INVESTIGATIONS INTO PATTERNS OF INTERACTIONS INVOLVING SEQUENTIALLY NEIGHBORING AMINO ACIDS IN FUNCTIONAL PROTEINS

THESIS SUBMITTED TO
SAVITRIBAI PHULE PUNE UNIVERSITY

FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN BIOTECHNOLOGY

BY
MANAS SANJAY SULE

RESEARCH GUIDE
DR. C. G. SURESH

DIVISION OF BIOCHEMICAL SCIENCES
CSIR-NATIONAL CHEMICAL LABORATORY
PUNE – 411008.
MAHARASHTRA, INDIA

Dedicated to

Namrata Mami & Milind Mama

सीएसआईआर - राष्ट्रीय रासायनिक प्रयोगशाला
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
डॉ. होमी भाभा मार्ग, पुणे - 411 008. भारत
CSIR - NATIONAL CHEMICAL LABORATORY
(Council of Scientific & Industrial Research)
Dr. Homi Bhabha Road, Pune - 411 008, India

CERTIFICATE

This is to certify that the work incorporated in the thesis "**Investigations into patterns of interactions involving sequentially neighboring amino acids in functional proteins**" submitted **by Mr. Manas Sanjay Sule** was carried out by the candidate under my supervision/guidance. Such material as has been obtained from other sources has been duly acknowledged in the thesis.

Research Supervisor
Dr. C. G. Suresh
Division of Biochemical Sciences
CSIR-National Chemical Laboratory

Date: June 2016

Communication Channels
NCL Level DID : 2590
NCL Board No. : +91-20-2590 2000
EPABX : +91-20-2589 3300
: +91-20-2589 3400
FAX
Director's Office : +91-20-2590 2601
COA's Office : +91-20-2590 2660
COS&P's Office : +91-20-2590 2664
WEBSITE
www.ncl-india.org

DECLARATION BY THE CANDIDATE

I hereby declare that the thesis entitled "**Investigations into patterns of interactions involving sequentially neighboring amino acids in functional proteins**" submitted by me for the degree of Doctor of Philosophy is the record of work carried out by me during the time period from November 2010 to June 2016 under the guidance of **Dr. C. G. Suresh** and has not formed the basis for award of any degree, diploma, associateship, fellowship, titles in this or any other university or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in this thesis.

Manas Sanjay Sule
Division of Biochemical Sciences,
CSIR- National Chemical Laboratory,
Pune – 411008.

Acknowledgement

The writing of this thesis marks the culmination of my journey towards obtaining Ph.D. It would not have been possible to accomplish this without the help, support and encouragement from many people including my well-wishers, my friends and colleagues.

First and foremost I am extremely thankful to **Dr. C. G. Suresh**, my research guide, for giving me complete freedom to execute my work. This work would not have been possible without his guidance, support and encouragement. Ever since my joining his lab, he has truly taught me the actual nature of the subject structural biology and bioinformatics. I cannot forget the time when he first taught me how to look at a protein structure. It is under his guidance that I was able to gain in-depth knowledge of structural bioinformatics. His teaching not just limits to scientific knowledge but extends to many approaches in life. To this day he is one of the best teachers that I have ever had.

I would also like to thank **Dr. Vidya Gupta**, for her critical suggestions in the work. It was on the basis on many of her suggestions that I was able to give form to my thesis. I would also like to thank **Dr. J. K. Pal**, Department of Biotechnology, University of Pune for his suggestions during the evaluation seminars. I would also like to thank **Dr. Sushama Gaikwad** for her timely opinions in the course of my doctoral studies. I would also like to thank **Dr. Sureshkumar Ramasamy** for his insights during the development of the iRDP web server.

My entire journey here at CSIR-National Chemical Laboratory would not have been a smooth one but for the support of my friends and lab mates. My seniors **Dr.**

Nishant kumar Varshney and **Dr. Urvashi Sharma** have not just been my labmates but have always treated me like their own brother. Urvashi not only taught me a lot but also cared for me during my illness. She was the first person who introduced me to the field of protein crystallography by teaching me to setup crystallization plates. I will always remember Nishant teaching me to handle the X-ray diffraction instrument in the lab as well as learning from him how to solve diffraction data. I would also like thank **Dr. Poorva Dharker** for her advice during the initial stages of my work. I would also like to thank **Dr. Tulika Jaokar** and **Payal** for their support. **Dr. Priyabrata Panigrahi** has played a major role in the execution of my doctoral work by helping me develop more accurate scripts to carry out my work faster. It has been with his help and enthusiasm that the development and implementation of the iRDP web server has taken place. I would also like to thank my other lab-mates **Dr. Ranu Sharma, Deepak Chand, Manu M. S., Ruby Singh, Ameya Bendre, Deepanjan Ghosh, Yashpal Yadav, Vijay Rajput**, **Debjyoti Boral, Shridhar Chougule, Swati Sinha, T. Selvi, Tejashri Hingmire** and **Aditi Bhand** for their help and support.

At this point, I must acknowledge **Dr. Trupti Kotbagi** who has been one of my earliest friend's in CSIR-NCL. Many times it has been her dedication and determination towards research that has inspired me in my journey here.

This acknowledgement would indeed be incomplete without the mention of **Milind Gupte** and **Namrata Gupte**, my uncle and aunt. During my stay here in Pune, they have treated me as their own son and given me a home away from my home. For this I would be forever be indebted to them.

I would like to thank **my parents and my sister, Juili** for their help, support and their unwavering belief in me that has led to me being at this stage. Right from the beginning of my work here, they have stood steadfast behind me thus enabling me to complete my work. I would also like to thank to **my in-laws** for their support during my doctoral studies.

Words fail me in thanking my wife, **Sayli** for her support during my time here. The existence of this thesis is largely due to her help and her belief in my work and me. Only her trust, love and immeasurable understanding have helped me bring this thesis to its present form.

I would like to thank Director, CSIR-National Chemical Laboratory (CSIR-NCL, Pune) for giving me the opportunity to work in this great institution. I would like to thank Council of Scientific & Industrial Research (CSIR) for the fellowship and CoESC for funding.

Finally I would like to thank all those who have directly or indirectly supported me.

Manas Sanjay Sule

Table of Contents

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

ABSTRACT

Protein sequence is considered as the first level of protein structure. The principles of a linear polypeptide folding into an exquisite three-dimensional structure, possessing specific biological function, are not yet fully understood. The classic experiment by Anfinsen demonstrated that the conformation of a protein often depended on amino acid sequence (Anfinsen *et al*., 1961). Arthur Lesk quotes, one of the basic laws of molecular biology is that "The gene sequence determines amino-acid sequence, the amino-acid sequence determines the protein structure and the protein structure determines the protein function." The regular structures such as α-helices and β-sheets constitute the second level of protein organization.

Protein structural motifs are often described as a succession of secondary structures. They connect secondary structure elements. Structural motifs are short segments of protein 3D structure, which may be spatially close but not necessarily adjacent in sequence. Structural motifs can play functional or structural role. An example of a short structural motif that generally performs a structural role is a beta-turn. A turn is defined as consisting of four consecutive residues where the polypeptide chain folds back. Structural motifs like helix-turn-helix motif play an important functional role in DNA binding.

Sequence motifs or Short linear motifs (SLiMs) in proteins are functional microdomains of immense importance in biological systems. They usually consist of a 3 to 10 residue stretch of protein sequence. They can be associated with a specific structure or function. Many protein interactions are facilitated through short linear motifs. These motifs have been implicated in biological processes, such as sub-cellular targeting [KDEL: Golgi-to-Endoplasmic Reticulum retrieving signal], post-translational modification [WxxW: C-Mannosylation site] and protein–protein interactions [LxCxE: ligand motif for the B-domain of the retinoblastoma proteins].

Analysis of short linear motifs can reveal conformational preference of localized folds. Thus, the conformational analysis of these motifs would facilitate structure prediction and homology modeling

In the present thesis, *in-silico* structural studies have been conducted for the identification of sequence motifs involving interactions of polar charged residues, mainly aspartic acid and glutamic acid on one side and arginine on the other. Secondary structure preferences, conformations and interactions within the motifs between Asp or Glu on one side and Arg on the other, next to each other or separated by varying number of residues, such as D-X(0,8)-R

and E-X(0,8)-R and D or E and R reversed, have been studied. Since the analyses of short structural motifs are quite extensive and cumbersome, a web based tool named as iMotifs has been developed for faster and efficient analysis. Since amino acids repeats also constitute sequence motifs and have been implicated in several medical conditions, secondary structure analysis for 3-8 residue repeats was carried out in protein structures. The conformational parameters for amino acids in such repeats were calculated and compared with the general conformational parameters given by Chou and Fasman for secondary structure prediction.

The thesis is organized into the following chapters:

Chapter 1: Introduction.

Chapter 2: iMotifs: A web-based tool for the analysis of sequence motifs in protein structures.

Chapter 3: Identification of D-R, R-D, D-X-R, R-X-D, DK, KD, D-X-K and K-X-D motifs, analysis of their conformations and hydrogen bonded interactions.

Chapter 4: Identification of E-R, R-E, E-X-R, R-X-E, E-K, K-E, E-X-K and K-X-E motifs, analysis of their conformations and interactions.

Chapter 5: Identification and analysis of D-X(2,8)-R, R-X(2,8)-D, E-X(2,8)-R, R-X(2,8)-E motifs, their occurrence and interactions.

Chapter 6: Analysis of the conformational preferences of homo-polymeric amino acid repeats in known protein structures.

Chapter 7: Comparison of short sequence structural motifs and conclusions.

Chapter 1

Introduction

Proteins form a critical and integral part of all living organisms. They are the main workhorses to translate the genetic information in every living being. Thus, determination of the structure of any protein is vital to decoding not only its function but also to infer the intricate mechanism of its structure-function relation. Once the structure determination is carried out, the analysis turns to identification of further structural details such as folds, motifs, active sites, domains as well as interactions amongst them. These features help to not only compare proteins but also to identify the contribution of each residue to structural stability and function. The understanding of the detailed mechanism of action will depend on this information. The study here investigates the occurrences of short patterns of amino acid sequences, flanking oppositely charged amino acids, occurring in unrelated proteins as structural motifs and look for characteristic features of such identified motifs including a detailed study of the interactions of the residues involved.

1.1 Protein Structure

Proteins are basically long polymeric chains of amino acids linked by CO-NH peptide bonds. Even though all proteins are synthesized from same set of 20 amino acids, they exhibit diverse biological functions. Proteins primarily vary in sequence length, amino acid composition and assemble into polymeric chains with diverse arrangement of the twenty amino acids along the sequence. The functional diversity observed amongst these molecules is attributed to the chemical nature of the constituting amino acids and their arrangement along the sequence that leads to diversity in three-dimensional structures.

A peptide is formed by two amino acids linked by peptide bond while a polypeptide comprises of several such linked amino acids. The backbone of a linear polypeptide comprises of a repeating sequence of the three atoms in each residue, namely, the amide N, the C_α bearing the side chain and the carbonyl C. These atoms are usually shown as N_i, C_i^α and C_i, wherein i represents the residue number. The maximum distance between C_α atoms of adjacent residues is 3.80 Å. The existence of an asymmetric chiral center at the C_α atom introduces an inherent asymmetry to the polypeptide chain.

The peptide bond has partial double-bond character due to resonance, with bond length being 1.33 Å, shorter than the ideal C-N bond length of 1.45 Å. The N

and H atoms in the peptide unit have a negative and positive charge of 0.20 electron unit respectively. Similar negative and positive charges of magnitude 0.40 electron reside on O and C atoms, respectively. These separated opposite charges confer a dipole moment of 3.5 Debye units on each peptide unit (Poland and Scheraga, 1967). The backbone contains amide –NH– as hydrogen bond donor and carbonyl –CO– as hydrogen bond acceptor. These atoms are vital in forming hydrogen bonds that contribute to the stabilization of the three-dimensional structures, especially secondary structures of proteins.

There are twenty different types of side chains, which characteristically define the twenty naturally occurring amino acids. These side chains range from a single methyl group in alanine to long linear chain in arginine or cyclic group as in tyrosine. The reported bond lengths and angles in some of the side chains are shown in Figure 1.

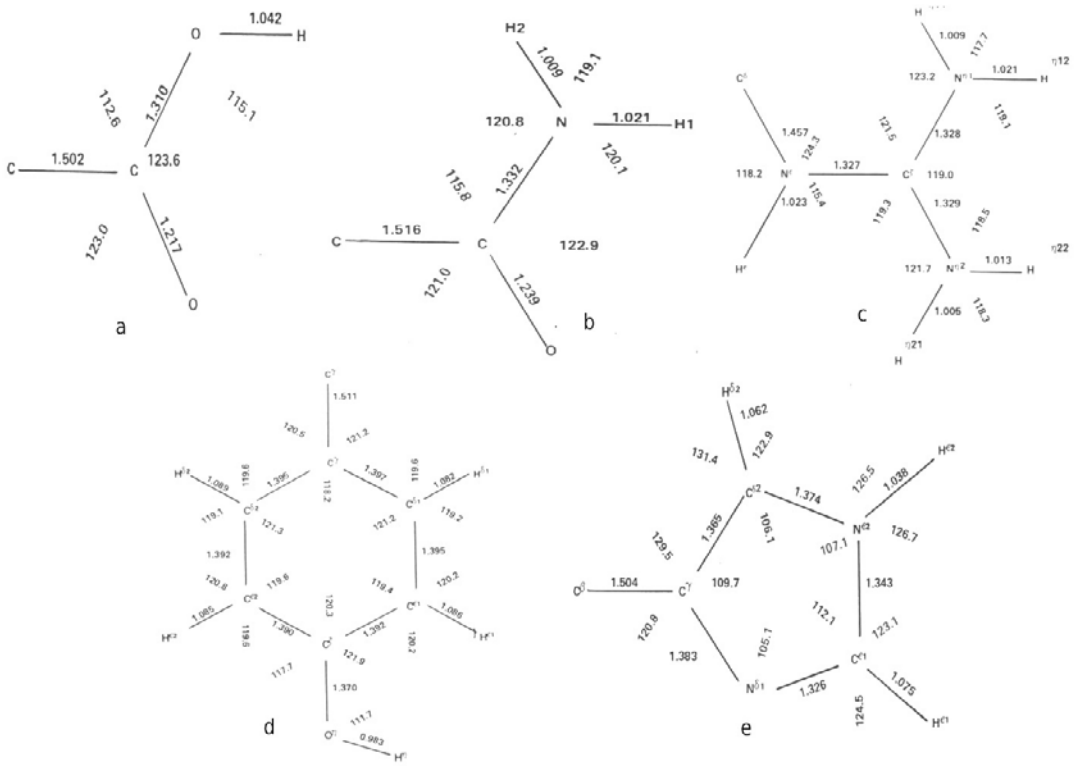

Figure 1. Dimensions of bond lengths and bond angles in some side chains of amino acids. (a) The carboxyl group in aspartic acid and glutamic acid. (b) The amide group in asparagine and glutamine. (c) The guanidium group in arginine. (d) The phenyl group in phenylalanine and phenol group in tyrosine. (e) The un-protonated imidazole group in histidine. (Vijayan, 1976)

1.1.1 The Main Chain Dihedrals

The rotations about the bonds of the peptide unit are referred to as torsions or dihedral angles. These are best described corresponding to a set of fully extended form of two peptide units (Figure 2).

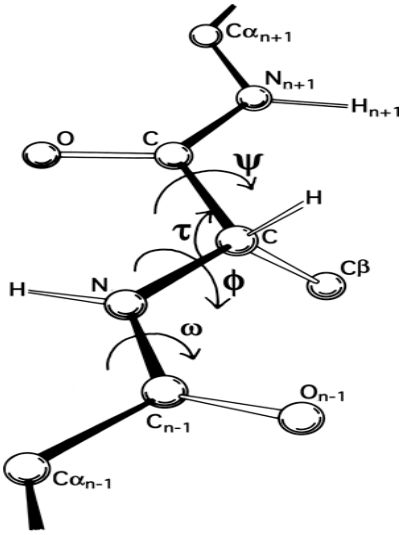

Figure 2. The dihedralangles of the main chain polypeptide (Adapted from http://kinemage.biochem.duke.edu)

The units are co-planar to the atoms N, C_α, C with the distance between the first and third C_α being 7.2 Å and the N-C_α-C angle correspondingly being 110°. The rotation about the N – C_α bond is defined as ϕ (phi) while rotation about the C_α – C bond is denoted as φ (psi). The angles are considered in the range of -180° to +180°. The rotation about the C– N bond is identified as ω (omega) and the values are distributed in the neighborhood of 0° and 180° corresponding to *cis* and *trans* peptide, respectively (Edsall, et al., 1966).

1.1.2 The Side Chain Dihedrals

The next set of angles for free rotations about chemical bonds assigned in amino acid is about bonds of the side chain atoms. This set of dihedral angles is denoted by $\chi_1 - \chi_5$ (Figure 3).

Based on the values of the angles, these dihedrals are assigned into four distinct conformations (Figure 4). Angles with values in the range of 0° – 30° and 0° – -30° are assigned the conformation cis (c). Similarly values in the range of 30° –

90° and -30° – -90° are assigned the conformation gauche+ (g+) and gauche- (g-) respectively. Angles in the range 90° to 120° and -90° to -120° as designated as c+ and c-, while angles with values in the range of 120° – 180° and -120° – -180° are assigned the conformation trans (t) (Chakrabarti and Pal, 2001).

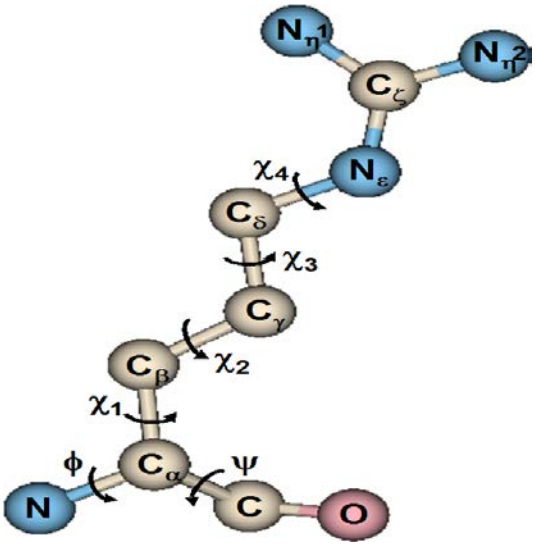

Figure 3. The conformations of side chain dihedral angles.

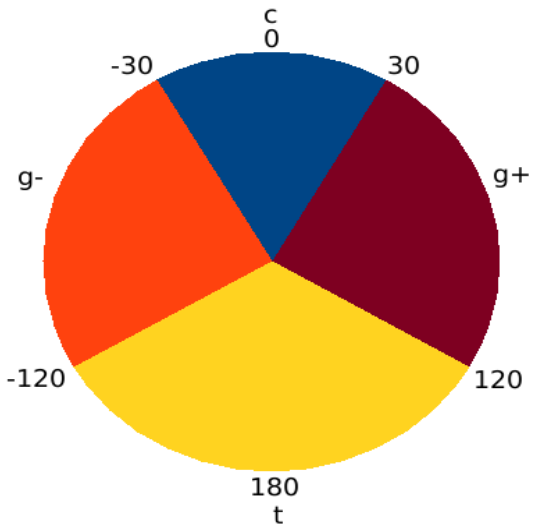

Figure 4. The conformations of side chain dihedral angles.

Along the side chain, the first dihedral angle about the $C_{\alpha} - C_{\beta}$ bond denoted as χ_1 is observed in all amino acids except glycine where side chain is absent. In the present work, the distribution of the values of χ_1–χ_5 has been studied using a set of high resolution protein structures obtained from Protein Data Bank (PDB; Figure 5). The local dataset contained 2602 unique structures yielding 2747 chains. In all

492309 residues from the PDB files were used to carry out this analysis. The distribution of the χ_1 was observed to have g-, g+ and t conformation. The χ_2 angle is observed in all amino acids except alanine and glycine. The distribution of χ_2 was observed to have g- and t conformation in majority. Alternatively in cystine, the conformation was observed to be close to g- (~90°) and g+ (70° to 80°).

The majority distribution of χ_3 angle is observed to have the t conformation with few having g- and g+ conformation. However, in cysteine the conformation was observed to be mainly g- and g+. The distribution of the χ_4 angle in arginine was found to have g+, g- and t conformation.

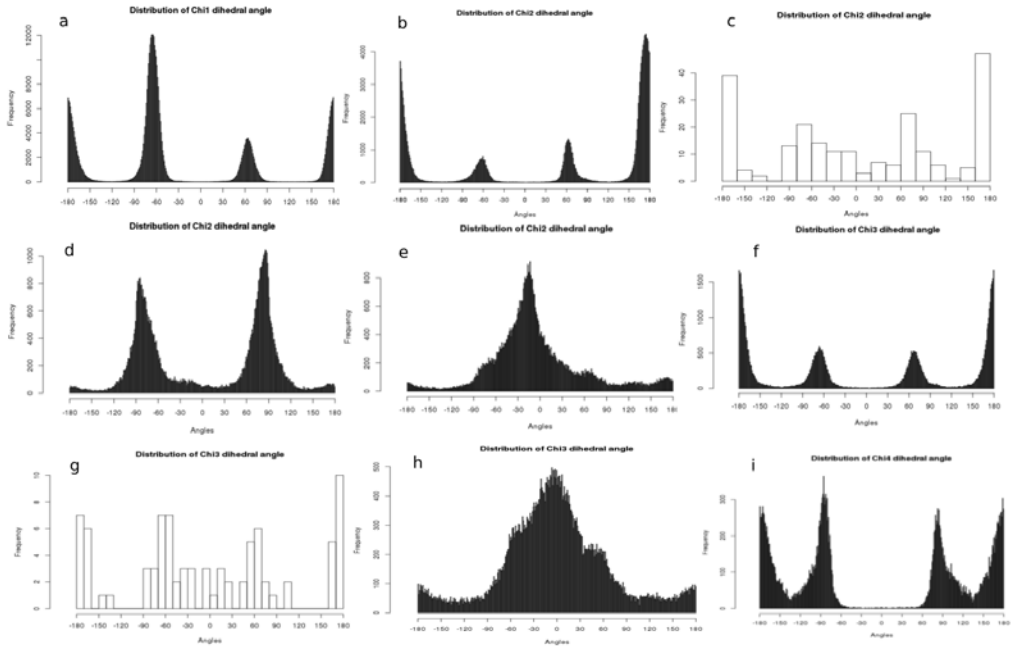

Figure 5. The distribution of side chain dihedral angle values ($\chi_1 - \chi_5$) in amino acids. (a) The distribution of dihedral angle χ_1 in all amino acids. (b) The distribution of dihedral angle χ_2 in leucine, isoleucine, methionine, lysine and arginine. (c) The distribution of dihedral angle χ_2 in cystine. (d) The distribution of dihedral angle χ_2 in phenylalanine, tyrosine, tryptophan and histidine. (e) The distribution of dihedral angle χ_2 in aspartic acid and asparagine. (f) The distribution of dihedral angle χ_3 in methionine, lysine and arginine. (g) The distribution of dihedral angle χ_3 in cystine. (h) The distribution of dihedral angle χ_3 in glutamic acid and glutamine. (i) The distribution of dihedral angle χ_4 in arginine.

1.2 Motifs in Proteins

Protein motifs are groups of structural elements in three-dimensional structures of proteins or a linear group of residues in a primary sequence, which can

be defined by a pattern. These motifs can occur once or in multiples in a structure, sequence or across proteins. The motifs can be broadly classified into structural motifs and sequence motifs. Both structural and sequence motifs provide a better understanding of the structural stability. Some sequence motifs represent structural motifs where, majority of the observed motifs prefer a specific backbone as well as side chain conformation.

1.2.1 Structural Motifs in Proteins

The elementary secondary structures, namely the α-helix, the β-sheet and unordered structures such as turns, coils and loops form the core structural motifs and considered as secondary structures. The α-helix was first defined by Linus Pauling in 1951 (Pauling, et al., 1951). The α-helices are stretches of amino acid residues with ϕ and φ angle pair being approximately -60° and -50°, respectively. The α-helix comprises of 3.6 residues per turn, with i –> i+4 hydrogen bonding and the ends being polar with a dipole moment of 0.5 – 0.6 unit charge (Figure 6). Variations in the α-helix are based on the change in the hydrogen bonding being i –> i+5 in π – helix and i –> i+3 in 3_{10} helix. While only α-helix provides a stable structure, both π – helix and 3_{10} helix are energetically not as favorable since the backbone atoms are tightly packed in 3_{10} helices and extremely loosely packed in π – helices, whereas they are favorably positioned for van der Waals interaction in α-helices (Branden, 1999).

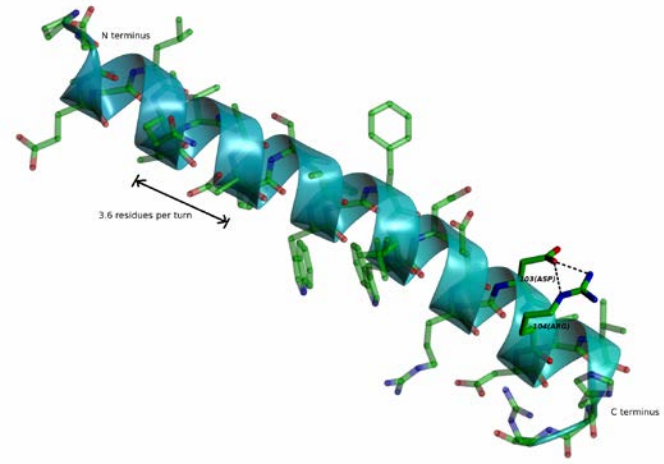

Figure 6. The alpha helix in PDB: 3KB9 with the side chain hydrogen bonding pattern between sequence neighbouring Asp and Arg residues is shown.

The β-sheet structure consists of a combination of multiple regions of the polypeptide sequence called the β-strands. These strands are usually 5-10 residues with fully extended conformations of ϕ and φ angles. The β-strands are aligned adjacent to each other with the C=O groups of one strand form hydrogen bonding interaction with the NH groups of the adjacent strand. The β-sheets can interact in two ways to form a pleated sheet. If the adjacent β-strands are in the same direction, in which case, they form a parallel β-sheet whereas if the successive strands are in opposite directions they form an anti-parallel β-sheet (Figure 7). The two forms have distinctive hydrogen bonding being between amide NH and carbonyl CO of off-placed residues in anti-parallel sheets whereas it is between NH and CO of oppositely placed residues of the two strands in parallel sheets (Branden, 1999).

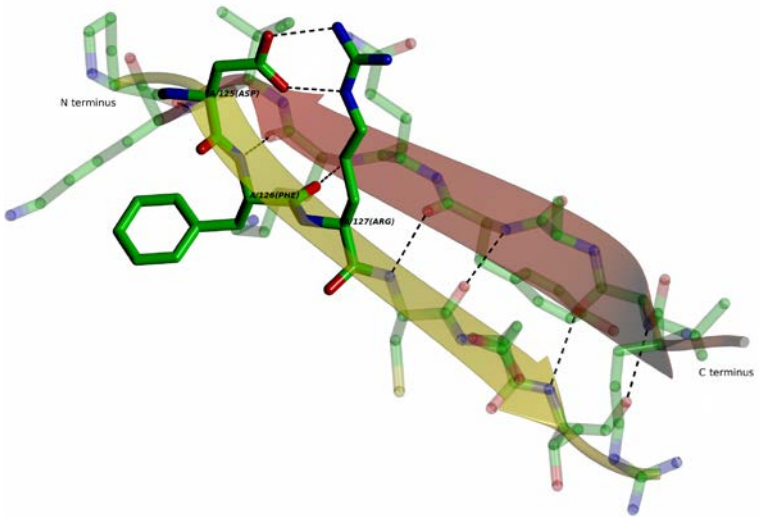

Figure 7. The anti-parallel β – sheet in 2PC1 with the main chain hydrogen bonding holding the sheet and the side chain hydrogen bonding pattern observed between the asparatic acid and arginine residues.

The unordered structures connecting ordered secondary structures are defined as turns or loop regions. These regions are of varied lengths and irregular shapes or belong to a classified β-turn. These stretches usually participate in the formation of ligand binding sites or part of enzyme active sites. An example of this is the antigen-binding site of antibodies, which consists of six loops of varied length and sequence.

Simple combinations of the basic secondary structure elements form geometric arrangements called super-secondary structures or simple motifs. One of the simplest structural motifs is the helix – turn – helix (HTH) motif that is specific

for DNA binding (Figure 8a). Robert Kretsinger first identified this motif (Moews and Kretsinger, 1975).

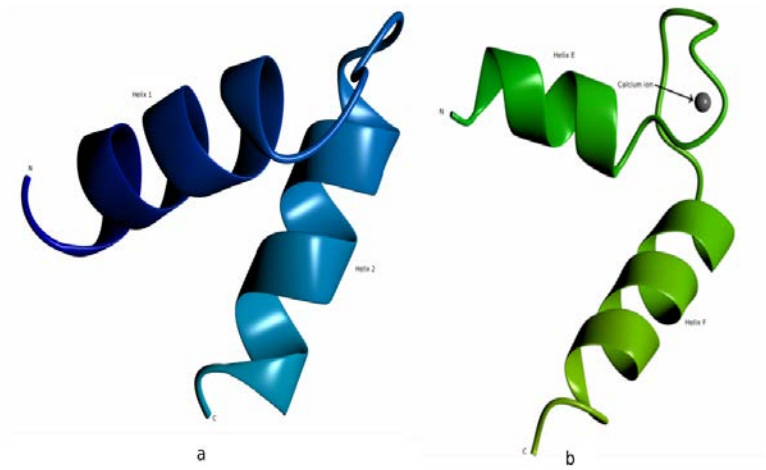

Figure 8. (a) The helix–turn–helix motif in 1B4A. (b) The EF – hand motif with the calcium ion in 4YI9.

Another version of the HTH motif is the EF–hand motif (Figure 8b) since the two helices were initially labeled as helix E and helix F in the parvalbumin structure. A calcium ion is bound in the loop between E and F helices. Calcium binding ligands bind to this motif at the loop regions resulting in the release of the calcium. Another simple motif involving helices is the α-helix hairpin or α-α hairpin (Figure 9), wherein a single loop region connects two successive α-helices. The two α-helices interact with each other through hydrophobic interactions.

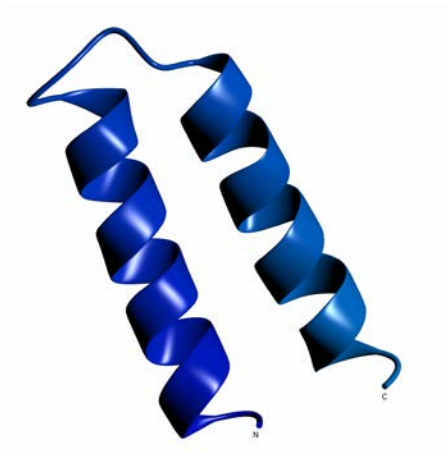

Figure 9. The α-helix hairpin motif in 2BPF.

The hairpin β motif or the β-β unit is the simplest motif involving anti-parallel β structures connected by a loop, where the β-strands can either be isolated or remain as part of β-sheets. The length of the loop region is usually between 3-5 residues. β-hairpins can be classified into two classes. Most β-hairpins are less than 7 residues in length and adopt the classical two residue reverse turn conformations for types I' and II'. For three residue loops in β-hairpins the first residue adopts the right-handed alpha-helical conformation while the second residue is in the region between alpha-helix and beta-sheet in a Ramachandran plot.

Either glycine, asparagine or aspartic acid are found at the last residue position adopting dihedral angles close to the left-handed helical conformation. In case of four residue turns, the first two adopt the right-handed alpha-helical conformation, the third in the region between alpha-helix and beta-sheet and the last being either glycine, asparagine or aspartic acid adopt the left-handed helical conformation (Sibanda, et al., 1991).

Another simple motif involving anti-parallel β-sheets is the Greek key motif (Figure 10a), where four adjacent β-strands are arranged in an ornamental pattern (Hutchinson and Thornton, 1993). In this motif, three anti-parallel strands are connected by β hairpins while the fourth is connected by a longer loop. Although no specific function is associated with this motif, it is found to occur frequently in protein structures. The β-meander motif (Figure 10b) consists of 2 or more successive anti-parallel β-strands connected by hairpin loops. This motif is commonly found in β-sheets and β-strands based architectures such as β-barrels and β-propellers. Multiple anti-parallel β-strands arranged in sheet, which coils and turns to form a β-barrel (Murzin, et al., 1994; Murzin, et al., 1994).

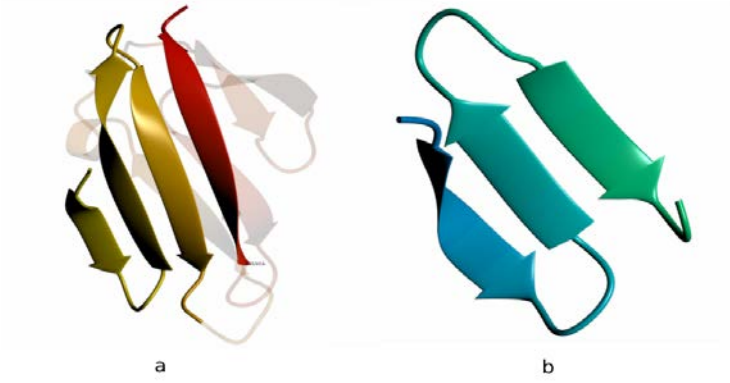

Figure 10. (a) The Greek Key Motif in 1UAS. (b) The β Meander Motif in 1JK4.

The strands comprise of polar and hydrophobic amino acids, in such a manner that hydrophobic residues make up the interior of the barrel to form a hydrophobic core while the polar residues are directed towards the outside of the barrel i.e. the solvent-exposed surface. They can be classified as up-and-down β-barrel or jelly roll β-barrel. A jelly roll β-barrel contains four pairs of anti-parallel beta sheets, wrapped to form a barrel shape of which only one is adjacent in sequence. The motif usually occurs in membrane proteins. The beta-propeller motif is made up of 4 to 8 blade-shaped β-sheets arranged toroidally about a central axis. Each sheet consists of four anti-parallel β-strands such that the first and fourth strands are nearly perpendicular (Murzin, 1992). The motif is commonly found in viral proteins such as neuraminidase as also is a prime component of low density lipoprotein receptor.

Combination of different secondary structure elements as well as simple motifs gives rise to complex structural motifs. The β-α-β motif consists of two parallel β-strands connected through an α-helix (Figure 11a). The motif comprises of a β-strand followed by a loop connecting α-helix and a second loop connecting the second β-strand. The motif is found as a part of most protein structures containing parallel β-sheets. Combinations of such motifs constitute the domain or tertiary structure. Domains are defined as parts of polypeptide chains that can fold independently into stable tertiary structures.

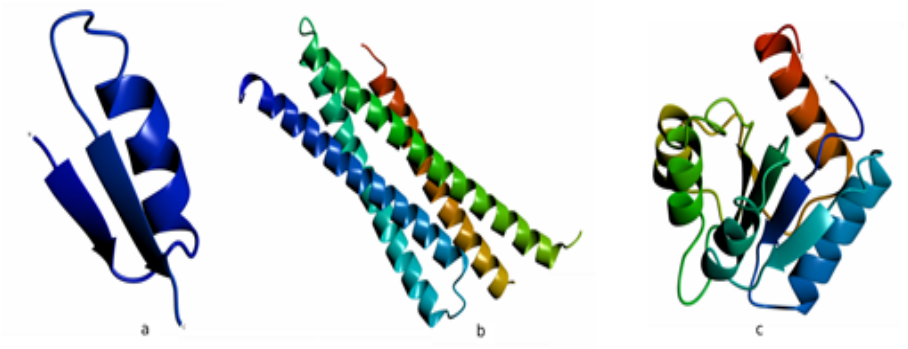

Figure 11. (a) The β-α-β motif in 4HA6. (b) The four helix bundle in 1FIO. (c) The Rossmann fold in 3CHY.

The four helix bundle is the most common α-helical domain in proteins (Figure 11b). The motif consists of four α-helices arranged into a bundle with coiled coil arrangement such that hydrophobic side chains are buried between the helices while hydrophilic side chain are exposed to the solvent (Weber and Salemme, 1980).

The helices are oriented about 20 degrees from their neighbours and the helices are stabilized by salt bridges (Kamtekar and Hecht, 1995). Protein with four helix bundles tend to show higher thermal stability (Harbury, et al., 1993). The motif has been detected in proteins such as Rop, cytochrome, ferrin and Lac repressor C-terminal. Another example of complex motif is the TIM barrel fold named after the triosephosphate isomerase enzyme. It is one of the most common and conserved protein folds. It consists of eight α-helices and eight parallel β-strands that alternate the protein backbone. The helices and strands form a solenoid that curves around in a doughnut shape, forming a toroid. The barrel core is tightly packed with bulky hydrophobic amino acids. The barrel is made up of 200-250 amino acids (Brändén, 1991). An example of motif containing open twisted β-sheets and helices is the Rossmann fold (Figure 11c). The motif contains alternating β-strand, α-helix, β-strand and hence is also called a β-α-β fold. The β-strands form β-sheet. The motif is largely found in proteins which bind nucleotides such as FAD, NAD and NADP. The FAD binding domain is also associated with the consensus sequence Gly-X-Gly-X-X-Gly (Schulz, et al., 1982) where X represents any of the 20 amino acids at that position. The NADPH binding β-α-β fold in NADP dependent enzymes contains the consensus sequence Gly-X-Gly-X-X-Ala, which differs only at the last position, where residue is Ala instead of Gly, thus allowing coenzyme specificity (Hanukoglu and Gutfinger, 1988).

While most motifs comprise primarily of regular secondary structures such as α-helices or β-sheets, the regions connecting these structural elements are regarded as irregular structure elements.

Based on domains in protein structures, they can be largely classified into four distinct structural classes (Chothia and Michael, 1976). The four classes are as follows:

(1) All α proteins: Proteins containing only α-helix secondary structures.
(2) All β proteins: Proteins having mainly β-sheet structures.
(3) α + β proteins: Proteins containing α-helical and β-sheet structures that do not combine and are segregated along the protein sequence.
(4) α/β proteins: Proteins containing alternating regions of α-helical and β-sheet secondary structures.

The MegaMotifBase is one of the most comprehensive collection of structural motifs for 1032 protein families and 1194 superfamilies (Pugalenthi, et al., 2008). The SCOP: Structural Classification of Proteins database (Release 1.75, June 2009) contains the distinct classification of proteins into these structural classes (Murzin, et al., 1995). Table 1 gives the statistics of currently available data in the SCOP database. SCOP2 is the knowledge-based derivative of the SCOP database. SCOP2 focuses not only on structural classes but takes into account both protein sequence as well as evolutionary relationships (Andreeva, et al., 2014).

1.2.2 Sequence Motifs in Proteins

Sequence motifs are short regions within in the whole sequence of proteins that have specific structural and/or functional interpretations attached to them. They are unique and have detectable sequence features, which helps to distinguish a set of protein sequences from the rest. These motifs echo specific functional or structural constraints that also help to imply common descent.

Table 1. Statistics of protein structural classes in SCOP2 database.

Protein Classes	Folds	Protein Superfamilies	Protein families
All α proteins	284	507	871
All β proteins	174	354	742
α/β proteins	147	244	803
α+β proteins	376	552	1055
Multi-domain proteins	66	66	89
Membrane & cell surface proteins	58	110	123
Small proteins	90	129	219
Total	**1195**	**1962**	**3902**

Thus, protein sequence motifs can ideally act as signatures of protein families; acting as instruments for protein structure-function prediction (Bork and Koonin, 1996).These motifs can either be of fixed length such as the cell signal recognition motif R-G-D or of variable length like P-X(2)-G-E-S-G(2)-[AS]-x(0,200)> motif, observed in mammalian Complement C2 in the complement system. Sequence motifs

can be used to define functional constraints such as phosphorylation site motifs, metal binding sites, enzyme active site motifs, nucleotide binding, covalent attachment sites as well as cellular regulation and targeting of proteins to particular subcellular locations. Cyclic-AMP dependent kinases have the consensus sequence motif R-X(1,2)-S/T or R-R/K-X-S/T-I/V/L while Cyclic-GMP dependent kinases have R/K(1,3)-X(1,3)-S/T-R/K(0,1). Ca^{2+} calmodulin dependent kinases have the consensus motif R-X-X-S/T-I/L/V. N-glycosylation sites have the sequence motif N-P-S/T-P (Gavel and von Heijne, 1990) whereas O-glycosylation sites have the motif S-X(2)-P (Gooley, et al., 1991). The metal binding calmodulin binding IQ motif has the 23 amino acid consensus sequence A-X(3)-I-Q-X(2)-F-R-X(4)-K-K (Rhoads and Friedberg, 1997). DNA binding zinc fingers have a wide range of motifs with the overall consensus sequence C-X(2)-C-X(17)-C-X(2)-C (Klug and Rhodes, 1987). Many proteases, esterases and serine active sites show the consensus G-X-S-X-G sequence.

Sequence based motifs can also be construed through a structural perspective. The D/N-P-D/N-G motif has been known to form a Type I β-turn in many globular proteins. The Asx (Asp/Asn) residue at the i position in the motif adopts a g+ conformation, allowing the side chain to form a hydrogen bond with the residue in i+2 or i+4 position (Lee, et al., 2008). The motifs G-X-X-X-G and A-X-X-X-A have been found to occur frequently in helices. The G-X-X-X-G motif contributes to helix-helix interactions while the A-X-X-X-A motif stabilizes the intermolecular as well as intramolecular interactions thus enhancing the thermostable nature of the protein containing them. These motifs are usually found in proteins present in extremophiles (Kleiger, et al., 2002).

Prosite is the most widely used tools for analysis of sequence based motifs. It is a database of protein families and domains. Prosite is based on a huge number of different proteins, most of them can be grouped on the basis of similarities in their sequences into a limited number of families. It currently contains patterns and profiles specific for more than a thousand protein families or domains (Sigrist, et al., 2002; Sigrist, et al., 2012). These signatures are coupled with documentation that provides background information on the structure and function of these proteins. The current release 20.121 (dated 2nd December, 2015) contains 1744 documentation entries, 1309 patterns and 1141 profiles. The ProRule section of Prosite constitutes manually

created rules for automated generation of annotation in the UniProtKB/Swiss-Prot format based on Prosite motifs (Sigrist, et al., 2005). In most cases rules are based on Prosite profiles as they are more specific than patterns, but occasionally also make use of patterns. Currently there are 1141 entries in ProRule. ScanProsite is the search and analysis utility of Prosite which allows users to scan proteins for matches against the Prosite collection of motifs as well as against user-defined patterns (De Castro, et al., 2006). It is available as a web-based and standalone tool. The tools also allow for scanning of user-defined patterns in defined format against a variety of sequences and structures such as UniProtKB (Consortium, 2014), PDB (Berman, et al., 2000) or user-generated databases. Out of the 1309 patterns recorded, 1132 were found to occur in structures of the PDB Database.

1.3 Non-covalent Interactions in Proteins

The folding of a polypeptide chain into the functional three-dimensional protein structure involves a wide range of factors. One of the prime factors amongst these are the non-covalent interactions (Kollman, 1977). This folding of the polypeptide chain first into secondary structures and then into complex tertiary and quaternary structures gives rise to unique molecular environments which give the protein its biological function. These molecular environments are the result of the presence of a limited set of non-covalent interactions. These interactions can be broadly classified into short range interactions and long range interactions. While short range interactions are vital to the formation of secondary structures, long range interactions are crucial to stabilization of the native protein conformation (Go and Taketomi, 1978). The most prominent non-covalent interactions are van der Waals interactions, electrostatic and ionic interactions including salt bridges, hydrogen bonds, hydrophobic interactions, aromatic-aromatic interactions and cation-pi interactions (Figure 12) of which hydrogen bonds considered significantly in this work have been described below.

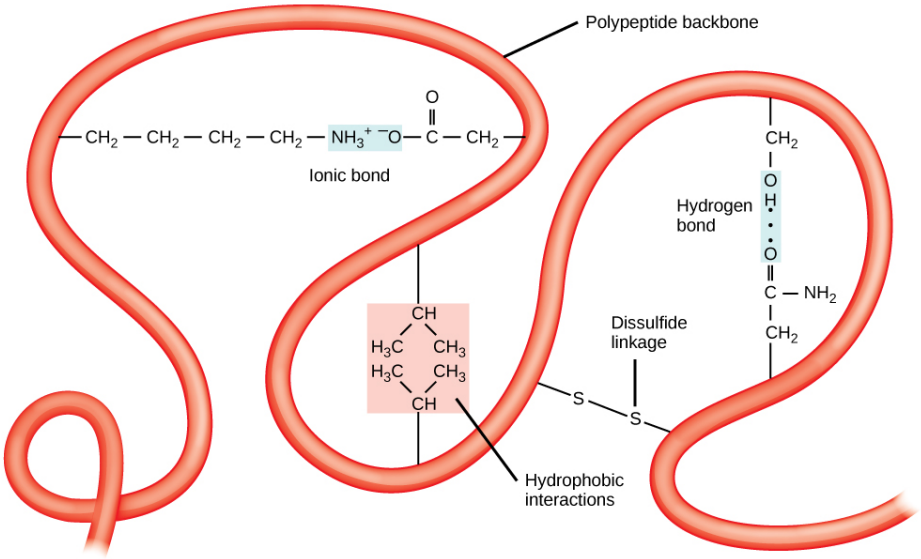

Figure 12: Non-covalent interactions in proteins (Adapted from http://biowiki.ucdavis.edu).

1.3.1 Hydrogen bonds

A hydrogen bond (or H-bond) is an attractive interaction between two electronegative atoms, a donor and an acceptor wherein a hydrogen atom lies aligned between them and covalently bound to the donor. The donor attracts the electron on the hydrogen from its orbital towards the donor itself. This leaves a partial positive charge on the hydrogen, which is electrostatically attracted towards the electronegative acceptor. The interaction is energetically favorable in a number of ways, including polarization energy and covalent energy, but particularly the electrostatic energy. H bonds are typically defined by a distance of less than 3 Å between the H donor and the H acceptor and by hydrogen-donor-acceptor angle below 90° (Baker and Hubbard, 1984; Huggins, 1971; Ippolito, et al., 1990; Latimer and Rodebush, 1920; McDonald and Thornton, 1994; Stickle, et al., 1992). In proteins hydrogen bonds usually involve the C = O and N – H atoms (Figure 13a-b). The H•••O distance is 1.9 – 2.0 Å. The covalent N – H distance is 1.03 ± 0.02 Å. The strength of a hydrogen bond lies within the broad range of 2 – 10 kcal/mol. Hydrogen bonds play a vital role in protein folding as well as structural stability. The ability of hydrogen bond formation amongst the atoms of amino acids allows the formation of secondary structures such as α-helix and β-sheets. They also play a formative role in protein function by interacting with ligands allowing the molecular reaction to proceed. The role of hydrogen bonds in protein stability has been widely studied

using site-directed mutagenesis (SDM), wherein the replacement of Asn to Ala generated a cavity that could be filled with water thus replacing the hydrogen bonding involving Asn (Harpaz, et al., 1994). Thus, it has been estimated that formation of a hydrogen bond results in positive contribution of 1.5 ± 1.0 kcal/mol (Fersht, 1987; Pace, 1994). Hydrogen bonds are also a major determinant of specificity in enzyme reactions, thereby assisting biological information transfer (Fersht, et al., 1985).

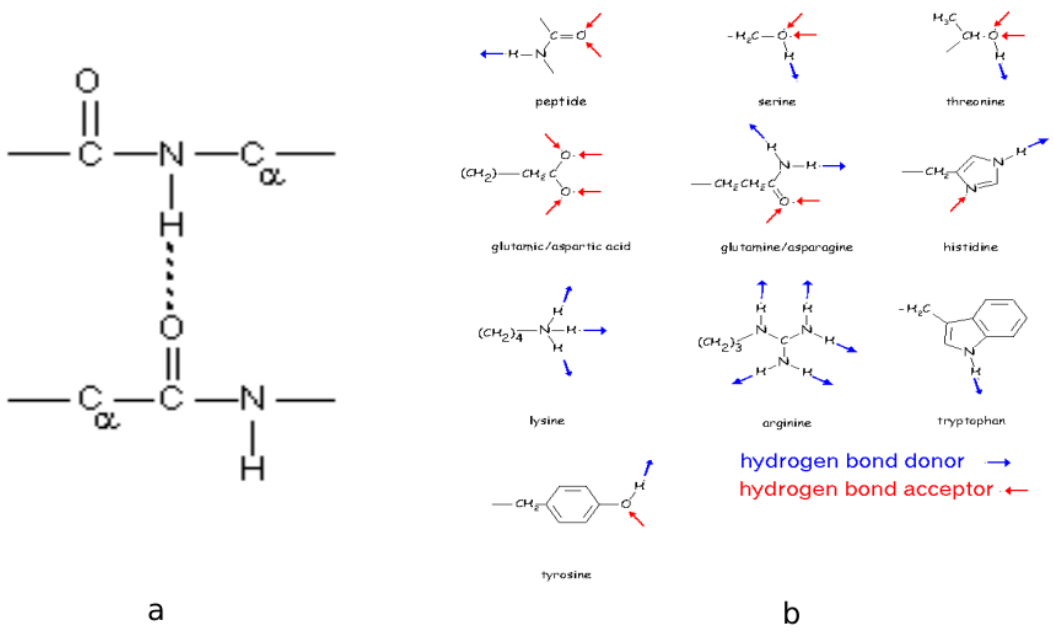

Figure 13. (a) The C = O and N – H atoms involved in a hydrogen bond. (b) Groups in amino acids that participate in hydrogen bond formation (Adapted from www.cryst.bbk.ac.uk).

Based on the atoms involved in the formation of the hydrogen bonds in proteins, they can be classified as main chain-main chain, side chain-side chain and side chain-main chain hydrogen bonds (Figure 14a-b). Main chain - main chain hydrogen bonds are a primary feature of regular secondary structures like α-helices and β-sheets (Kabsch and Sander, 1983). Side chain - side chain H bonds result from the tertiary organization of structural elements in proteins and are vital to protein function (Baker and Hubbard, 1984). Main chain - side chain hydrogen bonds are local in nature and have been found to involve upto five residues on either side of the reference residue including the residue itself (Eswar and Ramakrishnan, 2000; Stickle, et al., 1992).

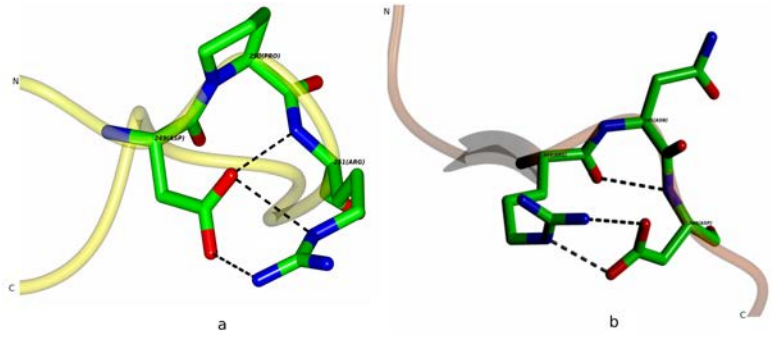

Figure 14: (a) The side chain – side chain and main chain – side chain hydrogen bonds in 1BT3. (b) The main chain – main chain and side chain – side chain hydrogen bonds in 3IKW.

1.4 Studying sequence motifs in protein structures.

Analysis of sequence motifs in protein structures begins with the creation of a study dataset followed by identification of the motif in the structures. The work presented here aims to study a multitude of sequence motifs in protein structures. The motif would be first located in the protein structures with the secondary structure. This would be followed by calculation of the main chain and side chain conformational properties. Finally hydrogen bonded interactions involving the motif residues would be calculated.

1.4.1 Generating the dataset.

In order to investigate sequence motifs in proteins structures, a local dataset of PDB database was generated. Based on the parameters of sequence identity and the crystallographic parameters resolution and R-factor, the dataset was created. R-factor is a measure of the agreement between the crystallographic model and the experimental X-ray diffraction data, i.e. it is a measurement of how well the predicted structure agree with the observed data (Morris, et al., 1992).

The parameter values used are as follows:

1. Sequence Identity: < 25%.
2. Resolution: < 3.0Å.
3. R-factor: < 0.25.

The sequence identity parameter was applied to identify unique protein structures, i.e. to avoid redundancy. The resolution and R-factor parameter values were set so as to ensure accurate computation of dihedral angles and hydrogen bonds.

Based on these parameters, the PISCES Sequence Culling server (Wang and Dunbrack, 2003) was used to prepare the list of PDB ids. The batch download mode of the PDB server was used to retrieve the pdb files identified in the list. The initial dataset comprised of ~5500 structures (November 2010). The dataset was regularly updated to include more unique protein structures. At the time of the last update (December 2015), the dataset contained 12872 files.

1.4.2 Identification of motif location, calculation of conformational parameters and secondary structure.

Identification of the motifs was carried out using the in-house developed iMotifs tool. Development and implementation of the web-based iMotifs tool is explained in Chapter 2. A standalone version of the tool was used to analyze the large number of PDB files in the local dataset. Results from the tool provided the conformational parameters namely the main chain ϕ and ψ dihedral angles along with side chain dihedral angles ($\chi_1 - \chi_5$) for each of the motif residues. The secondary structure assignment was also extracted from the results where residue-wise assignment was provided by DSSP (Kabsch and Sander, 1983). Based on the above parameters and visual inspection of motif conformations, the overall secondary structure assignment of the motif was decided and classified into three distinct groups, namely Helices, Sheets and Irregular structures.

1.4.3 Calculation of hydrogen bonded interactions.

The standalone version of iMotifs was modified in order to concentrate solely on hydrogen-bonded interactions by restricting the calculation of the interactions to hydrogen bonds by HBPLUS. The results thus obtained were sifted through to identify interactions involving the terminal residues of each motif. Here the terminal residues chosen were the aspartic acid and glutamic acid containing anionic carboxylate groups in their side chains and arginine and lysine containing cationic guanidinium and ammonium groups, respectively, in their side chains. These oppositely charged amino acid side chains are expected to interact between them or with the main chain atoms forming hydrogen bonds. Thus, assuming Asp, Glu, Arg and Lys side chains are charged in proteins, by identifying the hydrogen bonds between the oppositely charged groups in them we are also accounting for the salt bridges between these residues.

The hydrogen bonded interactions of guanidinium group with carboxylate group could be broadly classified into four types (Figure 15). **Type A interaction** occurs between the side chain nitrogen atoms (NH1 and NH2) of the guanidinium group and the carboxylate side chain oxygen atoms. **Type B** is found to exist between the side chain nitrogen atoms (NE and NH2) of the guanidinium group and the carboxylate side chain oxygen atoms. **Type C** occurs between the side chain nitrogen atoms (NH1 and NH2) and only one of the carboxylate side chain oxygen atoms while **Type D** is between side chain nitrogen atoms (NE and NH2) and only one of the carboxylate side chain oxygen atoms (Salunke and Vijayan, 1981).

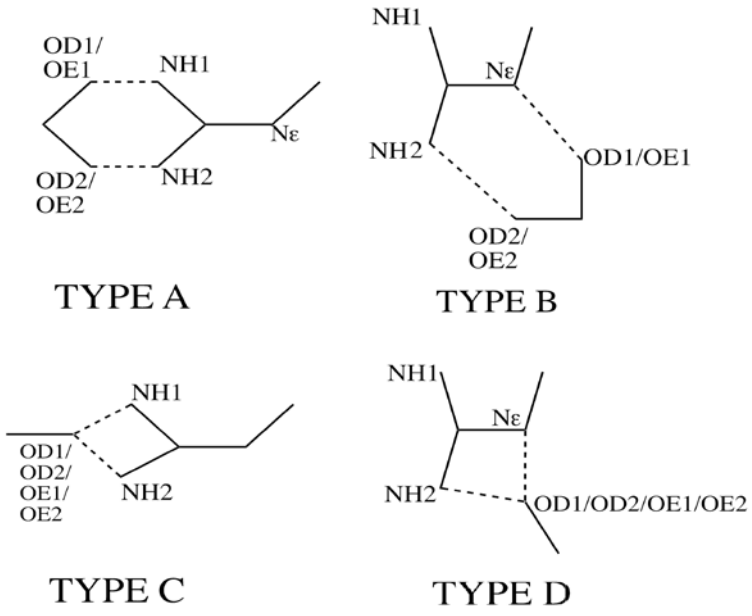

Figure 15. The classification of guanidyl hydrogen bonding.

Based on the type of hydrogen bonding, number of bonds, and secondary structure, groups were considered within each set of identified motifs. These groups were then used for further analysis.

1.4.4 Superimposition analysis of identified motifs.

Motifs within each group were superimposed to understand the similarity within the backbone folding of the motifs. The Root Mean Square Deviation (RMSD) value calculated were then plotted as an n X n matrix (where: n is the number of the motif in each group) (Figure 16).

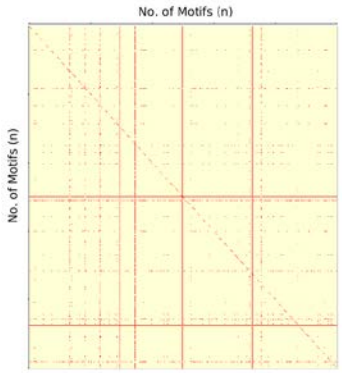

Figure 16. An n X n RMSD matrix.

A cut-off value of 0.3 Å was fixed for plotting. Values ≤ 0.3 Å were plotted as a white square while those > 0.3 Å were plotted as red. Pair of motifs with RMSD values ≤ 0.3 Å were considered to have a similar backbone conformation while those > 0.3 Å were deemed to have varied backbone conformations.

1.4.5 Ramachandran Plots.

The main chain dihedral angles of motif residues in each group were assessed by plotting them on the Ramachandran plot. The script for generating the Ramachandran plot was developed by Peter N. Robinson (version 0.31) in Java (Figure 17). This freeware script was modified to improve the output graphics. The definitions of allowed regions in the plot were taken as those given in Lovell *et. al.* (Lovell, et al., 2003). Ramachandran maps were plotted for each residue of the motif in order to verify the secondary structure in case of helices and sheets. In case of Irregular structural regions, Ramachandran plot allowed the verification of accuracy of the groups formed.

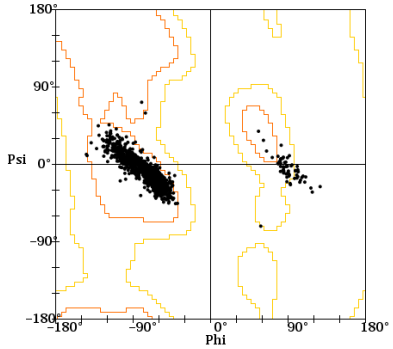

Figure 17. A normal Ramachandran Plot of the X residue in the DXR motif.

1.4.6 Calculation of $C_\alpha - C_\alpha$ distances.

Another method employed for studying the folding of the motif backbone was the calculation of $C_\alpha - C_\alpha$ distance of motif residues. An estimation of the distance between the C_α atoms of the terminal residues reveals the extent of backbone folding. A R-script that employed the Bio3d module was designed to calculate these distance parameters.

1.4.7 Calculation of $C_\beta - C_\beta$ angles.

For two-residue motifs (Example: Asp-Arg motif) the backbone fold was also studied by estimating a virtual dihedral angle using the C_β atoms in the side chains of the motif residues. The Bio3d based R-script was provided with four co-ordinates, namely C^{i}_{β}, C^{i}_{α}, C^{i+1}_{α} and C^{i+1}_{β} used for calculating the angle. The virtual dihedral angles thus obtained for each group were then plotted using a histogram (Figure 18).

1.4.8 Visualization of motifs.

The identified motifs were visualized using QUANTA Modeling Environment (Accelrys Inc.) and CCP4 Molecular Graphics (CCP4MG) (McNicholas, et al., 2011) software. Visualizing the motifs allowed for verification of the secondary structure estimated by DSSP as well as the existence and calculation of the hydrogen bonds carried out by HBPLUS (McDonald, et al., 1993).

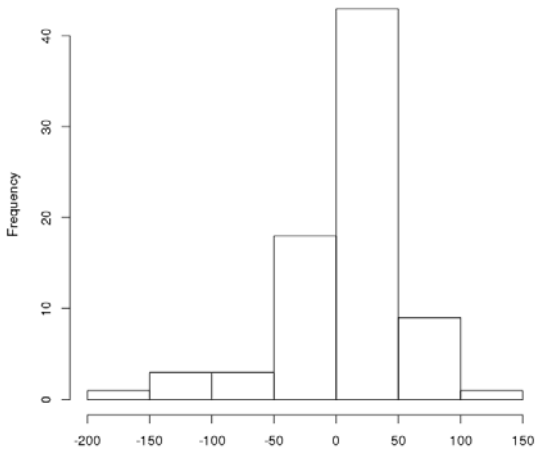

Figure 18. Histogram showing Cβ – Cβ virtual torsion angle in the DR motif.

1.4.9 Distribution of amino acid occurrence in the dataset.

The dataset used for the entire study was first analyzed for the distribution of the amino acid residues. The total number of residues considered for the analysis was found to be 3,561,909, which was used to normalize the occurrence of each amino acid in the dataset (Figure 19).

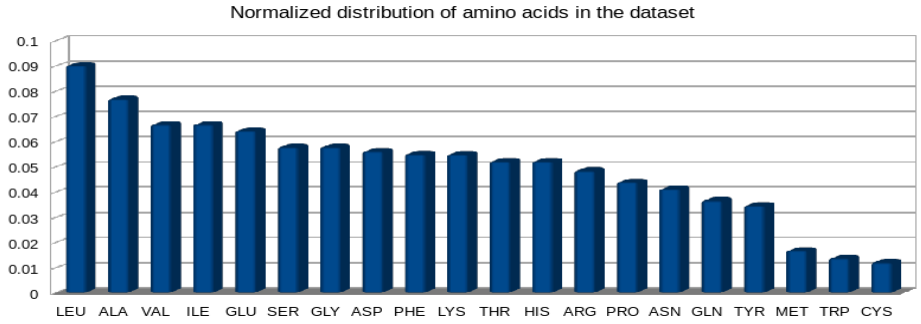

Figure 19: Normalized occurrence of 20 amino acids in the local PDB dataset.

1.5 Organization of the thesis

The remaining thesis after this introductory chapter is organized as follows:

Chapter 2: iMotifs: A web-based tool for the analysis of sequence motifs in protein structures.

This chapter describes the development and working of the iMotifs, a tool used further in all analyses. The currently available tools for analysis of short motifs do not focus comprehensively on features such as secondary structure, main-chain and side-chain dihedral angles, Ramachandran plot and non-bonded interactions. Hence, this is a new tool for thorough analysis of SLiMs developed in the form of iMotifs module. The tool can analyze a wide range of sequence-based motifs in protein structures through its ability to work with Perl-like regular expressions. The tool can analyse motifs in upto 100 protein structures at a time. The tool provides the results in a comparative format, making it simpler for the user to gain a wide overview of structural features. The comparative interaction profile is unique feature, as it provides a comparison of 7 different non-bonded interaction types involving the motif residues. The results can also be downloaded in zipped format which comprises of detailed reports in a tab-delimited format. The tool is freely available as part of the iRDP web server (http://irdp.ncl.res.in).

Chapter 3: Identification of D-R, R-D, D-X-R and R-X-D, K-D, D-K, D-X-K, K-X-D motifs, analysis of their conformations and hydrogen bonded interactions.

The chapter describes the conformational and interaction analysis of four sequence motifs viz. DR, RD, D-X-R and R-X-D. A total of 9351 occurrences of the pattern DR were found. They were classified by the secondary structure in which they occurred. Majority of them were found in loops and turns, which is grossly referred as random coil or irregular structure. 9% of this pattern had interactions involving Asp and Arg. Maximum (584) occurrences with interactions were in random coil. These motifs were classified by the presence of the number of interactions in them. Only eight of them possessed three interactions. They were all found in random coil region. The side chain of Arg showed conformation g+ t g+ t for χ1- χ4 while the Asp χ1 belonged to g+. For motifs with two interactions and occurred in random coil, the side chain conformations remained the same. However, the backbone in this case was found to vary in conformation showing that the loss of one hydrogen bond allowed more backbone flexibility. Next the pattern D-X-R was analyzed. In this case also a significant number of motifs possessing two and three interactions occurred in random coil conformation. The side chain analysis showed a specific folding of the backbone as well as side chain i.e Arg showed the conformation g- t g- t for χ1- χ4 while the Asp χ1 showed t conformation. All occurrences in the random coil with three interactions had the same hydrogen bonding pattern which was observed to stabilize the local fold as compared to the motifs in random coil with 2 interactions. In case of those with 2 interactions, the increase in the average Cα (D) ... Cα (R) distance caused the Asp side chain to slightly move away from the Arg side chain causing the loss of a hydrogen bond. In order to study the importance of the direction of sequence, the position of residues in the sequence were reversed and the patterns RD and R-X-D were analysed. Analysis of the RD motifs clearly exhibited the change in preference of the motifs for secondary structure as a considerable number of motifs with 2 interactions were found to occur in helices. The presence in helix caused the Arg and the Asp side chains to fold with conformation t t g+ t for Arg χ1- χ4 and g- for Asp χ1 so as to establish interactions. Motifs occurring in random coil region also locally folded like helices, but the side chain conformation was highly flexible and hydrogen bonding pattern different. The secondary structural preference for the R-X-D motif with interactions was again random coil region. In both the above patterns,

where 3 and 2 interactions present and occur in random coil, the Arg showed conformation g- g- g- g- for χ1- χ4. On comparison with D-X-R, the overall backbone was found to be constricted since the average Cα(D) ... Cα(R) distance was reduced. The fold with 3 interactions was more uniform in this case also compared to those with only 2 interactions. Similar analysis was carried out for K-D, D-K, D-X-K, K-X-D motifs.

Chapter 4: Identification of E-R, R-E, E-X-R and R-X-E, E-K, K-E, E-X-K, K-X-E motifs, analysis of their conformations and interactions.

The chapter discusses the conformation and interaction analysis of ER, RE, E-X-R and R-X-E sequence motifs in protein structures. Motifs involving glutamic acid in the place of aspartic acid were found to occur in comparatively higher numbers. Total 12533, 10991, 12132 and 10761 occurrences for ER, E-X-R, RE and R-X-E motifs were recorded. In ER, the maximum number of motifs with interactions was found in helices. In such motifs the Arg side chain showed conformation g- g- g+ t and Glu χ1 took t conformation. However, in the case of motifs in random coil, the side chains of Glu and Arg were found to be on the opposite sides of the backbone, bringing the Glu side chain closer to the Arg main chain nitrogen while the Arg side chain lies nearer to the Glu carbonyl oxygen. E-X-R motifs displayed preference for helices; those with interactions occur more in random coils. Motifs involved in helices were found to have the side chains of Glu and Arg in a position not conducive for side chain interactions, resulting in formation of main chain - side chain hydrogen bonds. Though motifs with interactions showed very high presence in helices, motifs with 3 interactions were found mainly in random coil. In such motifs Arg showed the conformation g- g- g- g- for χ1- χ4 while Glu χ1 was g+. This demonstrated that since the Glu side chain was longer, a highly folded conformation of Arg was necessary for the interactions between Arg and Glu. Analysis of motifs in random coil showed some cases where the fold behaved like that in helix having the same side chain conformations as well as hydrogen bonding pattern, while in others the g+ t g- t for Arg and g+ for Glu caused a change in the hydrogen bonding. While motifs in random coil with 2 interactions had exactly the same side chain conformations, hydrogen bonding was found to be different. Interestingly, RE motifs were found to have highly conserved backbone conformation as compared to ER displaying the role of positional interchange. With variations from the first 3 motifs, R-X-E motifs with

interactions showed a greater preference for random coil. Motifs in random coils with 3 interactions showed Arg side chain with highly folded conformation g- g- g- g- and Glu χ1 being g+. On comparison with E-X-R, it was observed that the change in position of R and E, resulted in an increase of the average Cα-Cα distance. Motifs in random coil with 2 interactions were either found to have g- t g+ g+ or g- t g- g- for Arg and g- or g+ for Glu, respectively, resulting in the folding of side chain such that only one of the side chain oxygen atoms of Glu could interact with the Arg side chain. The average Cα (R) ... Cα (E) distance of motifs in random coils was found to be reduced in case of R-X-E, resulting in a much more folded backbone. Similar to the RE motif, it was observed that the backbone conformation showed higher conservation. Similar analysis was carried out for E-K, K-E, E-X-K, K-X-E motifs.

