ± 128	14.43 ± 0.08	95.52
1LYY	D67H	1.30 ± 0.17	1.36 ± 0.08	21.66 ± 0.59	7550 ± 158	14.52 ± 0.09	100
1W08	T70N	1.09 ± 0.25	1.13 ± 0.28	21.60 ± 0.58	7383 ± 140	14.44 ± 0.09	97.44

5.3.3. B-factor values:

B-factors are the simplest method to analyze local deformability in a protein chain. The B factor values for the backbone C-alpha atoms in wild type and mutants are calculated from MD simulations and are plotted against their residue number as shown in **Figure 5.3 A, B, C, and D**. It has been observed that B factor values of C-alpha atoms in the beta domain (43-80) region shows appreciable difference for the wild type (**Figure 5.3 A**) and mutants (**Figure 5.3 B, C, and D**). This is mainly due to the presence of mutation sites in this beta domain region. In D67H (**Figure 5.3 C**), the residues around the point of mutations are disturbed noticeably resulting in the higher B factor values. While in T70N (**Figure 5.3 D**), the residues around the point of mutation are marginally perturbed. All mutants reveal reduced B-factor values around 100-110 in contrast to wild type. The above results reveal that mutations in the beta domain region make this region and other regions (100-110) more flexible and accounts for frequent change in the structure leading to the formation of diverse conformers.

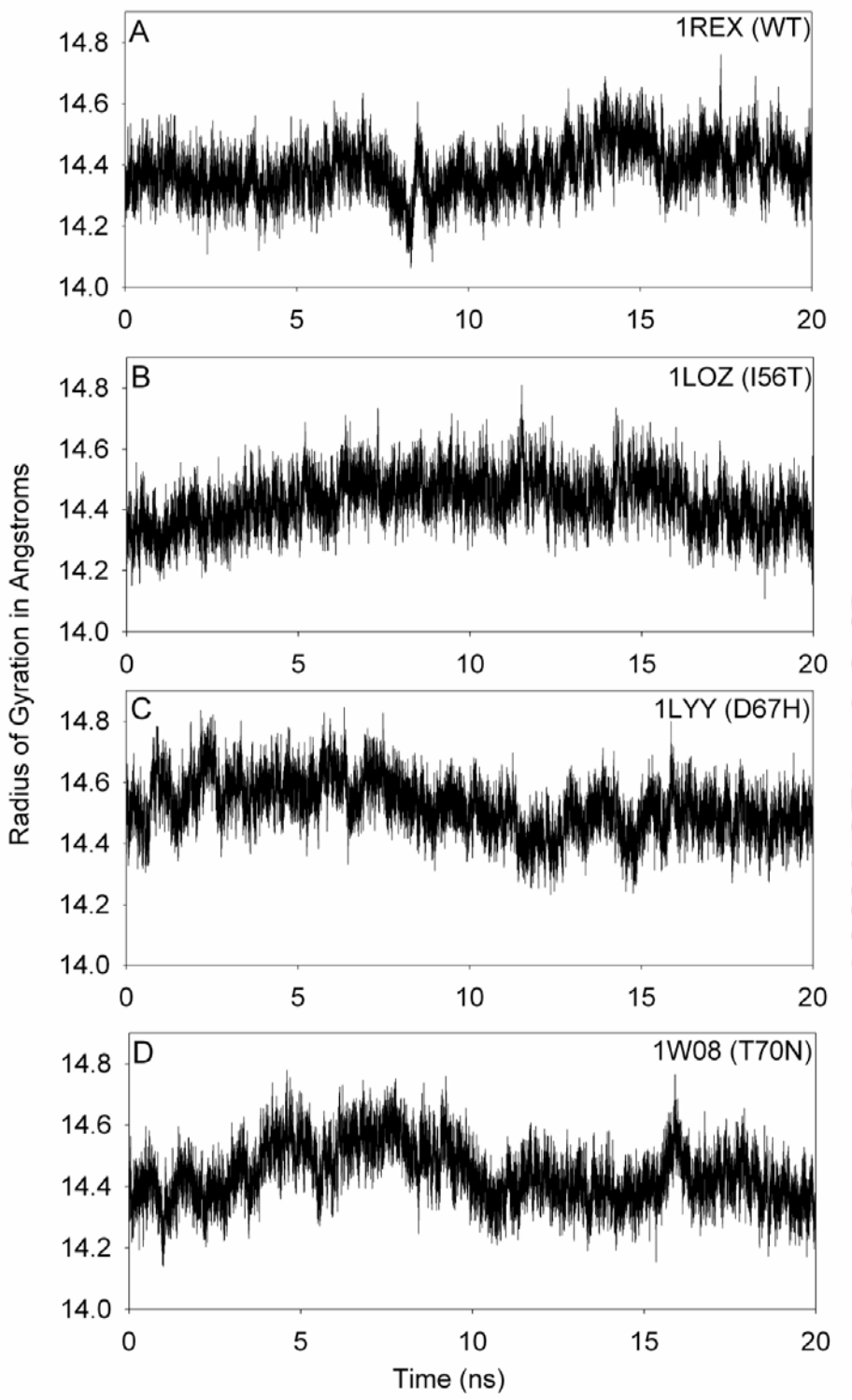

Figure 5.2:`*Radius of gyration of α-carbon atoms as a function of time is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

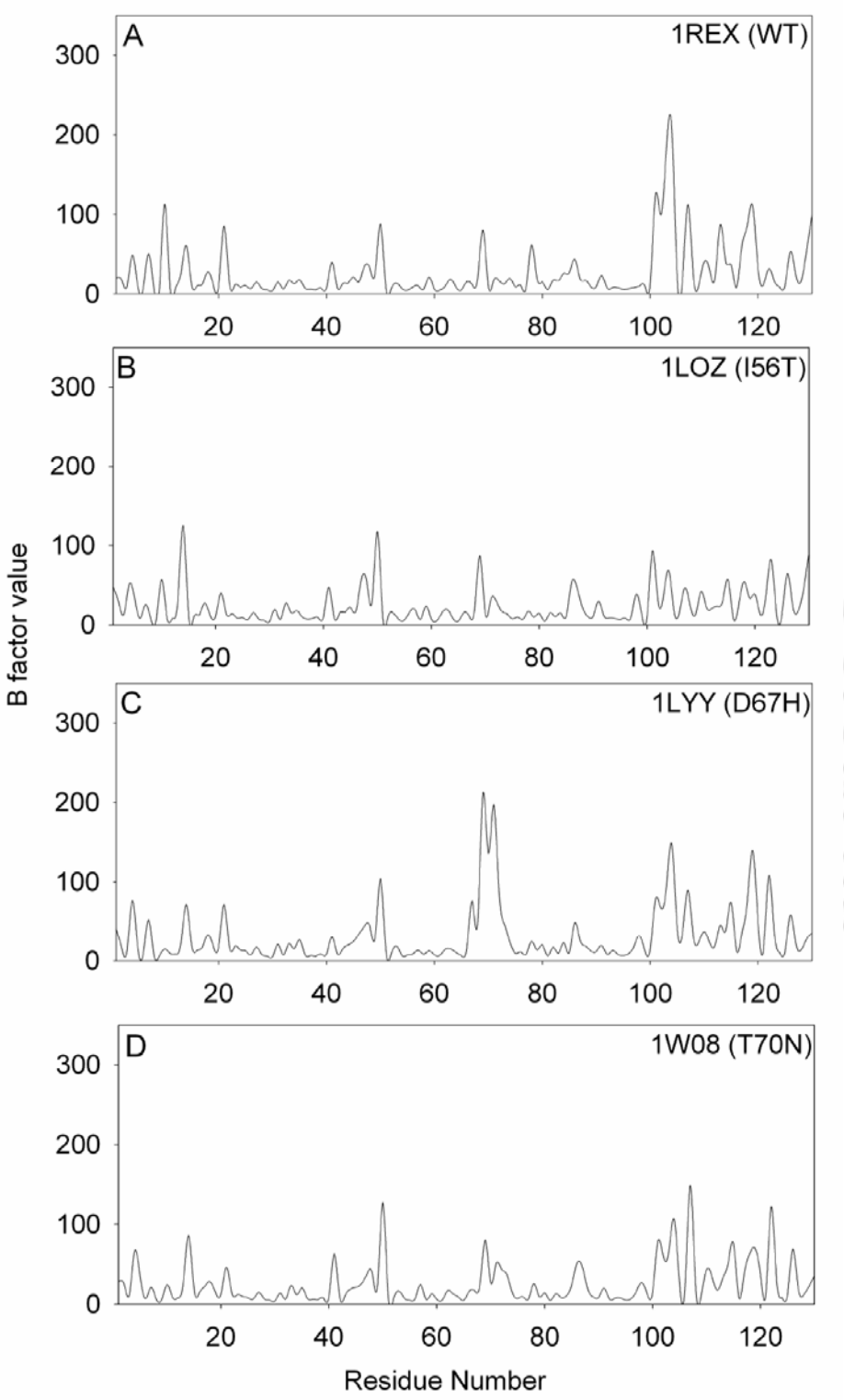

Figure 5.3:`*B-factor values of α-carbon atoms as a function of residue number is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

5.3.4. Solvent Accessible Surface Area (SASA):

In order to investigate the finer details about the mobility of flexible regions in the wild type and mutants, we calculated the SASA. The results are depicted in **Figure 5.4 A, B, C, and D**. The degree of mobility of the flexible hydrophobic regions in the wild type and mutants can be obtained by observing the time course of SASA. The average SASA values (see **Table 5.1**) for mutants are found to be larger than the wild type and follows the order D67H > T70N > I56T > wild type. This is especially because of the flexible and unfolded regions in the beta domain region in mutants. This trend is consistent with the trend observed for RMSD and radius of gyration. The results reflect that the flexible and unfolded regions in the beta domain region of mutants often get exposed to the solvent owing to their mobility (which can be inferred from the above B factor analysis). In wild type we see high fluctuations in SASA (**Figure 5.4 A**) but at the same time the average SASA value was found to be lower (see **Table 5.1**). This is so because the hydrophobic groups in the wild type are folded inside and shows flexibility even in buried conditions. As a whole we observe the existence of flexible and exposed hydrophobic regions in the beta domain of mutants unlike the wild type.

5.3.5. End to End chain distance (between first C-alpha atom and last C-alpha atom):

This is another trajectory parameter that gives information about the structural integrity of the protein during the simulation period. The distance between the first and last C^{α} atom was calculated during the time course of simulation for all the proteins under study and the results are depicted in the **Figure 5.5 A, B, C, and D** for wild type and mutants of human lysozyme. We observe a similar trend with wild type and all mutants except I56T. In I56T, we noticed large changes in end to end distance values during the time course of simulation. **Table 5.1** clearly reflects this observation where the std. dev. for I56T is significantly more in comparison to wild type and other mutants.

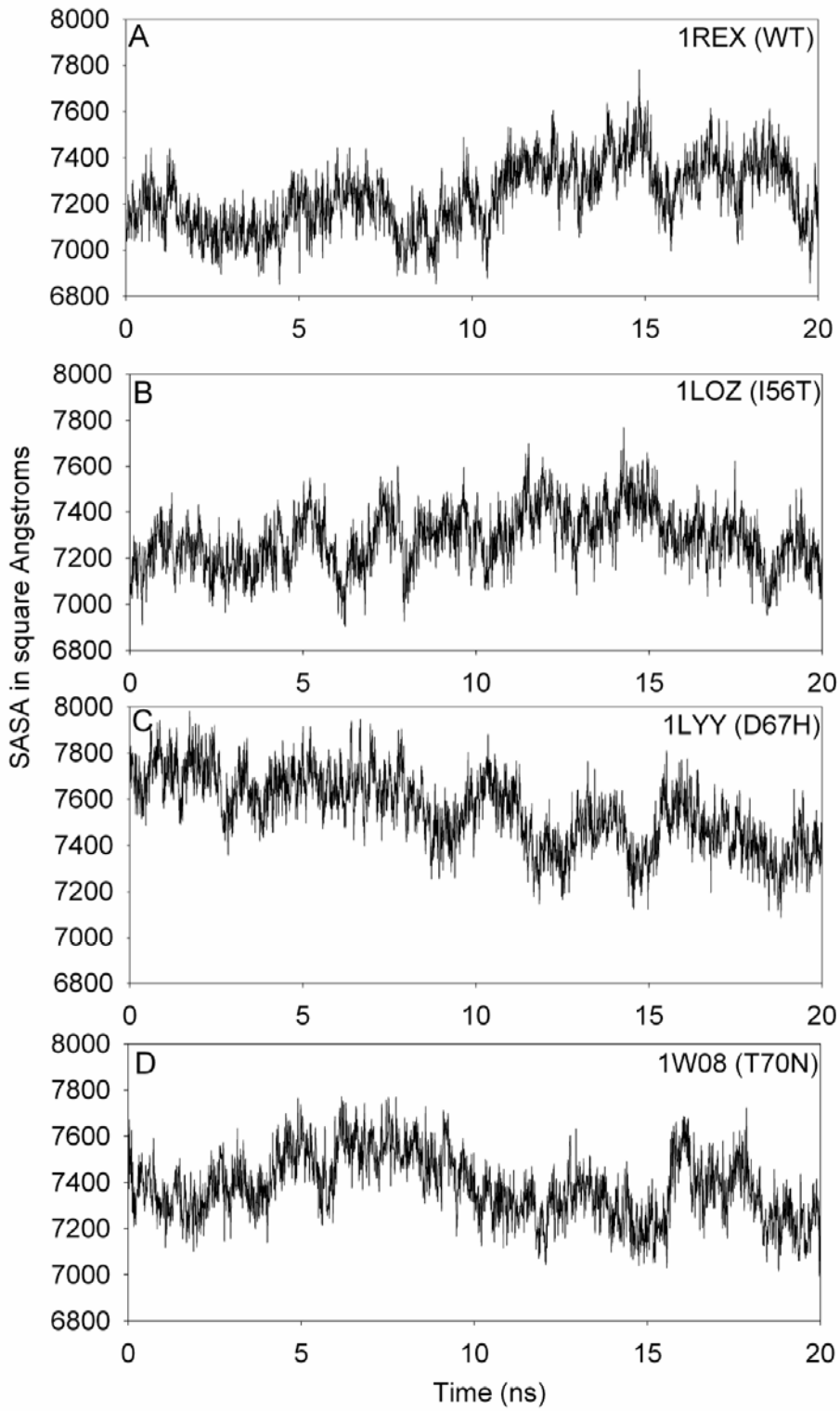

Figure 5.4:`*The solvent accessible surface area (SASA) of entire protein during the time course of simulation is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

