Structure and Sequence Level Analysis of βαβ Motifs and Loops in TIM Barrel Proteins: Implications for Protein Folding, Design and Engineering

THESIS

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

K.RAJA SHEKAR VARMA

ID. No: 2011PHXF404H

Under the supervision of

Prof. Ramakrishna Vadrevu

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

HYDERABAD CAMPUS

2017

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE,

PILANI - HYDERABAD CAMPUS

CERTIFICATE

This is to certify that the thesis entitled ***"Structure and Sequence Level Analysis of βαβ Motifs and Loops in TIM Barrel Proteins: Implications for Protein Folding, Design and Engineering"*** was submitted by **K.Raja Shekar Varma, ID.No. 2011PHXF404H** for the award of Ph.D. degree of the Institute embodies the original work done by him under my supervision.

Signature in full of the supervisor: _____

Name in capital block letters: **RAMAKRISHNA VADREVU**

Designation: **Associate Professor,**

Dept. of Biological Sciences

Date: _____

Acknowledgement

The work was carried out during 2012-2017 at the BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE-Pilani, Hyderabad Campus.

Firstly, I would like to express my gratitude to my advisor ***Prof. Ramakrishna Vadrevu****, for his valuable guidance and supervision throughout my research. He has been a tremendous mentor for me. I would like to thank him for encouraging me in my research and for allowing me to mature as a researcher.*

My deep sense of gratitude to ***Prof. Souvik Bhattacharyya****, Vice Chancellor (BITS) and* ***Prof. G. Sundar****, Director (BITS-Pilani, Hyderabad Campus), for permitting me to carry out my research work in the campus.*

I would also like to thank ***Dr. Debashree Bandyopadhyay****,* ***Dr. Durba Roy*** *for serving as my committee members.*

I also acknowledge ***Dr. Hanudatta S Atreya*** *for his valuable time and resources that helped me to progress in my research.*

*I am grateful to all the Faculty and HOD****,*** *Biological Sciences Department,* ***Dr. Naga Mohan****, BITS, Pilani, Hyderabad campus for their valuable support. I would also like to thank* ***Prof. Vidya Rajesh (ARD)*** *for their support.*

I would like to acknowledge all my friends who have taken part in this journey and in some way helped me along. Especially, ***Ravichand Palakurti****,* ***Shiva Prasad Bitragunta, Naresh and all co-scholars from Department****. I would also like to acknowledge all the lab technicians for their backing.*

I acknowledge University Grants Commission (UGC) and BITS, Pilani, Hyderabad campus for funding my fellowship.

At the end, a special thanks to my family. Words cannot express how grateful I am to my ***Mother (Parvathi Devi)*** *and* ***Father (Rama Raju)*** *for all of the sacrifices they have made on my behalf.*

Abstract

The $(\beta\alpha)_8$ / TIM barrel is one of the most common folds of known protein structures facilitating diverse catalytic functions. The fold is formed by the repetition of the basic βαβ building block in which the β-strands are followed by α-helices eight times alternating in sequence and structure. αβ and βα loops connecting α-helices to the β-strands and the β-strands to the α-helices contribute to stability and function respectively, an inherent imposition by the TIM barrel architecture itself. In this study αβ and βα loops from a dataset of 430 non redundant, high-resolution TIM barrels bearing sequence homology of <30% were analyzed for their amino acid propensities, sequence profiles and positional preferences of amino acids followed by loop dynamics analysis from tryptophan synthase alpha sub unit (αTS), a TIM barrel fold. The observed diversity, perhaps, dictates the distinct role of αβ and βα loops in stability and function respectively.

Given the distinct role of αβ and βα loops, the TIM barrel folds are aggressively pursued for functional and stability engineering. In this context, we developed, LoopX, a comprehensive database consisting of ~ 7,00,000 loop candidates of 3-14 residues in length, mined from non-redundant protein structures with <90% sequence similarity for comprehensive analysis of target and candidate loops for engineering. In addition an attempt was also made to identify the independently folding βαβ units from existing TIM barrel proteins.

In summary, the overall observations and reasoning will in addition to steering protein engineering efforts on TIM barrel design and stabilization can provide the basis for incorporating consensus loop sequences for designing independently folding βαβ modules and design/engineer novel or existing protein conformations.

Table of contents

LIST OF FIGURES

LIST OF TABLES

Abbreviations

α	alpha
β	beta
γ	gamma
α–α	alpha-alpha
α–β/α+β	alpha-beta
β–α–β/βαβ	beta-alpha-beta
Cα	C-alpha
α/β, αβ	alpha/beta, alpha beta
β/α, βα	beta/alpha, beta alpha
ϕ	phi
ψ	psi
RNase H	Ribonuclease H
Dnase	Deoxyribonuclease
DSSP program	Dictionary of Secondary Structure of Proteins
DSSP database	Database of secondary structure assignments of protein
SCOP	Structural Classification of Protein
CATH	Class Architecture Topology and Homologous superfamily
TIM	Triosephosphate Isomerase
αTS	alpha subunit of tryptophan synthase

PDB	Protein Data Bank
HTML	Hyper Text Markup Language
Perl	Practical Extraction and Reporting Language
GUI	Graphical User Interface
^{15}N	^{15}Nitrogen
^{13}C	^{13}Carbon
HET NOE/ [^{15}N-^{1}H] NOE	Heteronuclear Overhauser Effect
HSQC	Heteronuclear single quantum coherence
RMSF	Root Mean Square Fluctuation
RMSD	Root Mean Square Deviation
K	Kelvin

Abbreviations for Amino acids

Amino acids	3 letter code	1 letter code
Alanine	Ala	A
Arginine	Arg	R
Aspargine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoluecine	Ile	I
Luecine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Chapter 1

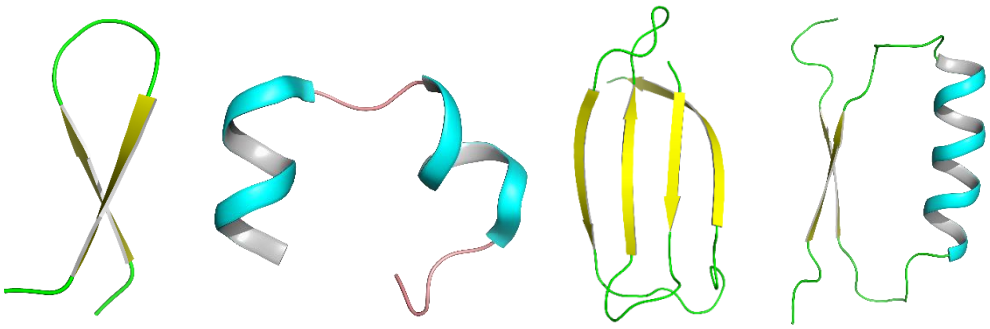
INTRODUCTION

Proteins are the essential biological macromolecules of life, made up of one or more polypeptide sequences comprising amino acid residues. The primary structure of any protein constitutes of linear amino acid residues linked by peptide bonds. Given, the amino acids preferences to adopt specific secondary structure conformations, the primary structure will fold into alpha helices or beta sheets or irregular random coils/loops. Alpha helices typically exist in right-handed spiral form, stabilized by i+4 -> i backbone (N-H->C=O) hydrogen bond interactions. Based on the parallel and anti-parallel arrangement of beta strands, beta sheets exist as parallel or antiparallel structures. The regular (α helices & β sheets) secondary structures of proteins are connected by irregular random coils/loops reversing the polypeptide chain, loops are often exposed to solvent and contribute to the enzymatic active sites and binding sites.

Loops are the highly variable flexible coil regions of proteins, which links regular secondary structures (α-helices & β-sheets) to form α–α, β–β, α–β and β–α units that make up all the known protein architectures. Despite their irregular structure, loops are known to acquire ordered conformations called turns. Based on their sequence length, amino acid preferences, backbone phi, psi values and Cα end-end residues distance, turns are classified into α, β, γ and π turn types. Loops are usually studied by grouping based on length (short <6 residues, medium 6-10 residues and longer loops >10residues) or based on the secondary structures they connect (α–α, β–β, α–β and β–α), many attempts were made to study loops by classifying based on various conserved features (Burke, Deane, & Blundell, 2000; Donate, Rufino, Canard, &

Blundell, 1996; Efimov, 1991; Espadaler et al., 2004; Kwasigroch, Chomilier, & Mornon, 1996; W. Li, Liu, & Lai, 1999; Oliva, Bates, Querol, Aviles, & Sternberg, 1997)

The secondary structure elements are further combined to form specific geometrical patterns called motifs/super secondary structures (Chothia & Finkelstein, 1990; Levitt & Chothia, 1976). The varying combinations of secondary structures have led to the formation of specific super secondary structures like beta hairpins, helix-loop-helix, Greek key motif, beta-alpha-beta motif (**Figure 1.1**) etc. Motifs/super secondary structures act as building blocks to form the larger three-dimensional conformations of structures called domains.



(A) **(B)** **(C)** **(D)**

Figure 1.1 *Super-secondary structures of proteins. A, B, C and D represent β-Hairpin motif, Helix-Loop-Helix motif, Greek Key motif and β–α–β motif structures respectively. The strands are indicated in yellow color, the helix in cyan and the loops in green.*

Sequence and structural analysis combined with computational methods have revealed that protein domains share structurally similar fragments like α–α, β–β, α–β and β–α units (Grishin, 2001). Studies have also stated that formation of tertiary conformation is preceded by native local structures due to folding initiation (Baldwin & Rose, 1999; Ramakrishna & Sasidhar, 1997). Further analysis also revealed that the intrinsic stable subunits of protein

domains are evolved by combinatorial assembly of smaller gene fragments (Lupas, Ponting, & Russell, 2001; Soding & Lupas, 2003) which have led to the divergence of domain architecture.

Pertaining to the concept that a complex protein structure is evolved from initial smaller independently folding units to favor stability, functionality and folding free energy. Several polypeptides with stable secondary and super-secondary structures like alpha helices, beta sheets, beta hairpins, alpha-helical motifs and beta alpha beta unit are designed and expressed in in-vitro systems and evaluated for their competency to fold independently (Blanco, Rivas, & Serrano, 1994; Cochran, Skelton, & Starovasnik, 2001; Dahiyat & Mayo, 1997; Ihalainen et al., 2008; Liang et al., 2009; Marqusee, Robbins, & Baldwin, 1989; Petukhov et al., 2009; Religa et al., 2007; Sadqi, de Alba, Perez-Jimenez, Sanchez-Ruiz, & Munoz, 2009; Searle, Williams, & Packman, 1995; Stanger et al., 2001; Struthers, Cheng, & Imperiali, 1996; Yakimov, Rychkov, & Petukhov, 2014; Zeng, Jiang, & Wu, 2016). Recently attempts were also made to build completely novel proteins by assembling stable structural units from varying proteins (Hocker, 2014; Nagarajan, Deka, & Rao, 2015; Soding & Lupas, 2003).

Keeping the evolutionary design of protein architecture in mind and the ability to design independently folding super secondary structures along with the availability of cutting-edge computational and experimental techniques, studies were instigated to pursue the design of novel protein structures with desired architecture, function, and stability.

Even though, four decades of protein folding studies have provided an understanding of the protein folding and unfolding for initial designing strategies, studies related to designing of new functional proteins with defined architecture has been dawdling. However, recent advances in computational and experimental techniques have led to the initial design of small α/β proteins, small Rossman fold and 4 helix bundles of approximately 100 residues in length (Eisenberg et

al., 1986; Harbury, Plecs, Tidor, Alber, & Kim, 1998; Kamtekar & Hecht, 1995; Regan & DeGrado, 1988) using computationally derived principles (Koga et al., 2012) and designing of larger proteins has been very challenging and pursued rigorously. Recently some success has been achieved in designing larger proteins using sub-domain structures and few attempts were also made to design proteins from scratch (Huang et al., 2016; Koga et al., 2012; Nagarajan et al., 2015). Attempts were even made to build new scaffolds using fragments from contemporary proteins but the creation of proteins that can host desired functionality and conformational stability has been long way demanding further study.

Role of Loops in Protein Folding, Stability & Function

Loop regions are known to be the critical determinants in protein folding, acting as protein folding initiation sites. The folding topology of a protein is often dominated by loops, The non-repetitive loops/reverse turn geometries connecting α helices and β strands provide compactness by bringing distant regions closer in space to ensue protein folding (Anderson et al., 2016; Colombo, De Mori, & Roccatano, 2003; Dyson & Wright, 1991; Lewandowska, Oldziej, Liwo, & Scheraga, 2010; Munoz, Thompson, Hofrichter, & Eaton, 1997; Ramirez-Alvarado, Blanco, Niemann, & Serrano, 1997; Richardson, 1981). Thus loops can also play not only a contributing role in stability but also in guiding the folding process (Fetrow, 1995; Hsu et al., 2006; Jager et al., 2008; Lewandowska et al., 2010; Marcelino & Gierasch, 2008; Ramakrishna & Sasidhar, 1997; Wright, Dyson, & Lerner, 1988; A. S. Yang, Hitz, & Honig, 1996).

Loops despite their irregularity in conformation contribute significantly to protein stability. More recent studies implicate the role of loops and turn conformations in reducing flexibility and encouraging stabilizing interactions between secondary structures of proteins (Anderson et

al., 2016; Balasco, Esposito, De Simone, & Vitagliano, 2013; Nagi AD, 1997; Predki, Agrawal, Brunger, & Regan, 1996; Simpson et al., 2005). Experimental data from representative proteins suggest that decreased loop length and increased turn propensity contribute to the protein stability (Anderson et al., 2016; Balasco et al., 2013; Fu, Grimsley, Razvi, Scholtz, & Pace, 2009; Nagi AD, 1997; Predki et al., 1996; Simpson et al., 2005), implying that loop length and turn propensity is one of the important determinants of protein stability. In fact, studies hint at the plausibility of the evolutionary requirement to select for turn sequences that confer thermodynamic stability (H. X. Zhou, Hoess, & DeGrado, 1996) and such turns can modulate the stability by their intrinsic preference to sample favorable ϕ, φ space (Fu et al., 2009; Predki et al., 1996; Simpson et al., 2005; Trevino, Schaefer, Scholtz, & Pace, 2007).

The conserved pattern of the configuration of short loop connections in β hairpins has indicated that loop curtailment is an adopted strategy by thermophilic RNaseH and thioredoxin for improved stability over their mesophilic counterparts (Balasco et al., 2013). Modeling of the effects of loop truncations revealed that increased folded state entropy can contribute significantly to stability (Gavrilov, Dagan, & Levy, 2015). A comparison study between mesophilic and thermophilic protein has revealed the role of loop dynamics in thermal stability (Vemparala, Mehrotra, & Balaram, 2011).

In proteins, the loops often define catalytic activity/function in protein framework (Fetrow, 1995; Fiser, Do, & Sali, 2000; L. N. Johnson, Lowe, Noble, & Owen, 1998) they mostly make up the active pocket of a protein. The higher flexibility of loops in proteins allows the active pocket to undergo conformational changes to facilitate efficient harboring and catalysis of a substrate. Many studies have shown that deletion or alteration of active site loops alters the protein catalytic activity and substrate binding capability (Fuller-Schaefer & Kadner, 2005; T.

A. Johnson & Holyoak, 2012). The Loops due to their higher diversity also plays a significant role in the immune responses as the six hypervariable loops (complementary determining regions) of antibodies are involved in effective antigen-antibody recognition and binding (Kim, Shirai, Nakajima, Higo, & Nakamura, 1999).

Apart from function, stability and recognition sites, loops are also involved in signaling cascades promoting protein-protein interactions (Bernstein et al., 2004; Zomot & Kanner, 2003), dimerization of proteins (Feng, Shi, Li, & Zhang, 2003; Fritz-Wolf, Schnyder, Wallimann, & Kabsch, 1996), as binding loops (Kawasaki & Kretsinger, 1995; Wierenga, Terpstra, & Hol, 1986) and are also helpful in membrane insertion of proteins(Benson, Huynh, Finkelstein, & Collier, 1998; Iacovache et al., 2006).

Loop Grafting/Engineering in Proteins

Due to their efficiency for frequent mutations that bring out new features to adapt with evolution and their diversification in folds for functional and stabilizing roles without compromising the structural integrity, loops have been aggressively targeted for engineering/grafting studies. Loop Grafting/engineering serves as an efficient technique to alter the enzymatic activity, conformation stability, and foldability of proteins, it is even used to create novel chimeric proteins. Loop grafting has been a resourceful technique in both industrial and medicinal field delivering proteins with altered features. In recent studies, Loop grafting/re-engineering was extensively used to alter the complementary determining regions (CDRs) to tune foldability, stability and function of antibodies (Ewert, Honegger, & Pluckthun, 2004; Jung & Pluckthun, 1997; Saerens et al., 2005; Sormanni, Aprile, & Vendruscolo, 2015; Xiong et al., 2014) and for effective vaccine development (Manoutcharian et al., 1999), using loop grafting technique many attempts were made to design novel binding proteins from non-immunoglobulin proteins

(Binz, Amstutz, & Pluckthun, 2005; Richter, Eggenstein, & Skerra, 2014; Smith, Tachias, & Madison, 1995). Designed chimeric proteins with extracellular loops grafted from receptors on to stable scaffolds were also efficiently used for protein-ligand interaction studies (Walser, Kleinschmidt, Skerra, & Zerbe, 2012). Loop grafting/re-engineering was also used to modify the catalytic activity of the existing scaffolds, for efficient molecular recognition (Boersma et al., 2008; Park et al., 2006) and functionality to create novel enzyme structures (Ochoa-Leyva et al., 2009).

TIM Barrel Fold & Architecture

Given the super secondary structure composition/fold architecture, proteins are mainly categorized into four classes (Knudsen & Wiuf, 2010; Lo Conte et al., 2000) I. Alpha/Beta (α/β) proteins II. Alpha-Beta (α–β) proteins III. All Alpha (α) proteins and IV. All Beta (β) proteins. Based on the super secondary structure composition and their arrangement in three-dimensional space the protein structures are further assigned to diverse folds like TIM barrel, flavodoxin fold, SH3 domain, immunoglobulin fold etc.

Alpha beta (α/β) class of proteins are the most frequent domain class in protein structures. A noticeable observation is the fact that the β-strands arrangement is mainly parallel unlike the anti-parallel arrangement of all β-sheet proteins. The β–α–β super secondary motif is the building blocks of α/β proteins. A regular repetition of repeating β–α–β super secondary motif in which helices and strands alternate with each other results in the formation of a stable inner central core of β-sheet and an outer layer α-helices packing against the central core as shown in **Figure 1.2**. Loops join the helices and strands as shown in **Figure 1.2**. A majority of α/β

proteins are cytosolic and are enzymes. Rossman, flavodoxin, TIM barrels and Leucine-rich horse-shoe folds fall in this category of proteins.

TIM barrel or $(\beta/\alpha)_8$ fold first observed and named after Triosephosphate Isomerase (TIM) in 1975 (Gracy, 1975), is one of the most ancient and frequently observed motif in proteins as evident from the fact that they constitute about 10% of all known protein structures in Protein Data Bank(H.M. Berman et al., 2000; Nagano, Orengo, & Thornton, 2002; Wierenga, 2001). TIM barrel folds with structural integrity and functional diversity catalyzes versatile enzymatic reactions among 5 of 6 enzyme classes of proteins, which include Oxidoreductases, Transferases, Lyases, Hydrolases and Isomerases (Anantharaman, Aravind, & Koonin, 2003; Nagano et al., 2002; Sterner & Hocker, 2005; Wierenga, 2001).

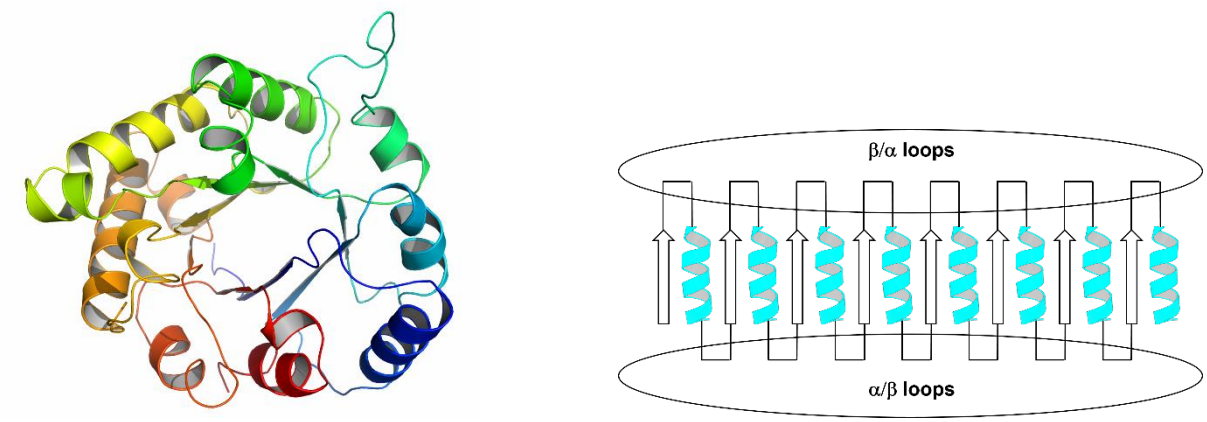
A TIM barrel fold is typically of >250 amino acid residues in length, can be monomeric or multimeric. These proteins are also called as α/β protein folds as they comprise of eight parallel β-strands (β1–β8) alternating with 8 alpha helices (α1–α8) in sequence and structure and interconnected by loops (Wierenga, 2001) as shown in **Figure 1.2**. The eight beta strands interconnected by alpha helices form the hydrophobic core of the protein giving a closed barrel architecture. The beta strands making up the barrel are further connected and stabilized by hydrogen bonds, each strand is hydrogen bonded to succeeding and preceding strands in order and the N-terminal strand is hydrogen bonded to C-terminal strand sealing the beta barrel. The eight amphipathic alpha helices form outer rim of the barrel which is maintained by helix-helix or helix-strands interactions shielding the core and considered more hydrophobic region than the core.

In the fold, the adjacent parallel beta strand with intervening alpha helix makes a β–α–β super secondary structure. Theoretical and experimental studies have revealed that the repeating β–α–β motif serves as a minimal unit of stability (Frenkel & Trifonov, 2005; Nagano et al., 2002; X. Yang, Kathuria, Vadrevu, & Matthews, 2009; Zitzewitz, Gualfetti, Perkons, Wasta, & Matthews, 1999) in TIM barrel folds. Further, the gene duplication of this fundamental building block has led to the proposal that higher order β/α structures might have evolved from this basic folding unit (Gerstein, 1997).

αβ and βα loops of TIM barrel proteins

In TIM barrel proteins loops connecting beta strands to alpha helices are referred as βα loops and loops connecting alpha helices to beta strands are referred as αβ loops (**Figure 1.2 B**). The active sites of all TIM barrel proteins are known to be harbored at the C-terminal ends of the β-strands by eight βα loops (Sterner & Hocker, 2005), as shown in **Figure 1.3 B**. The other end of barrel (N-terminal) comprises of αβ loops connecting alpha helices to beta strands which are known to contribute to conformational stability (Sterner & Hocker, 2005) as shown in **Figure 1.3 B**. The βα loops usually tend to be longer than αβ loops as observed in some proteins like TIM, HisA and HisF (Lang, Thoma, Henn-Sax, Sterner, & Wilmanns, 2000; Wierenga, 2001). Recently, Matthews and his co-workers also attributed the role of short ordered αβ loops to explain the increased stability at the N-termini of the strands in the *E.coli* α- subunit of tryptophan synthase (αTS), a TIM barrel protein (Vadrevu, Wu, & Matthews, 2008). This unique division of loops and their roles in function and stability makes them good targets for enzyme re-engineering to alter the function and stability without compromising each other. Recent studies have demonstrated that loop grafting is an efficient technique to alter the loops

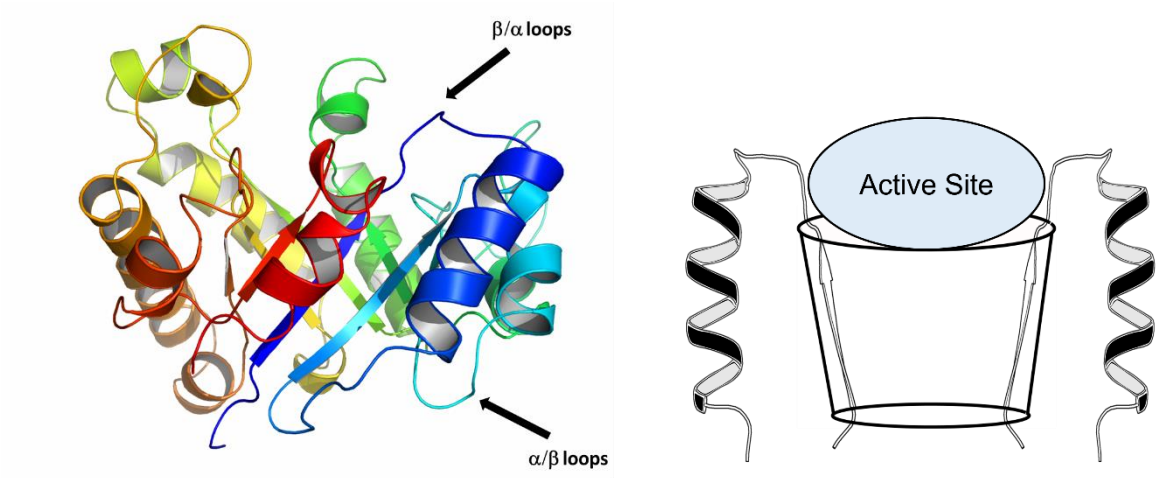
of a TIM barrel enzyme without altering its scaffold nature (Boersma et al., 2008; Ochoa-Leyva et al., 2011; Ochoa-Leyva et al., 2009).



(A) **(B)**

Figure 1.2: *(A) Representative Crystal structure of TIM barrel protein. PDB code: 8TIM.*

(B) Topology diagram of TIM barrel protein structure. Arrows and boxes represent β-strands and α-helices respectively. The black lines connecting the strands and helices constitute the top and bottom βα and αβ loops respectively.



(A) **(B)**

Figure 1.3: *(A) Crystal structure of TIM barrel showing βα and αβ loops. PDB code: 8TIM.*

(B) Active site geometry of TIM barrel proteins.

GAPS IN EXISTING RESEARCH

As discussed above TIM barrels are the most common folds observed with catalytic versatility spread over 5 of 6 enzyme classes. Given their loops segregation and importance in folding, stability, and function, TIM barrel loops are a good target for protein engineering to create proteins with new and improved function, stability and folding. Although researchers have managed to perform loop grafting to alter enzyme specificity and stability, difficulties in compatible loop selection for grafting/design and production of completely functional proteins still remains a challenge.

A systematic sequence and structural level study about loop length frequency, amino acid preferences, turn distributions, dynamics and conformational preferences of TIM barrel loops, provides a comprehensive knowledge of loops contribution to stability, folding and function. This knowledge will be useful for successful loop grafting/designing strategies or to fine tune TIM barrel proteins for desired features like improved folding, stability and enzymatic activity using rational design strategies. Though several studies are available on loops from proteins in literature, only couple of studies (Edwards, Sternberg, & Thornton, 1987; Scheerlinck et al., 1992) are available on 7 TIM barrel proteins and 17 α/β class of proteins dating back to more than two decades, considering the rapid growth in structural biology and discovery of new proteins and availability of high resolution protein structures, there is a compelling need for extensive sequence and structural level analysis of $\alpha\beta$ and $\beta\alpha$ loops from TIM barrel proteins for efficacious loop grafting/design and protein engineering studies.

Apart from TIM barrel proteins, recently grafting of loops between proteins has been aggressively pursued to incorporate novel features without compromising the structural integrity. But due to increase in a large amount of structural information, selection of

compatible loops from the existing protein structural data for efficient loop grafting/design has been taxing work and can be illustrated as finding a needle in a haystack. Although few databases and web servers (Espadaler et al., 2004; Knudsen & Wiuf, 2010; Ko et al., 2011; Lo Conte et al., 2000; Michalsky, Goede, & Preissner, 2003; Vanhee et al., 2011) are available for loop classification, modeling and design of new loops or missing loops in PDB crystal structures, no dedicated loop database with efficient search algorithms/tools is available to provide desired compatible loop candidates to aid loop grafting/design in proteins. Hence there is a need for dedicated comprehensive loop database with efficient search algorithms/tools and worldwide accessibility, to provide detailed sequence and structural level information to assist loop grafting/design, engineering, and modeling.

The de-novo design of proteins has always been a fantasy and few studies are carried out to design novel TIM barrel protein structures based on knowledge-based principles (Huang et al., 2016; Koga et al., 2012). The de-novo design of proteins not only generates proteins with new features but also sheds light on protein folding (Butterfield, Cooper, & Waters, 2005; Qi, Huang, Liang, Liu, & Lai, 2010). As discussed earlier, studies have shown that repeating βαβ units are the key part of TIM barrel scaffold. Therefore, identification of independently folding βαβ units from existing protein structures can aid in designing/building of novel protein structures and provides a deep understanding of TIM barrel architecture formation.

Although no natural independently folding β-α-β motifs were observed in isolation, recent studies have shown that highly stable independently folding β-α-β motifs can be designed by selecting amino acids with requisite propensities to favor beta strand, alpha helix and loop interactions for proper folding and stability of β-α-β motif (Butcher & Moe, 1996; Liang et al., 2009; Qi et al., 2010).

OBJECTIVES

1. Identification, extraction and structural and sequence level analysis of βα & αβ loops from non-redundant TIM Barrel proteins using computational and data mining techniques.
2. Development of a comprehensive loop database with a web-based graphical user interface (GUI) comprising sequence and structural level information of loops from all proteins with efficient query algorithms to aid engineer/graft and model loops in protein conformations.
3. Prediction and evaluation of independently folding βαβ units from TIM barrel proteins.

OVERVIEW OF THESIS

The thesis is focused on the identification, extraction, and analysis of αβ and βα loops from TIM barrel proteins, followed by the development of a loop database with efficient query algorithms to aid loop grafting/engineering. The thesis also focuses on identification evaluation of independently folding βαβ motifs from TIM barrel proteins.

i) TIM barrel is the most commonly observed scaffold and aggressively pursued protein engineering due to their common architecture and the distinguishable role of βα and αβ loops in protein function and stability. A non-redundant TIM barrel data set is considered for the analysis. Python and Perl programming languages are used for data mining segregation and extensive sequence and structure level analysis of TIM barrel loops.

ii) To analyze and evaluate the αβ and βα loop dynamics of TIM barrel proteins NMR experiments and molecular dynamics simulation techniques are used to study the alpha-subunit of tryptophan synthase a good TIM barrel motif model, which has been extensively studied for more than two decades.

iii) A comprehensive database is developed to assist protein engineering. A web-based graphical user interface with efficient search algorithms has been designed and implemented for easy asses and comprehensive analysis and identification of desirable candidate loops for grafting/engineering.

iv) βαβ units are known to be the building blocks of TIM barrel architecture. Molecular dynamic simulations alongside with additional computational approaches were implemented to identify and propose the independently folding βαβ units from naturally existing TIM barrel motifs.

A chapter-wise elaboration of the thesis is presented below:

In Chapter 2 the data mining, segregation and analysis of αβ and βα loops from TIM barrel proteins are carried out. A dataset of ~420 non-redundant TIM barrel proteins with less than or equal to 30% sequence similarity and x-ray crystallographic resolution of less than or equal to 3 are downloaded from the Protein Data Bank. An extensive sequence and structural level analysis of loops is carried out to assess loop lengthwise distribution, amino acid propensities, turn distribution, structural diversity, hydrophobic residues distribution to gain sequence and structural level features specifics.

In chapter 3, Molecular dynamics simulations techniques were implemented to analyze the loop dynamics of the alpha subunit of tryptophan synthase (αTS) in addition to an attempt to derive backbone dynamics properties from NMR. As the first step in that direction, the backbone NMR assignments were revisited. In a collaborative effort, the backbone assignments that were unavailable from an earlier study were obtained. NMR relaxation experiments were attempted to derive the dynamics of αβ and βα loops from αTS. Molecular dynamics simulation analysis was also carried out at varying temperatures to computationally evaluate the differences between αβ and βα loops dynamics and their contribution to the protein stability.

Chapter 4 deals with the development of a web-based comprehensive relational database housing protein loops of 3-14 residues length. A graphical user interface has been developed using HTML, Perl, Python and JavaScript programming languages to assist in easy access of the database. In-house developed query tools were also incorporated into the database GUI for flexible querying.

The fifth chapter exclusively deals with proposing potential of certain βαβ units from the existing TIM barrel motifs that may be stable and fold independently in isolation as an implication for protein engineering. Observation of secondary and super-secondary structures in isolated peptide fragments form proteins and in de novo designed sequences underscores the role of independently folding domains and their role in assembly of higher order structures. In this chapter, we address the finding of the needle(s) in haystack scenario, i.e., proposing/identifying likely βαβ candidates from the existing TIM barrel proteins that can fold independently. The likely βαβ candidates that could potentially fold into stable and well folded structures were explored based on certain sequence/structural features like loop length, propensity to fold into a helix, and stabilizing interactions, specifically the main chain to side chain hydrogen bonding interactions that clamp the βαβ units and play a significant role in stability of the fold. The short listed candidates were assessed for foldability and stability using Monte Carlo and molecular dynamics simulations approaches.

Chapter 2

Exploring Sequence and Structural Features of αβ and βα loop connections in TIM barrel Proteins

Introduction

Amino acid sequence dictates protein structure which in turn is responsible for function. The combination of the secondary structures, α-helices and β-sheets, connected by turns/loops form structural motifs such as β hairpins, helical hairpins coiled coils, βαβ etc., which serve as basic building blocks for the four major protein folds, α, β, α/β and α+β (Chothia & Finkelstein, 1990; Levitt & Chothia, 1976). While the α-helices and β-sheets provide the basic framework for structure and stability the intervening loops are the key participants in function.

The TIM barrel is one of the most commonly found folds of known protein structures (Nagano et al., 2002). The fold is formed by the repetition of the βαβ motif in which the β-strands are followed by α-helices alternating in sequence and structure. Basically, in this arrangement, two successive β-strands parallel to each other are joined by a α-helix via the loops. The loops connecting the β-strands to α-helices are referred to as βα loops while αβ loops correspond to the loops that connect the α-helices to the β-strands. The presence of eight parallel β-strands, each strand, hydrogen bonded to its preceding and succeeding strands forming a closed barrel is the defining feature of the TIM barrel $(\beta/\alpha)_8$ fold; hydrogen bonds between the N and C-terminal strands seal the barrel.

Experimental data (Zitzewitz et al., 1999) and bioinformatics analyses (Frenkel & Trifonov, 2005; Gerstein, 1997; Nagano et al., 2002) have indicated that a pair of adjacent parallel β-

strands and the intervening anti-parallel α-helix, i.e., the βαβ module, serves as the minimal unit of stability. Further, it has been suggested that gene duplication of this basic building block resulted in several common βα repeat structures, such as TIM barrel folds (Gerstein, 1997). Analysis of sequence and structural features, especially, of the loops that connect the αβ, βα hairpins are very limited and even that is available date back to more than two decades (Edwards et al., 1987; Scheerlinck et al., 1992). A systematic analysis of the sequence and structural features of the loops will provide sequence and structural patterns for successful grafting/design of loops or aids rational design of proteins.

TIM barrel scaffold displays an ability to facilitate the catalysis of a wide variety of reactions (Anantharaman et al., 2003; Sterner & Hocker, 2005; Wierenga, 2001). The active sites of TIM barrel enzymes are invariably comprised of the loops protruding from the C-termini of the β-strands contributing to the function of all TIM barrel enzymes. In contrast, the αβ loops at the opposite end of the barrel joining the C-termini of the α-helices and the N-termini of β-strands are linked to stability (Urfer & Kirschner, 1992). Given this distinct role of loops in function and stability, TIM barrels proteins offer a unique opportunity for engineering loops to create new and or improved functions without compromising the stability of the fold (Koga et al., 2012; Ochoa-Leyva et al., 2011; Ochoa-Leyva et al., 2009; Petsko, 2000).

Grafting of loops between proteins serves as a strategy to create new or alter existing catalytic function. In fact, alterations in active site residues and loop exchanges in TIM barrels (Ochoa-Leyva et al., 2011; Ochoa-Leyva et al., 2009), and other folds have lead to altered enzymatic activity. Specific to TIM barrels, two recent studies have demonstrated that loop exchange could serve as an efficient technique for tailoring enzyme activity without altering the overall

fold (Ochoa-Leyva et al., 2011; Ochoa-Leyva et al., 2009). Although some success could be achieved, selection of conformationally compatible loops in place of the target loop posed a challenge to match the native stability and activity (Ochoa-Leyva et al., 2011; Ochoa-Leyva et al., 2009). Therefore, for successful loop exchanges between TIM barrels, it is highly desirable to have a compilation of sequence and geometrical features of the βα and αβ loops which will be useful in guiding the selection of appropriate loops and assessment of their exchange potential via computational approaches.

Although numerous studies on loops from proteins, in general, are available in literature (Donate et al., 1996; Fernandez-Fuentes et al., 2004; W. Li, Liang, Wang, Lai, & Han, 1999; Thornton, Sibanda, Edwards, & Barlow, 1988; Vanhee et al., 2011), a dedicated compilation of the sequence and geometrical features of loops from TIM barrels is limited to a couple of studies (Edwards et al., 1987; Scheerlinck et al., 1992). Even these analyses were carried out on a very limited set of 7 TIM barrel proteins and 17 α/β class of proteins (Edwards et al., 1987). With the availability of a large number of high-resolution TIM barrel structures, there is now a compelling need for an exclusive analysis of αβ and βα loops. Further, it may be noted that, unlike anti-parallel β hairpins which are held by inter-strand hydrogen bonds, αβ/βα hairpin motifs are stabilized by the non-bonded interactions between the residues of strands and helices. As a result, sequence, size and geometry of the intervening connections between the strands and helices may be different from those that connect two anti-parallel strands in β hairpins. Therefore, for the successful design of αβ/βα hairpins and loop grafting, it is preferred to conduct a segregated analysis of loops from TIM barrel structures.

The desired outcome of this study is a compilation of αβ and βα loops from TIM barrel proteins, a comprehensive analysis of their size, composition and structural features. It is expected that such a study will not only offer clues for loop exchanges between TIM barrels but will also provide insights for rational design and or identification of independently folding βαβ motifs.

Methods

Identification and segregation

A set of 430 non-redundant TIM barrel PDB (H.M. Berman et al., 2000) structures with < 30% sequence similarity (the empirical cutoff value used to identify homolog sequences) and resolution ≤ 3.0 Å were shortlisted for this study using PDB advanced search option. The secondary structure information of the selected TIM barrel structures was extracted using the DSSP (Dictionary of Secondary Structure of Proteins) program (Joosten et al., 2011; Kabsch & Sander, 1983) which assigns the secondary structure of the protein. The intervening segments between the α-helices and β-strands were treated as loops and extracted using an in-house built Perl program and segregated as βα and αβ loops for analysis. Due to missing coordinates in some crystal structures, 6100 (95%) loops are successfully extracted from the expected 6450 loops. Out of the total 6100 loops extracted, 2819 and 3281 αβ and βα loops respectively, were segregated for analysis. The protocol adopted for the extraction, segregation, and analysis is summarized in the form of a flowchart and depicted in (**Figure 2.1**).

Propensities of amino acids

The propensity of each amino acid residue, ε_a, to occur in the TIM barrel loops was calculated based on the Chou-Fasman algorithm (Chou PY, 1978) using the equation

$$\varepsilon_a = \frac{a_s / n_s}{a_p / n_p}$$

where, a_s is the number of times residue *a* occurs in loops dataset, n_s, is the total occurrence of residues in the loop dataset, a_p is the total number of times residue *a* occurs in the TIM barrel dataset, and, n_p, the total residues in the TIM barrel data set. Propensity value of a residue > 1 indicates favorability of an amino acid to form the respective secondary structure and the propensity < 1 indicates disfavor (Gunasekaran, Ramakrishnan, & Balaram, 1997; Wang & Feng, 2003).

Position wise amino acid preferences analysis of αβ & βα loops

Position wise amino acid preferences for similar length αβ & βα loops were investigated using WebLogo (Crooks, Hon, Chandonia, & Brenner, 2004). WebLogo is a web-based application, used to generate a graphical representation of sequence logos. A sequence logo is the graphical representation of aligned amino acid sequences, sequence logo consists of stacks of amino acid single letter code symbols at every position. The overall size of the stack (symbols) at a position gives the amino acids conservation and size of the each symbol within the stack indicates the relative frequency of amino acids. WebLogo is featured in more than 2000 scientific publications.

Mining β-turn information from PDBsum

For the extracted αβ and βα loops, beta turns information is extracted from PDBsum (de Beer, Berka, Thornton, & Laskowski, 2014) a secondary structural database. In-house developed WebCrawler python scripts were deployed to automatically mine β-turns position, sequence, type, dihedral angles and hydrogen bonding information from PDBsum for respective loop candidates. The mined turn information is further sorted based on the type of β turn and analyzed.