Chapter 5: Identification and analysis of D-X(2,8)-R, R-X(2,8)-D, E-X(2,8)-R, R-X(2,8)-E motifs, their occurrence and interactions.

This chapter deals with the structural and interaction analysis of 4-10 residue motifs involving Asp and Arg or Glu and Arg. The first set of motifs analyzed conformed to the pattern D-X(2,8)-R. A total of 7 motifs resulting from the pattern were analyzed. Overall structural analysis of the motifs revealed the preference of the motifs to occur in random coil. A gradual decrease in the number of motifs involved in interactions was found, as the number of residues separating Asp and Arg increased. In all motifs it was observed that, majority of occurrences with interactions were found in random coil. Only in case of D-X(3)-R, it was seen that the number of motifs with interactions in helices was greater than that of random coil. The next set analysed constituted the reversal of the positions of Asp and Arg; the pattern being R-X(2,8)-D. The preference of the secondary structure even for this motif was observed to be random coil. Similar to first set, an unusually high occurrence of motif with interactions was found in helices in case of R-X(3)-D. Very low occurrence of motifs with interactions was recorded in sheets since side chains of both Arg and Asp would mostly be on opposite sides of the peptide plane. Motifs conforming to the pattern E-X(2,8)-R showed considerably higher presence in helices than random coil in E-X(2)-R and E-X(3)-R. A large number of E-X(3)-R motifs with interactions were in helices. With three residues in between Glu and Arg, they complete a full turn in helix, thereby bringing Glu and Arg side chains closer. Interacting motifs in case of E-X(8)-R also showed increased occurrence in helices. The reversal of E and R

positions was found to alter the preference of the secondary structure to random coil, for the set conforming to the pattern R-X(2,8)-E. Again R-X(3)-E motifs with interactions showed greater preference for helices along with significant numbers in R-X(2)-E. The occurrence of motifs with interactions was found to be very low in sheets. While a gradual decrease of interacting motifs was observed from X(2) to X(8) for random coil, a sudden drop in the occurrence of interacting motifs was observed for R-X(5)-E over all secondary structures, making it the least preferred motifs for the pattern. Patterns involving Asp/Glu and Arg with interactions when separated by 3 residues tend to show a preference for helix as the completion of the turn in the helix allows for the Asp/Glu and Arg side chain to come closer spatially in order to interact.

Chapter 6: Analysis of the conformational preferences of homo-polymeric amino acid repeats in known protein structures.

3-8 residues repeats of all 20 amino acids were analysed in this chapter for their conformational and structural preferences. Alanine showed the highest number of repeats ranging from 3-7 residue repeats. Structural analysis of all the repeats found was carried out. The overall distribution of the repeats in secondary structures revealed specific preferences for the amino acids. Ala, Arg, Glu, Gln, Leu and Lys were found to show higher preference for helices. The repeats of Val, Ile, Thr and Tyr occur in higher numbers in sheets. Amino acids such as Asn, Asp, Pro, Gly, Ser prefer random coil comprising of loops and hydrogen bonded turns. Based on the above results the conformational preference for amino acids in repeats was calculated. The original conformational parameters calculated by Chou-Fasman are widely used for secondary structure prediction. The conformational parameters calculated here for amino acids repeats were compared with those reported by Chou-Fasman. While most amino acid preferences given by Chou-Fasman were same in these repeats, subtle differences were also observed. While amino acids such as Phe, Ile and Met show considerable preference for helix, in repeats these amino acids showed very low preference. The amino acid Cys which usually is preferred in sheets showed equal preference for both helix and sheets on repetition.

Chapter 7: Comparison of short sequence structural motifs and conclusions.

This chapter compares the motifs identified. The chapter discuss the comparison of the secondary structure and side chain preferences along with the interactions between the motifs involving Asp, Arg and Glu, Arg. The chapter discusses the effect of change in the amino acid from Asp to Glu as well as the introduction of spacer residues in the patterns. Finally, the chapter summarizes the conclusions as well as lists the new findings of this research.

1.6. References

Andreeva, A., *et al.* SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research* 2014;42(D1):D310-D314.

Baker, E.N. and Hubbard, R.E. Hydrogen bonding in globular proteins. *Progress in biophysics and molecular biology* 1984;44(2):97-179.

Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Research* 2000;28(1):235-242.

Bork, P. and Koonin, E.V. Protein sequence motifs. *Current opinion in structural biology* 1996;6(3):366-376.

Brändén, C.-I. The TIM barrel—the most frequently occurring folding motif in proteins: Current Opinion in Structural Biology 1991, 1: 978–983. *Current Opinion in Structural Biology* 1991;1(6):978-983.

Branden, C.I. Introduction to protein structure. Garland Science; 1999.

Chakrabarti, P. and Pal, D. The interrelationships of side-chain and main-chain conformations in proteins. *Progress in biophysics and molecular biology* 2001;76(1):1-102.

Chothia, C. and Michael, L. Structural patterns in globular proteins. *Nature* 1976;261:552-558.

Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Research* 2014:gku989.

De Castro, E., *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research* 2006;34(suppl 2):W362-W365.

Edsall, J.T., *et al.* A proposal of standard conventions and nomenclature for the description of polypeptide conformations. *Journal of molecular biology* 1966;15(1):399-407.

Eswar, N. and Ramakrishnan, C. Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Engineering* 2000;13(4):227-238.

Fersht, A.R. The hydrogen bond in molecular recognition. *Trends in Biochemical Sciences* 1987;12:301-304.

Fersht, A.R., *et al.* Hydrogen bonding and biological specificity analysed by protein engineering. *Nature (London)* 1985;314:235-238.

Gavel, Y. and von Heijne, G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein engineering* 1990;3(5):433-442.

Go, N. and Taketomi, H. Respective roles of short-and long-range interactions in protein folding. *Proceedings of the National Academy of Sciences* 1978;75(2):559-563.

Gooley, A.A., *et al.* Glycosylation sites identified by detection of glycosylated amino acids released from Edman degradation: the identification of Xaa-Pro-Xaa-Xaa as a motif for Thr-O-glycosylation. *Biochemical and biophysical research communications* 1991;178(3):1194-1201.

Hanukoglu, I. and Gutfinger, T. cDNA sequence of adrenodoxin reductase-identification of a consensus sequence that distinguishes between type-i nad and nadp binding-sites. *FASEB Journal*. Federation Amer Soc Exp Biol 9650 Rockville Pike, Bethesda, MD 20814-3998; 1988. p. A356-A356.

Harbury, P.B., *et al.* A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* 1993;262(5138):1401-1407.

Harpaz, Y., Gerstein, M. and Chothia, C. Volume changes on protein folding. *Structure (London, England : 1993)* 1994;2(7):641-649.

Huggins, M.L. 50 Years of hydrogen bond theory. *Angewandte Chemie International Edition in English* 1971;10(3):147-152.

Hutchinson, E.G. and Thornton, J.M. The Greek key motif: extraction, classification and analysis. *Protein engineering* 1993;6(3):233-245.

Ippolito, J.A., Alexander, R.S. and Christianson, D.W. Hydrogen bond stereochemistry in protein structure and function. *Journal of molecular biology* 1990;215(3):457-471.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577-2637.

Kamtekar, S. and Hecht, M. Protein Motifs. 7. The four-helix bundle: what determines a fold? *The FASEB journal* 1995;9(11):1013-1022.

Kleiger, G., *et al.* GXXXG and AXXXA: common α-helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry* 2002;41(19):5990-5997.

Klug, A. and Rhodes, D. 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends in Biochemical Sciences* 1987;12:464-469.

Kollman, P.A. Noncovalent interactions. *Accounts of Chemical Research* 1977;10(10):365-371.

Latimer, W.M. and Rodebush, W.H. Polarity and ionization from the standpoint of the lewis theory of valence. *Journal of the American Chemical Society* 1920;42(7):1419-1433.

Lee, J., *et al.* A logical OR redundancy within the Asx-Pro-Asx-Gly type I β-turn motif. *Journal of molecular biology* 2008;377(4):1251-1264.

Lovell, S.C., *et al.* Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* 2003;50(3):437-450.

McDonald, I., *et al.* HBPLUS, a computer program for calculating potential hydrogen bonds in protein structures. *Dept. of Biochemistry, University College London, UK* 1993.

McDonald, I.K. and Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* 1994;238(5):777-793.

McNicholas, S., *et al.* Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica Section D: Biological Crystallography* 2011;67(4):386-394.

Moews, P. and Kretsinger, R. Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *Journal of molecular biology* 1975;91(2):201-225.

Morris, A.L., *et al.* Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics* 1992;12(4):345-364.

Murzin, A.G. Structural principles for the propeller assembly of β- sheets: the preference for seven- fold symmetry. *Proteins: Structure, Function, and Bioinformatics* 1992;14(2):191-201.

Murzin, A.G., *et al.* SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* 1995;247(4):536-540.

Murzin, A.G., Lesk, A.M. and Chothia, C. Principles determining the structure of β-sheet barrels in proteins I. A theoretical analysis. *Journal of molecular biology* 1994;236(5):1369-1381.

Murzin, A.G., Lesk, A.M. and Chothia, C. Principles determining the structure of β-sheet barrels in proteins II. The observed structures. *Journal of molecular biology* 1994;236(5):1382-1400.

Pace, C. Evaluating contribution of hydrogen bonding and hydrophobic bonding to protein folding. *Methods in enzymology* 1994;259:538-554.

Pauling, L., Corey, R.B. and Branson, H.R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* 1951;37(4):205-211.

Poland, D. and Scheraga, H.A. Energy Parameters in Polypeptides. I. Charge Distributions and the Hydrogen Bond*. *Biochemistry* 1967;6(12):3791-3800.

Pugalenthi, G., *et al.* MegaMotifBase: a database of structural motifs in protein families and superfamilies. *Nucleic acids research* 2008;36(suppl 1):D218-D221.

Rhoads, A.R. and Friedberg, F. Sequence motifs for calmodulin recognition. *The FASEB Journal* 1997;11(5):331-340.

Salunke, D. and Vijayan, M. Specific interactions involving guanidyl group observed in crystal structures. *International journal of peptide and protein research* 1981;18(4):348-351.

Schulz, G., Schirmer, R. and Pai, E. FAD-binding site of glutathione reductase. *Journal of molecular biology* 1982;160(2):287-308.

Sibanda, B.L., Sibanda, L. and Thornton, J. Conformation of beta hairpins on protein structures: classification and diversity in homologous structures. *Methods in enzymology* 1991.

Sigrist, C.J., *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics* 2002;3(3):265-274.

Sigrist, C.J., *et al.* New and continuing developments at PROSITE. *Nucleic acids research* 2012:gks1067.

Sigrist, C.J., *et al.* ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 2005;21(21):4060-4066.

Stickle, D.F., *et al.* Hydrogen bonding in globular proteins. *Journal of molecular biology* 1992;226(4):1143-1159.

Vijayan, M. CRC Handbook of Biochemistry and Molecular Biology. *Proteins* 1976;3:742-759.

Wang, G. and Dunbrack, R.L. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19(12):1589-1591.

Weber, P.C. and Salemme, F.R. Structural and functional diversity in 4-alpha-helical proteins. *Nature* 1980;287(5777):82-84.

Chapter 2

iMotifs: A web-based tool for the analysis of sequence motifs in protein structures

Understanding the sequence-structure relationships in proteins holds prime importance in the field of structural biology. This protein folding problem is based on the crucial question i.e. how a linear polymer chain composed of specific sequence of amino acid residues encodes an unique three-dimensional structure upon folding. While current computational approaches for structure prediction are widely based on the three-states i.e. helix, strand and loop that are dominant in proteins structures, the relationships between the sequence and local structure elements remain poorly explored. Most attempts to establish such relationships have been based on identification of structural motifs and characterization of amino acid frequencies at each position in the motif. The prime challenge for such local structure prediction lies in the identification and analysis of sequence patterns with characteristic structural features. The chapter presents the development and implementation of a web based motif analysis tool – iMotifs that can search for specific amino acid sequences, their frequency and structural features in a database of protein structures.

2.1 Tools for analysis of sequence motifs

Currently a wide range of tools are available for the identification and analysis of sequence based motifs. Prosite is the most widely used tools for motif analysis. Prosite is a database of protein families and domains. Prosite is based on a huge number of different proteins, which can be grouped on similarities in their sequences into a limited number of families. It currently contains patterns as well as profiles that are specific for nearly a thousand protein families or domains (Sigrist, et al., 2002; Sigrist, et al., 2012). These signatures are along with documentation that give information on the structure as well as the function of the proteins. The current release 20.114 contains 1722 documentation entries, 1309 patterns and 1115 profiles. The ProRule section of Prosite constitutes manually created rules for automated annotation in the UniProtKB or Swiss-Prot format which are based on Prosite motifs (Sigrist, et al., 2005). Although the rules are dependent on Prosite profiles which are more precise than patterns, occasionally patterns are also used. Currently there are 1115 entries in ProRule.

ScanProsite is the search and analysis tool of Prosite which allows users to scan proteins for matches against the collection of motifs available in as well as against user-defined patterns (De Castro, et al., 2006). It is available as a web-based and standalone tool. It allows submission of protein sequences to scan against the Prosite collection of motifs. The tools allow for scanning of user-defined patterns in defined format against a variety of sequences and structures such as UniProtKB (Consortium, 2014), PDB (Berman, et al., 2000) or user-generated databases. The format of the input motif can be either Prosite accession or identifier, user-defined pattern such as; S-x(2,4)-T-H-A-x-[FG] or combination of Prosite accessions/identifiers and patterns. Users can also submit protein sequences and patterns to scan them against each other. The Motif search server using DBGET and LinkDB identifies sequence motifs in query sequence and also provides functional and genomic information of the identified motifs (http://www.genome.jp/tools/motif/). The tool also allows for search of protein sequence libraries with user-defined patterns. Sequence pattern to be searched must be specified in the format of the Prosite pattern. eMOTIF Database is a collection of highly specific and sensitive protein sequence motifs which characterize conserved biochemical properties, biological functions while carrying out sequence analysis through the application of Pareto-optimal Discrete Motifs (Huang and Brutlag, 2001) (http://motif.stanford.edu/distributions/emotif/). The eMOTIF database comprises of eMOTIF-SEARCH, eMOTIF-MAKER and eMOTIF-SCAN tools which are also available for download. The eMOTIF-SEARCH program on provided with a protein sequence proposes families of proteins to which the query sequence might belong. The program considers a protein sequence as a potential member of a protein family, if the sequence contains discrete motif. The eMOTIF-MAKER program works by converting an ungapped multiple sequence alignment into a set of Pareto-optimal discrete motifs for use by the eMOTIF-SEARCH and eMOTIF-SCAN. The result reveals sub-families within the cluster of proteins sharing sequence similarity covered by the given ungapped multiple sequence alignment. Given a regular expression, eMOTIF-SCAN program discovers occurrences of that regular expression in any of the specified sequence databases. 3motif carries out visualization of discrete protein sequence motifs and their properties in three dimension (Bennett, et al., 2003). The tool is flexible in that the users can enter sequences, keywords, structures or sequence motifs to generate visualizations. Users can search using discrete sequence motifs

such as PROSITE patterns or any other regular expression-like motif. Properties of motifs such as sequence conservation and solvent accessible surface area are also displayed. The interacting motif database or iMOTdb (Bhaduri, et al., 2004; Pugalenthi, et al., 2006), lists interacting motifs that are identified for all structural entries in the PDB.

The conserved patterns or finger prints are recognized for individual structural entries and then grouped together for identifying the common motifs shared among all superfamily members. Interacting motifs can assist in understanding the structure-function relation of proteins. Information on such motifs is valuable in understanding protein folding, molecular modeling and protein engineering experiments. The iMOT DB provides links to aid sequence search protocol using PHI-BLAST (Zhang, et al., 1998) and SCANMOT (Chakrabarti, et al., 2005) employing the interacting motifs. The interacting motifs from superfamilies of proteins are derived using structural alignments obtained from PASS2 (Bhaduri, et al., 2004). These motifs represent a given protein family and provides useful insights regarding the structural and functional role of the protein. SCANMOT carries out multiple pattern search in protein sequences and structures. It also carries out recording of the inter-motif spacing, alignment and searching for statistically significant sequence similarities using sequence and structure databases.

2.1.1 iMotifs : *in silico* Structural Analysis of Sequence Motifs in Proteins

Even with the existence of the above-described tools for motif analysis it was realized that very few tools focus on the use of protein structures for the analysis of motifs. Also, additions of more structural features were necessary to make studies involving motifs meaningful. One of the main features that have not been explored extensively is molecular interaction. With this aim the iMotifs (*in silico* Structural Analysis of Sequence Motifs in Proteins) tool was developed. iMotifs comprises of two distinct sections, namely, the iMotifs tool and iMotif interaction database. iMotifs tool not only identifies sequence-based motifs in protein structures but also evaluates them further in terms of various structure dependent features such as solvent accessibility, secondary structure, location of residues in the Ramachandran plot, main chain and side chain dihedral angles and the occurrence of interactions such as disulphide bonds and 6 other non-covalent interactions. The results are presented to

the user in two-level format. The summary page offers a bird's-eye view into the location and other structural features, while the detailed report provides a more comprehensive insight into the residues in the motif and bonded and non-covalent interactions involving the motif residues. The iMotif interaction database comprises of a detailed structural analysis of 1132 patterns derived from Prosite. iMotifs tool was used to identify the patterns from the PDB database and interaction profile was calculated along-with other structural features. The detailed analysis for these motifs has been compiled and recorded in iMotif interaction database. The iMotifs tool is currently available through the iRDP web server (http://irdp.ncl.res.in).

2.2 Methodology

The iMotifs tool is developed in a Linux platform and utilizes R, Perl, HTML and PHP. The iMotifs tools is primarily based on the Bio3d (Grant, et al., 2006) and iGraph (Csardi and Nepusz, 2006) packages. The structural and interactions analysis carried out by iMotifs employs both in-house developed scripts and established tools (Table 1). Figure 2 gives is the detailed workflow of iMotifs.

Table 1. Public domain tools used by iMotifs for estimation of parameters.

Tools	Purpose	Reference
DSSP	For assignment of secondary structures	(Kabsch and Sander, 1983)
NACCESS	For estimation of residue solvent accessibility	(Hubbard and Thornton, 1993)
Procheck	For calculation of conformational parameters and assignment of Ramachandran Plot regions	(Laskowski, et al., 1993)
HBPLUS	For identification of hydrogen bonds	(McDonald and Thornton, 1994)
SSBOND	For identification of residue pairs for disulfide bond insertion	(Hazes and Dijkstra, 1988)

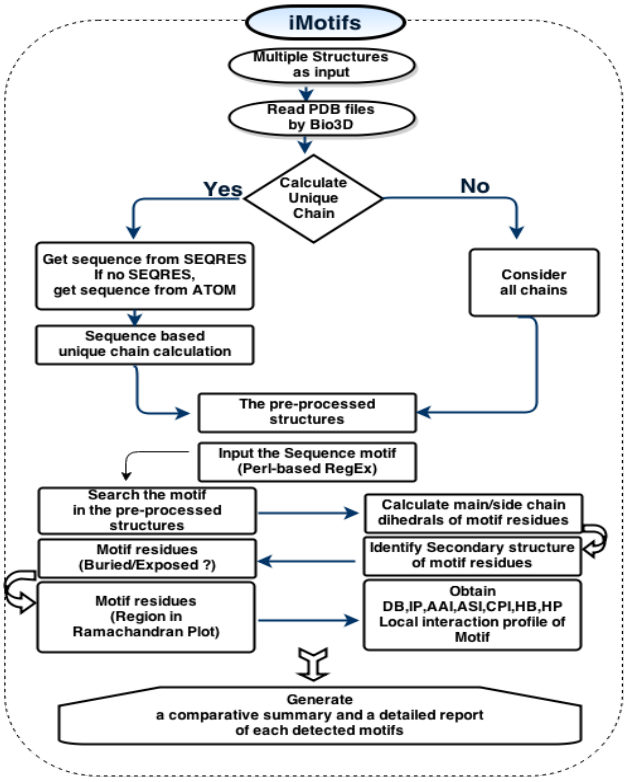

Figure 1. The workflow describing the identification and analysis of motifs in the iMotifs tool.

The input to iMotifs comprises of comma-separated list of PDB ids (Figure 2) and the search pattern in Perl-like regular expression (Figure 3). Apart from these, users can also modify other interactions related parameters such as the distance cut-off for ionic interactions, aromatic-sulphur interactions, aromatic-aromatic interactions, cation-pi interactions, hydrophobic interactions as well as relative solvent accessibility values (Figure 4).

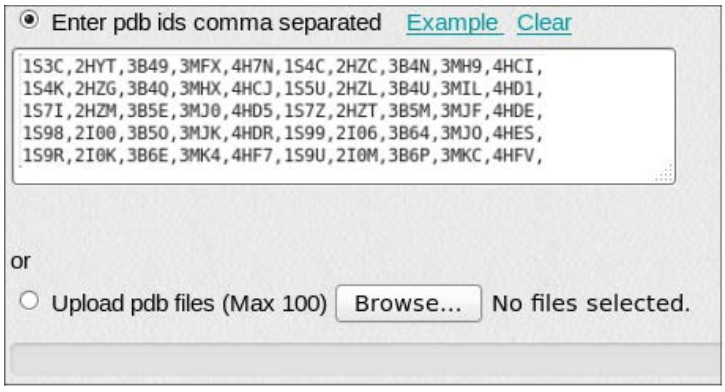

Figure 2. The PDB input that must be provided in the form of a comma-separated list.

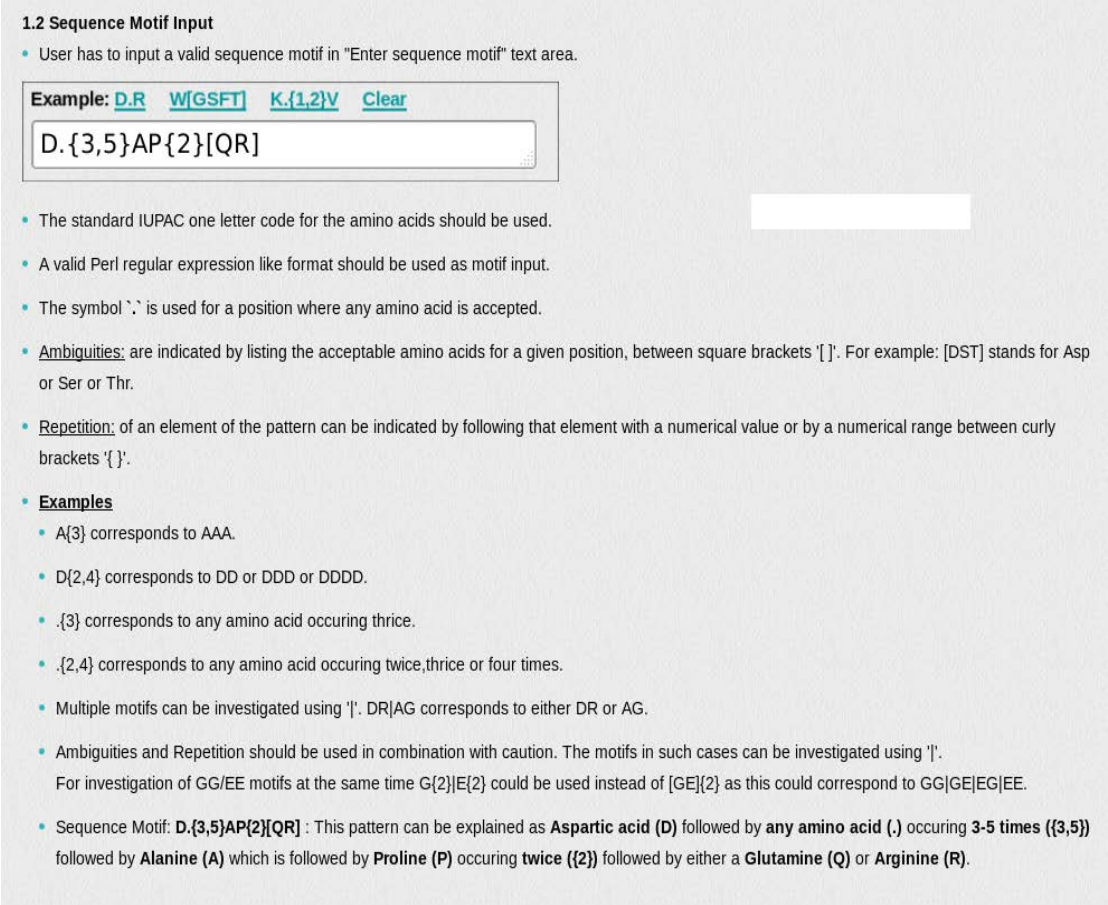

Figure 3. The Perl-like regular expression which serves as the search pattern input to iMotifs.

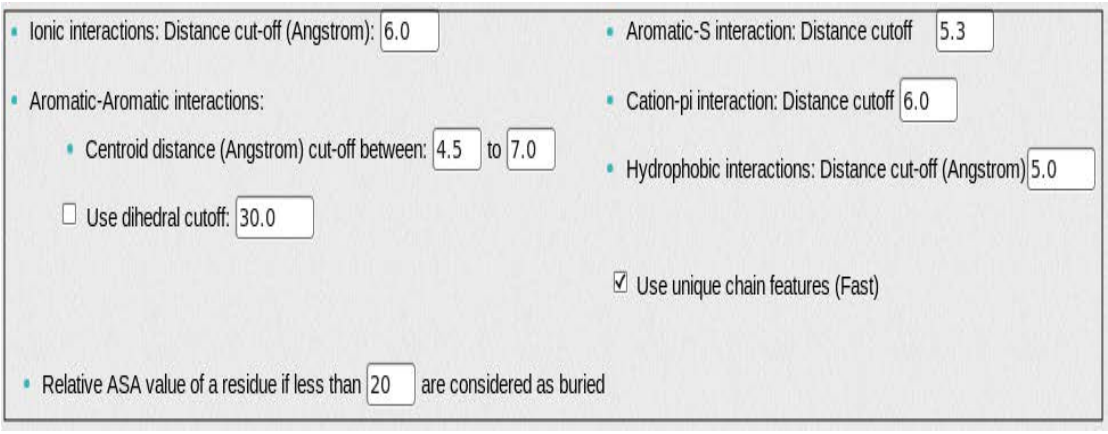

Figure 4. The parameters related to non-covalent interactions and solvent accessibility that can be modified by the users.

Once the PDB files are read by Bio3d the analysis in iMotifs begins. Chains representing nucleic acid part are removed from the file. The unique chain feature then estimates the choice provided by the user. If analysis is to be carried out on unique chains, the sequence is obtained from the SEQRES tag of the PDB file. If the sequence tag cannot be obtained from SEQRES tag, it is obtained from the ATOM

tag. Based on the sequence, unique chains are determined. As an alternative, if the unique chain feature is not selected analysis is carried out on all protein chains of PDB file. Once the selection of the chains is complete, the occurrence of the motif based on the pattern is searched. On identification of the pattern, conformational parameters are calculated for residues in the motif. DSSP is then used to identify the secondary structure assignment. Next is the calculation of solvent accessibility of the motif residues using NACCESS. This is followed by placing motif residues in Ramachandran plot regions by Procheck. Once this structural analysis is complete, iMotifs then turns to identification of disulphide bonds and non-covalent interactions involving the motif. In-house developed scripts were used to calculate ionic, aromatic-aromatic, aromatic-sulphur, cation-pi and hydrophobic interactions. Disulphide bonds were calculated using SSBOND while hydrogen bond calculation was carried out with HBPLUS.

The results obtained from the analysis are first recorded in individual files generated for each identified motif. The synopsis of these results is then presented to the users as a formatted webpage. Users can also download the results in the form of a zipped file.

JSmol is used to visualize the identified motifs (Gezelter, et al., 2013). The identified motifs in the summary are hyperlinked to JSmol web application. JSmol uses a HTML5 JavaScript. It is used in conjunction with the Java applet to provide an alternative to Java when the platform does not support it or does not support applets. On clicking the link, the JSmol is initiated, wherein the protein structure containing the motif is first displayed after which focus shifts to highlight the detected motif.

2.3 Description and Validation

iMotifs provides a comparative analysis of the motifs detected in protein structures. The detailed file generated for each motif provides a comprehensive analysis (Figure 5) while the summary represents the comparative analysis of all motifs (Figure 6). The unique feature of iMotifs is the interaction profile provided in the summary and the extensive details of covalent and non-covalent interactions involving the motif residues

a

Chain	Res.No	Res.ID	Phi	Psi	Chi1	Chi2	Chi3	Chi4	Chi5
A	164	D	-119.77	23.08	53.84	17.99	NA	NA	NA
A	165	S	-86.57	157.64	-74.29	NA	NA	NA	NA
A	166	R	-136.66	162.31	-54.36	171.87	-179.97	-176.38	NA
A	167	I	-120.37	124.34	-61.54	-175.78	NA	NA	NA

b

Acidic			Basic			ASA		SS	
Chain	Res.No	Res.ID	Chain	Res.No	Res.ID	Acidic	Basic	Acidic	Basic
B	544	GLU	A	8	LYS	42.6	28.5	H	E
A	38	ASP	A	11	ARG	7.7	15.1	H	E
B	551	GLU	A	11	ARG	35.8	15.1	E	E
A	12	ASP	A	18	HIS	0.4	0	E	E
A	23	ASP	A	26	HIS	39	0.6	S	H

c

Donor				Acceptor				Bond_type	Distance_DA	Distance_HA	DHA_Angle	HAAA_Angle	DAAA_Angle
Chain	Res.No	Res.ID	Atom	Chain	Res.No	Res.ID	Atom						
A	22	ASN	ND2	A	3	SER	O	SM	2.8	2.08	127.5	146.9	137.4
A	3	SER	OG	A	5	SER	OG	SS	3.41	2.42	175.2	111.1	111.7
A	5	SER	OG	A	3	SER	OG	SS	3.41	2.41	177.5	106.9	106.6
B	555	VAL	N	A	5	SER	O	MM	2.81	1.82	167	138.8	139.3
B	554	HIS	NE2	A	6	GLU	OE1	SS	3.12	2.35	132.8	147.6	139.1

d

N.Res	N.Int	Buried	Exposed	Network_Details
3	2	0	3	B544GLU-A8LYS B544GLU-B547LYS
3	2	0	3	A68ASP-A64LYS A68ASP-A71ARG
3	2	0	3	A76ASP-A79ARG A130GLU-A79ARG

e

Cys1		Cys2		ASA		Secstr	
Chain	Res1_no	Chain	Res2_no	Cys1	Cys2	Cys1	Cys2
B	492	B	525	3.6	24.2	E	C

Figure 5. Some components of the detailed file generated for each detected motif. (a) The main chain and side chain dihedrals for the motif residues. (b) Details of ionic interactions detected. (c) Details of hydrogen bonds calculated. (d) Details of the ionic networks detected. (e) Details of disulphide bonds detected.

Filename	Chain	Start	Motif	Acc	SS	R.plot_region	Interaction profile										
							IP	IP.Net	AP	AP.Net	AS	AS.Net	HB	Disul	Cat-pi	Cat-pi.Net	Hphob
1pnk	A	68	DIR	EBE	HHH	A,A,A	5	2	0	0	0	0	10	0	2	1	3
1pnk	B	204	DPR	EBE	CTT	B,-,A	1	0	0	0	0	0	8	0	0	0	1
1pnk	B	314	DPR	BEE	CHH	B,-,A	2	1	0	0	0	0	7	0	1	0	1
3k3w	A	90	DER	EEE	HHH	A,A,A	2	1	0	0	0	0	4	0	1	1	0
3k3w	B	204	DIR	EBB	CTT	B,A,A	3	1	0	0	0	0	6	0	1	1	1
3k3w	B	534	DIR	EBE	HHH	A,A,A	1	0	0	0	0	0	8	0	0	0	2

Figure 6. The summary file for the D.R pattern analyzed by iMotifs. Acc: Solvent Accessibility. SS: Secondary Structure. R.plot_region: Ramachandran Plot Region. IP: Ionic Pairs. IP.Net: Ionic Pair Interaction Network. AP: Aromatic Pairs. AP.Net: Aromatic Pair Interaction Network. AS: Aromatic-Sulphur Pairs. AS.Net: Aromatic-Sulphur Pair Interaction Network. HB: Hydrogen Bonds. Disul: Disulphide Bonds. Cat-pi: Cation - π Interaction. Cat-pi.Net: Cation - π Interaction Network. Hphob: Hydrophobic Interaction.

2.3.1 Description of parameter-analysis for the detected motifs by iMotifs.

iMotifs has been designed to carry out a comparative analysis of motifs conforming with the search pattern in protein structures in terms of structural features and interactions.

1. ***Secondary structure information*:** iMotifs provides the details of secondary structure assignment for each residue in each motif detected. The notations used by DSSP for secondary structure assignment are:

    - H = alpha helix
    - B = residue in isolated beta-bridge
    - E = extended strand, participates in beta ladder
    - G = 3-helix (3/10 helix)
    - I = 5-helix (pi helix)
    - T = hydrogen bonded turn
    - S = bend
    - C = random coil

2. ***Conformational parameters*:** The detailed file for each detected motif gives a report of conformational features such as Chain, Residue number, Residue ID, main chain and side chain dihedral angles of the residues in the motif.
3. ***Solvent accessibility*:** Solvent accessibility of residues is calculated using NACCESS. The notations in the summary are based on the relative solvent accessibility values. If the value is less than 20, the residue is considered buried, otherwise it is considered exposed.
4. ***Assignment of Ramachandran plot regions*:** The residues of motifs are assigned locations in Ramachandran plot using Procheck. The different regions on the Ramachandran plot are as described in Morris et al. (Morris, et al., 1992). The regions are labeled as follows:

    A: Core alpha

    L: Core left-handed alpha

    B: Core beta

    a: Allowed alpha

    l: Allowed left-handed alpha

~a: Generous alpha

~l: Generous left-handed alpha

p: Allowed epsilon

b: Allowed beta

~p: Generous epsilon

~b: Generous beta

5. ***The Interaction Profile:*** The interaction profile generated is an unique feature of iMotifs tools. It analyzes the motif residues for their involvement in non-covalent interactions as well as disulphide bonds.

a. Ionic interaction and networks: An ion pair consists of a positive and negative ion temporarily bonded together by the electrostatic force of attraction between them. In proteins, ion pairs are electrostatic interactions between the nitrogen atoms of basic residues (Arg/Lys/His) and the carboxylate oxygen atoms of acidic residues (Asp/Glu). A salt bridge is considered to be formed if the distance between any of the oxygen atom of acidic residues and the nitrogen atom of basic residues are within the cut-off distance. Ion pairs play important roles in protein structure and function. The ion pairs form large networks that criss-cross the protein surface and the subunit interfaces. Ion pair networks are energetically more favorable than an equivalent number of isolated ion pairs, because for each new pair the burial cost is cut in half: only one additional residue must be desolvated and immobilized. The default cut-off distance is set at 6Å (Perutz, 1978; Yip, et al., 1995).

The summary gives the number of ion-pairs the motif residues form and number of ion-pair networks in which the motif residues are involved. In the detailed file, the first table shows the details of ion pairs segregated by their nature as acidic and basic along-with their individual relative solvent accessibility and their secondary structure. The next table gives the details of each network observed as the number of residues involved, the number of interaction formed, the distribution of the residues by their solvent accessibility as buried and exposed, and finally the residues involved in the interaction network.

b. Aromatic-aromatic interactions and networks: Aromatic residues (F,W,Y) are considered to interact with each other if the distance between their phenyl

ring centroids lies between 4.5 Å & 7.0 Å. A cut-off of dihedral angle between the planes of such interacting aromatic residues can be set between 30° and 90°. The default centroid distance is set at minimum 4.5 Å and max 7.0 Å (Burley and Petsko, 1985).

The summary gives the number of aromatic-aromatic pairs the motif residues form and the number of aromatic-aromatic networks in which the motif residues are involved. The detailed file contains two tables. The first table lists the details of aromatic pairs along-with their Centroid Distance, Dihedral Angle, individual relative solvent accessibility and their secondary structure. The next table gives the details of each network observed as the number of residues involved, the number of interactions formed, the distribution of the residues by their solvent accessibility as buried and exposed and finally the residues involved in the observed interactions.

c. ***Aromatic-sulphur interactions and networks:*** A sulphur atom can interact with an aromatic ring through a S-pi interaction. Interactions between the sulphur atoms of cysteine and methionine and the aromatic rings of phenylalanine, tyrosine and tryptophan within 5.3Å account for aromatic-sulphur interactions. The default cut-off distance is set at 5.3 Å (Reid, et al., 1985).

The summary gives the number of aromatic-sulphur pairs the motif residues form and number of aromatic-sulphur networks in which the motif residues are involved. The detailed file contains two tables. The first table shows the details of aromatic residue and the sulphur contributing residue in the pair along-with their distance, individual relative solvent accessibility and their secondary structure. The next table gives the details of each network observed as the number of residues involved, the number of interaction formed, the distribution of the residues by their solvent accessibility as buried and exposed and finally the residues involved in the interaction networks.

d. ***Hydrogen bonds:*** A hydrogen bond (or H-bond) is an attractive interaction between two electronegative atoms, a donor and an acceptor. A hydrogen atom lies aligned between them and covalently bound to the donor. The donor attracts the electron on the hydrogen from its orbital towards the donor itself. This leaves a partial positive charge on the hydrogen, which is electrostatically attracted towards the electronegative acceptor (Figure 7). The interaction is energetically

favorable in a number of ways, including polarization energy and covalent energy, but particularly the electrostatic energy. H-bonds are typically defined by a distance of less than 3 Å between the H donor and the H acceptor and by donor-hydrogen-acceptor angle below 90° (Baker and Hubbard, 1984; Huggins, 1971; Ippolito, et al., 1990; Latimer and Rodebush, 1920; McDonald and Thornton, 1994; Stickle, et al., 1992). The hydrogen bonds have been calculated using the HBPLUS software. A charged-neutral hydrogen bond (CNHB) has a side chain atom of a charged residue (D/E/R/K/H) as one partner, while the other partner is either a main chain atom of any residue or a side chain atom of a neutral residue. A neutral-neutral hydrogen bond (NNHB) is defined as hydrogen bond between a side chain atom of a neutral residue and either a main chain atom of any residue or a side chain atom of another neutral residue.

Legend:

- −: Covalent Bond
- H: Hydrogen
- [DD1, DD2]: Donor Antecedents
- : : Hydrogen Bond
- D: Donor
- [AA1, AA2] Acceptor Antecedents
- A: Acceptor

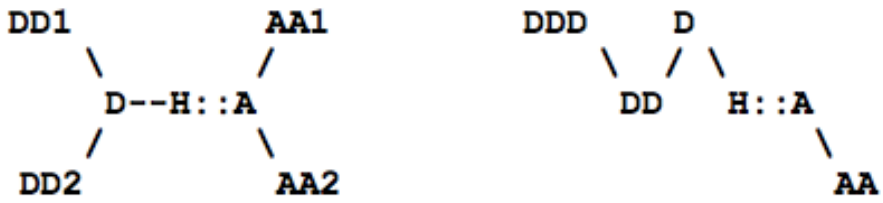

Figure 7. The hydrogen bonds definition used by HBPLUS.

Criteria:

Maximum Distances (D-A of 3.9 Angstroms, H-A of 2.5 Angstroms)

Minimum Angles (D-H-A of 90.0 degrees, D-A-AA of 90.0 degrees, H-A-AA of 90.0 degrees

Maximum Angles (D-A-AX of 20.0 degrees, H-A-AX of 20.0 degrees for amino-aromatic interactions (AX is at L to aromatic plane)

The summary gives to total number of hydrogen bonds calculated by HBPLUS involving the motif residues. The table in the detailed file displays the details of each hydrogen bond classified as donor and acceptor. The next column classifies the hydrogen bonds as main chain - main chain, main chain - side chain, side chain - main chain and side chain - side chain. The table then gives the D - A distance and the H - A distance. Next is the D - H - A angle, H - A - AA angle and the D - A - AA angle.

e. Disulphide bonds: Disulphide bonds or bridges stabilize proteins mostly through an entropic effect, by reducing the entropy of the unfolded state of the protein. Pairs of cysteines (sulphur atoms) within 2.2Å are considered for disulphide bridges. The entropy is known to increase in proportion to the logarithm of the number of residues separating the two cysteines forming the disulphide bridge (Matsumura, et al., 1989).

The summary gives the total number of disulphide bridges identified by SSBOND involving the motif residues. The table displays the details of the cysteines forming the disulphide bond (cystines), their individual relative solvent accessibility and the secondary structure.

f. Cation-pi interactions and networks: The cation-pi interactions are observed between side chains carrying positive charge such as Arg, Lys or side chains carrying partial charge such as Asn or Gln and aromatic rings of Phe, Tyr, Trp, His within 6 Å separation. The default cut-off distance is set at 6 Å (Sathyapriya and Vishveshwara, 2004).

The summary gives the number of cation-pi pairs the motif residues form and number of cation-pi networks in which the motif residues are involved. In the detailed file the first table shows the details of cation-pi pairs as individual cationic and aromatic residue, their individual relative solvent accessibility and secondary structure. The next table gives the details of each network observed in terms of the number of residues involved, the number of interaction formed, the distribution of the residues by their solvent accessibility as buried and exposed and finally the residues involved in the interaction networks.

g. ***Hydrophobic interactions:*** Hydrophobic interactions are widely considered to be important for protein structure, aggregation, and function. An average increase in stability of 1.3 (+/- 0.5) kcal/mol was calculated for each additional methyl group buried in protein folding. The residues (Ala, Val, Leu, Ile, Met, Phe, Trp, Pro, Tyr) are considered to participate in interactions if they fall within the default distance of 5 Å range (Hummer, et al., 1996; Kyte and Doolittle, 1982; Pace, 1992).

Summary contains the total number of hydrophobic interactions encountered. The table in the detailed file displays residues in the hydrophobic interaction pairs along with their individual secondary structure and relative solvent accessibility.

2.3.2 Visualization of motifs detected by iMotifs.

iMotifs not only offers identification and analysis of sequence-based motifs but also allows the user to visualize the motifs in three-dimensional structure (Figure 8). For this the JSmol web application has been implemented. JSmol is the extension of the Java-based molecular visualization applet Jmol. Jmol is a free, open source molecule viewer for students, educators, and researchers in chemistry and biochemistry. JSmol seamlessly offers alternatives to Java on these non-Applet platforms. JSmol is an interactive JavaScript framework that allows web developers to create pages that utilize either Java or HTML5 (no Java), at will. This enables Jmol to display interactive 3D molecular structures. The summary presented to each user is designed such that the user can view the identified motif. Each motif in the summary is designed as a hyperlink. On initiation, first the protein containing the motif is displayed in cartoon representation. The motif is displayed in ball and stick representation and the molecule is moved to zoom upon the motif. Finally, the motif residues are labelled at user's ease. Once the applet is loaded, on clicking the next identified motif, the new PDB is directly loaded into JSmol and new motif is displayed without reloading of the applet.

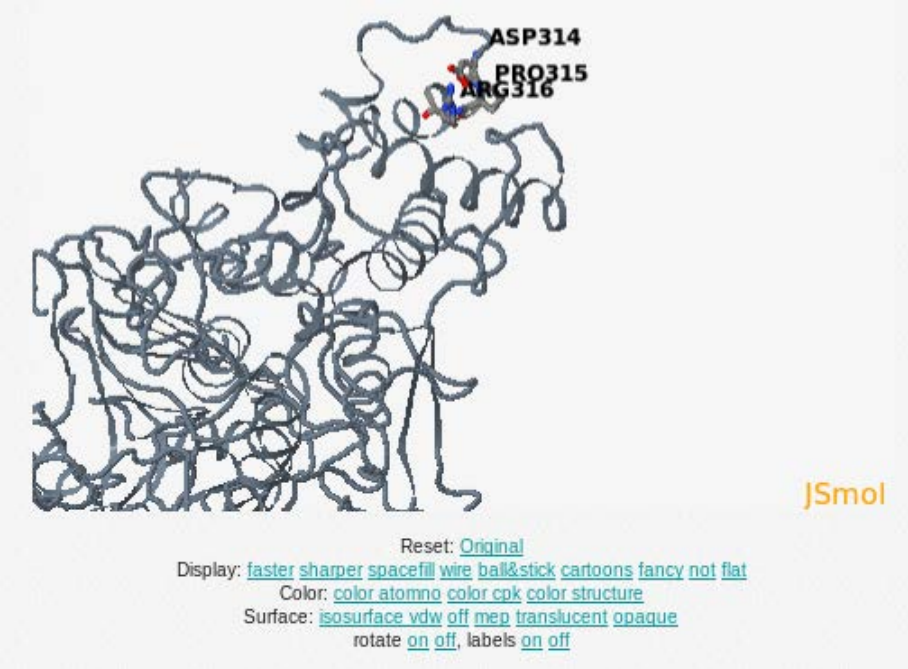

Figure 8. The DPR motif displayed in JSmol by iMotifs.

2.3.3 Structural Analysis of known motifs from Prosite with iMotifs tool.

Of the 1309 patterns currently present in the Prosite database, a total of 1132 patterns were analysed in the PDB database. The motifs were analysed for various structural parameters. In terms of length of the pattern searched the shortest motif contained four residues while the largest had 105 residues. The entire dataset was then classified by the occurrences of motifs in particular secondary structures. The secondary structure states were considered for the analysis viz. helices (H), Sheets (S) and Irregular structural Regions (IR). The IR state comprised of coiled structures, loops as well as hydrogen-bonded turns. Based on the secondary structure assignment by DSSP, the secondary structure composition of each motif was determined and converted to the percentage for all motifs in each pattern.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{Number\ of\ residues\ belonging\ to\ a\ particular\ SS\ state}{Motif\ length}$$

where:

SS state: Helices (H), Sheets (S) and Irregular structural Regions (IR)

Using a percentage cut-off of 50%, patterns with more than 50% helical content were identified. These were assigned to the H state. Similarly patterns with

more than 50% sheet content were identified and assigned to S state. Finally patterns with more than 50% secondary structure belonging to irregular structural regions were assigned to IR state.

From the 1132 patterns, 927 could be classified into the three states. While 3 patterns (PS00267, PS00465, PS00747) were found to lie at the interface of H and IR, 6 (PS00129, PS00726, PS00803, PS00159, PS01268, PS01436) were found at the S-IR interface. The remaining 205 patterns were found to have a combination of all the three states (Figure 9). In these patterns the longer length of the motifs allowed for the combination of all the three states.

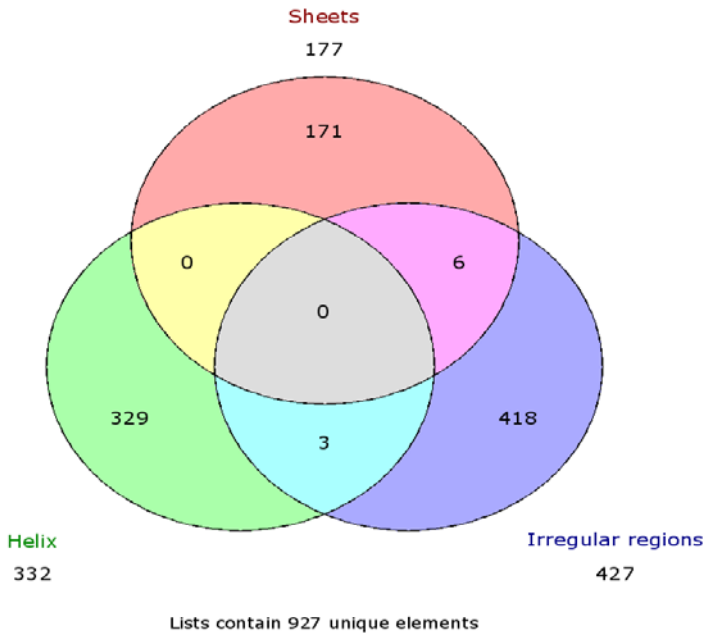

Figure 9. The Venn diagram describing the distribution of patterns in the three defined secondary structure states.

The interaction profile generated for the patterns were then analysed in terms of the 6 non-covalent interactions as well as the presence of disulphide bonds. First the presence of the 7 interactions was estimated in each detected motif of each pattern. Using the formula given below an interaction score (IntScore) was calculated.

$$IntScore = \frac{1}{N}\sum_{i=1}^{N}\frac{IP + AP + AS + CP + HB + HP + DB}{Motif\ Length}$$

where:

N: number of motifs identified for each Prosite pattern.

IP: number of Ionic interaction calculated for each motif in a pattern.

AP: number of Aromatic pair interaction calculated for each motif in a pattern.

AS: number of Aromatic-sulphur interaction calculated for each motif in a pattern.

CP: number of Cation-π interaction calculated for each motif in a pattern.

HB: number of Hydrogen bonds calculated for each motif in a pattern.

HP: number of Hydrophobic interactions calculated for each motif in a pattern.

DB: number of Disulphide bond interaction calculated for each motif in a pattern.

An example is the pattern PS00010 that has been described as Aspartic acid and asparagine hydroxylation site. The consensus sequence is C-x-[DN]-x(4)-[FY]-x-C-x-C. It is the site for post-translational hydroxylation of Asn or Asp, forming erythro-β-hydroxyasparagine or erythro-β-hydroxyaspartic acid. The site has been identified at the N-terminal of proteins having domains homologous to epidermal growth factor (EGF) (Stenflo, et al., 1988). The pattern was identified in 98 protein structures. 130 motifs were identified from the structures. The pattern was found to be involved in forming sheets. The pattern was found to have an interaction score of 1.604. Using a 25% sequence identity cut-off, four structures with unique sequences were identified. The sequence alignment shown below (Figure 10) shows the conservation of Cys residues surrounding the pattern as well as the observed disulphide bonds. In all cases disulphide bonds involving the Cys residues in the pattern were found to be conserved.

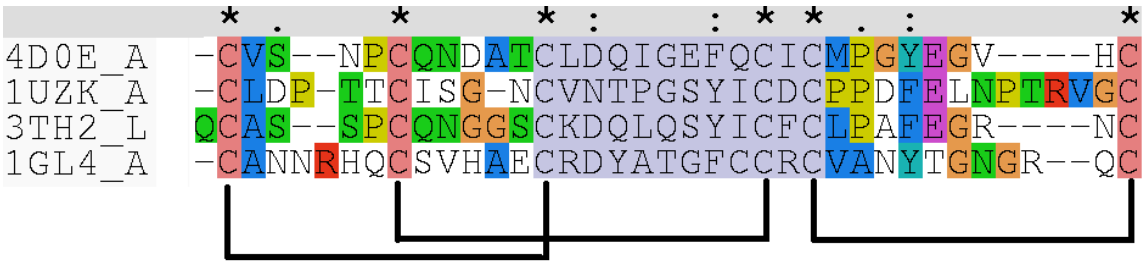

Figure 10. The sequence alignment of the patterns and surrounding residues in the unique structures analyzed. The detected motifs have been displayed in grey.

2.4 Summary

Although a large number of analysis tools for protein motifs are existent, very few focus on a wide range of parameters as iMotifs does based on protein structures. The web platform iRDP hosts the iMotifs analysis tool and iMotifs interaction database. The iMotifs tools aims to identify and analyse sequence defined patterns in protein structures followed by analysis of structural features such as secondary structure and solvent accessibility. Lastly, the interaction profile, a unique feature of iMotifs, is assembled from the 7 types of molecular interactions computed involving the motif residues. Another unique feature of iMotifs tool is its ability to analyse 100 protein structures simultaneously and the comparative representation of the results in the summary. iMotifs thus allows users to study sequence motifs not just as part of the protein primary sequence but also can be extended to the analysis of their structural features. The iMotifs interactions database explores 1132 known sequence patterns from the Prosite. It records the detailed analysis of each pattern carried out using iMotifs analysis tool.

2.5 References

Attwood, T.K., *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucleic acids research* 2003;31(1):400-402.

Bailey, T.L., *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 2009:gkp335.

Baker, E.N. and Hubbard, R.E. Hydrogen bonding in globular proteins. *Progress in biophysics and molecular biology* 1984;44(2):97-179.

Bennett, S.P., Nevill-Manning, C.G. and Brutlag, D.L. 3MOTIF: visualizing conserved protein sequence motifs in the protein structure database. *Bioinformatics* 2003;19(4):541-542.

Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Research* 2000;28(1):235-242.

Bhaduri, A., *et al.* iMOT: an interactive package for the selection of spatially interacting motifs. *Nucleic acids research* 2004;32(suppl 2):W602-W605.

Bhaduri, A., Pugalenthi, G. and Sowdhamini, R. PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC bioinformatics* 2004;5(1):35.

Burley, S. and Petsko, G. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* 1985;229(4708):23-28.

Chakrabarti, S., *et al.* SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs. *Nucleic acids research* 2005;33(suppl 2):W274-W276.

Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Research* 2014:gku989.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* 2006;1695(5).

Davey, N.E., *et al.* SLiMSearch: a webserver for finding novel occurrences of short linear motifs in proteins, incorporating sequence context. In, *Pattern Recognition in Bioinformatics*. Springer; 2010. p. 50-61.

De Castro, E., *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research* 2006;34(suppl 2):W362-W365.

Dinkel, H., *et al.* The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic acids research* 2013:gkt1047.

Edwards, R.J., Davey, N.E. and Shields, D.C. CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics* 2008;24(10):1307-1309.

Gaulton, A. and Attwood, T.K. Motif3D: Relating protein sequence motifs to 3D structure. *Nucleic acids research* 2003;31(13):3333-3336.

Gezelter, D., Smith, B. and Willighagen, E. Jmol: an open-source Java viewer for chemical structures in 3D. In.; 2013.

Grant, B.J., *et al.* Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 2006;22(21):2695-2696.

Gutman, R., *et al.* QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic acids research* 2005;33(suppl 2):W255-W261.

Hazes, B. and Dijkstra, B.W. Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein engineering* 1988;2(2):119-125.

Huang, J.Y. and Brutlag, D.L. The EMOTIF database. *Nucleic acids research* 2001;29(1):202-204.

Hubbard, S.J. and Thornton, J.M. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 1993;2(1).

Huggins, M.L. 50 Years of hydrogen bond theory. *Angewandte Chemie International Edition in English* 1971;10(3):147-152.

Hummer, G., *et al.* An information theory model of hydrophobic interactions. *Proceedings of the National Academy of Sciences* 1996;93(17):8951-8955.

Ippolito, J.A., Alexander, R.S. and Christianson, D.W. Hydrogen bond stereochemistry in protein structure and function. *Journal of molecular biology* 1990;215(3):457-471.

Jonassen, I. Efficient discovery of conserved patterns using a pattern graph. *Computer applications in the biosciences: CABIOS* 1997;13(5):509-522.

Jonassen, I., Collins, J.F. and Higgins, D.G. Finding flexible patterns in unaligned protein sequences. *Protein science* 1995;4(8):1587-1595.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577-2637.

Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* 1982;157(1):105-132.

Laskowski, R.A., *et al.* PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography* 1993;26(2):283-291.

Latimer, W.M. and Rodebush, W.H. polarity and ionization from the standpoint of the lewis theory of valence. *Journal of the American Chemical Society* 1920;42(7):1419-1433.

Matsumura, M., Signor, G. and Matthews, B.W. Substantial increase of protein stability by multiple disulphide bonds. *Nature* 1989;342(6247):291-293.

McDonald, I.K. and Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* 1994;238(5):777-793.

Morris, A.L., *et al.* Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics* 1992;12(4):345-364.

Pace, C.N. Contribution of the hydrophobic effect to globular protein stability. *Journal of molecular biology* 1992;226(1):29-35.

Perutz, M. Electrostatic effects in proteins. *Science* 1978;201(4362):1187-1191.

Pugalenthi, G., Bhaduri, A. and Sowdhamini, R. iMOTdb—a comprehensive collection of spatially interacting motifs in proteins. *Nucleic acids research* 2006;34(suppl 1):D285-D286.

Reid, K.S.C., Lindley, P.F. and Thornton, J.M. Sulphur-aromatic interactions in proteins. *FEBS Letters* 1985;190(2):209-213.

Rice, P., Longden, I. and Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in genetics* 2000;16(6):276-277.

Sathyapriya, R. and Vishveshwara, S. Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res* 2004;32(14):4109-4118.

Sigrist, C.J., *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics* 2002;3(3):265-274.

Sigrist, C.J., *et al.* New and continuing developments at PROSITE. *Nucleic acids research* 2012:gks1067.

Sigrist, C.J., *et al.* ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 2005;21(21):4060-4066.

Stenflo, J., *et al.* beta-Hydroxyaspartic acid or beta-hydroxyasparagine in bovine low density lipoprotein receptor and in bovine thrombomodulin. *Journal of Biological Chemistry* 1988;263(1):21-24.

Stickle, D.F., *et al.* Hydrogen bonding in globular proteins. *Journal of molecular biology* 1992;226(4):1143-1159.

Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 2000;28(6):1102, 1104-1102, 1104.

Yan, T., *et al.* PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic acids research* 2005;33(suppl 2):W262-W266.

Yip, K.S., *et al.* The structure of Pyrococcus furiosus glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure (London, England : 1993)* 1995;3(11):1147-1158.

Zhang, Z., *et al.* Protein sequence similarity searches using patterns as seeds. *Nucleic acids research* 1998;26(17):3986-3990.

Chapter 3

Identification of D-R, R-D, D-X-R, R-X-D, D-K, K-D, D-X-K and K-X-D motifs, analysis of their conformations and hydrogen bonded interactions.

This chapter focuses on the structural analysis of short sequence motifs containing charged amino acids, Aspartic acid and Arginine or Lysine as immediate neighbours along the sequence. The detected motifs were analyzed for their secondary structure preference, the localized fold and the patterns of hydrogen bonding interactions between the two oppositely charged residues. Any motifs with specific patterns of interactions and local conformations that repeat across several unrelated structures have been analyzed in more detail. The analysis was carried out in the order in which they appear in secondary structures such as helices and sheets and in the irregular structures. When the oppositely charged amino acids occur in regular secondary structures ionic interactions or hydrogen bonds or salt bridges involving their side chains provide additional stability to the structure. However, when they occur in irregular regions they adapt specific local fold with specific backbone and side chain conformations. Our effort was to identify and highlight such local fold or short structural motif.

3.1 Structural Analysis of Short Sequence Motifs containing Asp and Arg/Lys.

Motifs wherein Asp was present with either Arg or Lys as sequence neighboring amino acids or separated by a single spacer residue were studied. Total of eight motifs were analyzed for their secondary structure preferences and the presence of patterns of hydrogen-bonded interactions amongst the residues. Backbone conformation of the motif sequence and side chain conformational preferences of the oppositely charged sequence neighbouring residues engaged in interactions have been analyzed.

3.1.1 Analysis of sequence motifs containing Asp (D) and Arg (R).

The first set of sequence motifs analyzed comprised of four patterns namely D-R, R-D, D-X-R and R-X-D (Table 1, 2). A search of the selected structural database of unique proteins provided 9351 occurrences of the D-R motif. A major part of the identified D-R motifs had irregular structures. Among the 879 motifs with interactions only 128 had two or more hydrogen bonds. Sequences with reverse order of residues (R-D motifs) had 10094 occurrences. 4358 motifs were identified in the helix region. 960 of them were with interaction, which comprise of 505 occurrences having two or more hydrogen bonded interactions. Although a majority of the motifs

were found to occur as irregular structures 372 of them were identified with interactions, out of which 93 were observed with two or more H-bonds.

The sequence with a spacer residue (indicated as X) between Asp and Arg (D-X-R) displayed 9407 occurrences. Majority of the identified D-X-R motifs were having irregular structures while significantly fewer motifs were observed to occur in regular secondary structures of helices and sheets. 3331 motifs with irregular structures had hydrogen-bonded interactions; a significant number i.e 1843 motifs were detected to have two or more H-bonds. In the reversed sequence where Arg and Asp interchanged places (R-X-D), lesser number of motifs i.e. 8649 was observed. The almost equitable preference of the R-D for helices and irregular structures was shifted to more of irregular structures in the case of R-X-D motif. Of all the motifs in the irregular structures with interactions, 930 were found to involve two or more H-bonds whereas 180 were observed in sheets and 5 in helices with two or more H-bonds.

3.1.2 Analysis of motifs containing Asp (D) and Lys (K) as sequence neighbors.

This set of motifs analyzed comprised of four patterns namely D-K, K-D, D-X-K and K-X-D (Table 1, 2). A total 10874 occurrences of the D-K motif were detected. Table 1 gives the details of this motif. While a majority of the motifs were found to occur as irregular structures without interactions, 801 were observed to be involved in hydrogen-bonded interactions amongst the charged motif residues. Out of these 801 only 51 have two hydrogen bonds whereas the remaining motifs have only one. For the reversed sequence order of the motif residues, the search gave 12011 occurrences of the K-D motif. For this set there was a significant increase in the occurrence of the motif in helices. Out of the 724 occurrences of the motif with interactions, only 5 have two or more H-bonds.

There are 11068 occurrences of motif with a residue separating the reference amino acids Asp and Lys (D-X-K) as shown in Table 1. A comparatively large number of occurrences of the motif with interactions were observed among irregular structures. However, 2538 occurrences from the irregular structures showed only one interaction. Of the total 2891 occurrences with interactions, a mere 211 showed 2 or more H-bonds. With a spacer residue between Lys and Asp (K-X-D) the number of occurrences present was 9656 and the secondary structure preference shifted from

helix to irregular structure. While a total of 523 motifs with interactions were observed, only 21 were found to involve two or more H-bonds.

Table 1. The number of occurrences of the different types of motifs such as D-R, R-D, D-X-R, R-X-D, D-K, K-D, D-X-K, K-X-D present in the selected local PDB dataset of 12,872 unique (less than 25% sequence identity) proteins. Numbers in brackets indicate the total occurrence of the particular motif in the dataset. Numbers highlighted in red, repeated significantly across structures, have been studied in detail for interactions and conformations.

Motif (Total)	Helix		Sheets		Irregular regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
D-R (9351)	198	3048	97	334	584	5008
Total	3246		431		5592	
R-D (10094)	960	3398	33	639	372	4690
Total	4358		672		5062	
D-X-R (9407)	65	1857	314	519	3710	2900
Total	1922		833		6610	
R-X-D (8649)	45	1777	299	485	1529	4514
Total	1822		784		6043	
D-K (10874)	203	3623	60	391	538	6058
Total	3826		441		6596	
K-D (12011)	251	3969	27	646	446	6672
Total	4220		673		7118	
D-X-K (11068)	31	2117	114	662	2746	5398
Total	2148		776		8144	
K-X-D (9656)	16	2000	61	674	446	6486
Total	2016		735		6905	

Motifs involving Asp and both Arg as well as Lys show more preference for irregular structures and least preference for β-sheets. However, motifs with interactions were more amongst motifs of Asp and Arg. The Asp side chain with carboxylate group and the long side chain of Arg with the guanidinium group present allow for better interactions amongst the residues in the motif, thereby stabilizing the

local folding of the backbone. Based on the above results motifs with two or more interaction were studied in motifs involving Asp and Arg while motifs with single interactions were studied in motifs with Asp and Lys.

Table 2. The number of occurrences with interactions in the three secondary structure classes of D-R, R-D, D-X-R, R-X-D, D-K, K-D, D-X-K, K-X-D motifs present in the local PDB dataset. Numbers highlighted in red have been studied further.

Motif	Helix		Sheet		Irregular Structural Regions		Total with ints.
	H-bonds (ints.)		H-bonds (ints.)		H-bonds (ints.)		
	1	>1	1	>1	1	>1	
D-R	165	33	88	9	498	**86**	879
Total(ints.)	198		97		584		
R-D	455	**505**	30	3	279	**93**	1365
Total(ints.)	960		33		372		
D-X-R	65	0	159	**155**	1488	**2222**	4089
Total(ints.)	65		314		3710		
R-X-D	40	5	119	**180**	599	**930**	1873
Total(ints.)	45		299		1529		
D-K	**181**	22	59	1	**509**	29	801
Total(ints.)	203		60		538		
K-D	**251**	0	25	2	**443**	3	724
Total(ints.)	251		27		446		
D-X-K	31	0	111	3	**2538**	**208**	2891
Total(ints.)	31		114		2746		
K-X-D	16	0	58	3	**400**	46	523
Total(ints.)	16		61		446		

3.2 Detailed analysis of sequence motifs containing neighbouring Asp and Arg/Lys.

Extensive analysis of the 8 motifs mentioned above was carried out. The analysis focused on the local fold, side chain conformations of Asp and Arg/Lys and

the hydrogen bonding interaction involving the charged reference residues. It is observed that when the side chains are involved in more than one interaction then the motif tend to assume particular conformation either as part of a known secondary structure or when present in irregular structure it can be considered as a short structural motif. In secondary structures the interaction between Arg/Lys and Asp stabilizes the secondary structure whereas in the irregular structure the interaction stabilizes the short stretch of sequence as a structural motif. We have tried to characterize such motifs occurring in comparatively large numbers as part of alpha-helix, beta-sheet or irregular structure which are highlighted in tables 1 and 2.

3.2.1 Analysis of motifs in helices.

The eight sequence motifs identified above were studied for their occurrences in helices (Table 3). In case of D-R, out of the 198 with interactions, 165 involved single interaction while 33 showed two or more interactions. For the R-D motifs, out of the 960 occurrences with interactions, 455 showed single hydrogen bonds whereas 505 showed two or more. For the D-X-R motifs only 65 occurrences were identified with interaction while for the R-X-D motifs merely 45 were observed. In the D-K motif 203 occurrences were observed with interaction of which 181 showed a single H-bond. The reverse K-D sequence motif showed 251 with one hydrogen bond.

Table 3. The number of occurrences with 2 and 3 interactions for the three secondary structure classes of D-R, R-D, D-X-R, R-X-D motifs present in the local PDB dataset.

Motif	Helix		Sheet		Irregular Structural Regions		Total with ints.
	H-bonds (ints.)		H-bonds (ints.)		H-bonds (ints.)		
	2	3	2	3	2	3	
D-R	33	0	9	0	**78**	8	128
Total	33		9		86		
R-D	**503**	2	3	0	**90**	3	601
Total	505		3		93		
D-X-R	0	0	**143**	12	**505**	**1717**	2377
Total	0		155		2222		
R-X-D	3	1	**180**	0	**743**	**187**	1114
Total	4		180		930		

Table 4: The number of occurrences with 1 and 2 interactions for the three secondary structure classes of D-K, K-D, D-X-K, K-X-D motifs present in the local PDB dataset.

Motif	Helix		Sheet		Irregular		Total
(Total)	H-bonds (ints.)		H-bonds (ints.)		H-bonds (ints.)		with
	1	2	1	2	1	2	ints.
D-K	**181**	22	59	1	**508**	29	790
Total	203		60		527		
K-D	**251**	0	25	2	**443**	3	724
Total	251		27		446		
D-X-K	31	0	**111**	3	**2538**	**208**	2891
Total	31		114		2746		
K-X-D	16	0	58	3	400	19	496
Total	16		61		419		

3.2.1.1 Analysis of R-D motif with two hydrogen bonds in helices.

Among the 4 motifs involving Asp and Arg present in helices R-D is the only one having more than one interaction showing significant number of occurrences. As can be expected the pairwise superimposition analysis of the motifs belonging to the helix group showed all the motifs to have similar backbone conformation. The Cβ - Cβ angle distribution for motifs of this set was found to lie in the range of 50-70° (Figure 1). Based on analysis of the Arg side chain conformations, the partially extended conformation t t g+ t was found to be the most favorable (Figure 2).

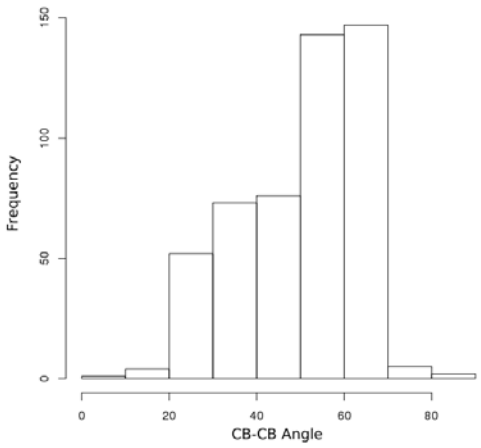

Figure 1. Histogram showing the distribution of the C_β-C_β orientation angle of the motifs.