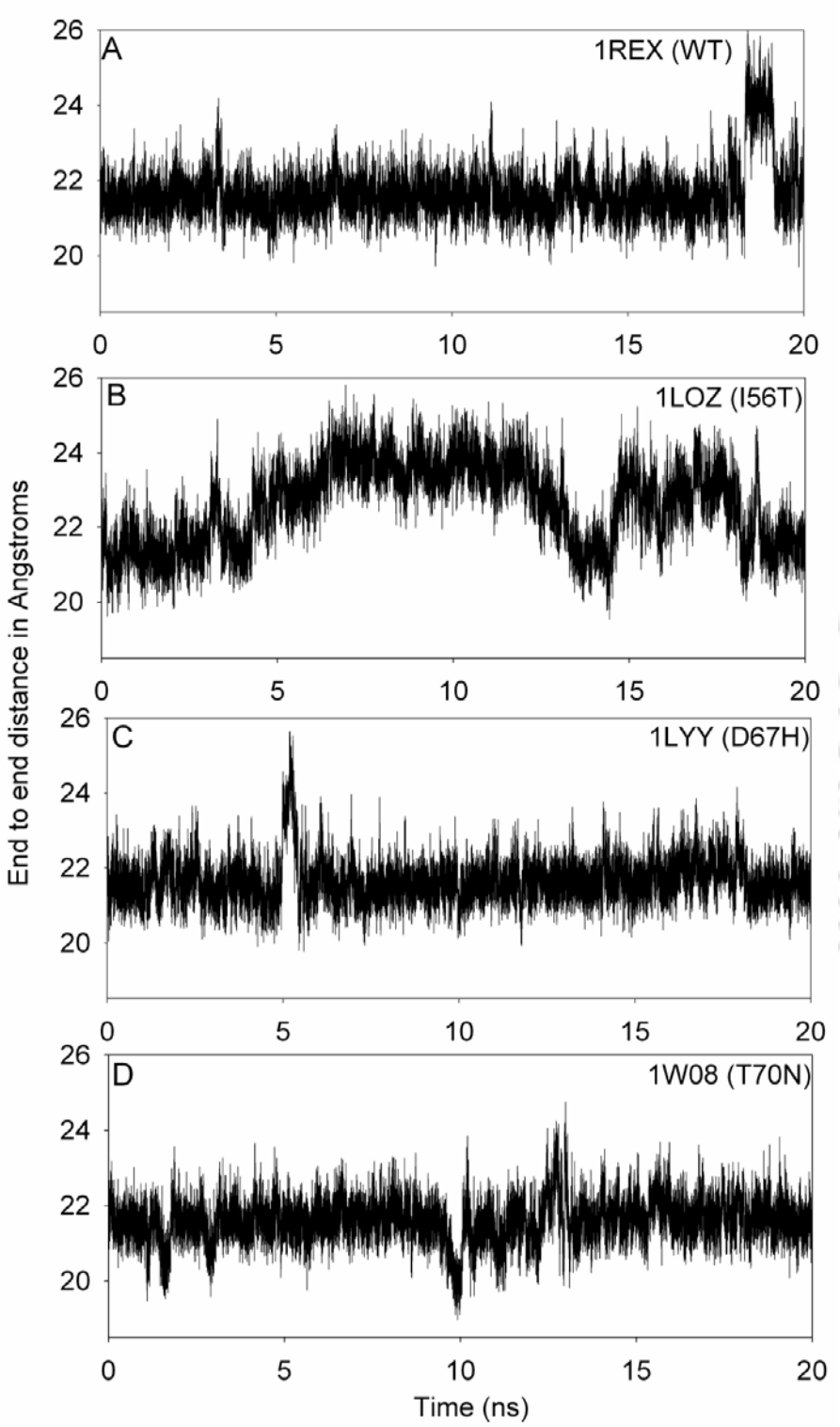

Figure 5.5:`*End to end chain distance of α-carbon atoms as a function of simulation time period is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

5.3.6. Analysis of Secondary structure:

The secondary structure analysis was carried out using the Kabsch and Sander algorithm incorporated in their DSSP (Dictionary of secondary structure for proteins) program (**Kabsch and Sander, 1983**). The results are plotted in **Figure 5.6 A, B, C, and D.** The plots show the structural variation of each residue during the time course of simulation. In the case of wild type, 1REX (**Figure 5.6 A**), most of the regions in the protein remains unchanged in the secondary structure during time course of simulation, although some transition in certain regions (Residue 38-42, 70-72, 80-85) appear that oscillates between turn and irregular structure, and turn and 3_{10} helix. In the case of mutants, the region near the point of mutation shows high degree of change in the secondary structure. In I56T (**Figure 5.6 B**), we observe many residues showing frequent change in secondary structure during the course of simulation. The regions and the different secondary structural transitions observed are: Residue No. 30-35 (alpha helices/turns); Residue No. 38-40 (strands/irregular structure); Residue No. 40-43 (turns/irregular structure); Residue No. 46-50 (turns/3_{10}helices); Residue No.56-58 (turns/3_{10}helices); Residue No. 80-85 (3_{10} helices/alpha helices/turns). In D67H (**Figure 5.6 C**), the region and the different secondary structural transitions observed are: Residue No. 60-70 (turns/alpha helices/irregular structure); Residue No. 80-85 (3_{10} helices/alpha helices/turns). In the case of T70N (**Figure 5.6 D**), we observe majority of the residues showing relatively lesser degree of change in secondary structure during the simulation period. The trend observed is similar to that of wild type. The above analysis reveal that in wild type and T70N, the secondary structure profile remains more or less same during the time period of simulation while in mutants (I56T and D67H), there is significant change in the secondary structure profile. In mutants, the beta domain region is observed to be more flexible and thus most of the residues in this region undergo frequent change in secondary structure and exist predominantly in irregular structure; and regular secondary structures are eventually replaced by irregular structure. So the existence and rapid change in structural dynamics of mutants is clearly visible from the secondary structural analysis. We also calculated the percentage of individual secondary structure content in wild type and mutants across all conformations sampled during the trajectories and the results are shown in **Table 5.2**. From the **Table 5.2**, we observe that mutants

D67H, I56T, and T70N contain higher content of beta-turn than the wild type. The order of % strand follows WT < I56T < T70N < D67H.

Table 5.2: *Secondary structure content in lysozyme during the simulation*

PDB code	Variant	% strand*	% alpha helix*	% 3_{10} helix*	% others*
1REX	WT	**6.6**	20.7	13.2	59.5
1LOZ	I56T	**8.3**	20.7	12.4	58.7
1LYY	**D67H**	**22.5**	13.3	5.0	59.2
1W08	T70N	**14.0**	16.5	9.9	59.5

*The percentage is calculated from the secondary structure content across all conformations sampled during the trajectories.

5.3.7. Distance Matrix Analysis:

To investigate the proximity of C^{α}-atoms in the protein structure, we analyze the average minimum distance matrix. The distance between the C^{α}-atoms for all pairs of residues in the wild type and mutants are depicted in the **Figure 5.7 A, B, C, and D**. The points near the diagonal represent distances between adjacent residues along the protein backbone. We can infer the details about the regions containing standard secondary structure and irregular structure in the protein. Helical elements are indicated by thick bulges (cylinders) along the diagonal. While the lines that are parallel and perpendicular to the diagonal represents other secondary structures (β-strands) in the protein. In the case of wild type (**Figure 5.7 A**) and mutants (**Figure 5.7 B, C, and D**), we can clearly see the structural difference that exists in the beta domain region (43-80). Further it can be seen that in the mutants, there is abnormal variation in the distance between C^{α}-atoms in comparison with wild type.

5.3.8. S^2 order parameter:

The generalized order parameter S^2 calculated from MD simulation trajectory using time dependent correlation motion function generally agrees well with NMR S^2 values (**Chandrasekhar *et al.*, 1992**). This is often useful to validate the conformational dynamics of proteins. The S^2 order parameter is an indicator of protein backbone motions in computationally feasible time scales. We have computed the S^2 values for the wild type and mutant of human lysozyme from the corresponding MD simulation trajectory. The results are shown in **Figure 5.8 A, B, C, and D**.

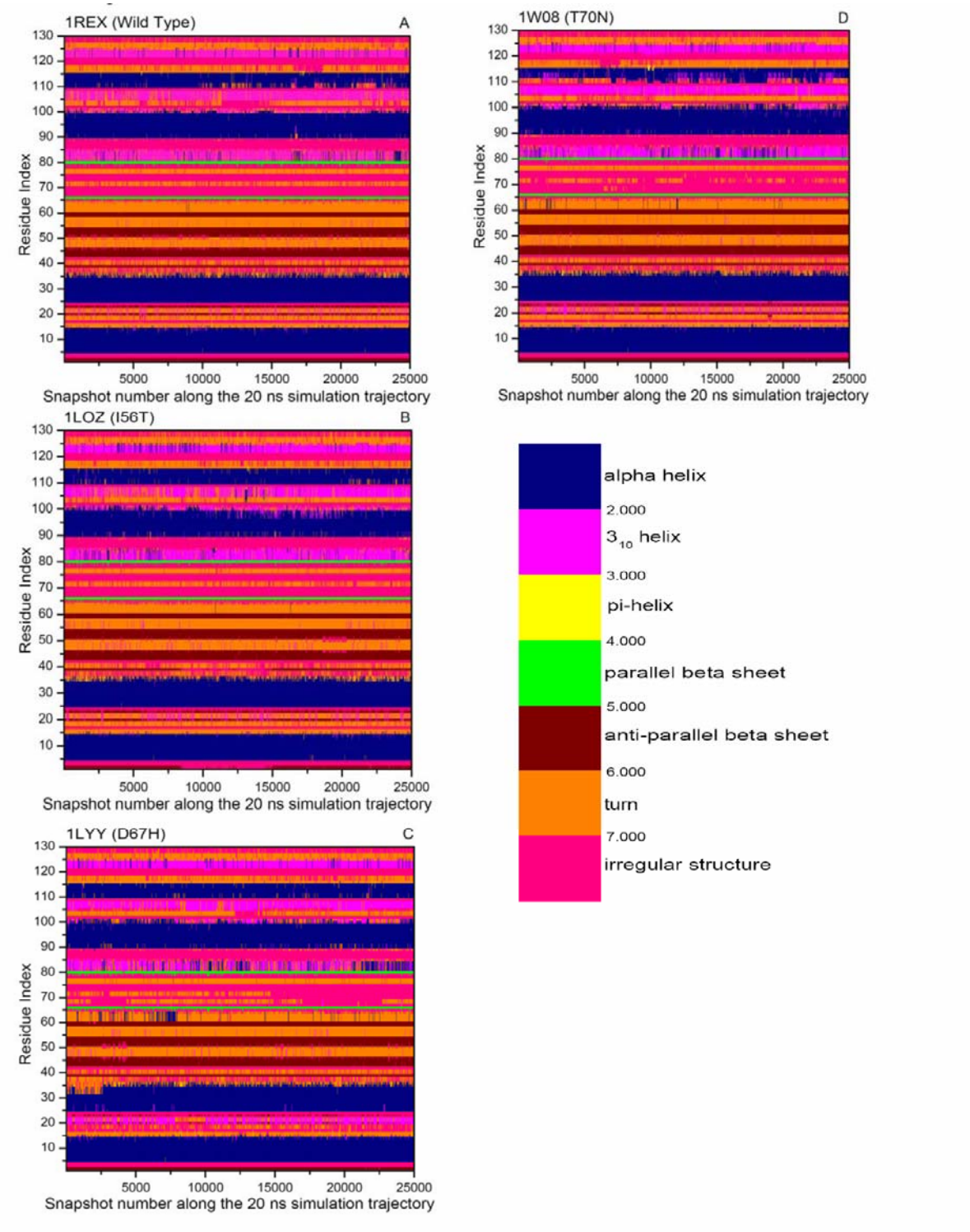

Figure 5.6:`*Detailed secondary structure data for each residue along the complete trajectory is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N). Colour code may be interpreted from the legend.*