Clustering

To observe conformational profiles of αβ and βα loops, loops of similar lengths were fitted and backbone atoms Root Mean Square Deviation (RMSD) is calculated using McLachlan algorithm (McLachlan, 1982) implemented in ProFit v3.1 program. Superimposed loops for a given loop length from the fits were clustered based on the backbone atoms RMSD matrix difference of ≤ 2 Å.

Analysis of Hydrophobic residue interaction & distribution in αβ & βα units

αβ and βα units connected by loops less than 7 residues in length were investigated for distribution of hydrophobic residues and their interaction. In-house built Python scripts are used to analyze the hydrophobic residues distribution in helix regions of αβ and βα units, for distribution analysis the region of the αβ and βα units helices are normalized to 0-1 where 0 is the N-terminal of the helix and 1 is the C-terminal of the helix, followed by identification of neighboring hydrophobic residues interactions within $4.5A^{0}$.

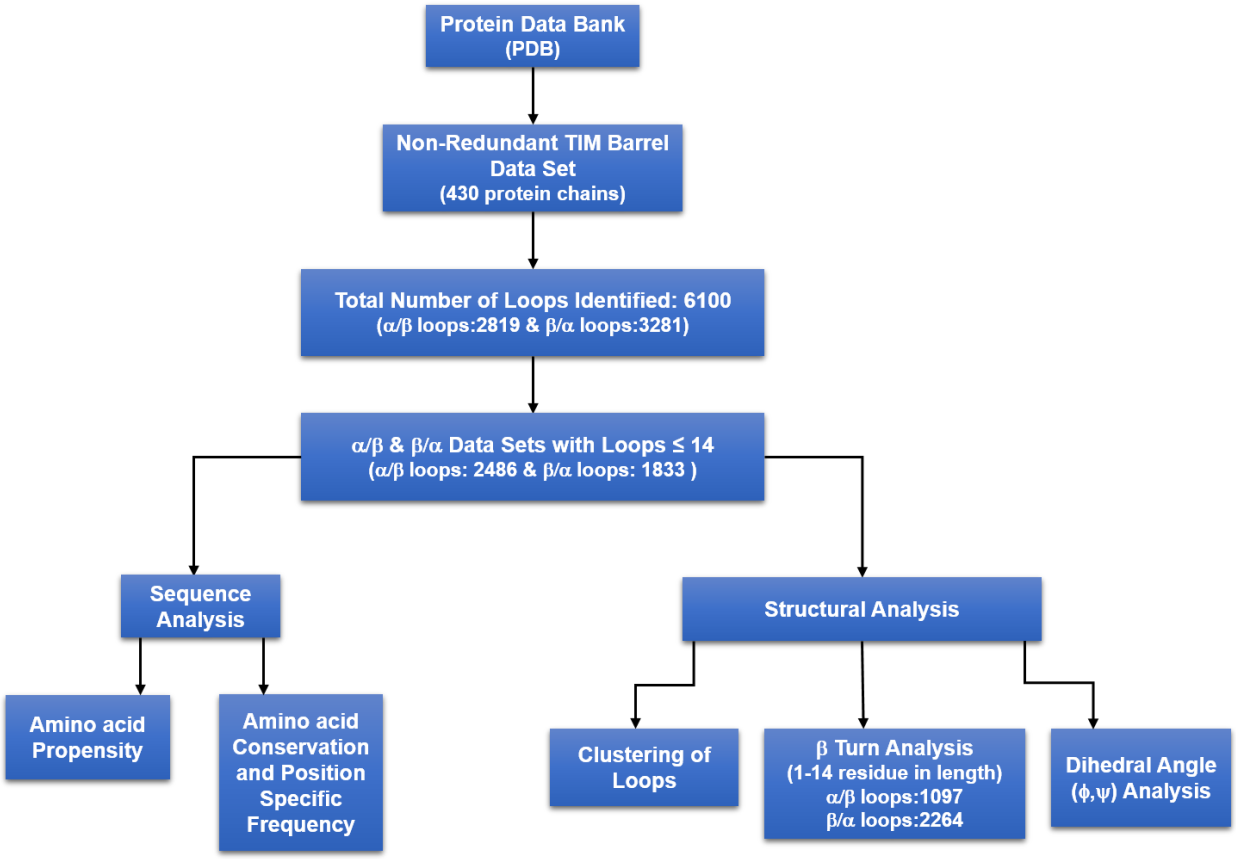


Figure 2.1: *Flowchart depicting the protocol and summary of the analysis of loops in TIM barrel proteins.*

Results and Discussion

Length distribution

The distribution of the loops lengths for the αβ and βα TIM barrel loops is shown as histograms in (**Figure 2.2**). It is clear that distribution of αβ loops is dominated by short connections, 2-6 residues long, increasing from 1-4, and steadily declining. In comparison, the distribution of βα loops is not as uniform as observed in αβ loops. Loops of length 1-4 residues are relatively fewer in number than the longer loops (>4 residues). In summary, more than 90% of the 2819 αβ loops are less than 10 residues long whereas 70% of the 3281 βα loops fall between 1-14 residues indicating that short loops are preferred in αβ loops.

This observed preference is in agreement with the view that the αβ loops are relatively shorter in length and contribute to the stability of the fold (Urfer & Kirschner, 1992). David Baker and his co-workers, using a combination of protein structural analysis and de novo simulations derived fundamental rules that threw light on the dependency of not only the lengths of the secondary structural elements but also on the length and chirality of the loops that connect them. It was found that foldability of the βαβ motifs is strongly dependent on the intervening loops that connect the αβ and βα structures (Koga et al., 2012). Guided by some of these tenets Huang et al., arrived at a basic βαβα (β_1-loop$_{(\beta\alpha1)}$-α_1-loop$_{\alpha\beta1}$-β_2-loop$_{(\beta\alpha2)}$-α_2-loop$_{(\alpha\beta2)}$) designed repeat unit that adopted a stable TIM barrel fold. The authors attributed the contribution of shorter loops to the observed robustness of the fold (Huang et al., 2016). In fact, experimentally it was observed that short loops increase the stability of not only α/β barrels (Urfer & Kirschner, 1992; Vadrevu et al., 2008) but also four helix bundle proteins (Nagi AD, 1997).

Another feature that is noteworthy is the prevalence of four residue connections (4 residues loops that connect the adjacent helices and strands) in αβ hairpins as against the preponderance of two residue connections β hairpins (Gunasekaran et al., 1997; Madan, Seo, & Lee, 2014). β hairpins in anti-parallel β-sheets are stabilized by the inter strand hydrogen bonding. Two residue connections, preferably the mirror image type I' and II', may be required for ideal orientation of the strands to be hydrogen bonded that run in opposite direction in β hairpins. On the contrary, αβ/βα hairpins, which are stabilized by the side chain packing between the hydrophobic residues of α-helices and β-strands, do not impose any such backbone hydrogen bonding constraint. Perhaps, loops longer than two residues have greater conformational flexibility which in turn can lead to optimum side chain interactions in αβ hairpins.

Based on the observed loops lengths distribution subsequent analysis has been restricted to ≤ 14 residues for extracting the amino acid propensities, sequence and structural patterns. Further, in view of the distinct role of loops in stability and function, αβ and βα loops have been considered separately rather than analyzing them together.

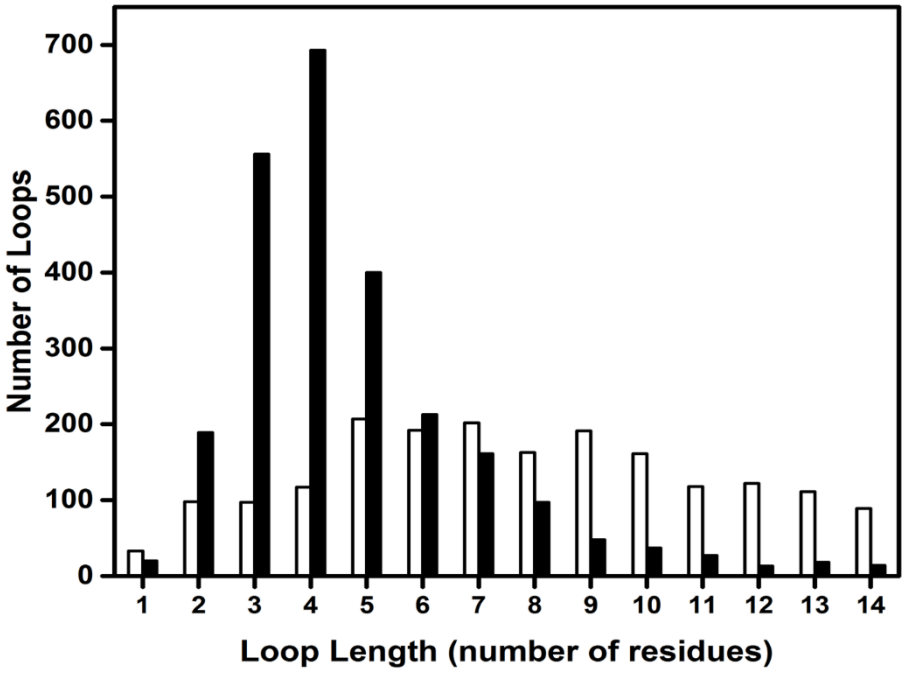


Figure 2.2: *Distribution of loops lengths in αβ (filled bars) and βα (open bars) TIM barrel loops.*

Amino acid propensities

For all the 20 amino acids the propensity to occur in the loops is shown in (**Figure 2.3**). A propensity value of ε_a >1 indicates higher tendency while a value of ε_a < 1 indicates lesser frequency to occur in loops (Gunasekaran et al., 1997; Wang & Feng, 2003). It is clear that residues glycine (G), proline (P), serine (S), threonine (T), asparagine (N), aspartic acid (D), histidine (H) possess greater propensity which is consistent with their normal tendency to occur more frequently in protein loops (Costantini, Colonna, & Facchiano, 2006; Gunasekaran et al., 1997). However, it may be noticed that for certain amino acids the preference to occur in αβ

and βα loops varies. Glycine, proline and the positively charged arginine, lysine residues show greater propensity to occur in αβ loops while serine, threonine, cysteine, histidine, tryptophan display tendency to occur more frequently in βα loops. Increased tendency to occur in the βα loops for serine, threonine, histidine, cysteine and tryptophan may be stemming from their role as catalytic residues in proteins (Bartlett, Porter, Borkakoti, & Thornton, 2002). Interestingly, the two small polar amino acids, asparagine and aspartic acid show similar preference to occur in αβ and βα loops.

The increased proline content observed in αβ is consistent with the occurrence of increased number of prolines, especially, in the loops of thermophiles (Bogin et al., 1998; C. Li, Heatwole, Soelaiman, & Shoham, 1999; Watanabe, Hata, Kizaki, Katsube, & Suzuki, 1997). Proline with its pyrrolidine ring is constrained to adopt limited configurations and can also restrict the backbone preferences for the preceding residue. Therefore, it possesses the lowest conformational entropy resulting in decreased entropy of the unfolded state (B. W. Matthews, Nicholson, & Becktel, 1987). Substitution of non-proline and non-glycine residues in turn conformations resulting in increased stability of the folded state (Hardy et al., 1993; Kimura, Kanaya, & Nakamura, 1992; Masumoto, Ueda, Motoshima, & Imoto, 2000; Stites, Meeker, & Shortle, 1994; Takano, Yamagata, & Yutani, 2001; Trevino et al., 2007), further is in support of the preferential distribution of glycine and prolines in αβ loops.

A perplexing observation, however, in contrast to the attributed role of charged residues in enzyme catalysis (Bartlett et al., 2002), is the distribution of the two positively charged residues arginine and lysine in αβ and βα loops. Given their higher distribution in enzyme active sites, it is rather, surprising to observe increased tendency for arginine and lysine to occur in αβ loops and not in the functional βα loops. In this context, it may be pertinent to note, from the

analysis of thermophilic proteins, that, the presence of higher ratios of charged amino acids located in the surface exposed loop regions increase the thermal stability due to salt bridge and ion pair interactions (Fukuchi, Yoshimune, Wakayama, Moriguchi, & Nishikawa, 2003; Nakashima, Fukuchi, & Nishikawa, 2003; Saunders et al., 2003; Suhre & Claverie, 2003; Szilagyi & Zavodszky, 2000; Xiao & Honig, 1999). In fact, mutational replacements of residues in loops to arginine, in some cases, provided an experimental demonstration of the enhanced thermal stability (Mortazavi & Hosseinkhani, 2011; Strub et al., 2004; Tanaka Y & I, 2004). These observations may provide a likely explanation for the rather unexpected distribution of arginine and lysine residues in αβ loops.

To corroborate this possibility, we have examined the interactions involving arginine and lysine residues with other amino acids arising from αβ and βα loops in the dataset of the TIM barrel proteins. From the examination of these interactions, we noticed that the number of interactions between the positively charged arginine/lysine side chains and the negatively charged aspartic/glutamic acid side chains are almost equal from αβ and βα loops. In contrast, we observed that the side chain (arginine/lysine) to main chain hydrogen bond interactions are about two times higher in αβ loops than in βα loops (**Figure 2.4**). And in most cases, involve the side chains of arginine or lysine in the αβ loops to the main chain carbonyl oxygen atom of residues in the preceding β-strand or the succeeding α-helix as shown in (**Figure 2.5**). This is reminiscent of the non-local side chain to main chain hydrogen bonding interactions, clamping βα hairpins and bracketing the βαβ modules in TIM barrels (X. Yang et al., 2009; X. Yang, Vadrevu, Wu, & Matthews, 2007). It was experimentally demonstrated that mutations leading to the disruption of these interactions lead to a dramatic loss in the stability of TIM barrel proteins (X. Yang et al., 2009; X. Yang et al., 2007). On a similar note, long-range side chain

to main chain hydrogen bonding interactions involving arginine and lysine resident in the αβ loops may contribute to the stability of the barrel. In fact from the study of seven TIM barrel proteins, it was speculated that similar long-range hydrogen bonds involving charged side chains of arginine and lysine with the backbone groups play an important role in stabilizing specific αβ loops (Scheerlinck et al., 1992). Such interactions that bestow rigidity are unlikely to be found in large numbers in βα catalytic loops where flexibility is obligatory for function (Dellus-Gur, Toth-Petroczy, Elias, & Tawfik, 2013; Pompliano, Peyman, & Knowles, 1990; Richard, Zhai, & Malabanan, 2014) .

In this context, it is interesting to note that the preference for the long chain positively charged amino acids in αβ loops can be recapitulated in the successful de novo design of TIM barrel structures. In the crystal structure variants of de novo designed TIM barrels, Huang et al.,(Huang et al., 2016) noticed long range hydrogen bonding interactions between the side chain of arginine located in αβ loops and backbone carbonyl oxygen of residues at the beginning of the immediately succeeding αβ loops. Based on their results from a series of designed TIM variants, together with a comparison of previous studies on designed TIM barrels (Nagarajan et al., 2015) such long range polar interactions were implicated not only to ensure strand registry but also critical for specifying the overall fold (Huang et al., 2016). That their role in important is further corroborated from the observed enhancement of thermal stability of the variants in which arginine was incorporated, presumably due to long range side chain to main chain hydrogen bonding interactions (Huang et al., 2016; X. Yang et al., 2009; X. Yang et al., 2007).

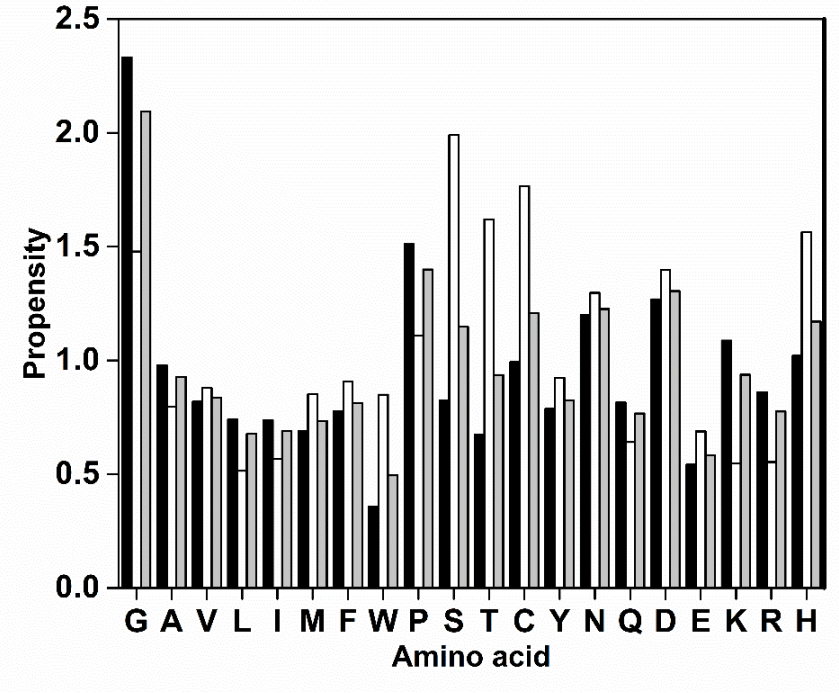


Figure 2.3: *Individual amino acid propensity to occur in αβ (filled bars) and βα (open bars) Shaded bars correspond to the overall propensity of amino acids in TIM barrel loops.*

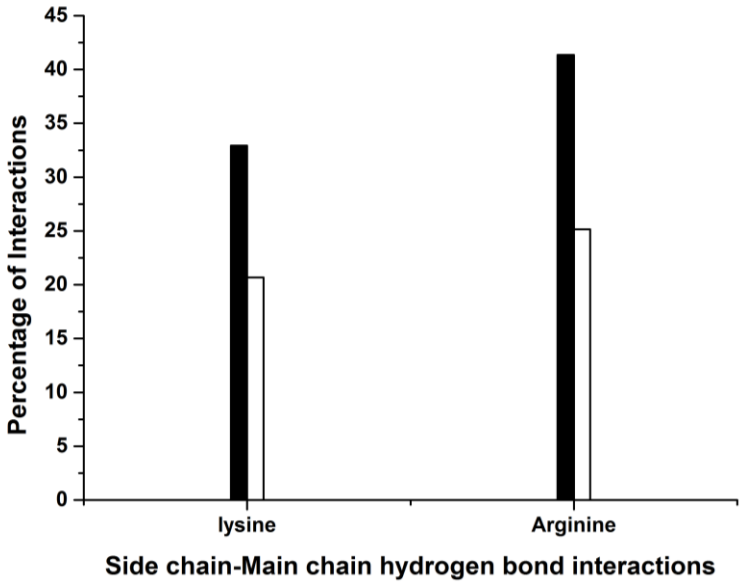


Figure 2.4: *Side chain to main chain hydrogen bonding interactions involving lysine and arginine residues. αβ loops (filled bars); βα loops (open bars).*

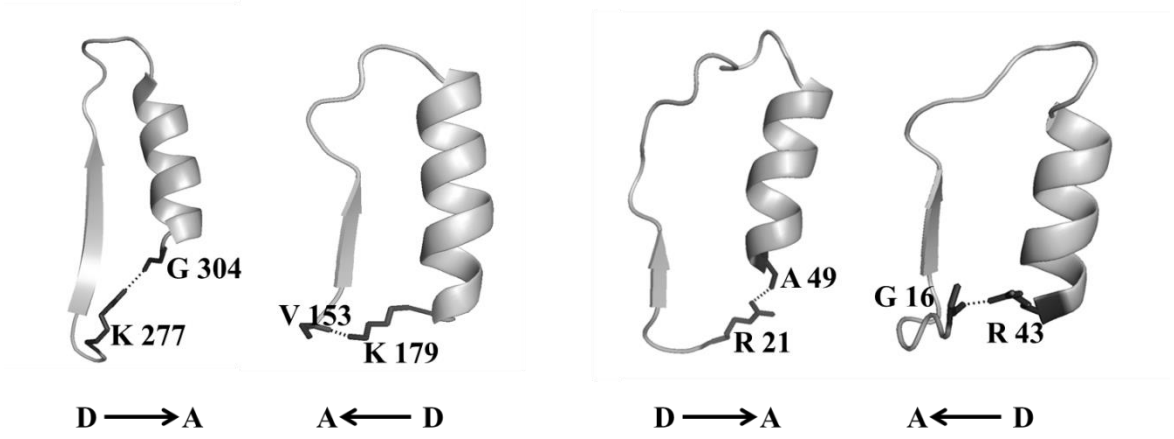


Figure 2.5: *Long-range side chain to main chain hydrogen bonding interactions. The side chains of arginine and lysine residues involved in the hydrogen bond interactions with the backbone carbonyl atoms are shown as sticks. D and A correspond to hydrogen bond donor and acceptor respectively.*

Sequence Features

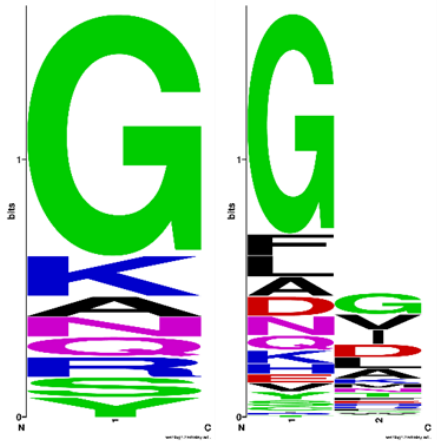
Sequence profiles of individual αβ and βα loops lengths showing position specific frequency and conservation of amino acids generated using WebLogo (Crooks et al., 2004) are shown in (**Figure 2.6**). Two observations are apparent from the profiles. First, the position-specific preference for amino acids appears to be different in αβ and βα loops. For instance, glycine, in particular, at the N-termini of short αβ loops is highly preferred. The obvious explanation for this observed preference for glycine at the beginning of αβ loops is that they serve as helix breaking signals preventing the continuation of the helix through the loop region. In addition to preventing the extension of the helix, side chain lacking glycines minimize potential steric clashes between the residues in the loop and the helix. Intriguingly, this consensus observed in a large body of naturally occurring TIMs appears to be recapitulated in the de novo design of a stable TIM barrel by Baker and his colleagues. In their βαβα (β_1-loop$_{(\beta\alpha1)}$-α_1-loop$_{\alpha\beta1}$-β_2-loop$_{(\beta\alpha2)}$-α_2-loop$_{(\alpha\beta2)}$) repeat unit that adopted a stable TIM barrel fold, the N-terminal residues

of the αβ loops 1 and 2 are glycines, an observation consistent with the preference for glycine at the N termini of the αβ loops.

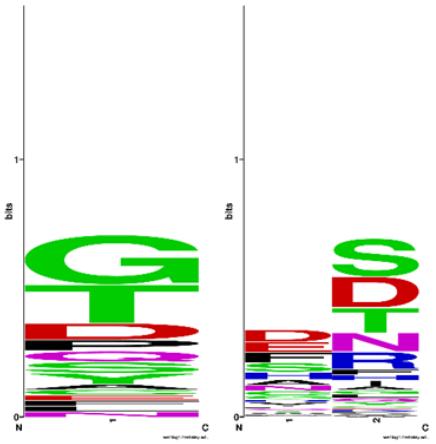
The second observation is the prevalence of short polar residues including aspartic acid (D), asparagine (N), serine (S), and threonine (T) at the C-terminal end of the β/α loops. Small polar residues are frequently observed at the beginning of α-helices in protein structures and serve as helix start signals by providing their side chains for hydrogen bonding with the backbone amide nitrogen of the first residue in the helix (Aurora & Rose, 1998; Bordo & Argos, 1994; Doig, MacArthur, Stapley, & Thornton, 1997). Therefore, the observed tendency for small polar residues to populate at the end of βα loops indicates their N-cap propensities for helix formation.

α/β

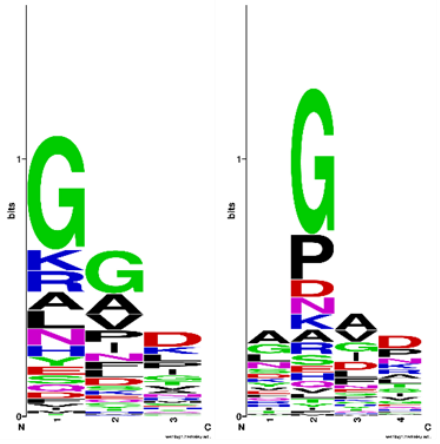
β/α



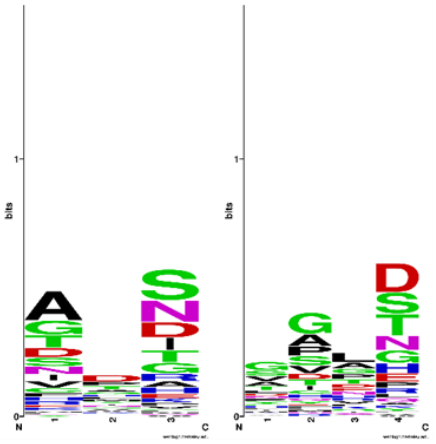
(i) (ii)



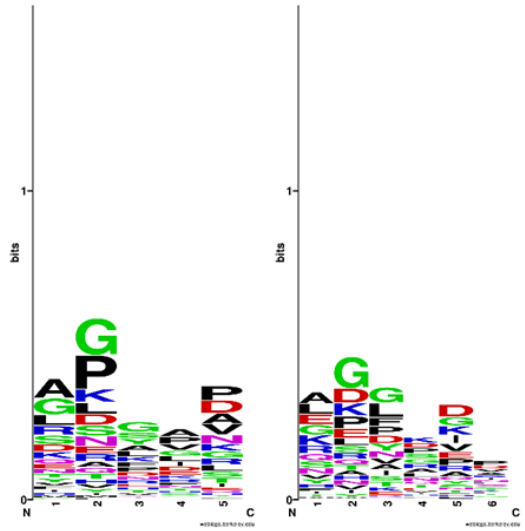
(xv) (xvi)



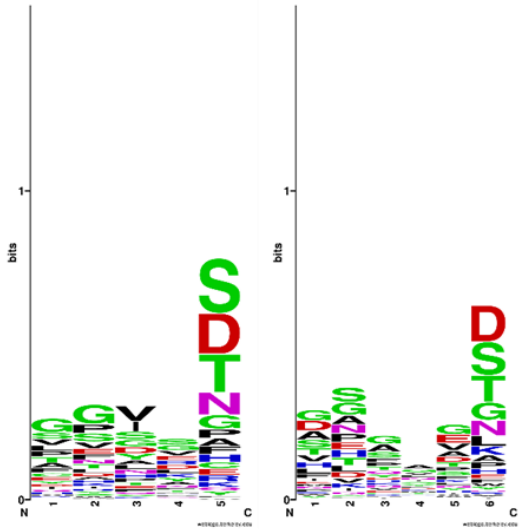
(iii) (iv)



(xvii) (xviii)

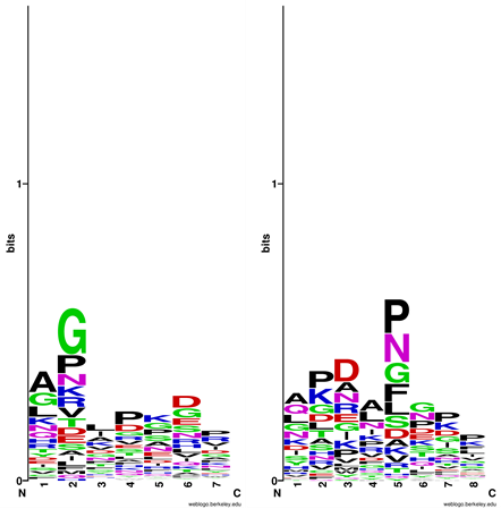


(v) (vi)



(xix) (xx)

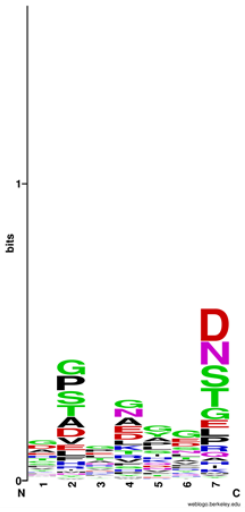
α/β



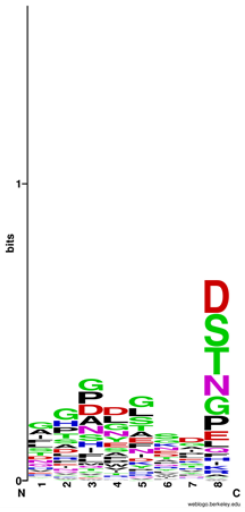
(vii)

(viii)

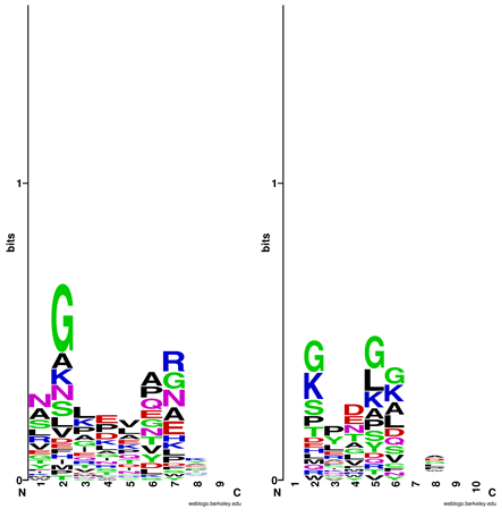
β/α



(xxi)

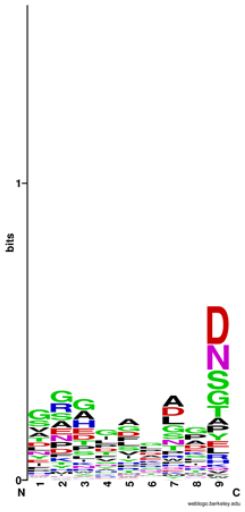


(xxii)

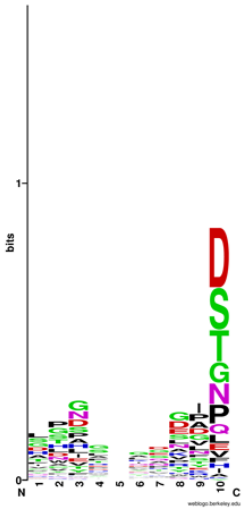


(ix)

(x)



(xxiii)



(xxiv)

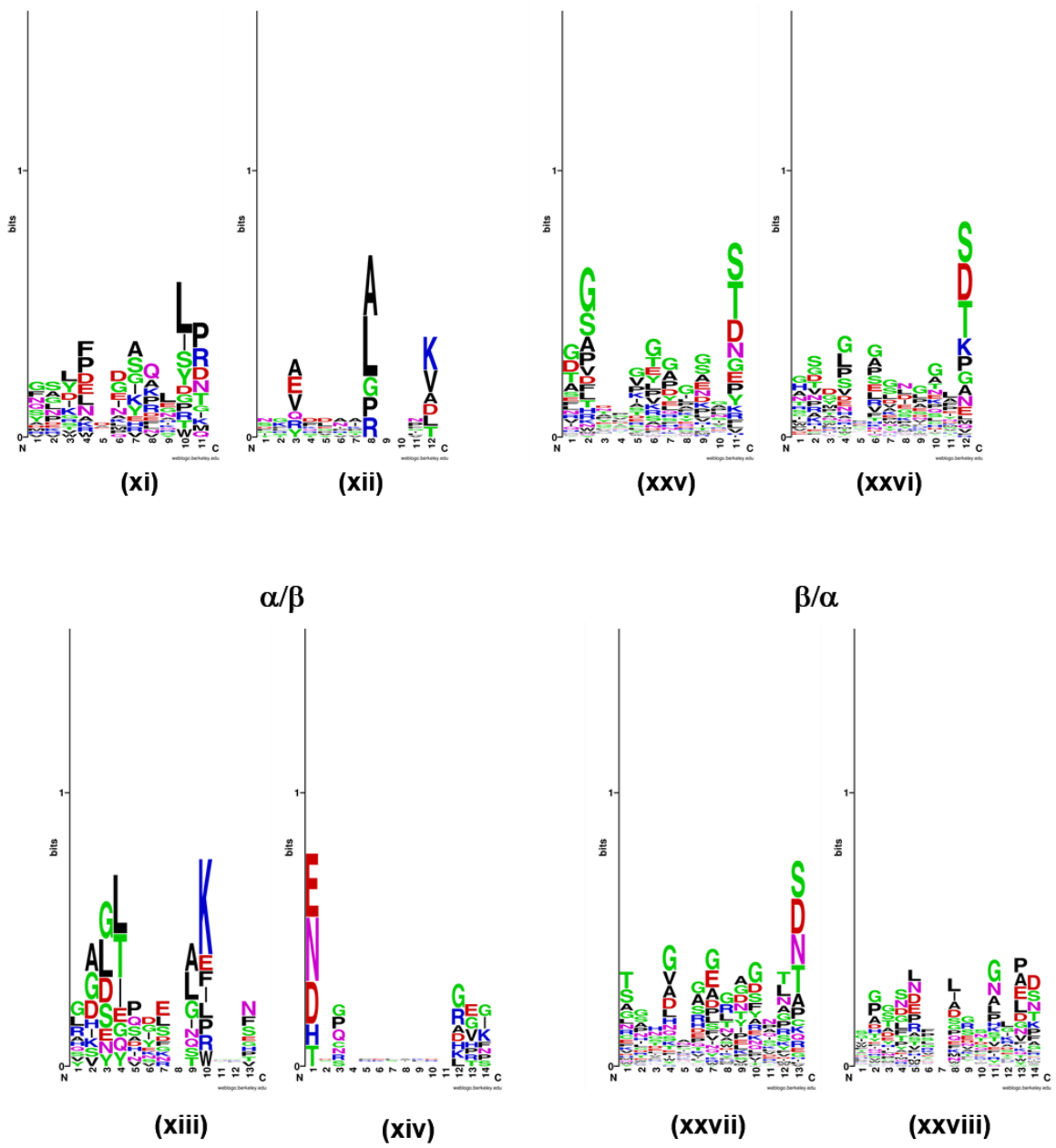


Figure 2.6: *Sequence profiles of αβ (i-xiv) and βα loops (x-xxviii) of length 1-14 residues showing the position specific preferences for amino acids. Stack represents conservation at the respective position and the height indicates the relative frequency of occurrence of the particular amino acid. Amino acids are represented as single letter abbreviations. Figure is generated using Weblogo (Crooks et al., 2004).*

Structural Features

Plots of the backbone dihedral angles, ϕ, ψ for the residues in short loops are shown in (**Figure 2.7**). The residues, mostly, glycines, in $\alpha\beta$ loops tend to occupy the left-handed α_L region of the Ramachandran ϕ, ψ map. This is consistent with the proposed role of the positive ϕ value in aiding the transition from helix to strand by changing the direction of the chain (Scheerlinck et al., 1992). Interestingly, the ϕ, ψ values of the last residue in the $\alpha\beta$ loops deviate from the trend of falling in the α_L region of the Ramachandran ϕ, ψ map. The distribution of the ϕ,ψ of the ultimate residue in the allowed regions will facilitate reentry of the chain into the immediate β-strand. Although, not the same extent, a small population of $\beta\alpha$ loop residues, especially from the longer loops, also fall in the α_L region of the Ramachandran ϕ, ψ map. This may be attributed due to the increased occurrence of type IV turns in $\beta\alpha$ loops and in some cases the residues adopting positive ϕ, ψ values in turn types such as inverse β turns, I' and II'.

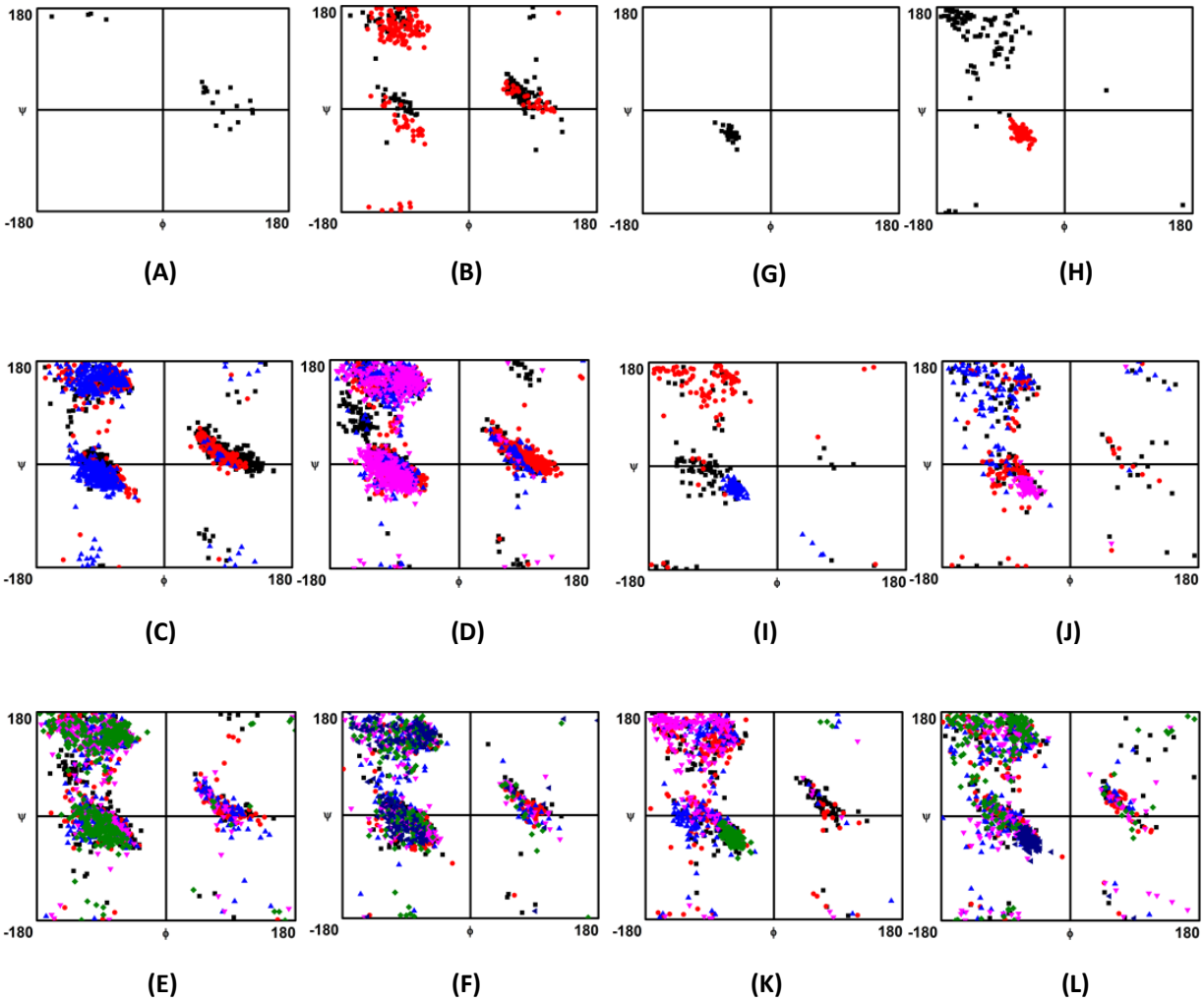


Figure 2.7: *Plots of the backbone dihedral angles, ϕ, ψ, in αβ and βα loops of length 1-6 residues. A-F correspond to the ϕ,ψ distribution of residues in the αβ loops. G-L correspond to the ϕ, ψ distribution of residues in βα loops. Color coding indicates residue position in the loops, black=residue1; red-residue2; blue=residue3; magenta=residue4; green= residue5; violet=residue6. Triangles correspond to the ϕ, ψ of glycines.*

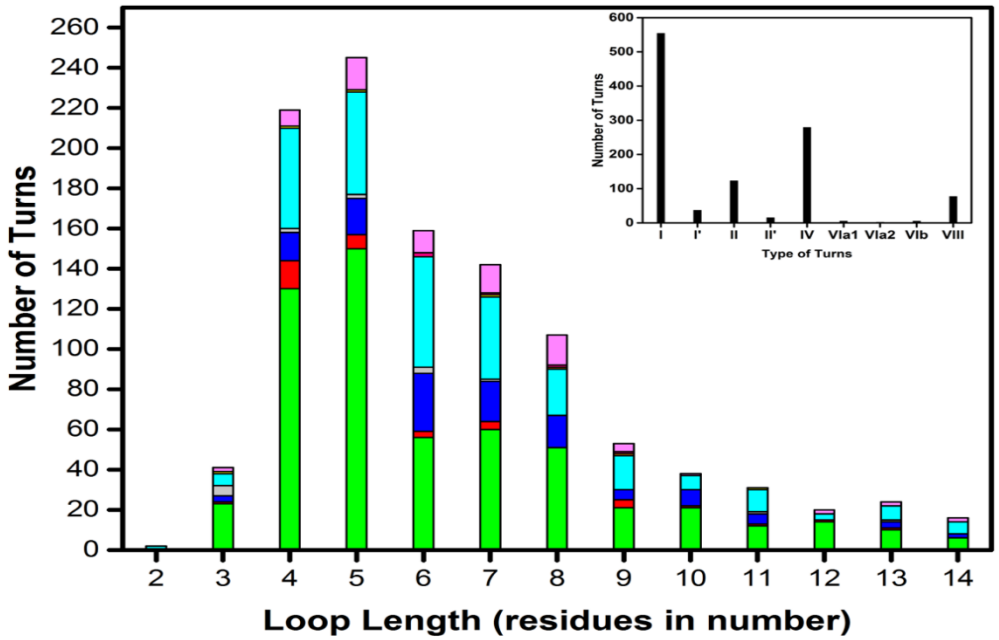
Distribution of turn types

Analysis of the backbone ϕ,ψ dihedral angles of the connecting residues in α helical and β hairpins in proteins indicated that they are not completely random and actually show a preference for specific turn conformations (Anderson et al., 2016; Guruprasad, Prasad, & Kumar, 2000; Guruprasad & Rajkumar, 2000). In order to examine the preferences for ordered

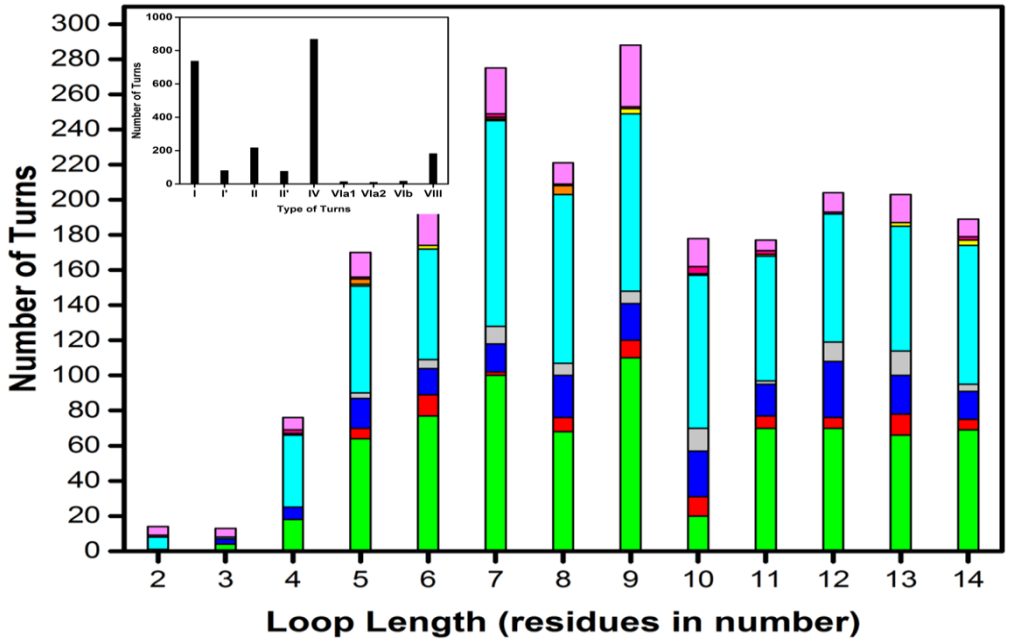
conformations, αβ and βα loops segments were examined for the occurrence of classical reverse turn conformations. The different types of β turns were extracted using PROMOTIF (Hutchinson & Thornton, 1996) data from PDBSUM database (de Beer et al., 2014). A total of 1097 and 2264 turns were identified from 1-14 residue long αβ and βα loops respectively. A majority of the turn conformations (60%) occur in short αβ loops of < 6 residues.