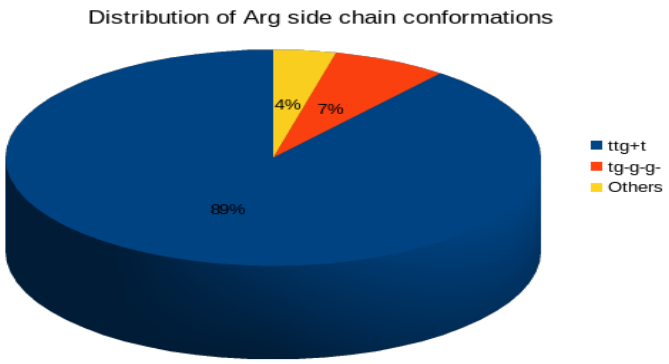

Figure 2. The distribution of Arg side chain conformation in R-D motif with two interactions in helices.

The Asp χ_1 was found to be exclusively g-.Of the total 449 with Arg t t g+ t conformation, 264 were observed to have Type D (Figure 3a) and 185 to have Type B (Figure 3b) hydrogen bonding interaction. The second conformation observed was the Arg t g- g- g- conformation. Even in this case the Asp χ_1 was found to be exclusively g-. All motifs belonging to this set were found to have only Type D interactions (Figure 3c). The above analysis highlighted the fact that with both extended and folded conformation states of the Arg side chain the Asp side chain χ_1 angle assumed a g- conformation in order to have Type D or Type B interactions between the residues.

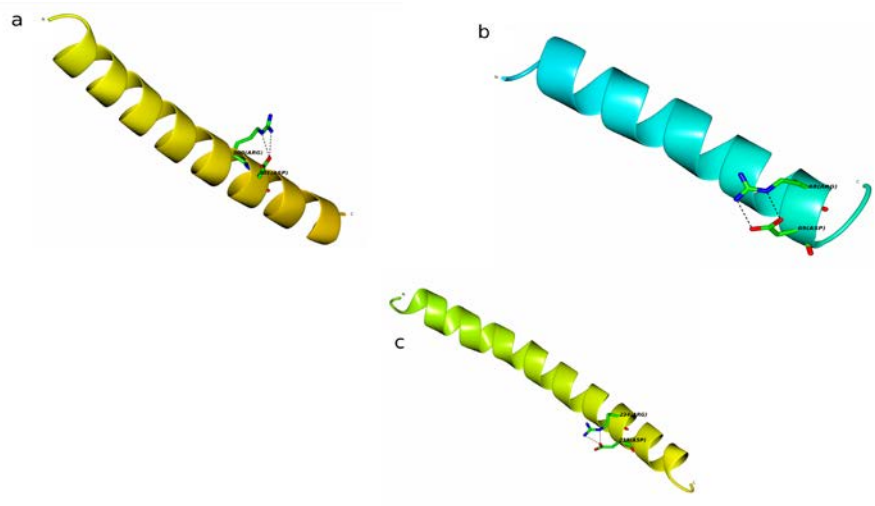

Figure 3. (a) The R-D motif in 1B5P occurring in helix group with Arg side chain conformation t t g+ t and Type D hydrogen bonding interaction. (b) The R-D motif in 4OHC occurring in helix group with Arg side chain conformation t t g+ t and Type B hydrogen bonding interaction. (c) The R-D motif in 3RKG occurring in helix group with Arg side chain conformation t g- g- g- and Type D hydrogen bonding interaction.

3.2.1.2 Analysis of D-K motif with one hydrogen bond in helices.

Total 203 occurrences of the D-K motif were identified with interactions. These could be further classified as 181 with one hydrogen bond and 22 with two. The side chain conformations for Lys residue showed mainly partially folded conformations. The most prominent conformation was g- t t t with 34 (19%) occurrences (Figure 4). In all cases The Asp χ_1 conformation was g-. The backbone conformation does not show any conservation. Only in two cases side chain side chain hydrogen bonding (Lys (NZ) – (OD1) Asp) was observed while in rest the hydrogen bonding was main chain-side chain (Asp (N) – (OD1) Asp). The next conformation studied was g- g- t g- with 32 (18%) occurrences (Figure 4). Here the Asp χ_1 conformation was again found to be g-. For this group all motifs were found to have a single side chain-side chain hydrogen bond Lys (NZ) – (OD2) Asp (Figure 5).

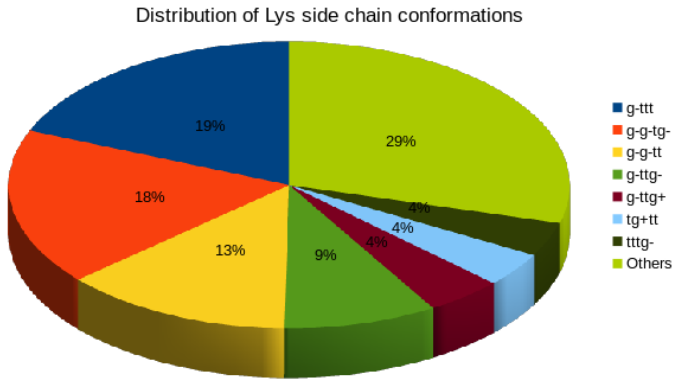

Figure 4. The distribution of Lys side chain conformation in D-K motif with one interaction in helices.

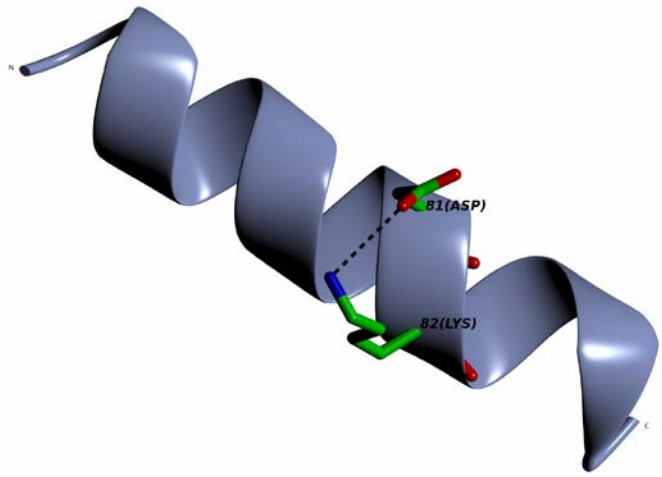

Figure 5. The D-K motif in 3VLD with g- g- t g- Lys side chain conformation and one side chain - side chain hydrogen bond.

3.2.1.3 Analysis of K-D motif with one hydrogen bond in helices.

All of the 251 occurrences of K-D motif were found to have a single interaction. 187 motifs had a single side chain – side chain hydrogen bond. Of the side chain conformations analyzed for Lys residue the partially folded g+ t t g+ side chain conformation was observed in majority of cases (Figure 6a). The Asp conformation angle χ_1 for the group was g+. All motifs were found to have a single side chain – side chain H-bond: Lys (NZ) – (OD1) Asp (Figure 6b).

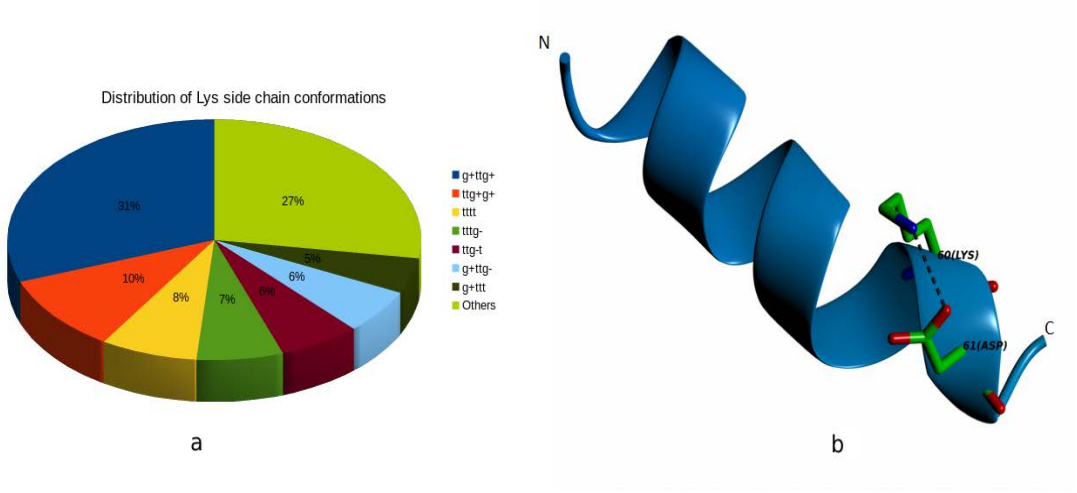

Figure 6. (a) The distribution of Lys side chain conformation in K-D motif with one interaction present in helices. (b) The K-D motif in 1XPM with g+ t t g+ Lys side chain conformation and one side chain - side chain hydrogen bond.

3.2.2 Analysis of motifs in sheets.

Only 97 occurrences of the D-R motif and 33 of R-D motif with interactions were recorded. Significant occurrences of motifs with interactions were recorded only for the D-X-R (314) and R-X-D (299) patterns.

3.2.2.1 Analysis of D-X-R motif with two hydrogen bonds in sheets.

143 occurrences of the D-X-R motif were found in sheets structure with two interactions. Pairwise superimposition carried out for the motifs showed considerable variation in the backbone conformation. The preferred Arg side chain conformation was g- t t t to be the most favorable one (53%) (Figure 7).

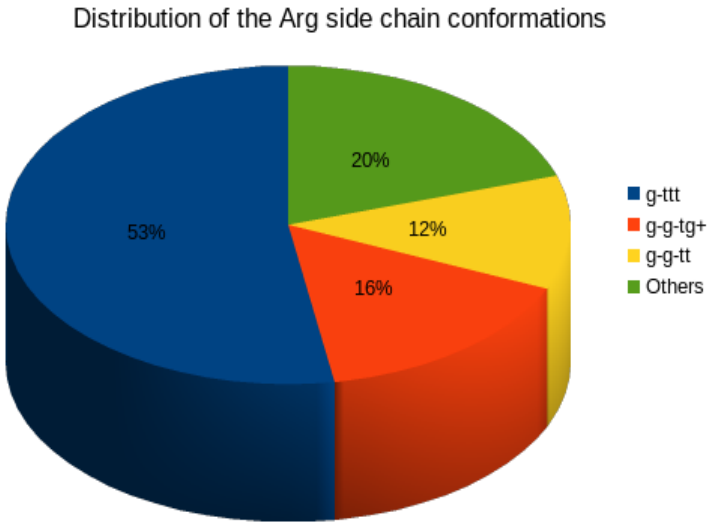

Figure 7. Distribution of Arg side chain conformations in D-X-R motifs found in sheet structures with two interactions.

Of the total 50 detected with this Arg side chain conformation 44 were found to have Asp χ_1 as t. Only in 6 cases the Asp side chain assumed g+ conformation for χ_1. The Ramachandran plot of the X residue show all motifs lie in the extended strand region.

The hydrogen bonding was dominantly found to be Type B (Figure 8a) in 35 cases while for 15 it was observed to be Type D (Figure 8b)

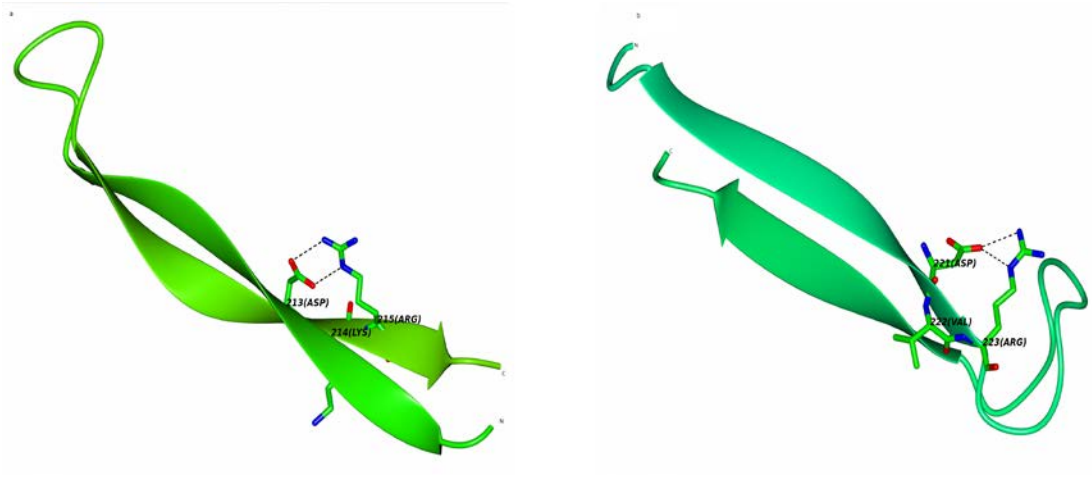

Figure 8. (a) The D-X-R motif in 1F8V occurring in sheet with g- t t t Arg side chain conformation, Asp χ_1 as t and Type B hydrogen bonding interaction. (b) The D-X-R motif in 4BBW occurring in sheet with g- t t t Arg side chain conformation, Asp χ_1 as g+ and Type D hydrogen bonding interaction.

3.2.2.2 Analysis of R-X-D motif with two hydrogen bonds in sheets.

As many as 17 varied conformations for the Arg side chain were observed among the 180 motifs in this group.

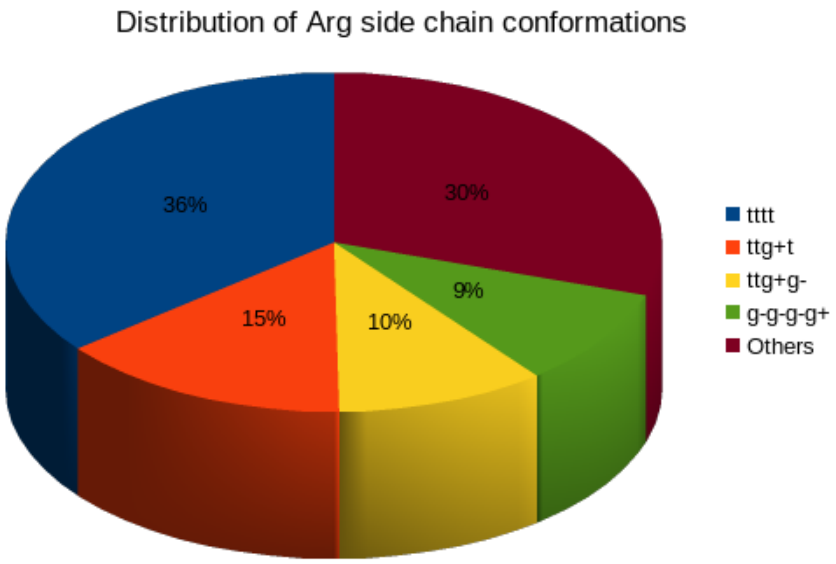

Figure 9. The distribution of Arg side chain conformation in R-X-D motif with two interactions in sheets.

The most dominant conformation observed was the extended rotamer t t t t (36%) (Figure 9). The superimposition analysis of the motifs belonging to this group showed variation of the backbone conformation.

The Asp side chain conformational angle χ_1 was found to be g- for all. The Ramachandran plot of X residue shows a single cluster with only one variant.

In 55 cases the hydrogen bonding was observed to be Type B (Figure 10a) while in 13 cases it was Type D (Figure 10b). In case of 3ILS/A/2082 (Format: PDBID/Chain/Residue number of first motif reside), the motif was found to lie at the C terminal end of a sheet while the ψ of X residue was observed to be less than 90° due to which the side chain-side chain H-bond was found to be replaced by a main chain-main chain H-bond namely, Asp (N) – (O) Arg (Figure 10c).

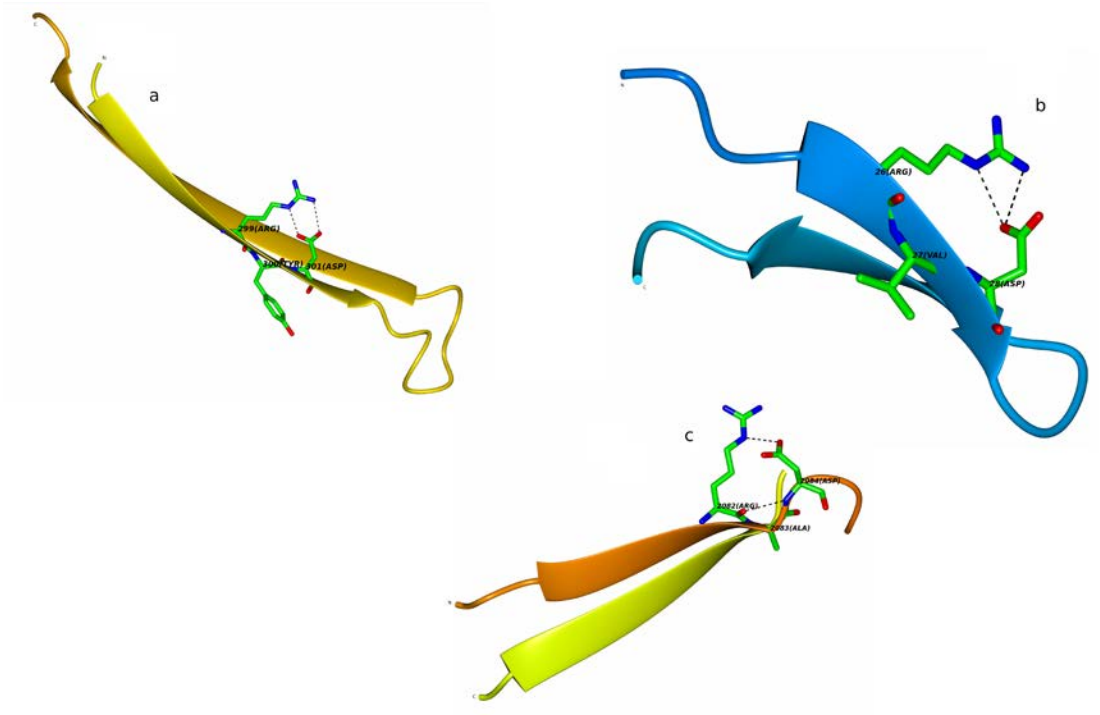

Figure 10. (a) The R-Y-D motif in 3SVZ occurring in sheet with t t t t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The R-V-D motif in 2ATZ occurring in sheet with t t t t Arg side chain conformation and Type D hydrogen bonding interaction. (c) The R-A-D motif in 3ILS occurring in sheet with t t t t Arg side chain conformation and one main chain – main chain and one side chain – side chain hydrogen bonding interaction.

3.2.3 Analysis of motifs in irregular structures.

Significant occurrences of motifs with interactions were observed for all motifs belonging to irregular structures. In the D-R motifs 584 occurrences were recorded with interactions of which 498 were with one H-bond and 86 were with two or three H-bonds. For the reverse R-D motifs 279 showed one hydrogen bond while 93 showed two or more. In case of the D-X-R, as high as 1488 motifs showed single H-bond while 2222 showed two or more. Even in the reversed positions of Arg and Asp, R-X-D, 599 were found to have one hydrogen bond and 930 having two or more. For the D-K and K-D motifs 509 and 443 motifs were found with one hydrogen bond. In the D-X-K motif 2538 motifs showed single H-bond and 208 showed two or more H-bonds. However, in the K-X-D motifs only 400 occurrences showed one H-bond and 19 showed two or more H-bonds.

3.2.3.1 Analysis of D-R motif with two hydrogen bonds in irregular structures.

Total 78 occurrences were observed for D-R motif belonging to irregular structure with two hydrogen bonds. The pairwise superimposition analysis of these motifs reveals a wide variation in motif backbone fold. The C_β-C_β virtual torsion angle was primarily found $< 50°$ (Figure 11).

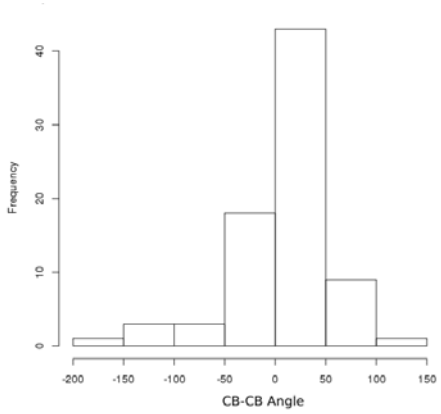

Figure 11. Histogram showing the distribution of the C_β-C_β orientation angle of the motifs.

This was followed by analysis of the side chain conformations of the Arg residues in the motifs. The graph shows the distribution of the Arg side chain conformations covering more than 65% of the cumulative number encountered in the analysis. The analysis of the side chains revealed g+ t g+ t to be the most favorable (40%) conformation (Figure 12). In all cases the Asp χ_1 was found to be consistently g+ except 3H2Z: A305 where the χ_1 was observed to be t. Of the 31 occurrences, 21 were found to have Type B hydrogen bonding (Figure 13a) while 10 were observed to have Type D (Figure 13b).

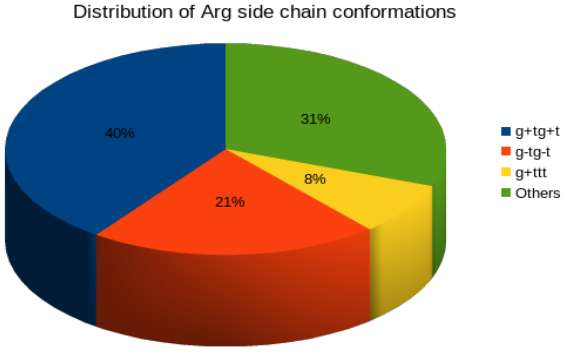

Figure 12. Distribution of Arg side chain conformations in D-R motifs in irregular structure with two interactions.

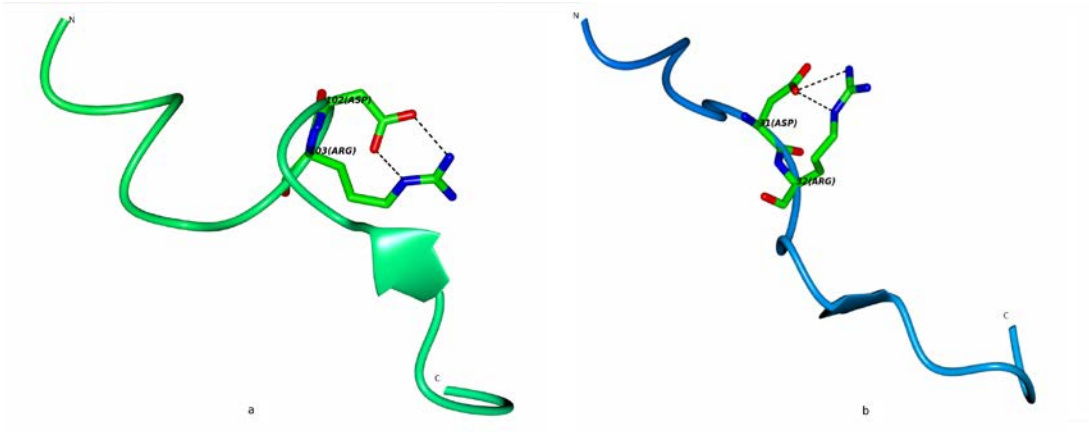

Figure 13. (a) The D-R motif in 4M8K occurring as irregular structure with g+ t g+ t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The D-R motif in 2BM8 belonging to irregular structure with Arg side chain conformation g+ t g+ t and Type D hydrogen bonding interaction.

In case of 3H2Z: A305, one main chain – side chain and one side chain – side chain hydrogen bond was found i.e. Arg (N) – (OD2) Asp and Arg (NE) – (OD2) Asp, a variation of Type D. In the other case 4JPX: A75, one main chain – side chain and one side chain – side chain hydrogen bond can be considered a variation of Type B.

For the second side chain conformation g- t g- t, of the 17 occurrences (Figure 12), nine were observed to have Type B interaction (Figure 14a) while eight were observed to have Type D interaction (Figure 14b). The Asp χ_1 in nine cases was observed to be t while in eight cases was found to be g+.

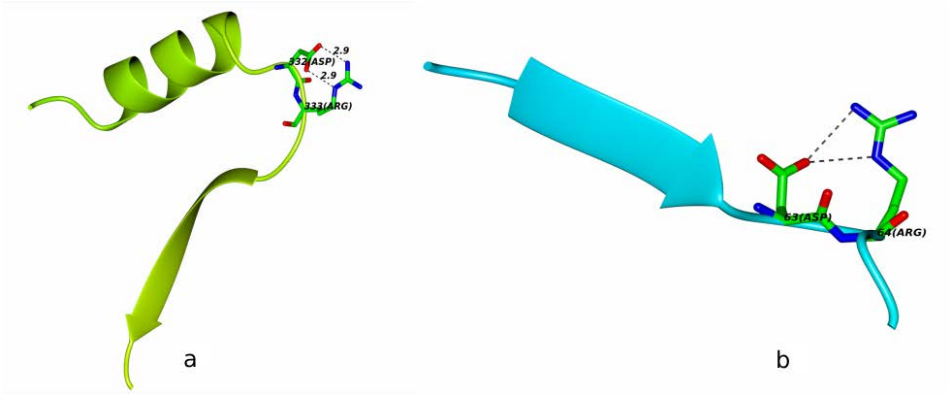

Figure 14. (a) The D-R motif in 1P9A occurring in irregular structure with g- t g- t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The D-R motif in 1FC9 occurring in irregular structure with g- t g- t Arg side chain conformation and Type D hydrogen bonding interaction.

3.2.3.2 Analysis of R-D motif with two hydrogen bonds in irregular structures.

The Cβ - Cβ angle distribution for motifs of this set was observed to lie in the range of 0-50° (Figure 15).

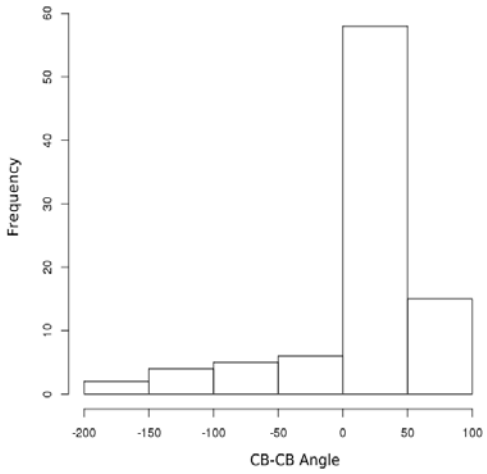

Figure 15. Histogram showing the distribution of the C_β-C_β orientation angle of the motifs.

The 90 occurrences observed were studied for their Arg side chain conformation. The partially extended conformation t t g+ t was observed to be the most favorable (Figure 16).

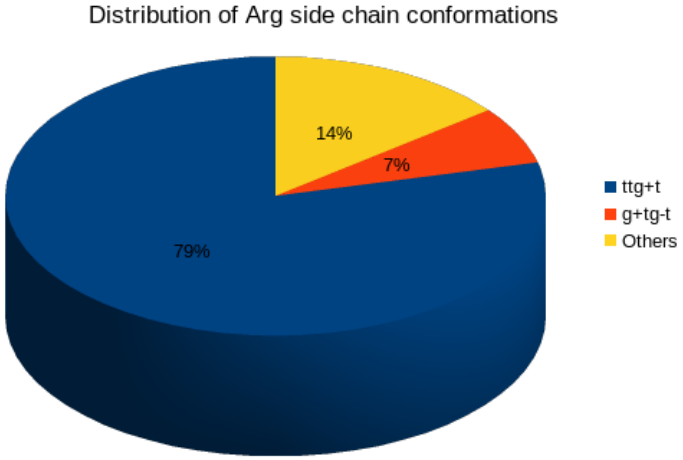

Figure 16. The distribution of Arg side chain conformation in R-D motif with two interactions in irregular structures.

The Asp χ_1 was found to be mainly g- except in few cases where it was observed to be t. The hydrogen bonding in 40 cases was found to be Type B (Figure

17a) while in 29 cases was observed to be Type D (Figure 17b). In case of 1OHE: A166, as the Asp side chain was found to lie above the plane of the peptide in the view shown, a side chain – main chain H-bond i.e. Arg (NE) – (O) Asp (Figure 17c) was observed while in case of 2GFF: A42, two side chain – main chain H-bonds namely Arg (NE) – (O) Asp and Arg (NH2) – (O) Asp were present.

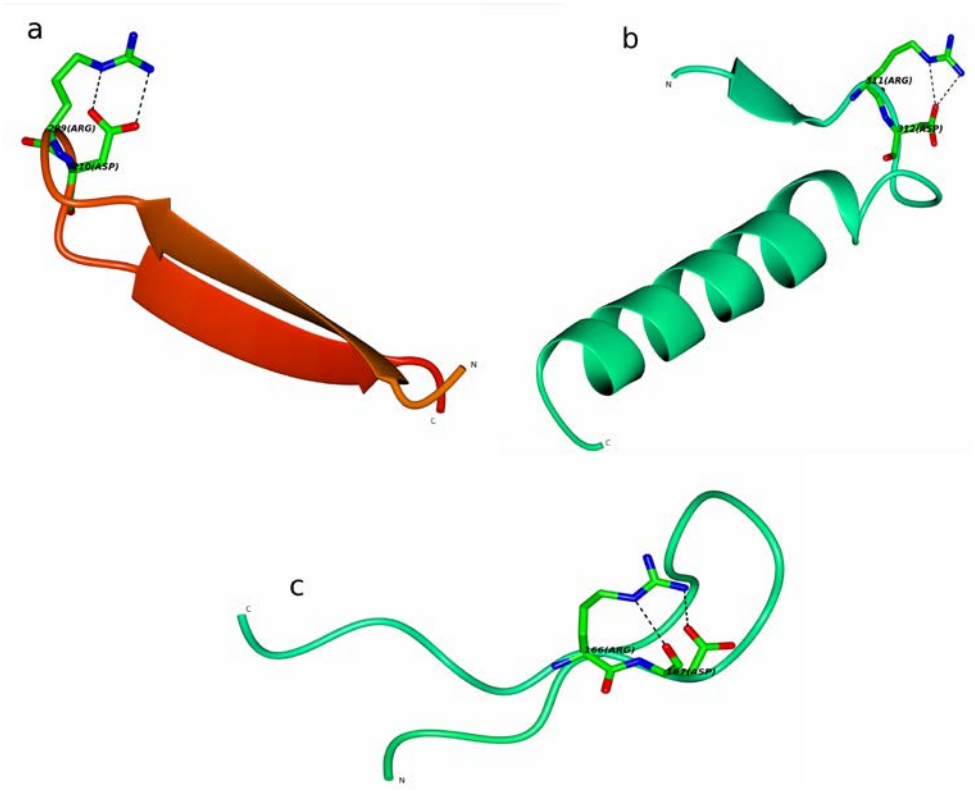

Figure 17. (a) The R-D motif in 4KFG occurring in irregular structures with t t g+ t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The R-D motif in 2DQ6 occurring in irregular structures with t t g+ t Arg side chain conformation and Type D hydrogen bonding interaction. (c) The R-D motif in 1OHE occurring in irregular structures with t t g+ t Arg side chain conformation and one side chain – main chain and one side chain – side chain hydrogen bonding interaction.

3.2.3.3 Analysis of D-X-R motif with two hydrogen bonds in irregular structures.

Total 505 occurrences of the D-X-R motif with two interactions in the irregular structures were observed from the local dataset. The pairwise superimposition of the motifs revealed them to have highly variable backbone (Figure 18).

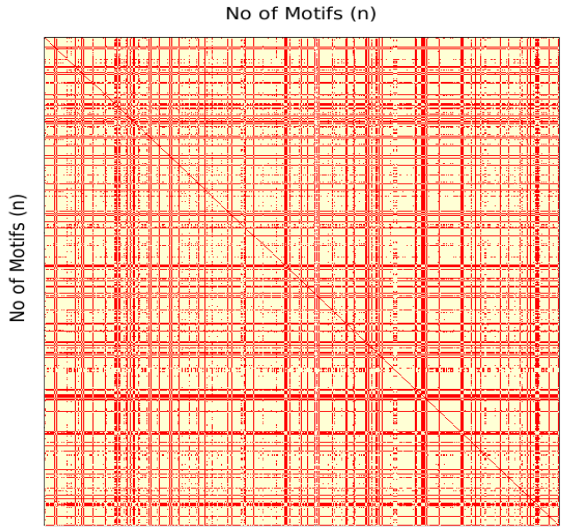

Figure 18. The pairwise superimposition graph of the D-X-R motifs in irregular structures with two interactions. A cut-off of 0.3Å was used.

Analysis of the Arg side chain rotamers revealed that they assumed a wide range of conformations ranging from a highly folded state to completely extended one. However, the partially folded side chain conformation g- t g- t was observed to be major one covering 37% of the total (Figure 19).

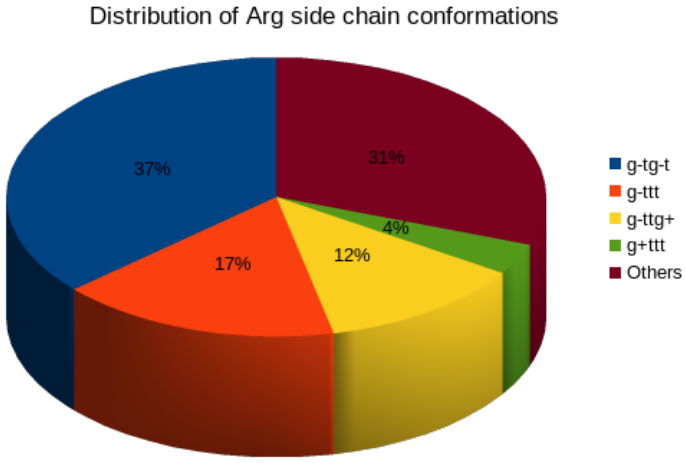

Figure 19. The distribution of Arg side chain conformation in D-X-R motif with two interactions in irregular structures.

The Asp χ_1 was found to have the extended conformation t in 114 cases while in 84 cases it assumed the folded conformation g+. The Ramachandran plot of the residues at the X position showed two distinct clusters; first near the α-helical region and the second at the left-handed helical region. Analysis of the residues at the X

position in motifs in left-handed helical region revealed them to be either Gly or Pro, which are allowed in this region (Figure 20).

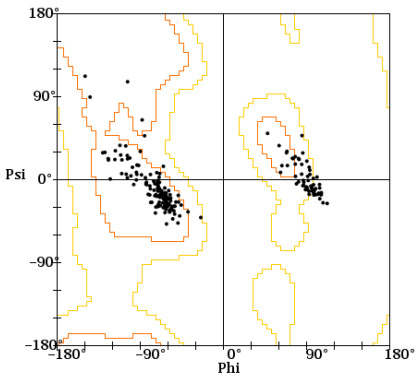

Figure 20. Ramachandran plot of residues in the X position in D-X-R motifs in irregular structures with two interactions and Arg side chain g- t g- t.

158 occurrences were observed to have a variant of Type B bonding where, two hydrogen bonds namely, Arg (N) – (OD1) Asp and Arg (NE) – (OD1) Asp (Figure 21a). In 39 cases Type B hydrogen bonding (Figure 21b) was observed while in 5 occurrences Type D bonding was observed. Only in case of 4H7U: A255, one main chain - side chain and one main chain – main chain hydrogen bonds were detected i.e. Asp (N) – (OD1) Asp and Arg (N) – (O) Asp (Figure 21c).

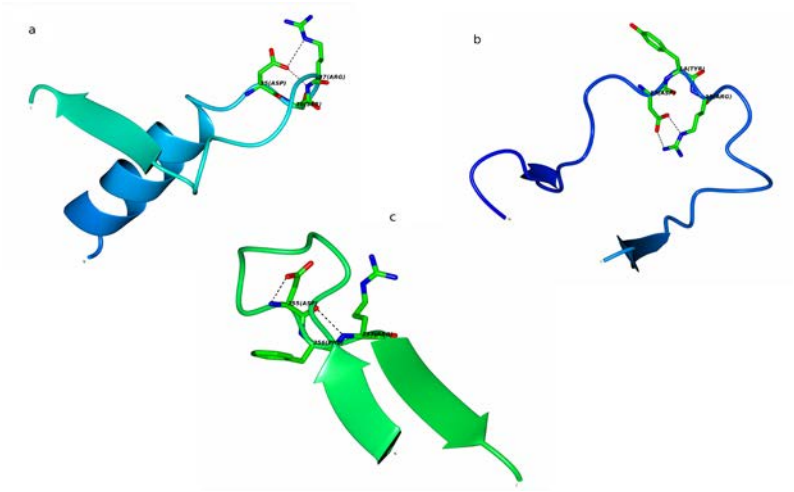

Figure 21. (a) The D-S-R motif in 1HKQ occurring in irregular structures with g- t g- t Arg side chain conformation and Type B variant hydrogen bonding interaction. (b) The D-Y-R motif in 2O34 occurring in irregular structures with g- t g- t Arg side chain conformation and Type B hydrogen bonding interaction. (c) The D-F-R motif in 4H7U occurring in irregular structures with g- t g- t Arg side chain conformation.

Motifs with the partially extended Arg side chain conformation g- t t t were found to contribute 17% of the total (Figure 19). Here the Asp χ_1 was found to have the extended conformation t in 74 cases while in 20 cases it assumed the folded conformation g+. The Ramachandran plot of X residues in the motifs showed segregation in three distinct clusters (Figure 22), the first two were similar to those discussed previously while a third with extended conformation was also observed. The only variation found was for 4JG3:A38 in the D-T-R motif where the Arg residue was found to have a positive ϕ value. Although a wide range of hydrogen bonding patterns were observed, a variant (40 occurrences) of Type B where a main chain - side chain interaction i.e. Arg (N) – (OD1) Asp was observed instead of Arg (NE) – (OD1) Asp (Figure 23a) and in 39 occurrences it was Type B (Figure 23b) and 8 occurrences were found with Type D interaction (Figure 23c).

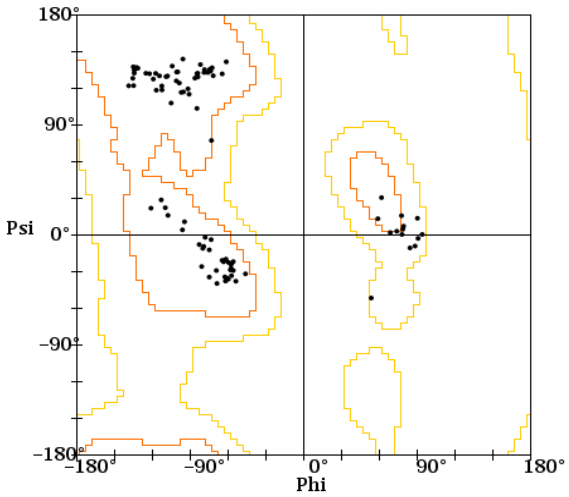

Figure 22. Ramachandran plot for residues in the X position in D-X-R motifs in irregular structures with two interactions and Arg side chain conformation g- t t t.

A variant of Type D was observed for 3TTC: A171, 4BB9: A356 and 4J18: A404 where Arg (N) – (OD1) Asp was observed instead of Arg (NE) – (OD1) Asp. In these cases it was observed that the residues in the X position had either long side chain or bulky side groups and the Asp and Arg side chains were found to lie on one side and above the peptide plane while the X residue side chain lie on the opposite side.

In case of 4MVH: A556, one side chain – main chain and one side chain – side chain, where both Asp (O) and Asp (OD2) were acceptors to hydrogen of Arg

(NH1) while hydrogen atoms of Arg (NH2) and Arg (NE) were found to point away from the Asp oxygen atoms (Figure 23d).

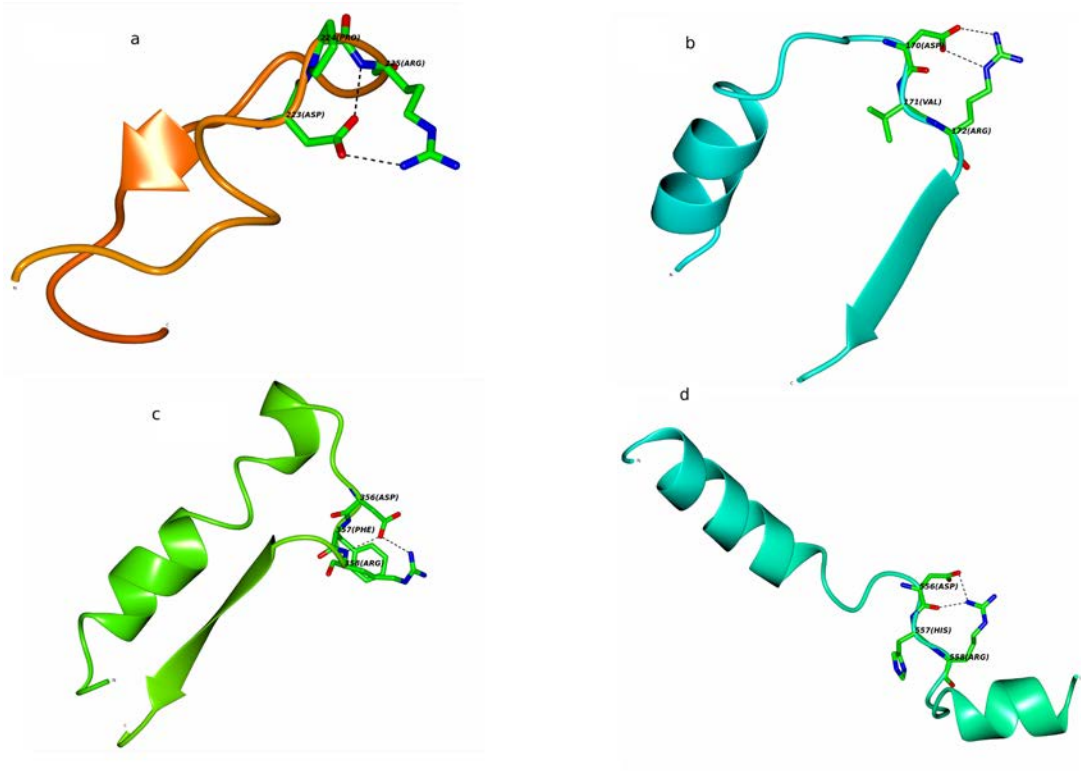

Figure 23. (a) The D-X-R motif in 2AMY occurring in irregular structures with g- t t t Arg side chain conformation and a variant of Type B hydrogen bonding interaction. (b) The D-X-R motif in 1O0S occurring in irregular structures with g- t t t Arg side chain conformation and Type B hydrogen bonding interaction. (c) The D-X-R motif in 4BB9 occurring in irregular structures with g- t t t Arg side chain conformation and Type D hydrogen bonding interaction. (d) The D-X-R motif in 4MVH occurring in irregular structures with g- t t t Arg side chain conformation and one side chain – main chain interaction; Arg (NH1) – (O) Asp and one side chain – side chain interaction; Arg (NH1) – (OD2) Asp.

3.2.3.4 Analysis of D-X-R motif with three hydrogen bonds in irregular structures.

D-X-R motifs with three interactions were found only in the irregular structure group. The superimposition analysis of this motif shows conservation of the backbone conformation across structures (Figure 24).

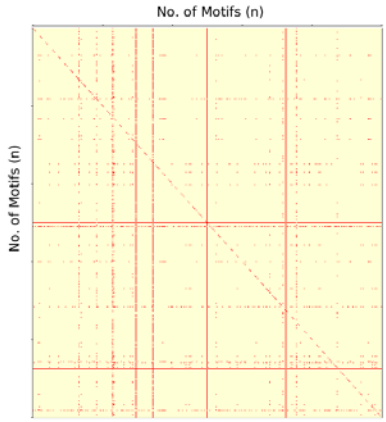

Figure 24. The pairwise superimposition graph of the D-X-R motifs in irregular structures with three interactions. A cut-off of 0.3Å was used.

Although about 20 different conformations for the Arg side chains were observed in this set the partially folded conformation g- t g- t was found to the most favorable one covering 95% of the data for this set (Figure 25). The Asp χ_1 conformation for the set was primarily found to be t.

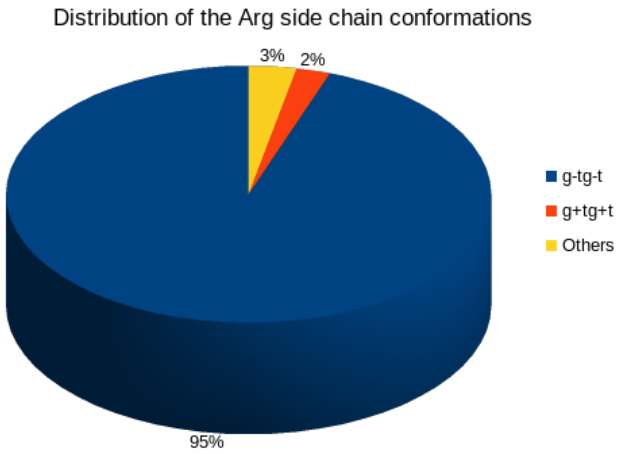

Figure 25. The distribution of Arg side chain conformation in D-X-R motif with three interactions in irregular structures.

The Ramachandran plot of the X residues of the motifs in this set shows exactly the same segregation as observed for the irregular structure group with two interactions. 1284 motifs were detected with their X residue lying in the α-helix region while 66 were found to lie in the left handed helix region (Figure 26).

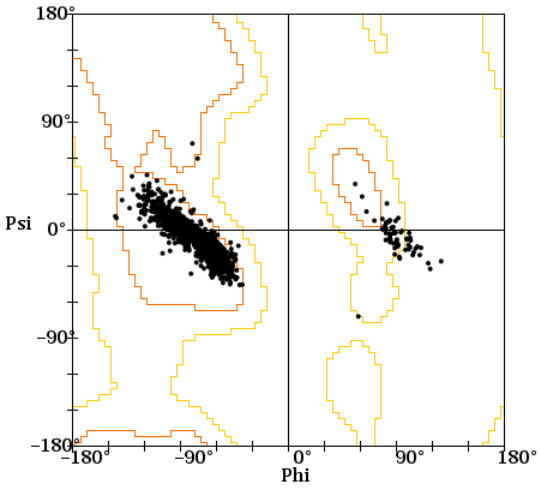

Figure 26. Ramachandran plot for residues in the X position in D-X-R motifs in irregular structures with three interactions.

All motifs were observed to have the same hydrogen-bonding pattern namely, Type B along with an additional main chain – side chain hydrogen bond Arg (N) – (OD1) Asp (Figure 27a). Only 31 occurrences were found where Arg side chain was observed to have the conformation g+ t g+ t (Figure 27b). The Asp χ_1 conformation for the set was found to be t. Also, the Ramachandran plots as well as the hydrogen bonding were observed to be similar to the previously analyzed set. The difference in side chain orientation with respect to the direction of the fold between Arg rotamers g- t g- t and g+ t g+ t can be seen in figures 27a & b, although the three hydrogen bonds remain same.

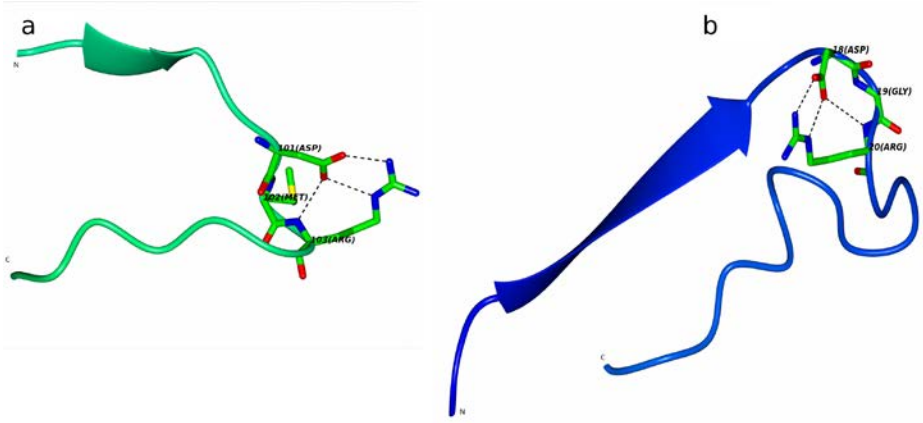

Figure 27. (a) The D-X-R motif in 3PFG occurring in irregular structures with g- t g- t Arg side chain conformation and Type B hydrogen bonding interaction with Arg (N) – (OD1) Asp. (b) The D-X-R motif in 1EOK occurring in irregular structures with g+ t g+ t Arg side chain conformation and Type B hydrogen bonding interaction with Arg (N) – (OD1) Asp.

3.2.3.5 Analysis of R-X-D motif with two hydrogen bonds in irregular structures.

743 motifs were detected with two interactions in fold belonging to irregular structures. A wide range of Arg side chain conformations were observed for the motifs belonging to this group.

Two rotamers of the Arg side chain conformations were observed in 159 (40%) occurrences. The most favorable conformation was found to be the folded g- t g+ g+ (Figure 28). While in 10 cases the Asp χ_1 was found to be g-, in rest of the 149

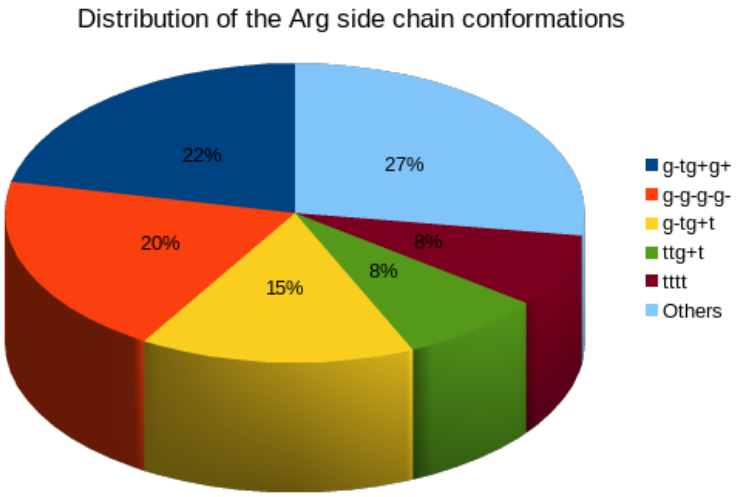

Figure 28. The distribution of Arg side chain conformation in R-X-D motif with two interactions in irregular structures.

cases the χ_1 angle was g+. The superimposition analysis of the motifs in this group revealed only two variations (Figure 29a). The Ramachandran plot for the X residue primarily occupies the α-helical region with few in the left handed helical region and the extended region (Figure 30a). For those in the left handed helical region, the X residue was observed to be Gly.

For others the backbone was found to fold differently which explained the variation observed in the Ramachandran plot. In 145 cases the Type B (Figure 31a) hydrogen bonding was observed while in the rest 12 cases the bonding was found to be Type D (Figure 31b). For 2P1M: A250, a variant of Type D wherein instead of a side chain-side chain, a main chain-side chain interaction, namely, Asp (N) – (OD2) Asp was observed. In case of 3BF0: A217, a side chain – main chain i.e. Arg (NE) –

(O) Asp as both Asp and Arg side chains were above the plane of the peptide in the view shown (Figure 31c).

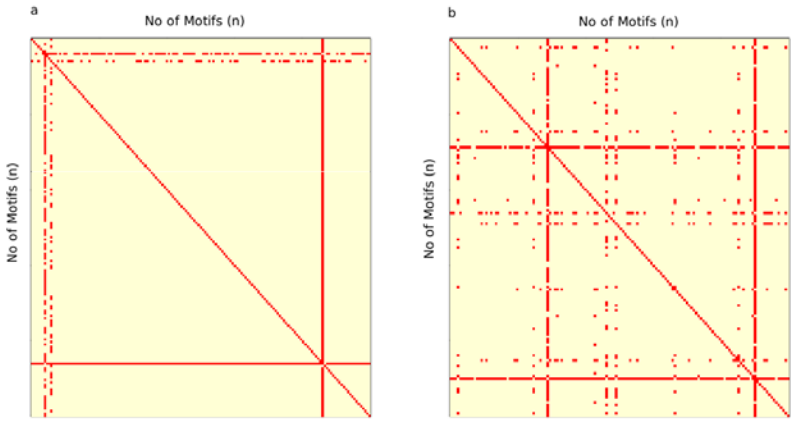

Figure 29. (a) The pairwise superimposition graph of the R-X-D motifs in irregular structures with two interactions and Arg side chain conformation g- t g+ g+ using a cut-off of 0.3Å. (b) The pairwise superimposition graph of the R-X-D motifs in irregular structures with two interactions and Arg side chain conformation g- g- g- g- using a cut-off of 0.3Å.

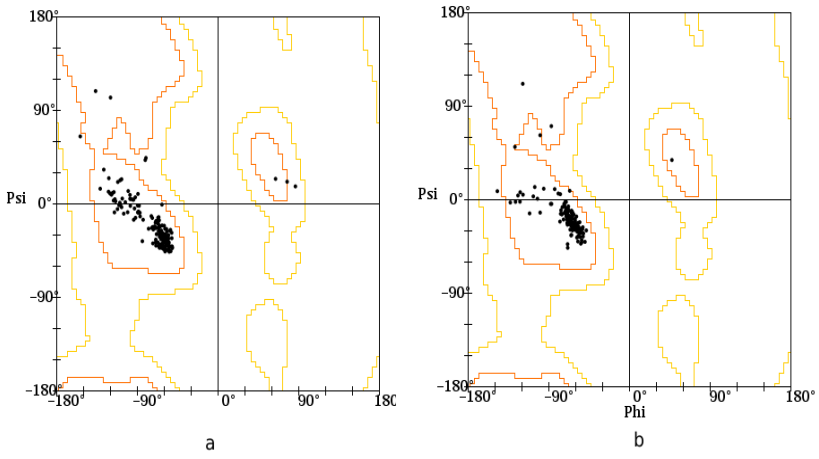

Figure 30. (a) Ramachandran plot for residues in the X position in R-X-D motifs in irregular structures with two interactions and Arg side chain conformation g- t g+ g+. (b) Ramachandran plot for residues in the X position in R-X-D motifs in irregular structures with two interactions and Arg side chain conformation g- g- g- g-.

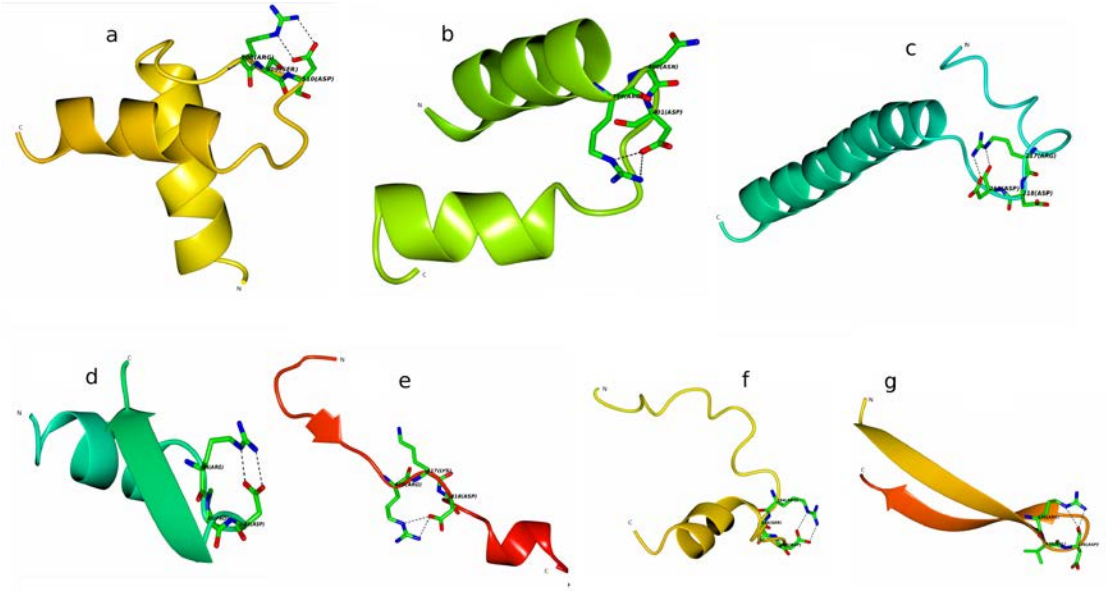

Figure 31. (a) The R-S-D motif in 4A57 occurring in irregular structures with g- t g+ g+ Arg side chain conformation and Type B hydrogen bonding interaction. (b) The R-N-D motif in 2EPL occurring in irregular structures with g- t g+ g+ Arg side chain conformation and Type D hydrogen bonding interaction. (c) The R-D-D motif in 3BF0 occurring in irregular structures with g- t g+ g+ Arg side chain conformation and a side chain – main chain H bond i.e. Arg (NE) – (O) Asp. (d) The R-P-D motif in 2Z7B occurring in irregular structures with g- g- g- g- Arg side chain conformation and Type B hydrogen bonding interaction. (e) The R-K-D motif in 1GSO occurring in irregular structures with g- g- g- g- Arg side chain conformation and Type D hydrogen bonding interaction. (f) The R-S-D motif in 4KQA occurring in irregular structures with g- g- g- g- Arg side chain conformation and variant of Type B hydrogen bonding interaction. (g) The R-V-D motif in 4HZ9 occurring in irregular structures with g- g- g- g- Arg side chain conformation and two side chain – main chain hydrogen bonds.

Next, the highly folded Arg side chain conformation g- g- g- g- was studied in the 144 motif occurrences. The superimposition analysis for the motifs in this group revealed two variations (Figure 29b). In 6 cases the Asp χ_1 was found to be g- and in the rest 138 it was found to be g+. The Ramachandran plot of X residue was observed to lie in the α-helical region with few variations (Figure 30b). In 139 cases the hydrogen bonding was Type B (Figure 34d) while in two cases it was observed to be Type D (Figure 31e). In case of 4HZ9: B134 two side chain – main chain hydrogen bonds were calculated namely Arg (NE) – (O) Asp and Arg (NH2) – (O) Asp (Figure 31g). For 4KQA: A394, a variant of Type B wherein Asp (N) – (OD1) Asp bond was observed (Figure 31f).

3.2.3.6 Analysis of R-X-D motif with three hydrogen bonds in irregular structures.

From the total 187 motifs detected with three interactions in irregular structures, 31% were found to have a highly folded Arg side chain conformation g- g- g- g- (Figure 32b). The superimposition analysis for the motifs shows almost similar backbone conformations (Figure 32a). The Asp χ_1 was found to be exclusively g+.

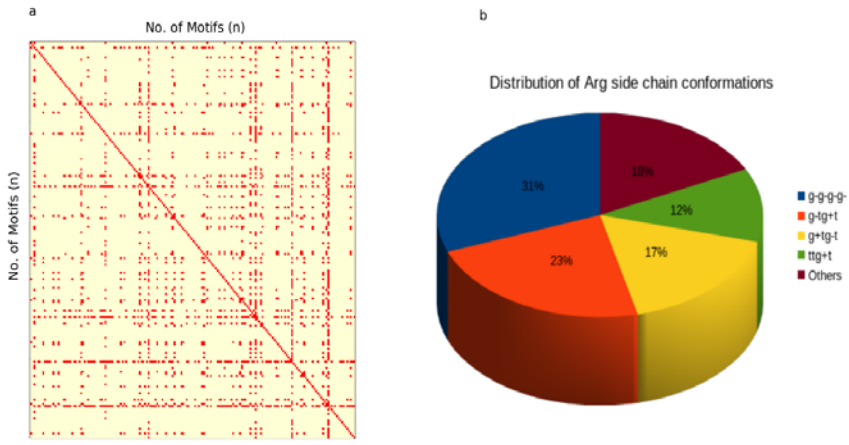

Figure 32. (a) The pairwise superimposition graph of the R-X-D motifs in irregular structures with three H bond interactions using a cut-off of 0.3Å. (b) The distribution of Arg side chain conformation in R-X-D motif with three interactions in irregular structures.

The Ramachandran plot for the X residue showed a single cluster (Figure 33). All the motifs in this group were found to have Type B hydrogen bonding along-with a main chain – side chain bond i.e. Asp (N) – (OD1) Asp.

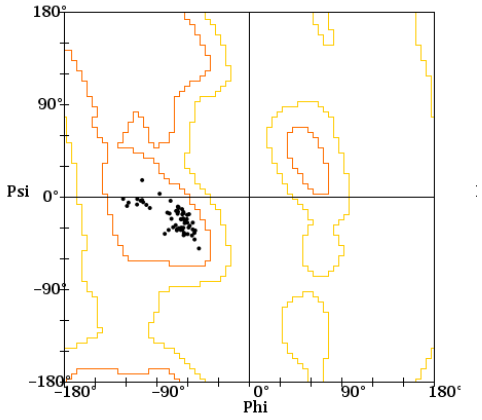

Figure 33. Ramachandran plot for residues in the X position in R-X-D motifs belonging to irregular structure with three interactions.

3.2.3.7 Analysis of D-K motif with one hydrogen bond in irregular structures.

Although 538 motifs were identified belonging to irregular structures with interaction, 508 were observed with one and 29 with two or more interactions. While a wide range of conformations of Lys side chain were identified, the conformations g- t t t was observed for 75 (15%) occurrences (Figure 34a). In all motifs the Asp χ_1 was g-. The hydrogen bonding in this group was mainly main chain – side chain, Lys (N) – (OD1) Asp. The backbone conformation showed no conservation across structures.

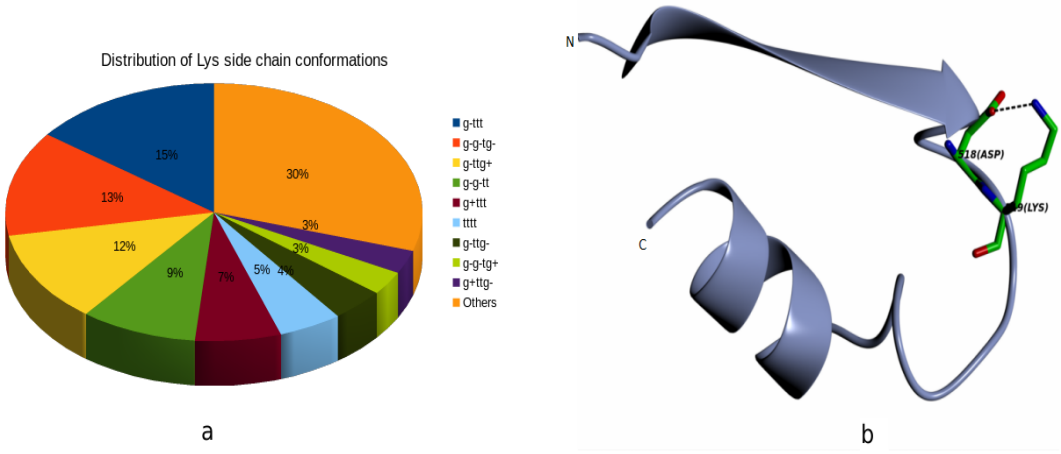

Figure 34. (a) The distribution of Lys side chain conformation in D-K motif with one interaction in irregular structures. (b) The D-K motif in 1C4O with g- g- t g- side chain conformation and one side chain – side chain hydrogen bond

The other conformation g- g- t g- was found in 67 occurrences (Figure 34a) where the Asp χ_1 was g-. Of the 67 occurrences, 48 were found to have side chain – side chain H bond, Lys (NZ) – (OD1) Asp (Figure 34b) while for the rest of the motifs the bonding was main chain - side chain. Even for this group no conservation of the backbone conformation was observed.

3.2.3.8 Analysis of K-D motif with one hydrogen bond in irregular structures.

Two side chain conformations of the Lys residue were studied in detail. In 81 occurrences the Lys side chain conformation was t t t g- (18%) (Figure 35). For this group, the Asp χ_1 was t. In 68 occurrences the hydrogen bonding was side chain - side chain (Figure 36a), side chain – main chain in 10 (Lys (NZ) – (O) Asp) and main chain - side chain in 3.

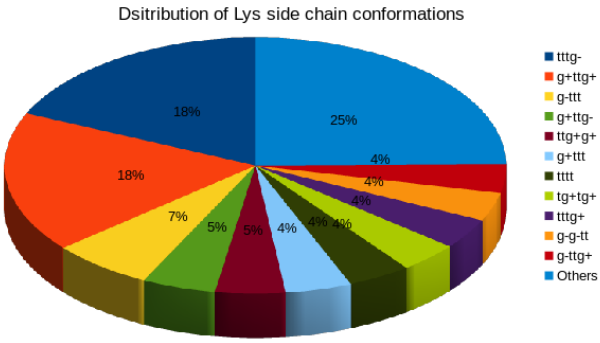

Figure 35. The distribution of Lys side chain conformation in D-K motif with one interaction in irregular structures.

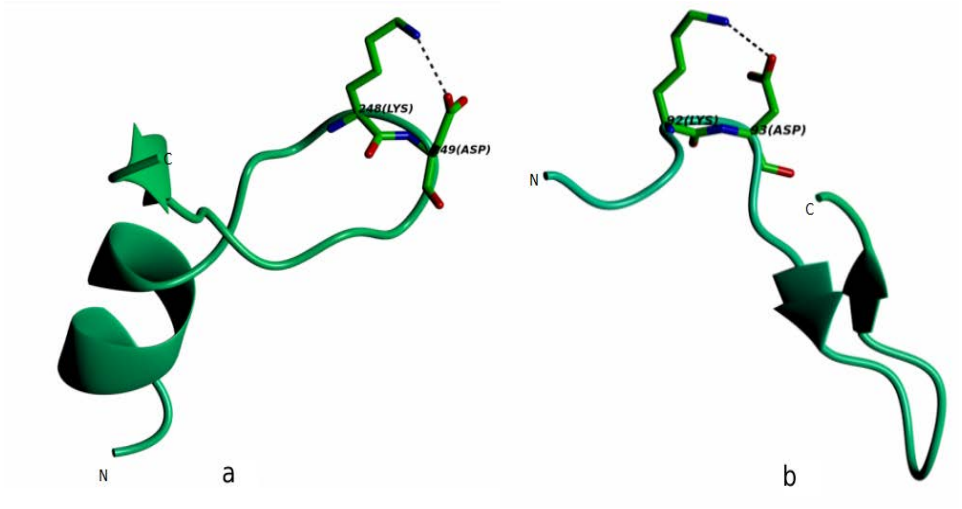

Figure 36. (a) The KD motif in 3NT1 with t t t g- side chain conformation and one side chain – side chain hydrogen bond in irregular structures. (b) The KD motif in 4H42 with g+ t t g+ sidechain conformation and one side chain – side chain hydrogen bond in irregular structures.

The second conformation identified was g+ t t g+ observed in 79 occurrences (18%) (Figure 35). In this case the Asp χ_1 was g+. Out of 79 motifs, in 74 motifs the hydrogen bonding was of the type side chain – side chain (Figure 36b) whereas it was main chain – side chain in 3 and side chain – main chain in two.

3.2.3.9 Analysis of D-X-K motif with one hydrogen bond in irregular structures.

Total 2538 motifs of the D-X-K pattern were observed with one interaction belonging to irregular structures. A very wide range of side chain conformations was observed for the Lys residue. The conformation g- t t t was found to occur in 610 (24%) occurrences (Figure 37a). The Asp χ_1 here also showed considerable variation with majority t and g+ (Figure 37b).

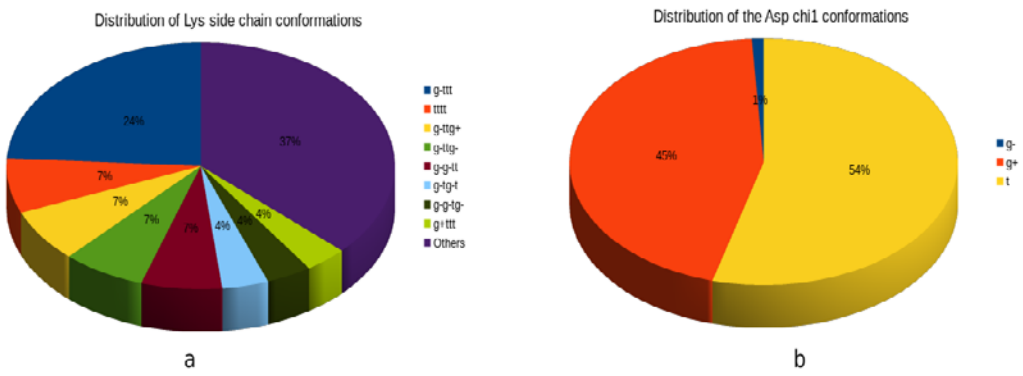

Figure 37. (a) The distribution of Lys side chain conformation in D-X-K motif with one interaction in irregular structures. (b) The distribution of Asp χ_1 conformation in D-X-K motif with one interaction in irregular structures.