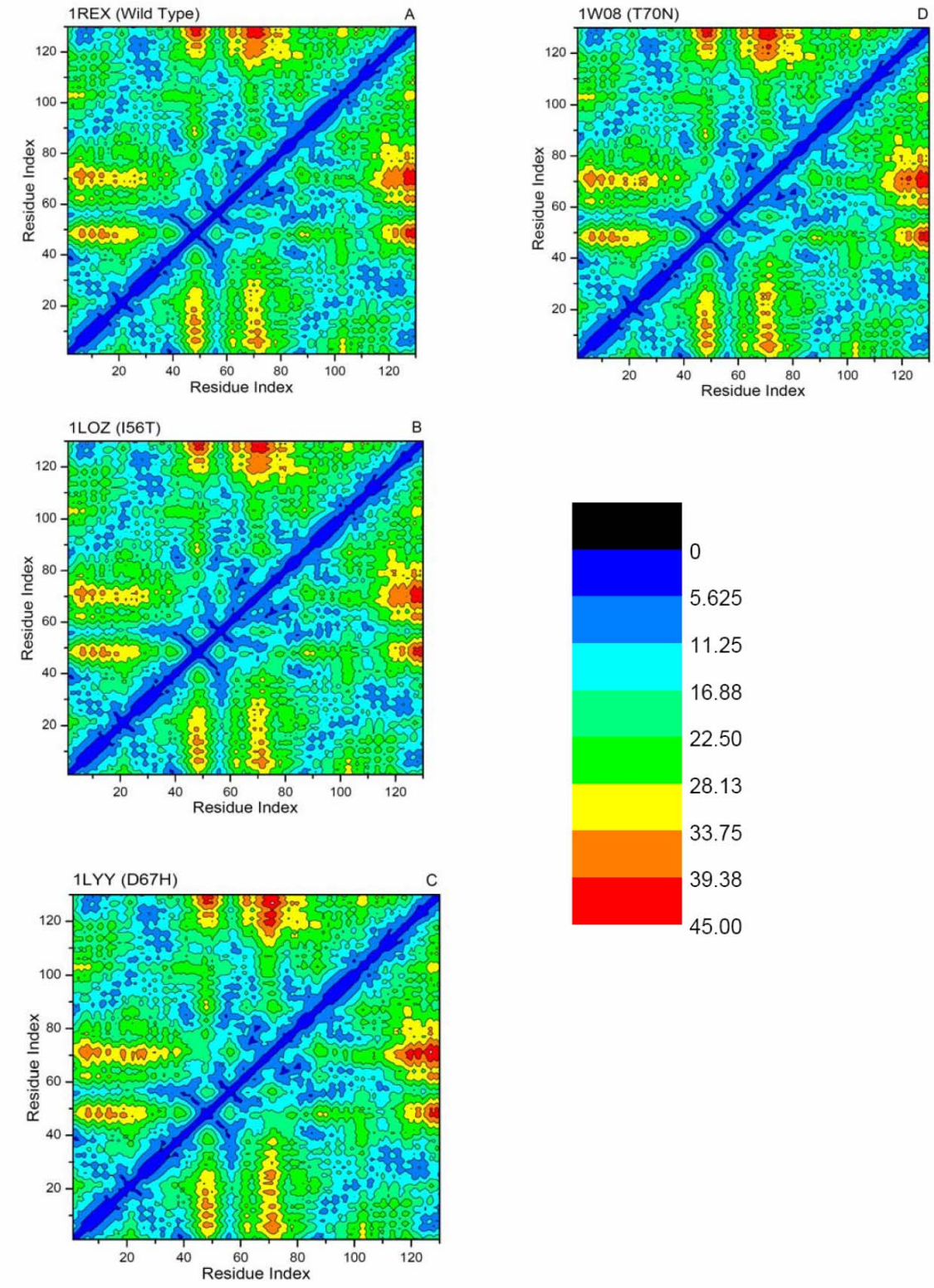

Figure 5.7:`*C^{α} atoms distance matrices are shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

The flexible regions or highly mobile regions in the protein can be inferred from the S^2 values (higher the S^2 value, lower will be flexibility and vice versa). It was observed that S^2 value of C-α atoms in the beta domain (43-80) region shows appreciable difference for the wild type (Figure 5.8 A) and mutants D67H, T70N (**Figure 5.8 C, and D**). This is mainly due to the presence of mutation sites in this beta domain region. In D67H (**Figure 5.8 C**) and T70N (**Figure 5.8 D**), the residues around the point of mutations are disturbed significantly resulting in the higher S^2 values. In addition S^2 values for all mutants reveal additional perturbations around the region 100-110 specifically for I56T. D67H also reveals a major perturbation in dynamics around region 15-25. The above results reveal that mutations can result in significant alterations in flexibility not only in the vicinity of mutations but also at regions far away.

5.3.9. Conformational Entropy:

To investigate the disorder associated with the structural arrangement of the protein, we analyzed the conformational entropy from the MD simulation trajectory. The results are shown in the **Figure 5.9**. The normalized conformational entropy per residue was observed to be slightly higher for the mutants. The calculated conformational entropy follows the order: Wild type < I56T < T70N < D67H. Thus the extent of conformational disorder present in the wild type and mutants can be seen from the magnitude of conformational entropy. The results obtained are in agreement with the trend observed in RMSD analysis. The degree of instability in mutants arises perhaps from flexible beta domain regions.

5.3.10. Analysis of water movement:

To investigate the behavior of water molecules around the residues in the core domain region (Residue 33 to 108) of the wild type and mutants of human lysozyme, we have analyzed the water movement. The number of water molecules within 8 Å around the residues in the core domain region along the time course of simulation for wild type and mutants are depicted in the **Figure 5.10 A, B, C, and D**. From the plots, it is visible that in comparison with wild type (**Figure 5.10 A**), the water movement in the mutants (**Figure 5.10 B, C, and D**), is fluctuating wildly around the residues in beta domain (especially between Residue No.67-80) and near the alpha/beta domain interface (especially between Residue No. 40-50).

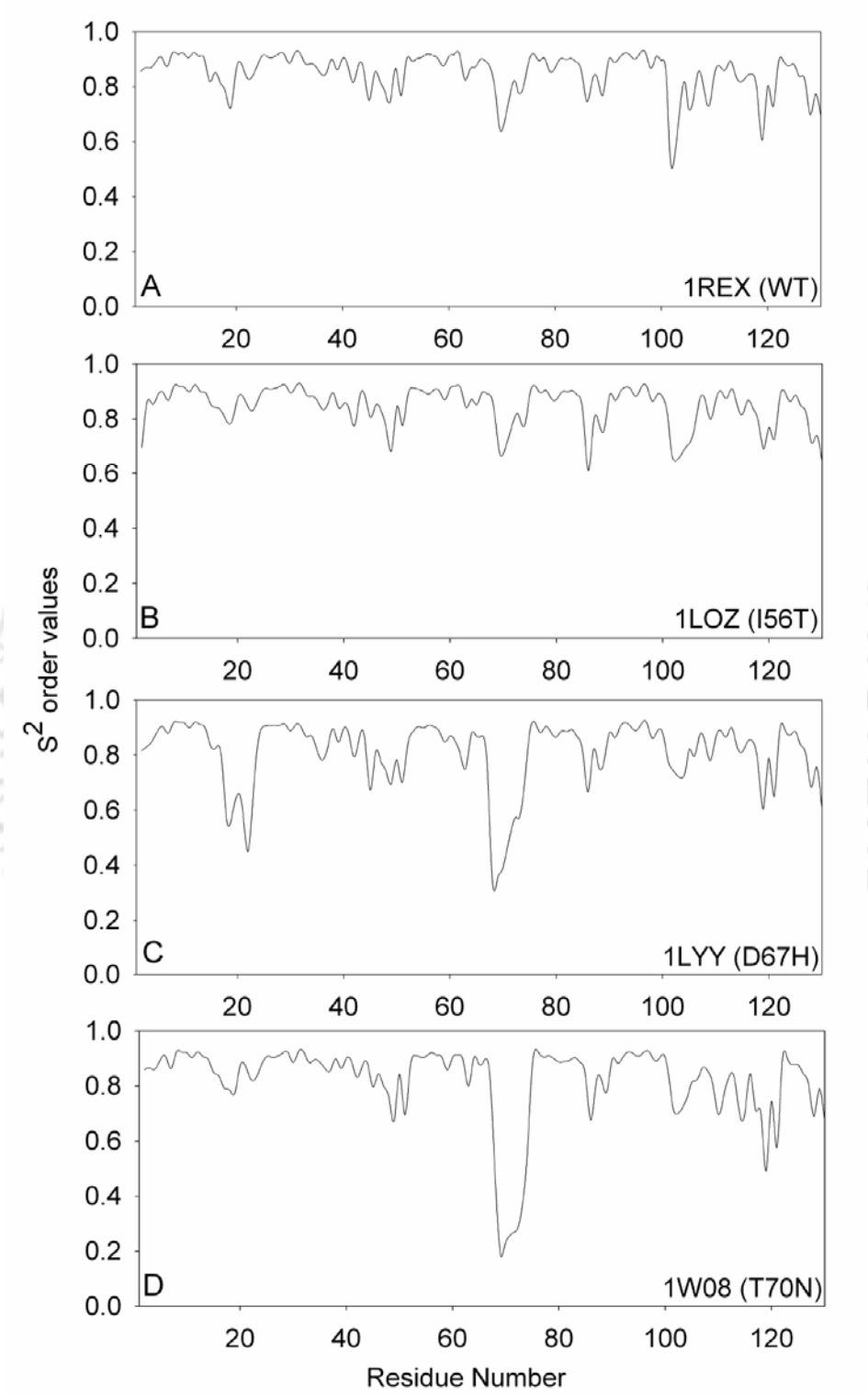

Figure 5.8:`*Calculated Generalized order parameter as a function of residue number is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N).*

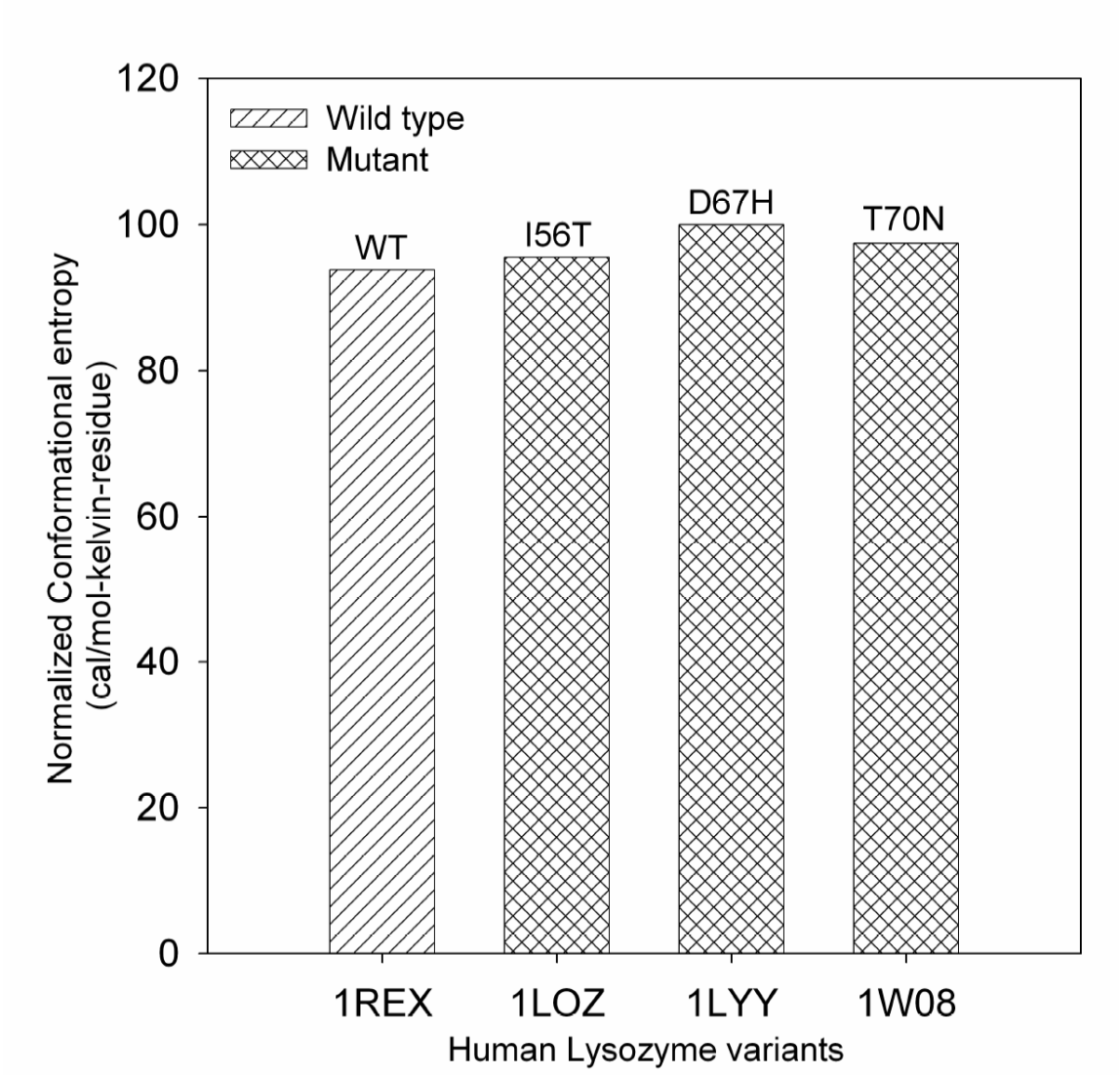

Figure 5.9:`*Normalized conformational entropy is shown for wild type (1REX) and mutants 1LOZ (I56T), 1LYY (D67H), (D) 1W08 (T70N).*

This indicates that beta domain and the alpha/beta domain interface regions in mutants are more flexible and often in contact with more number of water molecules. The degree of flexibility present in wild type and mutants can be known from the number of water molecules and its fluctuations around the residues in this region. The flexibility order was observed to be: D67H > T70N > I56T > Wild type. This observation reveals that the local structure in the beta domain and near the alpha/beta domain interface is somewhat different for wild type and mutants.

5.3.11. Hydrophobic contact Analysis:

Liu *et al.* (2006a) found 16 pairs of hydrophobic contact in the crystal structure of human lysozyme as follows: L31-F57, A32-I56, A32-F57, G37-I56, Y38-I56, Y38-F57, A42-G55, A42-I56, L84-G55, L84-I56, L85-G55, I89-I56, A92-G55, A92-I56, A92-F57, and A96-F57. The compactness of residues in this hydrophobic core accounts for the stability of the structure. The hydrophobic core was considered to be broken when the distance between the C-alpha atoms was greater than 8 Å for more time (~40 ps) during the simulation. Thus the degree of protein unfolding can be known from the lack of retention of these hydrophobic contacts. The results of the hydrophobic contact analysis of wild type and mutants during the MD simulation time course are given in the **Figure 5.11 A, B, C, and D**. Our analysis shows that hydrophobic core was much affected in the mutants I56T and D67H. In the wild type (**Figure 5.11 A**), it is observed that Y38-F57 hydrophobic contact was weakened quickly during the time course of simulation. In I56T (**Figure 5.11 B**), it is noticed that hydrophobic contacts of the following pairs: Y38-F57, I89-T56, A92-T56, A92-F57 are relatively weakened during the simulation period. In D67H (Figure 5.11 C), the following hydrophobic contacts Y38-F57, I84-I56, I89-I56, A92-I56, A92-G55, and A92-F57 are badly affected in the time course of simulation. While in T70N (**Figure 5.11 D**), the hydrophobic contact in the pairs Y38-F57 and I89-I56 are relatively weakened during the simulation. So as a whole we observe greater numbers of hydrophobic contact pairs are weakened in mutants I56T and D67H resulting in disruption of the hydrophobic core. Thus the structural integrity is relatively maintained in wild type and mutant T70N than the other mutants I56T and D67H. Moreover we see that hydrophobic contact pair Y38-F57 is very badly affected and disappeared quickly in wild type as well as in mutants. This observation predicts that Y38 may be the initiation site for the destruction of the hydrophobic core of human lysozyme leading to the formation of amyloid fibrils.