In general, it may be observed that type I and IV turns are highly populated over other turns in the loop segments (**Figure 2.8 A, B**). Type II and VIII turns occur to a lesser extent with a very low population of other turn types. The distribution of type I turns (60%) in short αβ loops is higher than type IV turns (30%). Whereas, in β/α loops, type IV turns are slightly higher. Type II and type VIII constitute roughly 10% with the other turn types occurring <3% in both types loops. We note that as the length of the βα loops increases, the distribution of the turn conformations increases significantly due to the presence of multiple and or consecutive turns in longer βα loops. This, in turn, may be leading to the over-estimation of turn types in βα loops. However, the overall preference for type IV turns is also reflected in short βα loops.

It is of interest to note that while type I' and II' turns are highly preferred in β-hairpin loops (Madan et al., 2014) the very low frequency of type I' and II' turns suggests that they are not favored in αβ/βα hairpins.



(A)



(B)

Figure 2.8: *Distribution of the turn types in 2-14 residue αβ (A) and βα loops (B). Colors indicate the respective turn type. cyan=type I, red=type I', blue=type II, magenta=type II', green=type IV, yellow=type VIa1, violet=type VIa2, orange=type VIb, brown=type VIII. The distribution of different turn types in αβ and βα loops are shown in the insets.*

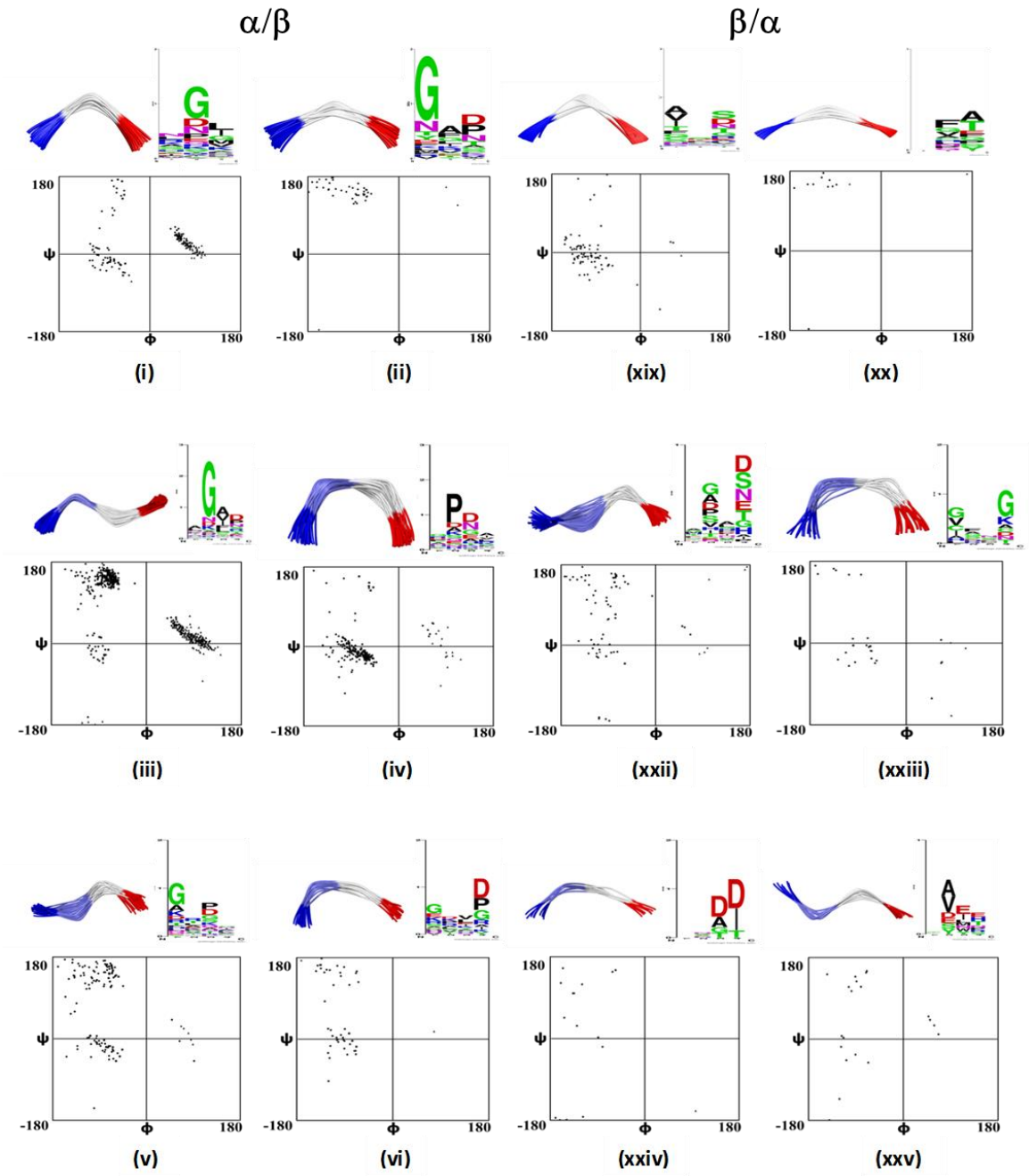
Sequence and Structural Diversity

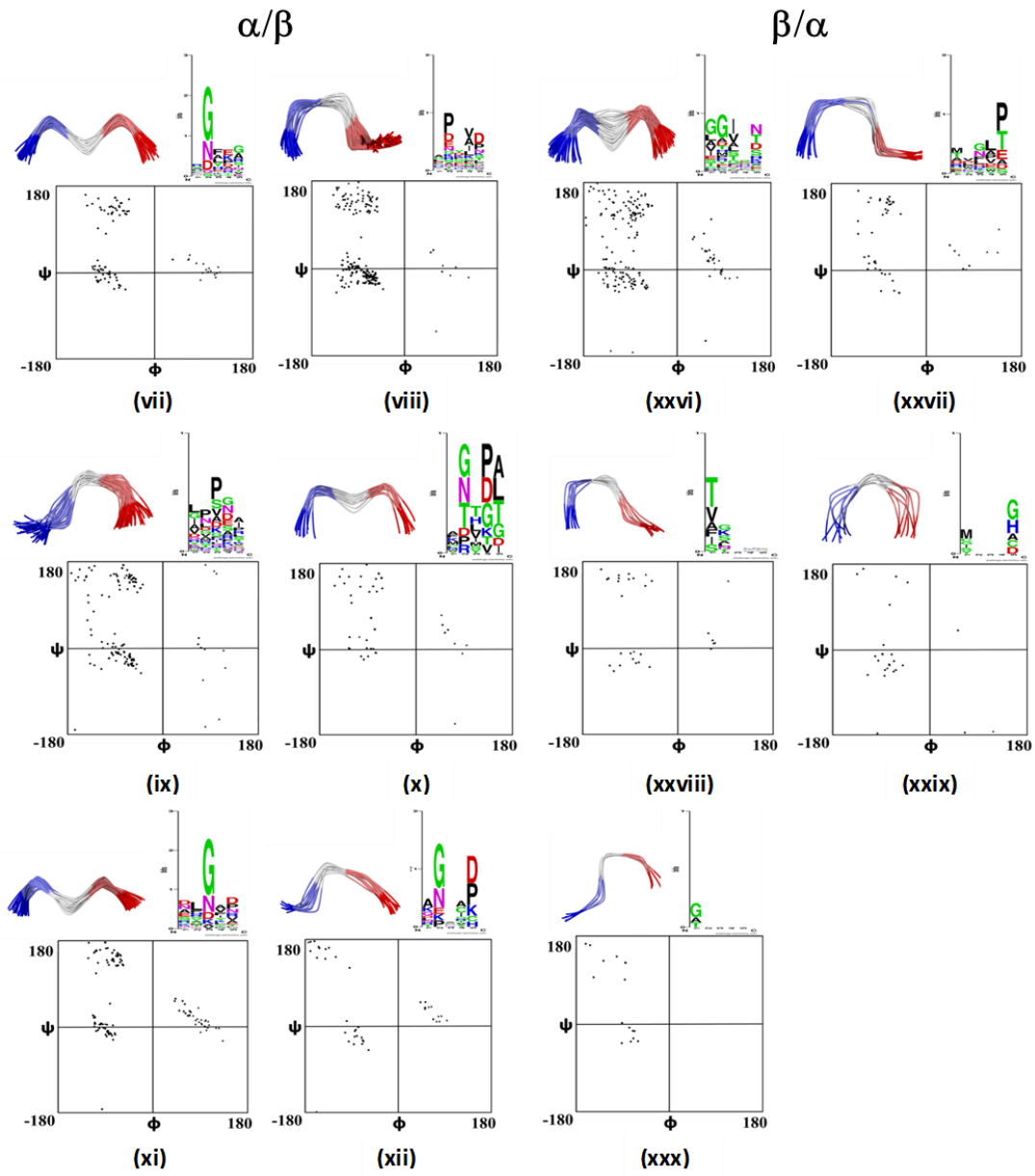
In order to identify similar conformations, short individual αβ and βα loops were clustered based on the root-mean-square deviations (RMSD) of the backbone atoms. Loops that could be clustered within RMSD of < 2 Å are shown in (**Figure 2.9**). Loops that did not fall in any cluster were omitted.

For any given loop length, clusters of different geometrical conformations are found and for each conformational cluster sequence conservation and ϕ, ψ predisposition are generated and shown in (**Figure 2.9**). It is interesting to observe cluster dependent sequence and conformational preferences, especially in 3-6 residue αβ loops. For instance, 3 residue αβ loops can be grouped into two major clusters, one with glycine preferentially occurring at position 1 (**Figure 2.9 ii**), and the other with a predisposition for glycines at position 2 (**Figure 2.9 i**). The backbone dihedral angles ϕ, ψ, for the cluster with glycines at second position tend to adopt α_L region.

Similarly, four residue loops can be grouped into four major conformational classes. One cluster shows a clear preference for glycine at position 2 with ϕ, ψ falling in the α_L region of the ϕ, ψ map (**Figure 2.9 iii**). For the cluster showing a preference for proline (P) and aspartate (D)/asparagine (N) at positions 2 and 3 respectively, the loop residue dihedral angles correspond to type I turn conformations (**Figure 2.9 iv**). The observation of the preference in four residue loops for proline and small polar amino acids, aspartate, asparagine is consistent with the observed residue preferences in type I turn conformations at second and third positions respectively (Guruprasad & Rajkumar, 2000). Further analysis may be required to explore the likely relation between the length and composition of the helices and strands with the

connecting loop length and geometry. Similar cluster specific sequence and corresponding dihedral angle preferences are observed in 5 and 6 residue loops.




α/β
β/α
180
Ψ
Φ
-180
180
(vii)
(viii)
(xxvi)
(xxvii)
(ix)
(x)
(xxviii)
(xxix)
(xi)
(xii)
(xxx)

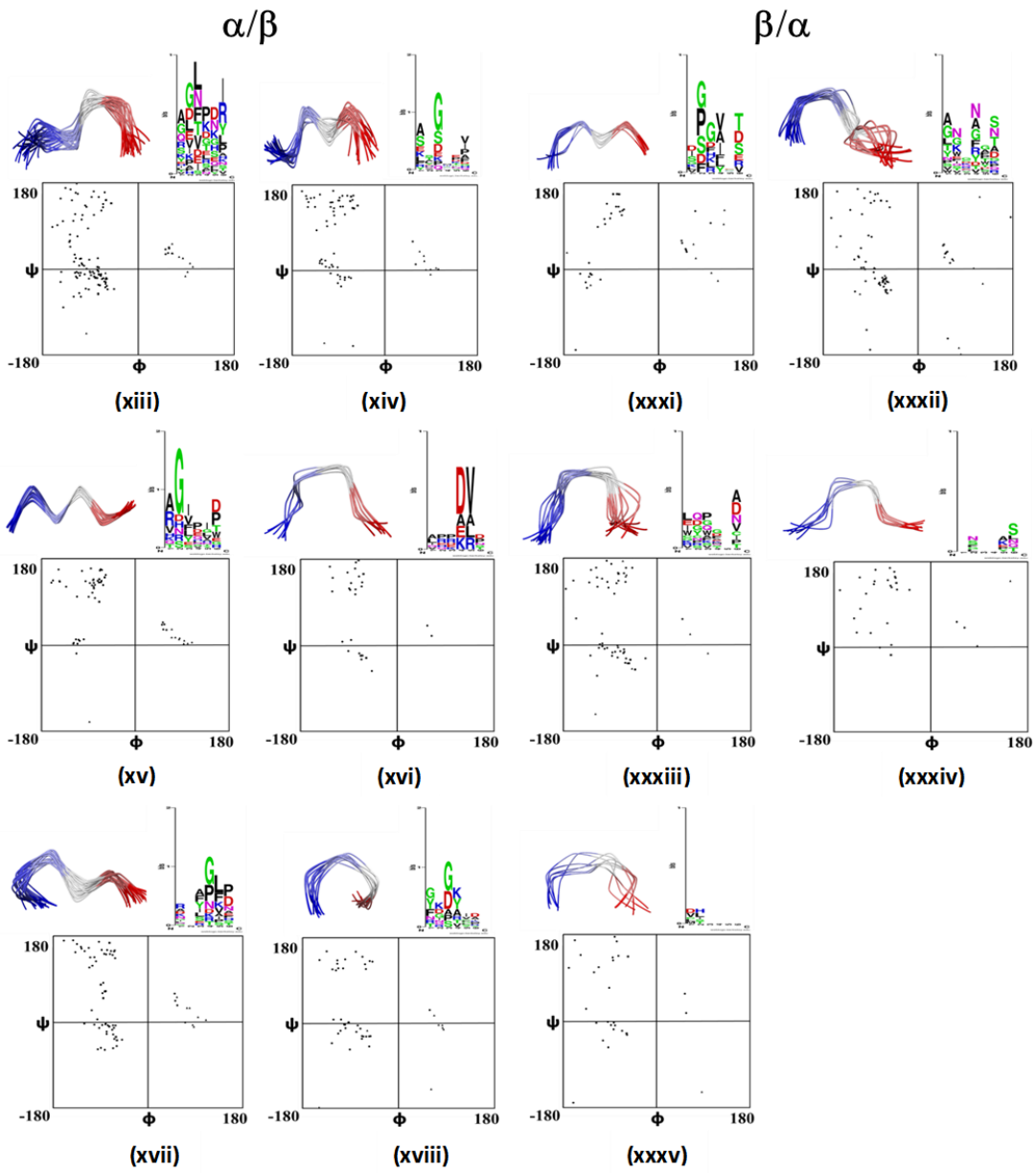


Figure 2.9: *Clusters of 3 residue αβ (i, ii) and βα (xix, xx) and 4 residue αβ (iii, iv, v, vi) and βα (xxii, xxiii, xxiv, xxv) and 5 residue αβ (vii, viii, ix, x, xi, xii) and βα (xxvi, xxvii, xxviii, xix, xxx) and 6 residue αβ (xiii, xiv, xv, xvi, xvii, xviii) and βα (xxxi, xxxii, xxxiii, xxxiv, xxxv) loops. Sequence profiles showing the position specific preferences along with the ϕ, ψ distribution for respective individual clusters are indicated.*

Hydrophobic clustering at the N-termini of the barrels connected by short loops

At this juncture, having established that αβ loops are short and favor constrained (turn) conformations, it becomes relevant to address the concomitant consequence(s). As a first step, we examined if the αβ loops promote interactions between the flanking helices and strands. **Figure 2.10 A**, reveals that the number of non-polar side chain interactions between the helices and strands connected by αβ loops is about two times higher than the non-polar interactions between the helices and strands connected by the βα loops. Further, it is interesting to note a gradual increase in the number of non-polar interactions from N to C termini of helices (**Figure 2.10 B, C**) indicating that αβ loops are more effective in promoting hydrophobic clustering towards the C-termini of the helices closer to the αβ loops. Assisted by favorable entropic considerations, the distance limitation imposed by short ordered αβ loops, in turn, can also lead to facilitating non-polar hydrophobic clusters as shown in **Figure 2.10 C**, leading to effective and increased reside-residue contacts, which make an important contribution to protein stability and folding (De Sancho & Munoz, 2011; Gromiha, 2001). Short constrained loops are implicated in reinforcing the interactions between the flanking secondary structural elements leading to increasing protein rigidity (Balasco et al., 2013; Vieille, Burdette, & Zeikus, 1996). Thus, predisposition for turn conformations observed in short αβ loops can make them more rigid and sampling limited conformational space, resulting in decreased conformational entropy. In fact, loop shortening as a means to enhance stability appears to be a general strategy adopted by proteins (Balasco et al., 2013; Nagi AD, 1997).

Overall, from the present analysis, it appears that interplay between the reduced conformational entropy, as a consequence of shortening and stiffness, can further promote enthalpic contributions from hydrophobic clustering of non-polar residues in αβ units connected by αβ

loops. Finally, from this analysis, it is clear that, at least in part, αβ and βα loops differ with respect to size, preference for turn conformations, residue composition (glycine, proline, arginine, and lysine). The differences in these properties may be responsible for their distinguished roles in stability and function. Future attempts may incorporate this knowledge, for designing proteins in general, and TIM barrels in particular.

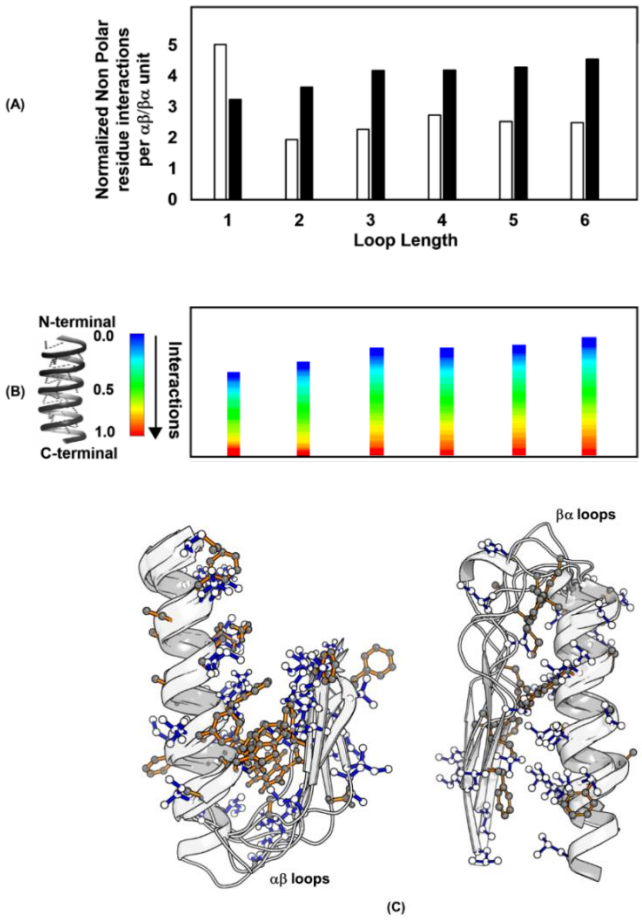


Figure 2.10: *(A). Hydrophobic clusters involving the residues of helices and strands for respective αβ (filled bars) and βα (open bars) loops length. Interactions to loop residues are also considered. (B). Extent of non-polar interactions along the length of the helices in αβ units of the TIM barrel proteins is illustrated using color coding. Blue to red indicates a gradual increase in the number of non-polar interactions along the length of the helix. The helical lengths (number of residues in helices) are normalized and designated as 0-1, representing the N-terminal middle and C-terminal residues of the helix respectively. (C). Superimposition of representative αβ and βα units from TIM barrel proteins. Non-polar isoleucine, valine, leucine residues are shown in blue, alanine, methionine, phenylalanine, tyrosine, tryptophan, proline residues are shown in orange.*

Conclusions

From this study, it is clear that preferences for size, conformation, and sequence differ between the αβ and βα loops of TIM barrel proteins. The distinctive role of αβ and βα loops in stability and function, respectively, perhaps is reflecting in their size, sequence profiles and conformations. The αβ loops are dominated by smaller loops in contrast to the longer loops in βα loops. It is expected that the longer βα loops can have a higher degree of freedom to adopt random conformations assisted by multiple turns, required for geometrical flexibility.

The presence of ordered conformations, in general, and type I turns, in particular, could be the contributing factors for the rigidity of αβ loops. The other likely factor that may also endow rigidity to the αβ loops is the higher proportion of long range side chain hydrogen bonding interactions involving the two positively charged residues, arginine and lysine (**Figure 2.4**)

Glycine residues appear to play an important role in adopting conformations that will result in tight turns leading to optimal packing between the helices and strands of the αβ hairpins. Moreover, lacking side chain, it can maximize the interactions between the helix and strand. A high preference for proline residues ensures reduced backbone conformational entropy in restricting the αβ loops flexibility. However, the observations are pertained to the loops with in the protein structures and may not be correlated with the study of TIM barrel loops in isolation, due to unavailability of experimental study.

In summary, the present study will help to facilitate short-listing of the candidate loops to be exchanged in place of a target βα loop. From our compiled information, loops of similar size with varied sequence composition can be identified, superimposed, and assessed for exchange potential using computational means. Apart from assisting engineering TIM barrel proteins, the

compilation of the αβ and βα loops shed light on the distribution of turn types and preferred template sequences, particularly, in short loops connecting the α and β elements in αβ/βα hairpins. This may encourage experimentalists in designing and or identifying sequences that can adopt α/β structures in solution. As per the findings in this study, four residue loops adopting type I turn conformation may be best suited for connecting the α helix to β strand in αβ hairpins.

Chapter 3

Structural and molecular dynamics analysis of the α subunit of tryptophan synthase provide clues for the role of αβ loops in the stability of the TIM Barrel fold

Introduction

Tertiary interactions lead to optimal packing and these long range contacts are vital for imparting stability to the final folded states of proteins (Chothia & Finkelstein, 1990). While the protein amino acid sequence determines the overall fold, the non-covalent interactions from within the sequence contribute to the stability in the final folded state. Experimental data clearly indicate that the non-covalent interactions arising from the inter residue proximities in the tertiary structures, including, ion-pair interactions, hydrogen and hydrophobic interactions, amino acid composition etc., contribute to protein stability (Baase, Liu, Tronrud, & Matthews, 2010; Gassner et al., 2003; Magyar, Gromiha, Savoly, & Simon, 2016; B. W. Matthews et al., 1987; Mooers, Baase, Wray, & Matthews, 2009; Szilagyi & Zavodszky, 2000; Tompa, Gromiha, & Saraboji, 2016; Vieille et al., 1996; Vieille & Zeikus, 2001; J. Xu, Baase, Baldwin, & Matthews, 1998; X. X. Zhou, Wang, Pan, & Li, 2008). The subtle balance between these interactions not only confers stability but also is responsible for flexibility and function. However, while, it is well established that loops facilitate function, their role in stability and folding is not very clear. Experimental demonstration of the sampling of turn conformations in solution by some small isolated turn sequences from proteins suggest that if not all, at least, certain turns in proteins play active role as folding initiation sites (Lewandowska et al., 2010; Marcelino & Gierasch, 2008; Ramakrishna & Sasidhar, 1997; Wright et al., 1988). In contrast, the turns appearing late in the process of structure acquisition are passive and may contribute

to the overall fold and stability. Although such studies have shed light on the relationship between loop sequences and turn forming tendency, the correlation between the loops and stability is not very clear. However, experimental data is providing information on the role of loop flexibility, reduced length and the interactions from within the loops in contributing to the overall stability of protein structures.

The role of the flexibility of loop conformations has been addressed in some cases. An enhanced propensity for ordered turn conformations in the loop regions play an important role in the overall structural stability. In fact, it was experimentally demonstrated that enhancing and or optimizing loop regions for increased turn propensity resulted in positive modulation of protein stability. Thus, these observations highlight the role of decreased loop flexibility in stability (Balasco et al., 2013; Nagi AD, 1997; Predki et al., 1996; Simpson et al., 2005). Interestingly, these studies hint at the plausibility of evolutionary requirement to select for turn sequences that confer thermodynamic stability (H. X. Zhou et al., 1996) and such turns can modulate the stability by their intrinsic preference to sample favorable ϕ, φ space (Anderson et al., 2016; Fu et al., 2009; Predki et al., 1996; Simpson et al., 2005; Trevino et al., 2007).

In some other instances, experimental data from representative proteins suggested that decreased loop length contributes to the protein stability (Balasco et al., 2013; Nagi AD, 1997; Predki et al., 1996; Simpson et al., 2005). Experimental data demonstrated that trimming of loops lead to an enhancement of stability, implying that loop length is one of the important determinants of protein stability (Collinet, Garcia, Minard, & Desmadril, 2001; Nagi AD, 1997). In fact, the conserved pattern of configuration of short loop connections in β hairpins has indicated that loop curtailment is an adopted strategy by thermophilic RNaseH and thioredoxin for improved stability over their mesophilic counterparts (Balasco et al., 2013).

Modeling of the effects of loop truncations revealed that increased folded state entropy can contribute significantly to stability (Gavrilov et al., 2015; H. X. Zhou, 2004). Thus it appears conclusive that as the length of the loop decreases stability enhances.

In addition to flexibility and reduced loop length, the sequence composition of loops can also greatly influence protein stability. Point mutations within the loops and in some cases loop replacements have been shown to substantially effect protein stability (Gekko, Kunori, Takeuchi, Ichihara, & Kodama, 1994; Hoedemaeker, van Eijsden, Diaz, de Pater, & Kijne, 1993; Parge, Hallewell, & Tainer, 1992). Close and optimal packing of secondary structures is critical for proper folding and stability and the interactions from within the loops can encourage stabilizing interactions between secondary structures (Chothia & Finkelstein, 1990; Ptitsyn & Finkelstein, 1989). For instance, in the case of four-helix bundle proteins, a significant contribution towards their stability arises from loop-helix interactions rather than from helix-helix interactions. It is also well known that hydrophobic clustering of residues from the adjacent strands and loops in β hairpins favors their formation (Colombo et al., 2003). The non-repetitive loops/reverse turn geometries connecting secondary structural elements provide compactness by bringing distant regions closer in space, (Anderson et al., 2016; Colombo et al., 2003; Dyson & Wright, 1991; Lewandowska et al., 2010; Munoz et al., 1997; Ramirez-Alvarado et al., 1997; Richardson, 1981) thus, reducing flexibility and promoting stabilizing interactions. Thus, despite their less organized conformation, loops can play not only a contributing role in stability but also in guiding the folding process.

Interestingly, in the context of the role of loops/turns in protein stability and function, TIM barrel architecture, imposes an inherent division of roles for its loops in stability and function. The fold is formed by the repetition of the basic βαβ motif in which the β-strands are followed

by α-helices alternating in sequence and structure. Basically, in this arrangement, two successive β-strands parallel to each other are joined by a α-helix via the loops. The loops connecting the β-strands to α-helices are referred to as βα loops while αβ loops correspond to the loops that connect the α-helices to the β-strands. The active sites are invariably comprised of the loops protruding from the C-termini of the β-strands contributing to the function of all TIM barrel enzymes. In contrast, the αβ loops at the N-terminal end of the barrels joining the C-termini of the α-helices and the N-termini of β-strands have been linked to stability (Sterner & Hocker, 2005; Urfer & Kirschner, 1992; Wierenga, 2001). Therefore, the $(\beta/\alpha)_8$ TIM barrel fold by providing an inherent distribution of responsibility of stability and function to loops can serve as a model to decipher the interactions that confer αβ loops role in the stability of the fold.

The alpha subunit of tryptophan synthase from *Escherichia coli* (αTS), a 29 kDa protein adopting TIM barrel fold, serves as an excellent model system for proposed study (**Figure 3.1 A, B**). Its structure, folding, dynamics and stability have been thoroughly investigated using a variety of experimental approaches (Finke & Onuchic, 2005; Gualfetti, Bilsel, & Matthews, 1999; Vadrevu, Falzone, & Matthews, 2003; Vadrevu et al., 2008; Wu, Vadrevu, Kathuria, Yang, & Matthews, 2007). Under equilibrium and kinetic folding and unfolding reactions of αTS partially folded intermediates are significantly populated (Finke & Onuchic, 2005; Gualfetti et al., 1999; Vadrevu et al., 2008; Wu et al., 2007). Further, measurements using NMR and MS approaches indicate that the N-terminal half of the barrel comprising of $(\beta\alpha)_{1\text{-}4}$ units is more stable than the C-terminal half of the barrel $(\beta\alpha)_{5\text{-}8}$ units. Native state hydrogen-deuterium exchange coupled with NMR spectroscopy identified that the most stable region is

comprised of α helix 1, β strand 2 and β strand 3. Interestingly a couple of amides contributed by the residues located in strands β_2 and β_3 are involved in long-range hydrogen bonding interactions between their main chain amide hydrogens and polar side chain acceptors that staple consecutive βα/αβ elements. Contributed by the residues located in the loop regions these non-local have been found to be contributing significantly to the stability of αTS.

Therefore, the barrel architecture provides a unique opportunity for attempting to address if properties such as loop flexibility, inter-residue interactions including hydrogen bonds, electrostatic and hydrophobic contacts, specifically, from within the loops with their flanking secondary structural elements are inherent to and dominating in αβ loops and therefore the reason for their role in the fold stability.

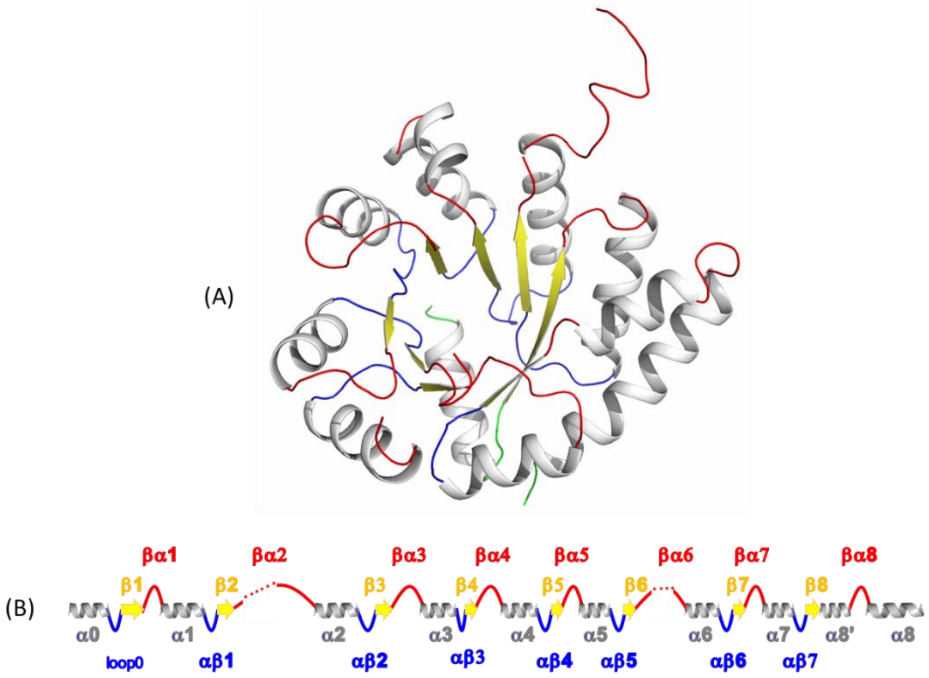


***Figure 3.1: A**. Crystal structure of the alpha subunit of tryptophan synthase from E.coli (PDB ID: 1v7y). **B**. Topology diagram of αTS protein structure showing β-strands and α-helices respectively connected by βα and αβ loops. The dotted regions indicate the missing structural information in PDB crystal structure.*

Nuclear Magnetic Resonance (NMR) spectroscopic and Molecular dynamics (MD) simulations techniques are being used as powerful techniques to establish fluctuations at the atomic level to pinpoint the flexible and rigid regions in proteins. The ^{15}N spin-lattice, spin-spin (T_1, T_2) relaxation measurements and [^{15}N,^{1}H] NOE measurements provide information about the fast and slow relaxation decay and tumbling of backbone N-H bond vectors for the individual residue of a protein that occurs in pico to nanosecond time scale.

Molecular dynamics (MD) simulations serve as a complimentary approach to NMR methods as a powerful technique to establish fluctuations at the atomic level by running the simulations for several lengths of time in nanoseconds to pinpoint the flexible and rigid regions in proteins. The temperature (B) factors derived from the crystallographic structures together with the average root mean square fluctuations (RMSF) obtained from the time-dependent unfolding simulations serve as direct qualitative measures for the assessment of stability. The relative RMS fluctuations along the protein backbone atoms can differentiate the regions sensitive to denaturation and the observed flexibility can be directly related to the stability in different parts of the protein (Khan, Farooq, & Kurnikova, 2016). Therefore, we have attempted NMR and MD simulations based analysis to address the role of flexibility and dynamics inherent to αβ and βα loops of αTS.

Materials and Methods

NMR Sample Preparation

Two samples of αTS were prepared for assignments process: a uniformly ^{13}C,^{15}N-labeled sample (control) and a sample with multiple amino acid types unlabeled simultaneously. The protocol for preparing the two samples for a given protein is similar except for the addition of unlabeled amino acids in the selectively unlabeled sample. Uniformly, ^{15}N, ^{13}C/^{15}N labeled αTS was expressed in *E.coli* BL21/DE3 cells transformed with plasmid pT7.WS1 (encoding αTS). Cells were grown at 37°C on M9 medium with [^{15}N] NH_4Cl (1g/L) and [^{13}C] glucose (2g/L).

Protein expression was induced by isopropyl β-D-thiogalactopyranosidase (IPTG) and protein was harvested after 4h. The resulting pellet was solubilized in potassium phosphate buffer (100mM, pH7.80) containing Ethylene diamine tetra acetic acid dipotassium salt (K_2EDTA, 5mM) and dithioerythritol (DTE, 2mM). After sonication, cell lysate was incubated for few minutes in potassium phosphate (100mM, pH7.8) containing K_2EDTA (5mM), DTE (2mM), $MgCl_2$ (10mM), NaCl (0.1M), Triton X-100 (0.01%), RNase A (10µg/mL), and DNase (10µg/mL). The protein was purified as described earlier (Vadrevu et al., 2003). A sample of αTS was also prepared: selectively unlabeled at Arg, Asn, Thr, Gly, Ser, and Ala (comprising ~38^% of total residues) by expressing the protein in *E.coli* BL21/DE3 cells as described earlier. Cells were grown in M9 medium containing [^{15}N] NH_4Cl with the desired unlabelled amino acid(s) (1.0g/L). For the ^{15}N spin-lattice/spin-spin relaxation (T_1/T_2) and HETNOE experiment [^{15}N,^{1}H] NOE), uniformly labeled ^{15}N αTS sample was prepared following the same protocol as described previously. The final NMR samples contained protein (~1mM) in potassium phosphate buffer (50mM, pH7.8) in $H_2O/^2H_2O$ (95:5).

Design and implementation of 2D NMR experiments for rapid NMR assignments

For the rapid NMR assignments process three 2D NMR experiments were recorded, one was regular 2D [^{15}N,^{1}H] HSQC; another was the 2D [^{15}N,^{1}H] projection of a 3D HNCO (denoted 2D HN(CO)) used routinely for protein resonance assignments, and the third was 2D HN-XU (Dubey, Kadumuri, Jaipuria, Vadrevu, & Atreya, 2016). In 2D HN(CO) peaks corresponding to both the unlabeled residue *i* and of residue *i*+1 are absent. Once the ^{15}N, ^{1}H shift correlations of the selectively unlabeled residues (*i*) are identified based on 2D ^{15}N, ^{1}H HSQC, those corresponding to *i*+1 were then identified.

^{15}N spin-lattice/spin-spin relaxation (T_1/T_2) and [^{15}N,^{1}H] Overhauser measurements

The [^{15}N,^{1}H] Heteronuclear Single Quantum Coherence (HSQC) spectra of ^{15}N spin-lattice/spin-spin and [^{15}N,^{1}H] Overhauser relaxation measurements were recorded on 800MHz spectrophotometer using Bruker Topspin software. The ^{15}N spin-spin spectra were recorded

with time delays of 10, 20, 30, 40, 50, 60, 70 and 90ms, and 100, 200, 300, 400, 500, 600, 700, 800 and 900ms time delays for the ^{15}N spin-lattice measurement. The pulse sequence used for the T_1 and T_2 measurements are hsqcct1etf3gpsi and hsqcct2etf3gpsi with 2048 and 440 complex data points along the ^{1}H and ^{15}N dimensions with 4 scans per measurement.

To measure the [^{15}N,^{1}H] Overhauser data, two different spectra were recorded without and with proton saturation using hsqcnoef3gpsi pulse program with 512 and 2048 complex data points along the ^{1}H and ^{15}N dimensions with 24 scans per measurement using Bruker Topspin software.

Data collection/analysis

All NMR measurements were recorded at 25 °C on a Bruker Avance 800 MHz spectrometer equipped with a cryogenic probe. Data were processed with NMRPipe (Delaglio et al., 1995) and analyzed using XEASY (Bartels, Xia, Billeter, Güntert, & Wüthrich, 1995) or SPARKY (Goddard & Kneller, 2006).