The Ramachandran plot was largely concentrated in the α-helix region with those having Gly as the X residue were occurring in the left – handed helix region while a few were observed in the β-sheet region (Figure 38a). The hydrogen bonding for the group was mainly main chain – side chain, Lys (N) – (OD1) Asp and a few with side chain – side chain: Lys (NZ) – (OD2) Asp (Figure 38b).

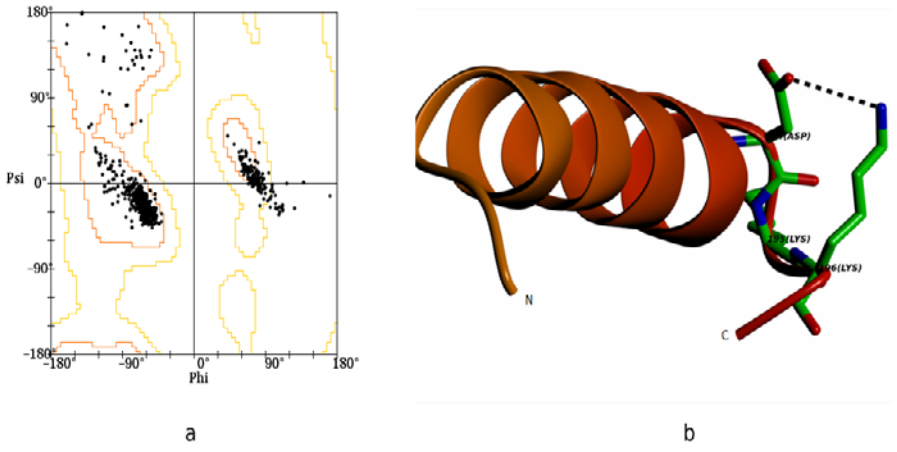

Figure 38. (a) Ramachandran plot for residues in the X position in D-X-K motifs belonging to irregular structure with one interaction. (b) The DKK motif in 4K2D with g- t t t side chain conformation and one side chain – side chain hydrogen bond in irregular structures.

3.2.3.10 Analysis of D-X-K motif with two hydrogen bonds in irregular structures.

208 motifs were observed with two interactions in the D-X-K motif in the irregular structures. The most prominent side chain conformation of Lys residue was

g- t g- g- (94) (Figure 39). The Asp χ_1 conformation for the motifs was observed to be t.

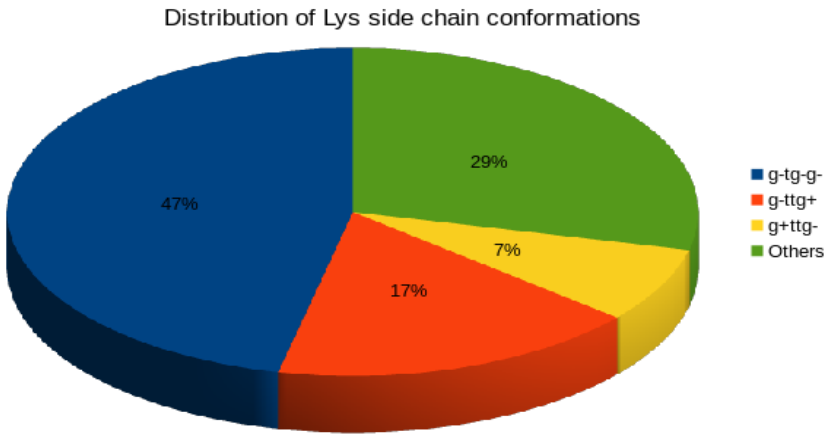

Figure 39. The distribution of Lys side chain conformation in D-X-K motif with two H-bonds in irregular structures.

The X residue occupied the α-helix region of the Ramachandran plot while 2 occurrences were observed to lie in the left handed region, which were found to be Gly and two in the sheet region (Figure 40a). From the 94 motifs, 92 were found to possess a side chain – side chain H-bond along-with a main chain – side chain H-bond (Figure 40b). The folding for this structural motif was observed to be nearly similar with D-X-R motif with three hydrogen bonds.

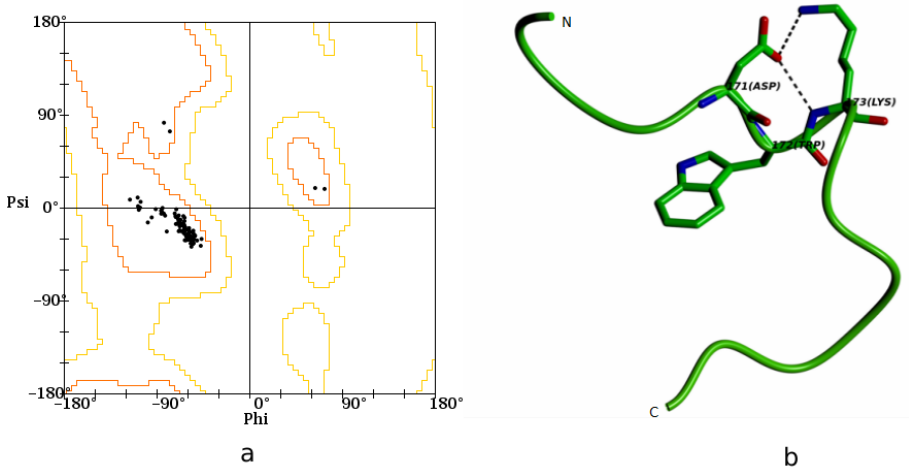

Figure 40. (a) Ramachandran plot for residues in the X position in D-X-K motifs belonging to irregular structures with two interactions. (b) The DWK motif in 1YG9 with g- t g- g- side chain conformation and one side chain – side chain hydrogen bond along-with a main chain – side chain bond in irregular structures.

3.2.3.11 Analysis of K-X-D motif with one hydrogen bond in irregular structures.

Total 400 K-X-D motifs were observed with one interaction in irregular structures. The side chain conformation of Lys was found to show a wide variation. There were 75 occurrences of the Lys side chain rotamer g- t t t (Figure 41). The Asp χ_1 conformation was found to be g+.

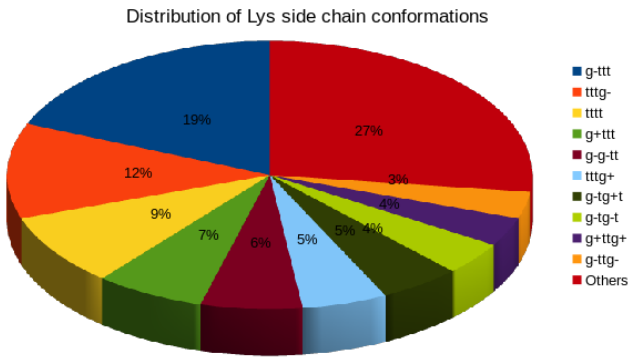

Figure 41. (a) The distribution of Lys side chain conformation in K-X-D motif with one interaction in irregular structures.

The Ramachandran plot shows a wide spread over the α-helix and β-sheet region (Figure 42a). 53 motifs were found to show a main chain - main chain H-bond, 21 showed a main chain – side chain H-bond while only 1 (1ASH:A15) was found to have a side chain – side chain hydrogen bond (Figure 42b).

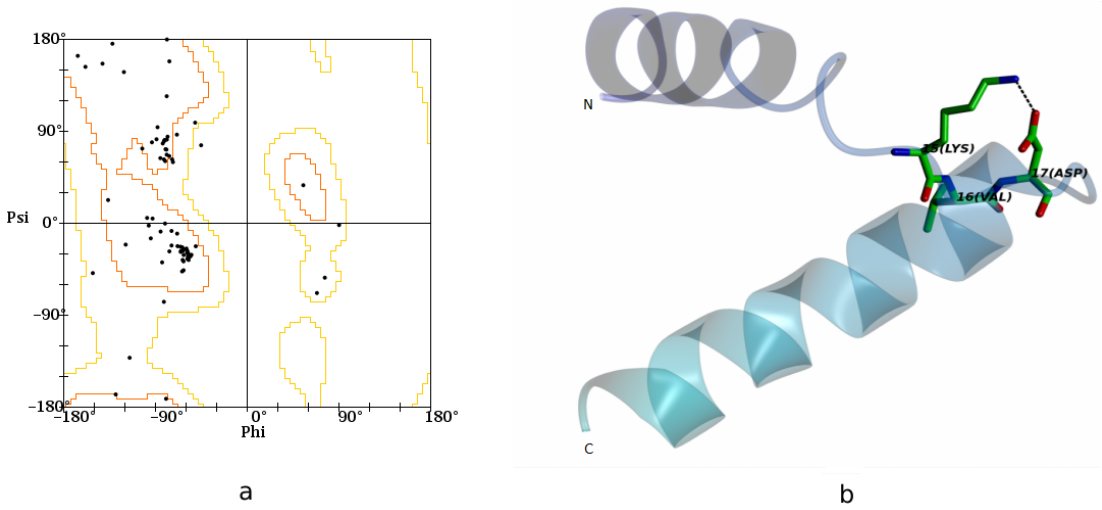

Figure 42. (a) Ramachandran plot for residues in the X position in K-X-D motifs belonging to irregular structure with two interactions. (b) The K-V-D motif in 1YG9 with g- t t t side chain conformation and one side chain – side chain hydrogen bond in irregular structures.

3.3 Comparative Analysis of the motifs

Based on the motifs explored above, the patterns involving Asp and Arg were compared in order to gain a better understanding of the characteristics of the motif due to such change of amino acid order in the sequence and introduction of residues separating the charged amino acids.

Table 4: Comparison of the D-R, R-D, D-X-R and R-X-D motifs.

Motif	SS	No. of Interactions	Type	Major Arg rotamer	Major Asp χ_1 Conf	Cα-Cα dist.	Fig No
D-R	H	2	Type D	g- g- t t	g+	-	45a
R-D	H	2	Type D	t t g+ t	g-	-	45b
D-X-R	S	2	Type B	g- t t t	t	6.502	48a
R-X-D	S	2	Type B	t t t t	g-	6.679	46c,48b
D-R	Irregular structures	2	Type B	g+ t g+ t	g+	-	43a,45c
R-D	Irregular structures	2	Type B	t t g+ t	t	-	45d
D-X-R	Irregular structures	2,3	Type B, Type B +1	g- t g- t	t	5.499	43b,48c
R-X-D	Irregular structures	2,3	Type B, Type B + 1	g- g- g- g-	g+	5.354	46d, 48d

3.3.1 Comparing the DR and D-X-R motifs.

The motifs DR and D-X-R were compared to study the change in the characteristics of the motifs with the introduction of a spacer residue. The motifs were compared for secondary structure preference, number and types of hydrogen bonding interactions, Arg and Asp side chain conformations and the average C_α (D/R)…C_α (R/D) distance where relevant. Table 4 gives a detailed report of the DR and D-X-R motifs.

The above analysis shows that both D-R and D-X-R have preference for irregular structures. However, between α-helix and β–sheet D-R is found more in α–helix and D-X-R more in β-sheet. For the D-R motif in helices, only two hydrogen bond interactions of Type D were found. The Arg side chain was found to be partially folded while the Asp side chain χ_1 was found to assume g+ conformation. D-X-R

motif in sheets again showed only two hydrogen bond interactions, however, these were observed to be of Type B. Since both the motifs were observed in the irregular structures, the comparison of their occurrence has been shown in Table 4. While motifs with two hydrogen bond interactions were mainly observed for the DR motifs, the introduction of a spacer residue in the D-X-R motifs allowed the motifs to have both two as well as three H-bonds amongst the motif residues. In both the cases the hydrogen bonding was found to be Type B, which in many of the D-X-R motifs in this group was supplemented by the presence of an additional bond Arg (N) – (OD1) Asp (Figure 43a,b). While in both motifs in this group the Arg side chain was partially folded, the conformation in D-R g+ t g+ t was changed to g- t g- t on introduction of the spacer residue. The Asp χ_1 was also found to change from g+ to t. While the folded side chain conformations of Arg and Asp allow for the hydrogen bonding with the residues as immediate neighbors in the D-R motif, for D-X-R motifs Asp conformation was t for the interaction with Arg due to the presence of the X residue between them. The D-X-R motifs observed with both two and three interactions were ideally found to belong to the same set wherein the Arg side chain conformation was observed to be g- t g- t. The Ramachandran plots for the X residues of both groups were found to occupy the same region and split in two groups. The backbone folding in both cases was also observed to be exactly similar (Figure 44a-d).

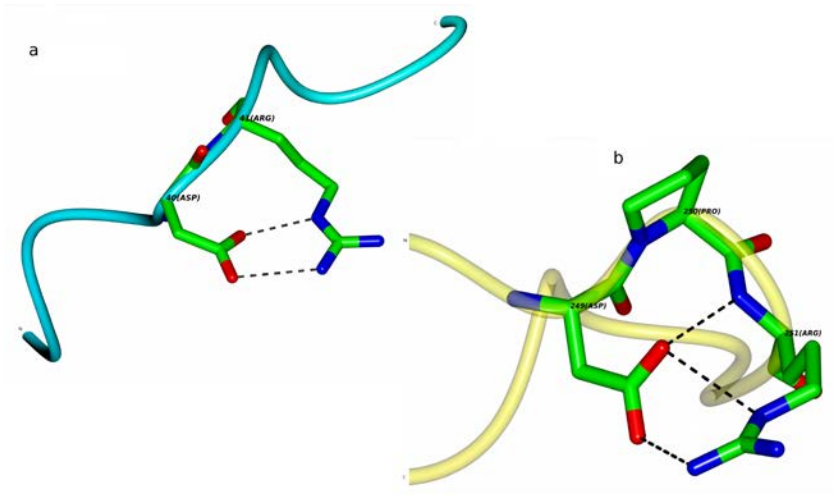

Figure 43. (a) The D-R motif in 4FZX occurring in irregular structures group with g+ t g+ t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The D-P-R motif in 1BT3 occurring in irregular structures group with g- t g- t Arg side chain conformation and Type B + 1 hydrogen bonding interaction.

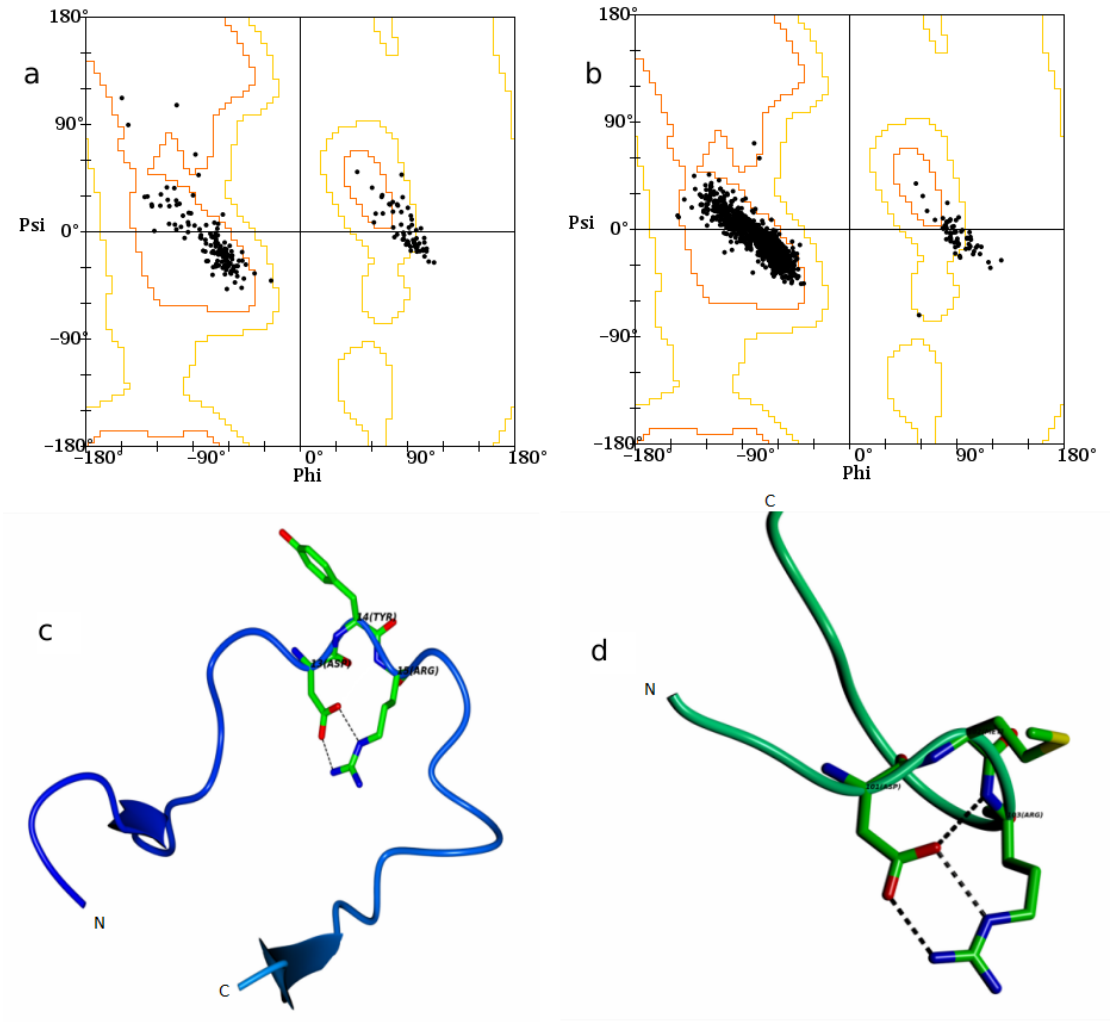

Figure 44. (a) The Ramachandran plot for the X residue for D-X-R motifs with Arg side chain g- t g- t and two interactions. (b) The Ramachandran plot for the X residue for D-X-R motifs with Arg side chain g- t g- t and three interactions. (c) The D-Y-R motif in 2O34 with Arg side chain g- t g- t and two interactions. (d) The D-M-R motif in 3PFG with Arg side chain g- t g- t and three interactions.

3.3.2 Comparing the DR and RD motifs.

The motifs DR and RD were compared to study the change in the features of the motifs with the reversal of the residue positions, although continuing as neighbors. Comparing the occurrence of the motifs in the secondary structure groups clearly reveals a minor shift of preference from the irregular structures group to helix group for the DR —> RD change. This shift of preference was found more evident for motifs with interactions where an increase of motif occurrence in helix group was noted along with a decrease in the irregular structures group. Table 4 gives a detailed report of the DR and RD motifs.

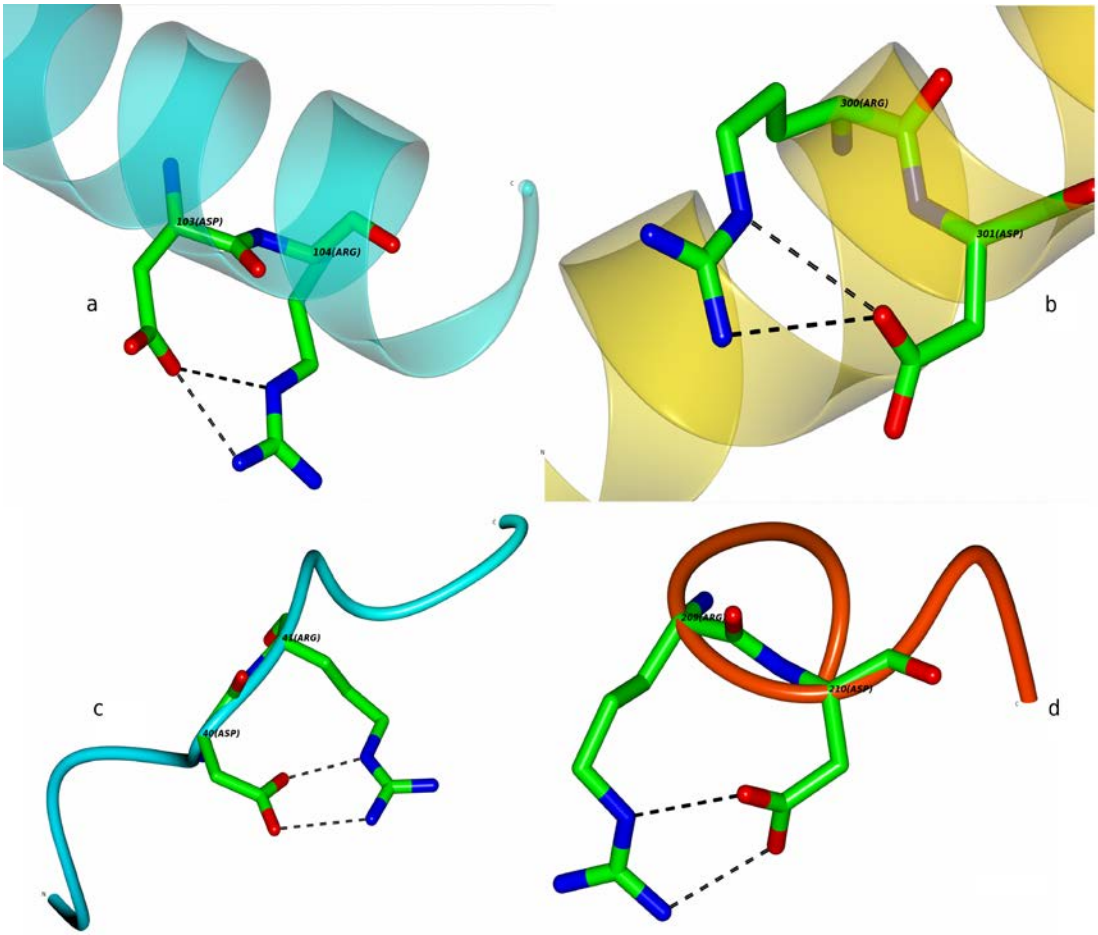

Figure 45. (a) The DR motif in 3KB9 occurring in helix structure with Arg side chain conformation g- g- t t and Type D hydrogen bonding interaction. (b) The RD motif in 1B5P occurring again in helix structure with Arg side chain conformation t t g+ t and Type D hydrogen bonding interaction. (c) The DR motif in 4FZX present in irregular structure with Arg side chain conformation g+ t g+ t and Type B hydrogen bonding interaction. (d) The RD motif in 4KFG which is also present in irregular structure and having Arg side chain conformation t t g+ t and Type B hydrogen bonding interaction.

Motifs primarily with only two interactions were observed for both DR and RD patterns. In case of both the motifs the hydrogen bonding was found to be conserved through the reversal of residue positions. While motifs in the helix group for both patterns were found to possess Type D hydrogen bonding, in case of those in irregular structures group the bonding was Type B. Distinctively the change observed was for the Arg and Asp side chain conformations, which for helix motifs in DR was found to have the partially folded g- g- t t conformation was on reversal of the residue positions assumed a more extended conformation t t g+ t (Figure 45a,b). A similar pattern was observed for the motifs in the irregular structures group wherein the partially folded conformation of g+ t g+ t changed to a more extended conformation t

t g+ t. The Asp side chain conformation for the helix motifs in both DR and RD though remained in gauche conformation, an evident change of g+ in DR to g- for RD was observed. However, for the motifs in irregular structures where Asp χ_1 assumed a g+ conformation, the residue position reversal required a more extended conformation t for the interaction amongst the residues to occur (Figure 45c,d).

3.3.3. Comparing the RD and R-X-D motifs.

The motifs RD and R-X-D were compared to gain a better understanding of the effect of introduction of spacer residue between Arg and Asp (Table 4). The preference of the RD motifs was for the helix group in the presence of interactions, however, when a spacer X residue was present the preference shifted to irregular structure and β-sheet.

In both the cases of RD and R-X-D motifs those with two interactions were mainly observed. For RD motifs in sheets, two side chain – main chain interactions viz. Arg (NE) – (O) Asp and Arg (NH1) – (O) Asp were observed since the Arg and Asp side chains were found to lie on opposite sides of the peptide plane (Figure 46a,b).

In this case the Arg residues assumed a partially folded side chain conformation g+ t g- t while the Asp χ_1 was observed to be g-. In the presence of a spacer residue between Arg and Asp both the side chains were found to lie on the same side of peptide plane. This allowed for the side chains to interact amongst themselves forming Type B hydrogen bonding. Here the Arg side chain assumed a completely planar extended conformation t t t t while the Asp conformation was g-. For the motifs in the irregular structures group, a slightly variant effect of the intervening residue was observed. While the hydrogen bonding of Type B remained same in both, the RD motifs in irregular structures assumed a more extended conformation t t g+ t for Arg which in case of the R-X-D motifs of the irregular structures group was in a highly folded state g- g- g- g- to allow for interaction amongst the side chains. The Asp χ_1 was g+ instead of g- (Figure 46c,d).

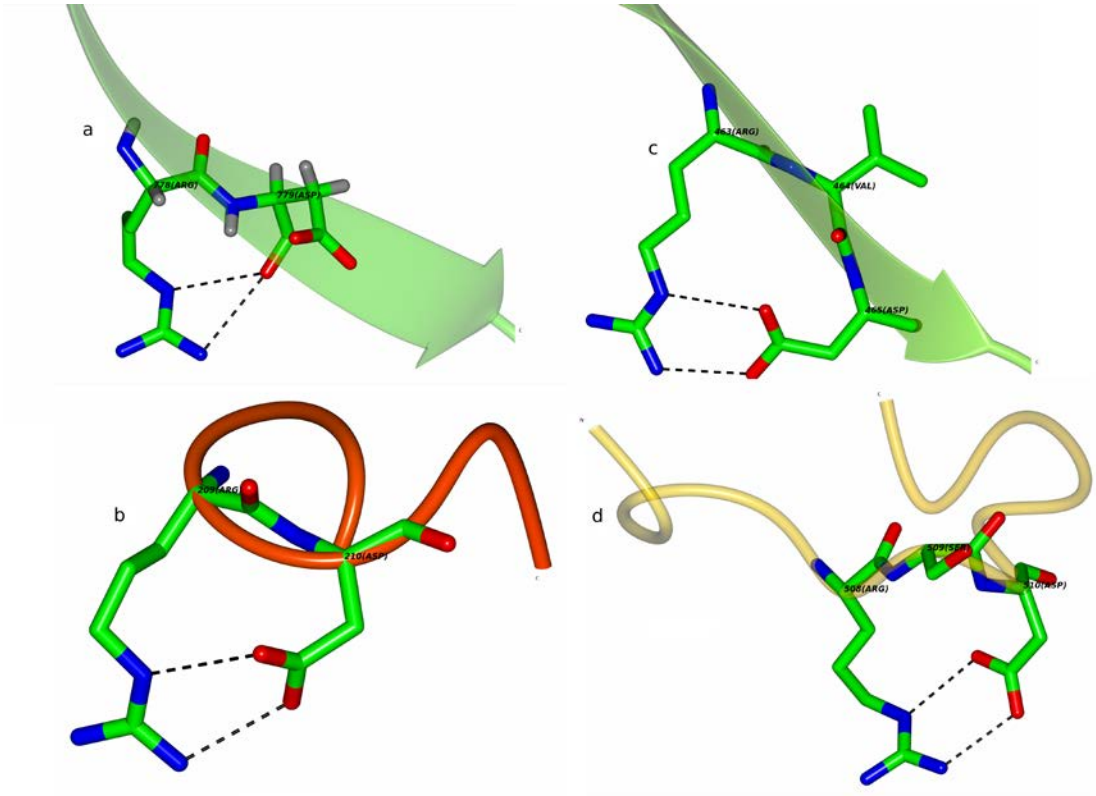

Figure 46. (a) The RD motif in 3PSF occurring in sheets group with g+ t g- t Arg side chain conformation and two side chain – main chain hydrogen bonds, namely, ARG (NE) – (O) ASP and ARG (NH1) – (O) ASP. (b) The RD motif in 4KFG occurring in irregular structures group with t t g+ t Arg side chain conformation and Type B hydrogen bonding interaction. (c) The R-V-D motif in 3BI1occurring in sheets group with t t t t Arg side chain conformation and Type B hydrogen bonding interaction. (d) The R-S-D motif in 4A57occurring in irregular structures group with g- g- g- g- Arg side chain conformation and Type B hydrogen bonding interaction.

3.3.4. Comparing the D-X-R and R-X-D motifs.

Comparative analysis of D-X-R and R-X-D highlighted the effects of both the presence of the spacer residue as well as the reversal of charged residue positions (Table 4). Both motifs D-X-R and R-X-D show low presence in regular secondary structures while shows considerably high occurrence in irregular structures. Presence of both motifs, with interactions, in irregular structures was also noted. The analysis carried out to estimate the preference of amino acids for the X position in D-X-R and R-X-D motifs has been compared below.

Apart from a comparatively high occurrence of Gly in D-X-R, at the X position there is no other significant difference. The non-polar uncharged amino acids such as Leu, Ala and Val were found in significant proportions along-with Glutamic acid, Isoleucine and Proline occurring at this position. In case of R-X-D motifs, high

occurrence of aliphatic amino acids such as Leu, Val and Ile was observed. Charged amino acids such as Glu and Lys along with Ala, Gly and Pro were found in considerable numbers at this position. Surprisingly, the aromatic amino acid Phe was equally preferred at this position (Figure 47).

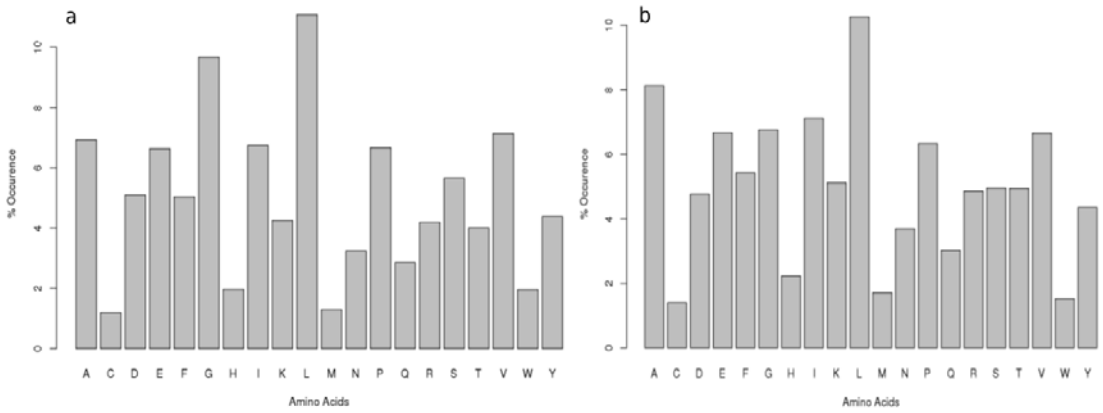

Figure 47. The frequency bar plot of the occurrence of all 20 amino acids at the X position in the D-X-R and R-X-D motifs.

In both D-X-R and R-X-D motifs belonging to sheet structure the hydrogen bonding was Type B. Similarly, for D-X-R and R-X-D motifs in irregular structures though Type B bonding was found conserved, D-X-R motifs with three H-bonds showed an additional Arg (N) – (OD1) Asp hydrogen bond. D-X-R motifs in sheets were observed to have a nearly planar conformation g- t t t for Arg, which in the reversed Asp and Arg residue positions viz. R-X-D, was a completely extended and planar Arg conformation t t t t. The Asp χ_1 was found to change from t to g- conformation. The D-X-R motifs in irregular structures, which assumed a partially folded conformation g- t g- t with Asp χ_1 as t, was found to change to a completely folded g- g- g- g- with Asp χ_1 also assuming a g+ conformation in R-X-D. However the superimposition analysis carried out for both D-X-R motifs with conformation g- t g- t and R-X-D motifs with conformation g- g- g- g- revealed the backbone to be nearly the same (Figure 48a-d).

With the spacer X residue being present, it was observed that the highly folded side chains of both Arg and Asp were critical for the hydrogen bonding interaction amongst them to occur. The R-X-D motif in sheets was found to be slightly widened with the average C_α (D/R)…C_α (R/D) distance being incremented from 6.502 Å for D-X-R to 6.679 Å.

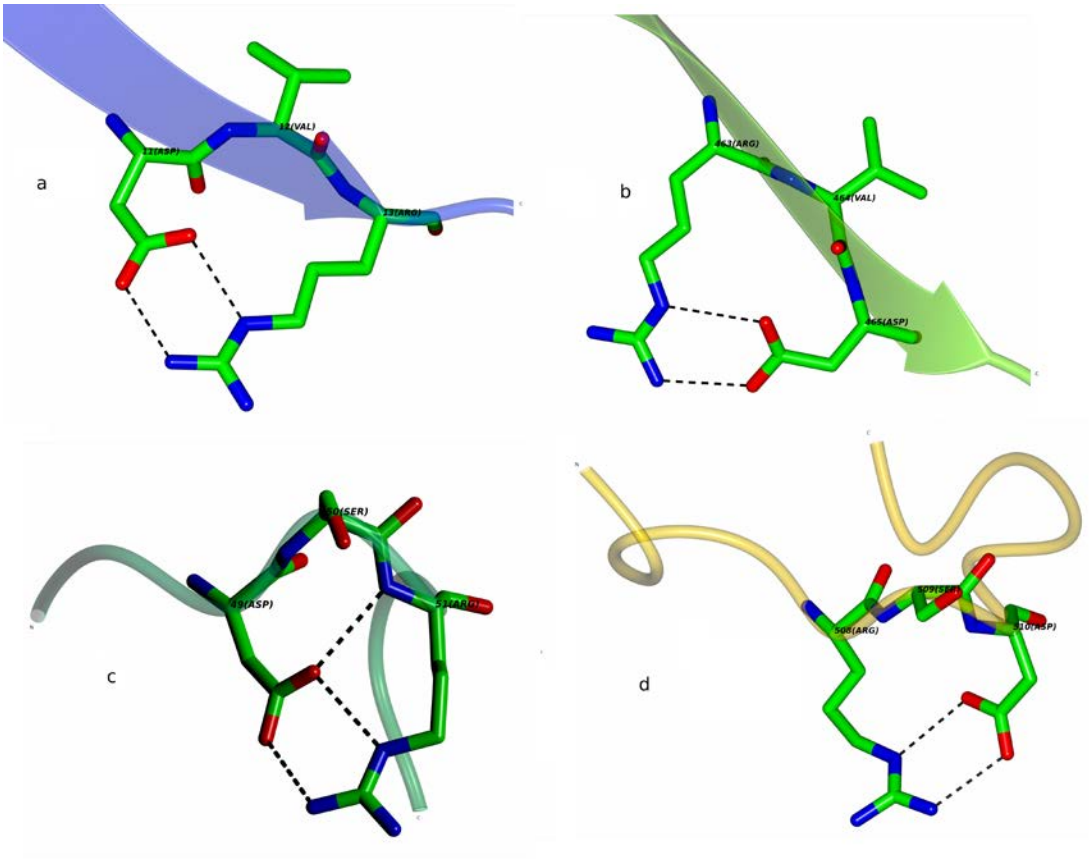

Figure 48. (a) The D-V-R motif in 2HXT occurring in sheet structure with g- t t t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The R-V-D motif in 3BI1 occurring in sheets group with t t t t Arg side chain conformation and Type B hydrogen bonding interaction. (c) The D-S-R motif in 1C7K occurring in irregular structures with g- t g- t Arg side chain conformation and Type B hydrogen bonding interaction. (d) The R-S-D motif in 4A57 occurring in irregular structures with g- g- g- g- Arg side chain conformation and Type B hydrogen bonding interaction.

3.4. Summary

Based on the initial observations, a detailed analysis was carried out for short structural motifs involving Asp and Arg/Lys. The motifs were analysed for the localized fold assumed by the residues and the interactions involved. The importance of the direction of the residues in the sequence was probed by first studying the motif residues Asp and Arg in N-terminal to C-terminal and then by reversing the positions of resides and again studying them in N-terminal to C-terminal direction. Next the role of a spacer residue was probed by analyzing motifs containing a residue in between Asp and Arg/Lys and then by reversing the Asp and Arg/Lys residue positions. All motifs were found to occur in large numbers in irregular structures both

with and without interactions. In case of motifs in helices significant numbers with interactions were observed for D-R, R-D, D-K and K-D motifs only. Motifs with interactions occurring in sheets were observed in substantial numbers only in D-X-R and R-X-D motifs. Motifs involving Asp and Arg involving two or more hydrogen bonds were studied in detail. In the case of Lys even those with one hydrogen bond were analyzed because Lys rarely formed parallel or divergent hydrogen bonds with the same carboxylate group like Arg did.

Chapter 4

Identification of E-R, R-E, E-X-R, R-X-E, E-K, K-E, E-X-K and K-X-E motifs, analysis of their conformations and interactions

This chapter deals with the structural analysis of short motifs involving the negatively charged amino acid Glutamic acid and positively charged sequence neighbour Arginine or Lysine. The detected motifs were analyzed for their occurrence in secondary structures; the localized fold and the hydrogen bonding interactions involving the motif residues were studied in detail in irregular structures.

4.1 Structural Analysis of Motifs involving Glu and Arg/Lys.

Motifs involving Glu with either Arg or Lys as neighboring amino acids in the sequence next to it or separated by a single spacer residue were studied. The resulting eight motifs were analyzed in their secondary structures where they can additionally stabilize the fold by side chain interaction. When present in the irregular structural fold they were analyzed for the presence as well as pattern of hydrogen-bonded interactions amongst these residues and local conformation resulting in specific folds found repeated in several unrelated proteins.

4.1.1 Analysis of motifs involving Glu and Arg.

The next set involving amino acids Glu and Arg analyzed comprised of four patterns namely E-R, E-X-R, R-E and R-X-E (Table 1,2). 12533 occurrences of the E-R motif were recorded from the local dataset of structures of unrelated proteins.

Most of the identified E-R motifs were found to belong to the helix group. While 1107 motifs were found with interactions, of these 96 were found to involve two or more hydrogen bonds. With a reversal of the order of Glu and Arg residues, 12132 R-E motifs were recorded. A significant increase in the number of motifs in the helix group was also observed. 1034 motifs of R-E type were observed with interactions.

For the E-X-R motif with a spacer residue between Glu and Arg, a total of 10991 occurrences were detected. Total 1789 motifs were observed with hydrogen bond interactions of which 492 were found to have 2 or more H-bonds. 268 motifs were observed in β-sheets with two interactions and 155 in irregular structures. 63 occurrences were noted in irregular structures with three interactions. With an intervening residue, lesser number of R-X-E motifs i.e. 10761 was observed.

Table 1. The number of occurrences of different types of E-R, R-E, E-X-R, R-X-E, E-K, K-E, E-X-K, K-X-E motif present in the local PDB dataset. Numbers in brackets indicate the total occurrence of the motif in the dataset.

Pattern (Total)	Helix		Sheets		Irregular Regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
E-R (12533)	592	6536	137	1536	378	3374
Total	7128		1673		3752	
R-E (12132)	656	6597	45	1016	333	3484
Total	7253		1061		3817	
E-X-R (10991)	354	4781	680	1289	756	3132
Total	5135		1969		3888	
R-X-E (10761)	117	3994	533	1292	1174	3651
Total	4111		1825		4825	
E-K (15756)	422	8350	19	1110	377	5478
Total	8772		1129		5855	
K-E (15195)	540	8137	26	1030	405	5057
Total	8677		1056		5462	
E-X-K (12449)	253	4601	137	1730	634	5093
Total	4854		1867		5727	
K-X-E (11622)	56	3753	121	1436	753	5503
Total	3809		1557		6256	

Table 2. The number of occurrences with and without interactions for the three secondary structure classes of E-R, R-E, E-X-R, R-X-E, E-K, K-E, E-X-K, K-X-E motifs present in the local PDB dataset.

Motif	Helix		Sheet		Irregular		Total with ints.
	Interactions (ints.)		Interactions (ints.)		Interactions (ints.)		
	1	>1	1	>1	1	>1	
E-R	535	57	130	7	346	32	1107
Total	592		137		378		
R-E	541	115	42	3	296	37	1034
Total	656		45		333		
E-X-R	349	5	407	273	538	218	1790
Total	354		680		756		
R-X-E	117	0	356	177	851	323	1824
Total	117		533		1174		
E-K	410	32	19	0	364	13	818
Total	422		19		377		
K-E	525	15	25	1	388	17	971
Total	540		26		405		
E-X-K	253	0	135	2	612	22	1024
Total	253		137		634		
K-X-E	56	0	121	0	712	41	930
Total	56		121		753		

The equitable preference for helices and irregular structures was found for the R-X-E motif. While 1824 motifs were found with interaction, most of them in irregular structure, 429 were found to have two interactions and 71 with three interactions. 70 motifs with three interactions were observed in irregular structure alone.

4.1.2 Analysis of motifs involving Glu and Lys.

This set analyzed comprised of four patterns namely E-K, E-X-K, K-E and K-X-E (Table 1, 2). Total 15756 occurrences of the E-K motif were detected. The highest

occurrence of E-K motifs was observed in helices. From the 818 motifs detected with interactions, merely 25 were observed with two interactions. With the reversed order of motif residues, 15195 occurrences of the K-E motif were observed. With the reversed sequence, a significant increase in the occurrence of the motif in helices was observed. While 971 occurrences were observed with interactions, only 24 were found to have two or more interactions.

With a spacer residue between Glu and Lys, 12449 occurrences of E-X-K were recorded. Nearly equal numbers of motifs were observed in both helices as well as irregular structural regions. While 1000 motifs were observed with one interaction, 23 were found to involve two interactions and only one was observed with three interactions. With a spacer residue in between and reversal of the positions of Glu and Lys, 11622 occurrences were recorded for the K-X-E motif. It was observed that the spacer residue in between leads to higher occurrence of the motif in the irregular structures. While a total 930 motifs with interactions were observed, only 41 were found to have two interactions and all these were in the irregular structures.

Motifs involving Glu and Arg or Lys are shown to have a preference for regular secondary structure such as helices. Only in case of the motif E-X-K and K-X-E this preference was found to shift to the irregular structure regions. Based on the preliminary analysis, it was observed that a detailed analysis of motifs involving Glu and Arg/Lys would be useful.

4.2 Detailed Analysis of motifs involving Glu and Arg/Lys.

Extensive analysis of the highlighted motifs belonging to E-R, E-X-R, R-E, R-X-E, EK, KE, EXK and KXE was carried out (Table 3, 4). The analysis focused on the side chain conformations of Glu and Arg/Lys as well as the hydrogen bonding interaction observed involving the motif residues.

Table 3. The number of occurrences with 2 and 3 interactions for the three secondary structure classes of E-R, R-E, E-X-R, R-X-E motifs present in the local PDB dataset. Those occurring more than 100 times in unrelated proteins have been further analyzed for patterns of hydrogen bonds and side chain conformations.

Motif	Helix		Sheet		Irregular Regions		Total
(Total)	Interactions (ints.)		Interactions (ints.)		Interactions (ints.)		with
	2	3	2	3	2	3	ints.
E-R	55	2	7	0	30	2	96
Total	57		7		32		
R-E	114	1	3	0	12	25	155
Total	115		3		37		
E-X-R	5	0	2	271	155	63	496
Total	5		273		218		
R-X-E	0	0	176	1	253	70	500
Total	0		177		323		

Table 4. The number of occurrences with 1 and 2 interactions for the three secondary structure classes of E-K, K-E, E-X-K, K-X-E motifs present in the local PDB dataset. Those occurring more than 100 times in unrelated proteins have been further analyzed for patterns of hydrogen bonds and side chain conformations.

Motif	Helix		Sheet		Irregular		Total
(Total)	Interactions (ints.)		Interactions (ints.)		Interactions (ints.)		with
	1	2	1	2	1	2	ints.
E-K	410	12	19	0	364	13	818
Total	422		19		377		
K-E	525	15	25	1	388	17	971
Total	540		26		405		
E-X-K	253	0	135	2	612	22	1024
Total	253		137		634		
K-X-E	56	0	121	0	712	41	930
Total	56		121		753		

4.2.1 Analysis of motifs in helices.

4.2.1.1 Analysis of R-E motif with two hydrogen bonds in helices.

The $C_\beta - C_\alpha - C_\alpha - C_\beta$ virtual torsion angle distribution for motifs of this set was found to lie largely in the range of 30-70° (Figure 1).

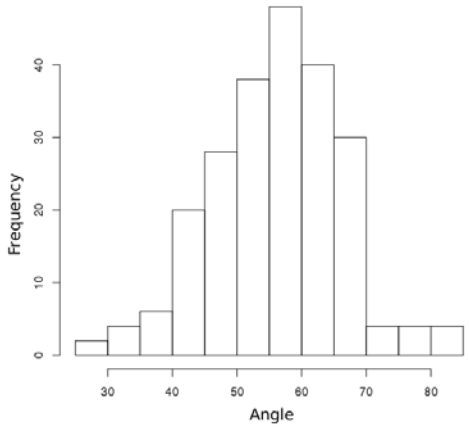

Figure 1. Histogram showing the distribution of the C_β-C_β virtual torsion angle of the motifs.

Based on the analysis of Arg side chain conformations, three major conformations were identified of which, the extended conformation t t g+ t was found to be the most abundant covering 58% of the total (Figure 2).

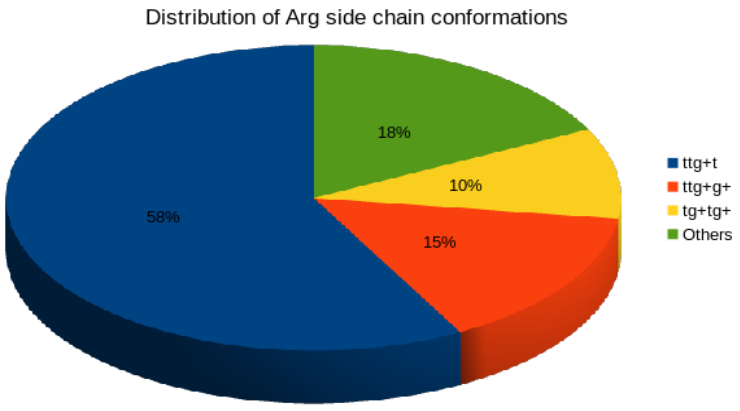

Figure 2. The distribution of Arg side chain conformation in R-E motif with two interactions in helices.

While the Glu χ_1 was found to be exclusively g-, while the χ_2 was observed to be mostly g- in 61 cases with the conformation being g+ in 5 cases. The hydrogen bonding in all cases was found to be type D (Figure 3).

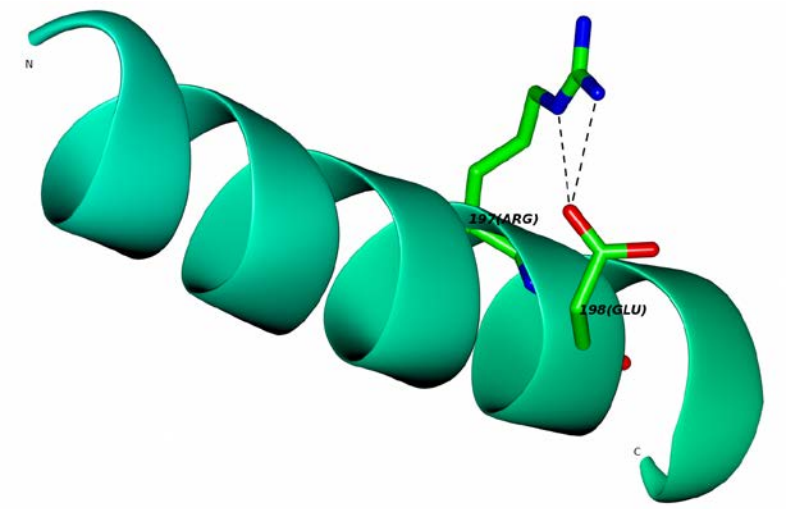

Figure 3. The R-E motif in 3C2Q occurring in helix group with t t g+ t Arg side chain conformation and Type D hydrogen bonding interaction.

4.2.1.2 Analysis of E-K motif with one hydrogen bond in helices.

410 occurrences of the E-K with one hydrogen bonded interaction were observed to occur in helices. While a wide range of side chain conformations was identified for the Lys residue, the nearly extended conformation g- t t t was found in 116 motifs covering 28% of the total (Figure 4).

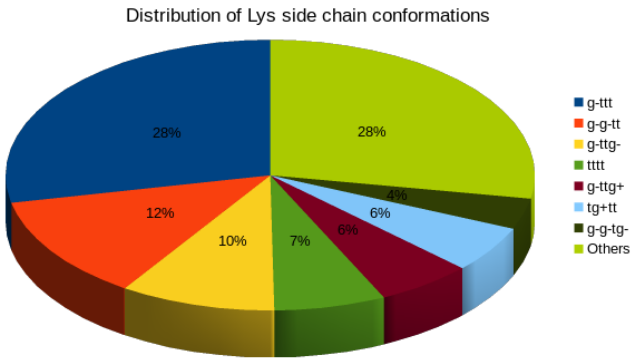

Figure 4. The distribution of Lys side chain conformation in E-K motif with one interaction in helices.

The Glu side chain χ_1 conformation was observed to be g+ for all while the χ_2 conformation was found to be g-. While for 115 motifs the hydrogen bonding interaction was found to be main chain - side chain in 4BJM:A83 which occurs at the C-terminal end of a helix the hydrogen bond was like side chain – side chain, Lys (NZ) - (OE1) Glu (Figure 5).

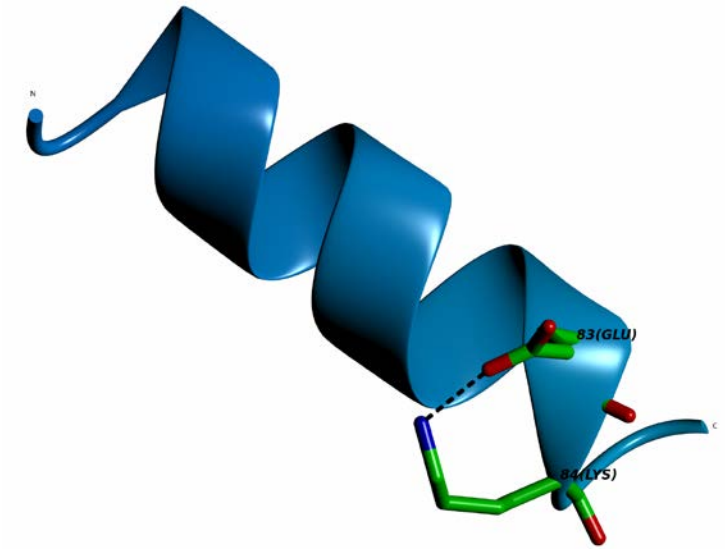

Figure 5. The E-K motif in 4BJM occurring in helix with g- t t t Lys side chain conformation and one side chain – side chain hydrogen bonding interaction.

4.2.1.3 Analysis of K-E motif with one hydrogen bond in helices.

Although 525 motifs of K-E pattern were found to belong to helices having one interaction, 409 showed main chain – side chain H-bond, while 107 showed side chain – side chain interaction and 7 had side chain – main chain interaction (Lys (NZ) – (O) Lys). The Lys side chain showed the completely extended conformation t t t t to occur in 80 occurrences (15%) as the highest with a wide variety of conformations (rotamers) occurring (Figure 6). The Glu side chain χ_1 conformation was observed to be g- for all while the χ_2 conformation was found to be g+.

For this set in 72 motifs the hydrogen bonding was main chain – side chain (Glu (N) – (OE1) Glu) while in 8 cases the bond was side chain – side chain, Lys (NZ) - (OE2) Glu (Figure 7). In all these 8 cases the Glu χ_2 conformation was found to be g-.

Again as observed in E-K motifs occurring in helices these motifs were found to occur at the C terminal end of helices

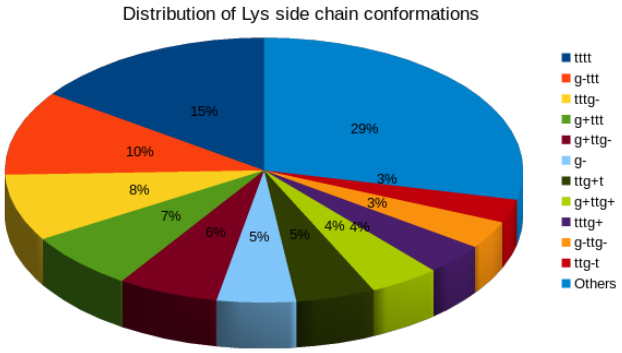

Figure 6. The distribution of Lys side chain conformation in K-E motif with one interaction in helices.

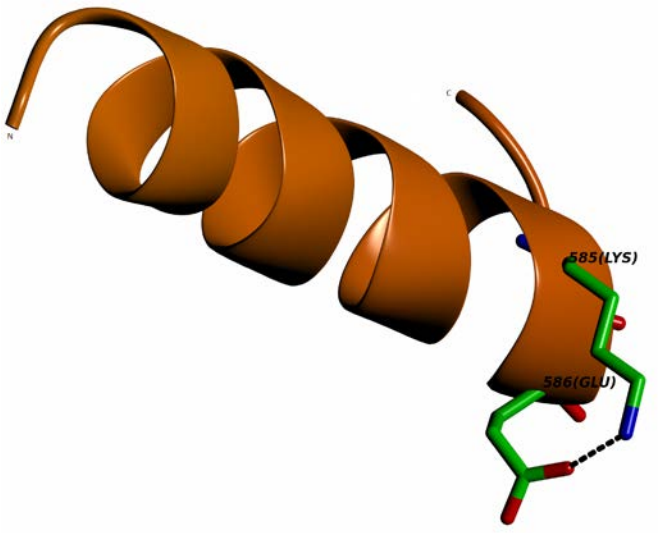

Figure 7. The K-E motif in 1PG4 occurring in helix with t t t t Lys side chain conformation and one side chain – side chain hydrogen bonding interaction.

4.2.1.4 Analysis of E-X-K motif with one hydrogen bond in helices.

The 253 motifs observed in helices were studied overall for the occurrences of hydrogen bond. In 246 cases one main chain – side chain bond was found while in 5 cases it was side chain – main chain and side chain – side chain in only two motifs. While a large variety of conformations was observed for the Lys side chain (Figure 8), the conformation g- t t t was observed in 64 cases (25%). For this set the Glu χ_1

conformation was g- while the Glu χ_2 conformation was observed to be g+. In all motifs the hydrogen bond was main chain – side chain bond (Figure 9).

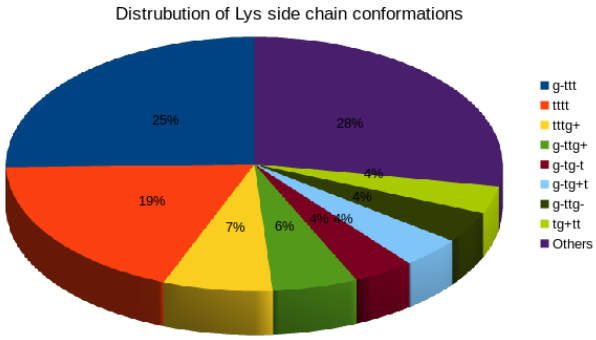

Figure 8. The distribution of Lys side chain conformation in E-X-K motif with one interaction in helices.

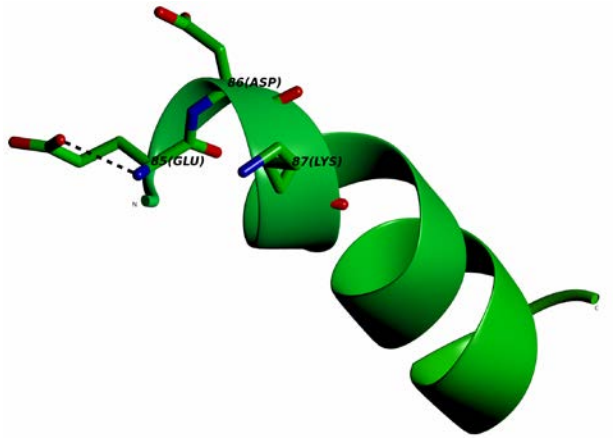

Figure 9. The E-D-K motif in 1FVK occurring in distorted helix with g- t t t Lys side chain conformation and one main chain – side chain hydrogen bonding interaction.

4.2.2 Analysis of motifs in sheets.

4.2.2.1 Analysis of E-X-R motif with two hydrogen bonds in sheets.

Total 268 occurrences of the E-X-R motif with two interactions in sheets were observed from the selected local dataset of unrelated proteins. Analysis of the Arg side chain conformation revealed that they assumed a wide range of conformations ranging from a highly folded state to completely extended one. However, the nearly extended side

chain conformation g- t t t was observed to be the most abundant one covering 50% of the total (Figure 10a).

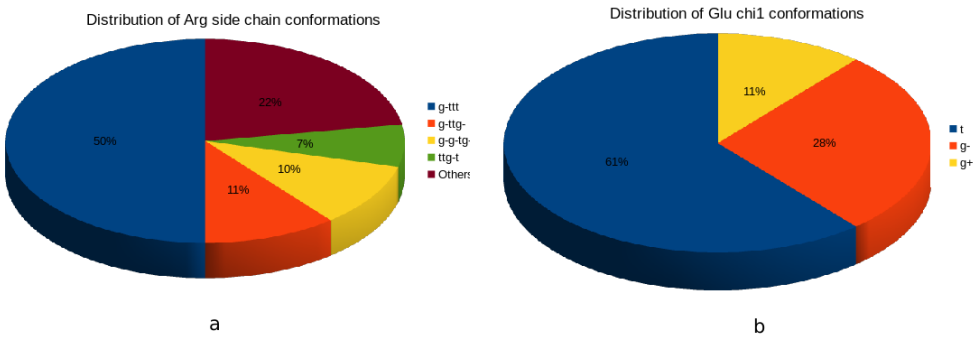

Figure 10. (a) The distribution of Arg side chain conformation in E-X-R motif with two interactions in sheets. (b) The distribution of Glu χ_1 conformations in E-X-R motif with two interactions in sheets and Arg conformation g- t t t.

The Glu χ_1 was found to have the extended conformation t in 82 cases while in 37 cases it assumed the folded conformation g- and g+ in the remaining 15 cases (Figure 10b). The Asp χ_2 was however found to be dominantly t in all cases except 1J9L:A188 where it was found to be g+. Analysis of the residues at the X position in motifs revealed Ile, Leu and Val to occur more (Figure 14).

Analyzing the hydrogen bonding in the motifs, revealed all motifs to have two side chain – side chain hydrogen bonds. 112 motifs were found to have Type D bonding (Figure 11) whereas 22 were observed having Type B bonding.

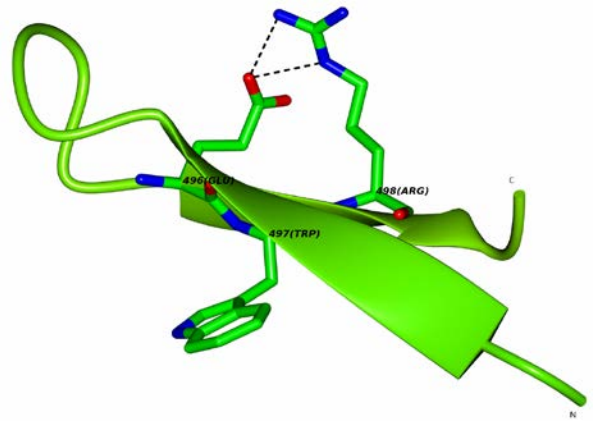

Figure 11. The E-W-R motif in 3WN7 occurring in sheets with g- t t t Arg side chain conformation and Type D hydrogen bonding interaction.

Motifs with the partially extended Arg side chain conformation g- t t g- were found to contribute 11% of the total. Here the Glu χ_1 was found to have g- conformation in 19 cases while in 10 cases it assumed the extended conformation t. The Glu χ_2 was found to be t for all motifs except 2YMZ:A68 where it was found to be g-. In 22 cases the hydrogen bonding was Type D (Figure 12) and only in case of 1R5M:A243 the bonding was Type B. In case of 2YMZ:A68, a variant of Type B was observed wherein one side chain – side chain bond was found to be replaced by a main chain – side chain bond while the other side chain – side chain bond Arg (NE) – (OE2) Glu remained.

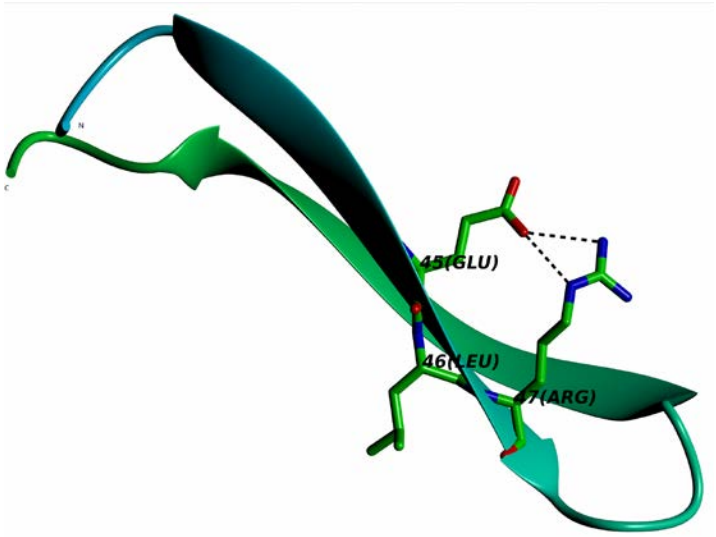

Figure 12. The E-L-R motif in 4ML5 occurring in sheets with g- t t g- Arg side chain conformation and Type D hydrogen bonding interaction.

4.2.2.2 Analysis of R-X-E motif with two hydrogen bonds in sheets.

176 motifs were detected with two interactions belonging to the sheets group. 20 different Arg side chain conformations were observed for the motifs belonging to this group.

The extended conformations t t t g+ and t t t t were found to cover more than 40% of the total. The most favorable conformation was found to be t t t g+ (26%) (Figure 13a). The Glu χ_1 was found to be g- in 37 cases and t and g+ in 3 cases each (Figure 13b). The Glu χ_2 was t in 42 cases and g- in 3LUQ:A159. The Ramachandran plot of X residue was observed to lie as single cluster. In terms of the hydrogen bonding 40 occurrences were

found to show Type D while only three had Type B (Figure 14a). Two notable variations observed were for 2CFE:A153 wherein the motif was at the N-terminal end of a strand (Figure 14b) and 4FW9:A383 where the motifs belonged to a distorted sheet.

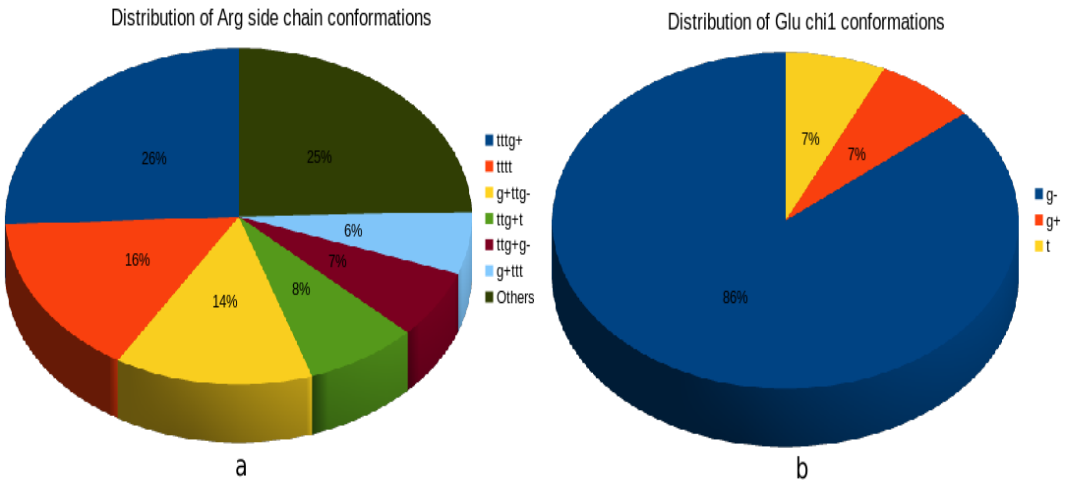

Figure 13. (a) The distribution of Arg side chain conformation in R-X-E motif with two interactions in sheets. (b) The distribution of Glu χ_1 conformations in R-X-E motif with two interactions in sheets with Arg side chain t t t g+.

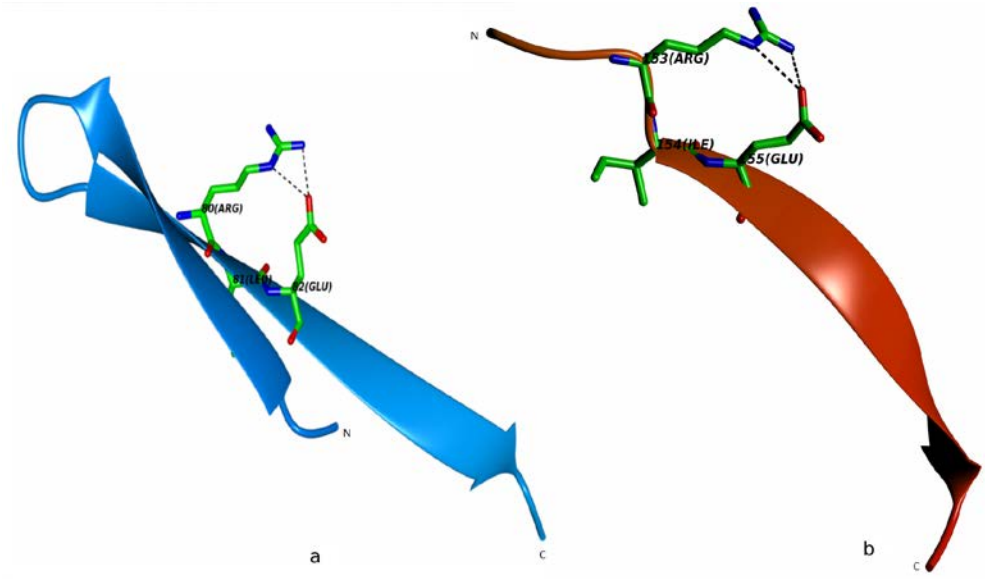

Figure 14. (a) The R-L-E motif in 2D5W occurring in sheets with t t t g+ Arg side chain conformation and Type D hydrogen bonding interaction. (b) The R-I-E motif in 2CFE occurring in sheets with t t t g+ Arg side chain conformation and Type D hydrogen bonding interaction.

Next, the fully extended conformation t t t t was studied that covered 16% of the total (Figure 16a). In 22 cases the Glu χ_1 was found to be g- and g+ and t in two cases each (Figure 15a). The Glu χ_2 also showed significant variation with 16 cases having the χ_2 conformation t, 9 with g- and in case of 3PFB:A17, the conformation was g+ (Figure 15b).

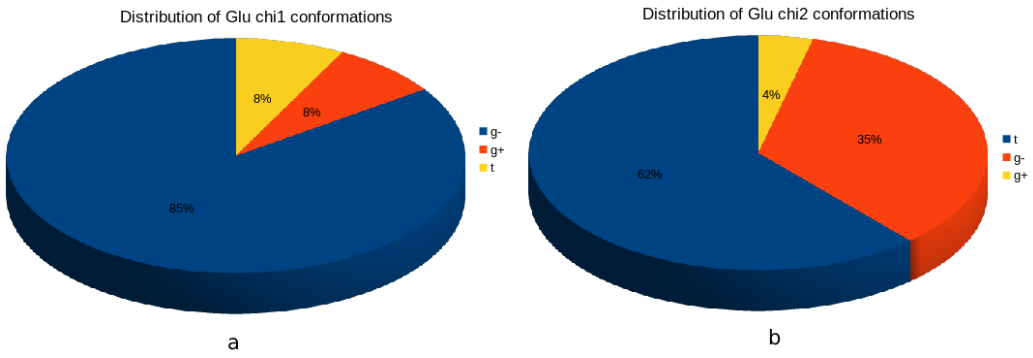

Figure 15. (a) The distribution of Glu χ_1 conformations in R-X-E motif with two interactions in sheets with Arg side chain t t t t. (b) The distribution of Glu χ_2 conformations in R-X-E motif with two interactions in sheets with Arg side chain t t t t.