5.4. Discussion:

Our work using MD simulations on wild type and mutant structures of human lysozyme show that a) Mutants D67H and T70N show marginally more deviations in RMSD values from wild type in comparison to mutant I56T.

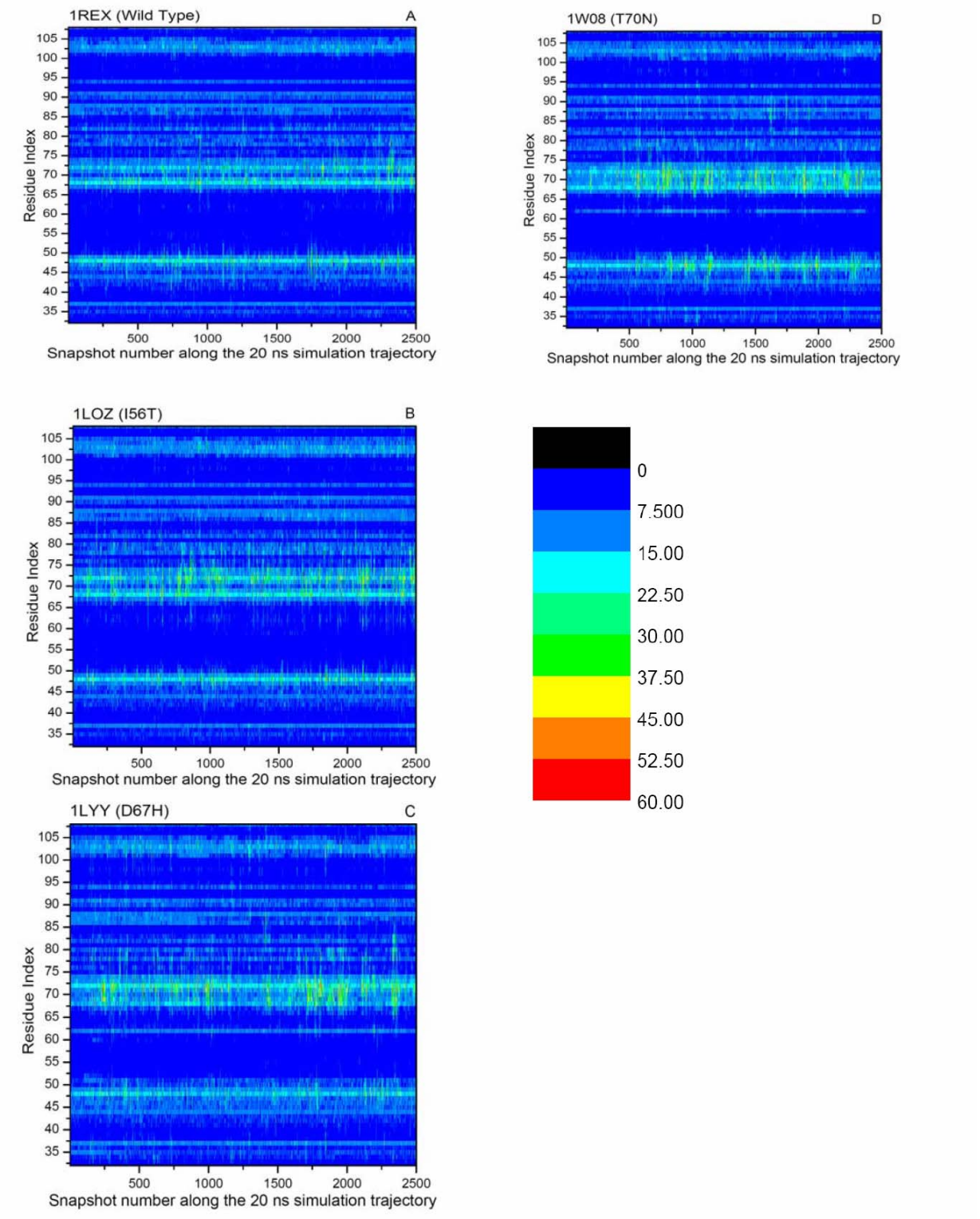

Figure 5.10:`*Water movement along the residues in the core domain is shown during the entire simulation period. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N). Color code may be interpreted from the legend.*

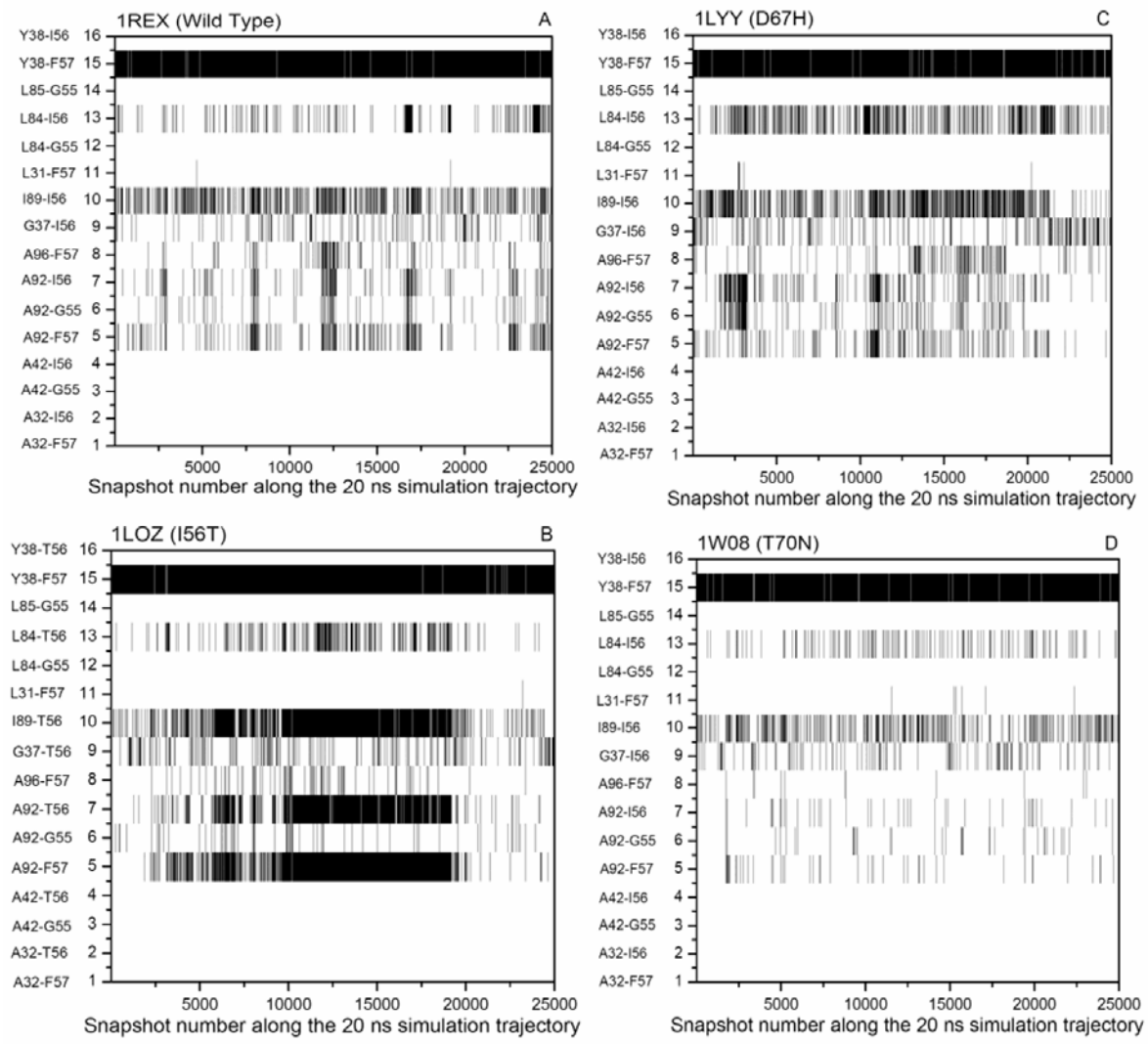

Figure 5.11:`*Hydrophobic contact analysis for core domain during the simulation period is shown. (A) 1REX (Wild type). (B) 1LOZ (I56T). (C) 1LYY (D67H). (D) 1W08 (T70N). The hydrophobic core was considered to be broken when the distance between the C-alpha atoms was greater than 8 Å. The black shade represents the distance between C-alpha atoms when it is greater than 8 Å.*

b) all three mutants possess R_g a shade higher than wild type. c) B-factor values reveal significant changes at site of mutation for D67H and T70N, while regions around 100-110 reveal reduced B-factor values in comparison to wild type. d) D67H and T70N mutants display higher SASA in comparison to I56T mutant. However, all mutants possess higher SASA in comparison to wild type. e) Mutant I56T shows larger fluctuations in end to end distance in comparison to wild type and other mutants. f) All mutants display significantly higher conformational entropy in comparison to wild type.

g) Analysis of secondary structures reveal, significant and large increase in β-sheet content among the mutants compared to wild type. h) S^2 order parameter was significantly low at site of mutation in D67H and T70N while all mutants showed higher S^2 values around 100-110 region. i) Increased water movement in the β-sheet regions, α/β domain interface regions of mutants in comparison to wild type and j) Hydrophobic contacts are significantly weakened in I56T and D67H causing disruption of hydrophobic core in these mutants. Overall from the work presented here one may conclude that all the mutants of human lysozyme display increased flexibility and dynamics, higher solvent exposure and weakening of crucial hydrophobic contacts during the simulation period. It is likely that these transient fluctuations in structure have a major role in the ***in vivo*** oligomerization of these mutant proteins under physiological conditions.

Structural analysis of human lysozyme has been performed in the past. **Moraitakis and Goodfellow (2003)** have employed an elevated temperature of 500 K to induce unfolding in human lysozyme and its mutant D67H in their MD simulations. Their studies reveal that while the mutant unfolds faster, both mutant and wild type lose their native secondary structure within first 2 ns. Like us they too observe higher atomic fluctuations and solvent accessibility in the β-domain for D67H in comparison to wild type although their simulations were run at 500 K. **Liu *et al.*, (2006a),** show the destabilization of human lysozyme in presence of ethanol at high temperatures. Their results reveal exposure of interior hydrophobic core at higher temperatures quite similar to our observations with I56T and D67H which appear transiently under ambient conditions. In a later work they employ the three mutants used in this study to demonstrate by MD simulations the disruption of hydrophobic core between α and β domains specifically for I56T and D67H at high temperature in presence of ethanol (**Liu *et al.*, 2006b**). Again these observations are consistent with our results on weakened hydrophobic contact in D67H and I56T. Recent (**Castillo and Ventura, 2009**) prediction of aggregation prone regions in human lysozyme has identified regions corresponding to residues 25-33, 57-66, 76-84 and 108-114. Interestingly we observe low S^2 order parameter values in D67H around 17-23 and high S^2 order parameter values around 100-110 in all mutants. Regions 108-114 also show a lot of water movement in our trajectories.

Chiti and Dobson (2009) have postulated that thermal fluctuations in native conditions can give rise to precursor states, that are prone to amyloid formation. To the best of our knowledge the studies here provide the first evidence confirming this for human lysozyme. It is thus not surprising that stabilizing the mutant lysozyme by Camelid VHH HL6 (**Dumoulin *et al.*, 2003**) antibody inhibits formation of aggregates by these mutants. It is likely that the antibody effectively reduces or even abolishes the transient formation of amyloid precursors. The studies presented here may therefore help in better strategies to prevent systemic amyloidoses associated with lysozyme.

5.5. Conclusions:

In this work, we have used the MD simulations to analyze and compare the conformational dynamics originating from wild type and mutants of human lysozyme. The analyses of backbone RMSD, B factor values, SASA, end to end chain distance, secondary structure content, distance matrix, S^2 order values, conformational entropy, water movement around residues in core domain and hydrophobic contacts all show the appreciable structural destabilization in mutants compared to wild type. The higher content of beta sheet secondary structure, increased flexibility and disruption in hydrophobic contacts near the alpha/beta domain interface and in beta domain in the case of mutants I56T, D67H perhaps leads to amyloidogenicity. Our results also to some extent confirms and indicates the residue involved in hydrophobic core Y38 (near the alpha/beta domain interface) to be the seed for fibril formation in mutants.

CHAPTER 6
Concluding Remarks

Concluding Remarks:

6.1. Summary:

The main theme of the current thesis is some investigations on protein structure, function and dynamics in disordered states and non-ideal conditions. In this thesis work we looked into the methods to identify and investigate functionally active loops, unstructured region in protein structures. We also identified the features unique to disordered regions and disordered proteins using Molecular dynamics simulations. In addition to this we have also looked into the effect of crowding on enzymatic reaction and insights about the amyloidosis of Human lysozyme from the conformational dynamics using MD simulations under native conditions.

In the first part of the thesis, we have demonstrated a method to identify and investigate functionally active loops and unstructured regions in protein structures using the MSRP parameter. Both the magnitude and the associated standard deviation of this parameter show unique characteristics for loop regions. This method is found to be a useful tool for identifying loop regions that are structurally perturbed after the protein is bound to a ligand. The MSRP parameter is shown to fluctuate more in unstructured regions like loops in comparison to regular regions like α-helices, during molecular dynamics simulations. It is envisaged that this method will a) enable a better categorization of loops and folds among proteins b) permit automated identification of functional loops in protein structures and c) provide clues on the diversity of conformations sampled by a disordered region during a molecular dynamics simulation.