Molecular Dynamics simulations

The starting crystal structure of tryptophan synthase alpha subunit (PDB ID: 1v7y) was obtained from the Protein Data Bank (PDB).(H.M. Berman et al., 2000) The crystal structure contains two missing β/α loops hence, modeling and refinement of the missing loops were carried out using MODELLER software (Eswar et al., 2006; Fiser et al., 2000; Sali & Blundell, 1993), the best model has been selected based on Dope score and submitted to ERRAT (Colovos & Yeates, 1993) to estimate the overall quality factor of the structure, it was observed that the modelled αTS structure has an overall quality of 98.4% whereas the initial PDB crystal structure has an overall quality factor of 96% with a Root Mean Square Deviation (RMSD) of

0.51 Angstroms(Å). The simulations were performed at 300, 400 and 500 K considering the observation that, typically, the unfolding of proteins occurs in the microsecond time scales (Duan, Wang, & Kollman, 1998). The difficulty of performing MD simulations for long timescales can be circumvented at higher temperatures as the unfolding is accelerated at elevated temperatures. It may be noted that the elevated simulation temperatures do not influence the unfolding pathways. (Chen et al., 2016; Daggett & Levitt, 1993). The chemical unfolding simulations were performed in 8M urea at 300 and 400 K simulated temperatures. The all-atom Molecular dynamics simulations were carried out using GROMACS 5.1 molecular dynamics simulation package for all the simulations the all-atom Gromos43a1 force field has been selected (Abraham & and Lindahl, 2015; Van Der Spoel et al., 2005; van Gunsteren, 1996). The protein was solvated in a cubic box with a distance cutoff 12 Å between the edge of the periodic box and surface of the protein with water and urea-water mixed molecules, respectively. The SPC (Simple Point-Charge) model was used for solvation, and this model has been successfully used for variety of thermal unfolding studies (Chen et al., 2016; Jiang, Chen, & Wang, 2016; Kundu & Roy, 2008; J. Li, Chen, Yang, & Hua, 2015; Paul, Hazra, Barman, & Hazra, 2014). In the mixed solvent system 3200 urea and 12085 of water molecules are included and Na^+ counter ions were added to neutralize the system, the urea topology files are prepared using PRODRG server (Schuttelkopf & van Aalten, 2004) for mixed solvent simulations. The system was subjected to energy minimization using steepest-descent algorithm down to a 1000 kJ/mol/nm till the energy get converged. Before production run all systems were equilibrated for temperature and pressure equilibration by position restraining the protein for 100ps using canonical NVT and NPT ensembles at respective temperatures and urea-water mixed solvents. The long-range electrostatic interactions are calculated using

Particle Mesh Ewald (PME) (Darden, York, & Pedersen, 1993; Essmann et al., 1995) method with a grid spacing of 0.16 nm and a cut-off of 1.0 nm was used for short-range electrostatic and van der Waals interactions, bond lengths were constrained using LINCS algorithm (Hess, 1997). All simulations were performed using a 2-fs integration time step, with a coupling coefficient of tT – 0.1 ps using modified Berendsen thermostat (Berendsen, Postma, Gunsteren, DiNola, & Haak, 1984), and Parrinello-Rahman pressure-coupling at 1bar with a coupling coefficient of tP = 1 ps.

To investigate the flexibility, stability and unfolding characteristics of the αTS the simulations were carried out at 300, 400 and 500 Kelvin (K) simulated temperatures for 25 ns to estimate the thermal unfolding and for chemical unfolding the simulations were carried out in 8M urea at 300 and 400 K simulated temperatures. All the results were analyzed using Gromacs 5.1 package and VMD timeline analysis tools (Humphrey, Dalke, & Schulten, 1996).

Analyzing the trajectory files

The Cα Root Mean Square Deviation (RMSD) values for respective timescales are calculated using gromacs inbuilt analysis tools. The Root Mean Square Fluctuations (RMSF) are also computed for the equilibrated trajectory timescale excluding the first 1ns and averaged for individual helix, strands, α/β and β/α loops secondary structure units. The long range side chain to main chain interactions and secondary structure was analyzed using VMD timeline tools. The simulation snapshots for converted trajectory files at required timescales were made using PyMOL (Schrodinger, 2015).

Results and Discussion

NMR assignments of αTS

Analysis of protein local flexibility that occurs in pico and nano second level has been possible using Nuclear Magnetic Resonance measurements including T_1, T_2 and heteronuclear NOE (Renner, Schleicher, Moroder, & Holak, 2002; Sahu, Bhuyan, Udgaonkar, & Hosur, 2000). The availability of the backbone NMR assignments of αTS, 29 kDa, TIM barrel protein prompted us to undertake this approach. However, only ~65% of the [^{15}N,^{1}H] HSQC spectrum was assigned. Peaks especially in the loops and at the termini of α helices. Given the importance of loops and terminal residues in the overall flexibility and dynamics, it was necessary to revisit achieving an increase in the number of the NMR assignments. Therefore, as a first step towards the goal of measuring the flexibility of the residues in αTS, we have attempted to assign the cross peaks that were otherwise unassigned in a previous attempt (Vadrevu et al., 2003). In order to correlate the backbone dynamics of αβ and βα loops using ^{15}N spin-lattice/spin-spin relaxation (T_1/T_2) and [^{15}N,^{1}H] NOE measurements were performed on tryptophan synthase alpha subunit (αTS). In an earlier study, a combination of selective labeling at Ile, Leu, Phe, and Val and analysis of 3D triple resonance spectra was used for sequential backbone resonance assignment (Vadrevu et al., 2003).

In the present study, a new combinatorial selective unlabeling strategy was used to achieve rapid NMR cross peak assignments. The strategy helps in identification of tripeptide sequence around the unlabeled residues and the labeled residues are assigned sequence specifically. The combinatorial selective unlabeling approach is a three steps process. In step 1, two αTS samples were made, one ^{13}C,^{15}N-labeled sample with simultaneous unlabeling of optimally chosen amino acids, and a control sample which is uniformly ^{13}C, ^{15}N-labeled. One 2D [^{15}N,^{1}H] HSQC and two 2D triple resonance experiments 2D HN(CO) and 2D HN-XU were performed on the unlabeled sample, 2D HN-XU is a new experiment which provides information of ^{15}N,^{1}HN

correlations of unlabeled C-terminal residue (U') of labeled amino acids (X). For the control sample, a 2D [^{15}N,1H] HSQC and 3D HNCACB/3D CBCA(CO)NH spectra are recorded for identification of amino acids. The assignment process will be initiated based on the labeled amino acids spin systems (denoted as X, Y and Z as shown in **Figure 3.2**). In the given protein the UXU', UXY, ZXU' and ZXY combinations are considered for sequential assignment process (Dubey et al., 2016). The amino acid selection for the protein of interest is based on $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ to distinguish different amino acids and the abundance and scrambling of the amino acids.

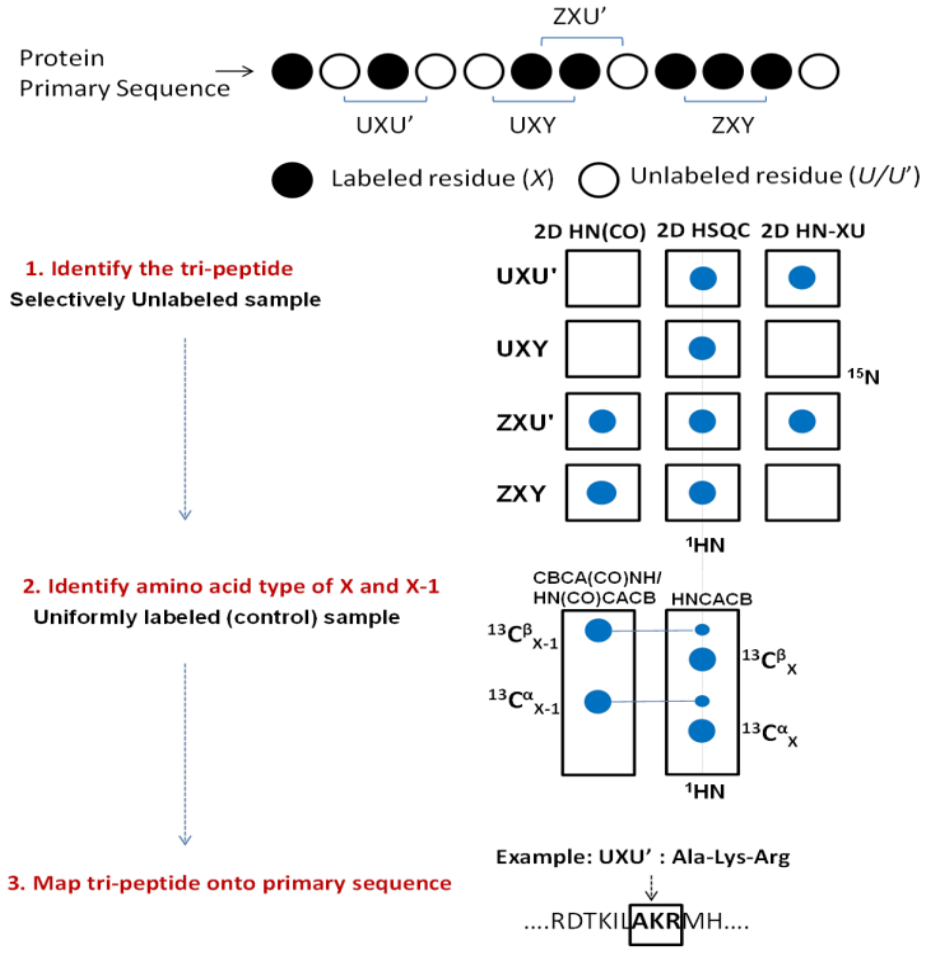


***Figure 3.2**. A schematic illustration of the assignment strategy (Dubey et al., 2016).*

The optimized set of amino acids selected for αTS are Ala, Arg, Asn, Gly, Ser, Thr which constitutes ~37% (99 residues) of the 268 residues, and these are distributed throughout the sequence. There are 38 UXY, 39 ZXU, 20 UXU', and 71 ZXY tripeptides in αTS (**Figure3.3**) effectively, excluding segments including at least one proline residue (i.e., tripeptides where X corresponds to proline), 149 tripeptides arise from the four possible categories. Of these, about 102 are unique, and this results in 102 labeled residues (X) that can be directly assigned by this methodology. This corresponds to 38% of total residues assignable in the protein.

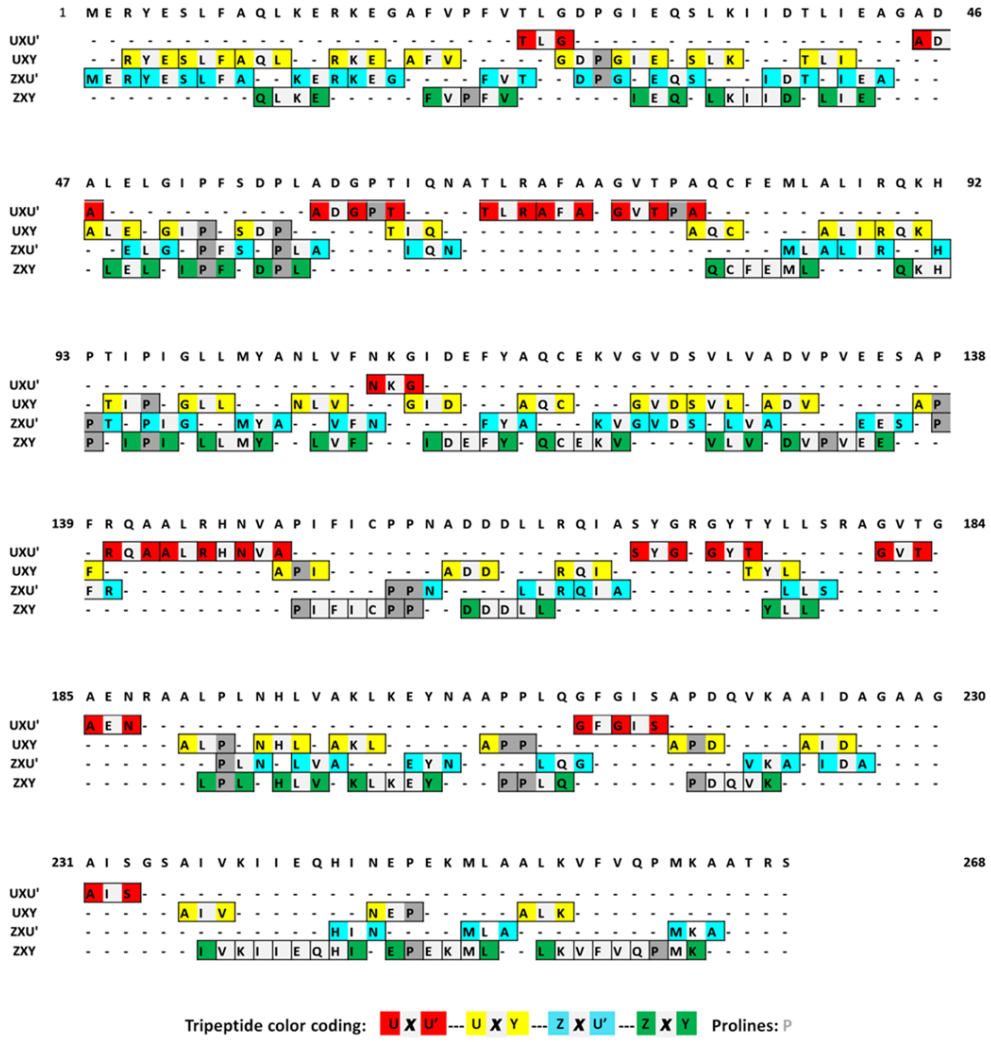
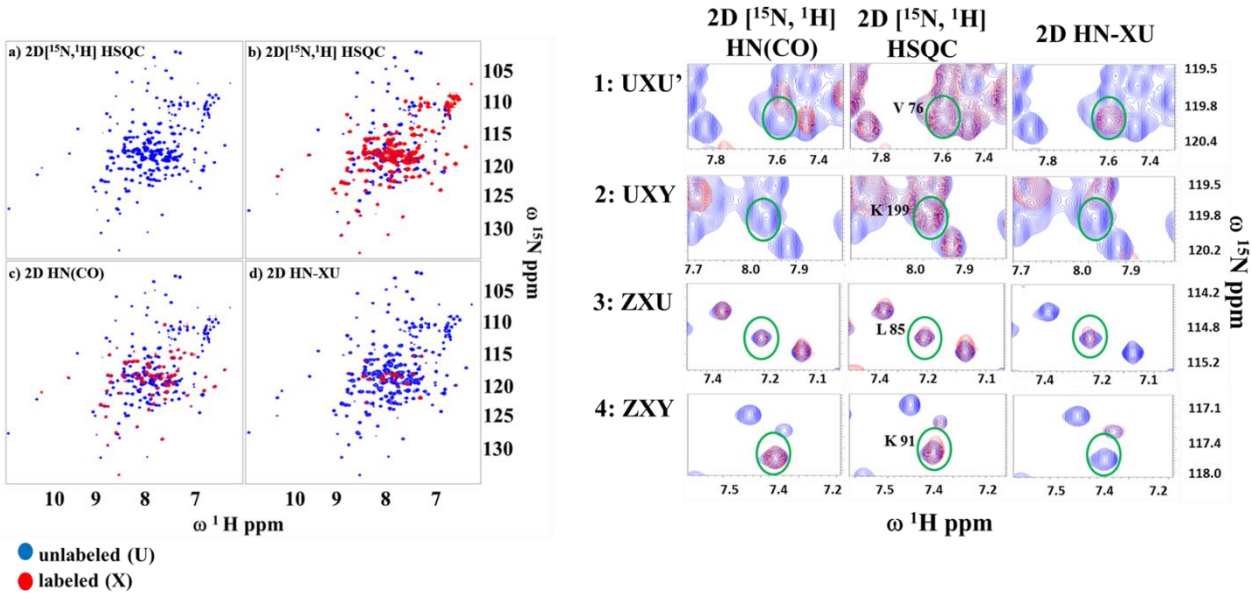


Figure 3.3*: The distribution of different tri-peptides in αTS protein across its sequence. The tri-peptides are highlighted in different colors*

Figure 3.4 a and **b** shows the three 2D spectra acquired for selectively unlabeled αTS overlaid with the 2D [^{15}N,^{1}H] HSQC spectrum of the uniformly labeled sample. Firstly, from the overlay of the HSQCs of the two samples (uniformly labeled and selectively unlabeled), 61 ^{1}H,^{15}N cross peaks were identified as belonging to one of the selectively unlabeled amino acids. The inability to find all the expected 99 ^{1}H,^{15}N cross peaks (corresponding to the sum of unlabeled amino acids) is presumable due to exchange broadening at pH 7.80 and/or spectral overlap.

In next step, tripeptide sequences including Ala, Arg, Asn, Gly, Ser, and Thr were identified (**Figure 3.3**). Starting with these residues and combining the pattern of correlations arising from the four possible categories (**Figure 3.2**), the four possible tripeptides patterns (UXY, ZXU, UXU', and ZXY) were identified and mapped onto the primary sequence of αTS for sequence-specific assignment. From the 149 triplets arising from the sequences, 66% of the X residues could be assigned by combining the 3D HNCACB data on the uniformly labeled αTS and the 2D spectra recorded on αTS (**Figure 3.2**). Twenty-four residues (~11% if prolines are excluded) that were unassigned in the earlier study (Vadrevu et al., 2003) could now be assigned from the observed *i*-1 and *i*+1 correlations, for example, the newly identified ^{75}GVT^{77}, ^{84}MLA^{86}, ^{198}AKL^{200}, and ^{90}EKH^{92} tripeptides (**Figure 3.4**). Taken together, the earlier assignments were augmented to result in 76% assignment.



(e)

Figure 3.4*: a. The 2D* [^{15}N,^{1}H] *HSQC spectrum of the control (uniformly) labeled sample of 29 kDa α-TS. (b)-(d) Overlay of the three 2D spectra acquired for the Ala, Arg, Asn, Gly, Ser and Thr selectively unlabeled sample of 29 kDa α-TS on the 2D ^{15}N-^{1}H HSQC spectrum of the control sample shown in (a). (e) An expanded portion of the three 2D spectra illustrating the peak patterns observed the four different tri-peptide categories.*

It is noteworthy that, for large proteins such as αTS, in addition to the challenges associated with size, reduced dispersion etc., sequence redundancy at the level of di- and tripeptides can be another limitation for speedy assignment. αTS contains about 70 and 20 redundant di- and tripeptide sequences, respectively.

^{15}N spin-lattice/spin-spin relaxation (T_1/T_2) and [^{15}N,^{1}H] Overhauser measurements on αTS

As mentioned the dynamics of protein backbone atoms can be measured using the steady state heteronuclear [^{15}N,^{1}H] Overhauser and ^{15}N spin-lattice/spin-spin (T_1/T_2) relaxation process and we have used these experiments to measure and understand the backbone N-H bond vector motion of residues in αTS. The measurements were carried out at 25^{o}C at a magnetic field of 800 MHz. The ^{15}N spin-lattice and spin-spin relaxation (T_1, T_2) data obtained for various delay times and the [^{15}N,^{1}H] NOE backbone amide resonance values are plotted against

the residue numbers (**Figure 3.6 and Figure 3.5**). As is observed from **Figure 3.5** and **Figure 3.6**, [^{15}N,^{1}H] NOE relaxation measurement values obtained are observed to be more than one which is unreliable for assessing the motion of backbone N-H vectors. This applies to the unreliable/inconsistent T_1, T_2 values throughout the sequence. One potential issue for this unreliable values could be arising from the fact that αTS is a relatively large protein (29 kDa). For large globular proteins, perhaps, it is required to try various relaxation delay times and collect numerous ^{1}H-^{15}H HSQC spectra for observing optimum signal to noise (S/N) ratio. In addition, particularly, the strand residues show relatively decreased intensity which can give erroneous values when measuring their intensities. Therefore, analysis of αβ and βα loops from αTS is pursued with the help of molecular dynamics simulation which provides reliable information on flexibility and rigidity at the atomic level.

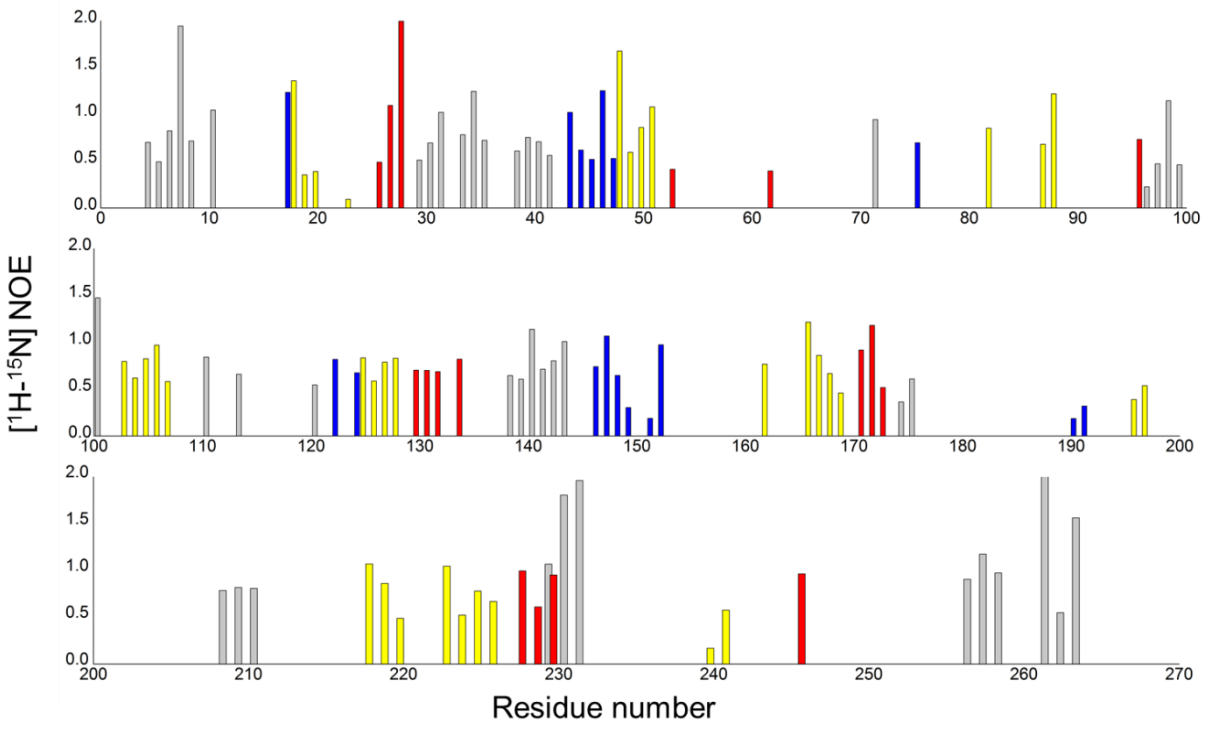


***Figure 3.5**: Backbone N-H vector [^{15}N,^{1}H] NOE values as a function of residue number for αTS*

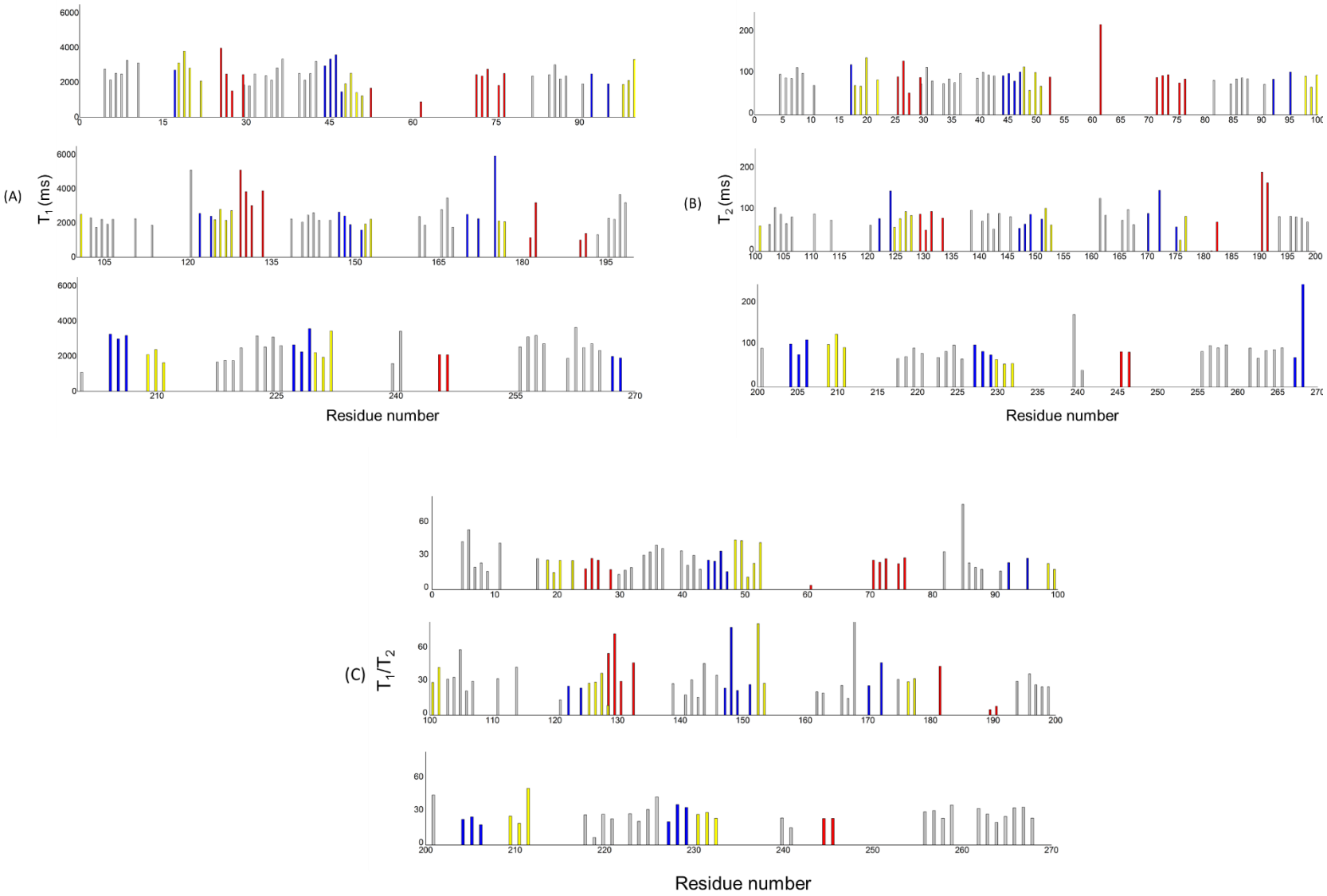


Figure 3.6: A.B*. T_1 and T_2 relaxation measurement values of αTS as a function of protein sequence residue number* ***C****. The T_1/T_2 ratio values plotted as a function of residue number.*

Backbone rigidity and flexibility analysis using Molecular Dynamics Simulations

B-factors in the protein X-ray crystallographic structures represent the spread of atomic electron densities around their equilibrium positions due to thermal motion and positional disorder and serve as indicators for regions of flexibility and stability in folded proteins (Parthasarathy & Murthy, 2000). Often it is noted that the regions of proteins and or group of residues possessing high B-factors exhibit increased flexibility and low stability. This is further corroborated in molecular dynamics studies wherein it is found that these flexible regions (regions with high B-factors) are the sites where protein unfolding is initiated (Daggett & Levitt, 1992; Lazaridis, Lee, & Karplus, 1997). **Figure 3.7** shows the pattern of the B-factors of individual residues obtained from the crystal structure of αTS. A closer examination of the pattern reveals three trends. First, the β strands show consistently low values throughout the sequence; (ii) the residues of βα loops show consistently higher values; (iii) residues of αβ loops show consistently low values than those in βα loops and some residues in certain helices. Consistently higher values αβ loops suggest that the βα loops may be less stable than the αβ loops. Yet another interesting feature within these regions is the gradation of the B-factor values from N to C termini of individual β stands and the αβ loops. The B-factors of these two elements of structures uniformly decrease from their N to C termini. On the other, a majority of α-helices in αTS, have B-factors reduced in the middle and more towards the C-termini. In contrast, the values, for the residues in the βα loops show a uniform increase from their N to C termini. By linking the three observations, it appears as if there is a rim of increased rigidity encompassing the middle towards to C-termini of α helices, N termini of β strands and αβ loops.

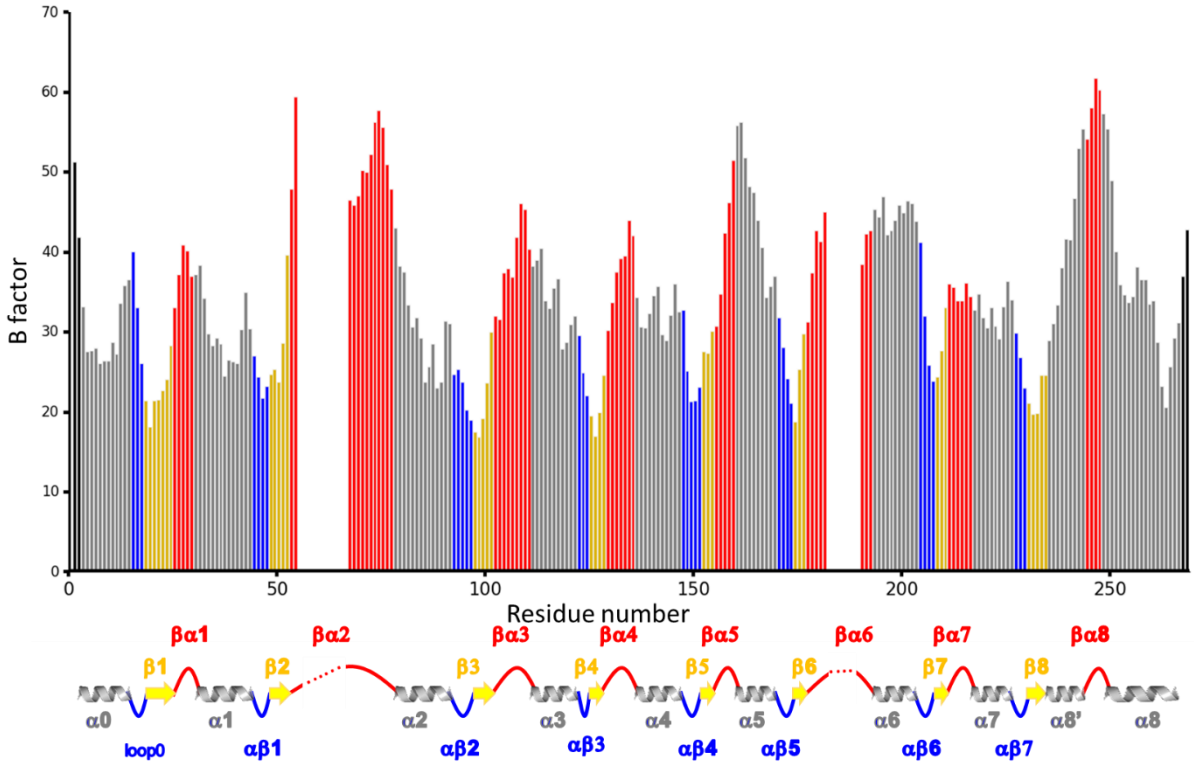
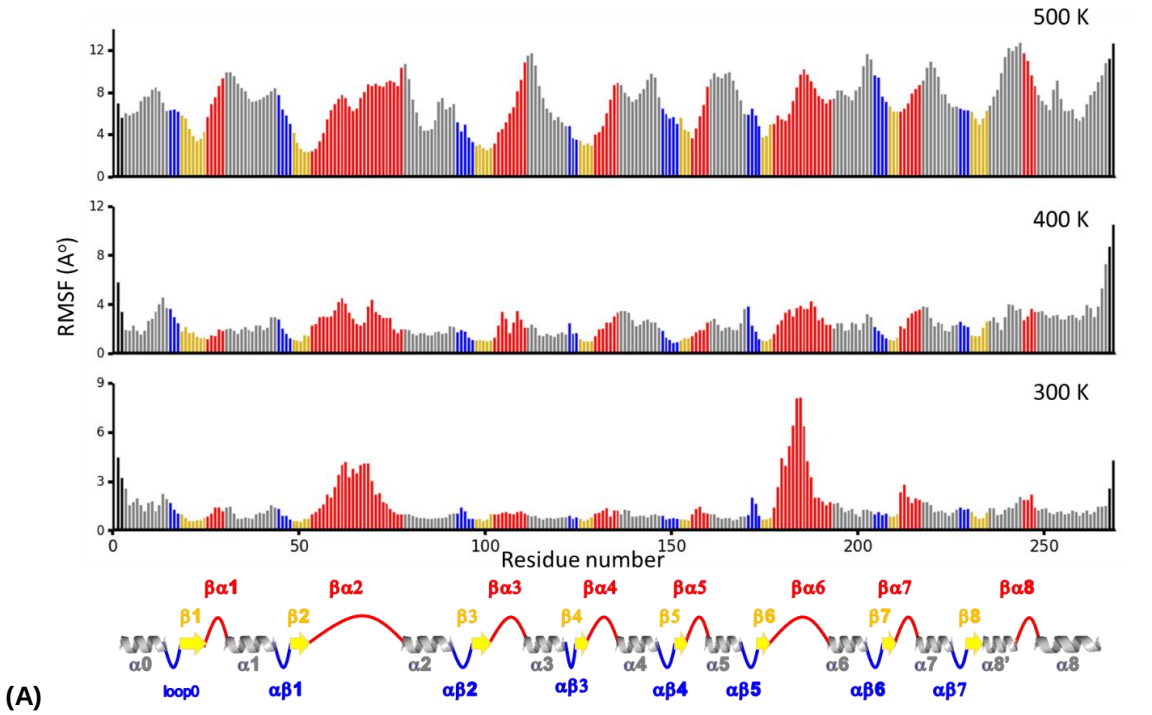


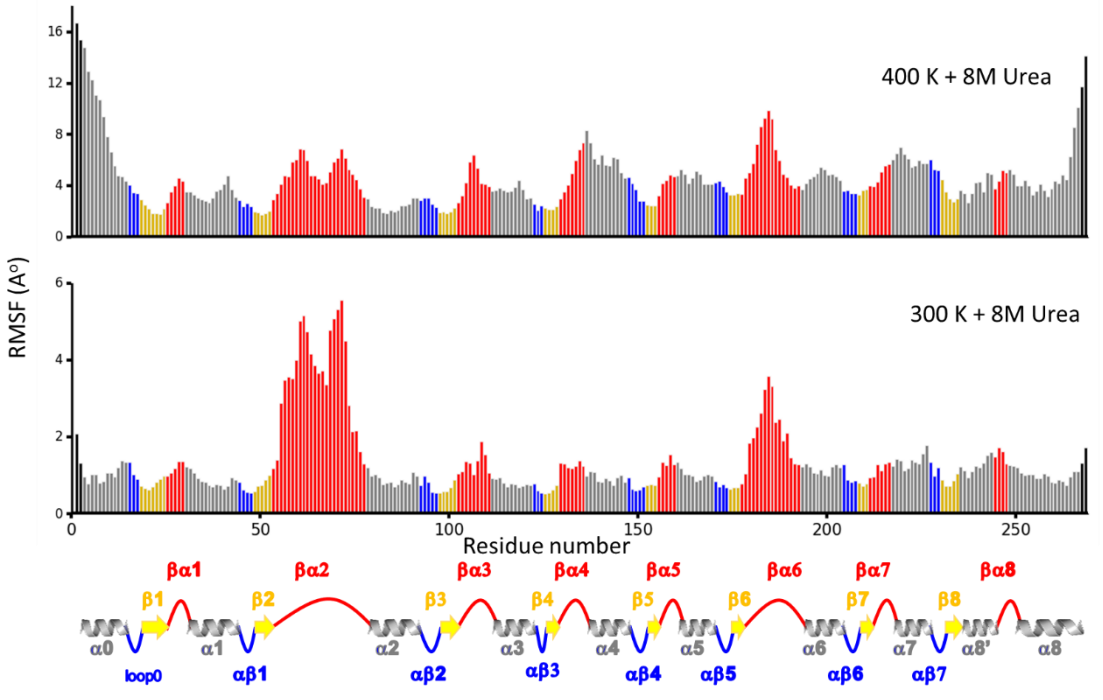
Figure 3.7*: B-factors of individual amino acids residues obtained from the crystal structure of αTS (PDB ID:1v7y). Helices are represented in grey, strands in yellow αβ loops in blue and βα loops in red. The dotted regions in topology indicate the missing structural information in PDB crystal structure.*

Another, generally, used indicator for flexible and stable (less flexible) regions along the polypeptide backbone is the root mean square fluctuations (RMSF) (Keskin, Jernigan, & Bahar, 2000; Maguid, Fernandez-Alberti, Parisi, & Echave, 2006; Papaleo, Riccardi, Villa, Fantucci, & De Gioia, 2006), a property comparable to the crystal structure B-factors, which basically is a measure of the average atomic mobility of backbone atoms calculated from the ensemble of conformations observed in molecular dynamics (MD) simulations. Atomistic motions of the Cα analyzed in terms of RMSF and their fluctuations at temperatures 300, 400 and 500 K are depicted in **Figure 3.8 A**. At 300 K the regions corresponding to two $\beta\alpha_2$ loops, 2 and 6 show significantly higher fluctuations. The same behavior is retained in the urea induced unfolding simulations

(**Figure 3.8 B**). With the increase in temperature to 400 K, although βα loop6 shows significantly higher RMSF values, no change in flexibility is apparent for βα loop2, however, an overall increase in flexibility throughout the entire length of the protein backbone may is apparent. Finally, at 500 K increased non-uniform RMS fluctuations along the entire length of the backbone can be observed. The fluctuations for the β strands and αβ loops, in general, are, however lower than α helices and βα loops. A closer examination of the RMS fluctuations revealed an interesting dynamic behavior for at least some of the α helices, especially at 500 K. The rigidity in at least some of the individual helices appears to be higher in the middle and steadily increases towards their N and C termini. This trend pertains to α1, α2, α4 located in the N-terminus segment of the protein. Further, it may also be seen that the residues of the helices show reduced flexibility closer towards the αβ loops, in general. This pattern of relative increase in rigidity in the middle of certain α helices, and at the N-termini of β strands and αβ loops is also reflected in the crystal structure B-factors, as discussed in the previous section. Overall, the combination of B-factors and RMS fluctuations suggest that the first four β strands ($\beta_1\beta_4$) and αβ loops 1-4 display more stability that the other four strands (β5-β8) and that the αβ loops are more rigid than their βα counterparts.