As can be expected the Ramachandran plot of X residue showed a single cluster in the sheet region. In 25 cases the hydrogen bonding was Type D (Figure 16a) while in case of 2R16:A775 it was observed to be Type B (Figure 16b). In case of 3PFB:A17 where the Glu χ_1 was t, the Glu χ_2 was g+ as mentioned before; the X residue was Glu.

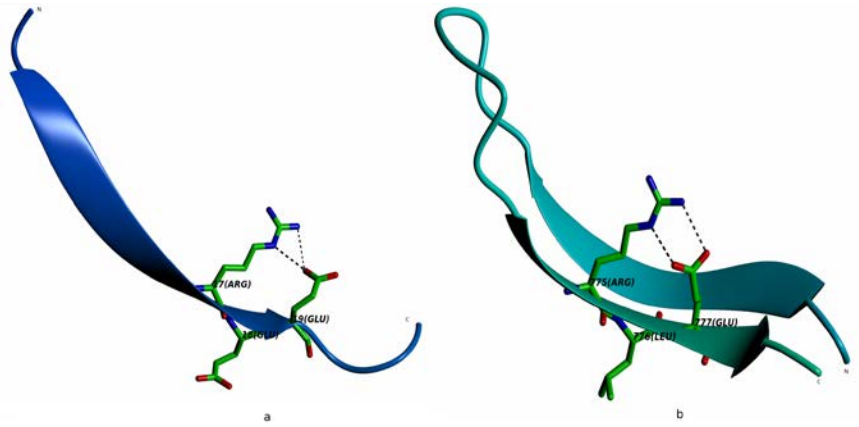

Figure 16. (a) The R-E-E motif in 3PFB occurring in sheets with t t t t Arg side chain conformation and Type D hydrogen bonding interaction. (b) The R-L-E motif in 2R16 occurring in sheets with t t t t Arg side chain conformation and Type B hydrogen bonding interaction.

4.2.2.3 Analysis of E-X-K motif with one hydrogen bond in sheets.

E-X-K motifs with one interaction in sheets were found to have 135 occurrences. In 116 motifs the hydrogen bonding was side chain – side chain, main chain –side chain in 17 and side chain – main chain in 2. From the very wide set of conformations observed for the Lys side chain, the partially folded conformation g- g- t t was found to occur in 33 (24%) motifs (Figure 17). The Glu χ_1 and χ_2 conformations both were t. The Ramachandran plot for the X residues showed all occurrences in the sheet region. In 29 occurrences the hydrogen bonding was side chain – side chain (Figure 18), while rest four it was main chain – side chain.

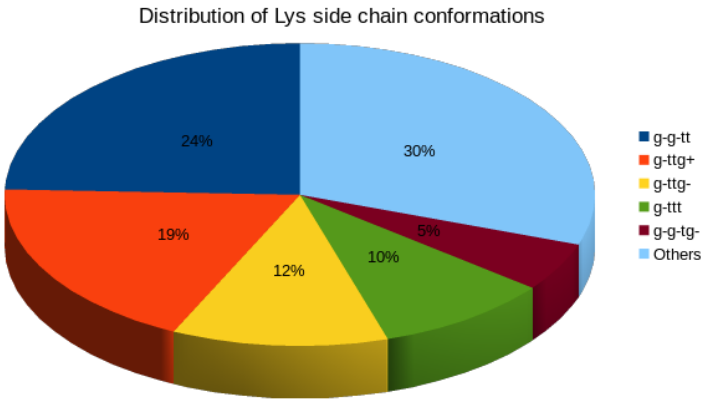

Figure 17. The distribution of Lys side chain conformation in E-X-K motif with one interaction in sheets.

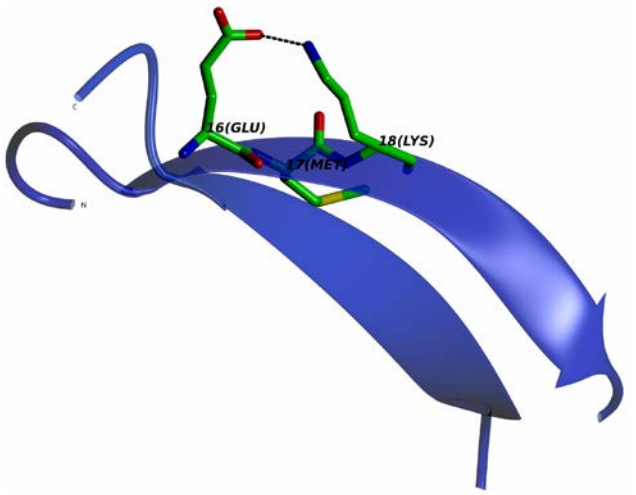

Figure 18. The E-M-K motif in 4KIK occurring in sheets with one hydrogen bonding interaction.

4.2.2.4 Analysis of K-X-E motif with one hydrogen bond in sheets.

The 121 occurrences were studied for the occurrence of hydrogen bonds. In 111 motifs side chain – side chain hydrogen bonds were observed, 7 had main chain - side chain and 3 had main chain – main chain (Glu (N) – (O) Lys). The analysis of side chain conformations for the Lys residue revealed the extended conformation t t t t to occur in 31 (26%) motifs (Figure 19a). The Glu χ_1 conformation was g- and χ_2 was t.

The hydrogen bonding in 30 motifs was side chain – side chain (Figure 19b), only in one it was main chain – side chain. In all these cases the side chains of K and E residues were on the same side of the strand allowing the side chain interaction.

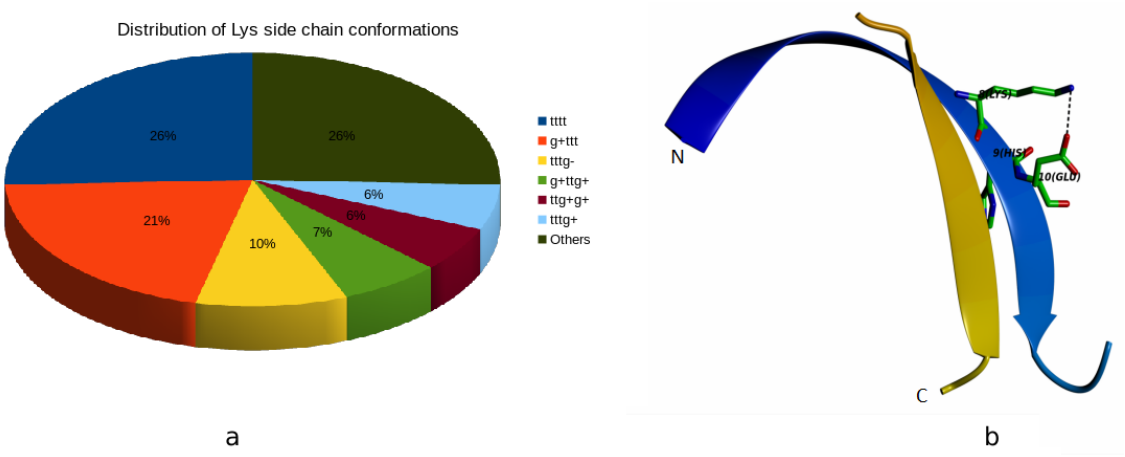

Figure 19. (a) The distribution of Lys side chain conformation in K-X-E motif with one interaction in sheets. (b) The K-H-E motif in 1BTN occurring in sheets with one hydrogen bonding interaction.

4.2.3 Analysis of motifs in irregular structures.

4.2.3.1 Analysis of E-X-R motif with two hydrogen bonds in irregular structures.

E-X-R motif with two interactions in the irregular structural regions showed 155 occurrences. The superimposition analysis of these motifs showed no conservation of the backbone conformation (Figure 20). About 35 different conformations for the Arg side chains were observed in this set. The partially folded conformation g- t g- t was found to

cover 14% of the data for this set (Figure 21a). The Glu χ_1 and χ_2 conformations also showed a wide variation for the set (Figure 21b-c).

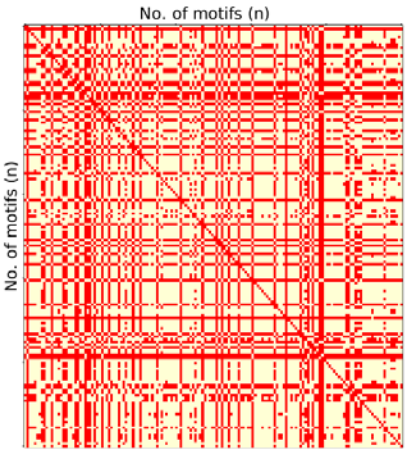

Figure 20. The pair-wise superimposition graph of the E-X-R motifs in irregular structural regions with two interactions. A cut-off of 0.3Å was used.

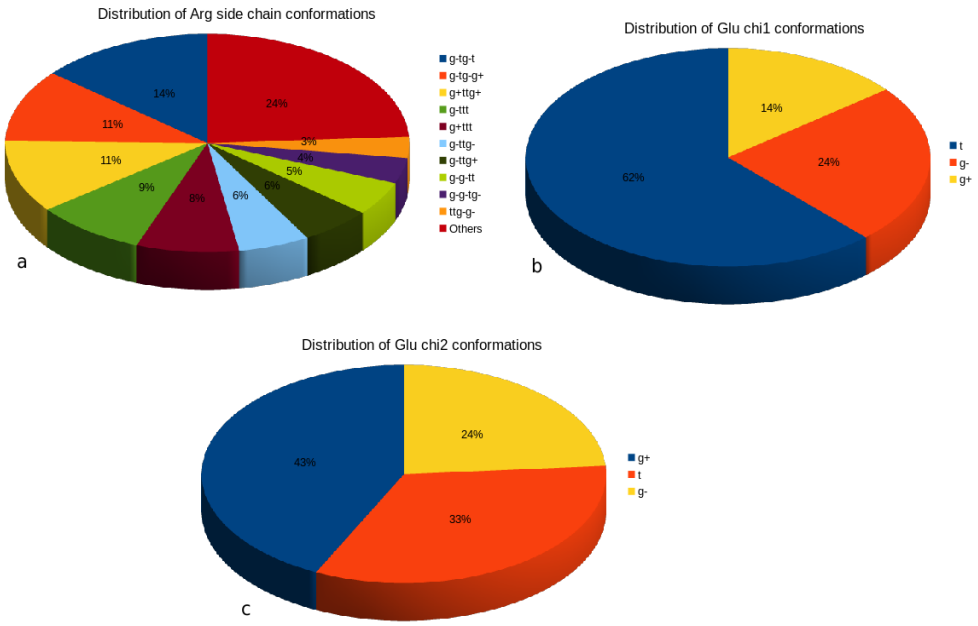

Figure 21. (a) The distribution of Arg side chain conformation in E-X-R motif with two interactions in irregular structural regions. (b) The distribution of Glu χ_1 conformations in E-X-R motif with two interactions in irregular structural regions and Arg side chain conformation g- t g- t. (c) The distribution of Glu χ_2 conformations in E-X-R motif with two interactions in irregular structural regions and Arg side chain conformation g- t g- t.

The Ramachandran plot of the X residues of the motifs in this set shows three clusters (Figure 22). For those in the left-handed helical region the X residue was found to be Gly. In case of the three variations, namely 2Z1D:A248, 3UD1:A1135 and 4H3W:A214, the motif residues were found to lie in an unstructured strand, at the interface of a strand and turn, respectively.

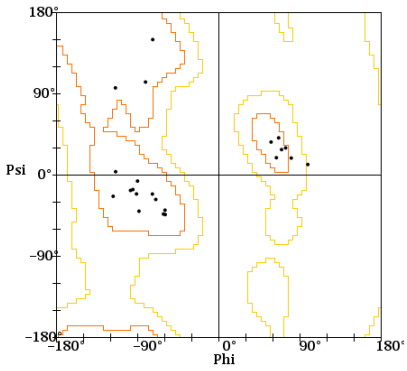

Figure 22. Ramachandran plot for residues in the X position in E-X-R motifs belonging to irregular structural regions with two interactions.

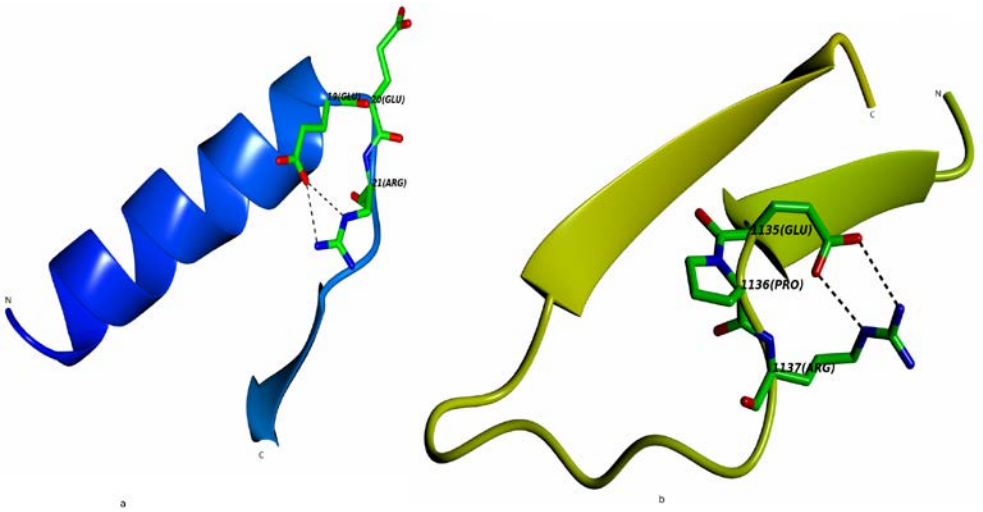

Figure 23. (a) The E-E-R motif in 1HRU occurring in irregular structural regions with g- t g- t Arg side chain conformation and Type D hydrogen bonding interaction. (b) The E-P-R motif in 3UD1 lying at the interface of a strand and a wide turn occurring in irregular structural regions with g- t g- t Arg side chain conformation and Type B hydrogen bonding interaction.

The hydrogen bonding in these motifs was also observed to show a high variance. In 8 cases the bonding was Type D (Figure 23a) while in two it was Type B (Figure 23b). In 10 cases one side chain – side chain H-bond observed before was substituted by a main chain – side chain bond. Only in two cases, both interactions were found to be side chain – main chain i.e. Arg (NE) – (O) Glu and Arg (NH2) – (O) Glu.

4.2.3.2 Analysis of R-X-E motif with two hydrogen bonds in irregular structural regions.

Total 253 RXE motifs detected with two interactions in the irregular structural regions. The superimposition analysis for the motifs shows the backbone having significant variation (Figure 24). From the 39 variant conformations, 19% constituted the partially folded Arg side chain conformation g- t g+ g+ (Figure 25a). Both Glu χ_1 and Glu χ_2 showed considerable variation. Glu χ_1 was found to be g+ in 22 cases, g- in 22 and t in 2 cases. The Asp χ_2 was observed to be t in 24 cases, g- in 21 and in 4 cases, t (Figure 25b-c).

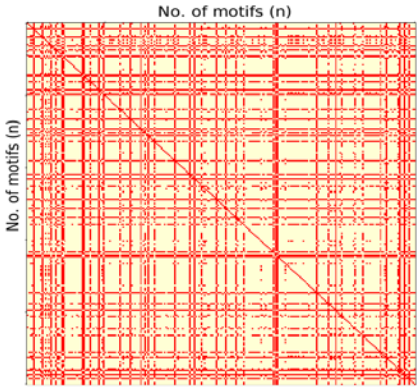

Figure 24. The pair-wise superimposition graph of the R-X-E motifs in irregular structural regions with two interactions using a cut-off of 0.3Å.

The Ramachandran plot for the X residue was found to be divided into 2 clusters with 4 variations. (Figure 26). In case of the variations, for 1H2W:A11, the motif was found to lie in a random coil leading to an isolated strand with the X residue being Asp. The motif 3G16:A133 was found to lie in a coiled region with a kink at the X residue.

For the motif 2QSU:A168 and 3WDL:B215, the motifs were found to be part of a distorted strand.

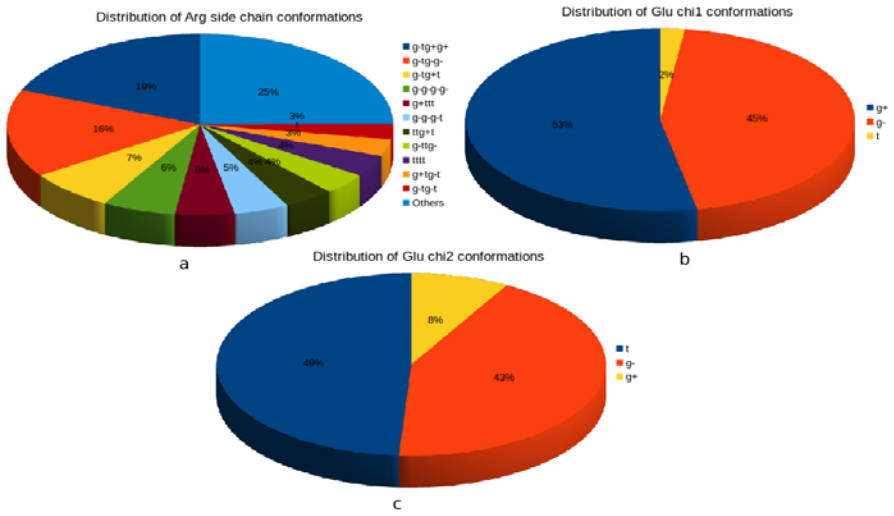

Figure 25. (a) The distribution of Arg side chain conformation in R-X-E motif with two interactions in irregular structural regions. (b) The distribution of Glu χ_1 conformations in R-X-E motif with two interactions in irregular structural regions with Arg side chain g- t g+ g+. (c) The distribution of Glu χ_2 conformations in R-X-E motif with two interactions in irregular structural regions with Arg side chain g- t g+ g+.

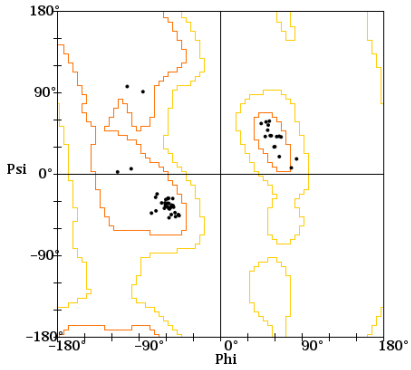

Figure 26. Ramachandran plot for residues in the X position in R-X-E motifs in irregular structural regions with two interactions and Arg side chain conformation g- t g+ g+.

In 19 cases the hydrogen bonding was Type D (Figure 27a) while in 7 it was Type B (Figure 27b). In two cases, namely 2QSU:A168 and 4G3J:A449, two side chain – main chain bonds i.e. Arg (NE) – (O) Glu and Arg (NH2) – (O) Glu were present.

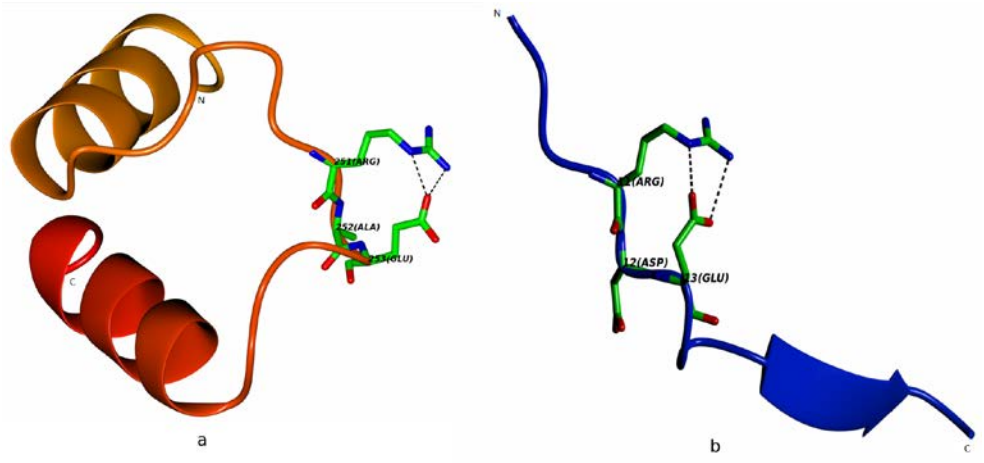

Figure 27. (a) The R-A-E motif in 3FUT occurring in irregular structural regions with g- t g+ g+ Arg side chain conformation and Type D hydrogen bonding interaction. (b) The R-D-E motif in 1H2W occurring in irregular structural regions with g- t g+ g+ Arg side chain conformation and Type B hydrogen bonding interaction.

The next conformation was g- t g- g-, which was observed in 16% of the total (Figure 25a). Here again considerable variation was recorded for both Glu χ_1 and χ_2 side chain dihedral angles. For 22 motifs the Glu χ_1 was g-, g+ for 22 and t in two cases. The Glu χ_2 was similarly observed to be t in 27 cases, g- in 13 and g+ only in one (Figure 28a-b).

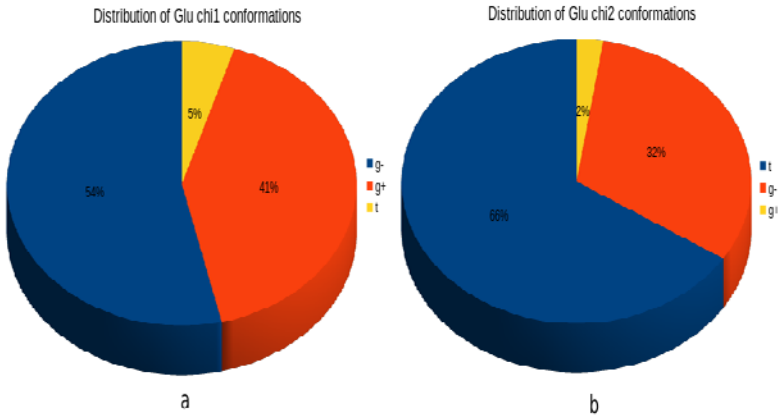

Figure 28. (a) The distribution of Glu χ_1 conformations in R-X-E motif with two interactions in irregular structural regions with Arg side chain g- t g- g-. (b) The distribution of Glu χ_2 conformations in R-X-E motif with two interactions in irregular structural regions with Arg side chain g- t g- g-.

The Ramachandran plot of the X residue of the motifs was observed to form two distinct clusters with only one outlier, namely in 2J0P:A80 (Figure 29). In 15 cases the hydrogen bonding was found to be Type D while in 11 it was Type B.

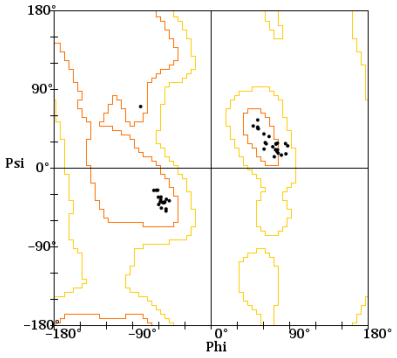

Figure 29. Ramachandran plot for residues in the X position in R-X-E motifs in irregular structural regions with two interactions and Arg side chain conformation g- t g- g-.

In 13 cases a variant of Type B was found where one side chain – side chain H-bond was found to be replaced by main chain – side chain bond. In case of 4NF9:A2096, the hydrogen bonding comprised of two main chain – side chain bonds, a variant of Type D with Arg (N) – (OE1) Glu bond (Figure 30a).

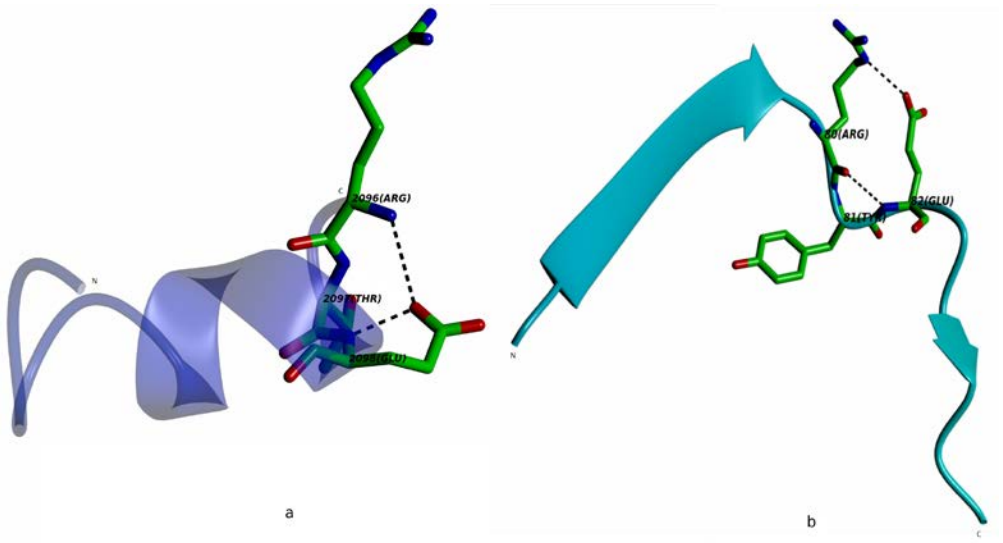

Figure 30. (a) The R-T-E motif in 4NF9 occurring in irregular structural regions with g- t g- g- Arg side chain conformation and described hydrogen bonding interaction. (b) The R-Y-E motif in 2J0P occurring in irregular structural regions with g- t g- g- Arg side chain conformation and described hydrogen bonding interaction.

The motif was found to lie at the N-terminal of the chain as part of a small helical region. For the motif 2J0P:A80, the bonding was found to have one main chain – main chain bond, Glu (N) – (O) Arg, instead of the side chain – side chain bond (Figure 30b).

4.2.3.3 Analysis of E-K motif with one hydrogen bond in irregular structural regions.

Total 364 occurrences of this motif were recorded with one interaction in irregular structural regions. 309 motifs were identified to have main chain – side chain interaction while 45 showed side chain – side chain and 10 had side chain – main chain bond. Of the wide range of conformation observed for Lys, 78 (21%) had the conformation g- t t t (Figure 31). The Glu χ_1 conformation was g+ and the χ_2 conformation was g-.

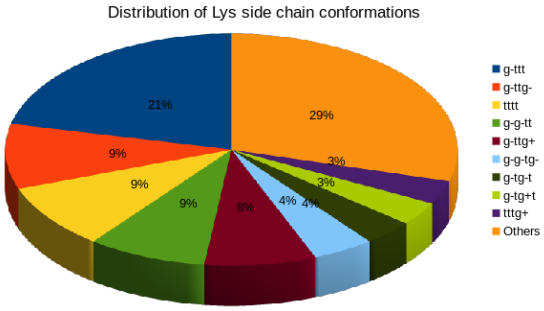

Figure 31. The distribution of Lys side chain conformation in E-K motif with one interaction in irregular structural regions.

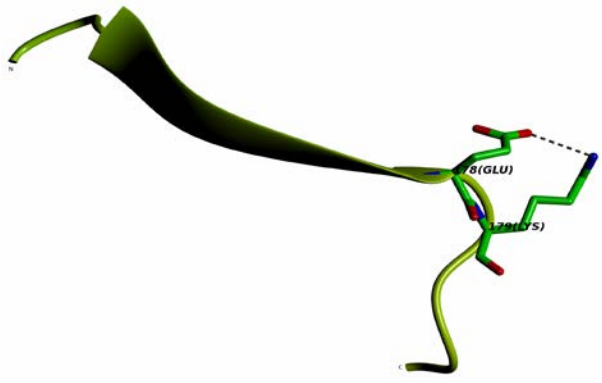

Figure 32. The E-K motif in 1TZF occurring in irregular structural regions with Lys side chain conformation g- t t t.

The hydrogen bonding in 5 cases was Lys (N) – (OE2) Glu, with only one showing side chain – side chain H-bond (Lys (NZ) – (OE2) Glu) (Figure 32) and in rest it was main chain – side chain.

4.2.3.4 Analysis of K-E motif with one hydrogen bond in irregular structural regions.

The occurrence of hydrogen bonds was studied in 388 occurrences of K-E motifs. Out of them 286 showed main chain – side chain bonds, 57 showed side chain – side chain and 45 showed side chain – main chain H-bonds. The study of Lys side chain showed the conformation g- t t t occurring in 58 (15%) motifs (Figure 33a). Here the Glu χ_1 conformation was g- and Glu χ_2 conformation was g+.

In 2 cases the bonding was Lys (N) – (OE1) Glu and in rest 55 cases the bonding was main chain – side chain. In the two cases mentioned above the side chains of the motif residues were found to lie on opposite sides of the peptide plane (Figure 33b).

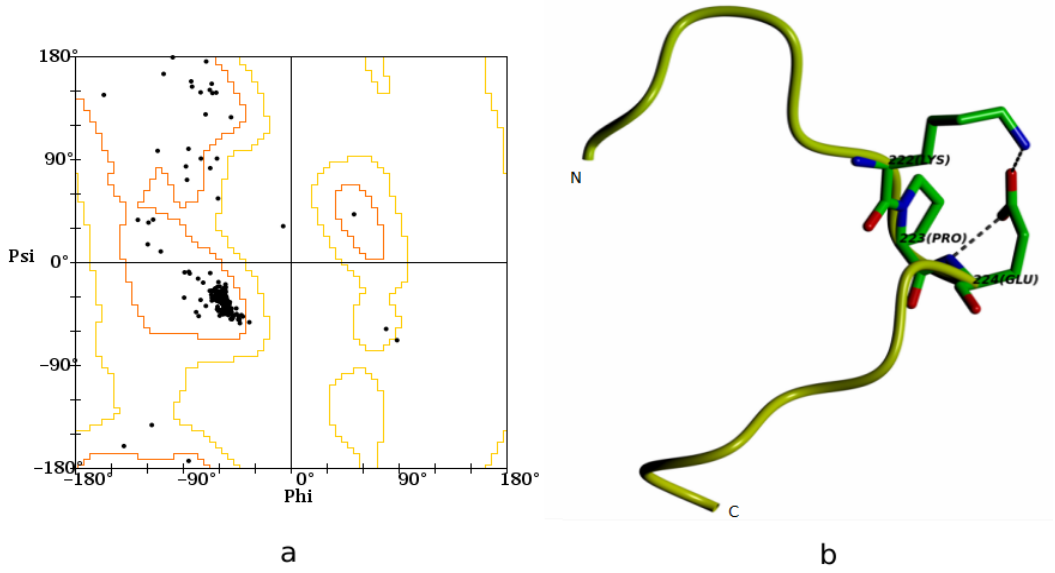

Figure 33. (a) The distribution of Lys side chain conformation in K-E motif with one interaction in irregular structural regions. (b) The K-E motif in 4MC5 occurring in irregular structural regions.

4.2.3.3 Analysis of E-X-K motif with one hydrogen bond in irregular structural regions.

Total 612 occurrences of the motif were identified. For this group 329 were found to have main chain – side chain, 144 showed side chain – side chain and 72 have side chain – main chain H-bonds. The conformation g- t t t for the Lys side chain was found to occur in 134 (22%) motifs (Figure 34). No specific conformation could be identified for Glu χ_1 and χ_2.

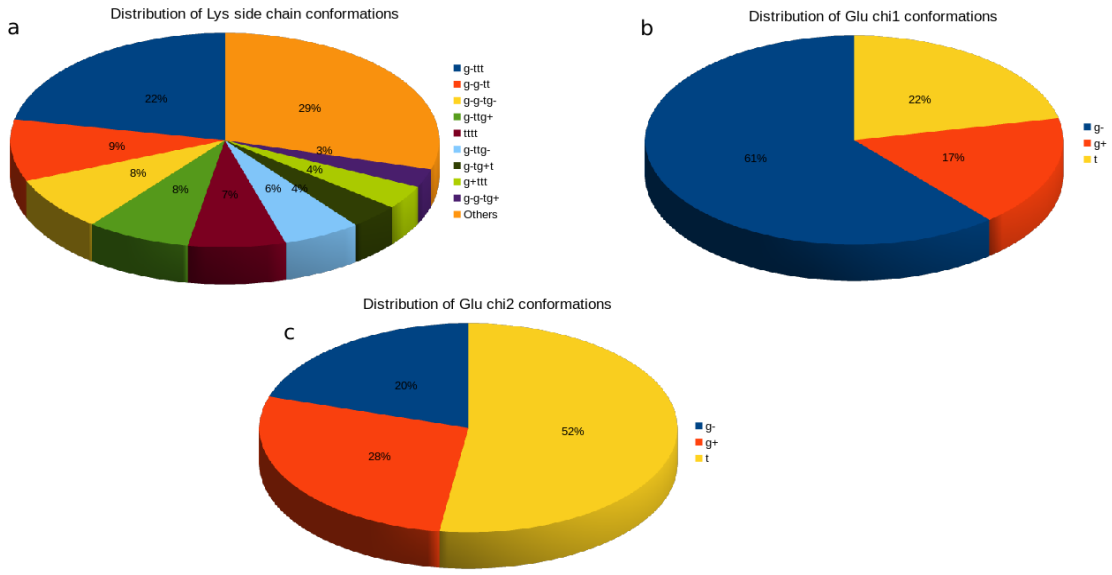

Figure 34. (a) The distribution of Lys side chain conformation in E-X-K motif with one interaction in irregular structural regions. (b) The distribution of Glu χ_1 conformations in E-X-K motif with one interaction in irregular structural regions. (c) The distribution of Glu χ_2 conformations n E-X-K motif with one interaction in irregular structural regions.

The Ramachandran plot for the X residue does not limit to a particular region and spread over all allowed regions (Figure 35a). In 74 cases the hydrogen bonding was main chain – main chain. Of them in 27 cases the bonding was Lys (N) – (OE1) Glu. In 45 cases the hydrogen bonding was side chain – side chain (Figure 35b) and in one it was side chain – main chain.

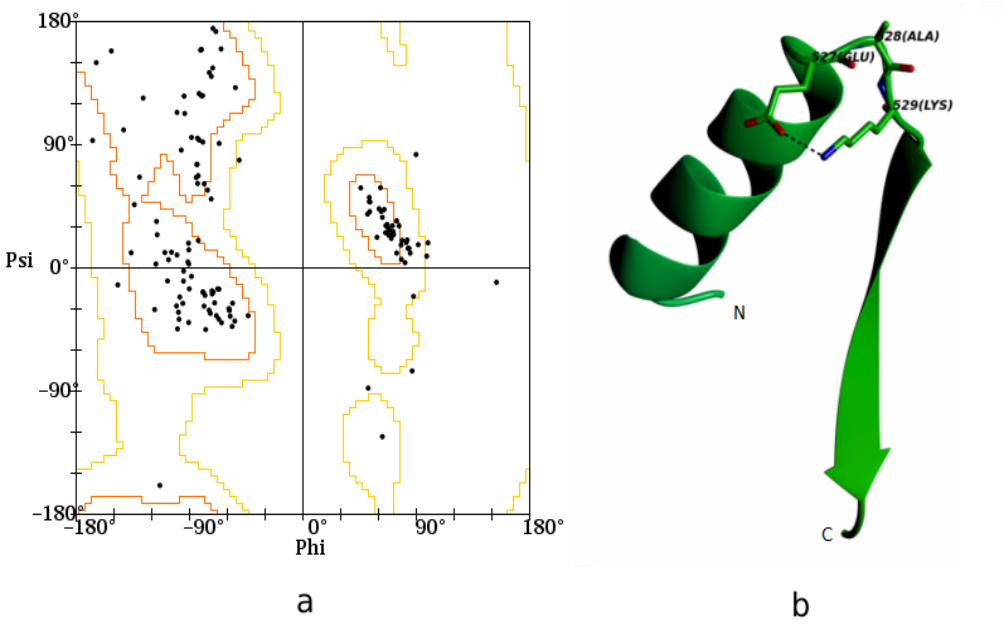

Figure 35. (a) The Ramachandran plot for E-X-K motifs with one interaction in irregular structural regions. (b) The E-A-K motif in 2B8E occurring in irregular structural regions.

4.2.3.3 Analysis of K-X-E motif with one hydrogen bond in irregular structural regions.

712 motifs of the K-X-E pattern were identified with one interaction belonging to irregular structural regions. From these 534 were found to have main chain – side chain hydrogen bond, 118 with side chain – side chain bond, 46 with main chain – main chain and 14 with side chain – main chain interaction. The side chain conformation analysis for the Lys residue exhibited more than 30 different conformations; the most prominent was g- t t t (Figure 36a). The Glu χ_1 and χ_2 conformation was observed to be varying (Figure 36b-c). The Ramachandran plot for the X residue shows a major cluster in α-helix region along with a spread into the β-sheet region (Figure 37a). Of the 181 occurrences, 169 were seen to have main chain – side chain bond, 10 had main chain – main chain bond, while only 2 had side chain – side chain bond (Figure 37b).

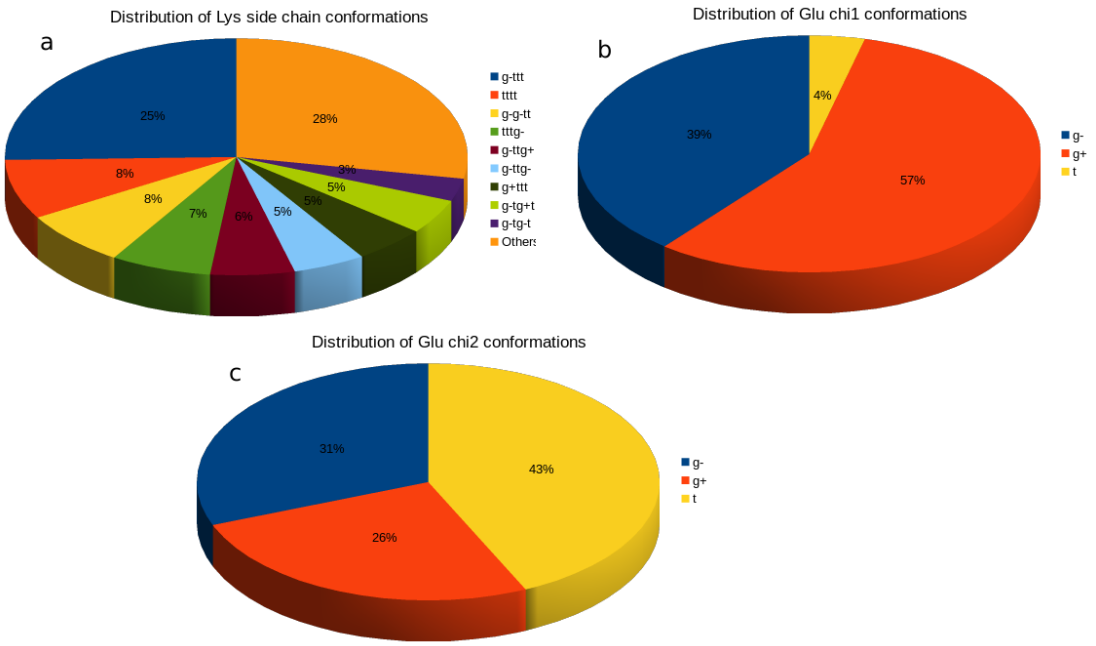

Figure 36. (a) The distribution of Lys side chain conformation in K-X-E motif with one interaction in irregular structural regions. (b) The distribution of Glu χ_1 conformations in K-X-E motif with one interaction in irregular structural regions. (c) The distribution of Glu χ_2 conformations in K-X-E motif with one interaction in irregular structural regions.

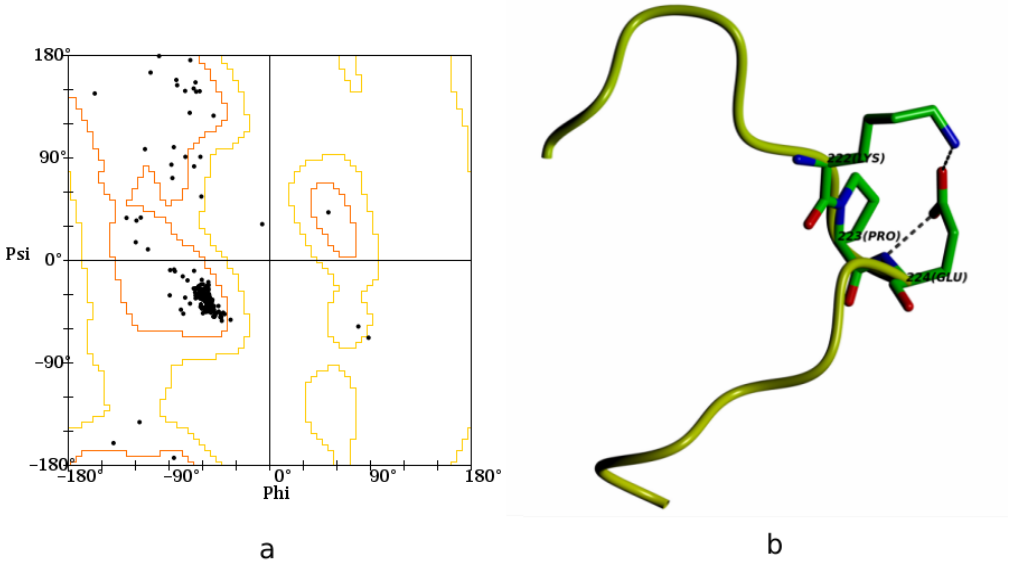

Figure 37. (a) The Ramachandran plot for K-X-E motifs with one interaction in irregular structural regions. (b) The K-P-E motif in 4I5X occurring in irregular structural regions.

4.3 Comparative Analysis of the motifs

Based on the motifs explored above, the patterns were compared to understand the role of factors such as change of sequence order and introduction of intervening residue on the fold and interactions.

4.3.1 Comparing the ER and E-X-R motifs.

The motifs E-R and E-X-R were compared to study the change in motifs with the introduction of a residue in between Glu and Arg. The motifs were compared for secondary structure, number and type of hydrogen bonding interactions, Arg and Glu side chain conformations.

The table shown reveals that for the E-R motifs, occurrences were only observed in helix and the irregular structure part while in case of E-X-R motif it was found in helices, sheets and irregular structural regions. E-R motifs in helices were found with two hydrogen bond interactions of Type B. The Arg side chain was found to have folded conformation while the Glu side chain χ_1 was found to assume the conformation t with Glu side chain χ_2 being g-. On introduction of the spacer residue, i.e. for E-X-R motifs in helices the Arg side chain was found to change to partially folded conformation t g+ t g+ and the Glu side chain χ_1 was g- and the Glu χ_2 was g+. The hydrogen bonding was found to involve main chain atoms thus different from Type B.

For E-R motifs in irregular structural regions, the Arg side chain was somewhat folded with the conformation g- g- t g- and the Glu χ_1 was g- and Glu χ_2 was g+. The hydrogen bonding in this case involved main chain atoms; Arg (N) – (OE1) Glu and Arg (NH1) – (O) Glu, since the side chain of the motif residues were found to lie on opposite sides of the peptide plane. In the E-X-R motif the Arg side chain assumed a partially folded conformation and the Glu χ_1 became t with the Glu χ_2 assuming g+ conformation. Here the motif residues were found to form 2 or 3 hydrogen bonds belonging to Type B, in case of the 3 interactions the third bond being Arg (N) – (OE1) Glu.

Table 4: Comparison of the E-R, R-E, E-X-R and R-X-E motifs.

Motif	SS	No. of Interactions	Type	Major Arg Conf	Major Glu Conf (χ1)	Major Glu Conf (χ2)	Cα-Cα dist.	Fig. No.
E-R	H	2	Type B	g-g-g+t	t	g-	-	38a
R-E	H	2	Type D	ttg+t	g-	g-	-	40a
E-X-R	H	2	Glu(N)-(OE1)Glu Arg(NH1)-(O)Arg	tg+tg+	g-	g+	5.407	38b
E-X-R	S	2	Type D	g-ttt	t	t	6.474	
R-X-E	S	2	Type D	tttg+	g-	t	6.665	41b
E-R	irregular structure parts	2	Arg(N)-(OE1)Glu Arg(NH1)-(O)Glu	g-g-tg-	g-	g+	-	39a
R-E	irregular structure parts	2	Arg(NE)-(O)Glu Arg(NH1)-(O)Glu	g+tg-t	g-	g-/t	-	40b
E-X-R	irregular structure parts	2, 3	Type B + 1/ Type D	g-tg-t	t	g+	5.787, 5.700	39b
R-X-E	irregular structure parts	2,3	Type B + 1, Type D	g-g-g-g-	g+	g-/t	5.532	41a

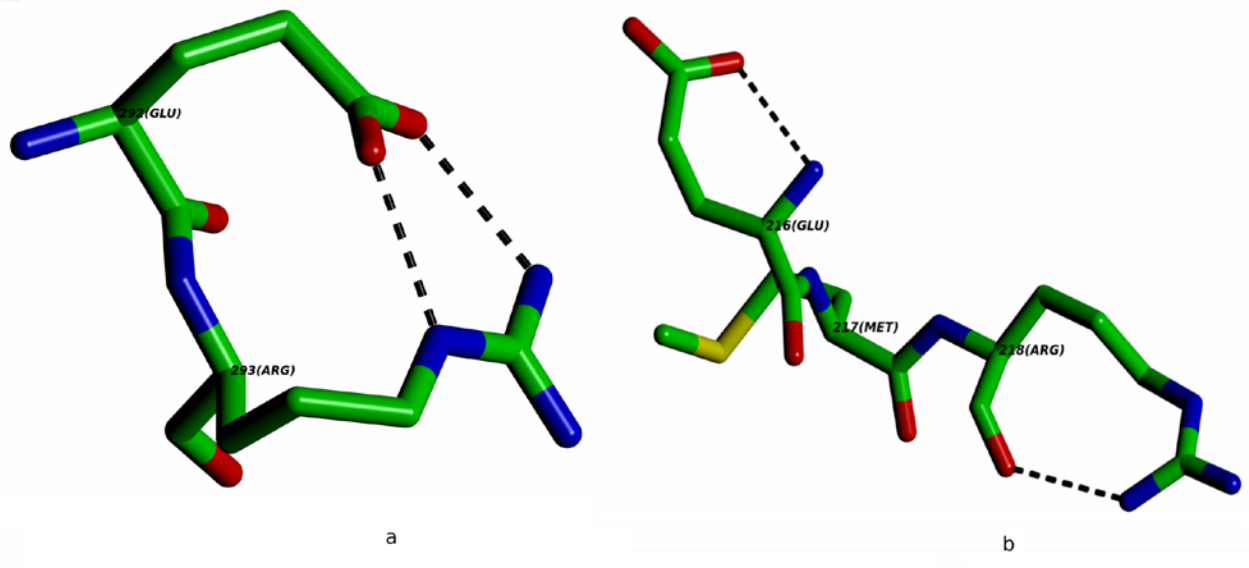

Figure 38. (a) The E-R motif in 2F7V occurring in helix group with g- g- g+ t Arg side chain conformation and Type B hydrogen bonding interaction. (b) The E-M-R motif in 4K0F occurring in irregular structural regions with t g+ t g+ Arg side chain conformation and hydrogen bonding interaction described above.

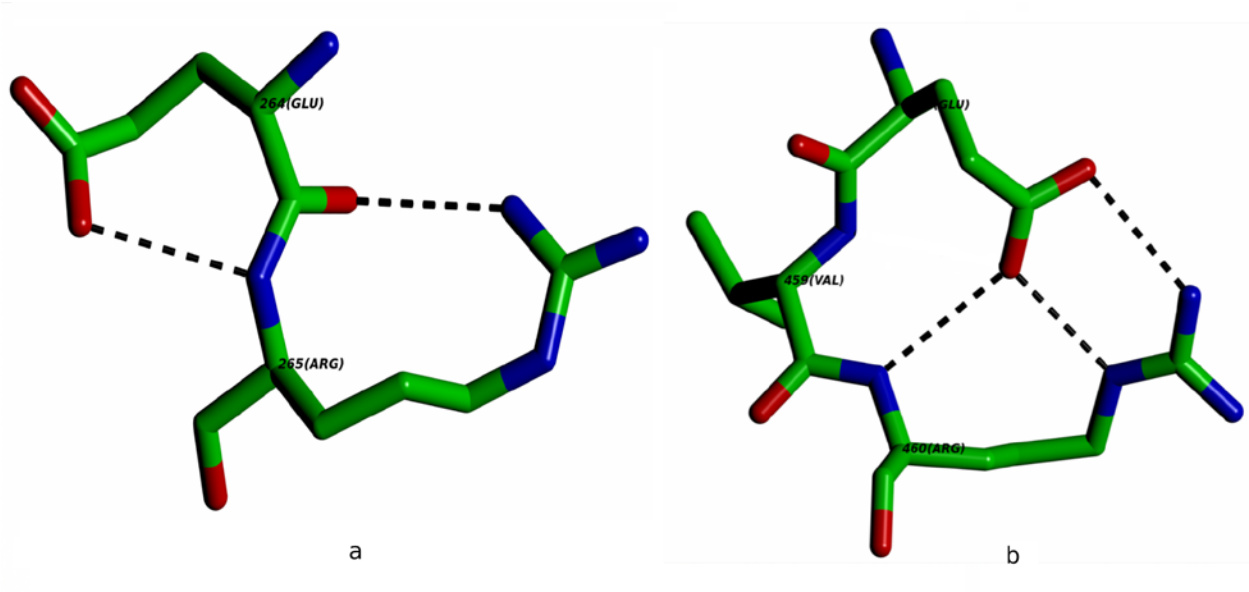

Figure 39. (a) The E-R motif in 2FN9 occurring in irregular structural regions with g- g- t g- Arg side chain conformation and the hydrogen bonding interaction as shown. (b) The E-V-R motif in 4BEW occurring in irregular structural regions with g- t g- t Arg side chain conformation and Type B hydrogen bonding, similar to that observed in D-X-R with three H-bonds discussed in chapter 3.

The E-X-R motifs as part of sheets were found to have the Arg extended side chain conformation. For interaction to take place, it was observed that the Glu χ_1 and Glu χ_2 also assumed the conformation t. The hydrogen bonding was observed to be Type D.

4.3.2 Comparing the E-R and R-E motifs.

The motifs E-R and R-E were compared to study the change in motifs structure with the reversal of the neighboring residue positions. Motifs with only upto two interactions were observed for both E-R and R-E patterns. In case of both the motifs the hydrogen bonding was to change by the reversal of residue positions. While motifs in the helix group possessed Type B hydrogen bonding in the case of E-R, in the case of R-E motifs the H-bonding was Type D. For E-R motifs in helix, the Arg side chain was found to be more folded which changed to a more extended conformation for the R-E motifs. The E-R helix motifs were found to have Glu χ_1, t and Glu χ_2, g- which changed for the RE motifs to a folded conformation with Glu χ_1 and Glu χ_2 being g-. In the irregular structural regions the Arg side chain was shifted from a folded state in E-R to a partially folded state in R-E. In both cases the Glu χ_1 was found to be g- whereas the Glu χ_2 changed from g+ to either g- or t. In both motifs the hydrogen bonding was observed to involve main chain atoms. The ARG (N) -(OE1) GLU and ARG (NH1) - (O) GLU bonding seen in ER motifs was found to shift to ARG (NE) – (O) GLU and ARG (NH1) – (O) GLU for the RE motifs.

4.3.3 Comparing the RE and R-X-E motifs.

The motifs RE and R-X-E were compared to know the effect of introduction of an intervening residue between Arg and Glu. The preference of the RE motifs for the helix group, on introduction of the X residue was found to shift significantly to the irregular structural regions, especially for motifs with interactions.

For RE motifs in irregular structural regions, the Arg side chain conformation was found to be partially folded which on addition of a spacer residue changed to a more folded conformation in the R-X-E motifs. While the Glu χ_1 for the RE motifs was found to be g-, the Glu χ_1 and Glu χ_2 for R-X-E motifs were found to assume a wide range of conformation. In both cases majority of the motifs were with two H-bonds. For RE motifs

in helices and R-X-E in sheets both have Type D hydrogen bonding. In RE motifs belonging to irregular structural regions, two side chain – main chain interactions viz. Arg (NE) – (O) Glu and Arg (NH1) – (O) Glu were observed since the Glu side chain was found to point away from the Arg side chain whereas in R-X-E Type D bonding was observed. R-X-E motifs belonging to sheets were found to have more extended Arg side chain conformation. The Glu χ_1 was found to be g- while χ_2 was t.

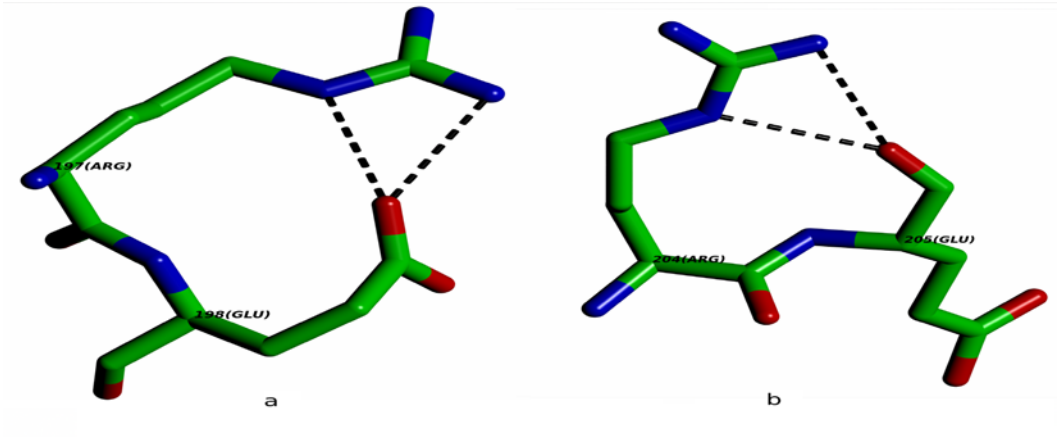

Figure 40. (a) The RE motif in 3C2Q occurring in helix group with t t g+ t Arg side chain conformation and Type D hydrogen bonding interaction. (b) The RE motif in 2IZR occurring in irregular structural regions with g+ t g- t Arg side chain conformation and described hydrogen bonding.

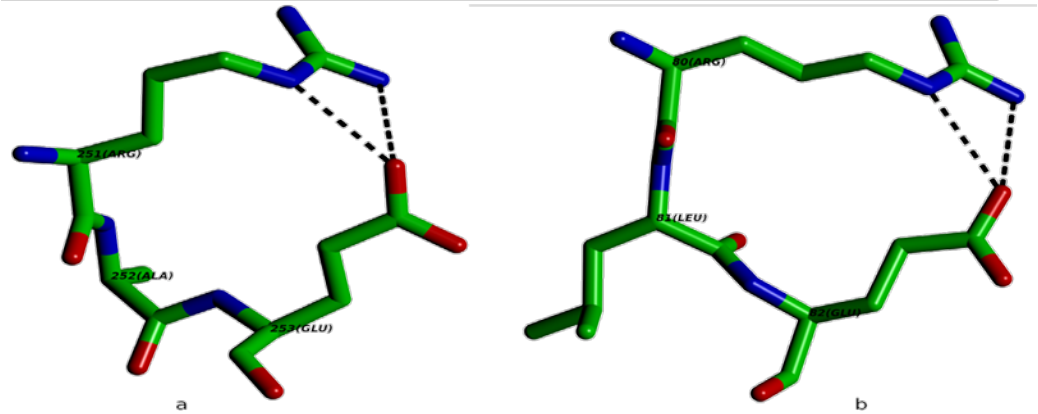

Figure 41. (a) The R-A-E motif in 3FUT occurring in irregular structural regions with Arg side chain conformation as g- t g+ g+ and Type D hydrogen bonding interaction. (b) The R-L-E motif in 2D5W occurring in β-sheet with t t t g+ for Arg side chain conformation and Type D hydrogen bonding interaction.

4.3.4 Comparing the E-X-R and R-X-E motifs.

Comparative analysis of E-X-R and R-X-E highlighted the effect of reversal of residue positions. The E-X-R motif was found to show an overall preference for helices. However, in terms of motifs with interactions, the preference shifts towards the irregular structural regions. In case of R-X-E motifs the occurrences was found to be almost equal in helices as well as irregular structural regions, whereas the motifs with interactions were more in irregular structural regions.

For both motifs similar trends were observed for occurrences of amino acids at the X position. The amino acids Leu, Ala and Val were occurring most frequently at the X position. Apart from these the charged residues Asp, Glu, Lys and Arg were observed to have significant occurrences with Gly and Ile at the X residue position (Figure 42).

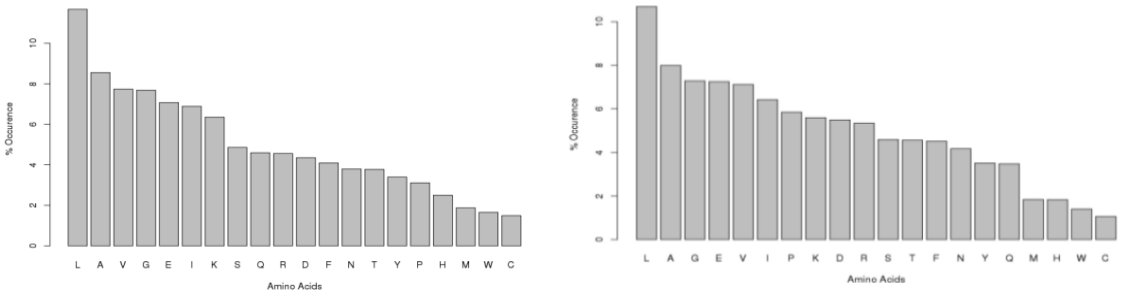

Figure 42. The frequency bar plot of the occurrence of all 20 amino acids at the X position in the E-X-R and R-X-E motifs.

Motifs with both two and three interactions were identified for E-X-R and R-X-E belonging to the irregular structural regions. Where E-X-R motifs were observed to have a partially folded conformation, on reversal of the residue positions, the conformation was found to change to a highly folded one. The Glu χ_1 was found to be t for E-X-R which reversed to g+ for R-X-E in one case and was found to vary in others. The Glu χ_2 was found to be mainly g+ for E-X-R while for R-X-E it was found to vary significantly. Comparison of these motifs revealed the backbone of the R-X-E motifs in irregular structural regions were highly folded compared to E-X-R motifs as evident not only from the Arg and Glu side chain conformations but also from the average C_α (E/R)…C_α (R/E)

distance being decremented from 5.787 Å for E-X-R to 5.525 Å for R-X-E. In case of motifs with three interactions for both motifs the hydrogen bonding was Type B with an additional hydrogen bond involving main chain atom also while in motifs with two interactions the bonding was of Type D.

Table 5. Comparison of the E-K, K-E, E-X-K and K-X-E motifs.

Motif	SS	No. of Interactions	Type of H-bond	Major Arg Conf	Major Glu Conf (χ1)	Major Glu Conf (χ2)	Cα-Cα dist.
E-K	H	1	MS	g- t t t	g+	g-	-
K-E	H	1	MS	t t t t	g-	g+	-
E-X-K	H	1	MS	g- t t t	g-	g+	5.49
E-X-K	S	1	MS	g- g- t t	t	t	6.46
K-X-E	S	1	SS	t t t t	g-	t	6.46
E-K	irregular structural regions	1	MS	g- t t t	g+	g-	-
K-E	irregular structural regions	1	MS	g- t t t	g-	g+	-
E-X-K	irregular structural regions	1	MS	g- t t t	various	various	5.93
K-X-E	irregular structural regions	1	MS	g- t t t	various	various	5.78

Footnote: MS: Main chain – side chain hydrogen bond. SS: Side chain – side chain hydrogen bond.

In case of E-X-R and R-X-E motifs in sheets, the Arg side chain conformations for both were observed to be nearly extended with the Glu χ_1 and Glu χ_2 also being t for E-X-R and g- and t for R-X-E, respectively. Compared to the E-X-R motifs, R-X-E motifs were found to be extended in the backbone conformations since the average C_α (E/R)…C_α (R/E) distance was 6.665 Å in R-X-E motifs while in E-X-R it was 6.474 Å. In both cases the observed hydrogen bonding was of Type D.

4.3.5 Comparing the E-K and K-E motifs.

The E-K and K-E motifs were found to occur in both helices and in irregular structures in significant numbers. For all cases, only one hydrogen bond was found to occur. The hydrogen bond in all of these was found to be main chain – side chain, Glu (N) - (OE1) Glu (Figure 43) in majority.

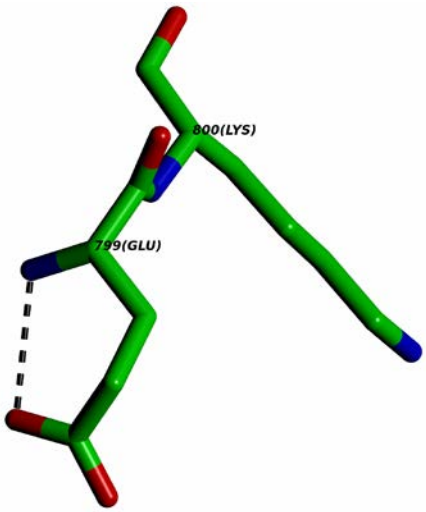

Figure 43. The E-K motif in helices with Glu (N) - (OE1) Glu hydrogen bond.

Very few cases were identified with side chain – side chain bonding. For the E-K motif in helices the Lys side chain was found to be almost extended with the conformation g- t t t while on reversal of the residue positions, for K-E in helices the conformation was fully extended t t t t. The Glu χ_1 and Glu χ_2 conformation was g+ and g- in E-K which reversed in K-E with Glu χ_1 being g- and Glu χ_2 being g+.

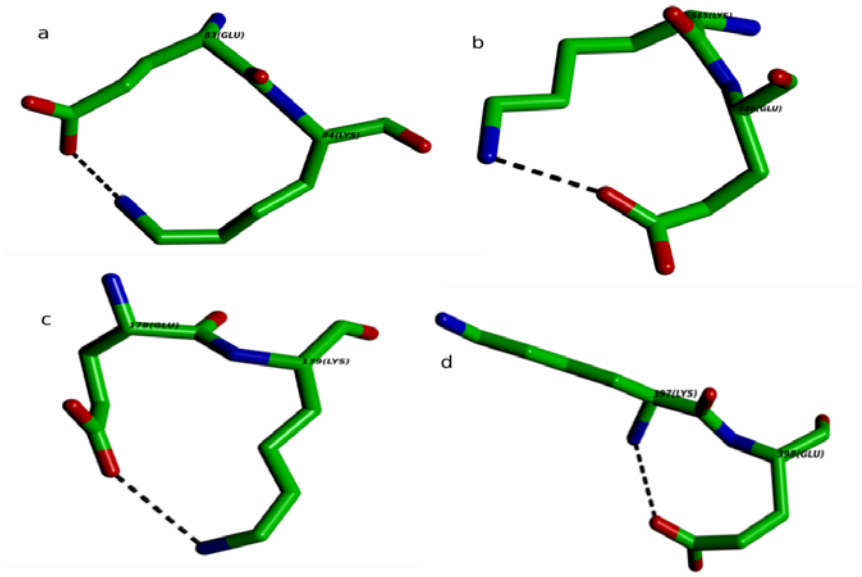

Figure 44. (a) The E-K motif in helices. (b) The K-E motif in helices. (c) The E-K motif in irregular structures. (b) The K-E motif in irregular structures.

For the motifs occurring in irregular structural regions the Lys side chain conformation was found to remain the same for the E-K and K-E motifs. However, the side chain conformation of Glu χ_1 and Glu χ_2 were found to be exactly reversed to Glu χ_1 and Glu χ_2 conformations as g+ and g- in E-K which were reversed in K-E as Glu χ_1 being g- and Glu χ_2 being g+.

4.3.6 Comparing the E-X-K and K-X-E motifs.

The E-X-K and K-X-E were observed to have significant occurrence in sheets and irregular structural regions. Only E-X-K motif was found in considerable numbers in helices, where the Lys side chain conformation was identified as g- t t t while the Glu χ_1 being g- and Glu χ_2 being g+. For E-X-K motifs occurring in the β-sheets, Lys side chain conformation was found to be partially folded, g- g- t t, which with reversal of Lys and Glu position; for the K-X-E motif the Lys side chain conformation was completely extended t t t t. The Glu χ_1 conformation was t along-with Glu χ_2 being t whereas for K-X-E the Glu χ_1 conformation was g- while the Glu χ_2 remained t. The average C_α (E/K)…C_α (K/E) distance was found to remain almost same for both.

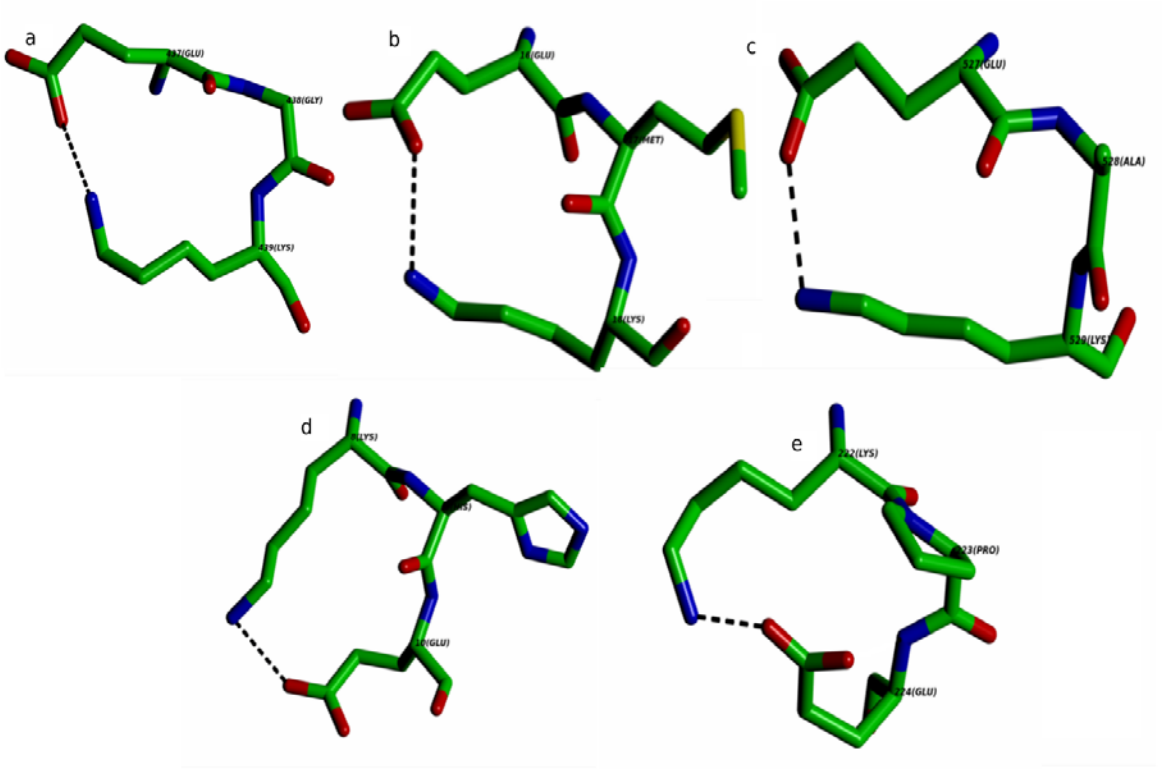

Figure 45. (a) The E-G-K motif in helices. (b) The E-M-K motif in sheets. (c) The E-A-K motif in irregular structures. (d) The K-H-E motif in sheets. (c) The K-P-E motif in irregular structures.

In irregular structural regions, both E-X-K and K-X-E were found to occur in considerable numbers. For both motifs the Lys side chain conformation was g- t t t. However, the Glu χ_1 and Glu χ_2 was found vary largely. In terms of the fold the K-X-E motifs were observed to be more compact with the average C_α (K)…C_α (E) distance being 5.78 Å as compared to 5.931 for E-X-K motifs. The oxygen atom of Glu involved in side chain – side chain H-bond is mainly OE1 in E-X-K while OE2 in K-X-E.

In the E-X-K motif the most frequently occurring amino acids at the X position were observed to be Leu and Val. These were followed by Gly, Ile, Ala and Glu in significant frequencies (Figure 3). With the interchange of the Glu and Lys positions, there was significant rise in occurrence of most amino acids. Although Leu still remained the most frequent, the occurrence of Ala, Glu, Gly and Val was found to be nearly in equal frequencies. This trend was closely followed by the occurrence of Ile and Lys (Figure 46).