In the second part of my thesis, we concentrated on disordered regions and intrinsically disordered proteins. Regardless of their abundance and importance, the challenges associated with the characterization of IDPs are not addressed to a greater extent. It is very difficult to characterize the disordered proteins because of heterogeneity and rapid inter-conversion of conformers leading to practical challenges. So we have attempted to get the characteristic features of intrinsically disordered Proteins using Molecular dynamics simulation as this method will yield a large ensemble of diverse structures and provide available tools to analyze structure and related dynamics. We have used the MD simulations to analyze and compare the conformational dynamics

originating from two ordered, one partially ordered and four disordered proteins. The observed conformational dynamics features of disordered proteins are unique and can be used as characteristic features to distinguish them from well ordered proteins. Particularly the structure instability feature in disordered protein due to disordered region is seen from high RMSD values, large fluctuations in radius of gyration, end to end C^{α} atom distance, SASA and RMSD, and higher conformational entropy. The other characteristic feature of disordered proteins is rapid change in conformational dynamics. This is incidental from the secondary structure analysis. Apart from this, the feature of high mobility or flexibility in disordered regions of disordered proteins is evident from lower generalized S^2 order parameter, and SASA analysis. In addition, the structural irregularity in disordered regions is reflected from distance matrix analysis and secondary structure analysis. Thus our results clearly suggest the various approaches employed to analyze MD simulation trajectories clearly bring out the dynamic nature of disordered regions.

In the other work, we investigated how the kinetics of an enzymatic reaction is dependent on size and concentration of crowding species. We have proposed a newer method to ensure efficient mixing of large molecular weight dextrans with enzyme. Our findings depicts that size and concentration of macromolecule play a crucial role in influencing the rate of an enzymatic reaction. The effect of crowding by smaller dextrans (40 kDa) showed minor decrease in rate in comparison with larger dextrans (500 and 2000 kDa) in the case of AP vs PNPP. We observed the effect of crowding by dextrans to have opposite effects on two different substrates IA and NA with acetyl cholinesterase. The effect of crowding by 200 kDa dextran size clearly stimulates the reaction at low concentrations with NA, while it inhibits it appreciably with IA unlike other dextrans of larger size. This finding shows that the increase in activity of $ES^{\neq}$ complex is selective on substrate. Thus our results reveal that the effect of crowding on enzymatic reactions is not simple as it appears and depends on crowder size and substrate nature.

In last part of my thesis, we have used the MD simulations to analyze and compare the conformational dynamics originating from wild type and mutants of human lysozyme. The analyses of backbone RMSD, B factor values, SASA, end to end chain distance, secondary structure content, distance matrix, S^2 order values, conformational entropy, water movement around residues in core domain, and hydrophobic contacts all

show the appreciable structural destabilization in mutants compared to wild type. The higher content of beta sheet secondary structure, increased flexibility and disruption in hydrophobic contacts near the alpha/beta domain interface and in beta domain in the case of mutants I56T, D67H perhaps leads to amyloidogenicity. Our results also to some extent confirms and indicates the residue involved in hydrophobic core Y38 (near the alpha/beta domain interface) to be the seed for fibril formation in mutants.

6.2. Scope of future works:

This thesis work gives some insights on the protein structure, function and dynamics in different conditions. The work also carries scope for future implications that includes

(a) Loops categorization with respect to the folds involved in the protein can be automated. MSRP method can be developed as a tool similar to any other secondary structure predictor to identify all the secondary structures in the protein.

(b) Functioning of IDPs can be studied using MD simulation

(c) Crowding effect can be studied on the functioning of IDPs

(d) The methods to prevent the Human lysozyme amyloidosis can be studied using MD simulations.

Bibliography

Bibliography:

Adler, A. J., Greenfield, N. J., and Fasman, G. D. (1973). Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol.* 27, 675-735

Alder, B. J., and Wainwright, T. E. (1957). Phase Transition for a hard sphere system. *J. Chem. Phys.* 27, 1208-1209.

Alder, B. J., and Wainwright, T. E. (1959). Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* 31, 459-466.

Alexandrescu, A. T., Gittis, A. G., Abeygunawardana, C., and Shortle, D. (1995). NMR structure of a stable "OB-fold" sub domain isolated from staphylococcal nuclease. *J. Mol. Biol.* 250, 134-143.

Al-Habori, M. (2001). Macromolecular crowding and its role as intracellular signaling of cell volume regulation. *Int. J. Biochem. Cell Biol.* 33, 844-864.

Anderson, C. M., Zucker, F. H., and Steitz, T. A. (1979). Space-filling models of kinase clefts and conformation changes. *Science* 204, 375-380.

Andricioaei, I., and Karplus, M. (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* 115, 6289-6292.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181, 223-230.

Anson, M. L., and Mirsky, A. E. (1925). On some general properties of proteins. *J. Gen. Physiol.* 9, 169-179.

Armstrong, R. N. (1998). Mechanistic imperatives for the evolution of glutathione transferases. *Curr. Opin. Chem. Biol.* 2, 618-623.

Benson, E. L., Huynh, P. D., Finkelstein, A., and Collier, R. J. (1998). Identification of residues lining the anthrax protective antigen channel. *Biochemistry* 37, 3941–3948.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684-3690.

Bernstein, L. S., Ramineni, S., Hague, C., Cladman, W., Chidiac, P. *et al.* (2004). RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling. *J. Biol. Chem.* 279, 21248-21256.

Berry, H. (2002). Monte Carlo simulations of enzyme reactions in two dimensions: fractal kinetics and spatial segregation. *Biophys. J.* 83, 1891-1901.

Berzofsky, J. A. (1985). Intrinsic and extrinsic factors in protein antigenic structure. *Science* 229, 932-940.

Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C., Phillips, D. C., and Sarma, V. R. (1965). Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 206, 757-761.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365-370.

Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V., Hutchinson, W. L. *et al.* (1997). Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* 385, 787-793.

Borchers, H., Ed., (1955). Landoldt-Bornstein Numerical Data and Functional Relationships in Physics, Chemistry, Astronomy, Geophysics, and Technology, 6th ed., Vol. 4, Part 1, Material values and mechanical behavior of non-metals.

Boyde, T. R. C. (1980). Foundation Stones of Biochemistry. *Voile et Aviron*, Hong Kong.

Bracken, C., Iakoucheva, L. M., Romero, P. R., and Dunker, A. K. (2004). Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.* 14, 570-576.

Briggs, G. E., and Haldane, J. B. (1925). A Note on the Kinetics of Enzyme Action. *Biochem. J.* 19, 338-339.

Brito, R. M. M., Damas, A. M., Saraiva, M. J. (2003). Amyloid formation by transthyretin: from protein stability to protein aggregation. *Curr. Med. Chem. Immunol. Endocr. Metab. Agents* 3, 349-360.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy minimization, and dynamics calculations. *J. Comp. Chem.* 4, 187-217.

Brooks, D. E. (2000). Can cytoplasm exist without undergoing phase separation? *Int. Rev. Cytol.* 192, 321-330.

Brown, A. J. (1902). "Enzyme action". *J. Chem. Soc. (Trans.)* 81, 373-388.

Burke, D. F., and Deane, C. M. (2001). Improved protein loop prediction from sequence alone. *Protein Engineering* 14, 473-478.

Carbajo, R. J., Silvester, J. A., Runswick, M. J., Walker, J. E., and Neuhaus, D. (2004). Solution structure of subunit F (6) from the peripheral stalk region of ATP synthase from bovine heart mitochondria. *J. Mol. Biol.* 342, 593-603.

Case, D. A., Darden, T. A., Cheatham, III, T. E., Simmerling, C. L., Wang, J. *et al.*(2002 and 2004). AMBER 7 and 8, *University of California*, San Francisco, USA.

Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R. *et al.* (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* 26, 1668-1688.

Castillo, V., and Ventura, S. (2009). Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comput. Biol.* 5, e1000476.

Cayley, S., Lewis, B. A., Guttman, H. J. and Record, M. T., Jr. (1991). Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo. *J. Mol. Biol.* 222, 281-300.

Chamberlain, A., MacPhee, C. E., Zurdo, J., Morozova-Roche, L. A., Hill, H. A. O., *et al.* (2000). Ultrastructural organization of amyloid fibrils by atomic force microscopy. *Biophys. J.* 79, 3282-3293.

Chan, H. S., and Dill, K. A. (1993). The protein folding problem. *Physics Today* (February), 24-32.

Chandrasekhar, I., Clore, G. M., Szabo, A., Gronenborn, A. M., Brooks, B. R. (1992). A 500 ps molecular dynamics simulation study of interleukin-1β in water. *J. Mol. Biol.* 226, 239-250.

Chen, H. F. (2009). Molecular dynamics simulation of phosphorylated KID post-translational modification. *PLoS One* 4(8), e6516.

Chiti, F., and Dobson, C. M. (2006). Protein msifolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* 75, 333-366.

Chiti, F., and Dobson, C. M. (2009). Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology* 5, 15-22.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., Dobson, C. M. (2003). Rationalisation of mutational effects on peptide and protein aggregation rates. *Nature* 424, 805-808.

Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M. *et al.* (1999). Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc. Natl. Acad. Sci.* USA 96, 3590-3594.

Chothia, C. and Leuzinger, W. (1975). Acetylcholinesterase: The structure of crystals of a globular form from electric eel. *J. Mol. Biol.* 97, 55-60.

Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 53, 537-572.

Chothia, C., Levitt, M., and Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proc. Natl. Acad. Sci.* USA 74, 4130-4134.

Chou, P. Y., Fasman, G. D. (1977). β-turns in proteins. *J. Mol. Biol.* 115, 135-175.

Clegg, J. S. (1984). Properties and metabolism of the aqueous cytoplasm and its boundaries. *Am. J. Physiol.* 246, R133-R151.

Cootes, A. P., Muggleton, S. H., and Sternberg, M. J. E. (2003). The Automatic discovery of structural principles describing protein fold space. *J. Mol. Biol.* 330, 839-850.

Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., and Dunker, A. K. (2005). Natively disordered proteins. In Handbook of protein folding, ed. J Buchner, T Kiefhaber, pp. 271-353. Weinheim, Germany: *Wiley-VCH, Verlag GmbH & Co. KGaA.*

De Leeuw, S. W., Perram, J. M., and Smith, E. R. (1986). Computer simulation of the static dielectric constant of systems with permanent electric dipoles. *Annu. Rev. Phys. Chem.* 37, 245-270.

Derham, B. K. and Harding, J. J. (2006). The effect of the presence of globular proteins and elongated polymers on enzyme activity. *Biochim. Biophys. Acta* 1764, 100-1006.

Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biochem. Sci.* 24, 329-332.

Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil Trans R Soc Lond B* 356, 133-145.

Dobson, C. M. (2004). Principles of protein folding, misfolding and aggregation. *Semin Cell Dev. Biol.* 15, 3-16.

Donate, L. E., Rufino, S. D., Canard, L. H. J., Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci.* 5, 2600-2616.

Dovidchenko, N. V., Bogatyreva, N. S., Galzitskaya, O. V. (2008). Prediction of loop regions in protein sequence. *J. Bioinform. Comput. Biol.* 6, 1035-1047.

DuBay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M., and Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* 341, 1317-1326.

Dumoulin, M., Last, A. M., Desmyter, A., Decanniere, K., Canet, D. *et al.* (2003). A camelid antibody fragment inhibits the formation of amyloid fibrils by human lysozyme. *Nature* 424, 783-788.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P. *et al.* (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26-59.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inf. Ser.* No. 11, 161-171.

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol.* 6, 197-208.

Dyson, H. J., and Wright, P. E. (2002). Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* 62, 311-340.

Dyson, H. J., and Wright, P. E., 2004. Unfolded proteins and protein folding studied by NMR. *Methods Enzymol.* 394, 299-321.

Edmond, E., Ogston, A. G. (1968). An approach to the study of phase separation in ternary aqueous systems. *Biochem. J.* 109, 569-576.

Efimov, A. V. (1979). Stereochemistry of α-helices and β-sheet packing in compact globule. *J. Mol. Biol.* 134, 23-40.

Efimov, A. V. (1993). Patterns of loop regions in proteins. *Curr. Opinion Struc. Biol.* 3, 379-384.

Eggers, D. K. and Valentine, J. S. (2001). Molecular confinement influences protein structure and enhances thermal protein stability. *Protein Sci.* 10, 250-261.

Eliezer, D. (2009). Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 19, 23-30.

Ellis, R. J. and Hartl, F. U. (1999). Principles of protein folding in cellular environment. *Curr. Opin. Struct. Biol.* 9, 102-110.

Ellis, R. J. (2001a). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* 11, 114-119.

Ellis, R. J. (2001b). Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* 26, 597-604.

Elowitz, M. B., Surette, M. G., Wolf, P. E., Stock, J. B., Leibler, S. (1999). Protein mobility in the cytoplasm of Escherichia coli. *J. Bacteriol.* 181, 197-203.

Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J. E., and Oliva, B. (2004). ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.* 32, D185-D188.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. *et al.* (1995). A smooth Particle Mesh Ewald method. *J. Chem. Phys.* 103, 8577-8593.

Ewald, P. (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* 64, 253-287.

Fandrich, M., and Dobson, C. M. (2002). The behavior of polyamino acids reveals an inverse side-chain effect in amyloid structure formation. *EMBO. J.* 21, 5682-5690.

Fandrich, M., Fletcher, M. A., and Dobson, C. M. (2001). Amyloid fibrils from muscle myoglobin. *Nature* 410, 165-166.

Fasman, G. D. (1996). Circular Dichroism and the conformational analysis of biomolecules. *Plenum Press*, Newyork.

Feng, W., Shi, Y., Li, M., and Zhang, M. (2003). Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. *Nat. Struct. Biol.*10, 972–978.

Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Ber. Dt. Chem. Ges.* 27, 2985-2993.

Flaugh, S. L., and Lumb K. J. (2001). Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27Kip1. Biomacromolecules 2, 538-540.

Fontana, A., de Laureto, P. P., de Filippis, V., Scaramella, E., and Zambonin, M. (1997). Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.* 2, R17-R26.

Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. (2004). Probing protein structure by limited proteolysis. *Acta Biochim. Pol.* 51, 299-321.

Forloni, G., Angeretti, N., Chiesa, R., Monzani, E., Salmona, M., Bugiani, O., Tagliavini, F. (1993). Neurotoxicity of a prion protein fragment. *Nature* 362, 543-546.

Fritz-Wolf, K., Schnyder, T., Wallimann, T., and Kabsch, W. (1996). Structure of mitochondrial creatine kinase. *Nature* 381, 341–345.

Fuller, R. S., Kaguni, J. M., Kornberg, A. (1981). Enzymatic replication of the origin of the Escherichia coli chromosome. *Proc. Natl. Acad. Sci.* USA 78, 7370-7374.