(A)



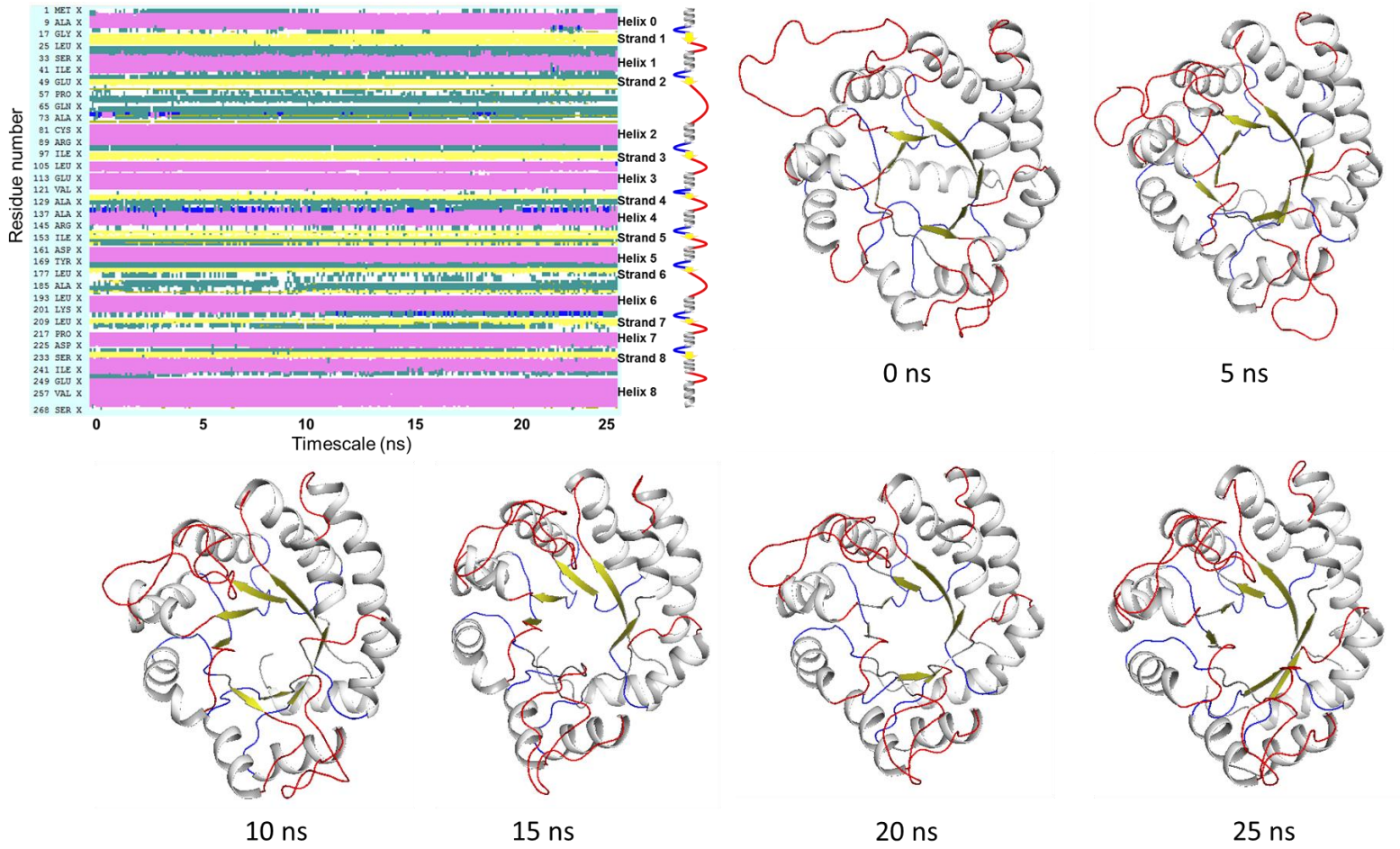
(B)

Figure 3.8: A*. RMSF of Cα atoms at simulated temperatures 300, 400 and 500 K.* ***B****. RMSF fluctuations at simulated temperatures and in the presence of 8M urea 300 and 400 K. Helices are represented in grey, strands in yellow, αβ loops in blue and βα loops in red.*

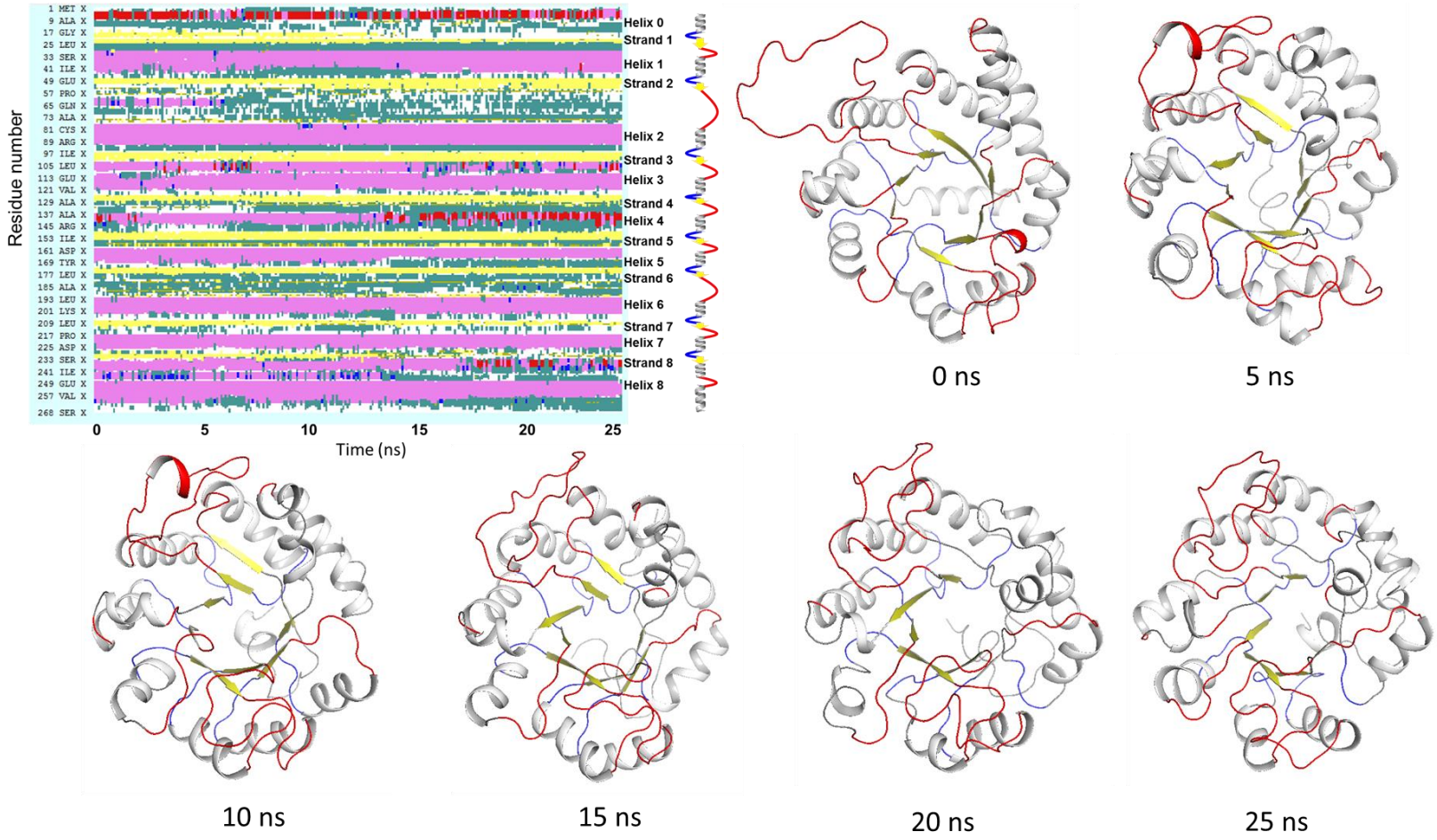
Secondary structure unfolding events

Figure 3.9 reveal the order of unfolding events of the secondary structure elements of αTS at 500 K. A continuous and progressive unfolding of α helices, in general, and some β strands, in particular, can be observed in the unfolding of the protein. The loss of structure within the first 1-2 ns at α0 and α8 suggests that the unfolding is initiated at these positions. Following this initial unfolding event a steady loss of structure can be observed for β strands 5-8 and helices α0, within 5 ns. α2, α3, β4 and β1 appear to be stable between 5- 15 ns. Clearly the most stable elements of secondary structure are β2 and β3 as they persist to the end of the simulation. Overall, it appears that from the time-dependent unfolding events the first four β strands (β1β4) and αβ loops flanking the β strands 2 and 3 along with the βα1 loop (the loop immediately succeeding β1) are more resistant to unfolding. The C-terminal part of the protein, starting from β strand 5 onwards is relatively less stable. Interestingly, combining all the information it also appears that αβ loops 2 and 3 that flank the β strands 2 and 3 are relatively more stable than the other loops.

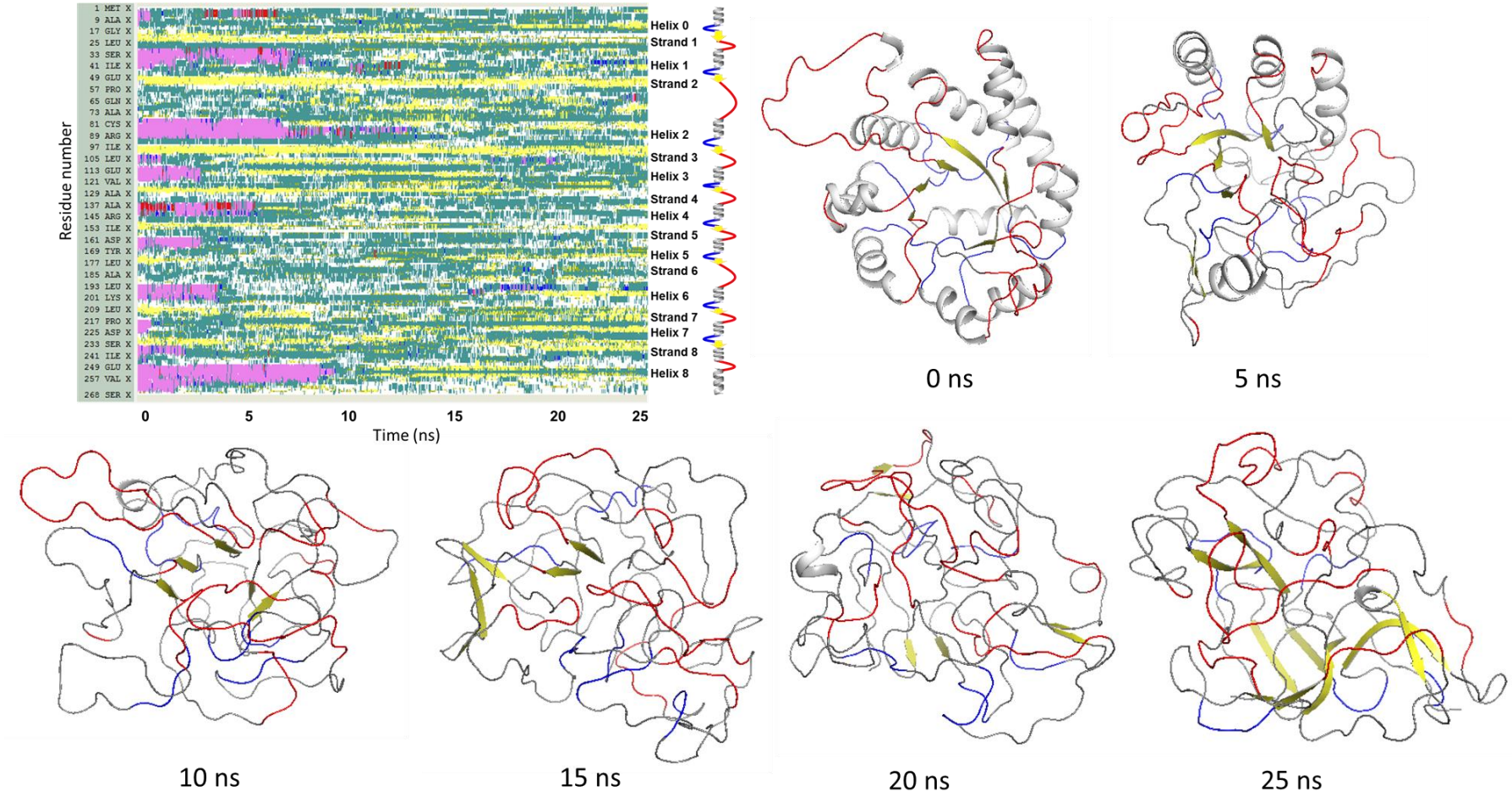
This general observation is consistent with the string of experiments on αTS which reveal that the N-terminal half comprising of regions up to the end of β strand 4 is more stable. Site specific information from NMR and HX further indicated that the most stable region is comprised of α1, β2 and β3, adjacent elements of secondary structures.


Residue number
Timescale (ns)
0
5
10
15
20
25
Helix 0
Strand 1
Helix 1
Strand 2
Helix 2
Strand 3
Helix 3
Strand 4
Helix 4
Strand 5
Helix 5
Strand 6
Helix 6
Strand 7
Helix 7
Strand 8
Helix 8
0 ns
5 ns
10 ns
15 ns
20 ns
25 ns

(A) 300 K


Residue number
Time (ns)
0
5
10
15
20
25
Helix 0
Strand 1
Helix 1
Strand 2
Helix 2
Strand 3
Helix 3
Strand 4
Helix 4
Strand 5
Helix 5
Strand 6
Helix 6
Strand 7
Helix 7
Strand 8
Helix 8
0 ns
5 ns
10 ns
15 ns
20 ns
25 ns

(B) 400 K



(C) 500 K

Figure 3.9: A.B.C*. Unfolding events of the secondary structure elements with the snapshots of αTS at 300, 400 and 500 Kelvin simulated temperatures for 0-25 nanosecond simulation time scale. Helices are represented in grey, strands in yellow αβ loops in blue and βα loops in red.*

Intramolecular non-covalent interactions

In some cases it has been shown that the interactions from within the loops can encourage stabilizing interactions between the flanking elements of secondary structures and can also contribute to the stability in certain cases (Chothia & Finkelstein, 1990; Gekko et al., 1994; Hoedemaeker et al., 1993; C. R. Matthews & Crisanti, 1981; Ptitsyn & Finkelstein, 1989). For instance, in the case of four-helix bundle proteins, a significant contribution towards their stability arises from loop-helix interactions rather than from helix-helix interactions (Kamtekar & Hecht, 1995). It is also well known that hydrophobic clustering of residues from the adjacent strands and loops in βhairpins favors their formation (Colombo et al., 2003). More recently, Matthews and his colleagues have experimentally demonstrated that mutations leading to the disruption of long-range main chain to side chain hydrogen bonding interactions involving the main chain amides of the N-terminal residue in β strands and the side chains of polar residues in the αβ loops in a couple of TIM barrel protein including αTS, lead to a dramatic loss of the stability of TIM barrel fold implying their key role in structural integrity (X. Yang et al., 2009; X. Yang et al., 2007). Further, a high distribution of side chain to main chain hydrogen bonding interactions involving the polar side chains of amino acids in the loops with the main chain amides and carbonyls of the N and C -terminal residues of the β strands, clamping βα hairpins and bracketing the βαβ modules in TIM barrels (X. Yang et al., 2009; X. Yang et al., 2007). With this in mind, we have examined the extent of non-covalent interactions including hydrogen bond, hydrophobic and ionic interactions which can influence the overall packing and stability. The residue-wise non-covalent interactions, hydrogen bonding, ionic and hydrophobic in nature are summarized in **Figure 3.10**. The residue-wise RMS fluctuations at 500 K are also included in the figure to investigate if there exists a correlation between increased non-interactions with increased rigidity. From the **Figure 3.10**, in general, it may be observed that the number

of per-residue interactions in the C-terminal region of the protein is relatively fewer in number. Also, β strands are involved in a maximum number of inter-residue non-bonded interactions, particularly in the first half of the protein and therefore contributing to the relatively higher rigidity observed for the β stands, particularly, β1–β4. The number of non-bonded interactions involving the residues resident in β strands 5, 6, 7 and 8 is fewer compared to the residues residing in βstrands 1-4, and, perhaps, the slight increase in flexibility and decrease in stability for these regions. This trend can also be observed in certain α helical regions. From **Figure 3.10**, it is interesting to note that the increased number of non-bonded interactions is also centered in the first half of the protein.

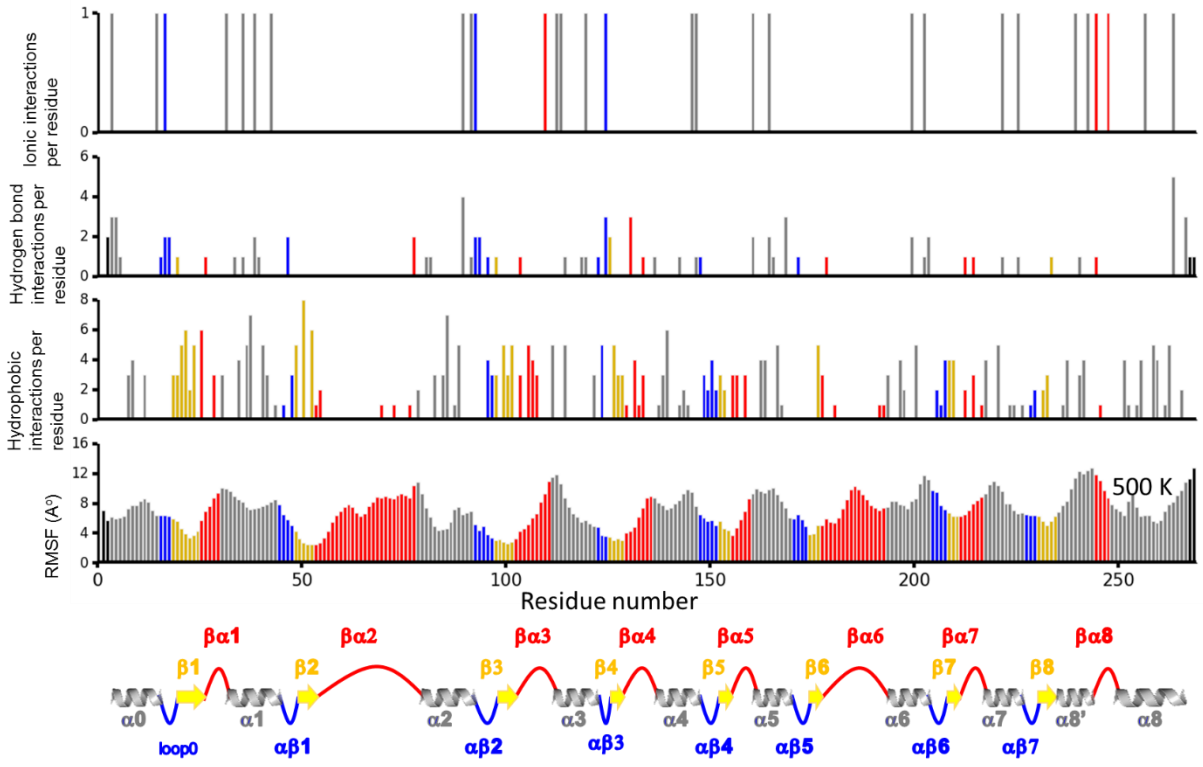


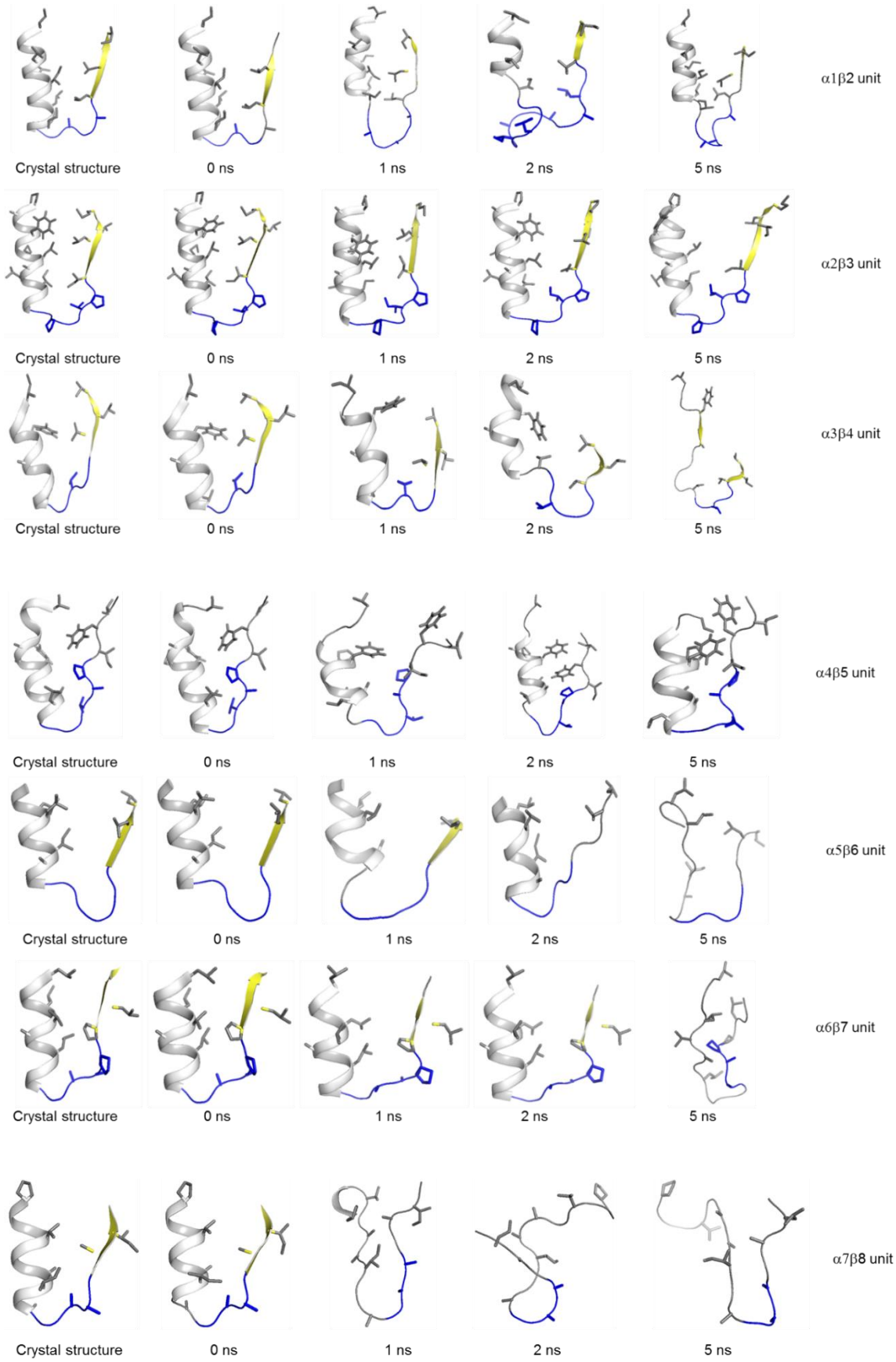
***Figure 3.10**: The hydrophobic, hydrogen bond and ionic interactions of individual residues obtained from the crystal structure of αTS (PDB ID:1v7y) and RMSF for the simulation timescale of 1-25 nanoseconds at 500 K. Helices are represented in grey, strands in yellow, αβ loops in blue and βα loops in red.*

Focusing on the loops, on comparing the per residue non-bonded interactions it is very clear that the αβ loops show increased number of contacts than the βα loops. It may noticed that αβ2 in the stable N-terminal region is involved in the long range main chain amide of F19 to side chain carboxyl group of Asp46 connecting the N-terminus of β1 and the subsequent αβ loop residue Asp 46 just before the beginning of β2 (X. Yang et al., 2009; X. Yang et al., 2007). Similar, long range main chain to side chain interactions also involve I97 back bone amide, of β3 and Asp 124 side chain carboxyl group in the αβ loop preceding β4, A103 backbone amide and Asp 130 side chain carboxyl group (X. Yang et al., 2009; X. Yang et al., 2007). These three interactions in αTS, between the main chain hydrogen bond donors and side-chain hydrogen bond acceptors, connect the N-terminus of one element of secondary structure, either β-strand or α-helix, to the C-terminus of the subsequent element of structure, either α-helix or β-strand, respectively. These non-local main chain to side chain interactions in αTS designated as βα-hairpin clamps have been experimentally shown to contribute significantly to the overall stability of the protein. Removal of these interactions by mutating these side chains to alanine resulted in the loss of overall stability by of 4-6 kcal mol^{-1}) (X. Yang et al., 2009; X. Yang et al., 2007). Therefore, it is likely that these, long range main chain to side chain hydrogen bond interactions be playing a role in the rigidity of αβ loops. Further, on closer examination, hydrophobic interactions between non-polar amino acids also appear to be higher, particularly in αβ loops 2, 3, 4 (**Figure 3.11A)**. Moreover, the interactions form a cluster of non-polar interactions primarily contributed by the residues in the C and N termini of α helices and β strands respectively (**Figure 3.11A**) The residues from the αβ loops are also a part of the cluster, mediating the interactions between the residues in helices and strands (**Figure 3.11A**). This probably provides an explanation for not only to the observed rigidity as reflected by the reduced B-factors an RMS fluctuations but also the experimentally observed stability. In this context,

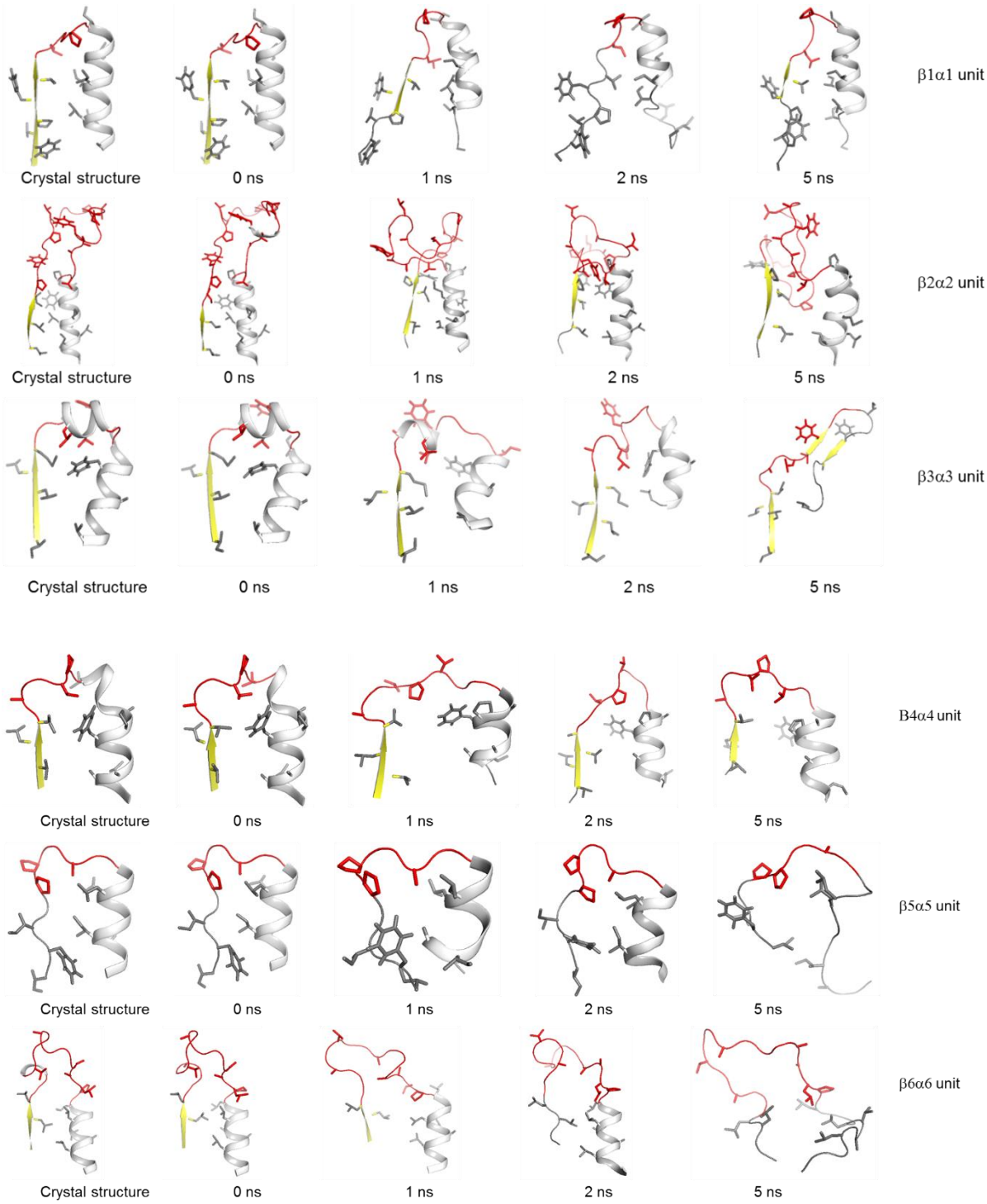
it is interesting to note that, native state HX data indicated enhanced protection at the N-termini of β strands in and the stability steadily decrease towards the C termini (Vadrevu et al., 2008). Further, based on thorough investigations using a variety of spectroscopic data Matthews and co-workers have identified a key role for the clusters of non-polar amino acids in stability and folding of β/α proteins (Gangadhara, Laine, Kathuria, Massi, & Matthews, 2013; Wu et al., 2007). The relatively smaller size together with the resident non-polar residues can, therefore, be more effective in promoting interactions between the flanking helices and strands.

In contrast, such interactions are sparse in βα loops **Figure 3.11 B**. Thus the αβ loops connecting α helices and β strands provide compactness by confining distant regions closer in space, thus, leading to reduced flexibility by promoting stabilizing interactions between the flanking elements of secondary structures. This possibility may also exist in short βα loops and or to the residues in the beginning of the βα loops. As can see, even in the case of βα loops residues at the N-termini of the loops show reduced rigidity, presumably, contributed by the stable β strands. However, the overall rigidity may be compromised by the lack of substantial stabilizing non-covalent interactions in addition to the increased length of the βα loops. Perhaps, the distinct role for αβ loops in stability is an outcome of optimization of different non-covalent interactions rather than any one or more specific interaction.

(A)


Crystal structure
0 ns
1 ns
2 ns
5 ns
α1β2 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α2β3 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α3β4 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α4β5 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α5β6 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α6β7 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
α7β8 unit

(B)


β1α1 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
β2α2 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
β3α3 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
B4α4 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
β5α5 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns
β6α6 unit
Crystal structure
0 ns
1 ns
2 ns
5 ns

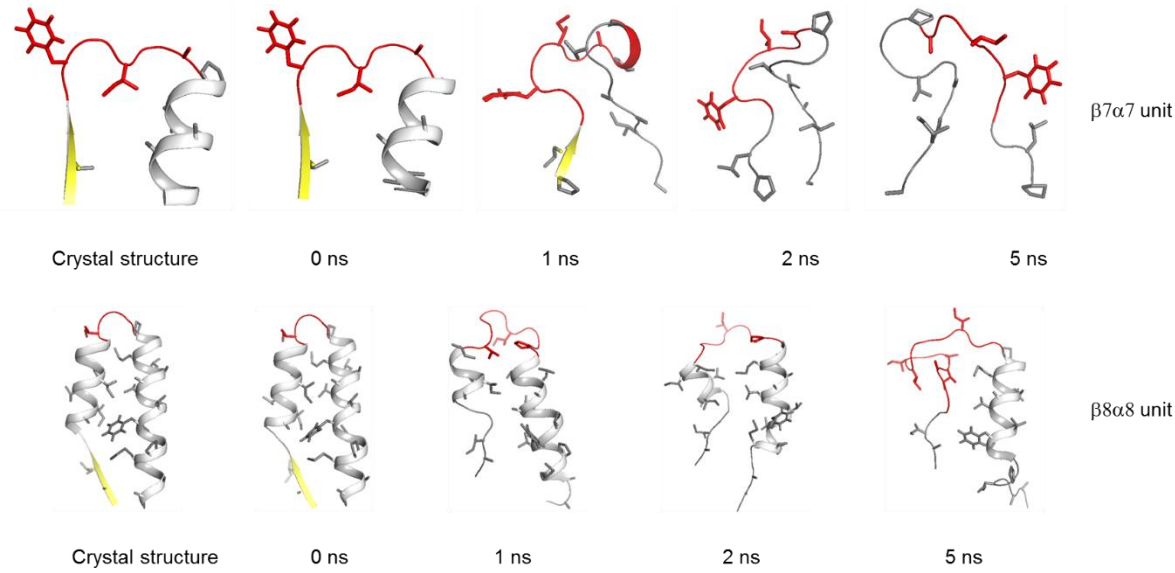


Figure 3.11: A. *The snapshots of αβ units from crystal structure and unfolding timescales of 0, 1, 2 and 5 nanoseconds depicting the cluster of non-polar interactions primarily contributed by the residues from the C and N termini of α helices, αβ loops and β strands of αTS respectively.* ***B.*** *The snapshots of βα units from crystal structure and unfolding timescales of 0, 1, 2 and 5 nanoseconds depicting the cluster of non-polar interactions primarily contributed by the residues from the C and N termini of β strands, βα loops and α helices of αTS respectively.*

Role of Non-Covalent Interactions in the stability of αβ loops in TIM Barrel Proteins

Analysis of structural determinants of αTS, whose unfolding and stability have been extensively studied, revealed that the β strands constituting the core of the fold are resistant to unfolding, in general, and strands 1-4 in particular (Rojsajjakul, Wintrode, Vadrevu, Robert Matthews, & Smith, 2004; Vadrevu et al., 2008). The α helical segments adjacent to β2 and β3 display higher rigidity and resistant to unfolding, is consistent with the experimental data. αβ loops show higher rigidity than the βα loops.

From this analysis, it is clear that, that there is a clear distinction between αβ and βα loops in their flexibility and their resistance to unfolding. The relatively higher rigidity of αβ loops is in concurrence with the increased number of stabilizing interactions. Overall, αβ

loops take part in an increased number of non-covalent interactions and it appears that this network of interactions bestow the rigidity and stability to the αβ loops. Added to this, the smaller size of the loops can assist establishing optimized packing and registry of β strands (X. Yang et al., 2009). Thus, a relatively increased number of inter-residue noncovalent interactions including hydrogen bonds, ionic and hydrophobic contacts from within the loops and their flanking secondary structural elements are inherent to and dominating in αβ loops and therefore the reason for their role in the overall fold stability.

Extending the role of non-covalent interactions arising from the αβ loops to the set of αβ loops from the TIM barrel proteins, it is interesting to note that the main chain to side chains clamping βαβ modules clamps are ubiquitously observed in TIM barrel proteins. The geometry of the barrel is slightly tilted with respect to the central axis of the barrel and this facilitates optimal situation for hydrogen bonding network between the strands of the barrel, long range main chain to side chain interactions and packing between the helices and strands. The short αβ loops, therefore, facilitate tighter interactions and increased rigidity.

Analyses from experimental and bioinformatics approaches clearly indicate that the βαβ modules, serve as the minimal unit of stability in β/α class of proteins. Our observations and reasoning will in addition to steering protein engineering efforts on TIM barrel design and stabilization can provide the basis for identifying and or designing stable βαβ modules.

Chapter 4

LoopX: a graphical user interface based database for comprehensive analysis and comparative evaluation of loops from protein structures

Introduction

Loops, the regions that connect secondary structure elements constitute the key components of protein function (Fetrow, 1995; Fiser et al., 2000). In addition, they also play a major role in protein stability and folding (Balasco et al., 2013; Fu et al., 2009; Lewandowska et al., 2010; Marcelino & Gierasch, 2008). The design of novel functions in proteins with increased stability has been a long-standing goal in the field of protein engineering. Recently, experimental endeavors indicate that grafting/engineering of loop regions in protein scaffolds can lead to novel activity and can alter existing functionality (Ochoa-Leyva et al., 2009; Park et al., 2006; Walser et al., 2012) of enzymes. In addition, the strategy of swapping loops/turns has also been exploited to enhance protein stability and foldability (Binz et al., 2005; Fu et al., 2009; Ochoa-Leyva et al., 2009; Tawfik, 2006; Walser et al., 2012; Wijma, Floor, & Janssen, 2013).

With the demonstrated role of loops in catalysis, molecular recognition, protein-protein interactions, antibody recognition, signaling etc., creating novel and or altering existing functionality, enhancing stability and foldability are being pursued aggressively via grafting/engineering of loops in protein scaffolds (Binz et al., 2005; Ochoa-Leyva et al., 2009; Tawfik, 2006; Walser et al., 2012; Wijma et al., 2013).The strategy of working with loops has the advantage of introducing a much more divergent sequence space in contrast to single amino acid substitutions.

However, the subtle balance between sequence, structure, stability and function imposes a high degree of constraint on choosing compatible loops for grafting/engineering that

preserve the overall fold and stability and yet deliver the desired functionality. Choosing compatible loop(s), to retain the overall stability for the expected functionality for a given target loop, to a large extent could be dictated by the similarity in size (end-to-end distance) or the shape (conformation) of the target and candidate loops. Intuitively, it is expected that for exchanging loops at a particular position in a protein, smaller the distortions between the target and candidate loops lesser will be the effect on stability and conformational distortion. In this context, a comprehensive analysis of both sequence and structural features of the target loops to be replaced will be required as a first step towards finding candidate loops. Following this, knowledge-based methods such as data mining, database searching aided by visualization will facilitate in assessing many different alternative candidates for the target loop grafting/engineering with minimal distortions. In addition to allowing searching and sampling candidate loops for a particular loop from structural databases such as Protein Data Bank (PDB) (H.M. Berman et al., 2000), database search methods may be appropriately suited as they depend on local sequence similarities and anchor geometries suggesting similar local structures. Given, the rapid growth and increase in protein structural information, flexible tools and web server based databases will greatly aid in a guided short-listing of most appropriate candidates. Therefore, the primary goal of our work is to facilitate the comparison of loops that resemble the target loop through a user defined criteria for conformation, sequence, or size, via flexible query tools powered by a graphical user interface and visualization.

Databases of protein loops have been developed and some of them are available and primarily serving as predictive tools for modeling purposes. For instance, graphical user interface (GUI) web servers such as FALC, LIP, ModLoop, and FREAD (Choi & Deane, 2010; Fiser & Sali, 2003; Ko et al., 2011; Michalsky et al., 2003), cater to the modeling of the missing loops in proteins structures. Slightly different, Brix database (Vanhee et al.,

2011), houses loops from proteins which are grouped as (i) loops that vary in their length but show similar end-to-end distance (ii) loops of identical length and similar structure. With the help of user-friendly GUI, it allows the users to upload their choice of PDB structure and find appropriate loops that are otherwise missing in the structure. Although some of the existing databases provide some options to users, it is highly desirable to have a comprehensive repository of loops from the currently available protein structures combining all query tools to extract and compare sequence, structure and other features of protein loops via the user-friendly graphical interface. It may be noted that there is no dedicated database with highly flexible query tools to allow researchers for efficient data mining and selection of desired loop candidates/sequences for loop grafting/engineering. In our current work we present, LoopX, a comprehensive database consisting of ~ 7,00,000 loop candidates of 3-14 residues in length, mined from non-redundant protein structures with <90% sequence similarity from Protein Data Bank (PDB) (Berman et al., 2000). The LoopX database has been first developed for only TIM barrel proteins comprising only 4320 loop candidates of 3-14 residue length, given the unavailability of a dedicated database for comprehensive and comparative analysis of loop conformations LoopX database has been extended to all available non-redundant protein structures with <90% sequence similarity from Protein Data Bank. The database has been equipped with a wide variety of query tools to mine loops based on their structural and sequence similarity. Two additional features are incorporated in LoopX. First, is the inclusion of an assessment of the polar and non-polar environment in the search for compatible loops based on the criterion of the root mean square deviation of the target loops. Second, LoopX stretches the flexibility to include the sequence and structural conservation in the search for structurally similar loops as a query option. Equipped with a graphical user interface, the database provides various rendering options to visualize sequence and structural level information along with hydrogen bonding

patterns, backbone phi (ϕ), psi (ψ) dihedral angles of both the target and candidate loops, Thus LoopX serves not only to list and visualize loops of a chosen protein but also allows the users to extract, compare and analyze the target and candidate loops, complementing to other databases on protein loops.

Methods

Non-redundant protein structures dataset with 90% sequence similarity and a resolution of ≤ 3 Å was mined from the Protein Data Bank (PDB) and segregated into the PDB90 dataset. In-house built Biopython scripts are used to extract individual non-redundant protein chains from the mined protein structures. Secondary structure information is obtained for all the proteins from the datasets using DSSP program (Joosten et al., 2011; Kabsch & Sander, 1983) developed by Wolfgang Kabsch and Chris to assign secondary structure conformation to a given 3D structure of a protein. DSSP works by reading the position of atoms in the protein followed by hydrogen bond energy calculation between all atoms and best two hydrogen bonds for each atom is used to assign the most likely class of secondary structure for a residue. Based on their secondary structure assignment from DSSP, the regions flanking the helical and strand regions were identified as loop conformations using in-house built python scripts. For the loops thus identified their sequence, and the DSSP assigned conformation for individual residues are extracted and stored in MySQL database. Considering the role of loop chirality in connecting different combinations of α helices and β strands as found in the four major protein structural classes (Koga et al., 2012), the loop database is further segregated and stored as loops from all α, all β, α+β and α/β proteins. This provides flexibility to allow a search for candidate loops based on the picked target loop from a particular structural class. In addition to the listed information, the database also houses the backbone ϕ, ψ dihedral angles of individual loop residues (**Figure 4.1**).

LoopX database will be updated periodically by extracting the loops from the new protein structures deposited in the PDB database, using in-house developed python scripts and following the protocol described above. The information on the last update will be displayed on the LoopX homepage.

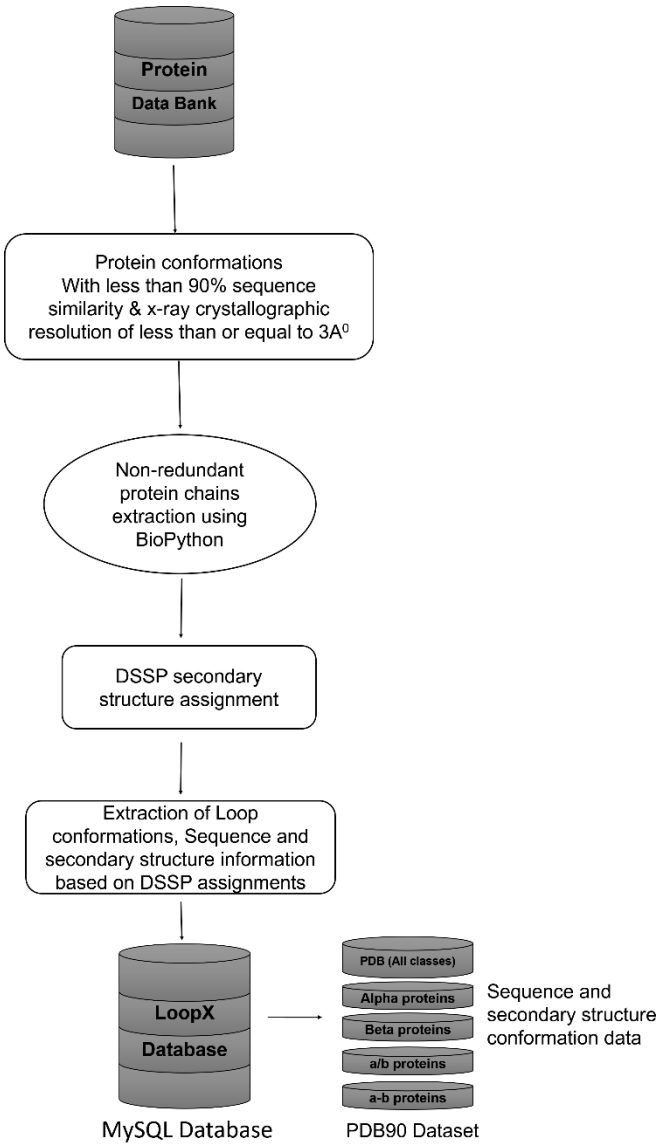


Figure 4.1: *Schematic representation of LoopX database construction.*

Search Tools/Algorithms

LoopX provides the users with three default search criteria and various options for searching and pooling of candidate loops for a chosen target loop. Similar target loops can be

identified based on backbone atoms root mean square deviation (RMSD), Cα end-to-end residue distance. The root mean square deviation (RMSD) based search criterion helps to retrieve loops based on backbone conformational similarity. The target and candidate loops are superimposed and fitted for root mean square deviations employing Kabsch algorithm (Wolfgang, 1978) using in-house developed python scripts. The RMSD based search can also be extended to include the similarity of the polar/non-polar environment ("V score") between the target and candidate loops

Root Mean Square Deviation (RMSD) based search criteria

RMSD based search criteria help to fetch candidate loop conformations with similar backbone scaffold as target loop. Root Mean Square Deviation (RMSD) is the average distance between two atomic positions of the superimposed 3-dimensional conformations. To calculate the RMSD between two loop conformations the two structures has to be in the same center and rotated on to one another, to ensue this a biopython script was written to find the centroid for both the conformations followed by stirring both the molecules to the center of the coordinate system. To align molecules, Kabsch algorithm(Wolfgang, 1978) is implemented to calculate the optimal rotation matrix and align the structures followed by calculation of RMSD for backbone atoms using the formula

$$\mathbf{RMSD\ (v,w)}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2+\left(v_{iy}-w_{iy}\right)^2+(v_{iz}-w_{iz})^2)}$$

Where i is the residue and v, w are the atoms for respective residue from two superimposed structures and x, y, z is the 3-dimensional spatial coordinates. To identify similar candidate loop conformation the RMSD of backbone atoms (Cα, C, N, and O) is considered then only Cα atoms. The RMSD values are usually given in Angstroms and the lower the RMSD

value the similar the two superimposed structures considered. **Figure 4.2** explains the detailed workflow of the RMSD based search criteria.