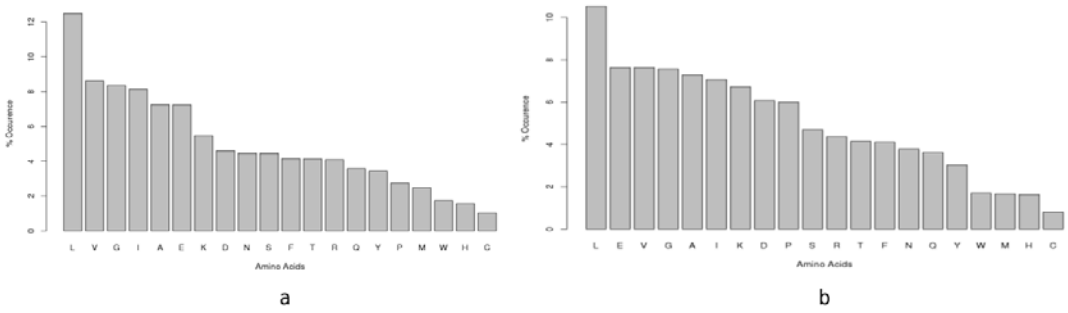

Figure 46. The frequency bar plot of the occurrence of all 20 amino acids at the X position in the E-X-K and K-X-E motifs.

4.4 Summary

Extending the initial analysis, further detailed analysis was performed for short structural motifs involving Glu and Arg/Lys. The motifs were analysed for the local fold assumed by the residues in motif sequence and the interactions involved along-with the role and the direction of the residues in the sequence. The sequence motifs involving Glu and Arg were found to show preference for regular secondary structures compared to irregular structural regions in both. A similar trend was observed for motifs with interactions in the case of E-R, R-E and E-X-R but not R-X-E where the preference was for irregular structural regions. It was noted that motifs with three hydrogen bonded interactions showed similar backbone conformation in all as compared to variation found in those with one or two or those without any hydrogen bonds. This highlighted the role of hydrogen bonds in steering the local folding of the motifs. In case of E -> R motifs the Arg side chain conformation were found to be partially folded in helices and irregular structural regions, whereas in R -> E motifs the conformation was found to be nearly extended in helices and highly folded in irregular structural regions. For sequence motifs involving Glu and Lys, with these residues as immediate neighbours the preference was observed for occurrence in helices whereas when a spacer residue was present the preference was observed to shift towards irregular structural regions. However, for motifs having interaction were found to occur in significant numbers for all E-K, K-E, E-X-K motifs in both helices and irregular structural regions but not in K-X-E where the occurrence was noted only in irregular structural regions. Motifs with interactions were

noted to have primarily only one hydrogen bond. In case of motifs with and without interaction the folding was observed to vary significantly.

Chapter 5

Identification and analysis of D-X(2,8)-R, R-X(2,8)-D, E-X(2,8)-R, R-X(2,8)-E motifs, their occurrence and interactions

From the analysis of two and three residue motifs in the previous chapters, we are now moving to the study of 4-8 residue motifs involving Aspartic acid or Glutamic acid on one end and Arginine on the other. From the previous analysis we could conclude that the interactions involving lysine were less significant and no specific pattern of conformation or H-bonding occurring in significant numbers could be detected. Hence we have considered only Arg as the cationic amino acid in further analysis. Thus, this chapter deals with the structural analysis of 28 motifs conforming to the general patterns of D/E-(2,8)X-R and R-(2,8)X-D/E. The table below lists all the 28 motifs analyzed using the local PDB dataset.

Table 1. List of 28 motifs analysed in this chapter.

D(2X)R	R(2X)D	E(2X)R	R(2X)E
D(3X)R	R(3X)D	E(3X)R	R(3X)E
D(4X)R	R(4X)D	E(4X)R	R(4X)E
D(5X)R	R(5X)D	E(5X)R	R(5X)E
D(6X)R	R(6X)D	E(6X)R	R(6X)E
D(7X)R	R(7X)D	E(7X)R	R(7X)E
D(8X)R	R(8X)D	E(8X)R	R(8X)E

5.1 Structural analysis of motifs in the pattern D-(2,8)X-R.

As more and more number of intervening residues occur the chances of interaction between Asp and Arg can be expected to reduce as the distance between these residues increases. This disposition will be most significant in the extended strands of β-sheets. It is indeed true can be seen from the numbers listed in table 2. What surprises is that even their number of occurrence without interaction in β-sheets is also low.

Total eight motifs conforming to this pattern were analyzed in terms of their secondary structure conformation in proteins (Table 2). The first motif analyzed was D-(2X)-R.

11068 motifs belonging to this pattern were identified in the dataset. While most motifs were without interactions between the reference amino acids Asp and Arg, 1953 motifs were observed with interactions in the helix group.

Table 2. Structural analysis of D-(2,8)X-R motifs.

Pattern (Total)	Helix		Sheets		Irregular Regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
D-(2X)-R (11068)	1953	2591	52	416	2515	3541
Total	4544		468		6056	
D-(3X)-R (10618)	3592	356	5	367	1970	4329
Total	3947		372		6299	
D-(4X)-R (9505)	68	2096	4	224	927	6186
Total	2164		228		7113	
D-(5X)-R (9774)	65	1834	6	270	615	6984
Total	1899		276		7599	
D-(6X)-R (9754)	93	2303	3	178	495	6682
Total	2396		181		7177	
D-(7X)-R (9444)	67	1753	2	110	480	7032
Total	1820		112		7512	
D-(8X)-R (9447)	50	1734	2	65	392	7203
Total	1784		67		7595	

Out of these, 1521 were found to have only one, 413 had two and 19 were with three hydrogen bonds, respectively. The second motif analyzed was of five residues, D-(3X)-R motif. Total 10618 occurrences were identified. Very high occurrences of motifs with interactions were observed in the helix group.

Of these 261 were found with two hydrogen bonds. In case of the irregular structures with specific conformation not repeated for consecutive residues 1370 occurrences were with one H-bond, 371 with two and 299 with three bonds. The next motif studied was the D-(4X)-R wherein 9505 occurrences were recorded. From the 927 motifs belonging to irregular structures with interactions, 603 were observed to have one hydrogen bond while 324 were observed with two or more hydrogen bonds. For the D-(5X)-R motif, 9774 occurrences were observed. Only 161 occurrences belonging to the irregular structural group were found to have two or more hydrogen bonds.

In case of the D-(6X)-R motif, 9754 occurrences were found of which 495 were found with interactions in the irregular structures. Only 93 were found with interactions in helices while 3 were noted in β-sheets. In the next motif D-(7X)-R, 9444 motifs were studied. Of the 480 motifs with interactions in irregular structures, 372 showed one H-bond while 101 had two H-bonds while only 17 were found with three H-bonds. The last motif analyzed in this section was D-(8X)-R. 9447 motifs were detected. In the irregular structure, out of the 392 with interactions, 300 had one hydrogen bond each while 82 had two and only 8 showed three hydrogen bonds.

5.2 Structural analysis of motifs in the pattern R-(2,8)X-D.

This section explores the structural analysis of eight motifs based on the consensus pattern R-(2,8)X-D (Table 3). The first motif studied was the R-(2X)-D. Total 9895 occurrences of this motif were recorded. From the 1786 identified with interactions in the irregular structures, 950 showed a single H-bond, 628 had two and 208 had three H-bonds.

Table 3. Structural analysis of R-(2,8)X-D motif.

Pattern (Total)	Helix		Sheets		Irregular Regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
R-(2X)-D (9895)	348	2061	13	734	1786	4953
Total	2409		747		6739	
R-(3X)-D (10000)	2719	276	7	655	839	5503
Total	2995		662		6342	
R-(4X)-D (9498)	20	1520	5	587	606	6640
Total	1540		592		7366	
R-(5X)-D (9124)	20	1200	12	397	632	6863
Total	1220		409		7495	
R-(6X)-D (9578)	32	1429	3	311	564	7239
Total	1461		314		7803	
R-(7X)-D (9528)	9	1324	3	165	485	7542
Total	1333		168		8027	
R-(8X)-D (9050)	16	1022	2	140	459	7411
Total	1038		142		7870	

The next motif analyzed was R-(3X)-D wherein, 10000 occurrences were noted. High number of occurrences with interactions was recorded for the helix group. Only 144 showed two H-bonds. For the irregular structures, of the 839 with interactions, 684 had one H-bond and 209 motifs showed two or more H-bonds. The third motif analyzed from this group was R-(4X)-D. Very few occurrences of motifs with interactions were recorded in α-helices and β-sheets. 450 were observed with single H-bond in irregular structural group while 156 showed two or more interactions in the same group.

For the next motif R-(5X)-D, 9124 occurrences were observed. Of these 632 were found involved in interactions in the irregular structure. Out of these, 405 had one H-bond while 213 showed two H-bonds and only 14 were found with three such bonds.

The fifth motif studied was R-(6X)-D. Here 9578 occurrences of the motif were identified. 409 motifs in the irregular structural group were found to have one hydrogen bond while 120 had two and merely 35 had three hydrogen bonds. The next motif analyzed was R-(7X)-D for which 9528 occurrences were estimated. Only 138 motifs were observed with two or more H-bonds in the irregular structure. The last motif studied in this group was R-(8X)-D. From the 9050 motifs identified, only 477 were found to have interactions. Most motifs with interactions were observed in the irregular structural group of which 343 had a single hydrogen bond. 116 were found to have more than one hydrogen bond.

5.3 Structural analysis of motifs belonging to pattern E-(2,8)X-R.

In this section eight motifs based on the consensus pattern E-(2,8)X-R were studied in terms of the structural analysis (Table 4).

The first motif analyzed was E-(2X)-R. A total of 14728 occurrences were noted. The motif was extensively found in helices. In this group 2036 were observed to have one H-bond. 1385 motifs were found to have two H-bonds while only 91 showed three H-bonds.

Table 4. Structural analysis of E-(2,8)X-R motif.

Pattern (Total)	Helix		Sheets		Irregular Regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
E-(2X)-R (14728)	3512	5014	62	612	1459	4069
Total	8526		674		5528	
E-(3X)-R (13945)	7088	574	2	596	996	4689
Total	7662		598		5685	
E-(4X)-R (10723)	262	3513	9	429	761	5749
Total	3775		438		6510	
E-(5X)-R (10920)	288	3404	1	372	678	6177
Total	3692		373		6855	
E-(6X)-R (11817)	454	3929	2	242	659	6449
Total	4383		244		7108	
E-(7X)-R (10799)	248	2971	6	185	787	6602
Total	3219		191		7389	
E-(8X)-R (10327)	172	3120	0	100	720	7115
Total	3292		100		7835	

E-(3X)-R was the next motif to be analyzed. 13945 occurrences were identified from the local dataset. A large number of motifs with interactions were found in the helices. Of these, 672 motifs were identified with two H-bonds occurring in helices. In the irregular structures 770 showed one H-bond and 226 had two or more H-bonds.

The third motif studied was E-(4X)-R wherein 10723 occurrences were recorded. Considerably less numbers of motifs with interaction were noted. In the irregular structural group 597 were found to have one H-bond. 164 motifs in this group had two or more H-bonds. In the helix structures 256 have a single H-bond while merely 6 show two H-bonds.

The fourth motif explored was E-(5X)-R. For this motif 10920 occurrences were identified. For the irregular structure, 576 motifs were found to have one H-bond and 93 with two and 9 with three H-bonds. In case of the E-(6X)-R, 11817 occurrences were recorded. In the helix structures 421 were found with single H-bond and 495 in irregular structure. 164 occurrences were observed with two or more H-bonds in irregular structural group. The next motif studied was E-(7X)-R. 10799 occurrences of this motif were observed. For the irregular structures, 519 occurrences with one hydrogen bond were identified. In 235 occurrences in this group two hydrogen bonds and in 36 three hydrogen bonds were recorded.

Total 10327 occurrences of E-(8X)-R motifs were found from the local PDB dataset. Of the 720 motifs in the irregular structures with interactions, 580 were found to involve a single H-bond while 140 involved two or more H-bonds.

5.4 Structural analysis of motifs based on the pattern R-(2,8)X-E.

This section deals with structural analysis of motifs belonging to the pattern R-(2,8)X-E, a reversal of those analyzed in the earlier section (Table 5). The first motif explored was R-(2X)-E.

For this motif, 11984 occurrences were recorded. Considerable number of motifs with interactions was observed for helix and irregular structural group. In the helix group 1048 occurrences were noted with a single H-bond, 220 motifs were found with two and 38 with three H-bonds. For the irregular structure 1048 motifs showed one hydrogen bond while 519 had two and 117 had three H-bonds. The R-(3X)-E motif was found to

Table 5: Structural analysis of R-(2X)-E motif.

Pattern (Total)	Helix		Sheets		Irregular Structural Regions	
	Interactions	No interactions	Interactions	No interactions	Interactions	No interactions
R-(2X)-E (11984)	1306	3904	32	614	1734	4394
Total	5210		646		6128	
R-(3X)-E (13269)	6030	428	5	532	1211	5063
Total	6458		537		6274	
R-(4X)-E (11341)	46	428	5	508	995	6420
Total	3412		513		7415	
R-(5X)-E (10601)	32	2452	3	390	1101	6623
Total	2484		393		7724	
R-(6X)-E (10996)	89	2885	4	280	810	6924
Total	2974		288		7734	
R-(7X)-E (10831)	30	2583	0	169	986	7063
Total	2613		169		8049	
R-(8X)-E (10356)	23	1854	2	143	742	7591
Total	1877		145		8334	

occur 13269 times in the local PDB dataset used here. In the helix very high numbers of motifs were found with interactions. 4570 were observed to have one H-bond. 282 had three H-bonds. For the irregular structure, 961 had one, 185 had two and 65 had three H-bonds, respectively. The third motif studied was R-(4X)-E. 11341 occurrences of this motif were detected. 823 occurrences of this motif were observed having single hydrogen bonded interaction while 156 had two and mere 16 had three H-bonds.

The fourth motif analyzed was R-(5X)-E for which 10601 occurrences were identified. Of the 1101 motifs occurrences observed with interactions in the irregular structure, 779 were found to have one H-bond and 332 were found to have two or more H-bonds. The next motif was R-(6X)-E wherein 10996 occurrences were noted. From these in the irregular structure, 649 showed involvement of a single hydrogen bond while 161 had two or more. The sixth motif explored was R-(7X)-E. 10831 instances of this motif were identified. 651 in the irregular structure were found to show one H-bond. 148 showed two while 87 had three H-bonds, respectively. The last motif studied was R-(8X)-E. For this motif 10356 occurrences were recorded. In the irregular structure, 622 had one hydrogen bond and 121 had two or more.

5.5 Comparative Analysis of the identified motifs.

The motifs identified in the sections above were analysed statistically for distribution and patterns. The graphs plotted for the occurrences of the motifs (Figure 1) show high presence of motifs with D/E and R separated by two and three residues (i.e. D/E-(2,3)X-R & R-(2,3)X-D/E). In case of D/E-(2,8)X-R, D/E-(2X)-R were observed to be the highest, while this trend was observed to change in case of reversal of the fringe residues i.e. R-(2,8)X-D/E, where R-(3X)-D/E was found to have the highest number of existence. In case of E-(2,8)X-R and R-(2,8)X-D/E a decreasing trend of occurrences observed for 3X to 5X accompanied by a sudden rise in 6X motifs again followed by a decreasing trend of occurrences for 7X and 8X. In case D-(2,8)X-R, a decreasing trend was identified for 2X to 4X with a rise at 5X again followed by a decreasing trend of occurrences for 6X to 8X.

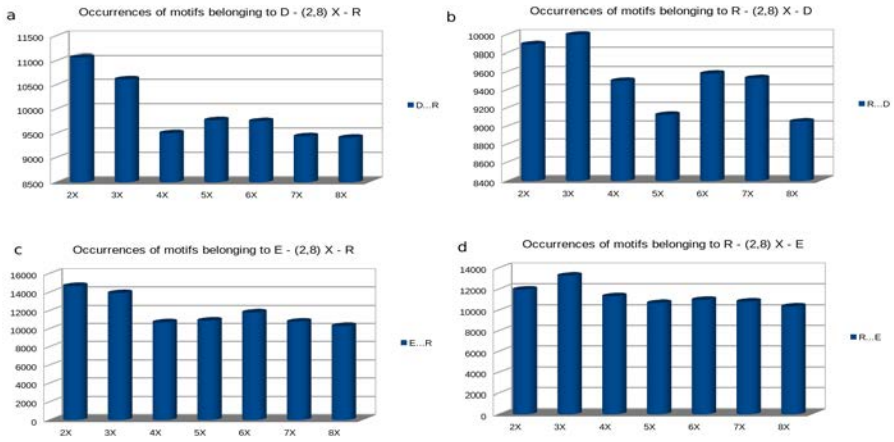

Figure 1. Comparative occurrences of the 28 motifs. (a) Occurrences of motifs belonging to D-(2,8)X-R. (b) Occurrences of motifs belonging to R-(2,8)X-D. (c) Occurrences of motifs belonging to E-(2,8)X-R. (d) Occurrences of motifs belonging to R-(2,8)X-E.

5.5.1 Analysis of motifs in helix group.

Motifs occurrences from above 28 patterns occurring in helix group were studied comparatively to assess their statistical significance (Figure 2).

For all the four types the highest occurrence of interacting motifs was detected in 3X. Higher occurrence of interacting motifs in 2X was observed for D-(2X)-R and E-(2X)-R as compared to their reverse position motif. Compared to others belonging to 8X category, higher percentage of interacting motifs were identified for E-(8X)-R.

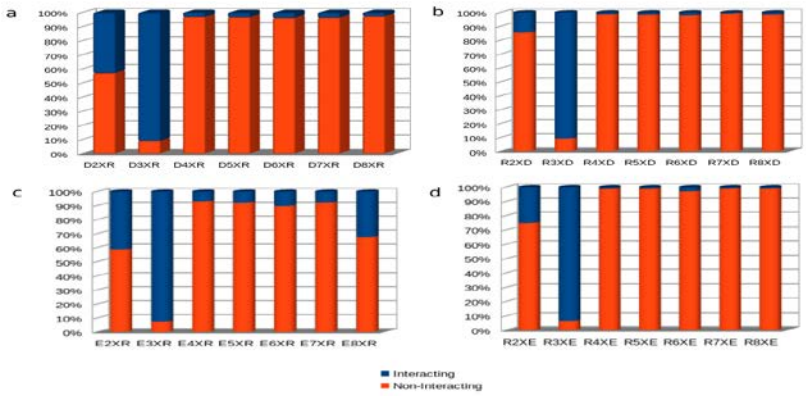

Figure 2. The distribution of interacting and non-interacting occurrence of the 28 motifs in helix group. (a) D-(2,8)X-R. (b) R-(2,8)X-D. (c) E-(2,8)X-R. (d) R-(2,8)X-E.

5.5.2 Analysis of motifs in sheets group.

Motifs occurrences from above 28 patterns belonging to sheets group were statistically analyzed.

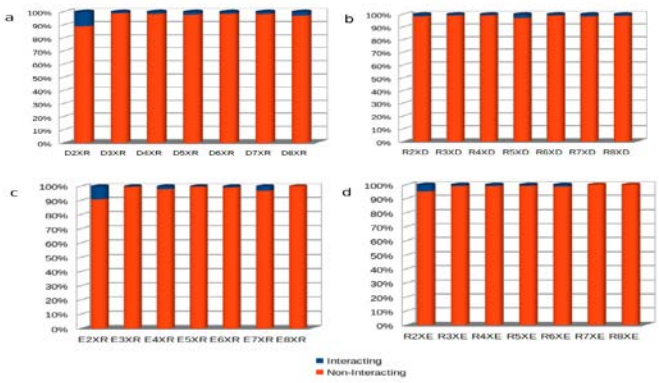

Figure 3. The distribution of interacting and non-interacting occurrence of the 28 motifs in sheets group. (a) D-(2,8)X-R. (b) R-(2,8)X-D. (c) E-(2,8)X-R. (d) R-(2,8)X-E.

As can be expected very few occurrences of interacting motifs were recorded due to separation of Asp or Glu from Arg present in stretched strands (Figure 3). Most occurrences of such motifs with interaction were observed for D-(2X)-R, E-(2X)-R and R-(2X)-E.

5.5.3 Analysis of motifs in irregular structural group.

Motifs occurrences from above 28 patterns in the irregular structural group were studied for the presence of interactions (Figure 4).

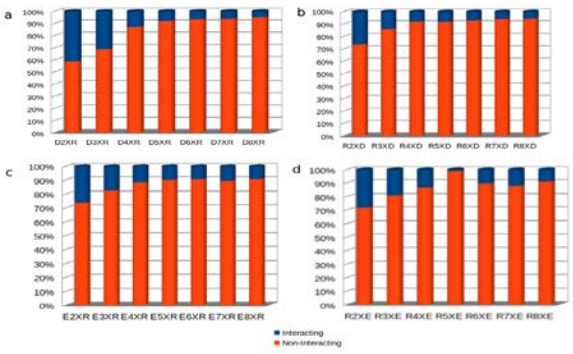

Figure 4. Distribution of interacting and non-interacting occurrence of the 28 motifs in irregular structural group. (a) D-(2,8)X-R. (b) R-(2,8)X-D. (c) E-(2,8)X-R. (d) R-(2,8)X-E.

An overall decreasing trend of interacting motifs from 2X to 8X was identified for D-(2,8)X-R, E-(2,8)-X-R and R-(2,8)X-D. In case of R-(2,8)X-E, this trend was observed only for 2X to 5X; higher percentage of interacting motifs were identified for R-(6X)-E, R-(7X)-E and R-(8X)-E as compared to R-(5X)-E.

Based on the analysis carried out above, it was realized that motifs wherein Asp/Glu and Arg were separated by 2 and 3 residues, show higher instances of local fold stabilization through interactions in all different structures. Hence it was decided to investigate these motifs in detail.

5.6. Detailed analysis of D/E-(2X)-R and R-(2X)-D/E motifs in helices.

Occurrences of four motifs D-(2X)-R, R-(2X)-D, E-(2X)-R and R-(2X)-E in the helix and irregular structures were studied in detail for conformational and interactions features (Tables 6 & 7).

Table 6. Detailed structural analysis of D/E-(2X)-R and R-(2X)-D/E motifs.

Motif	Helix		Sheet		Irregular Structures		Total with ints.
(Total)	Interactions (ints.)		Interactions (ints.)		Interactions (ints.)		
	1	>1	1	>1	1	>1	
D–(2X)–R	1521	432	32	19	1799	720	4523
Total	1953		51		2519		
R–(2X)–D	277	71	10	3	950	941	2252
Total	348		13		1891		
E–(2X)–R	2036	1476	22	40	987	472	5033
Total	3512		62		1459		
R–(2X)–E	1048	258	27	5	1098	636	3072
Total	1306		32		1734		

5.6.1 Analysis of D-(2X)-R motif with two H-bonds in helices.

For the D-(2X)-R motifs in helix, 1953 occurrences (Table 6) were identified with one or more H-bond interactions. These could be further classified as 413 having two hydrogen bonds and 19 with three hydrogen bonds (Table 7). The Arg side chain conformation analysis showed a wide variation in the conformations. The partially folded side chain conformation t t g+ g+ was found to occur in 22% (89 observations) (Figure 5). The Asp χ_1 was observed to be t. The hydrogen bonds in these motifs belonged to Type D (Figure 6).

Table 7. Detailed Interaction analysis of D/E-(2X)-R and R-(2X)-D/E motifs.

Motif	Helix		Sheet		Irregular		Total with ints.
(Total)	No. of H-bonds		No. of H-bonds		No. of H-bonds		
	2	3	2	3	2	3	
D – (2X) – R	413	19	16	3	566	154	1171
Total	432		19		720		
R – (2X) – D	59	12	3	0	628	313	815
Total	71		3		741		
E – (2X) – R	1385	91	40	0	372	100	1988
Total	1476		40		472		
R – (2X) – E	220	38	5	0	519	117	534
Total	258		5		636		

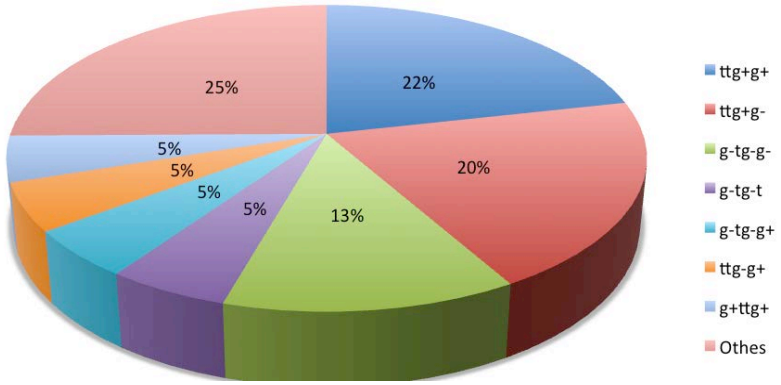

Figure 5. Distribution of Arg side chain conformations in D-(2X)-R motifs in helix group with two H-bonds.

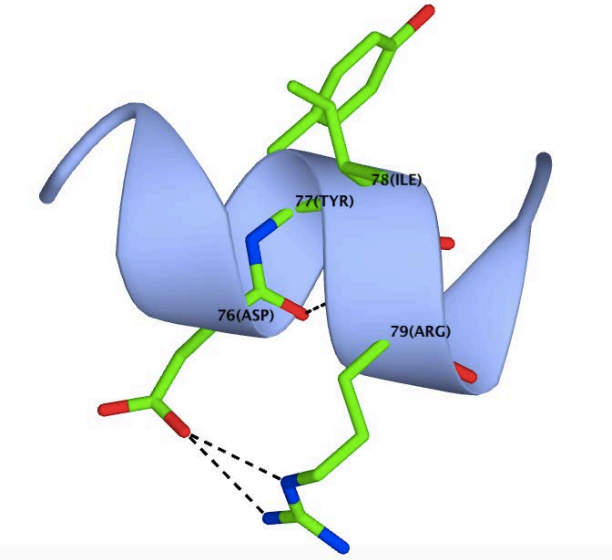

Figure 6. The DYIR sequence in 1RUT showing the Type D bonding.

5.6.2 Analysis of E-(2X)-R motif with two H-bonds in helices.

The 1385 observations (Table 7) identified were studied for the side chain conformations and hydrogen bonding. The most prominent side chain conformation was observed to be similar to D-(2X)-R i.e. t t g+ g+ in 64% (873 observations) of the total.

The Glu χ_1 in these motifs was t observed in 669 cases while in others it was g-, the Glu χ_1 was found to be g+ in 663 and g- in others. Although the Arg side chain conformation was same, the hydrogen bonding was found to be Type B (Figure 8) here in lieu of the Type D observed earlier.

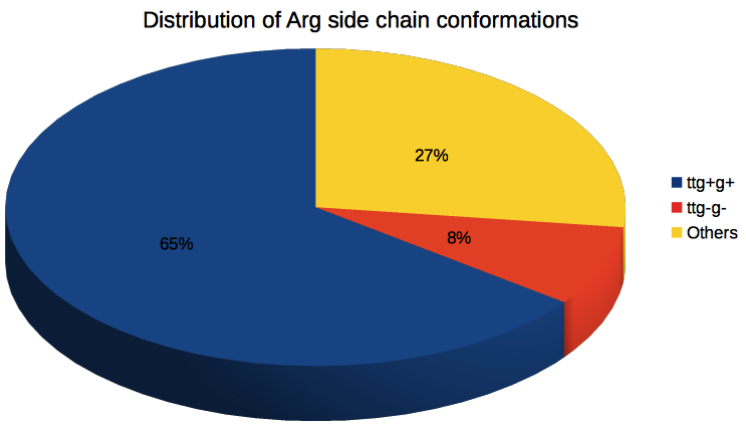

Figure 7. Distribution of Arg side chain conformations in E-(2X)-R motifs in helix group with two H-bonds.

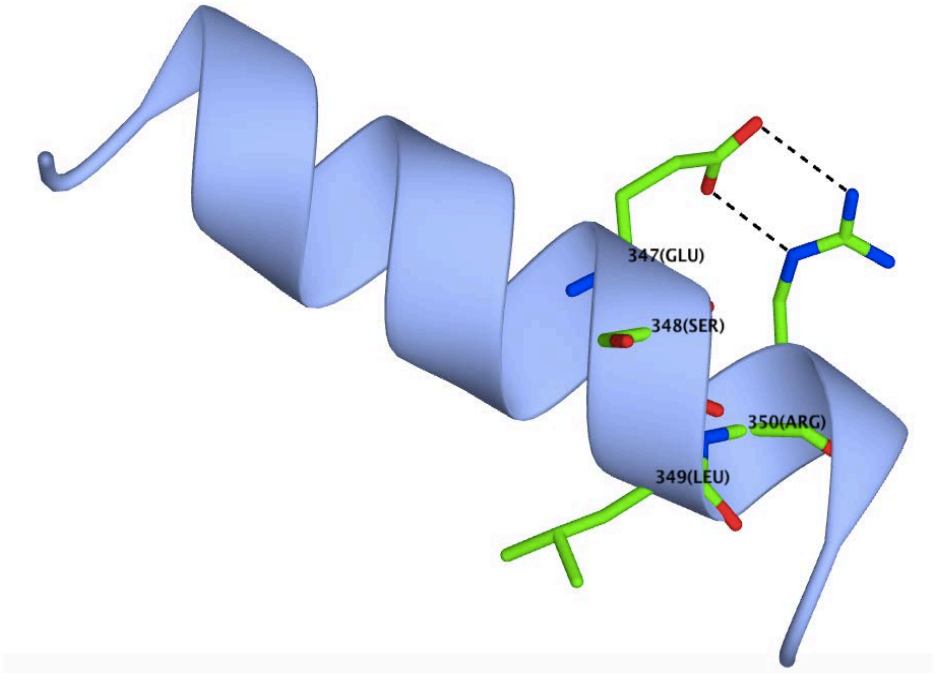

Figure 8. The ESLR sequence in 2IAG showing the Type B H-bonding.

5.6.3 Analysis of R-(2X)-E motif with two H-bonds in helices.

220 observations (Table 7) were studied in R-(2X)-E motif with two H-bonds occurring in helices. The Arg side chain analysis revealed the highly folded conformation g- g- g+ g+ (41%: 90 observations) to be the most prominent (Figure 9).

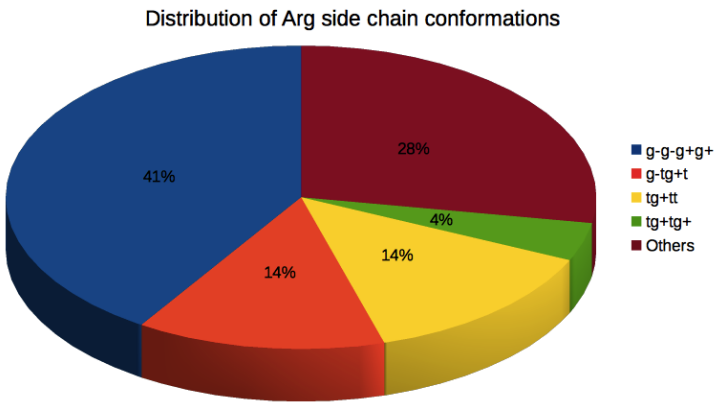

Figure 9. Distribution of Arg side chain conformations in R-(2X)-E motifs in helix group with two H-bonds.

The Glu χ_1 in these motifs was observed to be g-, while the Glu χ_2 here was found to be t. The hydrogen bonding was found to be Type B (Figure 10).

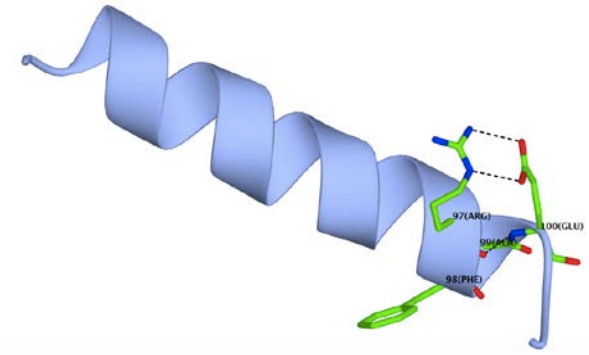

Figure 10. The RFAE sequence in 3PMM showing Type B hydrogen bonding.

5.7. Detailed analysis of D/E-(2X)-R and R-(2X)-D/E motifs in irregular regions.

For all the four patterns motifs were found to occur in irregular regions with two or three hydrogen bonds.

5.7.1 Analysis of D-(2X)-R motif with two H-bonds in irregular structural regions.

566 occurrences of this motif were identified having two hydrogen bonds. The motifs were studied for the Arg side conformation. The partially folded side chain conformation g- t t g+ was found to occur in 17% (95 observations) of the total

observations (Figure 11). The Asp χ_1 here was found to be g-. The Ramachandran plots of the X1 and X2 residues were found to lie in α-helix region (Figure 12a,b). The hydrogen bonding was found to be similar to Type B, but with one side chain – side chain bond being replaced by a main chain – side chain bond, Arg (N) – (OD1) Asp (Figure 13).

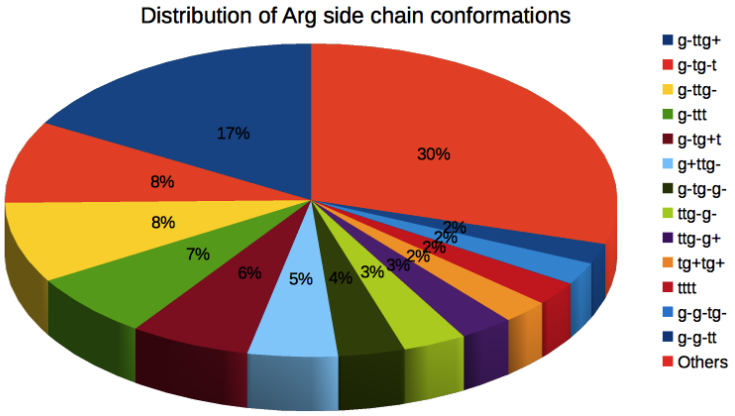

Figure 11. Distribution of Arg side chain conformations (in percentages) in D-(2X)-R motifs in irregular regions with two H-bonds.

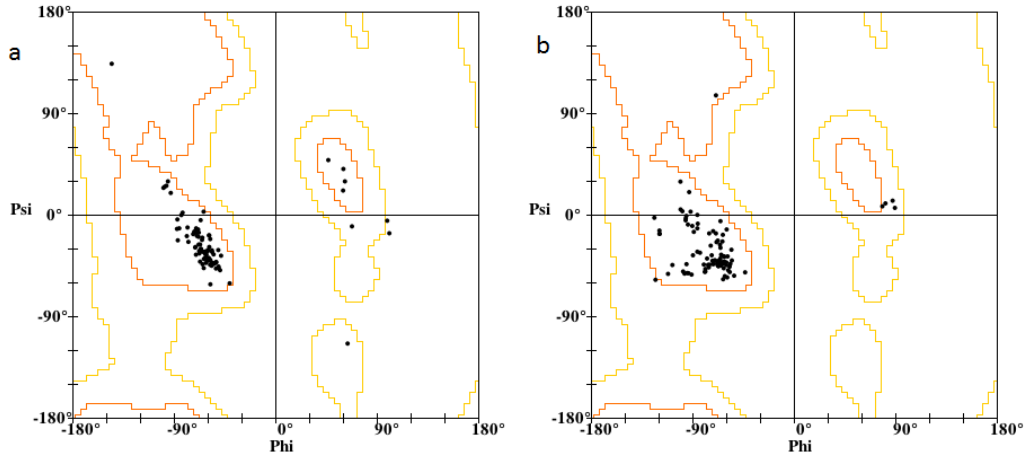

Figure 12. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

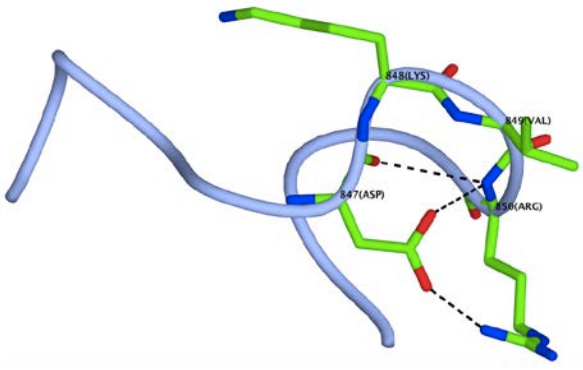

Figure 13. The DKVR sequence in 4IUG showing the two H-bonds described in Section 5.7.1.

The side chain conformation g- t g- t was found to occur in 48 observations. Here the Asp χ_1 here was also found to be g-. The Ramachandran plots the X1 and X2 residues were found to lie in α-helix region (Figure 14a,b). The hydrogen bonding was found to be similar to that described above.

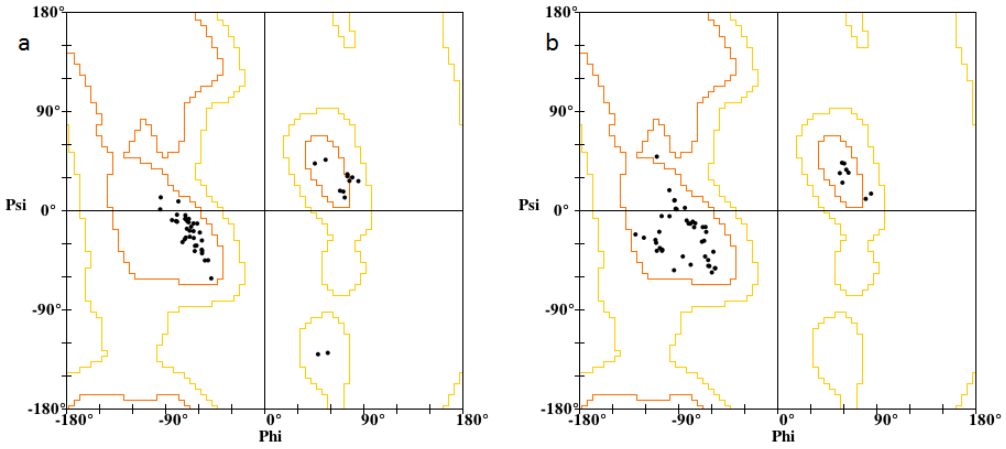

Figure 14. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

5.7.2 Analysis of D-(2X)-R motif with three H-bonds in irregular regions.

154 occurrences of this motif were identified having three hydrogen bonds. The partially folded Arg side chain conformation g- t g- t was found to occur in 34% (51 observations) of the total (Figure 15). The Asp χ_1 here assumed the conformation g- in order to interact with the Arg side chain.

The Ramachandran plots of the X1 and X2 residues were found to lie in α-helix region while a few were found to lie in left-handed helical region (Figure 16a,b). The hydrogen bonding was predominantly Type D with an additional main chain – side chain H-bond, while only in 6 cases it was found to be Type B. Although the hydrogen bonding is of Type D, the folding of the motif strikes a resemblance to that observed in D-X-R.

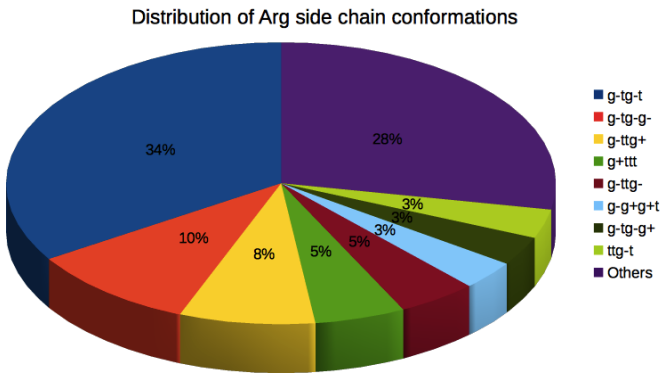

Figure 15. Distribution of Arg side chain conformations in D-(2X)-R motifs in irregular regions with three H-bonds.

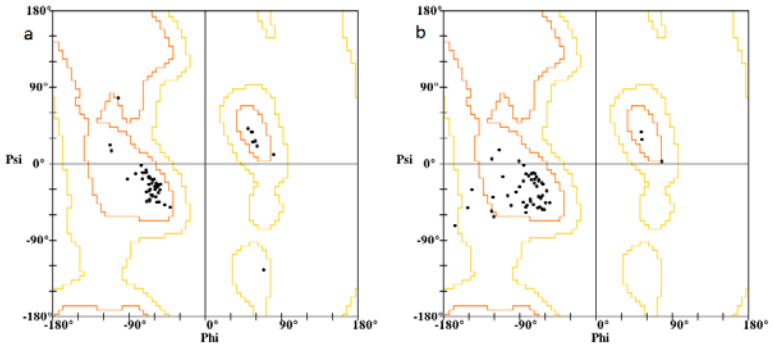

Figure 16. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

5.7.3 Analysis of R-(2X)-D motif with two H-bonds in irregular structural regions.

We have identified 628 occurrences of this motif with two H-bonds. The Arg side chain analysis identified the nearly extended g- t t t conformation to occur in 13% (79 observations) of the wide variety of conformations observed (Figure 17). The Asp χ_1 here was found to be g-.

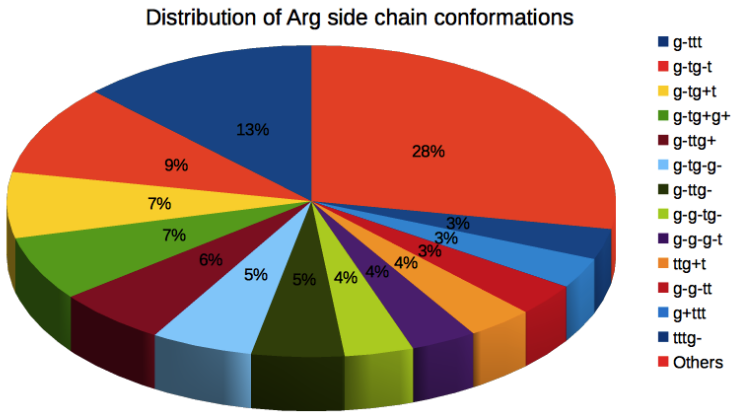

Figure 17. Distribution of Arg side chain conformations in R-(2X)-D motifs in irregular structural regions with two H-bonds.

The Ramachandran plot for X1 residue was found to cluster mainly in α-helix region with a second cluster in the β-sheet region suggesting a more extended backbone (Figure 18a). The X2 plot was found to cluster in the α-helix region with few occurring in left-handed helix region (Figure 18b).

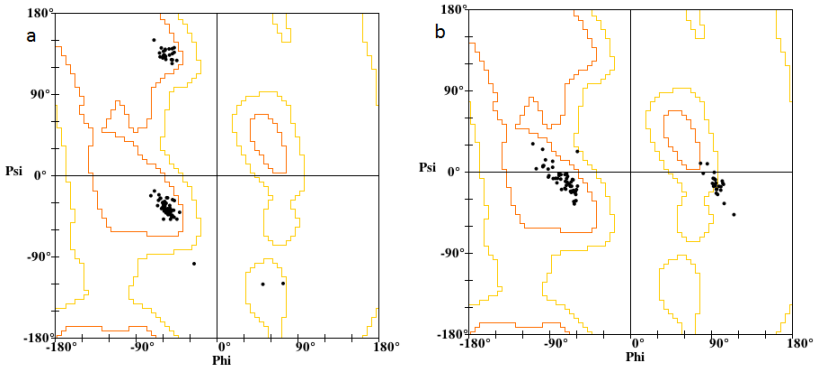

Figure 18. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

The hydrogen bonding for these motifs was observed to have one main chain – side chain bond, Arg (N) – (OD1) Asp and other being a main chain – main chain bond (Figure 19).

The Arg side chain conformation g- t g- t was found in 9% (59 observations) of the total (Figure 17). The Asp χ_1 here was again found to be g-. The Ramachandran plots were found to be exactly as explained for the earlier conformation (Figure 20a,b). The

hydrogen bonding was also found to remain the same. This revealed that even though the Arg side chain conformation varied the backbone folding remained conserved.

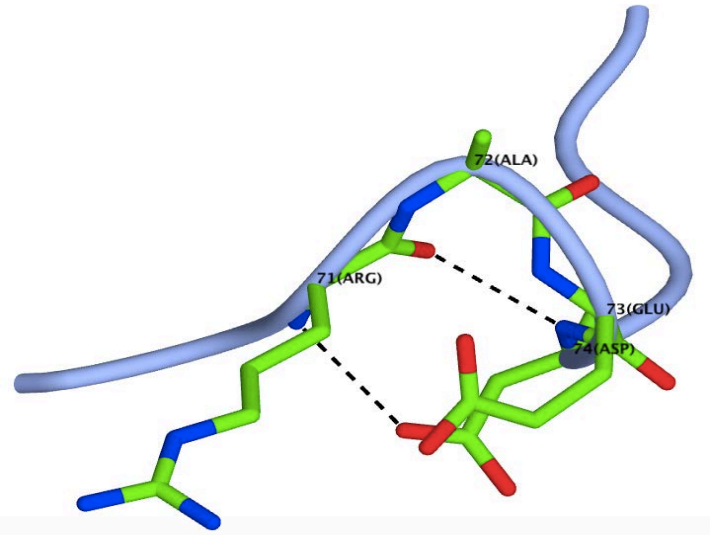

Figure 19. The RAED sequence in 1GNU showing the two H-bonds described in Section 5.7.3.

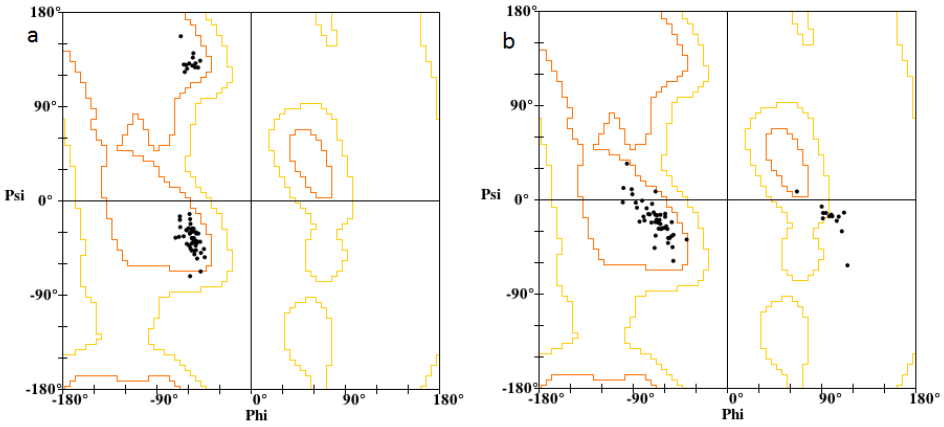

Figure 20. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

5.7.4 Analysis of R-(2X)-D motif with three H-bonds in irregular structural regions.

The 313 motifs identified were studied in detail. The partially folded conformation g- t g- t occurred in 64% (134 observations) of the total (Figure 21). The Asp χ_1 conformation was observed to be g-.

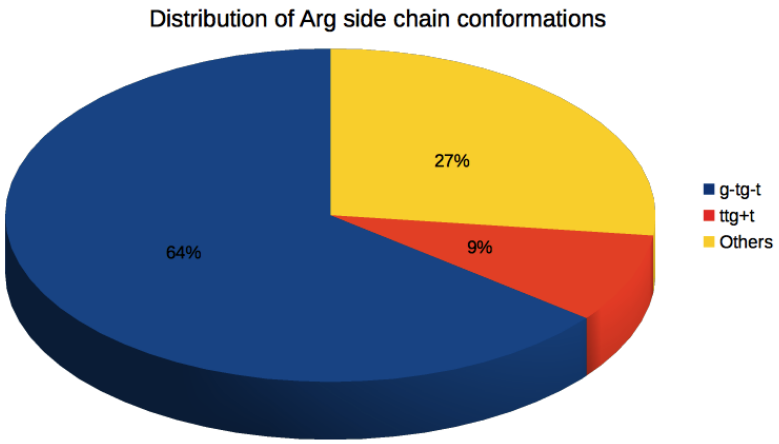

Figure 21. Distribution of Arg side chain conformations in R-(2X)-D motifs in irregular structural regions with three H-bonds.

The Ramachandran plot for the X1 and X2 residues were found to occupy exactly same regions as explained in Section 5.7.3 (Figure 22). This set of motifs was thus observed to be a part of the earlier set. However the hydrogen bonding here was found to be Type D along-with a main chain – main chain bond (Figure 23).

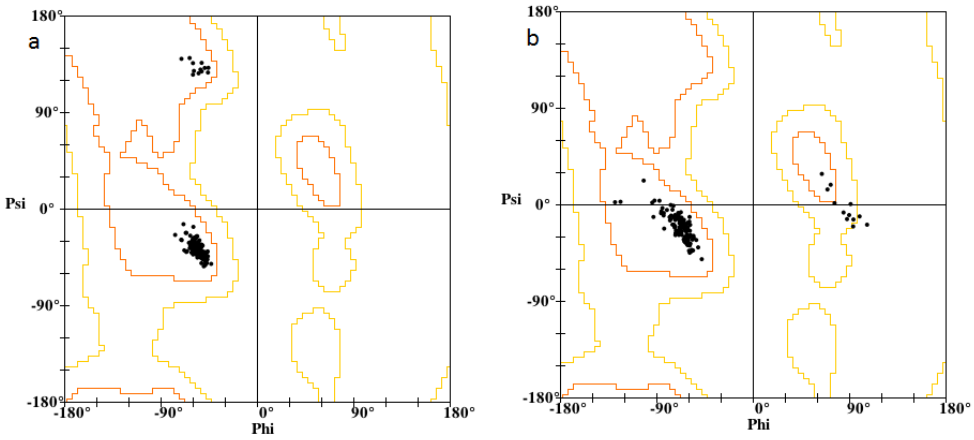

Figure 22. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

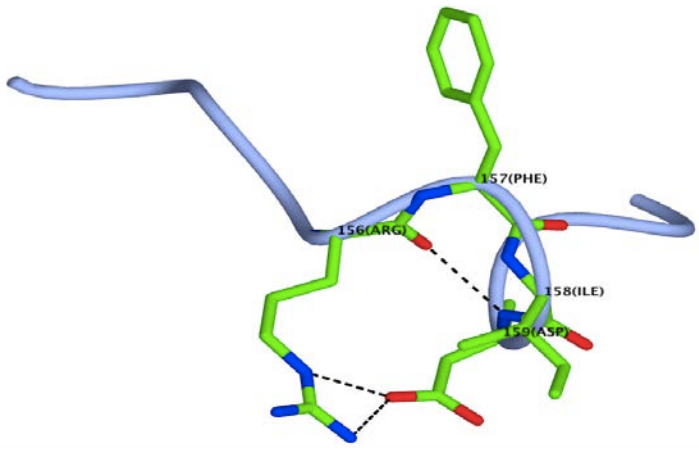

Figure 23. The RFID sequence in 1NKG showing Type D hydrogen bonding.

5.7.5 Analysis of E-(2X)-R motif with two H-bonds in irregular regions.

The next set of motifs was Asp replaced by Glu and separated by two amino acids from Arg occurring in irregular regions and having two hydrogen bonds. From the 372 occurrences, 10% (37 observations) were found to have the Arg side chain conformation g- t g+ t. The Glu χ_1 conformation was observed to be g-, while the χ_2 conformation was found to be variable.

The Ramachandran plot for the X1 and X2 residues were found to cluster in the α-helix region (Figure 25a,b). The hydrogen bonding was observed to be a variant of Type B with one side chain – side chain bond being replaced by a main chain – main chain bond (Figure 26).

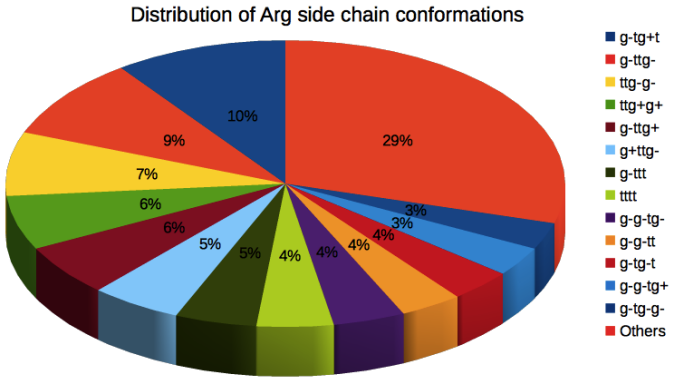

Figure 24. Distribution of Arg side chain conformations in E-(2X)-R motifs in irregular structural regions with two H-bonds.

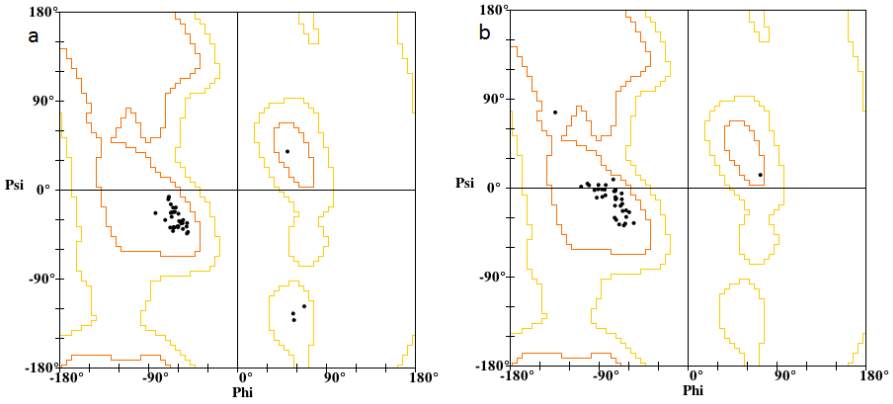

Figure 25. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

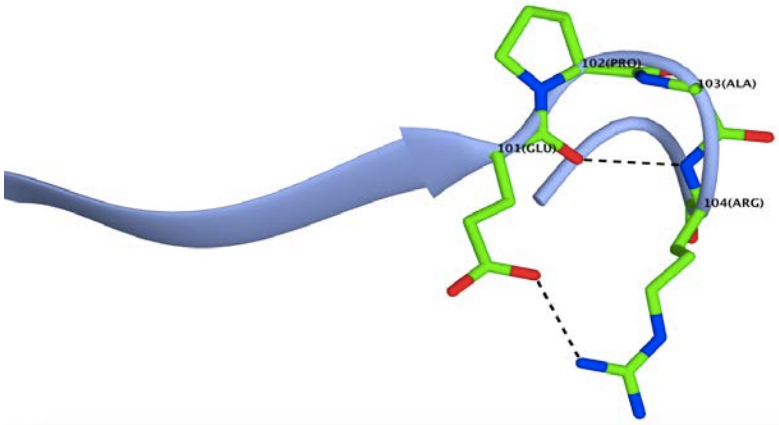

Figure 26. The EPAR sequence in 4MO4 showing hydrogen bonding described in Section 5.7.5.

5.7.6 Analysis of E-(2X)-R motif with three H-bonds in irregular structural regions.

The 100 motifs identified were studied in detail. The side chain analysis revealed 35% to have partial folded Arg side chain conformation t t g- g- (Figure 27). Again the Glu χ_1 conformation was observed to be g-, while the χ_2 conformation was found to be variable. The Ramachandran plot for the X1 residue clustered in α-helix region with few in the β-sheet region (Figure 28a). The plot for X2 was found to lie in α-helix and left-handed helix regions (Figure 28b). As previously, in most cases, residues lying in the left-handed region were identified to be Gly.

The hydrogen bonding in these motifs was found to be Type B along-with a main chain – main chain bond (Figure 29).

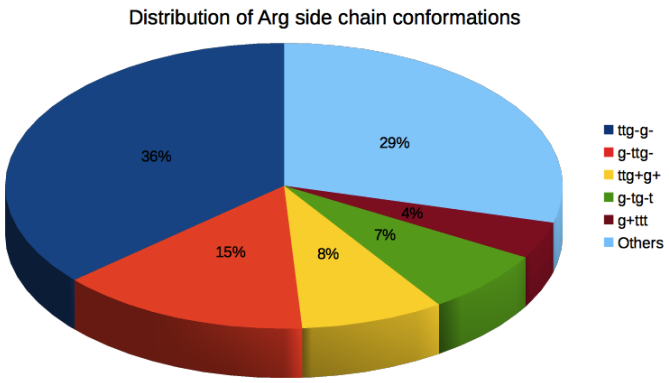

Figure 27. Distribution of Arg side chain conformations in E-(2X)-R motifs in irregular structural regions with three H-bonds.

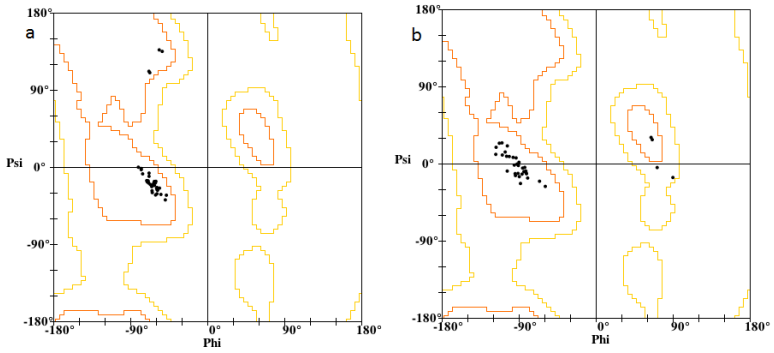

Figure 28. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

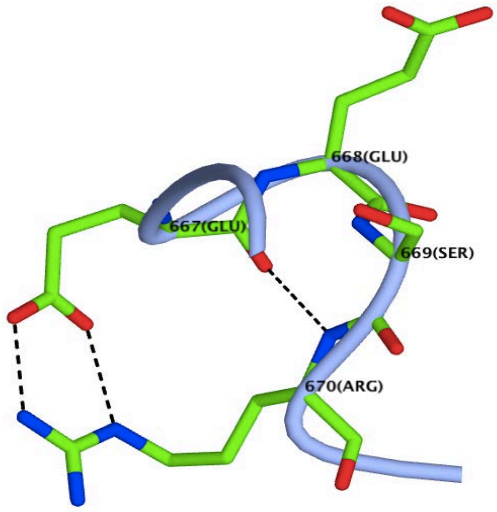

Figure 29. The EESR sequence in 3N75 showing Type B hydrogen bonding.

5.6.7 Analysis of R-(2X)-E motif with two H-bonds in irregular structural regions.

The 519 motifs belonging to this group were studied for side chain conformations and hydrogen bonds. In 14% (73 observations) of the cases, the Arg side chain conformation was g- t g+ t and the Glu χ_1 conformation was observed to be g- and χ_2 conformation was t (Figure 30).

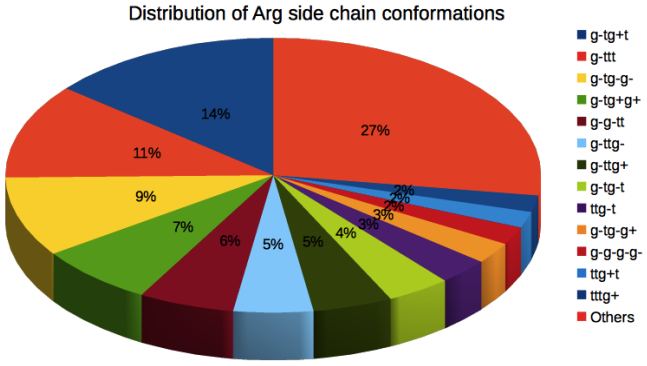

Figure 30. Distribution of Arg side chain conformations in R-(2X)-E motifs in irregular structural regions with two H-bonds.

The Ramachandran plot for the X1 residue was found to lie in the α-helix region while the X2 residue occupied the α-helix and left-handed helix regions (Figure 31a,b). The hydrogen bonding was found to be Type B (Figure 32).

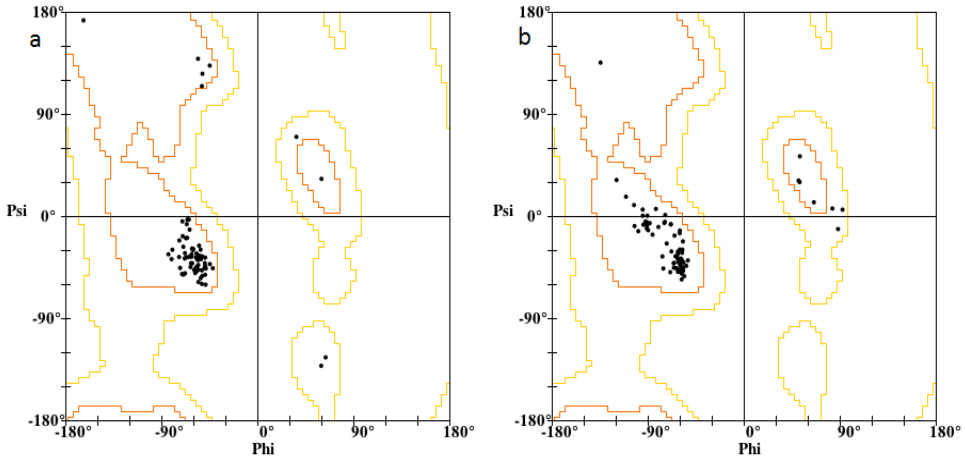

Figure 31. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

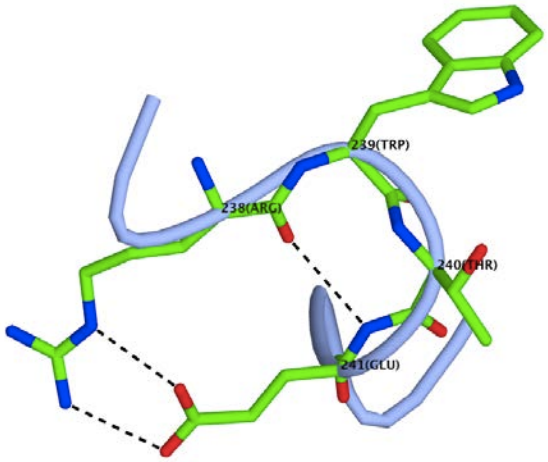

Figure 32. The RWTE sequence in 3EOJ showing Type B hydrogen bonding.

5.6.8 Analysis of R-(2X)-E motif with three H-bonds in irregular structural regions.

117 occurrences of the motif were observed with three hydrogen bonds occurring in irregular structural regions. Of these 31% (36 observations) were found to have the g- t g+ t with Glu χ_1 and χ_2 conformation being g- and t, respectively (Figure 33). The Ramachandran plot for the X1 residue was found to lie in the α-helix region while the X2 residue occupied the α-helix and left-handed helix regions (Figure 34a,b) The hydrogen bonding was found to be Type B (Figure 35). These motifs were found to be similar to the R-(2X)-E motifs with two interactions explained in Section 5.6.7.

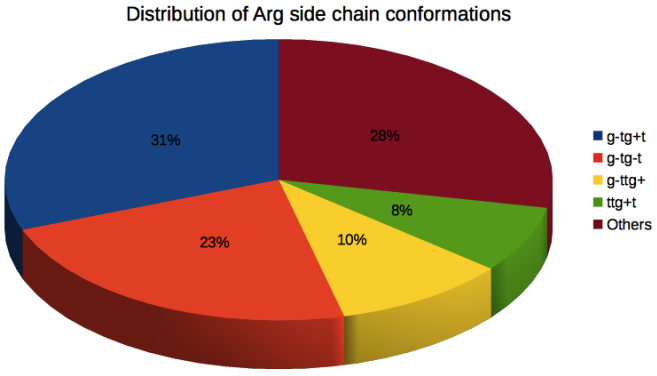

Figure 33. Distribution of Arg side chain conformations in R-(2X)-E motifs in irregular structural regions with three H-bonds.

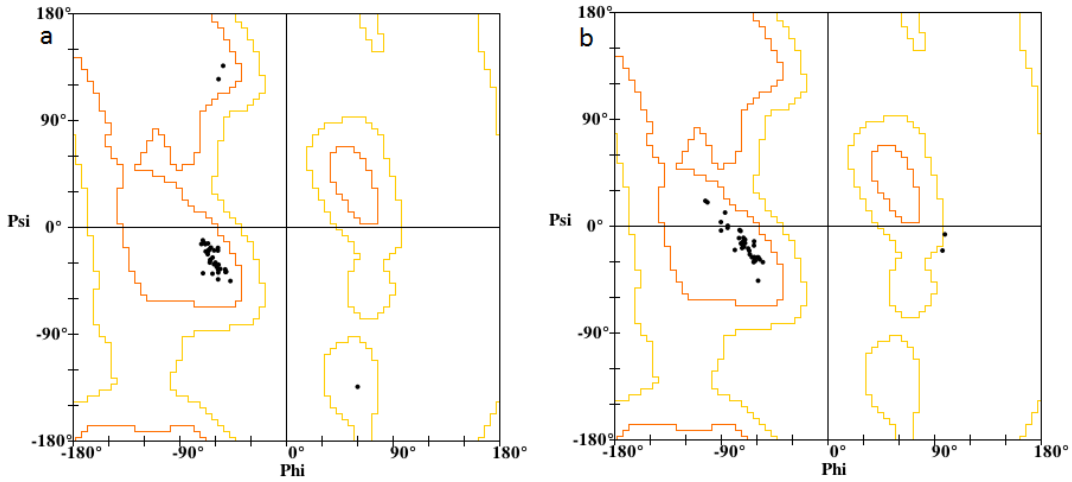

Figure 34. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue.

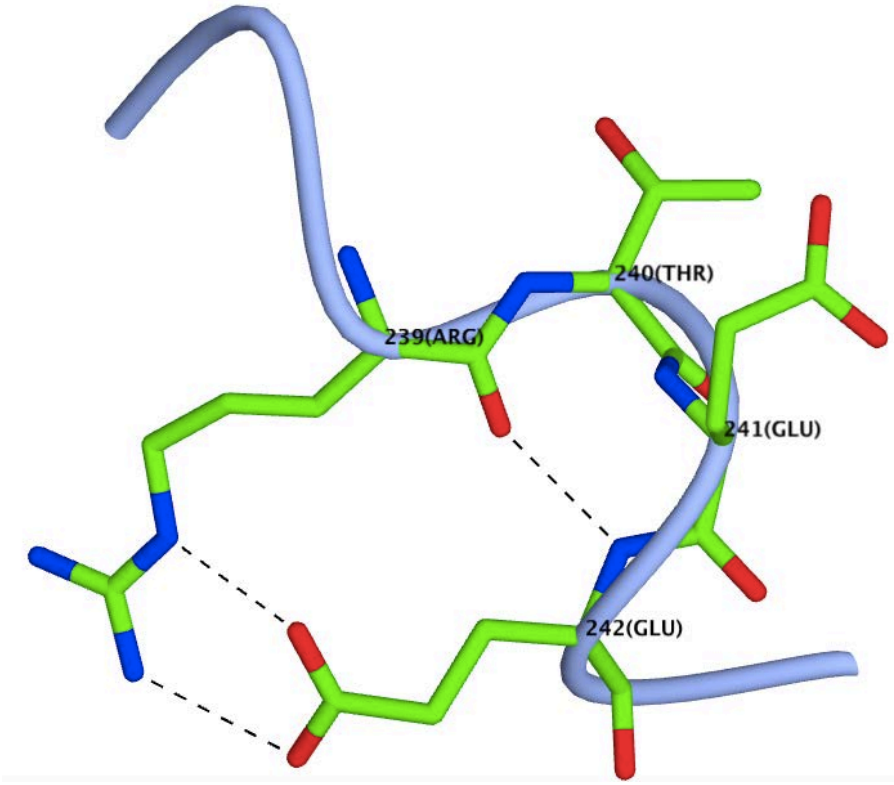

Figure 35. The RTEE sequence in 4JOQ showing Type B hydrogen bonding

5.7. Detailed analysis of D/E-(3X)-R and R-(3X)-D/E motifs in helix group.

Occurrences of four motifs namely D-(3X)-R, R-(3X)-D, E-(3X)-R and R-(3X)-E in the helix and irregular structures were studied in detail for conformational and interactions features (Tables 8 & 9).

5.7.1 Analysis of D-(3X)-R motif with two H-bonds.

3591 were found to be involved in one or more hydrogen bonded interactions. From these 261 occurrences were found to have two H-bonds. The occurrence of the characteristic helical i +4 $\rightarrow$ i, R (N) – (O) D bond was not considered in this analysis.

Table 8. Detailed structural analysis of D/E-(3X)-R and R-(3X)-D/E motifs.