Fulton, A. B. (1982). How crowded is the cytoplasm? *Cell* 30, 345-347.

Funahashi, J., Takano, K., Ogasahara, K., Yamagata, Y., Yutani, K. (1996). The structure, stability, and folding process of amyloidogenic mutant human lysozyme. *J. Biochem.* 120, 1216-1223.

Gao, M., and Skolnick, J. (2009). From Nonspecific DNA-Protein encounter complexes to the prediction of DNA-Protein interactions. *PLoS Comput. Biol.* 2009; 5(3):e1000341.

Gershon, N. D., Porter, K. R. and Trus, B. L. (1985). The cytoplasmic matrix: its volume and surface area and the diffusion of molecules through it. *Proc. Natl. Acad. Sci.* USA 82, 5030-5034.

Gerstein, M., and Krebs, W. (1998). A database of macromolecular motions. *Nucleic Acids Res.* 26, 4280-4290.

Giddings, J. C. (1970). Effect of membranes and other porous networks on the equilibrium and the rate constants of macromolecular reactions. *J. Phys. Chem.* 74, 1368-1370.

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physica A* 188, 404-425.

Goedert, M. (2001). Alpha-synuclein and neurodegenerative diseases. *Nat. Rev. Neurosci.* 2, 492-501.

Grimmler, M., Wang, Y., Mund, T., Cilensek, Z., Keidel, E. M. *et al.* (2007). Cdk-inhibitory activity and stability of p27Kip1 are directly regulated by oncogenic tyrosine kinases. *Cell* 128, 269-280.

Guan, Y., Manuel, R. C., Arvai, A. S., Parikh, S. S., Mol, C. D. (1998). MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nat. Struct. Biol.* 5, 1058-1064.

Guharoy, M., and Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics* 23, 1909-1918.

Gunasekaran, K., and Nussinov, R. (2007). How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.* 365, 257-273.

Gunasekaran, K., and Nussinov, R. (2004). Modulating functional loop movements: the role of highly conserved residues in the correlated loop motions. *Chembiochem* 5, 224–230.

Hall, D., and Minton, A. P. (2003). Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *BBA-Proteins Proteomics* 1649, 127-139.

Halling, P. J. (1989). Do the laws of chemistry apply to living cells? *Trends Biochem Sci.* 14, 317-318.

Heinrich, R., and Schuster, S. (1996). The regulation of Cellular Systems. *Chapman & Hall*, New York.

Henri, V. (1902). *Comptes rendues Acad. Sci.* 135, 916-919.

Hockney, R. W. (1970). The potential calculation and some applications. *Methods in Computational Physics* 9, 136-211.

Hockney, R., and Eastwood, J. (1981). Computer simulations using particles, *McGraw-Hill*, New York.

Homchaudhuri, L., Sarma, N. and Swaminathan, R. (2006). Effect of crowding by dextrans and Ficolls on the rate of alkaline phosphatase-catalyzed hydrolysis: a size-dependent investigation. *Biopolymers* 83, 477-486.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65, 712-725.

Huang, X., and Miller, W. (1991). A time-coefficient, linear-space local similarity algorithm. *Adv. Appl. Math* 12, 337-357.

Hubbard, S. J., Eisenmenger, F., and Thornton, J. M. (1994). Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci.* 3, 757-768.

Hunding, A., and Kaern, M. (1998). The effect of slow allosteric transitions in a simple biochemical oscillator model. *J. Theor. Biol.* 191, 309-322.

Iacovache, I., Paumard, P., Scheib, H., Lesieur, C., Sakai, N. *et al.* (2006). A rivet model for channel formation by aerolysin-like pore-forming toxins. *EMBO J.* 25, 457–466.

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell signaling and cancer associated proteins. *J. Mol. Biol.* 323, 573-584.

Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Smith, R. D., Dunker, A. K., and Ackerman, E. J. (2001). Aberrant mobility phenomena of the DNA repair protein XPA. *Protein Sci.* 10, 1353-1362.

Jackson, R. M., and Rusell, R. B. (2000). The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *J. Mol. Biol.* 296, 325–334.

Jiang, M., and Guo, Z. H. (2007). Effects of macromolecular crowding on the intrinsic catalytic efficiency and structure of enterobactin-specific isochorismate synthase. *J. Am. Chem. Soc.* 129, 730-731.

Johansson, H. O., Brooks, D. E., and Haynes, C. A. (2000). Macromolecular crowding and its consequences. *Int. Rev. Cytol.* 192, 155-170.

Johnson, L. N., Lowe, E. D., Noble, M. E., Owen, D. J. (1998). The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS. Lett.* 430, 1–11.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637.

Kaltashov, I.A., and Mohimen, A. (2005). Estimates of protein surface areas in solution by electrospray ionization mass spectrometry. *Anal. Chem.* 77, 5370-5379.

Kanagasabai, V., Arunachalam, J., Prasad, P. A., and Gautham, N. (2007). Exploring the conformational space of protein loops using a mean field technique with MOLS sampling. *Proteins* 67, 908-921.

Kawasaki, H., and Kretsinger, R. H. (1995). Calcium-binding proteins 1: EF-hands. *Protein Profile* 2, 297–490.

Kempner, E. S. (1993). Movable lobes and flexible loops in proteins: Structural deformations that control biochemical activity. *FEBS Lett.* 326, 4-10.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C., (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181, 662-666.

Kendrew, J. C., Dickerson, R. E., and Strandberg, B. E., 1960. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 206, 757-763.

Kepler, T. B., Elston, T. C. (2001). Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* 81, 3116-3136.

Kim, S. T., Shirai, H., Nakajima, N., Higo, J., and Nakamura, H. (1999). Enhanced conformational diversity search of CDR-H3 in antibodies: role of the first CDR-H3 residue. *Proteins* 37**,** 683–696.

Kinjo, A. R., and Takada, S. (2003). Competition between protein folding and aggregation with molecular chaperones in crowded solutions: insight from mesoscopic simulations. *Biophys. J.* 85, 3521-3531.

Kittel, C. (1986). Introduction to solid state physics. *John Wiley & Sons*, New York.

Kopelman, R. (1986). Rate-processes on fractals-theory, simulations, and experiments. *J. Stat. Phys.* 42, 185-200.

Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. USA* 93, 11504-11509.

Kuthan, H. (2001). Self-organisation and orderly processes by individual protein complexes in the bacterial cell. *Prog. Biophys. Mol. Biol.* 75, 1-17.

Lai, Z., Colon, W., and Kelly, J. W. (1996). The acid-mediated denaturation pathway of transthyretin yields a conformational intermediate that can self-assemble into amyloid. *Biochemistry* 35, 6470-6482.

Lanni, F., Waggoner, A. S. and Taylor, D. L. (1985). Structural organization of interphase 3T3 fibroblasts studied by total internal reflection fluorescence microscopy. *J. Cell Biol.* 100, 1091-1102.

Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A. *et al.* (1989). X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes. *Nature* 342, 299-302.

Laurent, T. C. (1995). An early look at macromolecular crowding. *Biophys. Chem.* 57, 7-14

Laurent, T. C. (1963). The interaction between polysaccharides and other macromolecules. 5. The solubility of proteins in the presence of dextran. *Biochem. J.* 89, 253-257.

Laurent, T. C. (1971). Enzyme reactions in polymer media. *Eur. J. Biochem.* 21, 498-506.

Laurent, T. C., and Ogston, A. G. (1963). The interaction between polysaccharides and other macromolecules. 4. The osmotic pressure of mixtures of serum albumin and hyaluronic acid. *Biochem. J.* 89, 249-253.

Laurent, T. C., Preston, B. N., and Carlsson, B. (1974). Conformational transitions of polynucleotides in polymer media. *Eur. J. Biochem.* 43, 231-235.

Liu, H. L., Wu, Y. C., Zhao, J. H., Fang, H. W., and Ho, Y. (2006a). Structural analysis of human lysozyme using molecular dynamics simulations. *J. Biomol. Struct. Dyn.* 24, 229-238.

Liu, H. L., Wu, Y. C., Zhao, J. H., Liu, Y. F., Huang, C. H., Fang, H. W., and Ho, Y. (2006b). Insights into the conformational changes of several Human Lysozyme variants

associated with hereditary systemic amyloidosis. *Biotechnology Progress* DOI 10.1021/bp060264a.

Lopez de la Paz, M., and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. USA* 101, 87-92.

Lubey-Phelps, K. (2000). Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. *Int. Rev. Cytol.* 192, 189-221.

Luby-Phelps, K., Castle, P. E., Taylor, D. L., and Lanni, F. (1987). Hindered diffusion of inert tracer particles in the cytoplasm of mouse 3T3 cells. *Proc. Natl. Acad. Sci. USA* 84, 4910-4913.

Mahoney, M.W., and Jorgensen, W.L. (2000). A five site model for liquid water and the reproduction of the density anomaly by rigid, non-polarizable potential functions. *J. Chem. Phys.* 112, 8910-8922.

Makowska, J., Rodziewicz-Motowidlo, S., Baginska, K., Vila, J. A., Liwo, A. et al. (2006). Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. *Proc. Natl. Acad. Sci. USA* 103, 1744-1749.

Markus, G. (1965). Protein substrate conformation and proteolysis. *Proc. Natl. Acad. Sci. USA* 54, 253-258.

Martin, J. (2002). Requirement for GroEL/GroES-dependent protein folding under nonpermissive conditions of macromolecular crowding. *Biochemistry* 41, 5050-5055.

Matouschek, A. (2001). Protein unfolding – an important process in vivo? *Curr. Opin. Struct. Biol.* 13, 98-104.

Mc Cammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature* 267, 585-590.

Merlini, G., and Bellotti, V. (2003). Molecular mechanisms of amyloidosis. *N. Engl. J. Med.* 349, 583-596.

Michaelis, L., and Menten, M. L. (1913). Die kinetic der invertinwirkung. *Biochem. Z.* 49, 333-369.

Mikhalyi, E. (1978). Application of proteolytic enzymes to protein structure studies. *CRC Press*, Boca Raton, FL.

Minton, A. P. (1981). Excluded volume as determinant of macromolecular structure and reactivity. *Biopolymers* 20, 2093-2120.

Minton, A. P. (2001). The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J. Biol. Chem.* 276, 10577-10580

Minton, A. P. (1983). The effect of volume occupancy upon the thermodynamic activity of proteins: some biochemical consequences. *Mol. Cell. Biochem.* 55, 119-140.

Minton, A. P. (1990). Holobiochemistry: the effect of local environment upon the equilibria and rates of biochemical reactions. *Int. J. Biochem.* 22, 1063-1067.

Minton, A. P. (1993). Molecular crowding and molecular recognition. *J. Mol. Recognit.* 6, 211-214.

Minton, A. P. (1998). Molecular crowding: analysis of effects of high concentrations of inert cosolutes on biochemical equilibria and rates in terms of volume exclusion. *Methods Enzymol.* 295, 127-149.

Minton, A. P. (2000). Implications of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.* 10, 34-39.

Mittag, T., and Forman-Kay, J. D. (2007). Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* 17, 3-14.

Moraitakis, G., and Goodfellow, J. M. (2003). Simulations of Human Lysozyme: Probing the conformations triggering amyloidosis. *Biophy. J.* 84, 2149-2158.

Moran-Zorzano, M. T., Viale, A., Munoz, F., Alonso-Casajus, N., Eydallin, G. *et al.* (2007). *Escherichia coli* AspP activity is enhanced by macromolecular crowding and by both glucose-1, 6-bisphosphate and nucleotide–sugars. *FEBS Lett.* 581, 1035-1040.

Morozova-Roche, L. A., Zurdo, J., Spencer, A., Noppe, W., Receveur, V., Archer, D. B. *et al.* (2000). Amyloid fibril formation and seeding by wild-type human lysozyme and its disease-related mutational variants. *J. Struct. Biol.* 130, 339-351.

Muchmore, S. W., Sattler, M., Liang, H., Meadows, R. P., Harlan, J. E. *et al.* (1996). X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature* 381, 335-341.

Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.

Ogston, A. G. (1958). The spaces in a uniform random suspension of fibres. *Trans. Faraday Soc.* 54, 1754-1757.

Ogston, A. G. (1962). Some thermodynamic relationships in ternary systems with special reference to the properties of systems containing hyaluronic acid and protein. *Arch. Biochem. Biophys.* 1 (Suppl.), 39-51.

Ogston, A. G. (1970). An open-ended tale. *Search* 1, 60-67.

Ogston, A. G., and Phelps, C. F. (1960). The partition of solutes between buffer solutions and solutions containing hyaluronic acid. *Biochem. J.* 78, 827-833.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005a). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 44, 12454-12470.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005b). Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 44, 1989-2000.

Ovadi, J., Batke, J., Bartha, F., Keleti, T. (1979). Effect of association-dissociation on the catalytic properties of glyceraldehyde 3-phosphate dehydrogenase. *Arch. Biochem. Biophys.* 193, 28-33.

Pauling, L., Corey, R. B., and Branson, R. H. (1951). The structure of proteins: two hydrogen-bonded configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* 37, 205-210.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham III, T. E., *et al.* (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.* 91, 1-41.

Pepys, M. B., Hawkins, P. N., Booth, D. R., Vigushin, D. M., Tennent, G. A. et al. (1993). Human lysozyme gene mutations cause hereditary systemic amyloidosis. *Nature* 362, 553-557.

Peters, C. W., Kruse, U., Pollwein, R., Grzeschik, K. H., Sippel, A. E. (1989). The human lysozyme gene. Sequence organization and chromosomal localization. *Eur. J. Biochem.* 182, 507-516.

Peterson, F. C., Thorpe, J. A., Harder, A. G., Volkman, B. F., and Schwarze, S. R. (2006). Structural determinants involved in the regulation of CXCL14/BRAK expression by the 26 S proteasome. *J. Mol. Biol.* 363, 813-822.

Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D. *et al.* (2002). A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. USA* 99, 16742-16747.

Privalov, P. L. (1979). Stability of proteins: small globular proteins. *Adv. Protein Chem.* 33, 167-241.

Provencher, S.W., and Glockner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* 20, 33-37.

Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* 47, 83-229.