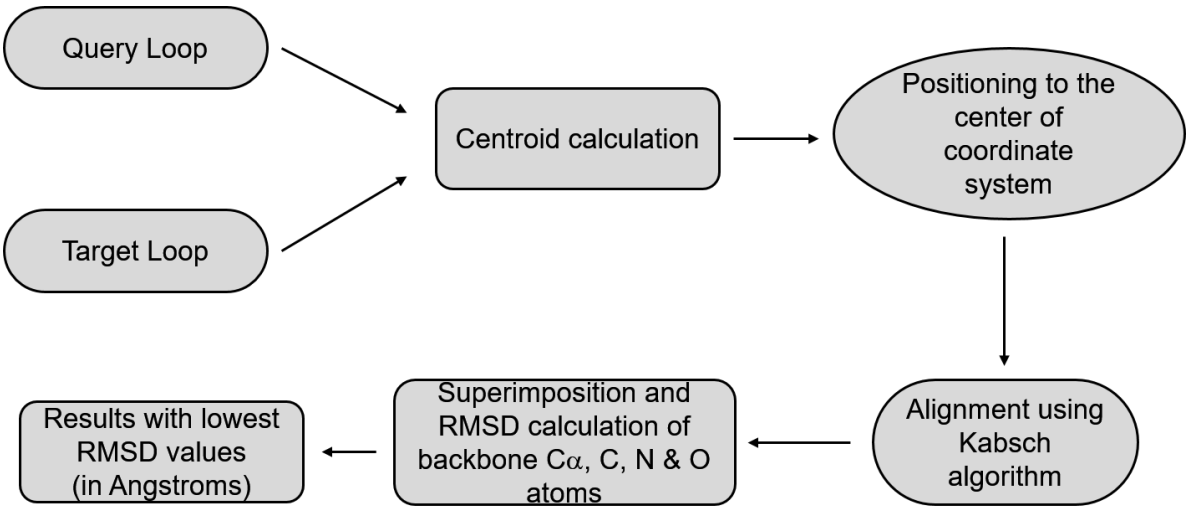


Figure 4.2: *Schematic representation of RMSD based search criteria.*

Cα end-to-end distance based search criteria

Cα end-to-end distance based search criteria is useful to find the candidate loops with similar end residue spatial conformations and of different length compared to target loop. In the Cα end-to-end distance based search, the in-house developed python script is used first to calculate the distance between end residues Cα atoms in space using 3-dimensional spatial coordinates (x, y and z) from PDB using the formula.

$$\mathbf{D}=\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Where D is the distance in Angstroms and x, y, z is the 3D spatial coordinates of atoms 1 and 2. The candidate loop conformations that are within the desired distance cutoff provided by the user will be further selected for superimposition of target and candidate loops. The fitting/superimposition of candidate loop on to the target loop for display is similar to that of the method used in RMSD calculation except that only backbone atoms of end residues

are used for fitting. **Figure 4.3** explains the detailed workflow of the Cα end-to-end distance based search criteria.

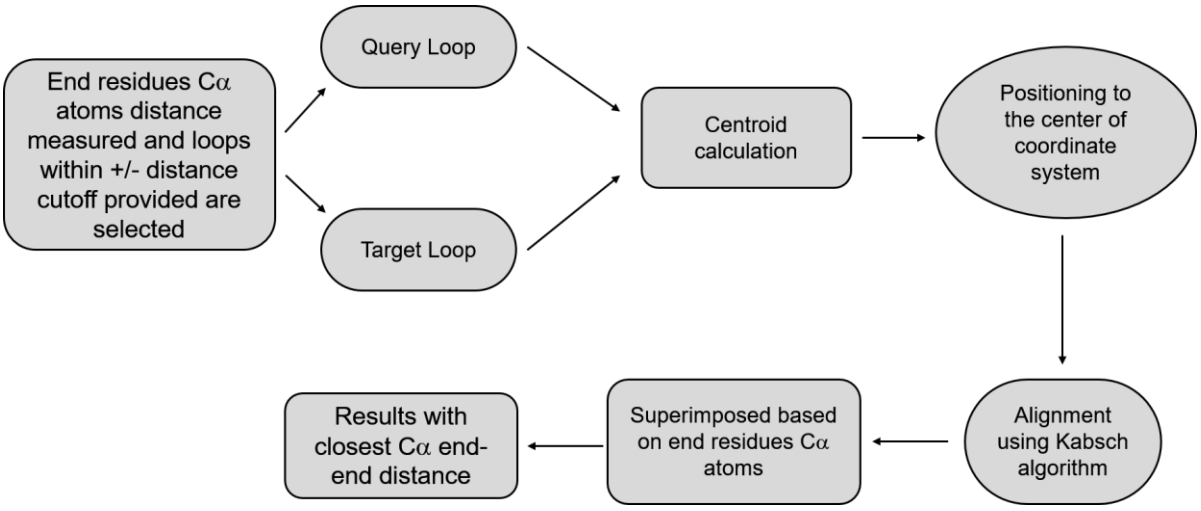


Figure 4.3: *Schematic representation of Cα based search criteria.*

RMSD and Similar Environment (V-score) based search

The RMSD based search has been extended to include the similarity of the polar/non-polar environment ("V score") between the target and candidate loops. To achieve this, we have developed an algorithm, which generates the polar/non-polar environment similarity working in two stages. First, the user-defined or the default backbone RMSD cutoff is applied to filter the loops which fall within the defined RMSD cutoff value. In the second step, hydropathy values are assigned for every amino acid residue side chain in a loop, along with the values for all the polar/non-polar side chains present within 6 Å of the loop residues. Following this, the overall average hydropathy value of all residues belonging to the loop and falling within 6 Å of the loop residues is calculated. The difference between the overall average hydropathy values of query and candidate loop can provide an assessment of the similarity of the two loops. A value of 0 indicates that the average hydropathy value is identical and increased values suggest a measure of the difference of environment between

the two loops. Thus, the V-score algorithm aids in arriving at loop candidates that can be accommodated in a similar environment as that of the target loop. **Figure 4.4** explains the detailed workflow of the "V-score" based search criteria.

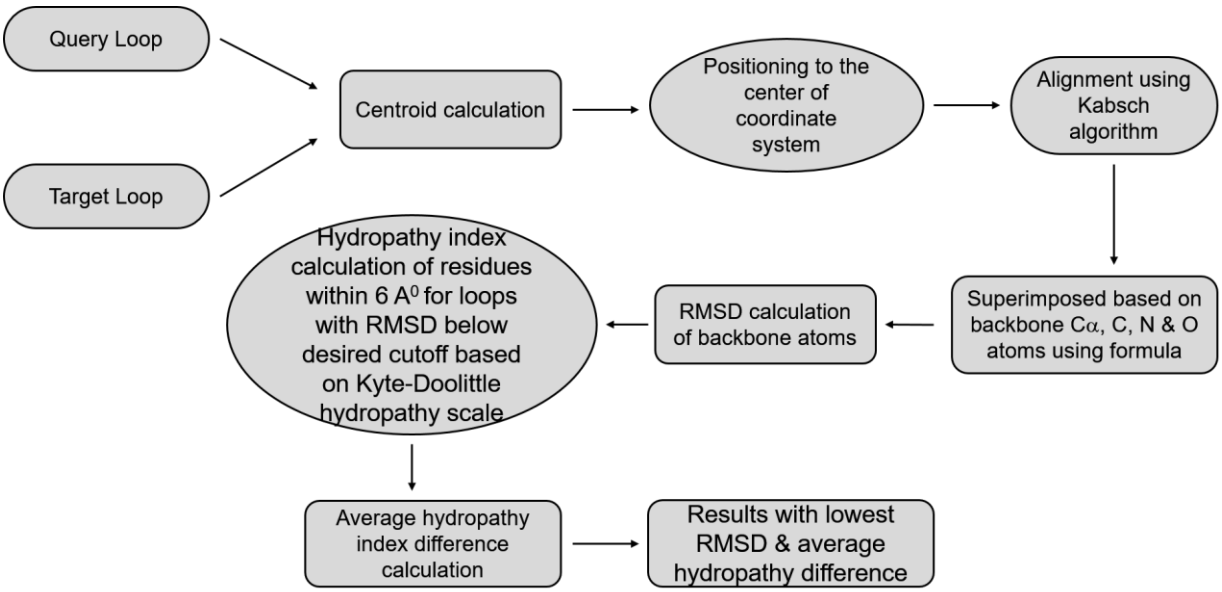


Figure 4.4: *Schematic representation of V-score based search criteria.*

Optional search tools

In addition to standard tools, another new feature made available to the users is the option for the conservation of a particular residue at specific position(s) and or the conformation of residue(s) of the target loop, in the search and retrieval process. Combined with default search criteria sequence and secondary structure conservation based search helps to fetch best compatible loops with desired features like amino acid sequence and secondary structure conformation.

Sequence conservation (amino acid residue) based search criteria

In loop grafting/design or in target-based loop design strategies, residues that play key roles in catalytic activity, binding or folding etc.) are retained on interest for desired features. Sequence conservation based search criteria combined with default search criteria helps to

conserve desired amino acid residues in compatible candidate loop hits, the high sensitivity level of sequence conservation search criteria allows the user to conserve the desired amino acids at specific positions.

Secondary structure conservation (as assigned by DSSP) based search criteria

Specific conformations like turns in loops are known to play a key role in folding and binding (Colombo et al., 2003; Dyson & Wright, 1991; Lewandowska et al., 2010; Munoz et al., 1997; Ramirez-Alvarado et al., 1997; Richardson, 1981). DSSP secondary structure assignments of loops give information regarding the possibility of turn and bend conformations in loops. In LoopX, an option to retrieve compatible candidate loops with desired secondary structure conformation as assigned by DSSP is provided. The secondary structure conservation based search also helps to retain specific secondary structure at desired positions.

LoopX GUI and Usage

LoopX Database web-based Graphical User Interface (GUI) is accessed online through http://182.75.45.10/loopx/loopx.html**.** LoopX GUI retrieves all the loops for a selected protein and enables users to search for structurally compatible loops for a chosen target loop mainly via a two-level user interface. **Figure 4.5** depicts the schema of overall database workflow.

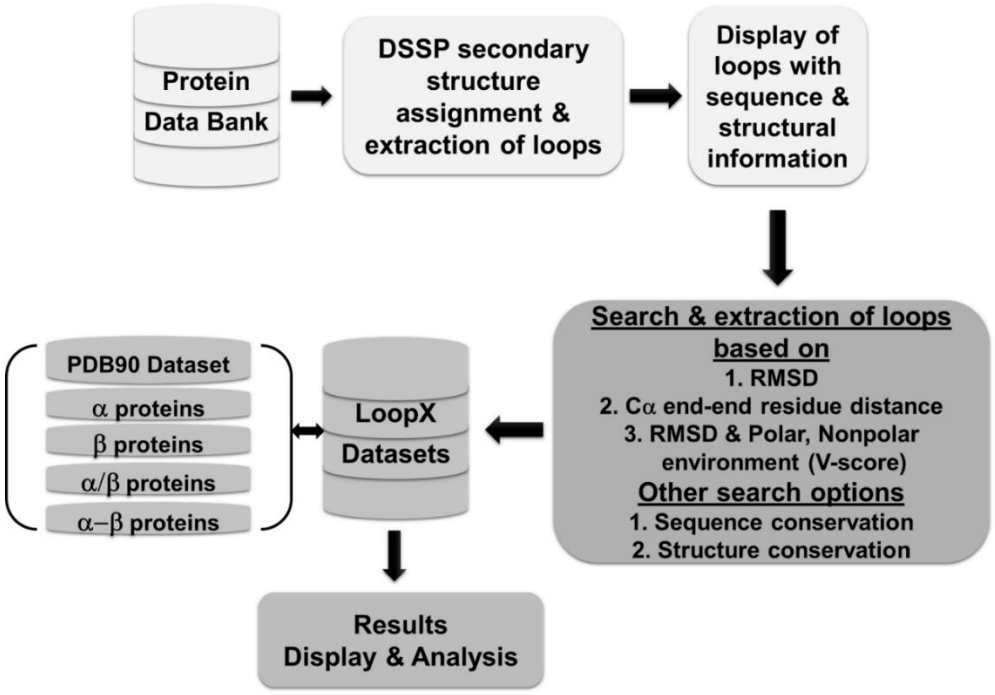


Figure 4.5: *Schematic representation of LoopX database workflow.*

Level 1 of Query Interface

At level 1, for a given PDB and chain ID, LoopX retrieves all the loops of the protein along with their sequence and structural details as assigned by DSSP. A molecular visualizer (Jsmol) with diverse rendering options allows visualizing loop conformation, hydrogen bonding interactions and the backbone ϕ, ψ dihedral angle plot of the loop residues.

At level 1, for a given PDB and chain ID, LoopX retrieves all the loops of the protein along with their sequence, and structural details as assigned by DSSP, a backend python script maintained by Perl CGI program downloads the respective PDB file of the provided PDBID and extracts the desired protein chain conformation and saves it in PDB file format. Followed by secondary structure assignments using DSSP program. Based on the DSSP secondary structure assignments complete loop conformations from the selected chain are extracted along with sequence and structure level information and presented to the user in level 2 for the selection of desired loop conformation for re-engineering/design. The

homepage page of LoopX is developed using HTML, CSS, JavaScript and linked to a Perl script for further processing.

Figure 4.6: *Screenshot of LoopX GUI homepage*

The second page for level 1 of LoopX provides the complete available loops for the selected protein chain along with the details regarding selected protein with a link to PDBsum database for more details. The starting and ending residue numbers of the every loop conformation are provided along with sequence and secondary structure (as assigned by DSSP) information. Jsmol a java web applet has been integrated into the second page with various display options for vigorous analysis of loop conformations for efficient selection of target loop to engineering/design. Various rendering options are provided for display of individual loops along with an option to display only loop and residues within $5A^0$, Hydrogen bonds display option is provided to monitor hydrogen bonding interaction within the loop and to the neighbor residues and Ramachandran plot analysis is also embedded to explore the Ramachandran plot dihedral angle distribution of loop residues. An option is also provided to select the loop region manually by defining the starting and ending residue numbers of desired loop region. The second page of level 1 is developed using HTML (for

web page output), CSS (to create buttons and tables), JavaScript (to integrate & pass commands to Jsmol), Perl (as CGI backend) & Python (to download PDB, run DSSP and extract loop information). After the selection of desired loop structure/region. The search button option provided for every loop is used to proceed to level 2 to select search criteria/tools for compatible loops search.

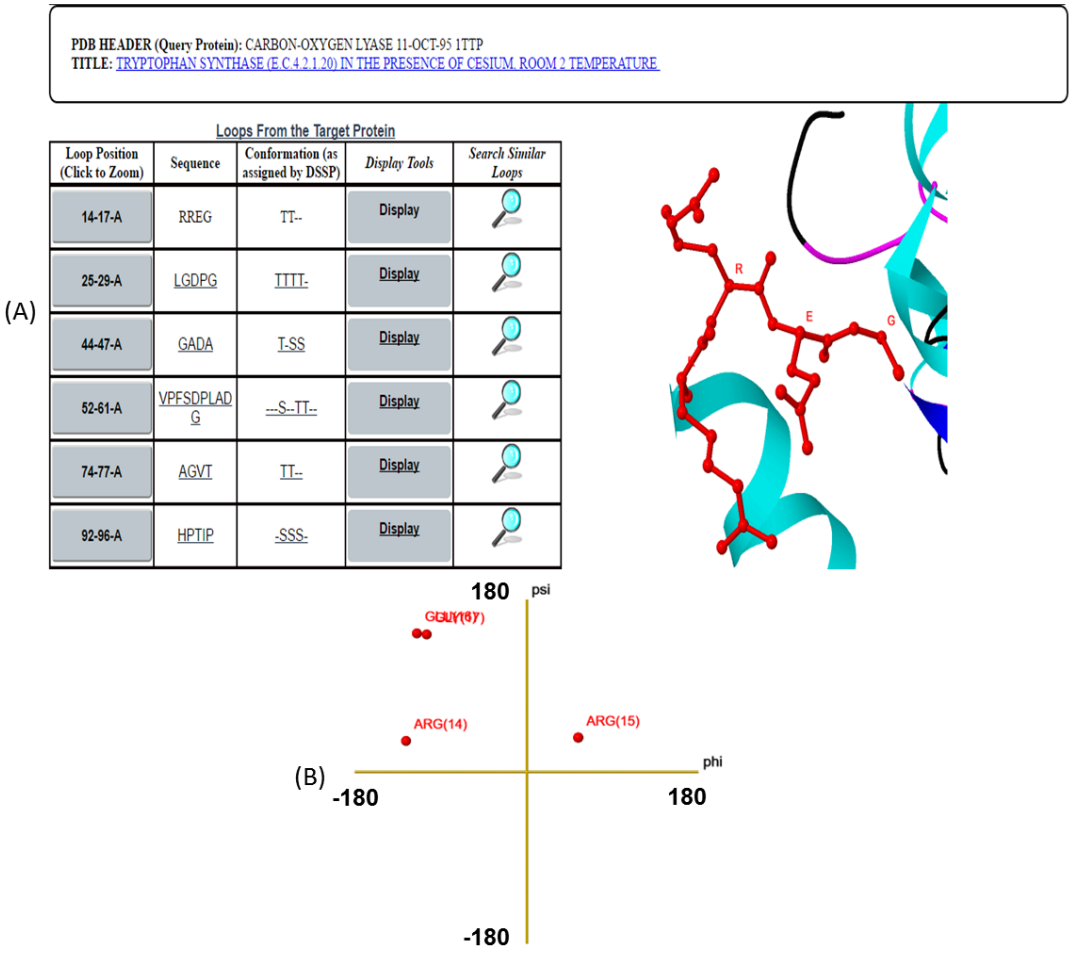


Figure 4.7 A. *A screenshot of the LoopX displaying some of the loops and their sequence and structural information for a chosen protein, 1TTP. On the right, is shown the part of the protein (ribbons) along with the selected loop, RREG, in red.* ***B.*** *Display of the backbone ϕ, ψ dihedral angles of the residues, selected RREG loop on the Ramachandran ϕ, ψ plot. The amino acids are indicated as three letter codes along with their residue number in parentheses. phi and psi on the x and y-axis denote ϕ, ψ respectively.*

Level 2 of Query Interface

At level 2, for a chosen target loop from the protein, the default or user-defined search criteria will identify similar loops based on either backbone conformational similarity or the Cα end-to-end distance. The option of assessing the polar and nonpolar environment of the target and candidate loops can be included in the search process (**Figure 4.8**). Additionally, options to search with the retention of amino acid sequence or conformation at a particular position(s) of the target loop will help to retrieve the loop candidates retaining the specified conformation of residues at the defined position(s) within the loop. **Figure 4.10** and **4.11** demonstrates the outcome of using this option. In **Figure 4.9**, candidate loops for RREG are retrieved without constraining the first two positions of RREG to "TT" conformation. While in **Figure 4.10**, only those four residue loops with the first two positions restricted to "TT" are retrieved. Further, the user can also avail the option of restricting the search to a particular class of proteins from the options given under the protein class wise dataset. In summary, the results present the loop conformations from the database based search criteria with sequence and structural level information. The target and candidate loops can be visualized, compared and analyzed using diverse rendering options, fit/overlay structures, and compare the hydrogen bonding pattern, polar/nonpolar environment, and Ramachandran ϕ, ψ plot for the residues of the target and loop candidates (**Figure 4.9 B**).

The level 2 query page is developed using HTML (for web page output), CSS (to create buttons and tables), Perl (as CGI backend) & Python (as search tools/algorithms). And the results page is developed using HTML (for web page output), CSS (to create buttons and tables), Perl (as CGI backend), Python (as search tools/algorithms) & SQL (to query and retrieve data from database). Due to script limitation, only top 50 results can be displayed in the results table hence to give access to complete results an option to download complete results as a text file is also provided.

Query PDB ID:	1ttp
Loop sequence selected:	RREG
Loop Conformation (as assigned by DSSP):	TT--
Starting residue of the loop:	14
Ending residue of the loop:	17
Chain ID:	A
RMSD & Environment (V-score) based search:	◉
RMSD (Angstrom) based search:	○
RMSD cutoff in Angstrom (default is 1):	1 (ex:0.5)
Restrict amino acid sequence at specific position(s) (or) anywhere in the loop (optional):	(ex:_XX_) (or) (ex:%XX%) (or) (ex:%X_X%) (hover on ? icon for more information)
Restrict conformation at specific position(s) (or) anywhere in the loop (optional):	(ex:_XX_) (or) (ex:%XX%) (or) (ex:%X_X%) (hover on ? icon for more information)
C-Alpha END-END distance based search:	○
+/- Distance cutoff in Angstrom:	(ex:0.5)
Select Protein Dataset :	Select Dataset Retrieve only representatives from structures at ◉ **30%** ○ **90%** sequence identity

GO Reset

Figure 4.8: *Level 2 of LoopX showing selected loop information and different search criteria options.*

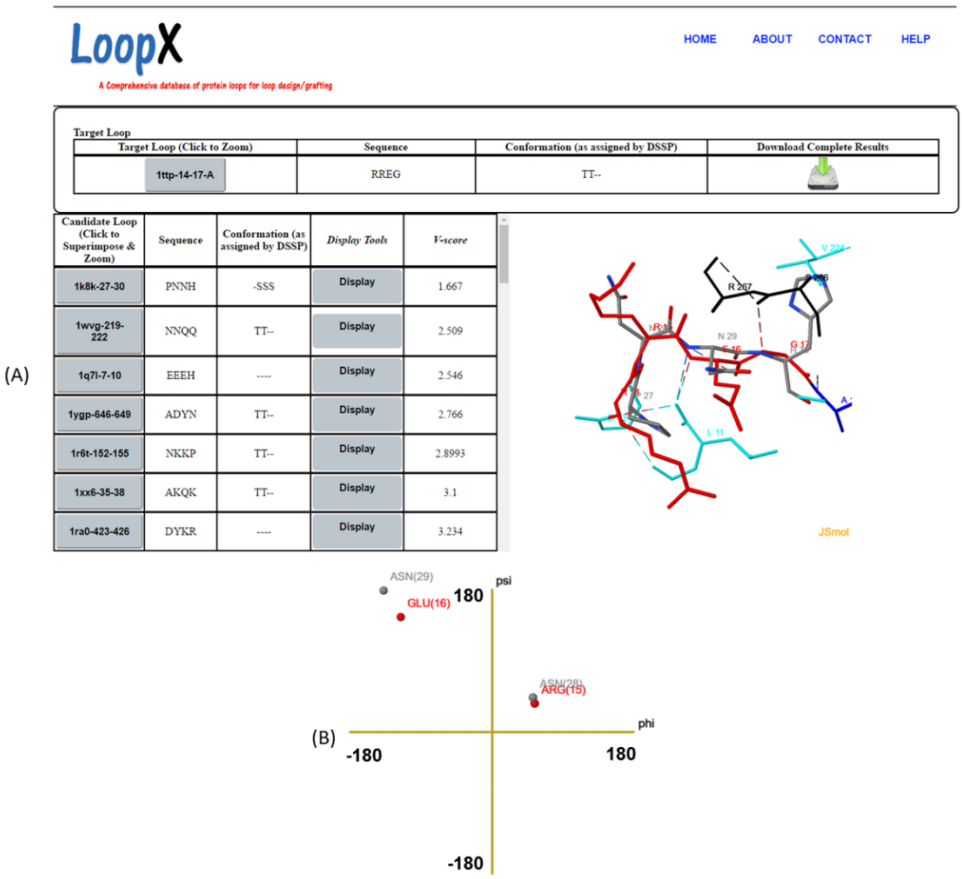


Figure 4.9: A*. A screenshot of the LoopX output displaying the summary of the extracted loops for a chosen target loop (target loop, RREG from 1ttp, 14-17). The display on the right, shows the overlay, hydrogen bonding interactions (dotted lines) of the target (red) and one of the extracted loop candidates (grey (PNNH).* ***B****. Superimposition of the backbone ϕ, ψ dihedral angles of the target (red) and candidate loop residues (PNNH; grey) on Ramachandran ϕ, ψ plot. The amino acids are indicated as three letter codes along with their residue number in parentheses*

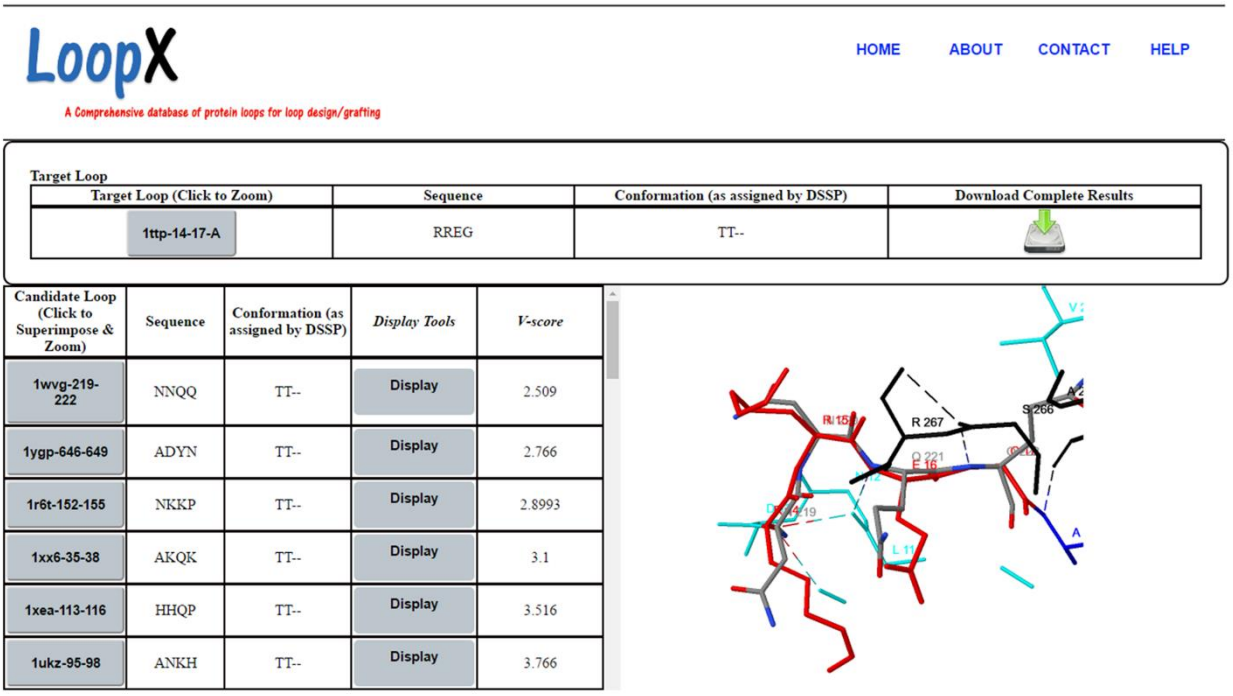


Figure 4.10: *A screenshot of the LoopX output showing an example of the extracted loops for a chosen target loop (target loop, RREG from 1ttp, 14-17) using the secondary structure conservation option. The extracted loops shows the conservation of a turn (TT--) conformation similar to target loop. The display on the right, shows the overlay, hydrogen bonding interactions (dotted lines) of the target (red) and one of the extracted loop candidates (grey (NNQQ).*

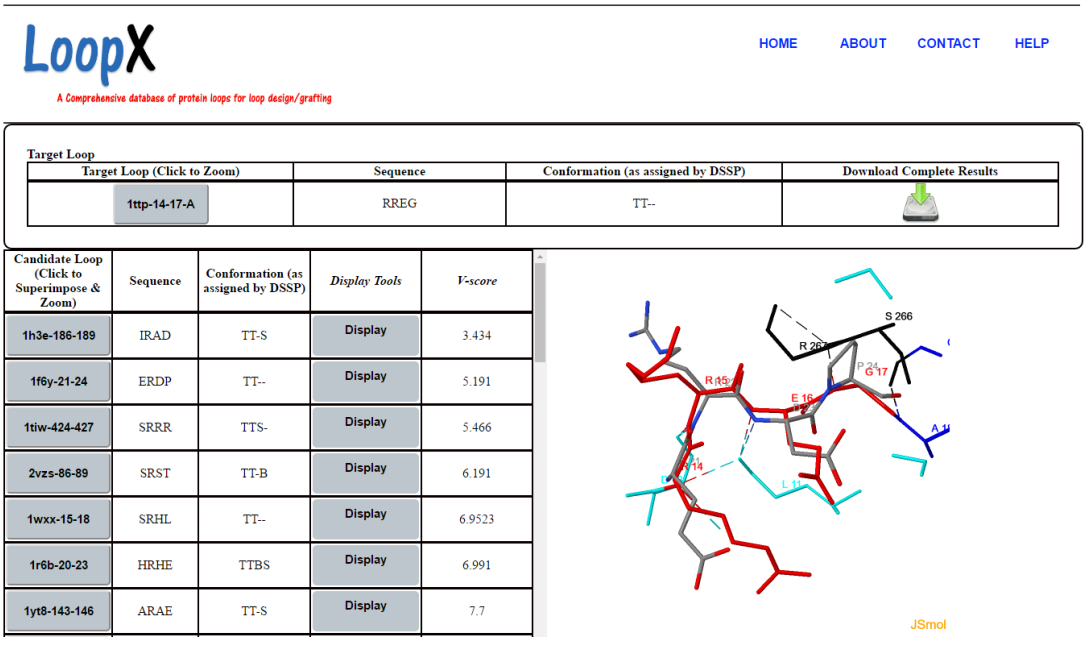


Figure 4.11: *A screenshot of the LoopX output showing an example of the extracted loops for a chosen target loop (target loop, RREG from 1ttp, 14-17) using the sequence conservation option. The extracted loops show the conservation of arginine (-R--) at second position similar to target loop. The display on the right shows the overlay of the target (red) and one of the extracted loop candidates (grey (ERDP).*

Loops retrieval and analysis: A case study

To demonstrate the efficiency of the database, a recent study on loop exchanges in TIM barrel protein reported by Ochoa et al (Ochoa-Leyva et al., 2011). In their study, loop 6 of phosphoribosyl anthranilate isomerase, a TIM barrel protein (PDBID: 1PII), was swapped with a set of 4 loops from diverse proteins. The selected loop candidates for swapping contained sequence identities between 0-30%, (**Table 4.1**) and cover a range of structural, functional and evolutionary relationships. Experimental evaluation of the study showed that the loops with 27, 25 and 17 percent sequence identity when swapped in place of the original loop, the resulting mutant proteins displayed foldability with decent functional activity. Using LoopX to retrieve compatible candidates for loop 6 of phosphoribosyl anthranilate, a database search was performed using Cα end-to-end distance criteria. In our search results two out of the three loops that showed the best structure and functional adaptability from the experimental results are retrieved (see **Table 4.1** and **Figure 4.12**).

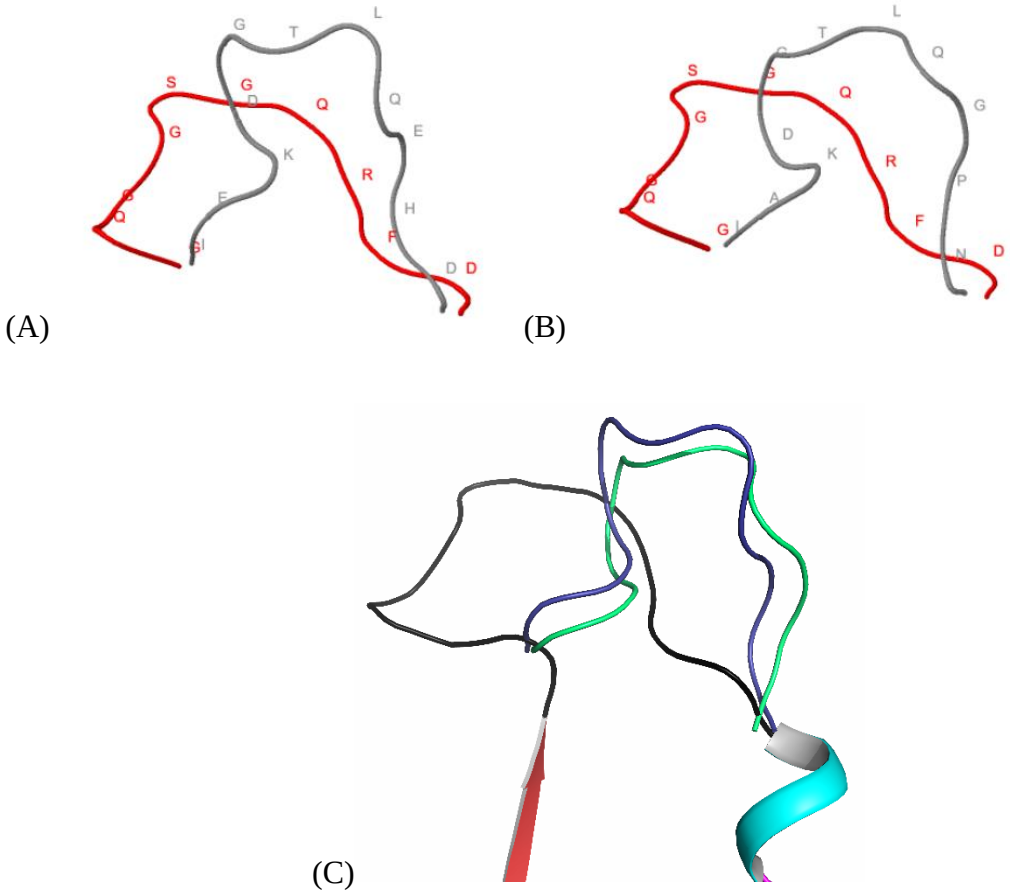


Figure 4.12: A*. Superimposition of target ((red) loop 6 from 1PII) and extracted candidate loop ((grey)IEKDGTLQEHD).* ***B****. Superimposition of target ((red) loop 6 from 1PII) and extracted candidate loop ((grey) IAKDGTLQGPN)* ***C****. Superimposition of target ((red)loop 6 from 1PII) and extracted candidate loops ((green) IAKDGTLQGPN, (blue) IAKDGTLQGPN)*

S.no	Sequence of target loop from phosphoribosyl anthranilate isomerase used in the previous experimental study	Candidate sequences used for swapping in the experimental study along with percentage of target-candidate sequence similarity	Sequence of target loop from phosphoribosyl anthranilate isomerase used in current study	Sequences fetched using LoopX
1	NGQGGSGQRFDW	DIAKDGTLQGPN (17%), EKDGTLQEHDF (27%), RAGVTGAENRAALP (25%), LILPANVA (0%)	NGQGGSGQRFDW	IAKDGTLQGPN, IEKDGTLQEHD

Table 4.1: *Comparison of compatible loop candidates extracted using LoopX for target sequence (*NGQGGSGQRFDW*) with loop candidates selected in experimental data.*

Discussion

LoopX database is specially developed with flexible query tools to aid loop grafting/design, by identifying compatible loop targets. The search tools employed in the database helps to identify target loops with similar backbone conformation or loops with similar Cα end-to-end conformational space. The environment charge based search algorithm along with RMSD aids in identifying loops with similar backbone and environment to preserve conformation of the loop and avoid side chains steric clashes of loop upon grafting/design, resulting in incorporation of new features without compromising the conformational integrity. However, the "V-Score" based search algorithm still lacks the feasibility of extracting loops with specific non-covalent interactions. The amino acid sequence conservation based search helps to retain important amino acid residues involved in the catalytic activity, active pocket formation or binding in target based loop designs, and the secondary structure conservation based search helps to retain vital turn conformations and bends that play a key role in folding and binding. We strongly believe that LoopX will reduce the taxing work of data mining for loop grafting, designing and re-engineering to alter the folding, binding and functional features of protein candidates.

Chapter 5

Stable βαβ modules from TIM barrel proteins

The acquired complex three-dimensional conformations of proteins is a culmination of simple structural fragments like α–α, β–β, α–β and β–α units (Grishin, 2001). The precise sequence and formation of local secondary /super secondary structures such as α–α hairpins, β hairpins, βαβ units and their optimal packing between them lead to thermodynamically stable tertiary structures. Repetition of these super secondary motifs lead to the formation of α-helical, β sheet β/α and α+β class of proteins. Thus, tertiary structures can be seen as a combination of basic building block motifs. Experimental observations clearly indicate that the formation of protein tertiary structure is preceded by the formation of native local structures. Such folding initiation sites populated along the polypeptide sequence can restrict the conformational search and guide the overall folding process (Baldwin & Rose, 1999; Lewandowska et al., 2010; Ramakrishna & Sasidhar, 1997; Shin et al., 1993; Wright et al., 1988). Interestingly, it may be noted that the divergence in domain architecture of proteins has been achieved by the combinatorial assembly of smaller gene fragments to form intrinsic stable subunits of protein domains (Lupas et al., 2001; Soding & Lupas, 2003). In recent studies from the comparison of domains from different proteins, it became obvious that the complex protein structures have evolved from a simple independently folding smaller super secondary fragments (Orengo & Thornton, 2005; Soding, Biegert, & Lupas, 2005). In a study conducted by Riechmann and Winter it was observed that new stable proteins can be designed by combinatorial assembly of fragments with a defined peptide (Riechmann & Winter, 2000, 2006). Thus, observation of secondary and super-secondary structures in isolated peptide fragments form proteins and in de novo

designed sequences underscores the role of independently folding domains and their role in assembly of higher order structures (Blanco et al., 1994; Cochran et al., 2001; Ihalainen et al., 2008; Liang et al., 2009; Marqusee et al., 1989; Petukhov et al., 2009; Ramakrishna & Sasidhar, 1997; Religa et al., 2007; Sadqi et al., 2009; Searle et al., 1995; Stanger et al., 2001; Yakimov et al., 2014; Zeng et al., 2016). Hence the design and or identification of small stable folding units from proteins can potentially assist not only in decoding the guiding principles of protein folding and stabilization but more importantly in contributing to the design and creation of proteins/enzymes with altered or new functions.

In this direction, the past decade has witnessed efforts from various research groups put into the design, discovery, and analysis of sequences that can adopt structures and acquire native-like stable conformations. Knowledge gleaned from the folded tertiary structures of proteins shed light on the role of (i) amino acid preferences to occur in helical, β sheet and reverse turn conformations, (ii) N and C-capping interactions in helical and β sheet structures, helped decipher the role of local interactions in folding and stability of secondary and super-secondary structural units including α helices, βsheets, and βhairpins, ββα motifs. Plethora of studies on peptide fragments that can fold into independent motifs like helix-loop-helix, beta-beta-alpha and beta sheets etc. were aggressively pursued leading to the accumulation of vast information on sequence and structural features contributing to the folding and stability of protein building block motifs (Blanco et al., 1994; Cochran et al., 2001; Dahiyat & Mayo, 1997; Ihalainen et al., 2008; Liang et al., 2009; Marqusee et al., 1989; Petukhov et al., 2009; Religa et al., 2007; Sadqi et al., 2009; Searle et al., 1995; Stanger et al., 2001; Struthers et al., 1996; Yakimov et al., 2014; Zeng et al., 2016).