Motif	Helix		Sheet		Irregular Structures		Total with ints.
(Total)	Interactions (ints.)		Interactions (ints.)		Interactions (ints.)		
	1	>1	1	>1	1	>1	
D–(3X)–R	3330	261	4	1	1370	600	5566
Total	3591		5		1970		
R–(3X)–D	2575	438	7	0	684	214	3624
Total	2719		7		898		
E–(3X)–R	5876	672	2	0	770	226	9085
Total	8087		2		996		
R–(3X)–E	5748	282	3	2	961	250	7246
Total	6030		5		1211		

For the 261 motifs showing two H-bonds, the Arg side chain analysis showed the folded conformation g- t g- g+ to dominate (Figure 36). In most cases, except four of

them, the Asp χ_1 was observed to be t. In the four cases namely 2XID:A264, 2YIE:A86, 4DOY:A379 and 4JEK:A379, the Asp χ_1 was g+.

The hydrogen bonding in 36 motifs was found to be of Type D (Figure 37). In 20 cases it was observed to be Type B with R (N) – (O) D bond. Only in three cases a side chain – side chain bond was found to be replaced by a main chain – side chain, D (N) – (OD1) D bond.

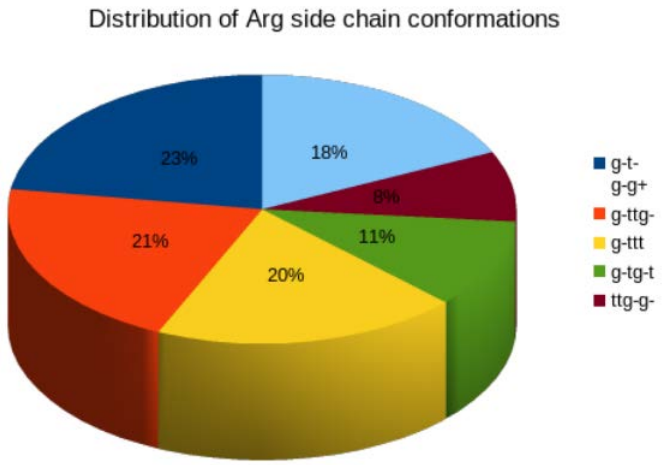

Figure 36. Distribution of Arg side chain conformations in D-(3X)-R motifs in helix group with two H-bonds.

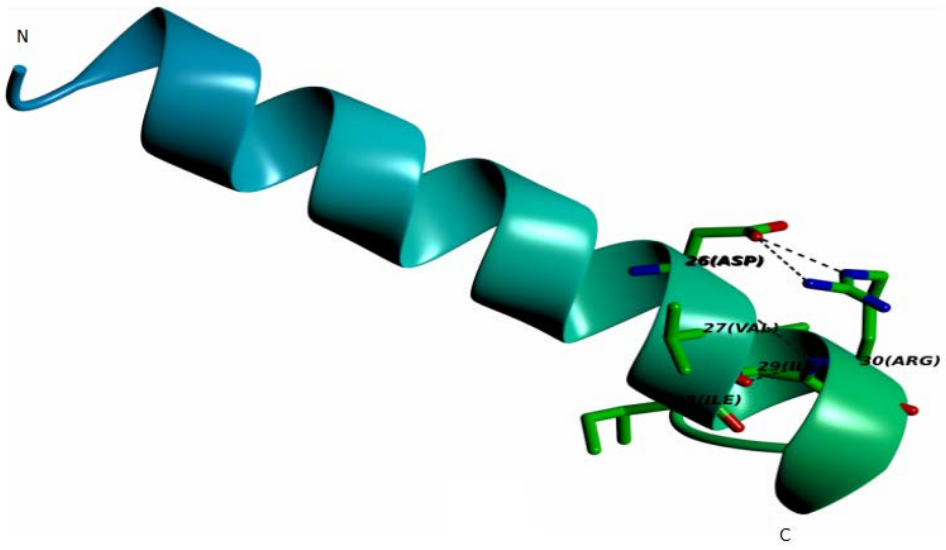

Figure 37. The DVIIR sequence in 1GYX showing the Type D bonding.

5.7.2 Analysis of R-(3X)-D motif in helix group with two interactions.

Of the 2996 occurrences, about 90% were found to involve one or more H-bonds. For the R-(3X)-D, 144 were found to have two H-bonds.

67 occurrences of the conformations were identified. The backbone fold of the motifs was found to be conserved The Arg side chain conformation was observed to be t g+ g- t.

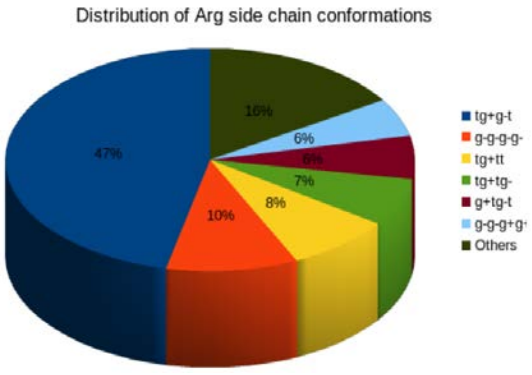

Figure 38. Distribution of Arg side chain conformations in R-(3X)-D motifs in helix group with two H-bonds.

Table 9. Detailed Interaction analysis of D/E-(3X)-R and R-(3X)-D/E motifs.

Motif	Helix		Sheet		Irregular		Total with ints.
(Total)	No. of H-bonds (ints)		No. of H-bonds		No. of H-bonds		
	2	3	2	3	2	3	
D – (3X) – R	261	0	0	0	371	229	861
Total	261		0		600		
R – (3X) – D	144	0	0	0	135	74	353
Total	144		0		209		
E – (3X) – R	672	0	0	0	191	35	898
Total	672		0		226		
R – (3X) – E	282	0	2	0	185	65	534
Total	282		2		250		

The Asp χ_1 for all the motifs was found to be g-. In 22 occurrences the hydrogen bonding was Type B while in 15 it was Type B. For 29 cases it was observed that one of the side chain - side chain bonds was replaced by a side chain - main chain bond, R (NE) – (O) R (Figure 39).

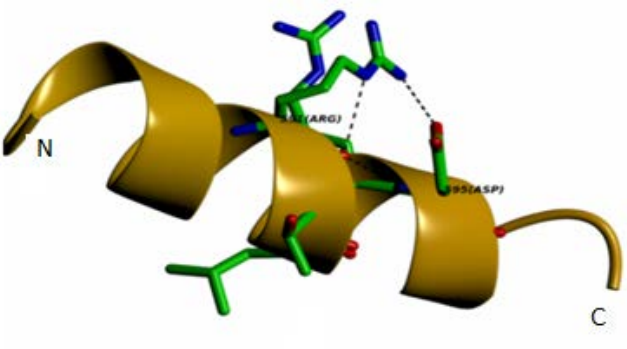

Figure 39. The R-T-L-R-D motif in 4O1P with two H-bonds.

5.7.3 Analysis of E-(3X)-R motif in helix group with two H-bonds.

66% of the motifs with the E-(3X)-R pattern in the helix group were found to be involved in one or more H-bonds. 672 motif had two hydrogen bonding interactions. On analysis of the 672 motifs with two H-bonds for their Arg side chain conformations, two distinct side chain conformations were revealed from a wide range of conformations (Figure 40).

For 322 motifs the Arg side chain conformations were found to be partially folded g- t t g-.

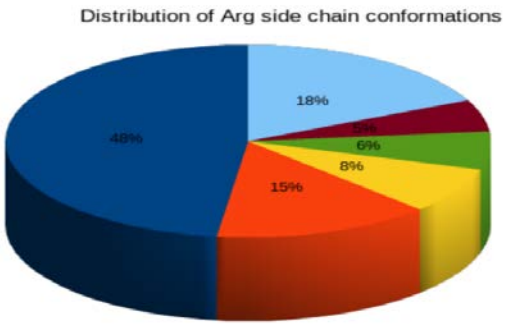

Figure 40. Distribution of Arg side chain conformations (in percentages) in E-(3X)-R motifs in helix group with three H-bonds.

The Glu χ_1 was t for 300 motifs while in 22 cases the angle was g-. On analyzing the 300 motifs for Glu, the χ_2 in 284 cases was t, 15 showed g+ while only in one case it was g-. In the 22 occurrences, 15 showed g-, while only 7 had t. The hydrogen bonding was observed to be Type D for all motifs (Figure 41).

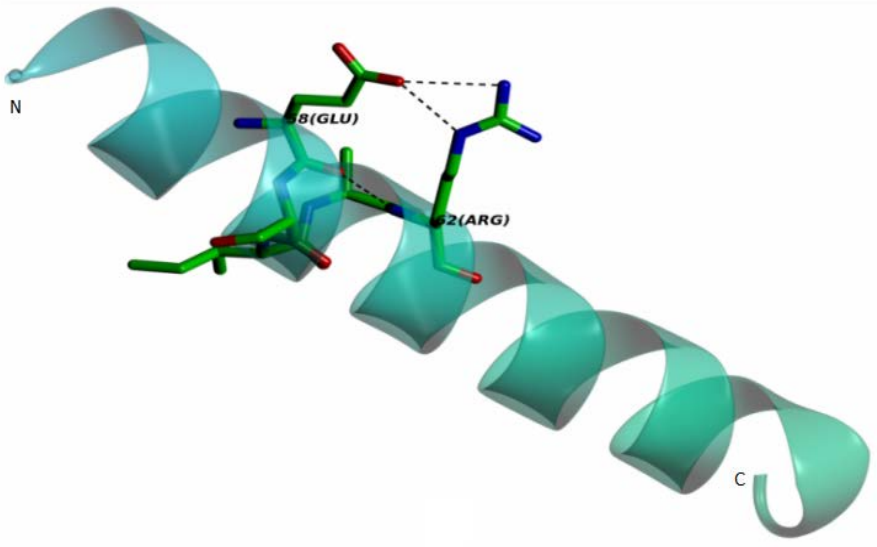

Figure 41. The ESIAR motif in 2AMH showing the hydrogen-bonding pattern.

5.7.4 Analysis of R-(3X)-E motif in helix group with two H-bonds.

While 6030 motifs in this group were identified with interactions ~76% were found to have one hydrogen bond. 282 motifs were found with two H-bonds.

The 282 motifs with two H-bonds were studied for their Arg side chain conformations. Two conformations were found to dominate (Figure 42). The partially folded conformation t g+ t g- was observed in 59 motifs.

The Glu χ_1 was g- in 52 cases while t in remaining 6. For 40 motifs the Glu χ_2 was g-, while it was t in 16 and g+ in 2 occurrences. The hydrogen bonding for all motifs was found to be Type D (Figure 43).

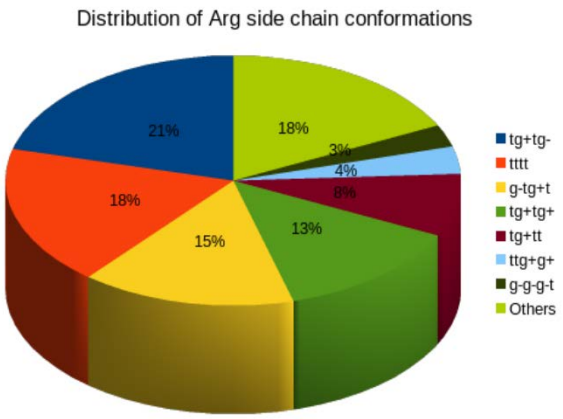

Figure 42. Distribution of Arg side chain conformations in R-(3X)-E motifs in helix group with two H-bonds.

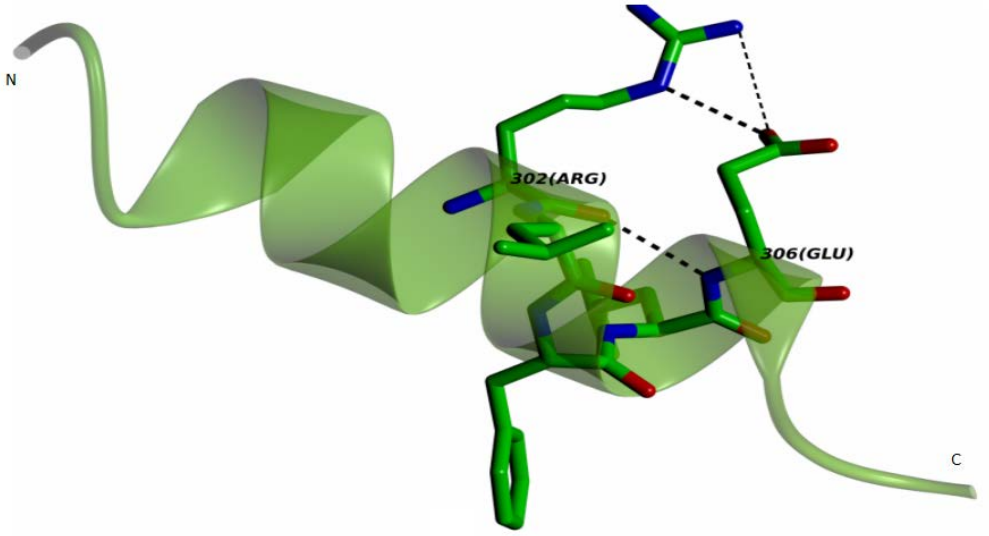

Figure 43. The RLFLE motif in 1M0W showing the hydrogen-bonding pattern.

The next conformation studied was the extended t t t t occurring in 18% of the total. The Glu χ_1 was g- in 44 cases while t in 7. In 47 cases the Glu χ_2 was t, while only in 4 cases it was g-. For 42 motifs the hydrogen bonding was observed to be Type D (Figure 44) and in 9 cases it was Type B with E (N) – (O) R bond. The Ramachandran plots of all the three X residues were in the helix region as a single group.

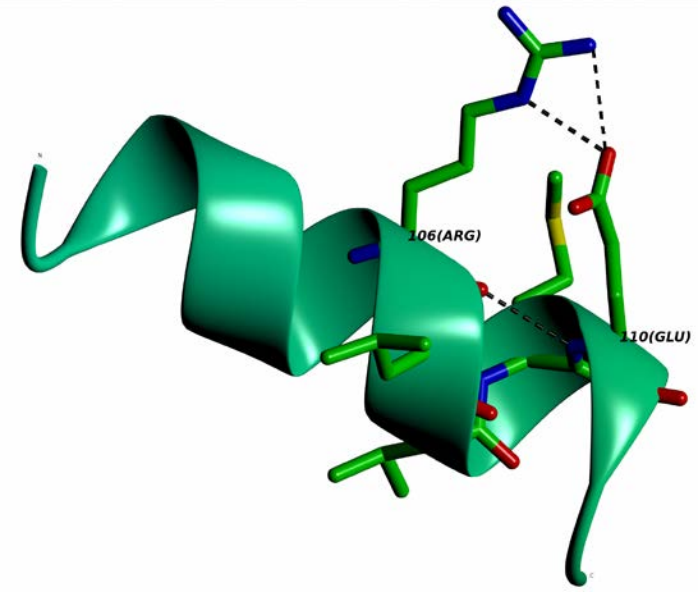

Figure 44. The RLVME in 4OSY showing the hydrogen bonding.

5.8 Detailed analysis of D/E-(3X)-R and R-(3X)-D/E motifs in irregular structures.

Motifs belonging to the four patterns but occurring in irregular structures were studied for their conformational and interaction features.

5.8.1 Analysis of D-(3X)-R motif with three H-bond interactions.

Of the 1970 motifs in irregular structures identified with interactions, 1370 were found to have a single H-bond, while 371 were found with two and 229 with three hydrogen bonded interactions.

The motifs identified with three hydrogen bonds were studied first for conformations of the Asp and Arg residues. While a variety of side chain conformations were identified the partially folded conformation g- t g- t covered 64% (147) of the total (Figure 45a). The Asp χ_1 conformation was found to be g+. The backbone was found to be conserved in some cases (Figure 45b).

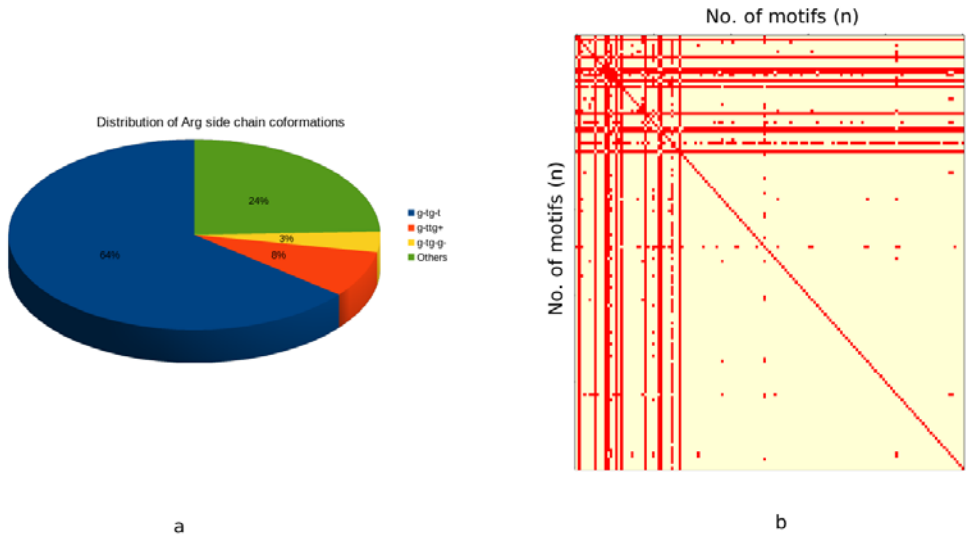

Figure 45. (a) Distribution of Arg side-chain conformations in D-(3X)-R motifs in irregular structures with three H-bonds. (b) The pairwise superimposition graph of the D-(3X)-R motifs in irregular structures with three H-bonds.

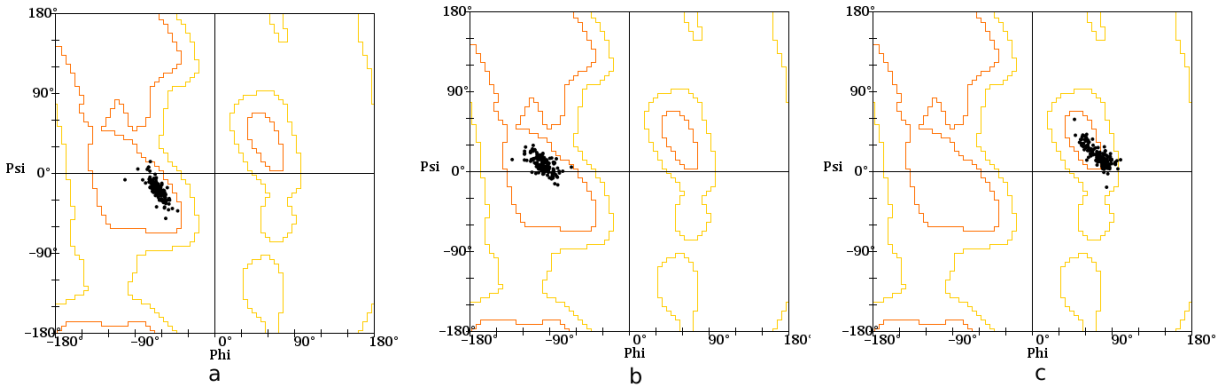

Figure 46. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue. (c) The Ramachandran plot for the X3 residue.

The Ramachandran plot for the X1 and the X2 residues were found to occupy the region of α-helix, while the X3 residue was observed as a single group in the left-handed helix region (Figure 46a-c). The hydrogen bonding in most cases was observed to be a variation of Type B with one side chain – side chain bond, namely Arg (NE) – (OD2) Asp along-with one main chain – side chain bond i.e. Arg (N) – (OD1) Asp and one main chain – main chain bond (Figure 47).

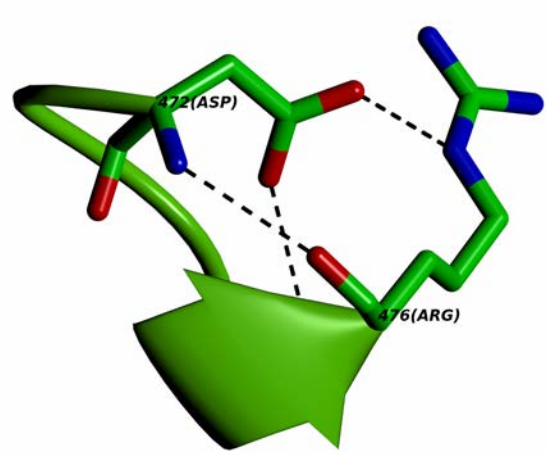

Figure 47. The DCLDR motif in 1EVL showing the hydrogen-bonding pattern.

5.8.2 Analysis of D-(3X)-R motif with two H-bonds.

Next motifs identified with two H-bonds were studied in detail. From the Arg side chain conformations identified for the 371 occurrences, H-bonds for two conformations were studied. The first conformation studied was g- t t g+ (17%) (Figure 48). In this case the Asp χ_1 conformation was g+. Only in one case the Asp χ_1 conformation was g-. The motif backbone conformation shows no conservation over all occurrences.

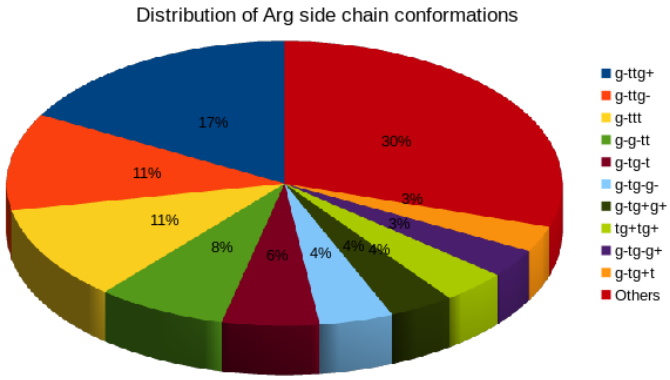

Figure 48. Distribution of Arg side-chain conformations in D-(3X)-R motifs in irregular structures with two H-bonds.

The Ramachandran plot for the X1 residue shows two distinct groups at the α-helix and left-handed helix region. The plot for X2 residue shows two groups, one in the α-helix region and the other in the extended region. Similarly, plot for X3 residue shows grouping in the left-handed helix and extended region (Figure 49a-c). The hydrogen

bonding observed was Type A, namely Arg (NH1) – (OD1) Asp and Arg (NH2) – (OD2) Asp (Figure 50).

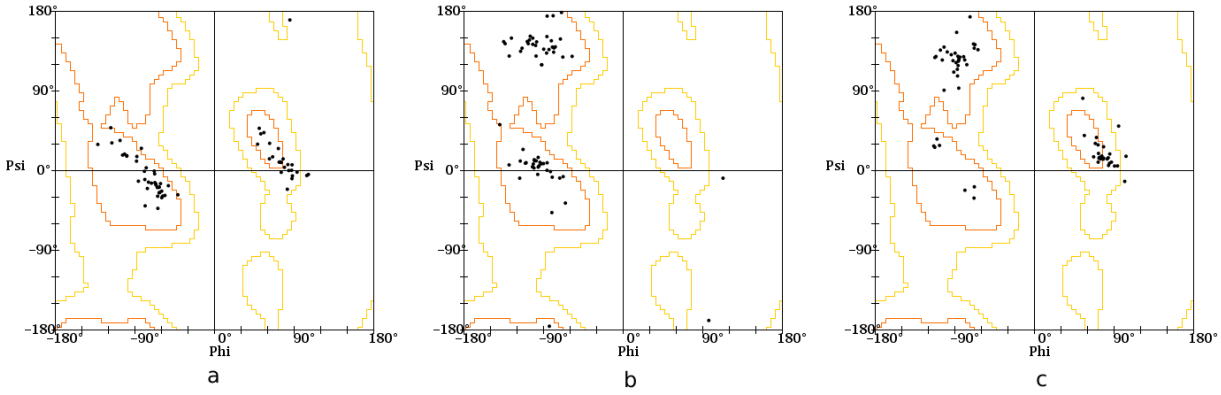

Figure 49. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue. (c) The Ramachandran plot for the X3 residue.

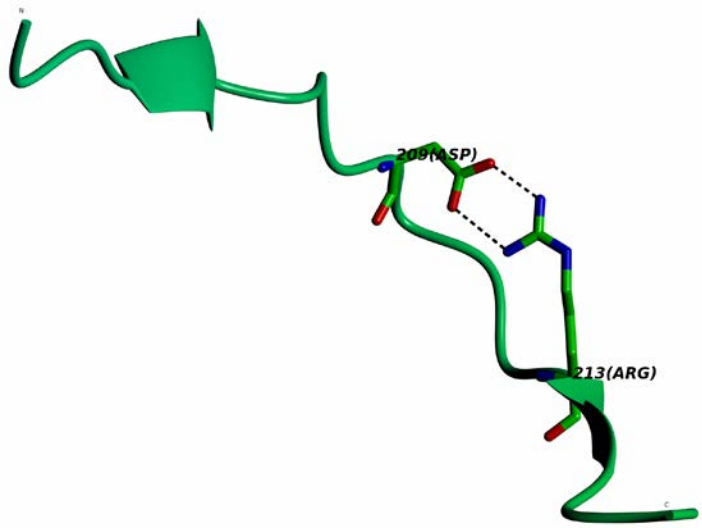

Figure 50. The DCLDR motif in 1EVL showing the hydrogen-bonding pattern.

The next conformation studied was g- t t g- (11%) (Figure 48). The Asp side chain Asp χ_1 conformation was g+. The Ramachandran Plot for the X1 residue shows a spread in the α-helix region while the X2 and X3 residue plots show considerable spread over the α-helix and left-handed helix and extended region (Figure 51). The backbone for the motifs shows no conservation. The hydrogen bonding in this case involved showed a wide variety.

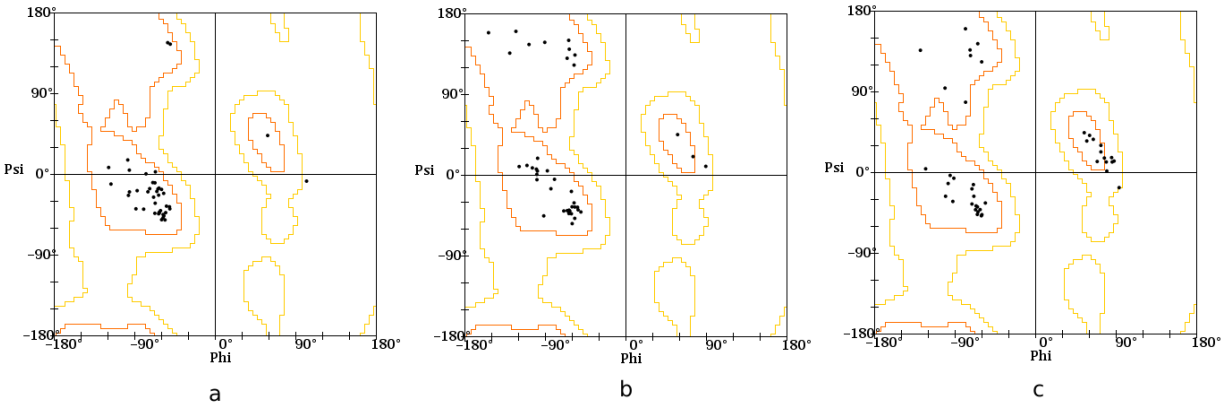

Figure 51. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue. (c) The Ramachandran plot for the X3 residue.

5.8.3 Analysis of R-(3X)-D motif with three H-bonds.

Motifs involving the reversal of positions of Asp and Arg occurring in irregular structures were studied for their conformations and the interactions involved. 74 occurrences were identified with three H-bonds while 135 were identified with two H-bonds.

The occurrences of the motif in irregular structures with three H-bonds were recorded. Studying the side chain conformations of the Arg residue revealed the conformation g- t g- t occurring in 21 motifs (Figure 52). Here the Asp χ_1 conformation was observed to be g-. Only in one case the χ_1 conformation was g+.

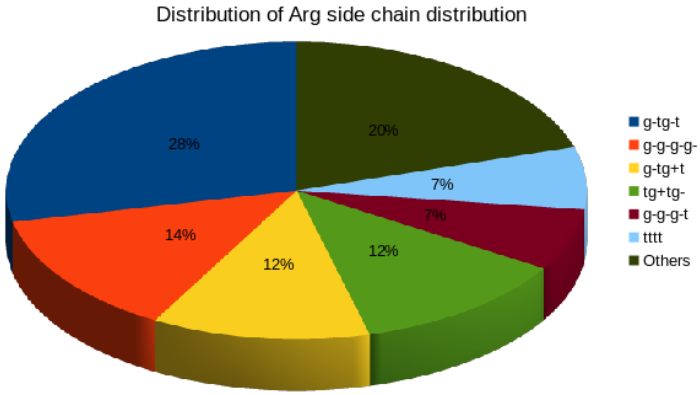

Figure 52. Distribution of Arg side-chain conformations in R-(3X)-D motifs in irregular structures with three H-bonds.

The motif backbone conformation shows no conservation. The Ramachandran plot for the X1 residue occupies primarily the extended region. The plot for X2 and X3 residues do not show clustering in any particular regions (Figure 53a-c). The hydrogen bonding here was observed to belong to Type B along-with a main chain –side chain bond, Arg (N) – (OD1) Asp (Figure 54). This is again similar to D-X-R motif with three H-bonds and Arg rotamer g- t g- t.

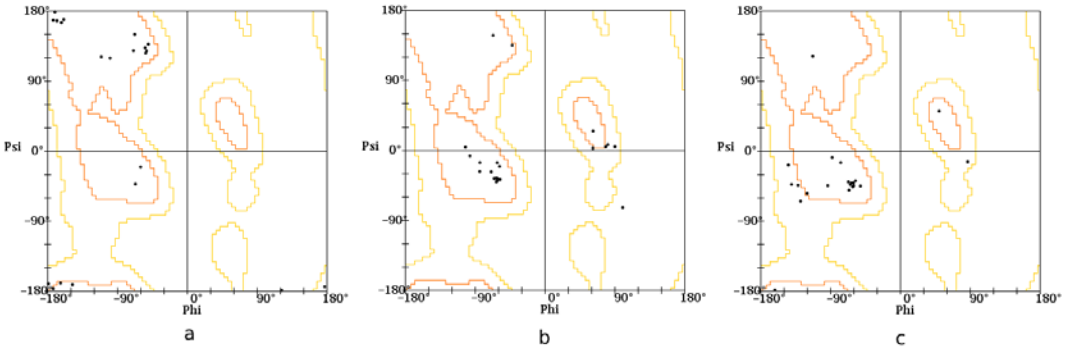

Figure 53. (a) The Ramachandran plot for the X1 residue. (b) The Ramachandran plot for the X2 residue. (c) The Ramachandran plot for the X3 residue.

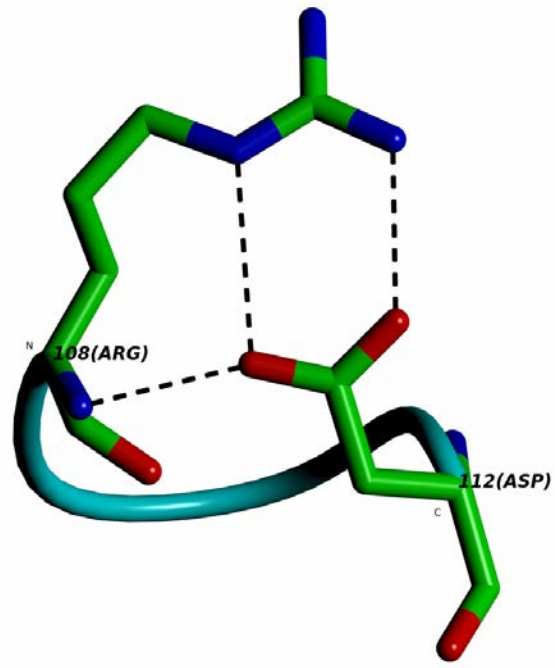

Figure 54. (a) The RPNAD motif in 1JX6 showing the hydrogen-bonding pattern. This H-bond interaction pattern was found to be comparable to that found in D-X-R with the same Arg rotamer.

5.8.4 Analysis of R-(3X)-D motif with two H-bonds.

Motifs in irregular structures with two H-bonds were studied in detail. 135 motifs were identified belonging to this group. Although a wide variety of side chain conformations were identified for the Arg residue, two conformations were observed to be major one. 18 instances of the Arg side chain conformation were being the completely extended t t t t (Figure 55). The Asp χ_1 conformation was found to be t in most cases and g- in few. The backbone conformation was observed to show no conservation.

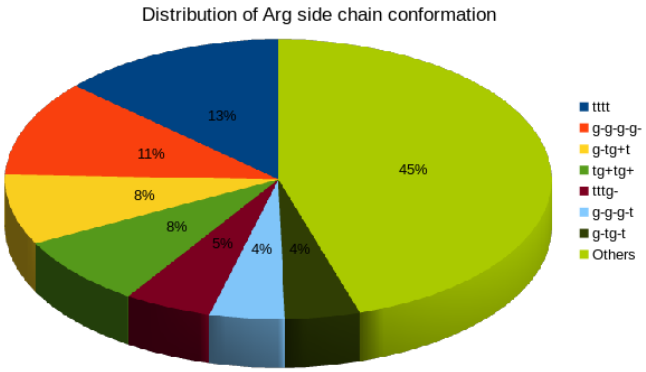

Figure 55. Distribution of Arg side-chain conformations in R-(3X)-D motifs in irregular structures with two H-bonds.

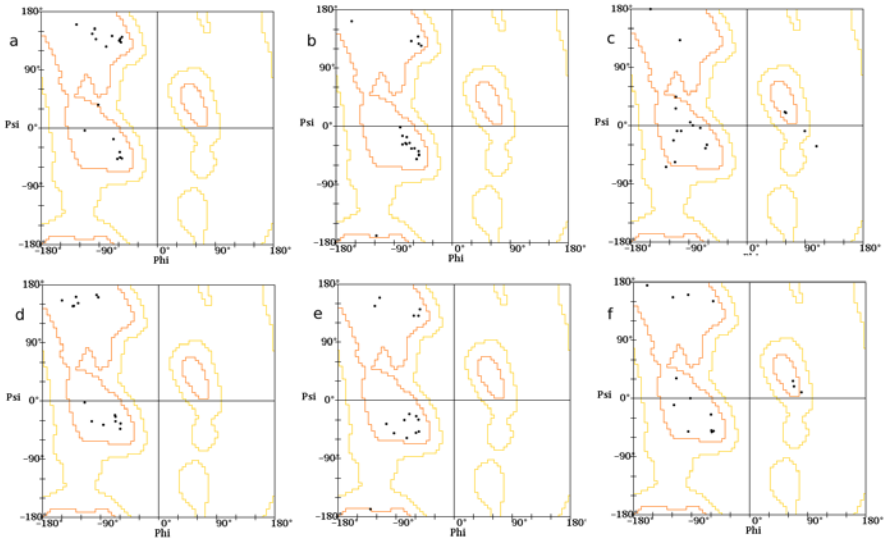

Figure 56. (a) The Ramachandran plot for the X1 residue with t t t t Arg side chain conformation. (b) X2 residue. (c) X3 residue. (d) The Ramachandran plot for the X1 residue with g- g- g- g- side chain conformation. (e) X2 residue. (f) X3 residue.

Even the Ramachandran plots for all the X residues showed spread over all regions (Figure 56a-c). In 6 cases the hydrogen bonding was found to be Type D (Figure 57a), Type B in 3 and in three cases one main chain – main chain bond was found in place of the side chain - side chain bond.

In 11% (15) of the cases the Arg side chain was completely folded g- g- g- g- (Figure 55). Here again the Asp χ_1 conformation showed both g- and g+ conformation in most cases. While X1 and X2 residues were found to spread over α-helix and extended regions in the Ramachandran plot, the plot for X3 was spread over all allowed regions (Figure 56d-f). The hydrogen bonding was observed to be Type C in 6 cases (Figure 57b) and Type B in 5.

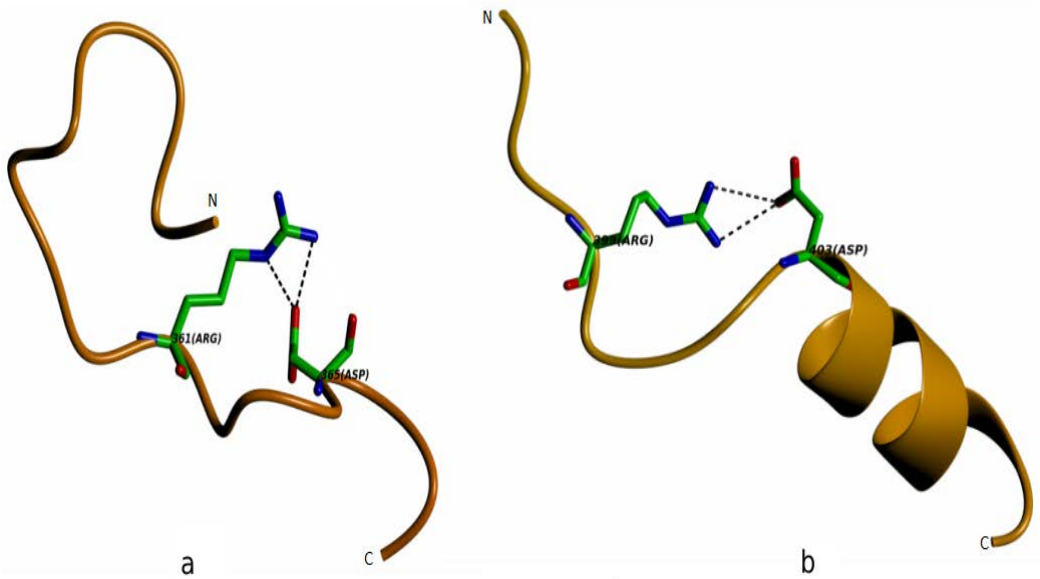

Figure 57. (a) The RPQFD motif in 2P3Z with t t t t side chain conformation showing Type D hydrogen-bonding pattern. (b) The RWTTD motif in 4B0T with g- g- g- g- side chain conformation showing Type C hydrogen-bonding pattern.

5.8.5 Analysis of E-(3X)-R motif with two H-bonds.

226 motifs were identified belonging to irregular structures with >1 hydrogen bonded interactions. These could be further classified as 191 occurrences with two H-bonds and 35 with three H-bonds.

The 191 motifs identified with 2 hydrogen bonded interactions were analyzed for the side chain conformation of the Arg residue. Wide range of conformations was observed for the motifs. The rotamer g- t t t constituted 18% (35) of the total (Figure 58).

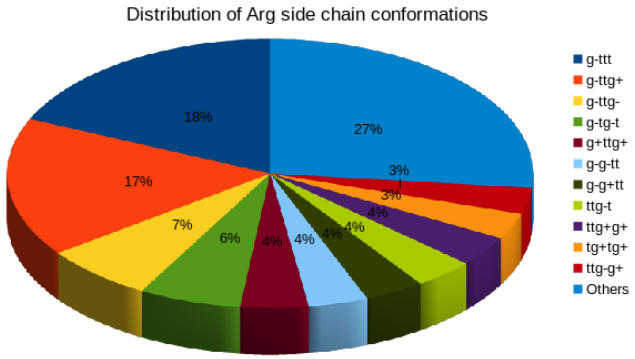

Figure 58. Distribution of Arg side-chain conformations in E-(3X)-R motifs in irregular structures with two H-bonds.

For these motifs the Glu χ_1 conformation was t while χ_2 conformation varied. Only in two cases the χ_1 conformation was g+. The backbone showed no conformational conservation. The Ramachandran plot for the X1 residue was found to show a cluster in the α-helix with three variations. The plot for X2 residue had spread over α-helix and extended regions with one in the left-handed helix region and one outlier. The X3 residue plot showed spread similar to X2 but had two variations in the left-handed helix region (Figure 59a-c). The hydrogen bonding was observed to be Type B in most cases (Figure 60).

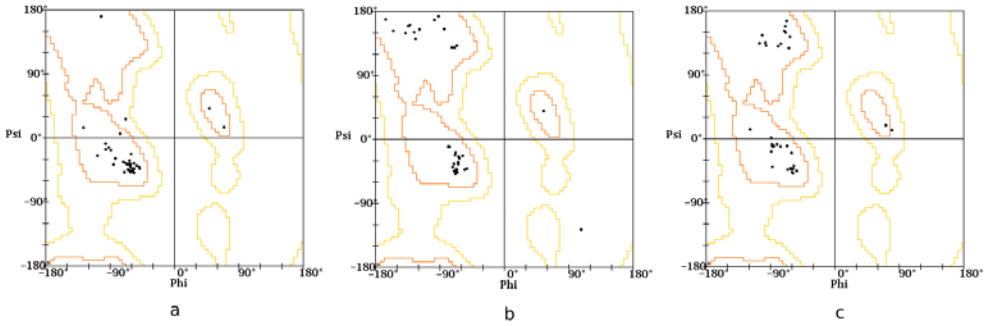

Figure 59. (a) The Ramachandran plot for the X1 residue for g- t t t Arg side chain conformation. (b) The Ramachandran plot for the X2 residue with g- t t t Arg rotamer. (c) The Ramachandran plot for the X3 residue with Arg rotamer g- t t t.

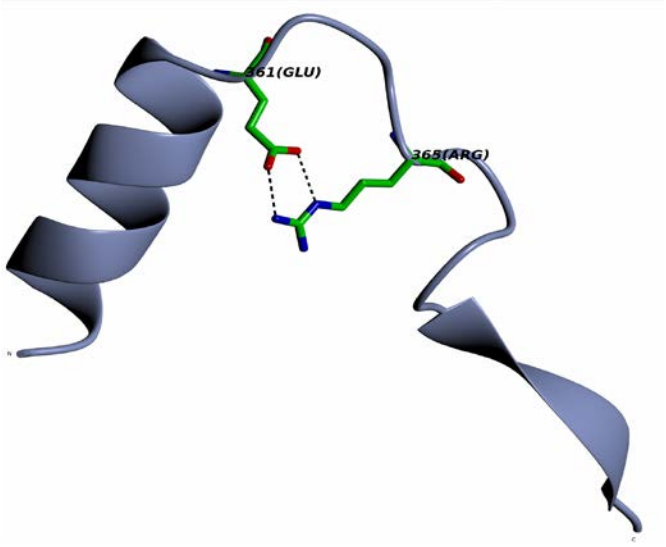

Figure 60. The EPFTR motif in 3BZM with g- t t t side chain conformation showing Type B hydrogen-bonding pattern.

The next conformation studied was g- t t g+. The Glu χ_1 conformation was g- while the χ_2 conformation was also found to be g-. The backbone conformation showed considerable conservation. The Ramachandran plot for all X residues showed single cluster in the α-helix region (Figure 61a-c). The hydrogen bonding for this group comprised a variant of Type C consisting of one side chain – side chain bond and one side chain – main chain bond.

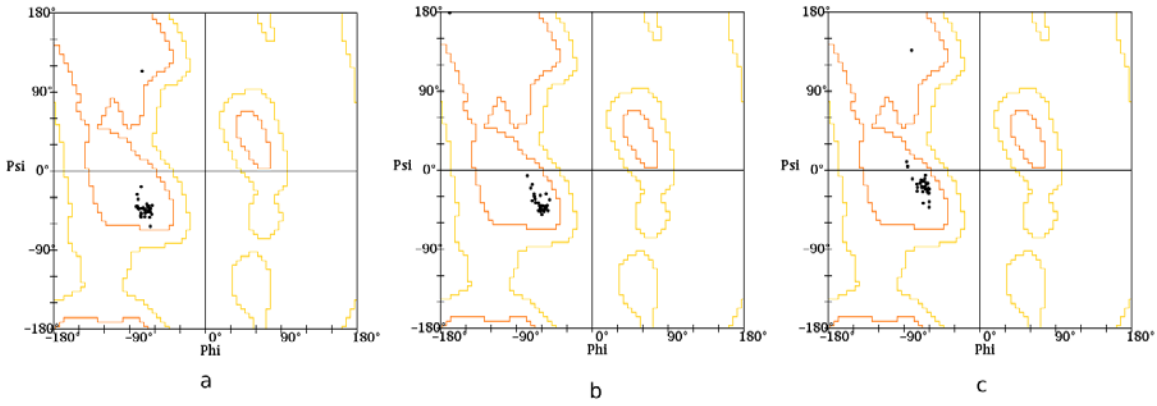

Figure 61. The Ramachandran plots with Arg g- t t g+ side chain conformation. (a) For X1 residue. (b) For X2 residue. (c) For X3 residue.

5.8.6 Analysis of R-(3X)-E motif with three H-bonds.

The total 250 occurrences identified with >1 H-bond comprised of 185 with two and 65 with three H-bonds.

For motifs with three interactions the most prominent Arg side chain conformation was found to be g- t g- t (19) where the Glu χ_1 conformation was g- and Glu χ_2 conformation was t. The Ramachandran plots for the X residues showed a cluster at the α-helix region (Figure 62). The hydrogen bonding observed here was Type B with a main chain – side chain bond (Figure 63). The hydrogen bonding pattern here was found to be similar to that in D-X-R with same Arg side chain conformation.

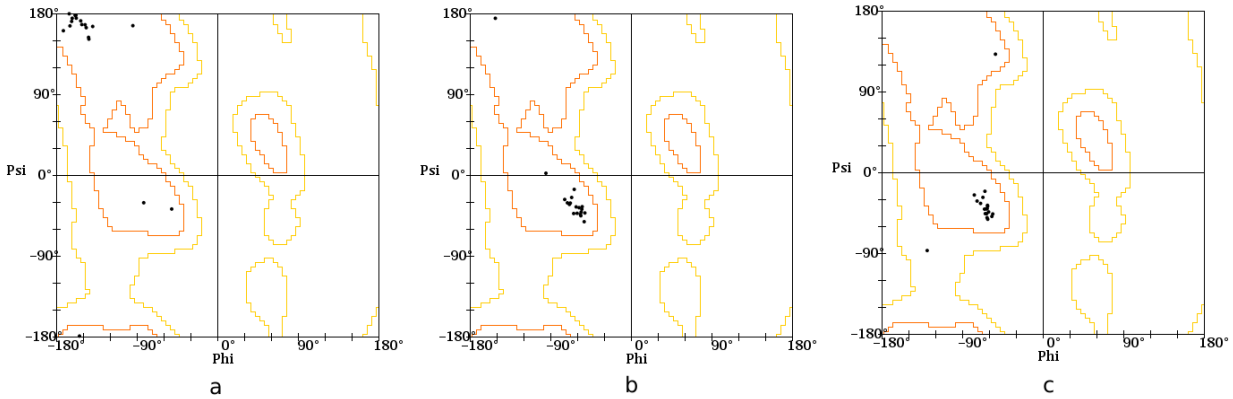

Figure 62. For R-(3X)-E motifs with Arg conformation g- t g- t the Ramachandran plots for (a) X1 residue (b) X2 residue and (c) X3 residue.

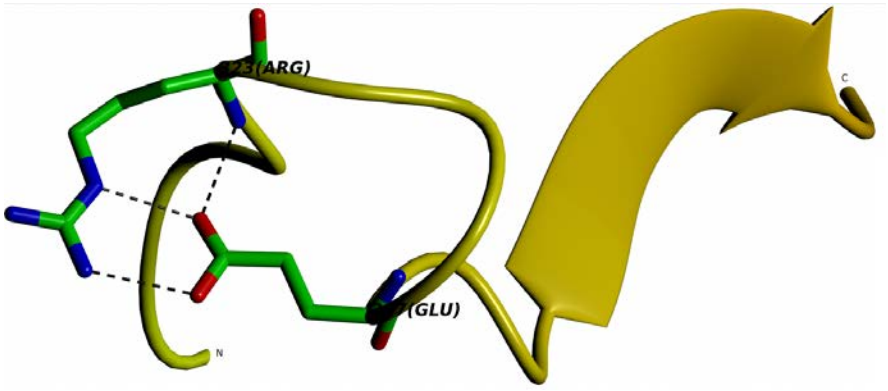

Figure 63. The RDDEE motif in 4C1Q with g- t g- t side chain conformation showing Type B hydrogen-bonding pattern. The interaction pattern was found to be similar to D-X-R with 3 H-bonds and same Arg rotamer.

5.8.7 Analysis of R-(3X)-E motif with two H-bonds.

Total of 185 occurrences of the motif with two H-bonds were identified. The Arg side chain analysis revealed three prominent conformations. The conformation g- t g+ t was observed in 23 (12%) motifs (Figure 64). The Glu χ_1 conformation was g- and Glu χ_2 conformation was t. The Ramachandran plot for the X residues showed two distinct groups (Figure 65a-c) and the hydrogen bonding was observed to be Type B (Figure 66a).

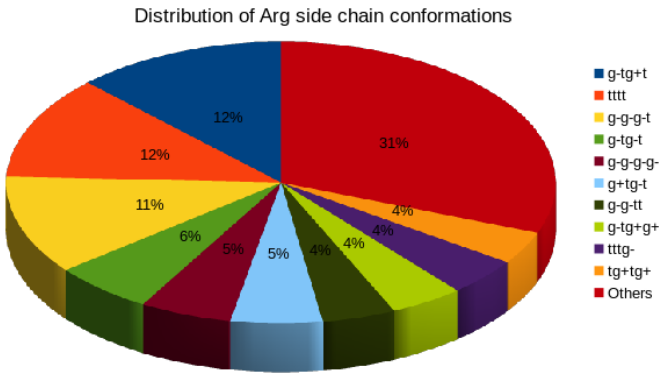

Figure 64. Distribution of Arg side-chain conformations in R-(3X)-E motifs in irregular structures with two H-bonds.

The next conformation group studied was t t t t in 22 occurrences (11%) (Figure 64). For this group the Glu χ_1 conformation was g- and Glu χ_2 conformation was t. The motif backbone was found to show no conservation. The Ramachandran plot for the X1 residue showed a cluster in the α-helix region and a spread in the β-sheet region. The plots for the X2 and X3 residues showed a single cluster in the α-helix region (Figure 65d-f). The hydrogen bonding here was found to have one side chain – side chain bond and one main chain – main chain bond.

The last conformation analyzed in detail was g- g- g- t occurring in 21 motifs (11%) (Figure 64). Here again the Glu χ_1 conformation was g- and Glu χ_2 conformation was t. The Ramachandran plot for the X1 residue mainly clustered in the β-sheet region while the X2 and X3 residues lie in the α-helix region (Figure 65g-i). The hydrogen bonding in this case was Type B (Figure 66b). The motifs were found to lie near the N-terminal end of α-helices.

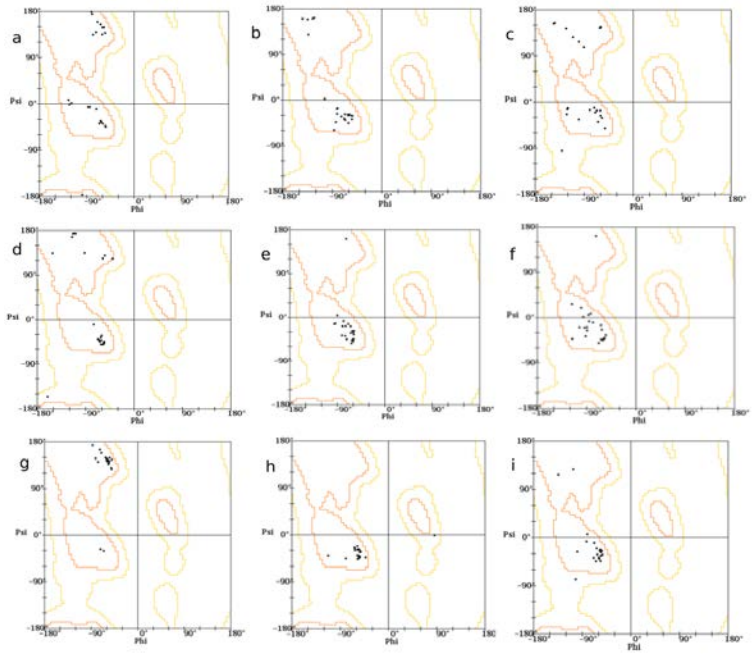

Figure 65. (a) The Ramchandran Plot for the X1 residue with g- t g- t side chain conformation. (b) X2 residue. (c) X3 residue. (d) The Ramchandran Plot for the X1 residue with t t t t side chain conformation. (e) X2 residue. (f) X3 residue. (g) The Ramchandran Plot for the X1 residue with g- g- g- t side chain conformation. (h) X2 residue. (i) X3 residue.

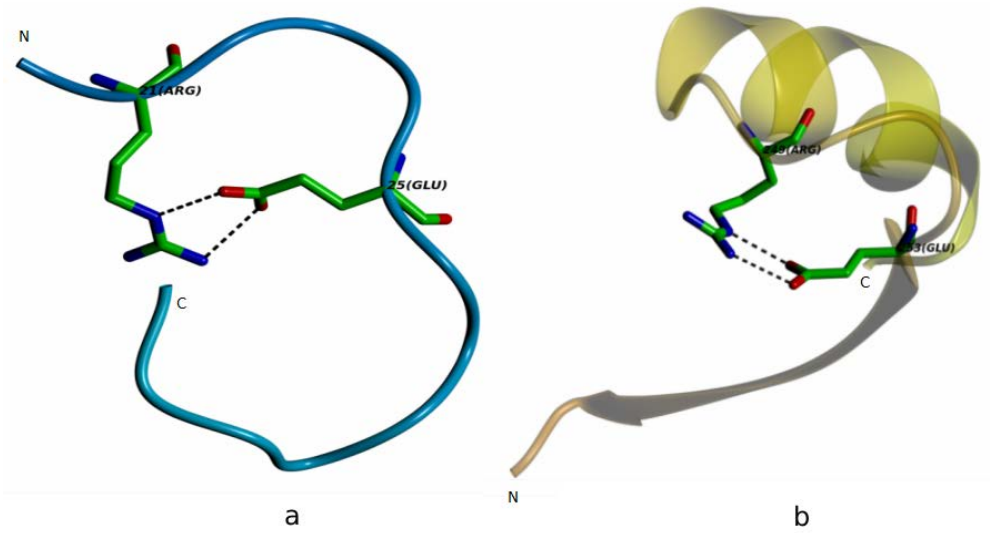

Figure 66. (a) The RPLAE motif in 3N0K with g- t g+ t side chain conformation showing Type B hydrogen-bonding pattern. (b) The RPSHE motif in 3GB5 with g- g- g- t side chain conformation showing Type B hydrogen-bonding pattern.

5.9. Comparative analysis of the motifs.

The four motifs identified and studied above revealed significant occurrences in helices. These motifs have been compared in this section with regard to their Arg and Asp/Glu side chain conformations and the hydrogen bonding patterns (Table 6). In all motifs, the, i+4 $\rightarrow$ i hydrogen bond, a characteristic of the helical hydrogen bonding remain common and hence was not separately considered during the analysis.

5.9.1 Comparing D-(3X)-R and R-(3X)-D motifs in helix.

The two motifs were compared to study the effect of the reversal of the Asp and Arg residue positions (Table 10). In both motifs, occurrences with two and three interactions were recorded. For the D-(3X)-R, the hydrogen bonding was found to be of Type D, which on reversal of the positions of R and D in R-(3X)-D motif, changed to Type B. In the first case the Arg side chain was found to be more folded with the Asp side chain being extended, for the R-(3X)-D motif the Arg side chain assumed a partially folded state with the Asp side chain changing to a more folded g- conformation in order to interact with each other.

Table 10. Comparison of the D-(3X)-R, R-(3X)-D, E-(3X)-R, R-(3X)-E motifs.

Motif	No. of Interactions	Type	Major Arg conformation (χ_1 - χ_4)	Major Asp/Glu (χ_1) conformation	Figure No.
D-(3X)-R	2	Type D	g- t g- g+	t	67a
R-(3X)-D	2	Type B	t g+ t g-	g-	67b
E-(3X)-R	2	Type D	g- t t g-	t	67c
R-(3X)-E	2	Type D	t t t t	g-	67d

5.9.2 Comparing D-(3X)-R and E-(3X)-R motifs in helix.

The D-(3X)-R and E-(3X)-R motifs were compared to assess the effect due to the change of Asp to Glu in the first flanking position. The E-(3X)-R was found to exceed D-(3X)-R both in terms of total numbers as well as those with interactions in helix group (Table 10). The change in residue at the first position, from Asp to Glu was found to have no effect since motifs with both amino acids and three interactions were identified. In both cases the hydrogen bonding was found to be Type D. Only in terms of the Arg side chain the more folded conformation of g- t g- g+ was found to change to a partially folded conformation g- t t g-. However, in the Asp and Glu side chain conformation were found to be exactly the same.

5.9.3 Comparing E-(3X)-R and R-(3X)-E motifs in helix.

Here the motifs E-(3X)-R and R-(3X)-E were compared to study the interchange of the Glu and Arg positions (Table 10). In both cases the motifs involving two and three H-bonds were recorded. The hydrogen bonding also was found to be the same i.e Type D. However, in case of the E-(3X)-R motifs the Arg side chain conformation was partially folded g- t t g- with the Glu χ_1 conformation being t, on reversal of the position, in the R-(3X)-E motifs the Arg residue assumed a completely extended conformation t t t t while the Glu χ_1 conformation was found to be folded with a g- conformation and participates in interaction.

5.9.4 Comparing R-(3X)-D and R-(3X)-E motifs in helix.

The motifs R-(3X)-D and R-(3X)-E were compared to understand changes when Asp was replaced by Glu at the last flanking position (Table 10). Significant increase in the number of motifs was found in case of Glu as compared to Asp at the last position, both in the total as well those with interactions (Table 10). Motifs with two H-bonds were identified in each case. However, a change in the hydrogen-bonding pattern was recorded i.e. type B to type D for the change from Asp to Glu. The Arg side chain conformation was also found to change from partially folded t g+ t g- in R-(3X)-D motifs to a completely extended t t t t in R-(3X)-E motifs. Both the Asp and the Glu χ_1 conformations were found to remain the same.

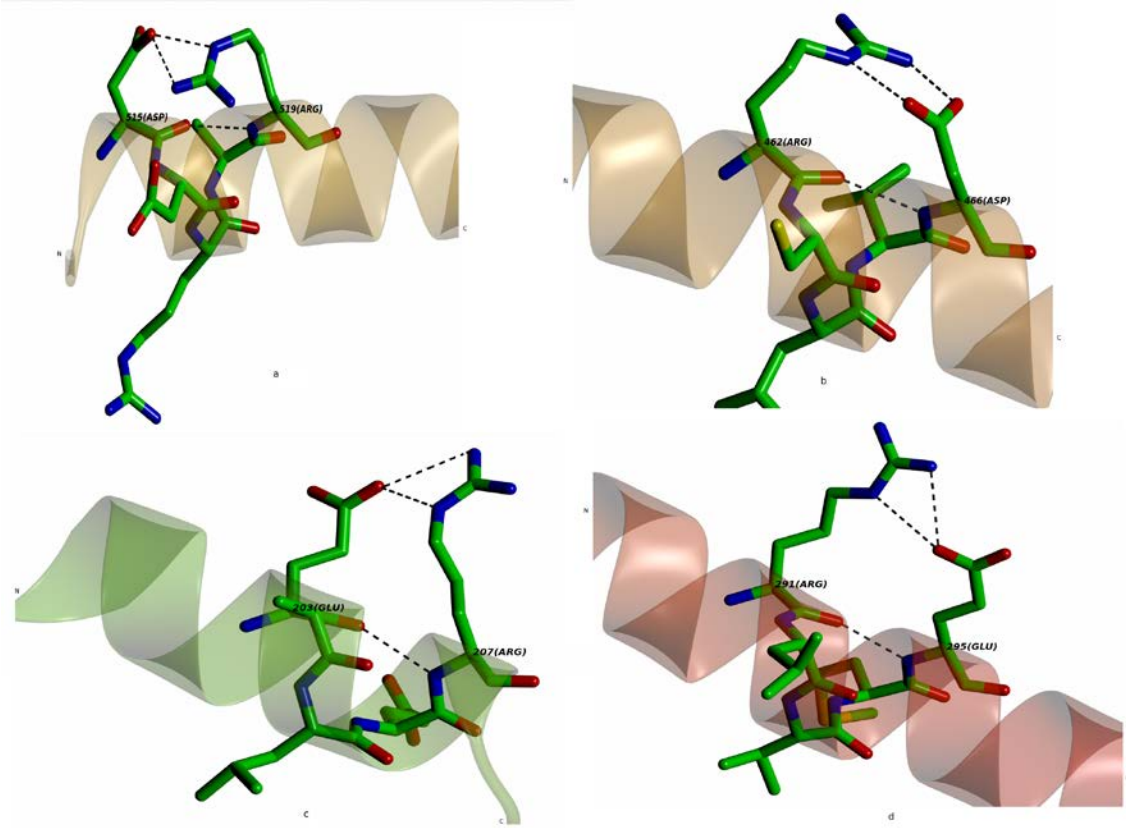

Figure 67. (a) The DAFVR motif in 4MIF showing the Type D hydrogen-bonding pattern. (b) The RCLVD motif in 2YU1 showing the Type B hydrogen-bonding pattern. (c) The EALER motif in 2ELC showing the Type D hydrogen-bonding pattern. (d) The RLVME motif in 1WOH showing the Type D hydrogen-bonding pattern.

5.10. Summary

Extending the studies carried out in the previous chapters, focus initially was on the structural analysis of 4- residue motifs with Asp/Glu and Arg as the flanking residues. The statistical analysis of the motifs in each of the three structural groups showed considerably few motifs with interactions in the helix and sheets classes with the exception of D/E-(2,3)X-R as well as R-(2,3)X-D/E wherein surprisingly increased numbers of motifs with interactions were observed. In the irregular structures decreasing trend of interacting motifs from 2X to 8X was identified for D-(2,8)X-R, E-(2,8)X-R and R-(2,8)X-D.

Extrapolating from the above studies four patterns namely D-(3X)-R, R-(3X)-D, E-(3X)-R and R-(3X)-E occurring the helix group were selected for detailed analysis as they exhibited very high numbers of motifs with interactions. A common interaction, the i+4 $\rightarrow$ i hydrogen bond, observed in these motifs as part of the intra-helical hydrogen bonding was not considered separately. In D-(3X)-R, the motifs with both two and three hydrogen-bonded interactions were observed. The side chain of the Arg residue at the last position was found to assume a more folded conformation with the Asp side chain assuming an extended conformation. The hydrogen-bonding pattern for motifs with three H-bonds was found to be of Type D while in motifs with two H-bonds, a modification of the same was recorded whereby one of the side chain – side chain H-bonds was found to be disrupted. On analyzing the R-(3X)-D, it was observed that the Arg side chain here had a partially folded conformation which was augmented by a similarly folded g- conformation of the Asp side chain thereby allowing the formation of the side chain - side chain hydrogen bonding pattern, Type B. For motifs with only two interactions it was realized that they too belonged to the same Type B, however in them one of the side chain – side chain interaction was lost. For the E-(3X)-R, the Arg side chain was partially folded while the Glu side chain had an extended conformation. Motifs with three H-bonds showed Type D hydrogen bonds, while those with two had one of the side chain H-bond missing. Finally in case of R-(3X)-E, the Arg side chain was found to have completely extended conformation, while the Glu side chain was folded. The hydrogen bonding in this case was the same as that found in E-(3X)-R.

Comparative study of these four motifs in helices revealed that reversal of the motifs' terminal residues resulted in change in the Arg and Asp/Glu side chain conformation while the hydrogen bonding was found to shift from Type D to Type B for D-(3X)-R and R-(3X)-D but remained exactly the same for E-(3X)-R and R-(3X)-E, i.e. Type D. The change from Asp to Glu in motifs D-(3X)-R and E-(3X)-R resulted in only a change in the Arg side chain conformation with Asp/Glu side chain conformation and hydrogen-bonding pattern remaining the same. In case of reverse motifs R-(3X)-D and R-(3X)-E, the Arg side chain changed from a partially folded conformation to a fully extended conformation while the Asp and Glu side chain conformation remained the same. The hydrogen bonding also was found to change from Type B to Type D.

Chapter 6

Analysis of the conformational preferences of homo-polymeric amino acid repeats in known protein structures

Single amino acid repeats, also known as homopolymeric amino acid (HPAAs) tracts, are usually known to exist in intrinsically unstructured regions (IURs) of proteins (Luo and Nijveen, 2014; Simon and Hancock, 2009). Although, such repeats need not fold into a specific 3D structure, it has been shown that homopolymers of certain amino acids can form specific secondary structures. Repeat structures in proteins are known to have functions related to molecular recognition and molecular assembly (Albà Soler and Guigó Serra, 2004; Faux, et al., 2005). Leu and Ala repeats being hydrophobic in nature, are commonly found in structured regions of proteins. Similarly, Glu repeats are often found within structured regions, although Glu is considered an unfavourable amino acid for forming ordered structures. Structural studies have revealed that while short poly-Ala peptides form alpha-helices, longer poly-Ala tracts are predicted to be predominantly beta-strands (Giri, et al., 2003). There has been a growing interest in single amino acid repeats in proteins ever since these are shown to be the cause of a variety of diseases. A number of experimental studies also suggest that they play an important role in protein function (Emili, et al., 1994; Kazemi-Esfarjani, et al., 1995; Lanz, et al., 1995; Mitchell and Tjian, 1989; Pinto and Lobe, 1996; Schwechheimer, et al., 1998).

It has long been observed that homopolymeric amino acid tracts are a very common feature of eukaryotic proteins (Green and Wang, 1994) and are present in nearly one-fifth of human gene products. They may be encoded by the repeat of a single codon or a mixture of synonymous codons. There is increasing biochemical evidence to show that repeats of amino acids such as Gln, Ala, Pro and Gly can modulate protein-protein interactions and regulate transcription (Gerber, et al., 1994; Perutz, 1994). HPAA tracts such as poly-Ala, poly-Gln, poly-Pro and poly-Ser are relatively abundant, especially among transcription factors (Faux, et al., 2005). The case of human Transcription factor II D (TFIID) is most striking as it contains a 34 residue Gln repeat which is absent in related proteins from other species (Subirana and Palau, 1999). Genome-wide studies of *Saccharomyces cerevisiae* (Young, et al., 2000) and mammals (Oma, et al., 2004) have shown a regular of association of HPAAs with transcription.

Forkhead box protein P2 (FOXP2) contains a 40-residue poly-Gln tract while the Myelin transcription factor 1 has a 32-residue poly-Glu tract. The Brain-2 (POU domain, class 3) transcription factor contains multiple tracts such as a 5-residue poly-Ala, 21-residue poly-Gly, 7-residue poly-Pro, and a 21-residue poly-Gln tract. Poly-Pro or poly-Gln can activate transcription by fusing to DNA binding domain of GAL4 factor (Gerber, et al., 1994). Lanz et al. have shown that replacement of glutamine tract with an artificial alanine tract by an out of frame mutation in the rat gluco-corticoid receptor (GR) results in termination of transcription. Interestingly, transcriptional activation has been found to be closely linked to the number of repeats of homopolymeric tract (both proline and glutamine tracts) (Lanz, et al., 1995). It has been suggested that some HPAAs have a positive role in evolution (Kashi and King, 2006) and are also involved in species-specific regulatory factor interactions. In addition, it has also been suggested that in protein kinases the repeats could play a significant role in the evolution of cellular signaling networks (Cox, et al., 1996). Signal peptides, situated at N-terminus of polypeptide chain of secreted and membrane proteins, have regions rich in amino acids with hydrophobic side chains essential for the interaction with the signal recognition particle and later on the translocase complex of the endoplasmic reticulum membrane region. Upon transit through the membrane, the signal peptide is cleaved off from the emerging polypeptide chain by signal peptidases and then rapidly degraded by proteases in most cases (Voss, et al., 2013). Based on an analysis of 20 different HPAA type repeats, it has been concluded that different HPAAs are able to affect intracellular localization in many cases (Oma, et al., 2004). Hydrophobic HPAA tracts exhibit a strong tendency for aggregation and hence produce severe cytotoxic effects when expressed in mammalian cells (Dorsman, et al., 2002; Oma, et al., 2004). HPAA tracts such as poly-Arg containing peptides have been used in protein engineering as a drug-delivery system owing to their ability to enter the cells (Tung and Weissleder, 2003). The metal-binding activity of six-residue poly-His tags has been widely exploited for purifying recombinant proteins. Poly-Lys provides a hydrophilic coating enabling cell adhesion and proliferation. A poly-Gly tract in plant protein Toc-75 has been shown to be vital for directing the protein to the chloroplast outer envelope (Inoue and Keegstra, 2003) while several viral poly-Arg rich proteins have been observed to participate in RNA binding (Calnan, et al., 1991; Nam, et al., 2001).