Ptitsyn, O. B., and Uversky, V. N. (1994). The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett.* 341, 15-18.

Radford, S. E., Dobson, C. M., 1999. From computer simulations to human disease: emerging themes in protein folding. Cell 97, 291-298.

Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradović, Z., Uversky, V. N. *et al.* (2007). Intrinsic disorder and functional proteomics. *Biophys. J.* 92, 1439-1456.

Raghava, G. P. S. (2002). APSSP2: A combination method for protein secondary structure prediction based on neural network and example based learning. *CASP5* A-132.

Rahman, A. (1964). Correlations in the motion of atoms in liquid argon. *Phys. Rev. A* 136, 405-411.

Ralston, G. B. (1990). Effects of crowding in protein solutions. *J. Chem. Educ.* 67, 857-860.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7, 95-99.

Ramachandran, G. N., Venkatachalam, C. M., and Krimm, S. (1966). Stereochemical criteria for polypeptide and protein chain conformations, III. Helical and Hydrogen-bonded polypeptide chains. *Biophys. J.* 6, 849-872.

Ramakrishnan, C., and Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations, II. Allowed conformations for a pair of peptide units. *Biophys. J.* 5, 909-933.

Ramakrishnan, B., Balaji, P. V., and Qasba, P. K. (2002). Crystal structure of beta 1,4-galactosyltransferase complex with UDP-Gal reveals on oligosaccharide acceptor binding site. *J. Mol. Biol.* 318, 491-502.

Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B., and Longhi, S. (2006). Assessing protein disorder and induced folding. *Proteins* 62, 24-45.

Rivas, G., Fernandez, J. A., Minton, A. P. (1999). Direct observation of the self-association of dilute proteins in the presence of inert macromolecules at high

concentration via tracer sedimentation equilibrium: theory, experiment, and biological significance. *Biochemistry* 38, 9379-9388.

Rivas, G., Fernandez, J. A., Minton, A. P. (2001). Direct observation of the enhancement of noncooperative protein self-assembly by macromolecular crowding: indefinite linear self-association of bacterial cell division protein FtsZ. *Proc. Natl. Acad. Sci. USA* 98, 3150-3155.

Rocken, C., Becker, K., Fandrich, M., Schroeckh, V., Stix, B. et al. (2006). ALys amyloidosis caused by compound heterozygosity in exon 2 (Thr70Asn) and exon 4 (Trp112Arg) of the lysozyme gene. *Hum. Mutat.* 27, 119-120.

Romero, P., Obradovic, Z., and Dunker, A. K. (2004). Natively disordered proteins: functions and predictions. *Appl. Bioinformatics* 3, 105-113.

Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* 23, 327-341.

Saraste, M., Sibbald, P. R., Wittinghofer, A. (1990). The P-loop a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* 15, 430–434.

Sasahara, K., Mcphie, P., Minton, A. P. (2003). Effect of dextran on protein stability and conformation attributed to macromolecular crowding. *J. Mol. Biol.* 326, 1227-1237.

Sasaki, Y., Miyoshi, D. and Sugimoto, N. (2007). Regulation of DNA nucleases by molecular crowding. *Nucleic Acids Res.* 35, 4086-4093.

Savageau, M. A. (1969). Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* 25, 365-369.

Savageau, M. A. (1976). Biochemical systems analysis: A study of Function and Design in Molecular Biology. *Addison-Wesley*, Reading. MA.

Savageau, M. A. (1992). A critique of the enzymologist's test tube. In: Bittar, E.E. (Ed.), Fundamentals of Medical Cell Biology, Vol. 3A. *Academic Press*, New York, pp. 45-108.

Savageau, M.A. (1995). Michaelis-Menten mechanism reconsidered: implications of fractal kinetics. *J. Theor. Biol.* 176, 115-124.

Schenk, P. W., and Snaar-Jagalska, B. E. (1999). Signal perception and transduction: the role of protein kinases. *Biochim. Biophys. Acta* 1449, 1–24.

Schlunegger, M. P., Bennett, M. J., and Eisenberg, D. (1997). Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.* 50, 61-122. Review.

Schnell, S. and Turner, T. E. (2004). Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. *Prog. Biophys. Mol. Biol.* 85, 235-260.

Schnell, S., Maini, P. K. (2000). Enzyme kinetics at high enzyme concentration. *Bull. Math. Biol.* 62, 483-499.

Schweers, O., Cchonbrunn-Hanebeck, E., Marx, A., Mandelkow, E. (1994). Structural studies of tau protein and Alzeihmer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.* 269, 24290-24297.

Sherr, C. J. (2006). Divorcing ARF and p53: an unsettled case. *Nat. Rev. Cancer* 6, 663-673.

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S. *et al.* (2007). DisProt: the database of disordered proteins. *Nucleic Acids Res* 35, Database issue, D786-D793.

Sinha, N., Kumar, S., and Nussinov, R. (2001a). Interdomain interactions in hinge-bending transitions. *Structure: Folding & Design* 9, 1165-1181.

Sinha, N., Tsai, C. J., and Nussinov, R. (2001b). A proposed structural model for amyloid fibril elongation: domain swapping forms an interdigitating beta-structure polymer. *Protein Eng.* 14, 93-103.

Spinelli, S., Liu, Q. Z., Alzari, P. M., Hirel, P. H., and Poljak, R. J. (1991). The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie* 73, 1391-1396.

Srere, P., Jones, M. E., and Mathews, C. (1989). Structural and Organizational Aspects of Metabolic Regulation. *Alan R. Liss*, New York.

Srinivasan, N., Sowdhamini, R., Ramakrishnan, C., Balaram, P. (1991). Analysis of short loops connecting secondary structural elements in proteins. In: Balaram P, Ramaseshan S, editors. Molecular Conformation and Biological Interactions. G N Ramachandran Festschrift. Bangalore: *Indian Academy of Sciences*, pp 59-73.

Stillinger, F. H., and Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.* 60, 1545-1557.

Sundberg, E. J., and Mariuzza, R. A. (2000). Luxury accommodations: the expanding role of structural plasticity in protein-protein interactions. Structure 8, R137-R142. Review.

Sunde, M., and Blake, C. C. F. (1997). The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Prot. Chem.* 50, 123-159.

Sunde, M., McGrath, K. C., Young, L., Matthews, J. M. et al. (2004). TC-1 is a novel tumorigenic and natively disordered protein associated with thyroid cancer. *Cancer Res.* 64, 2766- 2773.

Symth, E., Syme, C. D., Blanch, E. W., Hecht, L., Vasak, M., and Barron, L. D. (2001). Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers* 58, 138-151.

Tainer, J. A., Thayer, M. M., and Cunningham, R. P. (1995). DNA repair proteins. *Curr. Opin. Struct. Biol.* 5, 20–26.

Tan, S. Y., and Pepys, M. B. (1994). Amyloidosis. *Histopathology* 25, 403-414.

Thomas, P. J., Qu, B. H., Pedersen, P. L. (1995). Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* 20, 456-459.

Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527-533.

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579, 3346-3354.

Tramontano, A. (1996). The architecture of loops in proteins. In Advances in Computational Biology, ed. H. O. Villar, *JAI Press*, Greenwich; pp 239-259.

Trexler, A. J., and Nilsson, M. R. (2008). The formation of amyloid fibrils from proteins in the lysozyme family. *Curr. Protein and Pept. Sci.* 8, 537-557.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation, and cell signaling. *J. Mol. Recogn.* 18, 343-384.

Uversky, V. N. (1999). A multiparametric approach to studies of self organization of globular proteins. *Biochemistry (Mosc.)* 64, 250-266.

Uversky, V. N., and Ptitsyn, O. B. (1996). All-or-none solvent-induced transitions between native, molten globule and unfolded states in globular proteins. *Fold. Des.* 1, 117-122.

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41, 415-427.

Valleix, S., Drunat, S., Philit, J. B., Adoue, D., Piette, J. C., Droz, D. *et al.* (2002). Hereditary renal amyloidosis caused by a new variant lysozyme W64R in a French family. *Kidney Int.* 61, 907-912.

van den Berg, B., Ellis, R. J. and Dobson, C. M. (1999). Effects of macromolecular crowding on protein folding and aggregation. *EMBO J.* 18, 6927-6933.

van den Berg, B., Wain, R., Dobson, C. M., Ellis, R. J. (2000). Macromolecular crowding peturbs protein refolding kinetics: implications for folding inside the cell. *EMBO J.* 19, 3870-3875.

Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* 6, 1425-1436.

Verkman, A. S. (2002). Solute and macromolecule diffusion in cellular aqueous compartments. *Trends Biochem. Sci.* 27, 27-33.

Verlet, L. (1967). Computer "Experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Rev.* 159, 98-103.

Vinkemeier, U., Moarefi, I., Darnell, J. E., Jr. and Kuriyan, J. (1998). Structure of the amino-terminal protein interaction domain of STAT-4. *Science* 279, 1048-1052.

Voet, J. G., and Voet Donald Voet, J. G. (1990). Fundamentals of Biochemistry, *John Wiley & Son*, pp 109.

Vonrhein, C., Schlauderer, G. J., and Schulz, G. E. (1995). Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* 3, 483-490.

Walter, H. (2000). Consequences of phase separation in cytoplasm. *Int. Rev. Cytol.* 192, 331-343.

Wang, J., Cao, Z., and Li, S. (2009). Molecular dynamics simulations of intrinsically disordered proteins in Human diseases. *Curr. Comput. Aided Drug Des.* 5, 280-287.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337, 635-645.

Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., Lansbury, P. T. Jr. (1996). NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 35, 13709-13715.

Wells, M., Henning, T., Rutherford, T. J., Markwick, P., Jensen, M. R. *et al.* (2008). Structure of tumour suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA* 105, 5762-5767.

Wenner, J. R. and Bloomfield, V. A. (1999). Crowding effects on EcoRV kinetics and binding. *Biophys. J.* 77, 3234-3241.

Westhof, E., Altschuh, D., Moras, D., Bloomer, A.C., Mondragon, A. et al. (1984). Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature* 311, 123-126.

Wetzel, R. (1996). For protein misassembly, it's the "I" decade. *Cell* 86, 699-702.

Wickstrom, L., Okur, A., and Simmerling, C. (2009). Evaluating the performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* 97, 853-856.

Wierenga, R. K., and Terpstra, P. (1986). Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. *J. Mol. Biol.* 187, 101–107

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E. *et al.* (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* 245, 616-621.

Woody, R.W. (1995). Circular dichroism. *Methods Enzymol.* 246, 34-71.

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J. Mol. Biol.* 293, 321-331.

Yazaki, M., Farrell, S. A., and Benson, M. D. (2003). A novel lysozyme mutation Phe57Ile associated with hereditary renal amyloidosis. *Kidney Int.* 63, 1652-1657.

Zhou, H. X. (2004). Protein folding and binding in confined spaces and in crowded solutions. *J. Mol. Recognit.* 17, 368-375.

Zimmerman, S. B. and Minton A. P. (1993). Macromolecular crowding: biochemical, biophysical, and physiological consequences. *Annu. Rev. Biophys. Biomol. Struct.* 22, 27-65.

Zimmerman, S. B. and Trach, S. O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J. Mol. Biol.* 222, 599-620.

Zimmerman, S. B. (1993). Macromolecular crowding effects on macromolecular interactions: some implications for genome structure and function. *Biochim. Biophys. Acta* 1216, 175-185.

Zimmerman, S. B., Harrison, B. (1987). Macromolecular crowding increases binding of DNA-polymerase to DNA: an adaptive effect. *Proc. Natl. Acad. Sci. USA* 84, 1871-1875.

Zimmerman, S. B., Pheiffer, B. H. (1983). Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or Escherichia coli. *Proc. Natl. Acad. Sci. USA* 80, 5852-5856.

Zomot, E., and Kanner, B. I. (2003). The interaction of the gamma-aminobutyric acid transporter GAT-1 with the neurotransmitter is selectively impaired by sulfhydryl

modification of a conformationally sensitive cysteine residueengineered into extracellular loop IV. *J. Mol. Biol.* 278, 42950–42958.

Publications

List of Journal Publications:

1. **Mattaparthi Venkata Satish Kumar** and Rajaram Swaminathan (2010). A novel approach to segregate and identify functional loop regions in Protein Structures using their Ramachandran maps. *Proteins 78, 900*-916.

2. B. R. Meher, **Mattaparthi Venkata Satish Kumar**, Pradipta Bandyopadhyay (2009). Molecular Dynamics simulation of HIV-protease with polarizable and non-polarizable force fields. *Indian Journal of Physics,* 83, 81-90 (invited article).

3. B. R. Meher, **Mattaparthi Venkata Satish Kumar**, Kausik Sen (2008). Pressure Induced Conformational Dynamics of HIV-1 Protease: A Molecular Dynamics Simulation Study. *IEEE Society*, pp.118-122, doi: 10.1109/ICIT.2008.39.

List of Conference Publications:

1. Abstract submitted to the 54th Annual Meeting of Biophysical Society to be held on 20-24th February 2010 at San Francisco, California, USA.
Title: An Automated approach to segregate and identify functional or disordered loop regions in protein structures using their Ramachandran maps. Control/Tracking Number: 10-A-172-BPS
Authors: **Mattaparthi Venkata Satish Kumar** & Rajaram Swaminathan.

2. Abstract published in the proceedings of The National Symposium on Cellular and Molecular Biophysics (under the aegis of Indian Biophysical Society) held during 22nd to 24th January 2009 at the Centre for Cellular and Molecular Biology, Hyderabad.
Title: A Molecular Dynamics approach to Characterize Intrinsically Disordered Proteins
Authors: **Mattaparthi Venkata Satish Kumar** and Rajaram Swaminathan

3. Abstract published in the proceedings of the International Conference on Bioinformatics (BIOCONVENE-2007), held on 16th -22nd December, 2007 at Hyderabad, India.
Title: Flap dynamics of HIV-1 protease: A comparative study with multiple AMBER force fields.
Authors: B. R. Meher, **Mattaparthi Venkata Satish Kumar** and Pradipta Bandyopadhyay.

4. Abstract published in the proceedings of the 5th International Conference on Bioinformatics (INCOB-2006), held on 18th -20th December, 2006 at New Delhi, India.
Title: Effect of Simulation protocol and force field on the flap dynamics of HIV-1 protease.
Authors: B. R. Meher, **Mattaparthi Venkata Satish Kumar** and Pradipta Bandyopadhyay.