Given the knowledge of first principles of folding, stability, and sequence to structure relationship obtained from design and identification of independently folding units, further encouraged researchers to design/build completely new proteins (Eisenberg et al., 1986;

Fezoui, Weaver, & Osterhout, 1994; Harbury et al., 1998; Huang et al., 2016; Kamtekar & Hecht, 1995; Koga et al., 2012; H. Li, Helling, Tang, & Wingreen, 1996; Regan & DeGrado, 1988; Struthers et al., 1996). Initially, small α/β proteins, Rossman fold and 4 helix bundles of approximately 100 residues in length (Eisenberg et al., 1986; Harbury et al., 1998; Kamtekar & Hecht, 1995; Regan & DeGrado, 1988) were designed using computationally derived principles (Eisenberg et al., 1986; Harbury et al., 1998; Kamtekar & Hecht, 1995; Koga et al., 2012; Regan & DeGrado, 1988), but designing of larger proteins has been very challenging. However, recently many attempts were made to build new scaffolds using fragments from existing proteins like HisF, chemotactic Y, HisA, OmpX, TrpF and NarL. Attempts were also made to build new proteins by combining stable structural fragments from various proteins (Hocker, 2014; Nagarajan et al., 2015; Soding & Lupas, 2003).

TIM barrel fold belonging to the α/β class of proteins is a very frequent/common enzymes observed in more than 10 percent of the known protein structures. TIM barrel proteins are known to occur in five of the six enzyme classes. The TIM barrel proteins are made up of a regular repeating βαβ motif which acts as a building block, comprising a total of 8 strands and helices arranged in an alternating repetitive pattern. The repetitive arrangement of βαβ motif in TIM barrel fold forms a stable inner core made up of beta strands shielded by amphipathic alpha helices. The loops in TIM barrel fold are categorized into α/β and β/α loops as described earlier. The αβ loops connect α helices and the βstrands which are believed to be involved in stability while the βα loops connect β strands to the α helices participate in the enzymatic function. TIM barrel proteins have been the target of interest for protein manipulations and designing studies (Huang et al., 2016; Koga et al., 2012; Nagarajan et al., 2015; Ochoa-Leyva et al., 2011; Ochoa-Leyva, Montero-Moran, Saab-Rincon, Brieba, & Soberon, 2013; Ochoa-Leyva et al., 2009).

Experimental and theoretical studies have revealed that βαβ unit acts as a minimal unit of stability (Frenkel & Trifonov, 2005; Nagano et al., 2002; X. Yang et al., 2009; Zitzewitz et al., 1999). In a very recent study carried out by Li et al, it was observed that a de novo designed peptide sequence of 38 amino acids long, adopted a well folded and stable βαβ conformation in aqueous solution.(H. Li et al., 1996). In their design strategy, they have included the basic rules in design of α helices and β sheets from observations obtained from the survey of protein structures and βαβ units. Secondary structure preferences of amino acids, local (short range) ionic interactions between side chains of amino acids, capping preferences were included in the design of the βαβ motif. Sequence optimization from the initial sequence resulted in a successful design of the following sequence: GSGQVRTIWVGGTPEELKKLKEEAKKANIRVTFWGD (strand and helix regions of sequence are colored in red and cyan respectively), that adopted into well folded and stable βαβ structure in isolation. Attempts were also made to de novo design hyper stable super secondary structures with constraints using computational approaches (Bhardwaj et al., 2016).

The above argument underscores the success of designing and or identification of independently folding motifs either de novo designed or from the existing protein structures, However, intriguingly, so far with the exception of a couple of denovo designed βαβ units, (Bhardwaj et al., 2016; Liang et al., 2009) naturally occurring βαβ sequence from proteins that folds independently has not been identified, although other peptide sequences from proteins that adopt secondary and super-secondary structures have been observed to fold in isolation (Blanco et al., 1994; Lewandowska et al., 2010; Marqusee et al., 1989; Searle et al., 1995; Shin et al., 1993); but not a βαβ sequence so far.

In this chapter we address the finding of the needle(s) in haystack scenario, i.e., proposing/identifying likely βαβ candidates from the existing TIM barrel proteins that can fold independently. As the first step in this direction, we have taken the de novo designed βαβ sequence that is stable in isolation(Liang et al., 2009) and explored the features within it as a benchmark model.

Firstly, we assessed the propensity of the helix in the designed sequence GSGQVRTIWVGGTPEELKKLKEEAKKANIRVTFWGD, using AGADIR (Lacroix, Viguera, & Serrano, 1998; Munoz & Serrano, 1995a, 1995b, 1997), a program that predicts the propensity of a sequence to adopt a helical conformation. The program gives the percentage helicity of a given sequence under specified conditions of pH, ionic strength and temperature. The prediction from AGADIR is based on the experimentally observed percent helicity of a large number of peptide sequences that actually adopted helices in solution (Lacroix et al., 1998; Munoz & Serrano, 1995a, 1995b, 1997). It was observed from AGADIR that the part of the sequence that adopts helix conformation in the designed βαβ motif shows a very high percentage of helicity (38.36%). Therefore, a high helix propensity is considered as one of the requirements for the βαβ folding. Following this, we have assessed the foldability/folding tendency of the entire βαβ motif using Monte Carlo, REMD and Molecular dynamics simulations based approaches. Computational approaches like replica exchange and monte carlo simulations can shed more light on foldability and stability of peptide/proteins (Ho & Dill, 2006; R. Zhou, 2007). Foldability/structure prediction studies carried out by Dill and his coworkers on 133 peptide 8-mer fragments demonstrates the power of replica exchange molecular dynamics approaches in protein structure prediction (Ho & Dill, 2006). QUARK and PEPFOLD are the two monte carlo and replica exchange based methods to assess the foldability of protein sequences. This was further followed by subjecting the βαβ sequences identified from the above prediction to

unfolding molecular dynamics simulations. The longer the persistence of the structure in the unfolding simulations the more likely is its stability. The observations from these approaches for the βαβ sequences, were benchmarked for consideration in identifying the βαβ sequences from TIM barrels proteins that show similar behavior to that of the stable designed βαβ sequence. The likely βαβ candidates that could potentially fold into stable and well folded structures were explored based on certain sequence/structural features like loop length, propensity to fold into a helix, and stabilizing interactions, specifically the main chain to side chain hydrogen bonding interactions that clamp the βαβ units and play a significant role in stability of the fold. The details of the entire approach are presented below.

Methods

Identification and segregation of βαβ units from TIM barrels

A non-redundant dataset of 420 TIM barrel PDB (H.M. Berman et al., 2000) structures with < 30% sequence similarity and a structural resolution of ≤ 3.0 Å were shortlisted for this study. DSSP program (Joosten et al., 2011; Kabsch & Sander, 1983) was used to assign the secondary structure for the protein structures from the dataset. Based on the secondary structure information provided by DSSP program the consecutive beta-alpha-beta regions were identified, extracted and saved as individual βαβ units, using in-house python scripts. From this set of βαβ units, αβ and βα loops of ≤ 14 residues long were segregated for further analysis, using in-house built python scripts.

Alpha-helical propensity prediction

The propensity of a sequence to fold into a helical conformation has been predicted using AGADIR prediction algorithm. The AGADIR prediction program developed by Luis Serrano and Michael Petukhov (Lacroix et al., 1998; Munoz & Serrano, 1995a, 1995b, 1997) is a helix content prediction algorithm based on helix/coil transition theory and evaluated against the available experimental data. AGADIR calculates the helical propensity (% helicity) by taking into account the temperature and pH conditions of the peptide. Helical sequence regions from the shortlisted1608 βαβ units and from the designed βαβ, DS119 (PDB ID: 2KI0) are acetylated and amidated at terminal regions were considered for AGADIR prediction at temperature 25^{o}C and pH 7.

Foldability and stability assessments of βαβ candidates

Monte Carlo/Molecular dynamics simulations are further used to assess the foldability and stability of likely βαβ candidates that can fold independently. The likely candidate βαβ sequences shortlisted from AGADIR prediction for % helicity, and along with control sequence (2KI0) were considered for QUARK (D. Xu & Zhang, 2012) and PEPFOLD3 (Lamiable et al., 2016; Shen, Maupetit, Derreumaux, & Tuffery, 2014; Thevenet et al., 2012) folding prediction. These two approaches will predict the likeliness of peptide sequences to fold into independent βαβ units.

QUARK is an algorithm developed by Zhang lab for the abintio protein folding/protein structure prediction. QUARK can predict the likely foldable 3D models for the provided protein sequence of less than 200 residues long using replica-exchange Monte Carlo simulations (REMC)., the REMC simulations are guided by an optimized atomic-level knowledge-based force field (D. Xu & Zhang, 2012). QUARK provides 10 best likely

structure models which are selected based on clustering using revised SPICKER program (Zhang & Skolnick, 2004b) and template modeling score (TM-score) of models generated from the REMC simulations (Zhang & Skolnick, 2004a).

PEPFOLD3 is a de novo approach for predicting peptide structure for amino acid sequences of length 5-50 residues long, PEPFOLD3 works based on Hidden Markov Model sub-optimal conformation sampling approach (Altschul et al., 1997), describing the polypeptide chain conformation using a series of local overlapping canonical conformations of 4 residue amino acids fragments. PEPFOLD3 provides 5 best structure models for the given amino acid sequence based on sOPEP energy (optimized potential for efficient structure prediction) (Sippl, 1990).

Finally, unfolding simulations are used to assess the persistence of the shortlisted βαβ structures to unfolding as a function of time. Longer the retention of structural features, greater the stability of the particular βαβ module. Thermal denaturation of βαβ units was simulated at 300 and 400 K. All-atom Molecular dynamics simulations were carried out using GROMACS 5.1 molecular dynamics simulation software using Gromos43a1 force field (Abraham & and Lindahl, 2015; Van Der Spoel et al., 2005; van Gunsteren, 1996). The βαβ units were solvated in a cubic box with a distance cutoff 12 Å between the edge of the periodic box and surface of the motif with water molecules, respectively. The SPC (Simple Point-Charge) model was used for water molecules and Na^+ Cl^- counter ions were added in required simulations to neutralize the system. The system was subjected to energy minimization using steepest descent algorithm down to a 1000 kJ/mol/ nm till the energy get converged. Before Production run all the systems were equilibrated for temperature and pressure equilibration of the system for 100ps using canonical NVT and NPT ensembles at respective temperatures. The long-range electrostatic interactions are calculated using

Particle Mesh Ewald (Darden et al., 1993; Essmann et al., 1995) (PME) method with a grid spacing of 0.16 nm and a cut-off of 1.0 nm was used for short-range electrostatic and van der Waals interactions, bond lengths were constrained using LINCS algorithm (Hess, 1997). All simulations were performed using a 2-fs integration time step, with a coupling coefficient of tT – 0.1 ps using modified Berendsen thermostat (Berendsen et al., 1984), and Parrinello-Rahman pressure-coupling at 1bar with a coupling coefficient of tP = 1 ps. The simulations were carried out at 300 and 400 K simulated temperatures for 5 nanoseconds for the final likely βαβ units to access the stability by studying the unfolding evolution.

The overall approach is summarized below:

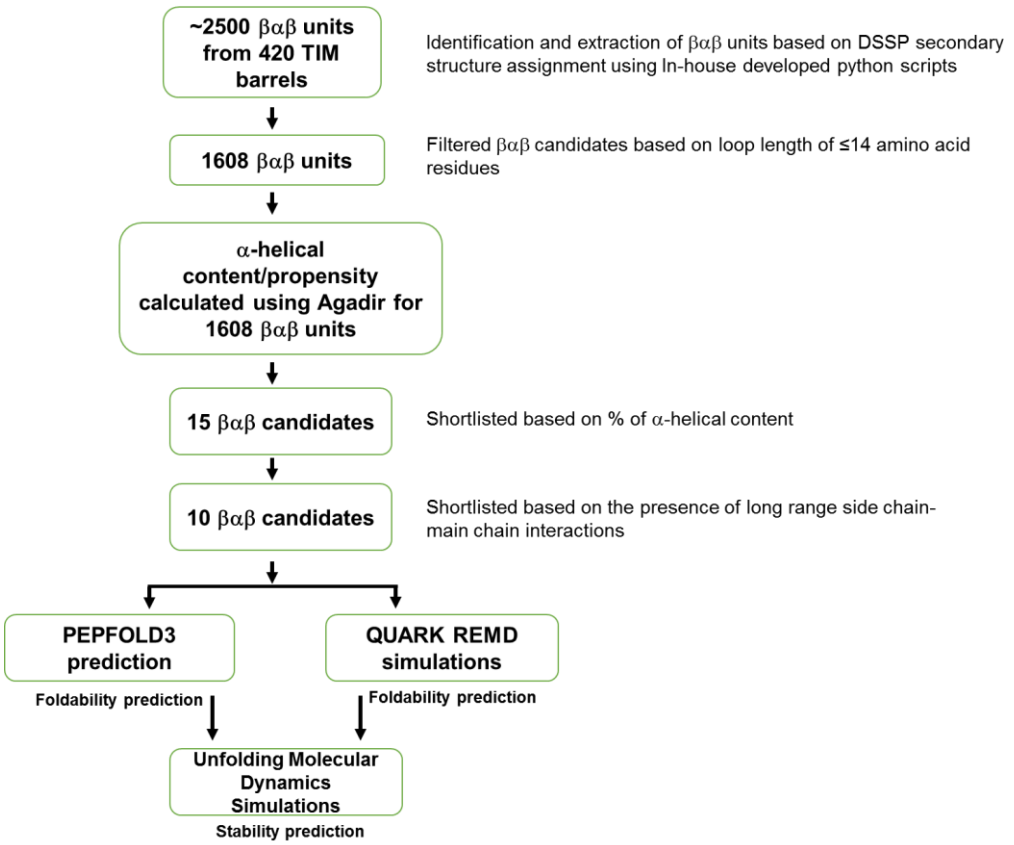


Figure 5.1: *Schematic representation of likely βαβ candidate's selection*

Results and Discussions

Identification and segregation of βαβ units

A total of 420 Non-redundant TIM barrel proteins were mined from Protein Date Bank with < 30% sequence similarity and structural resolution of ≤ 3.0 Å. Approximately 2500 βαβ units were extracted from 420 non-redundant TIM barrel proteins using in-house developed python scripts which identified and extracted the units based on the secondary structure information calculated by DSSP program. Following this, in order to identify the βαβ units with a potential to independently fold certain filters were used. As a first filter, a loop length restriction has been employed on βα and αβ loops connecting the beta-alpha and alpha-beta units of βαβ motif. Our studies as discussed in this thesis, shorter loops, particularly, αβ loops are important for stability. Long loops with higher conformational entropy may be detrimental for folding and stability. Therefore, βαβ units with αβ and βα loops less than or equal to 14 residues long are only considered for further process. A total of 1608 βαβ units are identified at the end of this step for further selection criteria.

Alpha-helical propensity prediction

The second criterion employed is the helical propensity. As discussed earlier that the designed βαβ peptide showed a high helix propensity for its helical region and therefore this is considered as one of the requirements for βαβ folding. The α helical regions of the 1608 βαβ peptide sequences were estimated for helix propensity (helix forming tendency) using AGADIR prediction algorithm. The predicted helix propensity was then compared with the AGADIR indicted helical propensity of the helix region of the denovo designed βαβ motif (DS119; PDB ID: 2KI0). From the AGADIR prediction, it was observed that the lowest AGADIR helical propensity score predicted for a βαβ helix sequence is 0.01%

whereas the highest AGADIR helical propensity score predicted is 53.14% at pH 7.0 and at 25^{o}C temperature. (**Figure 5.2**). The helical region of DS119 has been de novo designed based on the known principles of folding and been experimentally observed to fold. Therefore, the AGADIR helical propensity of DS119 has been used as a benchmark score to select the βαβ sequences with good helical propensity. For DS119 the AGADIR helical propensity calculated was 38.36% at pH 7.0 and at 25^{o}C temperature. In order to select the βαβ sequences with similar or better helical propensity than DS119 a helical propensity cutoff value of minimum 38.36% needs to be employed. The cut-off value has been relaxed to 28.77% to consider the βαβ sequences with at least 75% helical propensity that of DS119. For further analysis for likely βαβ candidate selection, AGADIR helical propensity of minimum 28.77% has been used as a cutoff. This resulted in 13 βαβ candidates showing AGADIR alpha-helical propensity value of equal to or greater than 28.77% (**Table 5.1**).

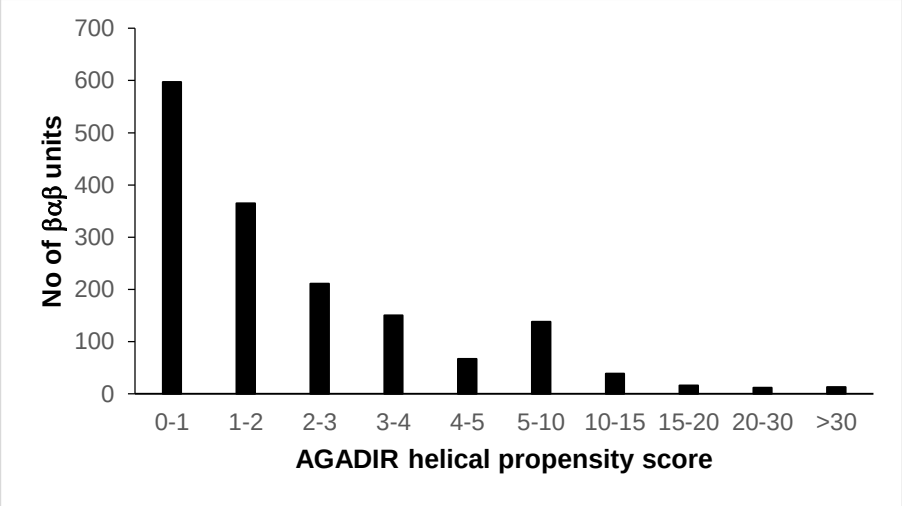


Figure 5.2: *Distribution of βαβ units based on their AGADIR helical propensity score.*

AGADIR prediction (%)	βαβ Sequences	PDBID	Presence of clamp interactions (long range side chain-main chain interaction)
53.14	PQAVFNVELKSFPGLGEEAARRLAALLRGREGVWVSSFDP	1VD6	Yes
38.42	GRSLFNSAKVDEEELEMKINLLKKYGGTLIVLLMG	3BOF	Yes
37.9	EVDGYRIDHIDGLFKPEEYLRRLKNKIGNKHIFVEKI	3HJE	Yes
37.88	NYVQVYVMLPLDAVSVNNRFEKGDELRAQLRKLVEAGVDGVMVDVWGL	2XFR	Yes
37.33	YIIEIKFSLIHFKNIPLEDLLLFIKAWANFAQKNKLDFVVEGIETK	3KZP	No
36.59	HGKRPYASLLFGDTPQETLERARAARRDGFAAVKFGWGP	2HZG	No
35.81	GFEIIITNSYIIYKDEELRRKALELGIHRMLDYNGIIEVDSGS	1IQ8	No
34.70	FADVLLDGEVTEASILRKLDDLERIARRNGQAIGVASAFD	2QV5	Yes
34.43	GQTTFKWKVGVMSPEEEQAILKALLAALPPGAKLRLDANGSWD	2OZT	Yes
33.95	GIVLNGENALSIGNEEEYKRVAEMAFNYNFAGFTLLRY	1VEM	Yes
33.21	AKPFDFVFHGGSGSLKSEIEEALRYGVVKMNVDTD	3ELF	Yes
32.76	HPLVGGLILFTRNYHDPEQLRELVRQIRAASRNHLVVAVDQEGGRV	4GVF	Yes
30.77	YPKFWGRYLSEVPNVSEGLTRDEIVRIRNYGVKVLLPIYNAAF	1SFS	Yes
38.36	GSGQVRTIWVGGTPEELKKLKEEAKKANIRVTFWGD	2KI0 (Designed sequence)	No

Table 5.1: *List of the likely βαβ candidates with predicted AGADIR helical propensity score and clamp interactions (long range side chain-main chain interaction), the de novo designed DS119 (PDBID: 2KI0) sequence is also included in the list. The helix and strand regions of the sequence are colored in cyan and red respectively.*

Long range side chain – main chain interactions analysis (clamps)

The long range side chain to main chain interactions are known to play a crucial role in stabilizing the TIM barrel proteins (X. Yang et al., 2009; X. Yang et al., 2007). The 13 selected βαβ units are further manually checked for the presence of long range side chain to main chain interactions using pymol (Schrodinger, 2015) molecular visualizer. It was observed that out of 13 only 10 βαβ are harboring clamps. These 10 βαβ units are short listed for further stability and foldability assessment using replica exchange molecular simulations and molecular dynamics unfolding simulation approaches.

Foldability assessments using QUARK and PEPFOLD prediction

The identified 10 βαβ units were analyzed for foldability. The foldability assessment has been carried out on final 10 βαβ candidate sequences along with DS119 sequence as a control. To assess the structure predictions, four βαβ sequences with AGADIR helical propensity score of 0.2, 2, 5 and 10 % were also submitted for foldability prediction as controls.

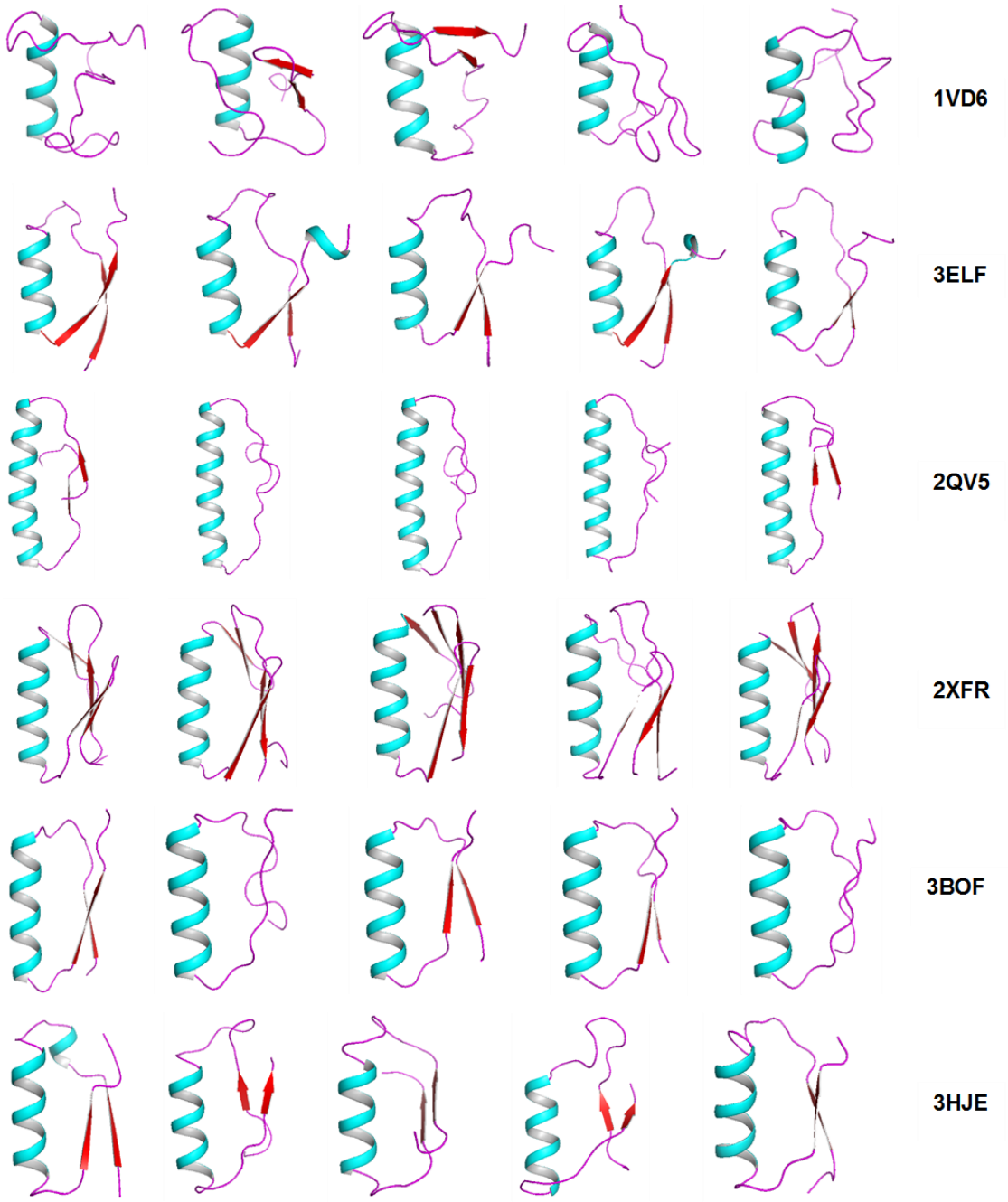
QUARK Fold prediction assessments

For a given peptide sequence QUARK algorithm predicts 10 best possible protein structure models. All the models are built from small fragments of 1-20 residues long without using any global template followed by assembling of fragments using 10 different replica exchange Monte Carlo (REMC) simulations each with around 200 cycles and the simulations are run with the help of an atomic-level knowledge-based force field. Finally, the models generated in REMC simulations are clustered using revised SPICKER program to select best top 10 models assisted by template modeling score function (TM-score).

In our current study, the final 10 βαβ candidate sequences identified, along with a sequence of DS119 (control (2KI0)) are subjected to simulations using QUARK server. The QUARK predictions for submitted sequences are shown in **Figure 5.3**. Although 10 submitted sequences have shown good helical region, only 4 (3ELF, 3HJE, 2OZT and 4GVF) out of 10 sequences submitted to QUARK showed more than 50% chances of foldability into proper βαβ units i.e. at least 5 out of 10 output structures were predicted to be folding into βαβ units (**Table 5.2**). And as for the control sequences with 0.2, 2, 5 and 10 AGADIR helical propensity scores, no proper βαβ structure was predicted by QUARK. However, it may be noted that for the de novo designed control βαβ sequence it was observed that 10 out of 10 structures predicted were βαβ units (**Figure 5.4**) suggesting the efficacy of the design by first principles of folding.

βαβ sequence submitted	AGADIR prediction(%)	Source PDB ID	βαβ Predictions out of 10
PQAVFNVELKSFPGLGEEAARRLAALLRGREGVWVSSFDP	53.14	1VD6	0 out of 10
AKPFDFVFHGGSGSLKSEIEEALRYGVVKMNVDTD	33.21	3ELF	7 out of 10
FADVLLDGEVTEASILRKLDDLERIARRNGQAIGVASAFD	34.70	2QV5	1 out of 10
NYVQVYVMLPLDAVSVNNRFEKGDELRAQLRKLVEAGVDGVMVDVWGL	37.88	2XFR	0 out of 10
GRSLFNSAKVDEEELEMKINLLKKYGGTLIVLLMG	38.42	3BOF	4 out of 10
EVDGYRIDHIDGLFKPEEYLRRLKNKIGNKHIFVEKI	37.9	3HJE	5 out of 10
GQTTFKWKVGVMSPEEEQAILKALLAALPPGAKLRLDANGSWD	34.43	2OZT	10 out of 10
GIVLNGENALSIGNEEEYKRVAEMAFNYNFAGFTLLRY	33.95	1VEM	2 out of 10
HPLVGGLILFTRNYHDPEEQLRELVRQIRAASRNHLVVAVDQEGGRV	32.76	4GVF	5 out of 10
YPKFWGRYLSEVPNVSEGLTRDDEIVRRIRNYGVKKVLLPIYYNAAFF	30.77	1SFS	0 out of 10
Control sequences	**AGADIR prediction(%)**	**Source PDB ID**	**βαβ Predictions out of 10**
HQIAWICGTAEKWAPFFWHAGAKGFTSGLV	0.2	3E96	0 out of 10
FDGVIFSDDLSMEGAAIMGSYAERAQASLDAGCDMILVCNN	2	4GVF	0 out of 10
GADVGLLEGFRSKEQAAAAVAALAPWPLLLNSVENG	5	3LYF	0 out of 10
GGSSIKYFPGGLKHRAEFEAVAKACAAHDFWLEPTGG	10	3M0Z	0 out of 10
GSGQVRTIWVGGTPEELKKLKEEAKKANIRVTFWGD (designed sequence)	38.36	2KI0	10 out of 10

Table 5.2: *List of the likely βαβ candidates along with control sequences submitted for QUARK fold predictions. The number of βαβ conformations out of 10 predicted structures by QUARK is also indicated.*


1VD6
3ELF
2QV5
2XFR
3BOF
3HJE

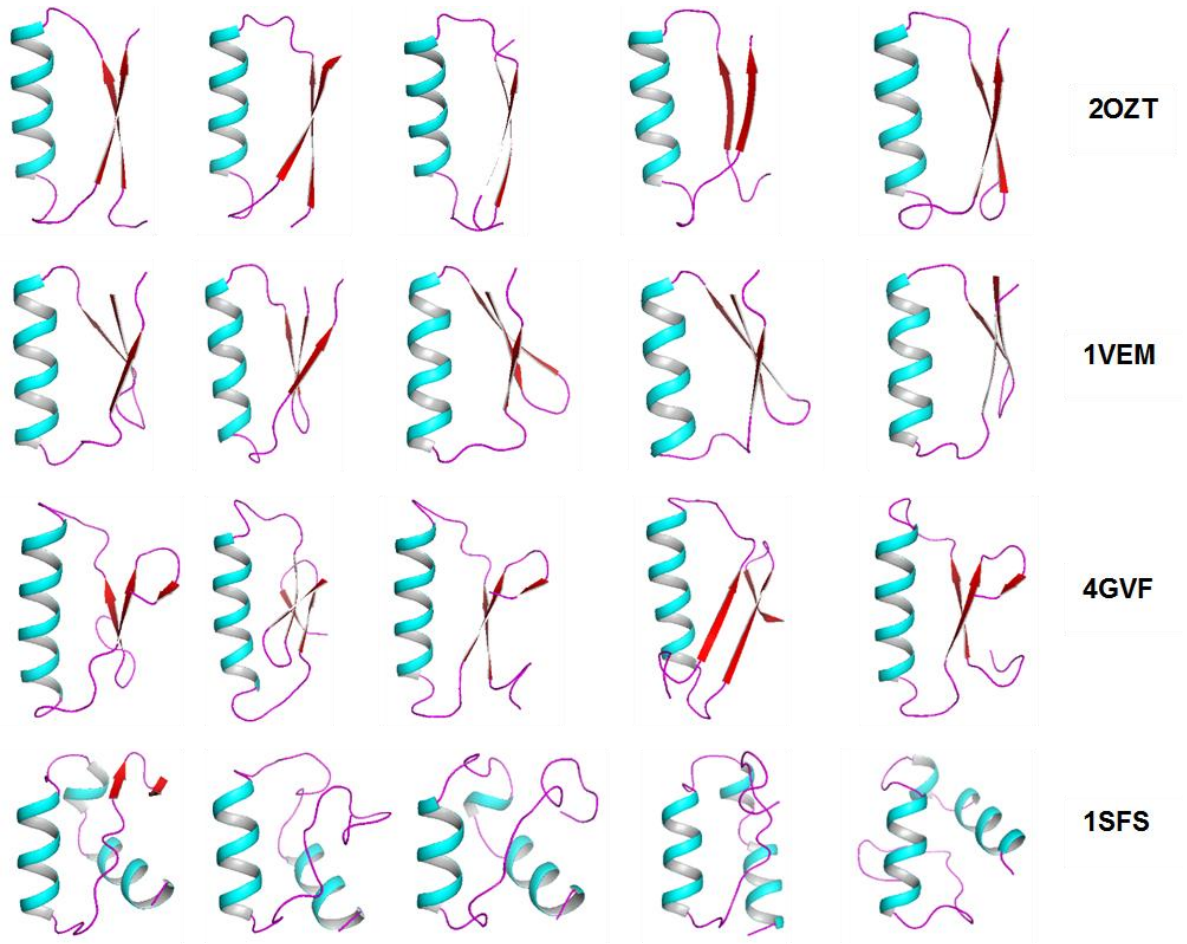


Figure 5.3: *Structural conformations predicted by QUARK for 10 likely $\beta\alpha\beta$ candidate sequences.*

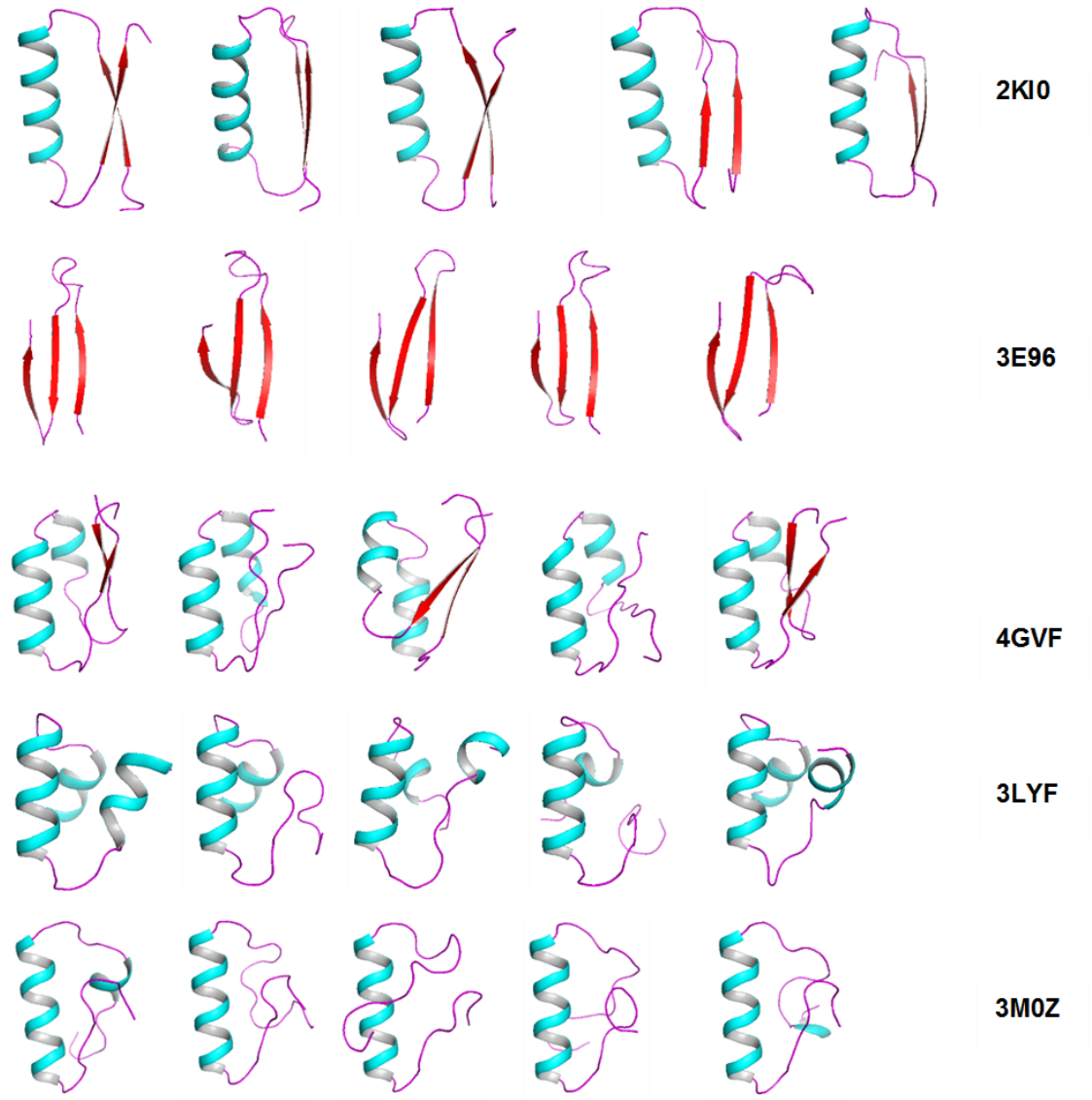


Figure 5.4: *Structural conformations predicted by QUARK for control βαβ candidate sequences.*

PEPFOLD3 Fold prediction assessments

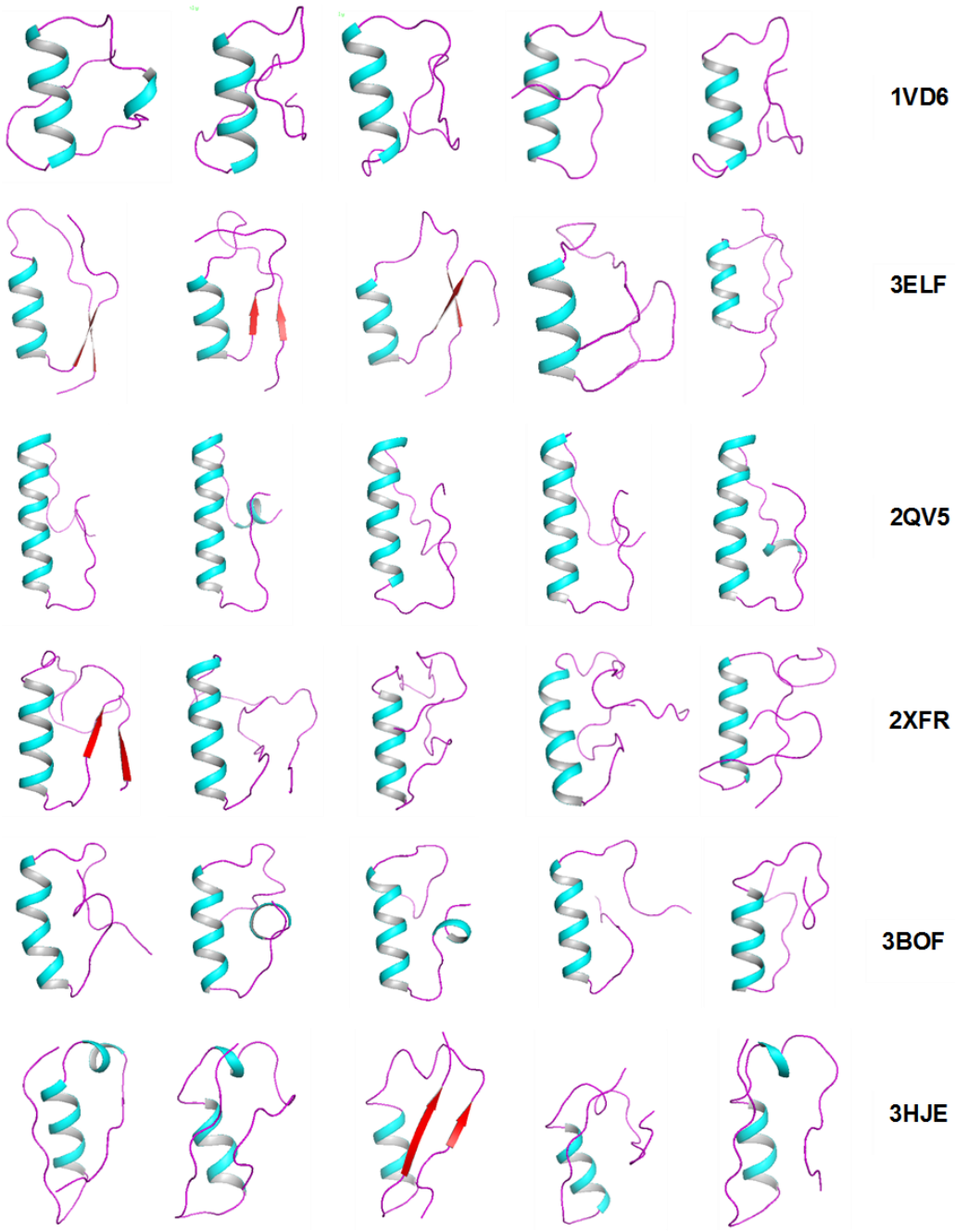
PEPFOLD3 is a de novo based method to predict the structure of amino acid sequences using Hidden Markov Model sub-optimal conformation sampling approach where the polypeptide conformation is obtained using a series of 4 residue fragments local overlapping canonical conformations. In PEPFOLD3 the completed peptide models are refined by 30,000 Monte-Carlo steps and energy minimized using gromacs 5 (Abraham & and Lindahl,

2015) for correct backbone geometry and finally, 5 best models are provided from clustering of models based on sOPEP (optimized potential for efficient structure prediction) energy.

The final 10 likely βαβ candidate sequences identified after long range side chain to main chain interaction feature cutoff were submitted to PEPFOLD3 along with control sequences. It was observed that out of 10 submitted potential candidate sequences only 2 sequences from 2OZT and 3ELF show strong foldability prediction than the designed DS119 sequence; 2KI0 (**Figure 5.5**). Interestingly the designed DS119 sequence also showed only 3 out of 5 predicted models as βαβ units, whereas the control sequences with an AGADIR score of 0.2, 2, 5 and 10 % showed no tendency for βαβ folding (**Figure 5.6**) (**Table 5.3**). It may be noted that for two βαβ sequences (2OZT and 3ELF) both predictions indicate high folding propensity.