While HPAA tracts are useful as peptides that can have multiple applications, their functional and medical roles have specially received much attention. Uncontrolled genetic expansions of HPAA tracts lead to the development of serious debilitating human diseases. Expanded poly-Gln and poly-Ala tracts have been associated with neurological disorders like Huntington disease or Oculopharyngeal Muscular Dystrophy (OPMD), respectively. Currently, nine poly-Gln-related diseases (Table 1) and nine poly-Ala related diseases (Table 2) have been identified (Riley and Orr, 2006) Out of all the polyamino acid repeats characterized till date, poly-Gln repeats are the most extensively studied ones. Trinucleotide repeat expansions that result in expanded poly-Gln tracts are the basis of several human neurogenetic diseases. In the nine poly-Gln linked diseases the proteins believed to be responsible for the disease contain expanded poly-Gln tracts that have been shown to aggregate and form fibrils both *in vitro* and *in vivo* (Faux, et al., 2005). The pathogenic length of the poly-Gln tracts are specific for each protein family (Cummings and Zoghbi, 2000). For example, Huntington's disease develops only when the poly-Gln repeat in the Huntington protein is 38 amino acids (generally encoded by 36 CAG repeats and one each CAA and CAG) whereas Machado-Joseph disease develops when the poly-Gln repeat in Ataxin-3 is 45 amino acids in length (Chow, et al., 2004; Cummings and Zoghbi, 2000). The accumulation of the aggregated protein correlates with cell death and the onset of degenerative disease.

Expansion of poly-alanine tracts, mainly in transcription factor genes, have been shown to cause mental retardation and deformities in brain, digits as well as other structures (Brown and Brown, 2004). Removing a poly-Ala tract from murine *HOXD13* has been shown to have a direct effect on bone phenotype, indicating the involvement of a HPAA in an important biological process (Albrecht, et al., 2004). Like poly-Gln repeats, many of the disease-linked poly-Ala tracts are transcription factors (Brown and Brown, 2004) with the proteins containing lengthened poly-Ala tracts (>10) result in an enhanced tendency to aggregate and form fibrils (Fan, et al., 2001).

Like poly-Gln repeats, many of the disease-linked poly-Ala tracts are transcription factors (Brown and Brown, 2004) with the proteins containing lengthened poly-Ala tracts (>10) result in an enhanced tendency to aggregate and form fibrils (Fan, et al., 2001).

Table 1. Known diseases caused by poly-glutamine expansion in humans.

Disease	Gene	Locus	Protein	CAG Repeat	
				Normal	Diseased
Spinobulbar muscular atrophy (Kennedy disease)	*AR*	Xq13–21	Androgen receptor (AR)	9–36	38–62
Huntington's disease	*HD*	4p16.3	Huntingtin	6–35	36–121
Dentatorubral-pallidoluysian atrophy (Haw–River syndrome)	*DRPLA*	12p13.31	Atrophin-1	6–35	49–88
Spinocerebellar ataxia type 1	*SCA1*	6p23	Ataxin-1	6–44	39–82
Spinocerebellar ataxia type 2	*SCA2*	12q24.1	Ataxin-2	15–31	36–63
Spinocerebellar ataxia type 3 (Machado–Joseph disease)	*SCA3* (*MJD1*)	14q32.1	Ataxin-3	12–40	55–84
Spinocerebellar ataxia type 6	*SCA6*	19p13	α_{1A}-voltage-dependent calcium channel subunit	4–18	21–33
Spinocerebellar ataxia type 7	*SCA7*	13p12–13	Ataxin-7	4–35	37–306
Spinocerebellar ataxia type 17	*TBP*	6q27	TATA-binding protein	25-42	47-63

Several studies have demonstrated that many nondisease-linked polyamino acid tracts are also toxic to cells, leading to protein aggregation or misfolding (Dorsman et al. 2002; Fandrich and Dobson 2002). While extensive studies have been carried out on HPAA tracts the analyses have been based on the DNA and protein sequences. However, the extension of these studies on HPAAs to 3D protein structures has been relatively meagre.

Table 2. Known diseases caused due to poly-alanine expansion in humans.

Condition	Gene	Gene type	Locus	Expansion size	Protein dysfunction
Synpolydactyly type II	*HOXD13*	Transcription factor	2q31-32	15A→22–29A	Dominant negative
Cleidocranial dysplasia	*RUNX2* (*CBFA1*)	Transcription factor	6p21	17A→27A	Loss-of-function
Oculopharyngeal muscular dystrophy	PABPN1	Polyadenylate-binding protein	14q11.2-13	10A→11–17A	Toxic protein aggregates
Holoprosencephaly (HPE5)	ZIC2	Transcription factor	13q32	15A→25A	Loss-of-function
Hand-foot-genital syndrome	HOXA13	Transcription factor	7p15-14.2	18A→24A or 26A	Unclear, might be dominant negative
Blepharophimosis, ptosis and epicanthus inversus	FOXL2	Transcription factor	3q23	14A→22–24A	Partial loss-of-function
Mental retardation; X-linked, with isolated growth hormone deficiency	SOX3	Transcription factor	Xq26.3	15A→26A	Unknown

Infantile spasm syndrome, X-linked; Partington syndrome; lissencephaly with ambiguous genitalia, X-linked; mental retardation X-linked 36 and 54	ARX	Transcription factor	Xp22.13	A-tract#1 (amino acids 100–115) 16A→18 or 23A; A-tract#2 (amino acids 144–155) 12A→20 A	Partial loss-of-function
Congenital central hypoventilation syndrome/Ondine curse	PMX2B (PHOX2B)	Transcription factor	4pl2	20A→25 –29A	Loss-of-function

With the availability of the vast expanse of 3D structures in the PDB database (Berman, et al., 2000), it is important to analyze the HPAAs in protein structures to understand their structural and functional roles. Initial part of this chapter focuses on the identification of 3-8 residue HPAAs of all 20 amino acids along with their classification into the defined secondary structure groups based on the local PDB dataset generated for the analysis. In the next step the generated data are analysed to calculate probability score that predicts the secondary structure preference of selected HPAA.

6.1 Secondary Structure Analysis of HPAAs for all 20 amino acids.

After identifying the HPAAs of all 20 naturally occurring amino acids in the local PDB dataset, analysis was carried out first on triple amino acid repeats and then extended progressively to higher length repeats that could be identified in the local dataset. Table 3 gives a report of the varied lengths of HPAAs identified. The HPAA tracts are classified based on their secondary structures into helix, sheets and irregular

structural regions as already defined. Those in the irregular structural regions are further classified into those belonging to well-defined hydrogen-bonded turns or to other irregular structures.

Table 3. Various HPAAs identified in the local PDB dataset.

AA	Tri-AA repeat	Tetra-AA repeat	Penta-AA repeat	Hexa-AA repeat	Hepta-AA repeat	Octa-AA repeat
Alanine	1230	130	15	4	3	-
Glutamine	168	7	-	-	-	-
Tyrosine	-	7	-	-	-	-
Valine	678	44	1	-	-	1
Isoleucine	231	12	1	-	-	-
Leucine	1020	76	3			-
Phenylananine	80	3	-	-	-	
Glutamic acid	614	26	4	-	-	-
Lysine	332	22	-	-	-	-
Arginine	173	21	1	-	-	-
Methionine	29	-	-	-	-	-
Tryptophan	3	-	-	-	-	-
Threonine	313	26	1	-	-	-
Histidine	35	3	-	-	-	-
Glycine	636	74	11	1		-
Serine	443	27	4	1	-	1
Aspartic acid	272	16	1	-	-	-
Proline	182	24	8	-	-	-
Asparagine	173	9	1	-	-	-
Cysteine	5	-	-	-	-	-

6.1.1 Alanine.

Ala HPAAs were dominantly observed in helices. While 852 occurrences were observed in helices, 235 were found in irregular structures with 75 (Figure 1) in turns for tri-amino acid (tri-AA) repeats (Table 4). The trend was found to continue for tetra-amino acid (tetra-AA) and penta-amino acid (penta-AA) repeats albeit no

occurrences were recorded for turns. Higher HPAAs such as hexa-amino acid (hexa-AA) and hepta-amino acid (hepta-AA) repeats were observed to occur exclusively in helices.

Table 4. Various HPAAs identified for alanine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular structural regions		Total
				Turns	Other Irregular structures	
Ala	3A	852	68	75	235	1230
	4A	93	7	0	30	130
	5A	14	0	0	1	15
	6A	4	0	0	0	4
	7A	3	0	0	0	3

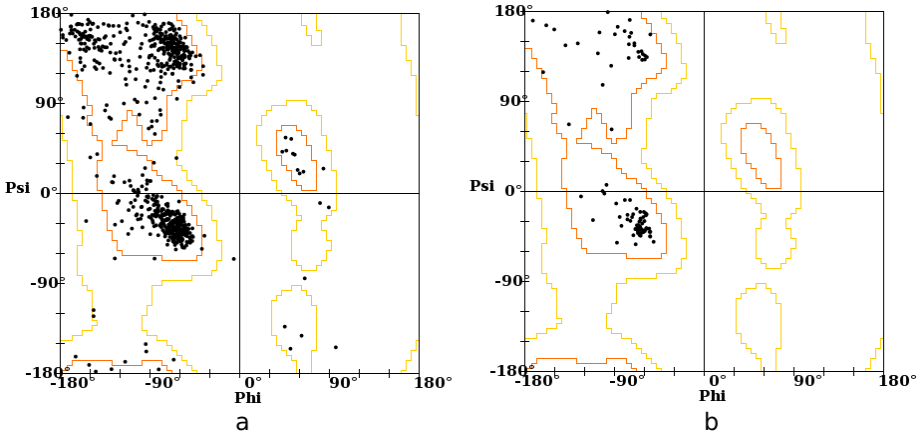

Figure 1. Ramachandran plots for alanine (a) tri- and (b) tetra-amino acid repeat HPAAs occurring in irregular structures.

6.1.2 Glutamine.

Only tri-AA and tetra-AA were found for glutamine HPAAs. The tri-AA HPAAs show high presence in helices. Only 45 tri-AA occurrences (Table 5) were found in irregular structures (Figure 2).

Table 5. HPAAs identified for glutamine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Gln	3Q	118	1	4	45	168
	4Q	6	1	0	0	7

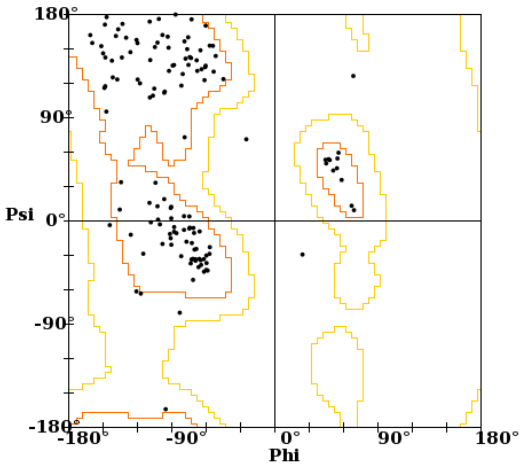

Figure 2. Ramachandran plot for glutamine tri amino acid repeat HPAAs occurring in irregular structures.

6.1.3 Tyrosine.

Only tetra-AA HPAAs were observed for Tyr. Of the seven occurrences, four were found to occur in helices whereas two existed in irregular structures and only one was in sheets (Table 6).

Table 6. HPAAs identified for Tyrosine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Tyr	4Y	4	1	0	2	7

6.1.4 Valine.

Tri-AA to penta-AA and octa-AA repeats of valine HPAAs were identified in the database. Tri-AA repeats have preference for beta-sheets. Only two occurrences of tri-AA are observed in turns. Higher HPAAs of valine such as penta-AA and octa-AA were found to occur only in irregular regions (Figure 3). Table 7 gives the distribution of the observed valine HPAAs in the secondary structure groups. The Ramachandran plot of valine HPAAs in irregular regions shows distribution in α-helix and β-sheet regions.

Table 7. HPAAs identified for valine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Val	3V	70	487	2	119	678
	4V	0	38	0	6	44
	5V	0	0	0	1	1
	8V	0	0	0	1	1

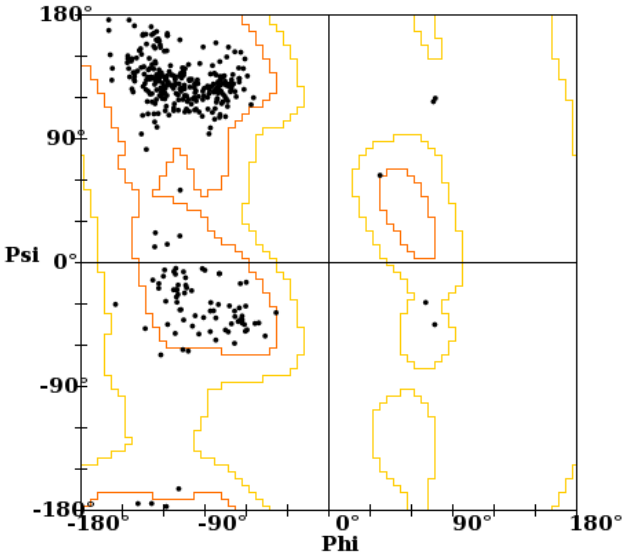

Figure 3. Ramachandran plot for valine tri amino acids HPAAs occurring in irregular structures.

6.1.5 Isoleucine.

The isoleucine HPAAs prefer sheet structures. Majority of HPAAs were observed as tri-AA only 12 were found as tetra-AA, all in sheets. One penta-AA identified was also in sheet. Only 37 Ile tri-AA were found in helix group and 36 in irregular regions (Figure 4). Table 8 given below classifies observed HPAAs in the secondary structure groups. Clearly Ile prefers sheet structure.

Table 8. HPAAs identified for isoleucine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Ile	3I	37	156	2	36	231
	4I	0	12	0	0	12
	5I	0	1	0	0	1

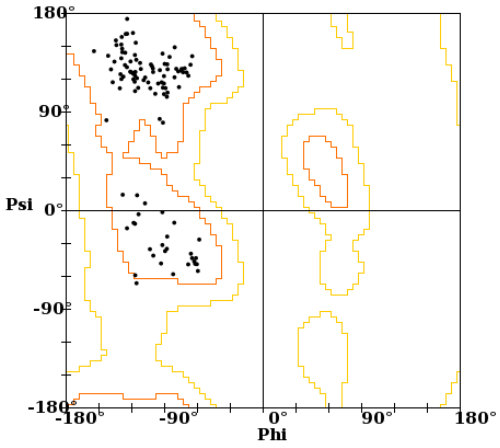

Figure 4. Ramachandran plot for isoleucine tri amino acids HPAAs occurring in irregular structures.

6.1.6 Leucine.

The amino acid leucine shows presence in all secondary structure groups (Figure 5). While maximum occurrence was noted in helices for tri-AA, its presence in other groups was also found to be notable (Table 9). In case of penta-AA, the presence of leucine HPAAs was found only in helices. The equitable preference of Leu in all secondary structures has already been noted in the amino acid propensity parameters developed by Chou-Fasman (Chou and Fasman, 1974).

Table 9. HPAAs identified for leucine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Leu	3L	558	284	24	154	1020
	4L	33	32	0	11	76
	5L	3	0	0	0	3

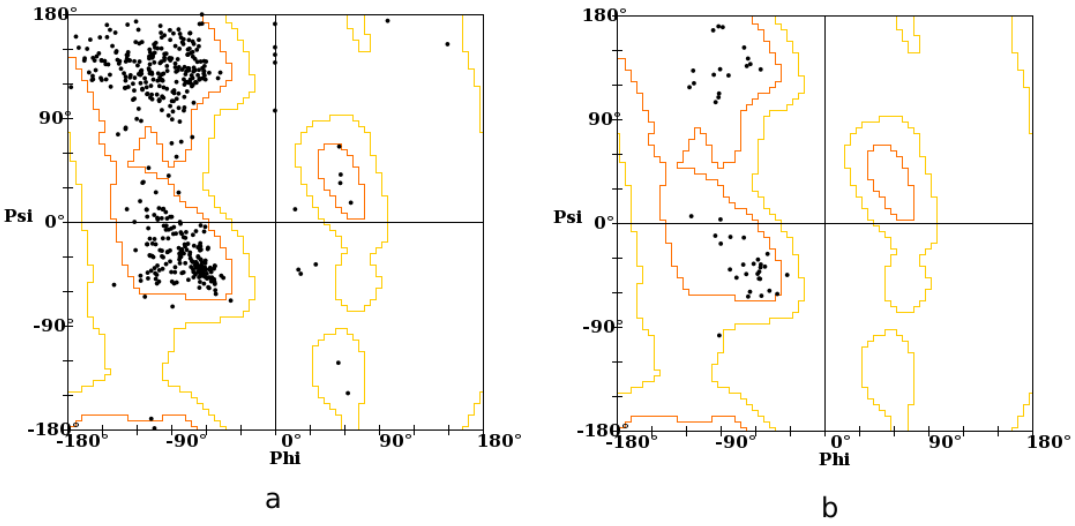

Figure 5. Ramachandran plots for leucine (a) tri- and (b) tetra- amino acids HPAAs occurring in irregular structures.

6.1.7 Phenylalanine.

Phe HPAAs were observed in tri-AA as well as tetra-AA repeat. In case of tri-AA, a nearly equitable distribution of HPAAs was observed in helices and sheets (Table 10).

Table 10. Various HPAAs of phenylalanine identified in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Phe	3F	27	36	3	14	80
	4F	2	1	0	0	3

6.1.8 Glutamic acid.

Tri-AA Glu HPAAs show highest occurrence in helices with 143 in irregular structures and 47 in turns (Figure 6a). However, for tetra-AA, 15 occurrences of HPAA were observed in irregular structures (Figure 6b) while only 9 were recorded in the helix group whereas for penta-AA two occurrences each were found in helix and irregular structures (Table 11).

Table 11. HPAAs identified for glutamic acid in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Glu	3E	405	19	47	143	614
	4E	9	2	0	15	26
	5E	2	0	0	2	4

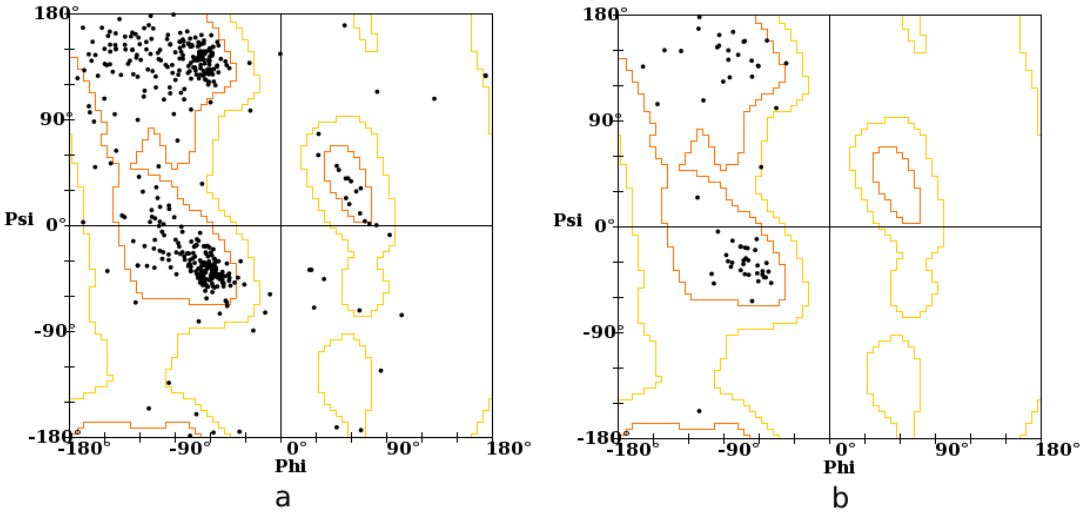

Figure 6. Ramachandran plot for glutamic acid (a) tri- and (b) tetra- amino acids HPAAs occurring in irregular structures.

6.1.9 Lysine.

Occurrences of Lys HPAAs as tri-AA and tetra-AA were recorded. Equitable occurrence of Lys HPAAs in tri-AA was noted to exist in the helix and irregular structures (Figure 7). 16 occurrences of tetra-AA Lys HPAAs were found in the irregular structures (Table 12).

Table 12. HPAAs identified for lysine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Lys	3K	132	24	41	135	332
	4K	3	3	0	16	22

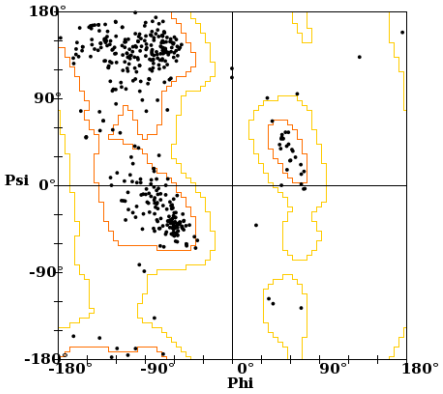

Figure 7. Ramachandran plot for lysine tri amino acids HPAAs occurring in irregular structures.

6.1.10 Arginine.

HPAAs for the amino acid arginine were found in helix and irregular structures (Figure 8). Only one penta-AA Arg HPAA was recorded in irregular structures (Table 13).

Table 13. HPAAs identified for arginine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Arg	3R	132	27	4	94	257
	4R	10	0	2	9	23
	5R	0	0	0	1	1

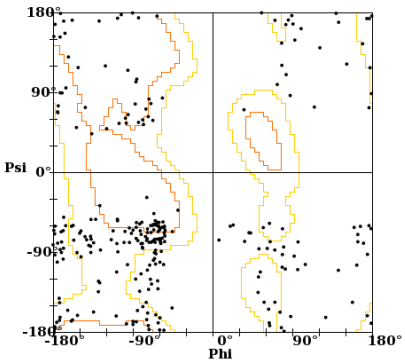

Figure 8. Ramachandran plot for arginine tri amino acids HPAAs occurring in irregular structures.

6.1.11 Methionine.

Only tri-AA HPAAs were observed for Met. The structural preferences were mainly for helices and irregular structures (Table 14).

Table 14. Various HPAAs for methionine identified in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Met	3M	14	4	0	11	29

6.1.12 Tryptophan.

Only three HPAAs for the amino acid Trp possessing large side chain were observed occurring as tri-AA (Table 15).

Table 15. HPAAs of tryptophan identified in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Trp	3W	2	0	0	1	3

6.1.13 Threonine.

Tri-AA, tetra-AA and penta-AA were found for Thr HPAAs. Most HPAAs were observed in the irregular structures (Figure 9a,b). In case of penta-AA only one occurrence was noted in sheets group (Table 16).

Table 16. HPAAs identified for threonine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Thr	3T	54	66	8	185	313
	4T	5	3	0	18	26
	5T	0	1	0	0	1

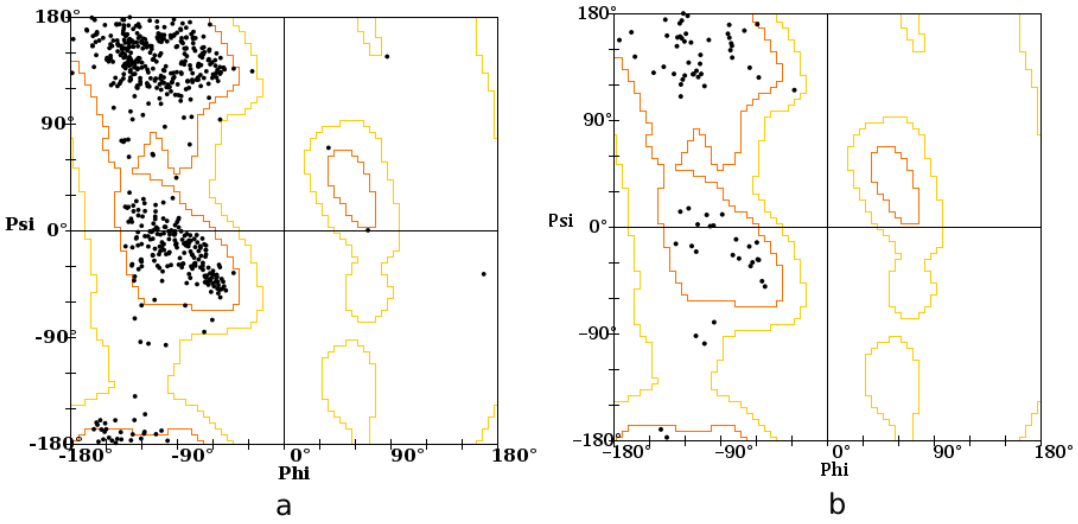

Figure 9. Ramachandran plot for threonine (a) tri- and (b) tetra- amino acids HPAAs occurring in irregular structures.

6.1.14 Histidine

Few occurrences of histidine HPAAs as tri-AA and tetra-AA were observed. Since the 6-residue His-tag is usually an artefact and not a part of the native protein sequence, such occurrences were omitted from the analysis. In case of tri-AA, most occurrences were noted to belong to irregular structures (Table 17).

Table 17. HPAAs identified for histidine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
His	3H	5	9	4	17	35
	4H	2	0	0	1	3

6.1.15 Glycine.

Occurrences of Gly HPAAs in irregular structures as tri-AA were observed to dominate amongst all HPAAs of Gly (Figure 10a). 68 occurrences of tetra-AA were observed in irregular structures (Figure 10b) while 3 each were found in helix and sheets. Gly HPAAs showed 8 occurrences in irregular structures (Figure10c) and 3 in sheets as penta-AA. Only single occurrence of hexa-AA HPAA was in irregular structures (Table 18).

Table 18. HPAAs identified for glycine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Gly	3G	26	34	144	432	636
	4G	3	3	0	68	74
	5G	0	3	0	8	11
	6G	0	0	0	1	1

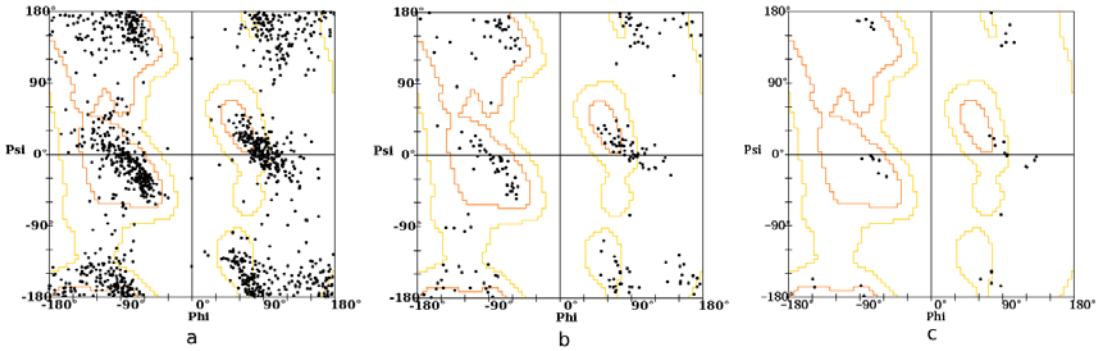

Figure 10. Ramachandran plots for glycine (a) tri-, (b) tetra-, (c) penta- amino acids HPAAs occurring in irregular structures.

6.1.16 Serine.

Considerable numbers of HPAAs were observed for Ser. Ser HPAAs were majorly observed to occur in irregular structures (Figure 11a,b). The HPAAs were found to exist as tri-AA to octa-AA. Only one occurrence was noted for hexa-AA and octa-AA with both observed in irregular structures (Table 19).

Table 19. HPAAs identified for serine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Ser	3S	50	40	50	303	443
	4S	2	0	2	23	27
	5S	1	1	0	2	4
	6S	0	0	0	1	1
	8S	0	0	0	1	1

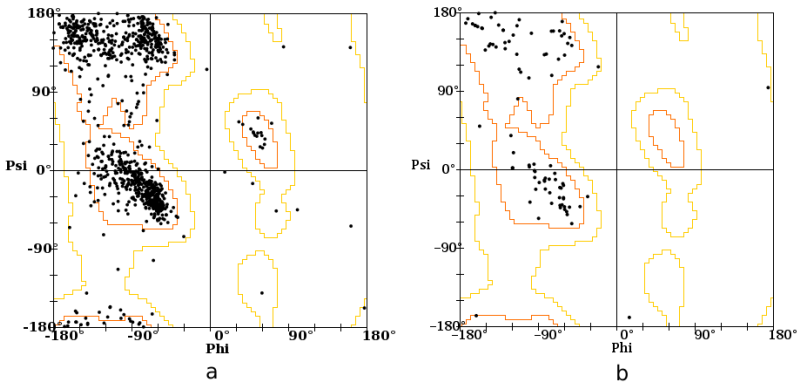

Figure 11. Ramachandran plots for serine (a) tri- and (b) tetra- amino acids HPAAs occurring in irregular structures.

6.1.17 Aspartic acid.

Asp HPAAs were observed as tri-AA, tetra-AA and penta-AA. Most occurrences were observed as tri-AA and majority in irregular structures (Figure 12). Of these 142 were in irregular structures while the remaining 98 were found to lie in turns (Table 20).

Table 20. Various HPAAs identified for aspartic acid in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular Regions	
Asp	3D	31	1	98	142	272
	4D	2	0	2	12	18
	5D	0	0	0	1	1

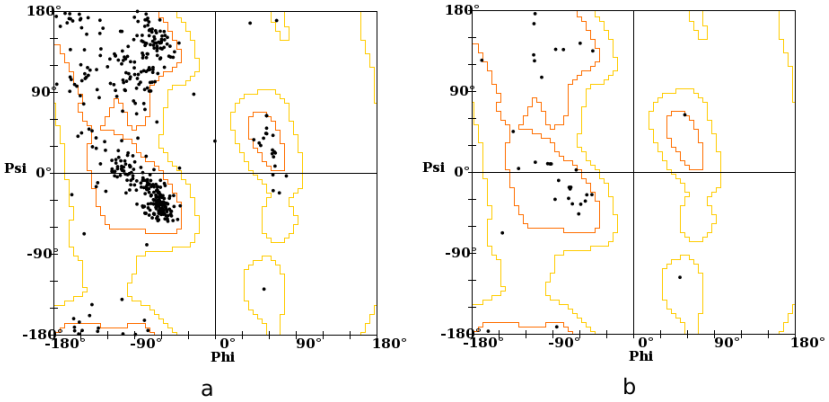

Figure 12. Ramachandran plots for aspartic acid (a) tri- and (b) tetra- amino acids HPAAs occurring in irregular structures.

6.1.18 Proline.

The existence of Pro HPAAs was observed exclusively in irregular structures (Figure 13). In case of tri-AA, 175 were observed in irregular structures, while 7 were found in turns. For tetra-AA and penta-AA occurrences of HPAAs were observed only in irregular structures (Table 21).

Table 21. HPAAs identified for Proline in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Pro	3P	0	0	7	175	182
	4P	0	0	0	24	24
	5P	0	0	0	8	8

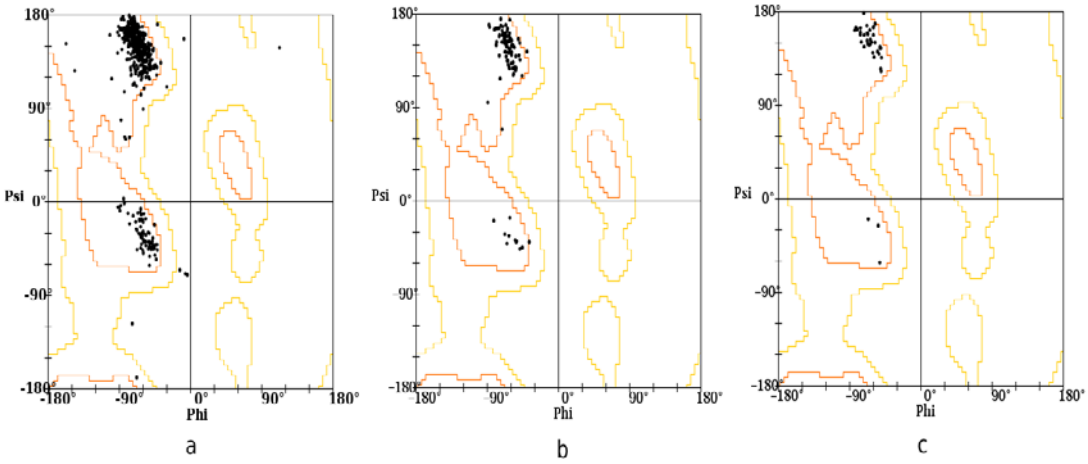

Figure 13. Ramachandran plots for proline (a) tri-, (b) tetra, (c) penta- repeat amino acids HPAAs occurring in irregular structures.

6.1.19 Asparagine.

Comparably higher numbers of tri-AA and tetra-AA HPAAs were found for asparagine in irregular structures than helices and sheets (Figure 14). Only one penta-AA HPAA was observed in irregular structures. About 40 observations of tri-AA HPAAs were noted as belonging to turns (Table 22).

Table 22. HPAAs identified for asparagine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Regions		Total
				Turns	Other Irregular structures	
Asn	3N	15	6	40	112	173
	4N	2	0	1	6	9
	5N	0	0	0	1	1

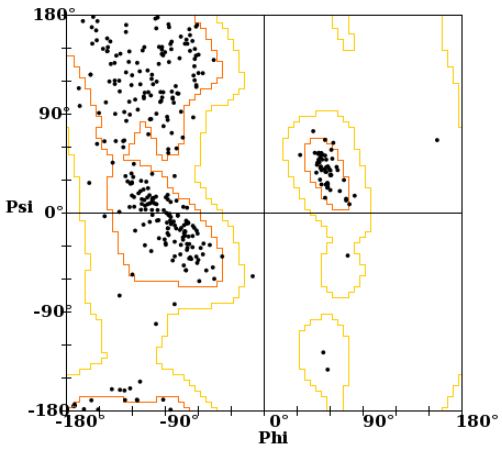

Figure 14. Ramachandran plot for asparagine triple amino acids HPAAs occurring in irregular structures.

6.1.20 Cysteine.

Only tri-AA was recorded for Cys HPAA. Out of the total five, two each were found in helices and sheets, while one was in irregular structure.

Table 23. HPAAs identified for cysteine in the local PDB dataset.

AA	HPAA	Helix	Sheet	Irregular Structural Regions		Total
				Turns	Other Irregular structures	
Cys	3C	2	2	0	1	5

6.2 Secondary structure prediction from amino acid sequences.

Protein three-dimensional structures are determined mainly using techniques such as X-ray crystallography and Nuclear Magnetic Resonance (NMR). While both methods are being increasingly used for structure determination, X-ray crystallography is limited by the difficulty for obtaining good quality crystals whereas

NMR can be carried out only for small proteins. Available studies have shown that the native conformation of the protein results from its amino acid sequence (Anfinsen, 1973), a wide range of computational methods have been researched for secondary structure prediction of proteins from the sequence. Earliest attempts were by focusing on the fact that proteins with high presence of proline residues possessed low helical content (Cohen and Szent-Gyorgyi, 1957), while Davies has shown a qualitative relationship between helix content and presence of residues such as Ser, Thr, Val, Ile, Cys (Davies, 1964) which are defined as helix breakers (Blout, et al., 1960). Edmundson, Braunitzer and Guzzo worked on an approach for predicting protein conformations from amino acid sequences based on helical regions in lysozyme (Braunitzer, et al., 1964; Edmundson, 1965; Guzzo, 1965). While Periti et al. have used helical and anti-helical pairs (Periti, et al., 1967), Low et al. used matching helical fragments of known structures (Low, et al., 1968). Wu and Kabat were the first to use dihedral angles of known tripeptide sequences to design 20 X 20 table of frequencies of occurrences of amino acids in helical and non-helical regions (Wu and Kabat, 1971). The three-state (α, β, coil) prediction algorithm was the first to predict β-regions (Finkelstein and Ptitsyn, 1971).

The method devised by Chou-Fasman for secondary structure prediction focused on use of amino acid propensities estimated by statistical procedures wherein conformational potentials or propensities were assigned to each amino acid in the protein (Chou and Fasman, 1974; Fasman, 2012). The propensities, one for each secondary structure type was obtained from the statistical analysis of proteins with known secondary structures. The ratio was calculated as the fractional existence of the residue in secondary structure element to the fractional occurrence in all structures. Several new methods based on algorithms such as multiple alignments, neural networks, information theory, nearest neighbor method, and hydrophobicity profiles have emerged over recent years. Although these methods have been observed to improve the accuracy of prediction as compared to the Chou-Fasman method, many researchers still prefer to use conformational potential for secondary structure predictions.

6.2.1 Estimating Amino acids propensity values for structure prediction in HPAAs.

The Chou-Fasman method originally developed was based on the analysis of only 15 proteins consisting of 2473 amino acids (Chou and Fasman, 1974). The data set was extended in 1989 to 64 proteins (Fasman, 1989) and 144 proteins in 1998 (Kyngäs and Valjakka, 1998) to improve the amino acid propensity values. Facchiano *et. al*. used the concept of structural classes to compute amino acid propensities for 2168 protein structures. Given below are the amino acids propensities for the three structural classes.

Table 24. The propensities of all 20 amino acids for three secondary structure states (Adapted from Facchiano *et. al.* (2006)).

Amino acid	Pα	Pβ	Pc
Ala	1.39	0.75	0.8
Cys	0.74	1.31	1.05
Asp	0.89	0.55	1.33
Glu	1.35	0.72	0.86
Phe	1.01	1.43	0.76
Gly	0.47	0.65	1.62
His	0.92	0.99	1.07
Ile	1.04	1.71	0.59
Lys	1.11	0.83	1
Leu	1.32	1.1	0.68
Met	1.21	0.99	0.83
Asn	0.77	0.62	1.39
Pro	0.5	0.44	1.72
Gln	1.29	0.76	0.89
Arg	1.17	0.91	0.91
Ser	0.82	0.85	1.24
Thr	0.76	1.23	1.07
Val	0.89	1.86	0.64
Trp	1.06	1.3	0.79
Tyr	0.95	1.5	0.78

Footnotes: Pα: refers to the propensity of each amino acid occurring in a helix. Pβ: refers to the propensity of each amino acid occurring in a beta-sheet. Pc: refers to the propensity of each amino acid occurring in coil region.

The methodology used by Facchiano *et. al*. for the calculation of the amino acid propensities was adapted for the calculation of propensity values for the occurrence of the amino acids in HPAA tracts. The residue propensity values calculated using the secondary structure content of proteins based on the secondary

structure assignment by DSSP (Kabsch and Sander, 1983). The propensity value for each residue is calculated as the ratio of frequency of occurrence of the residue as part of a HPAA tract in any particular (α-helix, β-sheet, coil) secondary structure to the frequency of occurrence in the protein subset.

$$\boldsymbol{P_{ij} = \frac{n_{ij}/n_i}{N_j/N_T}}$$

Where:

n_{ij}: The number of residues of type i occurring in HPAA tracts in structure of type j.

n_i: The total number of residues of type i occurring in HPAA tracts.

N_j: The total number of residues occurring in HPAA tracts structure of type j.

N_T: The total number of residues in the subset of PDB used in this analysis.

"i" is for each of the 20 naturally occurring amino acids.

"j" is for the three secondary structure state considered in the analysis, namely (α-helix, β-sheet, coil).

$$P_{ik}^{norm} = \frac{\left(P_{ik} - P_k^{min}\right)}{\left(P_k^{max} - P_k^{min}\right)}$$

The propensity values calculated for all 20 amino acids were then normalized as follows.
Where

P_{ik}: The propensity of each amino acid in secondary structure element of type k (α, β or coil).

P_k^{min}, P_k^{max}: The minimum and maximum values between the propensities P_{ik}.

6.2.2 Comparative analysis of Amino acids propensity values for structure prediction in HPAAs.

Based on the process described above propensity values were calculated for all the 20 amino acids.

Table 25: The propensities of all 20 amino acids in HPAA tracts for three secondary structure states.

Amino Acid	Pα	Pβ	Pc	Preference
Gln	**1**	0.018	0.146	Pα
Ala	**0.989**	0.074	0.107	Pα
Trp	**0.938**	0	0.213	Pα
Glu	**0.903**	0.046	0.204	Pα
Tyr	**0.804**	0.197	0.157	Pα
Arg	**0.714**	0.129	0.29	Pα
Met	**0.679**	0.19	**0.237**	Pα,Pc
Lys	**0.529**	0.107	**0.447**	Pα,Pc
Leu	**0.758**	**0.399**	0.022	Pα,Pβ
Phe	**0.497**	**0.612**	0.059	Pα,Pβ
Cys	**0.562**	**0.551**	0.056	Pα,Pβ
Val	0.133	**1**	0.032	Pβ
Ile	0.21	**0.963**	0	Pβ
Pro	0	0	**1**	Pc
Gly	0.055	0.078	**0.886**	Pc
Asp	0.161	0.005	**0.861**	Pc
Asn	0.133	0.044	**0.85**	Pc
Ser	0.156	0.117	**0.767**	Pc
Thr	**0.244**	**0.283**	**0.553**	Pα, Pβ, Pc
His	**0.325**	**0.318**	**0.495**	Pα, Pβ, Pc

Footnotes: Pα: refers to the propensity of each amino acid occurring in a helix. Pβ: refers to the propensity of each amino acid occurring in a beta-sheet. Pc: refers to the propensity of each amino acid occurring in a coil region. Significant propensity values in each secondary state have been shown in bold.

Seven amino acids namely Alanine, Glutamic acid, Leucine, Glutamine, Arginine, Tryptophan and Tyrosine were found to prefer the helix state in case of HPAA tracts. Comparing these with the results postulated by Facciano *et. al.*, similar trends were observed for the similar amino acids. However, an opposite trend was observed for the amino acids phenylalanine, isoleucine, lysine and methionine, which

showed preference for the helix structure generally, no such preferences were observed in terms of HPAA tracts. For the β-sheet state the preference of the amino acids isoleucine and valine could also be highlighted for the HPAA tracts. On the other hand propensity values for amino acids cysteine, phenylalanine and tyrosine that highlight their preference for β-sheets structure could not be observed in values for HPAA tracts. Interestingly the preference of tyrosine was found to shift from β-sheets to helices for the HPAA tracts. In term of coiled regions, the amino acids aspartic acid, glycine, asparagine, proline and serine were found to have a preference in HPAA tracts. These results correlated with those given by Facciano *et. al.*, except for cysteine in which the propensity was observed to be very low.

6.3 Summary.

This chapter has described the identification and structural analysis of HPAA tracts in local PDB dataset. The initial identification of HPAA tracts for all 20 amino acids revealed Ala tri-AA (three consecutive amino acid residues) to have the highest occurrence followed by Leu tri-AA. In terms of tetra-AA again Ala repeats were dominantly observed. Considerable occurrences of Leu and Gly tetra-AA repeats were also observed in the dataset. Penta-AA repeats were mainly observed for Ala and Gly. While Ala exhibited both hexa-AA and hepta-AA repeats, hexa-AA, hepta-AA and octa-AA repeats were identified only for Ser. The structural analysis carried out highlighted that amino acids Ala, Arg, Glu, Gln, Leu and Lys were preferred in repeats occurring in helices. Similarly, amino acids such as Val, Ile, Thr and Tyr were observed to have preference for sheets under HPAA tracts. Finally amino acids such as Asn, Pro, Ser, Asp and Gly were found to be preferred in Random Coil regions such as Loops, Coils and Turns as repeats.

The conformational values estimated for the HPAA tracts were mostly found to correlate with existing studies. However, for the amino acid tyrosine the structural preference was found to shift from sheets to helices while in case of cysteine a marginal preference was observed in Random Coil regions.

6.4 References.

Albà Soler, M. and Guigó Serra, R. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res. 2004; 14 (4): 549-54* 2004.

Albrecht, A.N., *et al.* A molecular pathogenesis for transcription factor associated poly-alanine tract expansions. *Human molecular genetics* 2004;13(20):2351-2359.

Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* 1973;181(4096):223-230.

Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Research* 2000;28(1):235-242.

Blout, E., *et al.* The dependence of the conformations of synthetic polypeptides on amino acid composition, *Journal of the American Chemical Society* 1960;82(14):3787-3789.

Braunitzer, G., *et al.* The hemoglobins. *Advances in protein chemistry* 1964;19:1-71.

Brown, L.Y. and Brown, S.A. Alanine tracts: the expanding story of human illness and trinucleotide repeats. *TRENDS in Genetics* 2004;20(1):51-58.

Calnan, B.J., *et al.* Analysis of arginine-rich peptides from the HIV Tat protein reveals unusual features of RNA-protein recognition. *Genes & Development* 1991;5(2):201-210.

Chou, P.Y. and Fasman, G.D. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* 1974;13(2):211-222.

Chou, P.Y. and Fasman, G.D. Prediction of protein conformation. *Biochemistry* 1974;13(2):222-245.

Chow, M.K., *et al.* Polyglutamine Expansion in Ataxin-3 Does Not Affect Protein Stability: implications for misfolding and disease. *Journal of Biological Chemistry* 2004;279(46):47643-47651.

Cohen, C. and Szent-Gyorgyi, A.G. Optical rotation and helical polypeptide chain configuration in α-proteins. *Journal of the American Chemical Society* 1957;79(1):248-248.

Cox, G.W., *et al.* Molecular cloning and characterization of a novel mouse macrophage gene that encodes a nuclear protein comprising polyglutamine repeats and interspersing histidines. *Journal of Biological Chemistry* 1996;271(41):25515-25523.

Cummings, C.J. and Zoghbi, H.Y. Fourteen and counting: unraveling trinucleotide repeat diseases. *Human molecular genetics* 2000;9(6):909-916.

Davies, D.R. A correlation between amino acid composition and protein structure. *Journal of Molecular Biology* 1964;9(2):605-609.

Dorsman, J.C., *et al.* Strong aggregation and increased toxicity of polyleucine over polyglutamine stretches in mammalian cells. *Human molecular genetics* 2002;11(13):1487-1496.

Edmundson, A.B. Amino-acid sequence of sperm whale myoglobin. *Nature* 1965;205:883-887.

Emili, A., Greenblatt, J. and Ingles, C.J. Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. *Molecular and Cellular Biology* 1994;14(3):1582-1593.

Fan, X., *et al.* Oligomerization of polyalanine expanded PABPN1 facilitates nuclear protein aggregation that is associated with cell death. *Human molecular genetics* 2001;10(21):2341-2351.

Fasman, G.D. The development of the prediction of protein structure. In, *Prediction of protein structure and the principles of protein conformation*. Springer; 1989. p. 193-316.

Fasman, G.D. Prediction of protein structure and the principles of protein conformation. Springer Science & Business Media; 2012.

Faux, N.G., *et al.* Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome research* 2005;15(4):537-551.

Finkelstein, A. and Ptitsyn, O. Statistical analysis of the correlation among amino acid residues in helical, β-stractural and non-regular regions of globular proteins. *Journal of molecular biology* 1971;62(3):613-624.

Gerber, H.-P., *et al.* Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* 1994;263(5148):808-811.

Giri, K., *et al.* Caspase 8 mediated apoptotic cell death induced by β-sheet forming polyalanine peptides. *FEBS letters* 2003;555(2):380-384.

Green, H. and Wang, N. Codon reiteration and the evolution of proteins. *Proceedings of the National Academy of Sciences* 1994;91(10):4298-4302.

Guzzo, A.V. The influence of amino acid sequence on protein structure. *Biophysical journal* 1965;5(6):809.

Inoue, K. and Keegstra, K. A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts. *The Plant Journal* 2003;34(5):661-669.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577-2637.

Kashi, Y. and King, D.G. Simple sequence repeats as advantageous mutators in evolution. *TRENDS in Genetics* 2006;22(5):253-259.

Kazemi-Esfarjani, P., Trifiro, M.A. and Pinsky, L. Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the (CAG) n-expanded neuronopathies. *Human Molecular Genetics* 1995;4(4):523-527.

Kyngäs, J. and Valjakka, J. Unreliability of the Chou-Fasman parameters in predicting protein secondary structure. *Protein engineering* 1998;11(5):345-348.

Lanz, R.B., *et al.* A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic acids research* 1995;23(1):138-145.

Low, B.W., Lovell, F. and Rudko, A.D. Prediction of alpha-helical regions in proteins of known sequence. *Proceedings of the National Academy of Sciences of the United States of America* 1968;60(4):1519.

Luo, H. and Nijveen, H. Understanding and identifying amino acid repeats. *Briefings in bioinformatics* 2014;15(4):582-591.

Mitchell, P.J. and Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* 1989;245(4916):371-378.

Nam, Y.-S., *et al.* Exchange of the Basic Domain of Human Immunodeficiency Virus Type 1 Rev for a Polyarginine Stretch Expands the RNA Binding Specificity, and a Minimal Arginine Cluster Is Required for Optimal RRE RNA Binding Affinity, Nuclear Accumulation, andtrans-Activation. *Journal of virology* 2001;75(6):2957-2971.

Oma, Y., *et al.* Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. *Journal of Biological Chemistry* 2004;279(20):21217-21222.

Periti, P., Quagliarotti, G. and Liquori, A. Recognition of α-helical segments in proteins of known primary structure. *Journal of molecular biology* 1967;24(2):313-322.

Perutz, M. Polar zippers: their role in human disease. *Protein science* 1994;3(10):1629-1637.

Pinto, M. and Lobe, C.G. Products of the grg (Groucho-related gene) family can dimerize through the amino-terminal Q domain. *Journal of Biological Chemistry* 1996;271(51):33026-33031.

Riley, B.E. and Orr, H.T. Polyglutamine neurodegenerative diseases and regulation of transcription: assembling the puzzle. *Genes & development* 2006;20(16):2183-2192.

Schwechheimer, C., Smith, C. and Bevan, M.W. The activities of acidic and glutamine-rich transcriptional activation domains in plant cells: design of modular transcription factors for high-level expression. *Plant molecular biology* 1998;36(2):195-204.

Simon, M. and Hancock, J.M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* 2009;10(6):R59.

Subirana, J.A. and Palau, J. Structural features of single amino acid repeats in proteins. *FEBS letters* 1999;448(1):1-3.

Tung, C.-H. and Weissleder, R. Arginine containing peptides as delivery vectors. *Advanced drug delivery reviews* 2003;55(2):281-294.

Voss, M., Schröder, B. and Fluhrer, R. Mechanism, specificity, and physiology of signal peptide peptidase (SPP) and SPP-like proteases. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 2013;1828(12):2828-2839.

Wu, T. and Kabat, E.A. An attempt to locate the non-helical and permissively helical sequences of proteins: application to the variable regions of immunoglobulin light and heavy chains. *Proceedings of the National Academy of Sciences* 1971;68(7):1501-1506.

Young, E.T., Sloan, J.S. and Van Riper, K. Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. *Genetics* 2000;154(3):1053-1068.

Chapter 7

Comparison of short sequence structural motifs and conclusions

Based on oppositely charged amino acids as flanking residues, sequence motifs involving Asp and Arg/Lys and Glu and Arg/Lys have been searched in a database of the structures of unique proteins and analysis presented in chapters 3,4 and 5. Several variations of the sequences containing a cationic amino acid (Arg or Lys) and anionic amino acid (Asp or Glu) having potential for mutual interaction by forming salt bridges when they are sequence neighbours or even when intervening residues are present between them in the sequence have been compared (Table 1).

Table 1. Comparative analysis of Asp and Arg/Lys and Glu and Arg/Lys as neighbours or when separated by one intervening residue in the sequence.

Motif	Secondary Structure	No. of Interactions	Type	Major Arg/Lys Conf	Major Asp/Glu Conf (χ1)	Major Glu Conf (χ2)
R-D	H	2	Type D	t t g+ t	g-	-
R-E	H	2	Type D	t t g+ t	g-	g-
D-K	H	1	MS	g-t t t	g-	-
K-D	H	1	MS	g+ t t g+	g+	-
E-K	H	1	MS	g- t t t	g+	g-
K-E	H	1	MS	t t t t	g-	g+
E-X-K	H	1	MS	g- t t t	g-	g+
D-X-R	S	2	Type B	g- t t t	t	-
R-X-D	S	2	Type B	t t t t	g-	-
E-X-R	S	2	Type D	g- t t t	t	t
R-X-E	S	2	Type D	t t t g+	g-	t
E-X-K	S	1	MS	g- g- t t	t	t
K-X-E	S	1	SS	t t t t	g-	t
D-R	Irregular structure	2	Type B	g+ t g+ t	g+	-
R-D	Irregular structure	2	Type B	t t g+ t	t	-
D-K	Irregular structure	1	MS	g- t t t	g-	-
K-D	Irregular structure	1	SS	t t t g-	t	-

E-K	Irregular structure	1	MS	g- t t t	g+	g-
K-E	Irregular structure	1	MS	g- t t t	g-	g+
D-X-R	Irregular structure	2,3	Type B, Type B +1	g- t g- t	t	-
R-X-D	Irregular structure	2,3	Type B, Type B + 1	g- g- g- g-	g+	-
D-X-K	Irregular structure	1	MS	g- t t t	variant	-
K-X-D	Irregular structure	1	MS	g- t t t	g+	-
E-X-R	Irregular structure	2, 3	Type B + 1/ Type D	g- t g- t	t	g+
R-X-E	Irregular structure	2,3	Type B + 1, Type D	g- g- g- g-	g+	g-/t
E-X-K	Irregular structure	1	MS	g- t t t	variant	variant
K-X-E	Irregular structure	1	MS	g- t t t	variant	variant

Footnote:. MS: Main chain – side chain hydrogen bond. SS: Side chain – side chain hydrogen bond.

7.1 Comparison of motifs in helices.

The motifs R-D, R-E, D-K, K-D, E-K, K-E and E-X-K were found to occur substantially in helices. The interactions remained favourable upto 3 intervening residues present. The motifs containing Arg were found to have two interactions of Type D with the Arg side chain conformation and the Asp/Glu χ_1 same. A similar observation was made for D-K, K-E and E-X-K motifs with one interaction. However, for the K-D and K-E, the Lys side chain was found to change from being partially extended (g+ t t g+) to fully extended (t t t t) while the Asp/Glu χ_1 changing from g+ to g-. For all motifs with Lys the hydrogen bonding was observed to be main chain – side chain.

7.2 Comparison of motifs in sheets.

Motifs were found to occur in significant numbers in sheets with interaction only when an intervening residue separating the terminal residues was present. Thus the presence of spacer residue allowed the terminal residue side chains to lie on the same side of the peptide plane. As the number of intervening residues increase the interactions also vanish. For motifs involving Arg the change from Asp to Glu resulted in the side chain bonding to shift from Type B to Type D. While in case of D-X-R and E-X-R the Arg side chain as well as the Asp/Glu χ_1 remained same, for the R-X-D and R-X-E motifs the Arg conformation changed slightly from t t t t to t t t g+ but the Asp/Glu χ_1 remained same. In motifs involving Lys (E-X-K and K-X-E) the hydrogen bonding was observed to shift from main chain – side chain to side chain – side chain along with the Lys side chain as well as the Glu χ_1 conformation.

7.3 Comparison of motifs in irregular structures.

Motifs with interactions were found to occur in considerable numbers in all motifs except E-R and R-E. While motifs involving Lys showed only a single interaction those involving Arg showed two or more interactions. The change from Asp to Glu did not show any major effect in D-K and E-K motifs with the Lys side chain and the hydrogen bonding remaining the same. In case of K-D and K-E the Lys side chain shifted from t t t g- to g- t t t along with the Asp/Glu χ_1 conformation changing from g- to g+. Even the hydrogen bonding observed to be side chain – side chain in K-D changed to main chain – side chain in K-E.

Motifs separated by a spacer residue and having Arg showed both two and three interactions. For the D-X-R and E-X-R motifs the Arg side chain, Asp/Glu χ_1 conformation and the hydrogen bonding remained the same. Further investigation into the backbone folding of these revealed them to be nearly the same. While similar observation was made for R-X-D and R-X-E, in few motifs belonging to R-X-E, Type D hydrogen bonding was also identified to occur along-with Type B. Finally for motifs having Lys (D-X-K, E-X-K and K-X-E), only one main chain – side chain bond was observed with the Lys side chain conformation remaining g- t t t but Asp/Glu χ_1 conformation varying significantly.

7.4 General Conclusions

Sequence motifs in proteins, is an arena that has been explored extensively over number of years. The advent of advanced bioinformatics tools for sequence analysis and the explosion of the number of full-length sequences added to databases like Uniprot has made this analysis easier and more accurate. This has led to annotation of many sequence motifs as signature for protein families as well as assignment of a functional importance. Prosite, which is the largest database of sequence motifs currently holds nearly 1300+ entries of sequence motifs that have been described as signature motif for protein families or some functional role. Structural motifs have largely been construed as combinations of regular secondary structures along-with some interconnecting irregular regions. Very few sequence motifs such as G-X-X-X-G and A-X-X-X-A that are known to occur in helices have been explored from a structural perspective.

With this in view, the work presented here endeavored to explore sequence-structure relationships through the analysis of a set of sequence motifs from a structural point-of-view. This included the identification of the motifs in the proteins, assignment of secondary structure, calculation of conformational angles and the calculation of the interactions involved. The first task in this study was to design a set of tools that could generate the necessary information accurately for a large numbers of structures in a short time. The iMotifs module designed through iDRP (http://irdp.ncl.res.in) was the result of these efforts. The tool allows identification of motifs in protein structures followed by secondary structure analysis, solvent accessibility and finally the calculation of non – covalent interactions and disulphide bonds if present. The development of this tool provided a major impetus to the further studies reported in this work. The selection of the motifs for analysis was based on the need to study the local folding assumed by the residues under consideration along-with the role of interactions. The analysis thus concentrated on the involvement of oppositely charged residues as sequence neighbors. This resulted in the Asp or Glu residues occurring at one end providing the anionic carboxylate groups in their side chains while the other terminus comprised of amino acids such as Arg or Lys contributing cationic guanidinium and ammonium group, respectively, in their side chain. Both salt bridges (ionic interactions) as well as hydrogen bonds contribute to protein stability through interactions between oppositely charged residues.

Based on the occurrence of oppositely charged residues motifs involving Asp/Glu at one end and Arg at the other separated by up to 8 residues were studied along with Asp/Glu at one end and Lys at the other separated by up to one residue. Unusually high numbers of motifs showing hydrogen-bonding interaction were found for Asp and Arg separated by one and three residues and Glu and Arg separated by three residues (Figure 1). The localization of the motifs in secondary structures revealed exceedingly high numbers of motifs in irregular structures for Asp and Arg separated by one residue and both Asp and Arg as well as Glu and Arg separated by three residues (Figure 2,4) in irregular structures and helices. Very low numbers of motifs were identified with hydrogen bonding interaction occurring in sheets (Figure 3). In motifs involving Asp and Arg, the D-X-R motif was observed to have an unusually high occurrence in irregular structures with both two and three interactions.

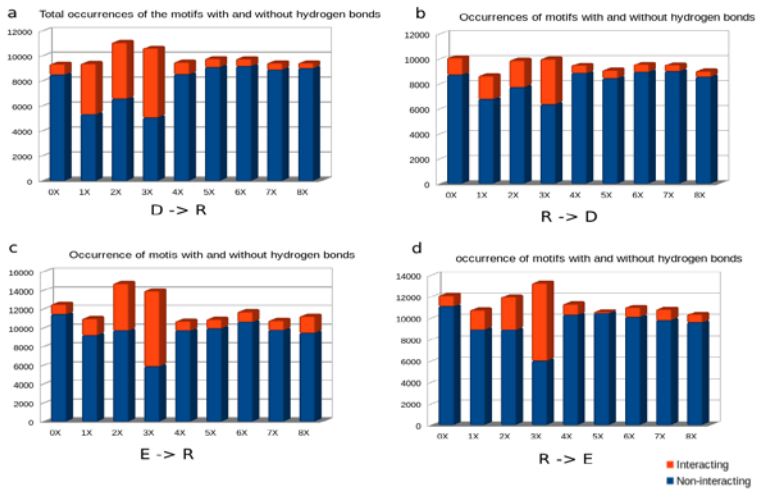

Figure 1. Occurrence of motifs involving Asp and Arg separated by up to 8 residues. Motifs have also been classified based to the presence or absence of hydrogen bonds.

For both groups the Arg side chain was observed to have a partially folded conformation g- t g- t while the Asp χ_1 conformation was observed to be t. The hydrogen bonding formed a pattern of three bonds wherein the two were involved in side chain – side chain interaction (Type B) with the third being a main chain – side chain: Arg (N) – (OD1) Asp. Although this bond was found to be missing in motifs with two interactions, the superimposition of the motifs revealed the backbone folding to remain the same. The X-residues in both groups were observed to occupy the same α-helix and left-handed helix region in the Ramachandran plot even though they occurred in irregular structures.

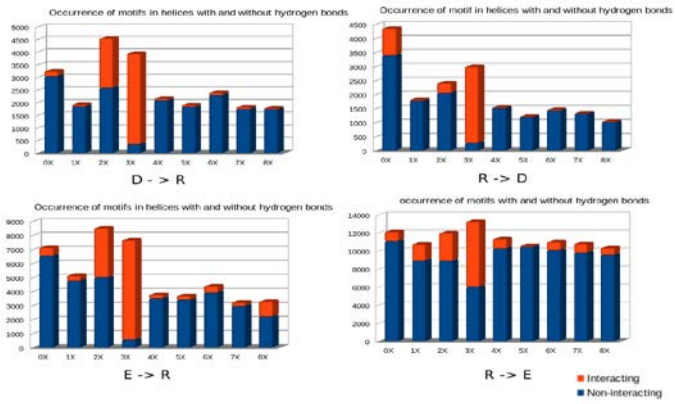

Figure 2. Occurrence of motifs involving Asp and Arg separated by up to 8 residues occurring in helices. Motifs have also been classified based to the presence or absence of hydrogen bonds.

Similarly important observation was the occurrence of the Arg side chain conformation g- t g- t with Glu χ_1 being t and χ_2 conformation being g+ in the E-X-R motif, where three interaction (Type B) with main chain – side chain: Arg (N) – (OE1) Glu was found. The backbone of this fold was found to be similar to that of D-X-R with three interactions already characterized along with the location of the motifs in the Ramachandran plot. Another important observation was the location of the motifs in case of D-X-R was usually in turns or loops connecting regular secondary structure elements. This observation was however found to be different in case of E-X-R where the motifs were identified to occur in longer loops.

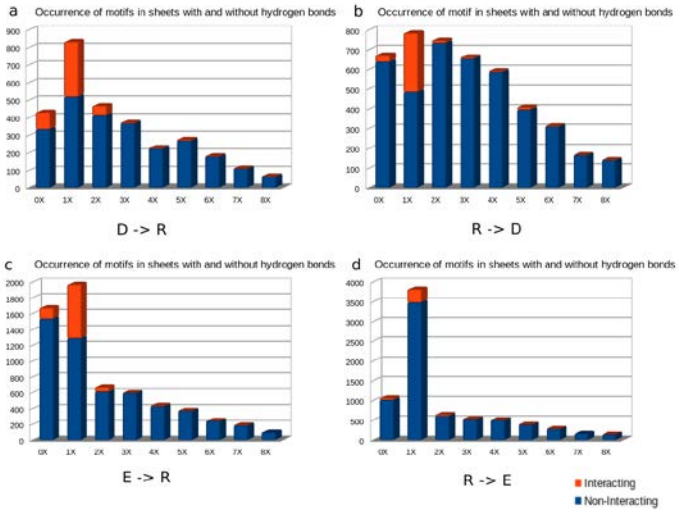

Figure 3. Occurrence of motifs involving Asp and Arg separated by up to 8 residues occurring in sheets. Motifs have also been classified based to the presence or absence of hydrogen bonds.

On extending the analysis carried out till now by introducing more spacer residues between the terminal residues Asp/Glu and Arg in both directions D/E -> R and R -> D/E (Figure 1) in case of motifs separated by three residues, exceedingly high numbers were found to occur in helices, which on further analysis also suggested that most of these H-bonds were the main chain – main chain i+4 -> I hydrogen bond, always found characteristic of α-helices. Thus this bond was disregarded during the analysis. In case of helices, for D–(3X)–R and R–(3X)–D motifs, the hydrogen bonding pattern varied from Type D to Type B, which in case of E–(3X)–R and R–(3X)–E remained the same (Type D). The Arg side chain conformation g- t g- t was observed for D–(3X)–R, R–(3X)–D and R–(3X)–E occurring in irregular regions with three hydrogen bonds. Contrary to the earlier observations, the Asp/Glu χ_1 which was found to be t, here the same was identified to be g-/g+. Even the regions occupied by these motifs were found to be nearly the same as previous observations for D-X-R and E-X-R motifs. These motifs were identified to occur in longer loops. In contrast, for these motifs in irregular regions with two hydrogen bonds, the Arg side chain conformation, Asp/Glu χ_1 conformation, location in the Ramachandran plot and hydrogen bonding was observed to be quite variant.

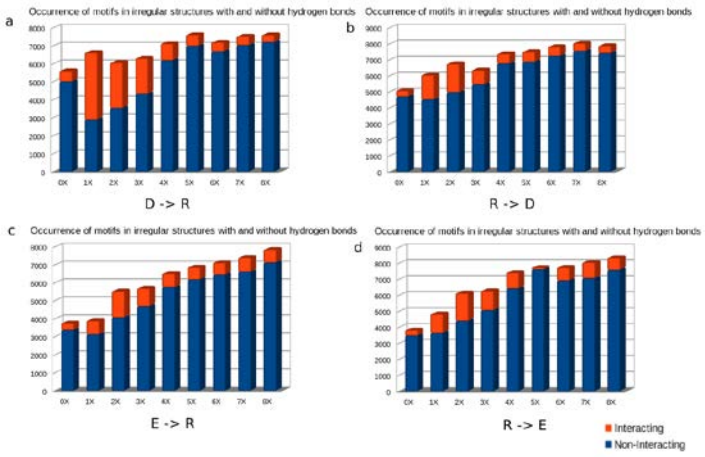

Figure 4. Occurrence of motifs involving Asp and Arg separated by up to 8 residues occurring in irregular structures. Motifs have also been classified based to the presence or absence of hydrogen bonds.

These results have led us to believe that the Arg side chain conformation may be an important factor in these type of structural motifs that involve oppositely charged residues; occurring in irregular structural regions such as turns and long loop regions and are involved in interactions. The interaction pattern observed here could thus provide additional stability to

the local structural folding of these motifs. When the positively charged amino acid was Lys rarely hydrogen bonded pattern could be identified. Thus, in further analysis sequence motifa involving Lys were not considered.

A very interesting observation is the extended type B interaction (in addition to type B, a H-bond between Asp/Glu side chain O and Arg amide N atoms) involving the flanking Asp/Glu and Arg residues was repeatedly found in different motifs having different number of spacer X-residues (1X to 3X) in between the oppositely charged amino acids. Very often the Arg rotamer involved is g- t g- t. So, this interaction may be playing an important role in imparting additional stability to local folds in irregular structural part in proteins.

Using the tools for identifying sequence motifs, in chapter 6 we studied the occurrence of Homopolymeric Amino Acids (HPAA) tracts in protein structures to look at their conformational preference for secondary structures. These tracts are known to play an important role in many functional proteins as well as have been implicated in many human diseases. Starting with a trimer of each of the 20 naturally occurring amino acids the study was extended to maximum of eight-residue HPAA.

While such eight-residue HPAAs were observed only for valine and serine, majority were identified as tri-AA for which, the highest were found in case of Ala. The occurrences were segregated by their presence in secondary structure groups of helices, sheets and irregular structural regions. Those occurring in irregular regions were further classified as those occurring in hydrogen bonded turns and other irregular regions (e.g. coils). Based on results obtained, the existing propensity values available for all the amino acids were updated, restricted to occurrence in HPAA. Here an α-helix preference was identified for Gln, Ala, Trp, Glu, Tyr and Arg while Val and Ile showed preference for β-sheets. The amino acids Pro, Gly, Asp, Asn and Ser were observed to have a preference for coils or irregular structures. Met and Lys were observed to have preference for both α-helices and irregular structures, where as Leu, Phe and Cys showed preference for both α-helices and β-sheets. Finally Thr and His were found to occur in any of the three secondary structure groups.