Comments raised by the Thesis External Examiner and my Response

Q. Chapter 4 describes experimental results on effects of crowding on kinetics of alkaline phosphatase and acetyl cholinesterase catalysis. This work has no apparent correlation with work described in other three chapters, which describe computational studies?

Ans. Though the chapter 4 is experimental work and remaining chapters describe computational studies, the common between all these chapters is that they are aimed at studying protein structure and dynamics in disordered conditions. In chapter 4, the proteins, Alkaline Phosphatase and Acetyl cholinesterase were studied under disordered conditions that is crowded media using dextrans whereas in chapter 2, the disordered regions namely loops, chapter 3, the structure of intrinsically disordered proteins and in chapter 5 the Human lysozyme amyloidosis disorder were studied. As a whole all the chapters discuss about the protein condition in disordered environments.

Q. The candidate does report some interesting new data, he finds that the effect of crowding by identical dextrans is different for the two substrates studied, but is unable to suggest a possible explanation or hypothesis for his findings. Perhaps he could be queried on this at the viva?

Ans. Our results imply that increase in activity of $ES^{\neq}$ complex is selective on the substrate. Such an event could entail different reaction mechanisms for the two substrates with acetylcholinesterase. Thus the effects exerted by macromolecular crowding on enzymatic reactions are dependent on both the size and concentration of the crowding agent.

Q. Chapter 2 claims to provide a novel approach to identify functional loop regions in protein structures and some MD studies. The whole approach is based on the assumption that by reducing the multiple torsion angle description to a single parameter, which incidentally is mentioned several times before being defined on p66, a better understanding is possible?

Ans. We have discussed about the details about the calculation of MSRP from Ramachandran Map on p65-66 although in earlier pages we touched the concept.

Q. The findings that helices have small spread, followed by strands in beta sheets, is to be expected for any parameter used, since phi-psi values are confined to a small region, for hydrogen bonds to form in these structures?

Ans. This is true for helices, the phi-psi values are confined to a small region. But our Method of MSRP, has the ability to assess the structural perturbation in those secondary structures also.

Q. The candidate does not seem to realize that a large scatter in phi-psi values (or MSRP) for a particular secondary structure does not necessarily imply a disordered structure (p77)?

Ans. We have mentioned this on p77, “Thus the magnitude of MSRP may not always reflect the structural disorder in the region”.

Q. A typical example is beta turns, wherein the two middle residues have completely different phi-psi values, giving a high MSRP value, but it is a very "structured" motif. Hence mean values for these are quite meaningless, but these should give small "sigma" (std dev) values. The high values in Table 2.7 and Fig 2.4 are therefore surprising-perhaps the candidate has combined the different types of beta turns into a single class-he should check this out by comparing his results with actual torsion angle values. A similar validation for the suggested new parameter should have been carried out when describing the MD results?

Ans. Actually we have considered different types of beta turns into a single category. We have observed for well structured beta turn, the sigma value is low. This I have observed during my earlier calculation itself. But to avoid complexity in classifying the beta turns further into different types, I have included all the beta turns into a single category.

Q. While MSRP may simplify the presentation of results, significant information will be lost since different phi-psi combinations could lead to same value for this Euclidean distance?

Ans. Some times, this is true. This is true when the length of secondary structure contains less than 3 or 4 residues. But the structural perturbation can be known along any secondary structure just by observing the jump in MSRP between any two residues within the secondary structure.

Q. A major problem I faced in evaluating the MD studies in a meaningful way, was in trying to find a rationale behind the choice of structures being simulated (both in Chapter 2 and Chapter 3)-they are not of similar size, secondary structure, resolution or function?

Ans. Our aim is compare dynamics of ordered and disordered proteins. So we looked for well ordered proteins and well disordered proteins. We have considered three sets of Proteins. First set contains ordered proteins (1BGF, 1MUN), second set contains partially ordered protein (2HDL) and the third set contains intrinsically disordered proteins (2SOB, 1LXL, 1VZS and 1JH3). First set, ordered proteins (1BGF and 1MUN) were taken from list of PDBSELECT with <25% similarity and with less than or equal to 2.5 Angstroms resolution determined by X-ray and without chain breaks. These two proteins are well ordered and belong to "all alpha proteins" category. Second set, partially ordered protein which is NMR determined structure. Third set, intrinsically disordered proteins (2SOB, 1LXL, 1VZS and 1JH3) were taken from DisProt database. We have taken ordered and intrinsically disordered proteins of more or less similar Amino acid chain length.

1BGF (124 AA); 2SOB (103 AA)
1MUN (225 AA); 1LXL (221 AA)
2HDL (78 AA); 1VZS (76 AA)

In addition to this the intrinsically disordered proteins (selected above) are involved in important biological function. 2SOB, 1JH3 and 1VZS are mentioned in DisProt database as completely intrinsically disordered proteins. And 1LXL to be partially disordered

protein. So we have selected these four IDPs for our study. And also for these IDPs the NMR structure is available for the entire protein chain.

Q. Their starting MSRP values and average phi-psi values are not listed in Table 2.8, so it is very difficult to understand how any parameter can be compared for such widely different starting structures and lead to biologically interesting results?
Ans. Now the starting MSRP values and average phi-psi values are listed in the Table 2.8.

Table 2.8: *Ordered and Unstructured proteins considered for molecular dynamics simulation*

PDB ID	Name of the Protein	Average MSRP	σ_{MD}*	Chain length	Starting MSRP	Average Phi/Psi angles
1BGF	STAT-4 N-Domain (ordered)	49	2.4	124	45.96	-72.80/-14.61
1MUN	Catalytic domain of MutY from *E coli* (ordered)	71	2.2	225	68.27	-68.97/8.96
2HDL	Brak/CXCL14 (unstructured)	95	6.3	78	104.79	-83.06/61.71
2SOB	Sub-domain of staphylococcal nuclease (unstructured)	108	5.7	103	125.17	-91.82/76.83
1LXL	Apoptosis regulator Bcl-x_L (unstructured)	84	2.9	221	89.28	-59.77/1.56
1LXL*	Apoptosis regulator Bcl-x_L (unstructured region 28-82 alone)	140	6.4	55	159.96	-49.92/49.43
1VZS	F6 subunit of ATP synthase (unstructured)	87	4.1	76	65.39	-70.42/4.91

Q. The finding that terminal loops are less flexible than intermediate loops also needs some explanation?
Ans. Actually, we have observed that the MSRP values do not differ appreciably for loops in between secondary structures and terminal loops (Page No 70).
But the Table 2.7 shows MSRP values a little higher for loops between secondary structures than terminal loops. The loops between secondary structures are smaller in size when compared with terminal loops. So the number of atoms in loops between sec structures are less than terminal loops. Loops with less number of atoms are known to be more flexible than the loops with larger number of atoms. The greater the number of atoms in the loops, the greater will be the restriction in flexibility due to steric hindrance (this is true when the residue in the loop contains bulky groups). As the length of loops (between secondary structures) in the case of All alpha proteins, All beta proteins, All alpha + Beta proteins and All Alpha/Beta proteins, are smaller in size when compared with the terminal loops, we observed higher MSRP values for loops between secondary structures than terminal loops. And also the nature of secondary structure on either side of loop also decides the fate of flexibility of the loop.
In the case of unstructured proteins, we observe more or less same MSRP value for loops between sec structures and terminal loops. This is because of presence of flexible atoms (as the protein is unstructured) in the terminal loops as well as loops between sec structures.

Q. In Chapter3, Fig 3.1E, the entire region RMSD is shown to be larger than disordered region alone, this cannot happen unless there is some large domain motion with respect to the disordered segment-but this will then raise questions

about the reduction in radius of gyration and end to end distance shown in Figures 3.2E and 3.3 E?

Ans. For this 1LXL structure, negative NOE values were observed for residues 35-78. This region is thus highly mobile in solution relative to the overall tumbling rate of the molecule. This region (28-80) is mentioned as disordered region in the DisProt database. We have thus considered this region alone as disordered region and plotted RMSD and compared with RMSD of entire protein. The other part of the protein chain has the following structural features.

The structure at Helix alpha 6 contains a kink at Pro 180 that causes a change in the direction of this helix.

The C-terminal helix (alpha 7) is connected to alpha 6 by an irregular turn composed of two glycines (186 and 187).

Also there is flexible hydrophilic loop between 101 and 103.

Thus the entire Protein chain shows larger RMSD when compared with disordered region alone. The hydrophobic core has lost its compressibility due to the high mobility of loop region between 28 and 80. This is reflected in downward trend in radius of gyration plot (Figure 3.2E). Figure 3.3 E represents actually the SASA of entire protein during time course of the simulation and not the end to end distance (as mentioned by the examiner).

Q. It is interesting that in Chapter3 and 5, the conventional analysis protocols have been followed for MD analysis. However some of the comments made on the basis of distance matrices are not really borne out by figures 3.6 and 5.7?

Ans. Here some of the statements are made related to well ordered or disordered structures throughout the protein chain in the case of Figure 3.6 and beta domain between 43-80 in the case of Fig 5.7. These statements are inferred keeping in mind the average structure obtained from the MD simulation and distance matrix plot. This thing cannot be inferred from distance matrix alone.

Q. Overall the large increase in strand and reduction in helix content reported in Table 5.2 for D67H mutant of lysozyme is not reflected as large differences in any of the other parameters- RMSD, end to end distance, radius of gyration etc shown in various figures as well as Table 5.1 Figures showing actual atomic structures or ribbon diagrams (of either snapshots or MD average) would be useful in visualizing these changes, rather than just the trajectory plots?

Ans. Yes, this is true. This can be shown using MD average structure. I have shown ribbon diagrams of MD average structure (Figure 5.12).

Q. On p 133 reduced B factor values are wrongly interpreted as being associated with more flexibility, whereas it should be opposite?

Ans. Yes, this is true. Reduced B factor implies lesser the flexibility. This is changed now. "The above results reveal that mutations in the beta domain region make this region ~~and other regions (100-110)~~ more flexible and accounts for frequent change in the structure leading to the formation of diverse conformers".

Q. Regarding methodology for MD- the number and type of cations used for neutralizing the system is not mentioned?

Ans. Now included.
1BGF, K+, 2 atoms
1MUN, Cl-, 9 atoms
2HDL, Cl-, 12 atoms
2SOB, Cl-, 9 atoms
1LXL, K+, 12 atoms
1VZS, K+, 1 atom
1JH3, Cl-, 2 atoms,
1REX, Cl-, 9 atoms
1LYY, Cl-, 11 atoms
1LOZ, Cl-, 9 atoms
1W08, Cl-, 9 atoms

Q. Size of TIP3P BOX is not mentioned?
Ans. Size of TIP3PBOX used for the simulation is 10.

Q. Also it is not clear on p68, whether crystallographic waters includes all water molecules in pdb file or only those in first hydration shell-generally MD simulations do not consider these waters?
Ans. We have considered the crystallographic waters mentioned in the pdb file.

Q. A few references are not included in reference list: example Lovell et al (2003)?
Ans. Now included

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M. *et al.* (2003). Structure validation by Calpha geometry: phi, psi and Cbeta deviation. *Proteins.* **50,** 437–450.

Oldfield et al., 2000 (Page No 89) is now changed to Oldfield *et al.*, 2005b

Symth et al., 2001(page No 175) is now changed to Smyth *et al.*, 2001

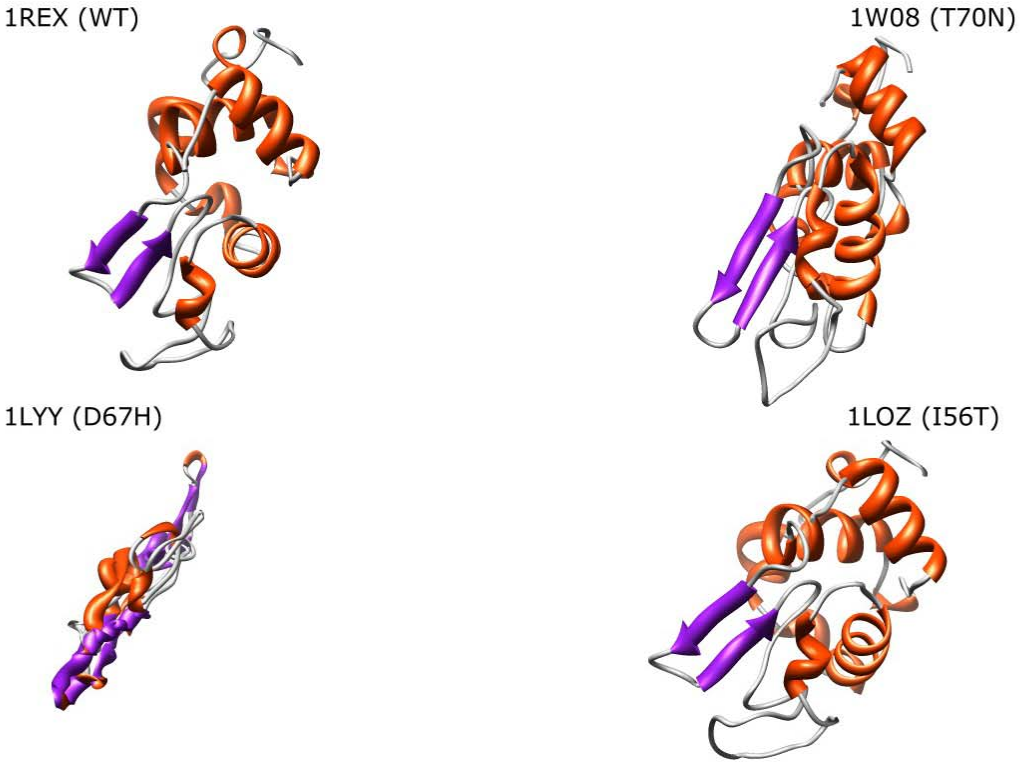

Figure 5.12: *Molecular Dynamics Average structure for wild type and mutants of Human Lysozyme. The conversion of some portion of helices to beta strand is seen clearly in the case of mutants especially 1LYY (D67H).*