βαβ sequence submitted	AGADIR prediction(%)	Source PDB ID	βαβ Predictions out of 5
PQAVFNVELKSFPGLGEEAARRLAALLRGREGVWVSSFDP	53.14	1VD6	0 out of 5
AKPFDFVFHGGSGSLKSEIEEALRYGVVKMNVDTD	33.21	3ELF	4 out of 5
FADVLLDGEVTEASILRKLDDLERIARRNGQAIGVASAFD	34.70	2QV5	0 out of 5
NYVQVYVMLPLDAVSVNNRFEKGDELRAQLRKLVEAGVDGVMVDVWGL	37.88	2XFR	1 out of 5
GRSLFNSAKVDEEELEMKINLLKKYGGTLIVLLMG	38.42	3BOF	0 out of 5
EVDGYRIDHIDGLFKPEEYLRRLKNKIGNKHIFVEKI	37.9	3HJE	2 out of 5
GQTTFKWKVGVMSPEEEQAILKALLAALPPGAKLRLDANGSWD	34.43	2OZT	5 out of 5
GIVLNGENALSIGNEEEYKRVAEMAFNYNFAGFTLLRY	33.95	1VEM	0 out of 5
HPLVGGLILFTRNYHDPEEQLRELVRQIRAASRNHLVVAVDQEGGRV	32.76	4GVF	0 out of 5
YPKFWGRYLSEVPNVSEGLTRDDEIVRRIRNYGVKKVLLPIYYNAAFF	30.77	1SFS	0 out of 0
Control sequences	**AGADIR prediction(%)**	**Source PDB ID**	**βαβ Predictions out of 10**
HQIAWICGTAEKWAPFFWHAGAKGFTSGLV	0.2	3E96	0 out of 5
FDGVIFSDDLSMEGAAIMGSYAERAQASLDAGCDMILVCNN	2	4GVF	0 out of 5
GADVGLLEGFRSKEQAAAAVAALAPWPLLLNSVENG	5	3LYF	0 out of 5
GGSSIKYFPGGLKHRAEFEAVAKACAAHDFWLEPTGG	10	3M0Z	0 out of 5
GSGQVRTIWVGGTPEELKKLKEEAKKANIRVTFWGD (designed sequence)	38.36	2KI0	3 out of 5

Table 5.3: *List of the likely βαβ candidates along with control sequences submitted for PEPFOLD3 fold predictions. The number of βαβ conformation predictions out of 5 predicted structures by PEPFOLD3 is also indicated.*


1VD6
3ELF
2QV5
2XFR
3BOF
3HJE

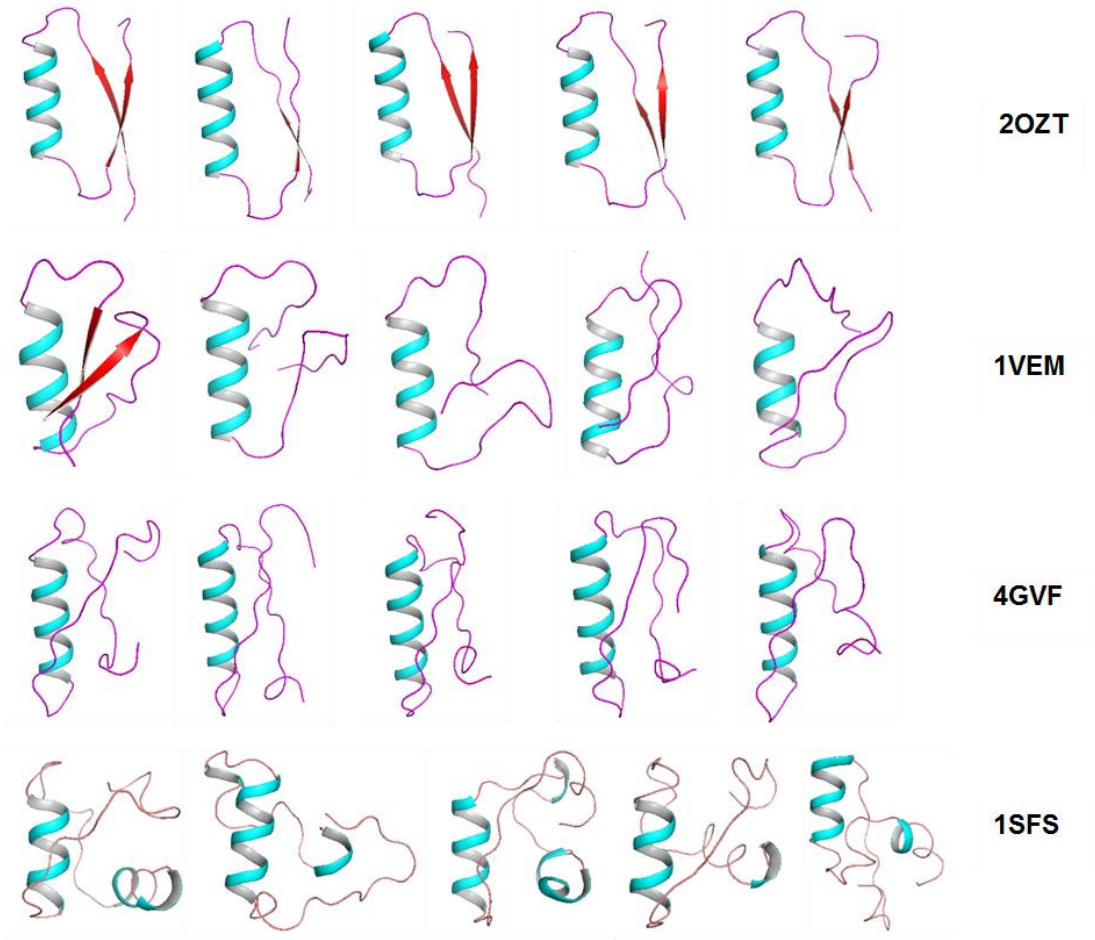


Figure 5.5: *Structural conformations predicted by PEPFOLD3 for 10 likely βαβ candidate sequences.*

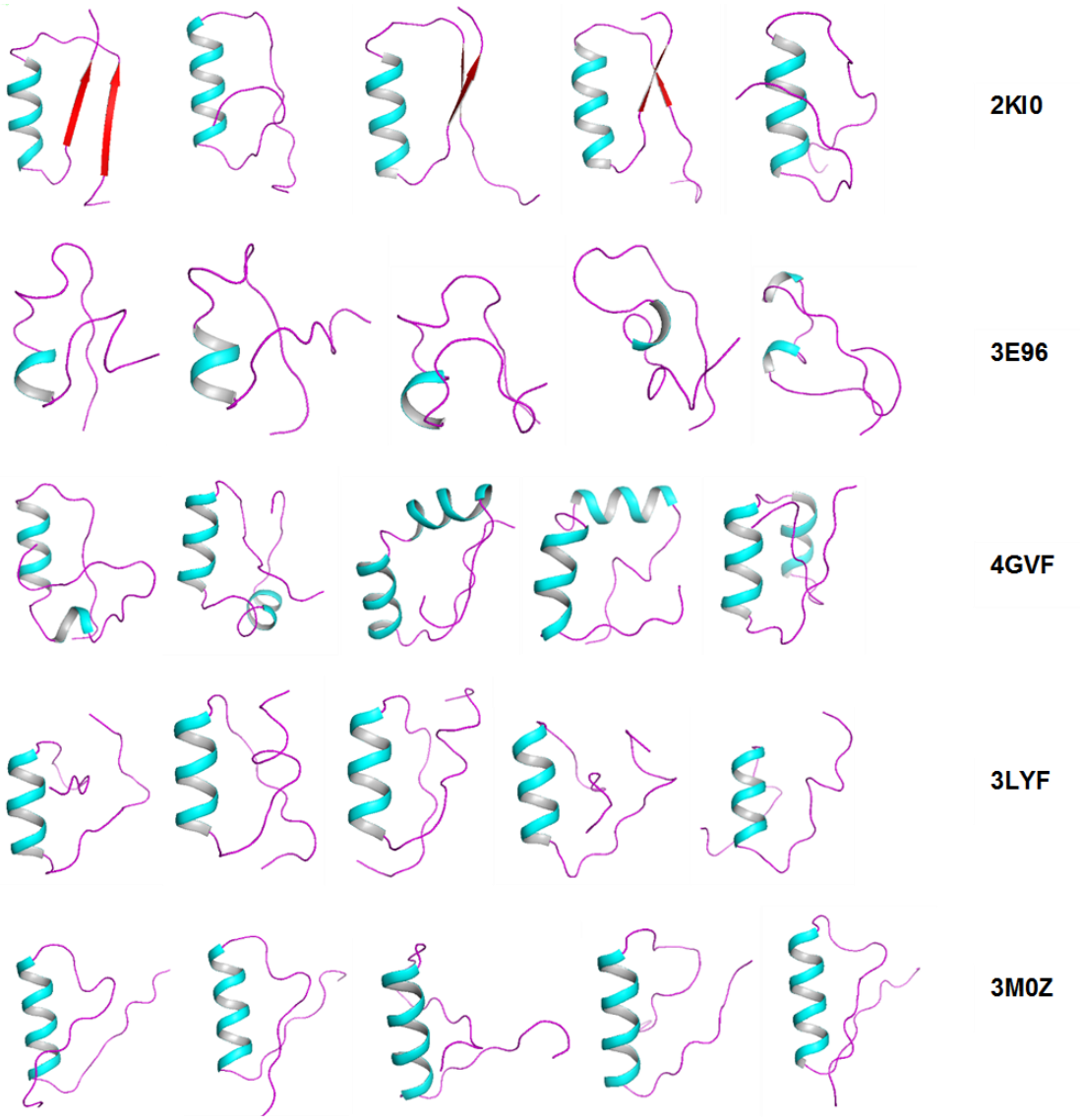


Figure 5.6: *Structural conformations predicted by PEPFOLD3 for control βαβ candidate sequences.*

Stability assessment using Molecular Dynamics Simulations

The stability assessment for the βαβ candidates was carried out using all-atom ~~Gromacs~~ unfolding MD simulations at 300 and 400 K simulated temperatures. The structures that resist unfolding on a longer time scales could indicate higher tendency to fold (Coincon, Heitz, Chiche, & Derreumaux, 2005; Pang & Allemann, 2007; Sham, Ma, Tsai, & Nussinov, 2001). Keeping in mind the time taxing computational requirement the final 10 likely βαβ candidates were further narrowed down to 3 based on QUARK and PEPFOLD3 fold

predictions. The βαβ candidates from 2OZT, 3ELF and 3HJE that have shown higher foldability to βαβ structures during fold predictions were selected for unfolding MD unfolding simulations using Gromacs. The details of the unfolding simulations with the persistence of structure during 0-5 ns simulations are shown below.

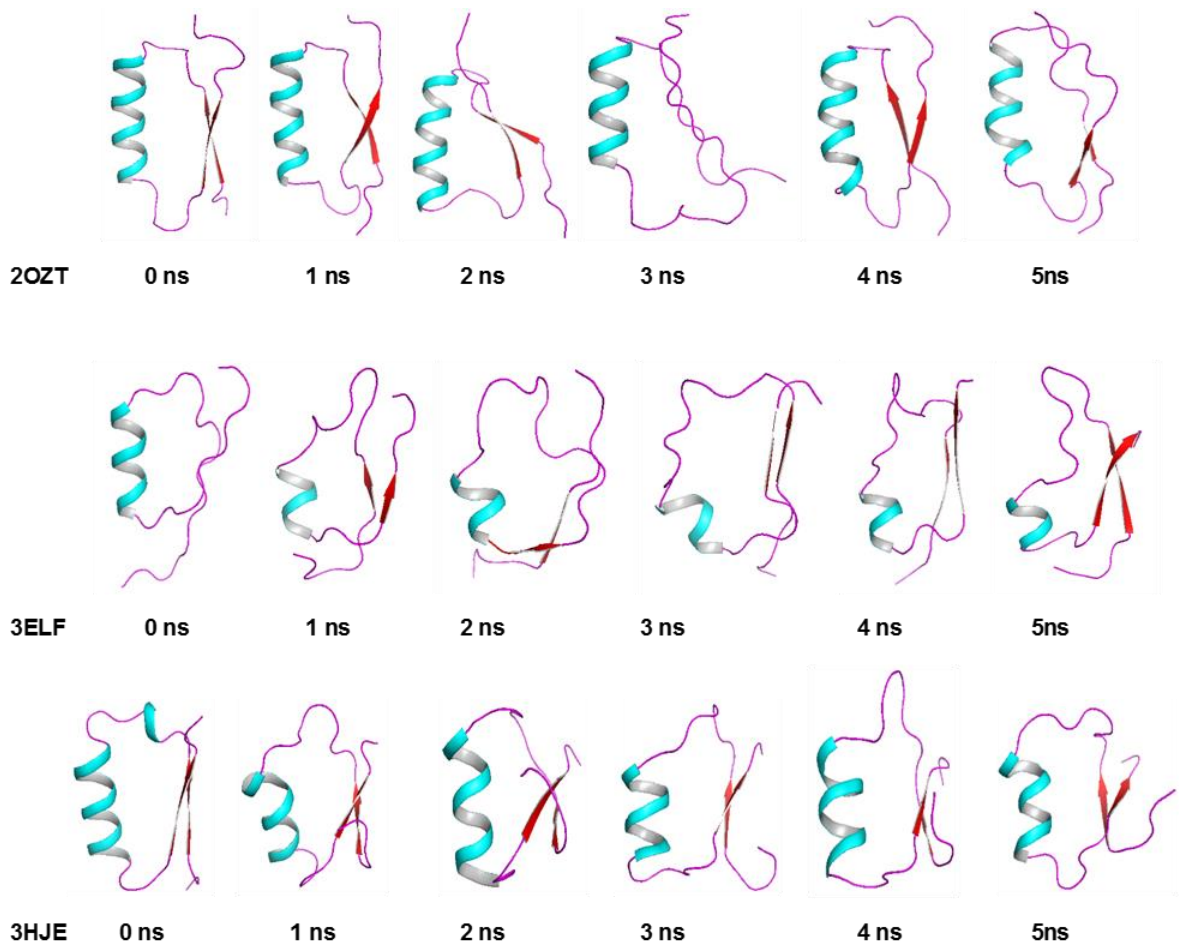


Figure 5.7: *Structural snapshots at 0, 1, 2, 3 4&5 nanoseconds from Gromacs unfolding simulations at 400 K simulated temperature for the βαβ candidates from 2OZT, 3ELF & 3HJE.*

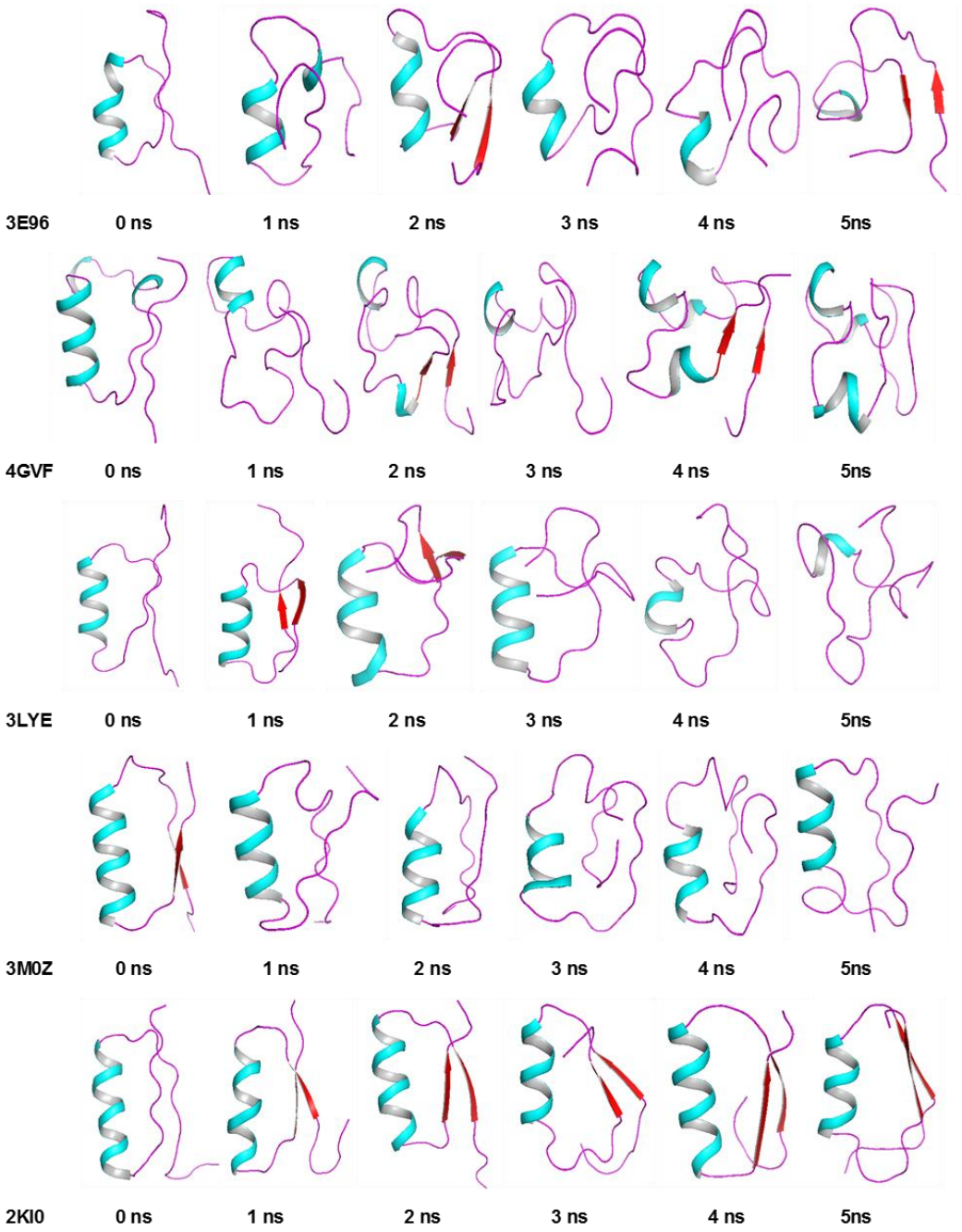


Figure 5.8: *Structural snapshots at 0, 1, 2, 3 4&5 nanoseconds from Gromacs unfolding simulations at 400 K simulated temperature for control βαβ candidates with AGADIR helical propensity of 0.2, 2, 5 & 10 along with designed DS119 (PDBID:2KI0) βαβ conformation.*

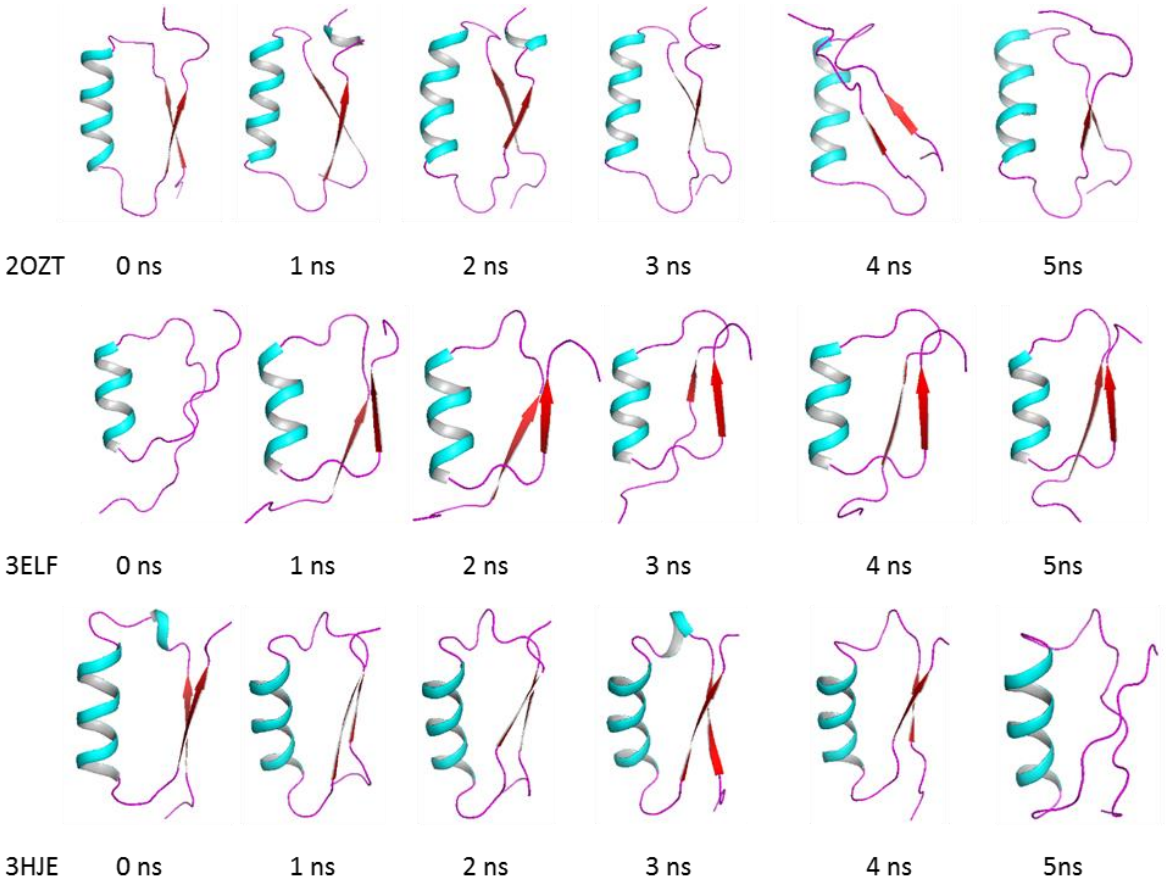


Figure 5.9: *Structural snapshots at 0, 1, 2, 3 4&5 nanoseconds from Gromacs unfolding simulations at 300 K simulated temperature for the βαβ candidates from 2OZT, 3ELF & 3HJE.*

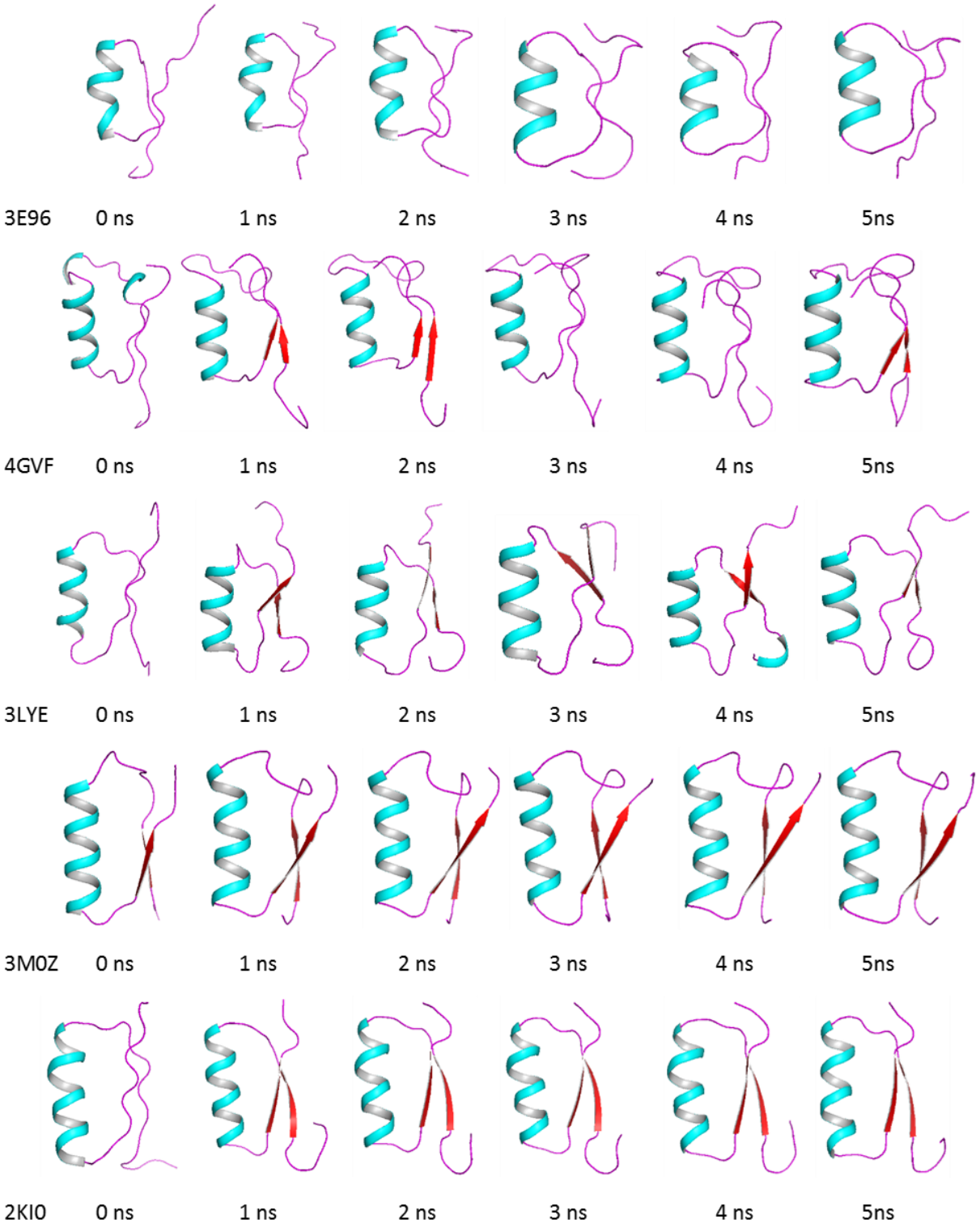


Figure 5.10: *Structural snapshots at 0, 1, 2, 3 4&5 nanoseconds from Gromacs unfolding simulations at 300 K simulated temperature for control βαβ candidates with AGADIR helical propensity of 0.2, 2, 5 & 10 along with designed DS119 (PDBID:2KI0) βαβ conformation.*

The simulations carried out at 400 K simulated temperature has shown shed more light on stability aspects of βαβ candidates from 2OZT, 3HJE and 3ELF. As seen in **Figure 5.7** the βαβ candidate from 2OZT, which has shown 100% βαβ foldability from QUARK and PEPFOLD3 predictions was observed to be highly stable without any noteworthy change in structural conformation followed by βαβ candidate from 3HJE which has also displayed a substantial stability during 5 nanosecond simulation. In contrast, the βαβ unit from 3ELF has shown loss of helical conformation (**Figure 5.7**) consistent with the foldability predictions from QUARK and PEPFOLD3. The designed DS119 βαβ unit conformation is also observed to be persistent during unfolding simulation at 400 K whereas the control sequences with AGADIR helical propensity score 0.2, 2, 5, 10 lost a significant βαβ conformation within 0-1 nanosecond timescale (**Figure 5.8**). Whereas, for the simulations carried out at 300 K simulated temperature it was observed that no unfolding transition with major differences in the conformational changes of βαβ units has been observed (**Figure 5.9 & 5.10**).

DISCUSSION

The analysis of certain βαβ units that exist naturally in TIM barrel proteins suggest that they display the substantial potential to adopt ordered and stable conformations. The approach for the search for the potential βαβ candidates with a tendency for folding is based on intrinsic tendency to fold, (alpha-helical propensity), and stabilizing interactions such as loop length, and long range side chain main chain interactions etc. The denovo designed βαβ unit (DS119) that has been benchmarked for foldability has been predicted to fold as an independent βαβ unit from the monte carlo based folding simulations in addition to displaying resistant to unfolding in the molecular dynamics simulations. Interestingly, at least a few βαβ units identified from TIM barrel structures has shown strong foldability

predictions providing an opportunity to test them experimentally. Expression, purification and structural studies of these candidates are likely to provide insights for optimizing their sequences for folding and stability. Apart from the identification of likely βαβ candidates that can fold independently this work will pave way for development of algorithms that can predict independently folding βαβ motifs which are the basic building blocks of TIM barrel proteins.

One possible reason to find only few independently folding βαβ candidates in this work could be due to evolution of some folding units to favor overall protein folding and stability by optimizing for cooperatively. From recent studies also it was observed that optimization of packing interactions in N-terminal half of the top7 protein enhances the cooperative folding of C-terminal intermediate.

Chapter 6

CONCLUSIONS AND FUTURE PERSPECTIVES

CONCLUSIONS

In summary, this thesis presented (i) a comprehensive analysis of loops from TIM barrel proteins (ii) compiled a detailed library of loops from protein structures in the form of a database, which is powered by a graphical user interface, facilitating users with a diverse set of query-based search options to extract and compare the sequence and structural features of loops (iii) detailed structural and dynamic properties of loops and their role in stability and function of TIM barrel proteins employing both structural and molecular dynamics simulations (iv) identification, analysis and assessment of plausible independently folding/stable βαβ structural motifs from TIM barrel proteins.

The thesis presents a comprehensive compilation and analysis of αβ and βα loops which indicate that size, conformation, and sequence features and preferences differ between the αβ and βα loops of TIM barrel proteins. The αβ loops are dominated by smaller loops in contrast to the longer loops in βα loops. Constrained, short αβ loops can assist establishing efficient hydrophobic packing between the flanking helices and strands. It is expected that the longer βα loops can have a higher degree of freedom to adopt random conformations assisted by multiple turns, required for geometrical flexibility leading to function.

Also, the presence of ordered conformations, in general, and type I turns, in particular, could be the contributing factors for the rigidity of αβ loops. The other likely factor that may also endow rigidity to the αβ loops is the higher proportion of long range side chain hydrogen bonding interactions involving the two positively charged residues, arginine, and lysine.

Glycine residues appear to play an important role in adopting conformations that will result in tight turns leading to optimal packing between the helices and strands of the αβ hairpins. Moreover, lacking side chain, it can maximize the interactions between the helix and strand. In summary, the distinctive role of αβ and βα loops in stability and function, respectively, perhaps is reflecting in their size, sequence profiles, and conformations.

In summary, the present study will help to facilitate short-listing of the candidate loops to be exchanged in place of a target loop. Apart from assisting engineering TIM barrel proteins, the compilation of the αβ and βα loops shed light on the distribution of turn types and preferred template sequences, particularly, in short loops connecting the α and β elements in αβ/βα hairpins. This may encourage experimentalists in designing and or identifying sequences that can adopt α/β structures in solution. As per the findings in this study, four residue loops adopting type I turn conformation may be best suited for connecting the α helix to β strand in αβ hairpins. Analyses from experimental and bioinformatics approaches clearly indicate that the βαβ modules, serve as the minimal unit of stability in β/α class of proteins. Our observations and reasoning will in addition to steering protein engineering efforts on TIM barrel design and stabilization can provide the basis for identifying and or designing independently folding stable βαβ modules.

The analysis from structures and molecular dynamics simulations further provided more insights on the flexibility and dynamic nature of αβ and βα loops in tryptophan synthase alpha subunit, a TIM barrel protein, in particular. Detailed residue level analysis, indicates that αβ loops are more rigid compared to βα loops and alpha helices and may thus play a significant contributing role in the overall stability of the fold. The analysis further indicates that the hydrophobic and long range side chain main chain interactions can play a crucial role in restricting the flexibility of the loops, in αTS, in particular, and TIM barrel proteins,

in general. These observations and findings from the current study will greatly assist in designing independently folding αβ and βαβ units, in addition, to assist engineering stability in TIM barrel proteins by manipulating and or introducing αβ loops at appropriate positions.

The LoopX database presented in this work provides comprehensive information on sequence, conformation, hydrogen bonding interactions of protein loops. Equipped with efficient search tools/algorithms and a visualizer, the database extracts compatible loop candidates for a chosen target loop, thus providing an opportunity for comparing their sequence and structural level information. To the best of our knowledge, LoopX is a comprehensive web-based archive of protein loops which also provides the facility to examine loops and analyze them for their structural similarity, hydrogen bonding interactions, the backbone ϕ, ψ plots. In summary, LoopX serves as a comprehensive graphical user interface driven database for both analysis and comparative evaluation of protein loops, thus adding and complementing to other active databases on protein loops.

The analysis and study of independently folding βαβ motifs from TIM barrels provide opportunities to test the shortlisted βαβ sequences to adopt ordered and stable structures. Given the fact that complex protein structures have evolved from small independently folding super secondary structures, such as βαβ motifs, the exhaustive sequence and structural features based prediction can lead to the identification likely βαβ candidates from TIM barrel proteins that can fold independently. The search methodology and the proposed strategy employed in this process are based on the analysis of loops and the contribution of stabilizing interactions arising from within the loops in addition to basic principles/theories of protein folding and stability. Apart from the identification of likely βαβ candidates that can fold independently this work will pave way for development of algorithms that can

predict independently folding βαβ motifs which are the basic building blocks of TIM barrel proteins.

FUTURE PERSPECTIVES

The design of novel proteins or engineering existing proteins for desired features has been the researchers' efforts. Insights from loops analysis and dynamics work will greatly assist in future protein folding/design/engineering studies.

Selection of compatible loop candidates for successful loop grafting in proteins has been taxing. The challenge of identifying compatible loops for grafting will be greatly aided of LoopX. The comparative analysis of loops at the level of sequence and structure will in addition, to guiding most promising candidates, will help in minimizing the number of candidates to be tested.

The prediction of promising βαβ sequences from TIM barrels will open up opportunities for expression and purification of the sequences and further designing super-secondary βαβ motifs. This also provides insights into exploring nature's secrets that assist the motifs to fold independently combining these motifs to build/engineer new proteins.

Limitations:

The efficiency of LoopX needs to be evaluated experimentally. The foldability and stability of the proposed βαβ candidates need experimental verification.

References

Abraham, M. J., Murtola,T., Schulz,R., Pall,S., Smith,J.C., Hess,B.´, & and Lindahl, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX, 1*, 19-25.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Anantharaman, V., Aravind, L., & Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr Opin Chem Biol, 7*(1), 12-20.

Anderson, J. M., Jurban, B., Huggins, K. N., Shcherbakov, A. A., Shu, I., Kier, B., & Andersen, N. H. (2016). Nascent Hairpins in Proteins: Identifying Turn Loci and Quantitating Turn Contributions to Hairpin Stability. *Biochemistry, 55*(39), 5537-5553. doi: 10.1021/acs.biochem.6b00732

Aurora, R., & Rose, G. D. (1998). Helix capping. *Protein Sci, 7*(1), 21-38. doi: 10.1002/pro.5560070103

Baase, W. A., Liu, L., Tronrud, D. E., & Matthews, B. W. (2010). Lessons from the lysozyme of phage T4. *Protein Sci, 19*(4), 631-641. doi: 10.1002/pro.344

Balasco, N., Esposito, L., De Simone, A., & Vitagliano, L. (2013). Role of loops connecting secondary structure elements in the stabilization of proteins isolated from thermophilic organisms. *Protein Sci, 22*(7), 1016-1023. doi: 10.1002/pro.2279

Baldwin, R. L., & Rose, G. D. (1999). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem Sci, 24*(2), 77-83.

Bartels, C., Xia, T. H., Billeter, M., Güntert, P., & Wüthrich, K. (1995). The Program Xeasy for Computer-Supported Nmr Spectral-Analysis of Biological Macromolecules. *Journal of Biomolecular Nmr, 6*(1), 1-10.

Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol, 324*(1), 105-121.

Benson, E. L., Huynh, P. D., Finkelstein, A., & Collier, R. J. (1998). Identification of residues lining the anthrax protective antigen channel. *Biochemistry, 37*(11), 3941-3948. doi: 10.1021/bi972657b

Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. v., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J Chem Phys, 81*(8), 3684-3690. doi: 10.1063/1.448118

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242.

Bernstein, L. S., Ramineni, S., Hague, C., Cladman, W., Chidiac, P., Levey, A. I., & Hepler, J. R. (2004). RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling. *J Biol Chem, 279*(20), 21248-21256. doi: 10.1074/jbc.M312407200

Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., . . . Baker, D. (2016). Accurate de novo design of hyperstable constrained peptides. *Nature, 538*(7625), 329-335. doi: 10.1038/nature19791

Binz, H. K., Amstutz, P., & Pluckthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol, 23*(10), 1257-1268. doi: 10.1038/nbt1127
Blanco, F. J., Rivas, G., & Serrano, L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Biol, 1*(9), 584-590.

Boersma, Y. L., Pijning, T., Bosma, M. S., van der Sloot, A. M., Godinho, L. F., Droge, M. J., . . . Quax, W. J. (2008). Loop grafting of Bacillus subtilis lipase A: inversion of enantioselectivity. *Chem Biol, 15*(8), 782-789. doi: 10.1016/j.chembiol.2008.06.009

Bogin, O., Peretz, M., Hacham, Y., Korkhin, Y., Frolow, F., Kalb, A. J., & Burstein, Y. (1998). Enhanced thermal stability of Clostridium beijerinckii alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic Thermoanaerobacter brockii alcohol dehydrogenase. *Protein Sci, 7*(5), 1156-1163. doi: 10.1002/pro.5560070509

Bordo, D., & Argos, P. (1994). The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. *J Mol Biol, 243*(3), 504-519. doi: 10.1006/jmbi.1994.1676

Burke, D. F., Deane, C. M., & Blundell, T. L. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics, 16*(6), 513-519.

Butcher, D. J., & Moe, G. R. (1996). Role of hydrophobic interactions and desolvation in determining the structural properties of a model alpha beta peptide. *Proc Natl Acad Sci U S A, 93*(3), 1135-1140.
Butterfield, S. M., Cooper, W. J., & Waters, M. L. (2005). Minimalist protein design: a beta-hairpin peptide that binds ssDNA. *J Am Chem Soc, 127*(1), 24-25. doi: 10.1021/ja045002o

Chen, L., Li, X., Wang, R., Fang, F., Yang, W., & Kan, W. (2016). Thermal stability and unfolding pathways of hyperthermophilic and mesophilic periplasmic binding proteins studied by molecular dynamics simulation. *J Biomol Struct Dyn, 34*(7), 1576-1589. doi: 10.1080/07391102.2015.1084480

Choi, Y., & Deane, C. M. (2010). FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins, 78*(6), 1431-1440. doi: 10.1002/prot.22658

Chothia, C., & Finkelstein, A. V. (1990). The classification and origins of protein folding patterns. *Annu Rev Biochem, 59*, 1007-1039. doi: 10.1146/annurev.bi.59.070190.005043

Chou PY, F. G. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol, 47*, 45-148.

Cochran, A. G., Skelton, N. J., & Starovasnik, M. A. (2001). Tryptophan zippers: stable, monomeric beta -hairpins. *Proc Natl Acad Sci U S A, 98*(10), 5578-5583. doi: 10.1073/pnas.091100898

Coincon, M., Heitz, A., Chiche, L., & Derreumaux, P. (2005). The betaalphabetaalphabeta elementary supersecondary structure of the Rossmann fold from porcine lactate dehydrogenase exhibits characteristics of a molten globule. *Proteins, 60*(4), 740-745. doi: 10.1002/prot.20507

Collinet, B., Garcia, P., Minard, P., & Desmadril, M. (2001). Role of loops in the folding and stability of yeast phosphoglycerate kinase. *Eur J Biochem, 268*(19), 5107-5118.

Colombo, G., De Mori, G. M., & Roccatano, D. (2003). Interplay between hydrophobic cluster and loop propensity in beta-hairpin formation: a mechanistic study. *Protein Sci, 12*(3), 538-550. doi: 10.1110/ps.0227203

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci, 2*(9), 1511-1519. doi: 10.1002/pro.5560020916

Costantini, S., Colonna, G., & Facchiano, A. M. (2006). Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem Biophys Res Commun, 342*(2), 441-451. doi: 10.1016/j.bbrc.2006.01.159

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190. doi: 10.1101/gr.849004

Daggett, V., & Levitt, M. (1992). A model of the molten globule state from molecular dynamics simulations. *Proc Natl Acad Sci U S A, 89*(11), 5142-5146.

Daggett, V., & Levitt, M. (1993). Protein unfolding pathways explored through molecular dynamics simulations. *J Mol Biol, 232*(2), 600-619. doi: 10.1006/jmbi.1993.1414

Dahiyat, B. I., & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science, 278*(5335), 82-87.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys, 98*(12), 10089-10092. doi: 10.1063/1.464397

de Beer, T. A., Berka, K., Thornton, J. M., & Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Res, 42*(Database issue), D292-296. doi: 10.1093/nar/gkt940

De Sancho, D., & Munoz, V. (2011). Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys, 13*(38), 17030-17043. doi: 10.1039/c1cp20402e

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., & Bax, A. (1995). Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular Nmr, 6*(3), 277-293.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., & Tawfik, D. S. (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J Mol Biol, 425*(14), 2609-2621. doi: 10.1016/j.jmb.2013.03.033

Doig, A. J., MacArthur, M. W., Stapley, B. J., & Thornton, J. M. (1997). Structures of N-termini of helices in proteins. *Protein Sci, 6*(1), 147-155. doi: 10.1002/pro.5560060117

Donate, L. E., Rufino, S. D., Canard, L. H., & Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci, 5*(12), 2600-2616. doi: 10.1002/pro.5560051223

Duan, Y., Wang, L., & Kollman, P. A. (1998). The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc Natl Acad Sci U S A, 95*(17), 9897-9902.

Dubey, A., Kadumuri, R. V., Jaipuria, G., Vadrevu, R., & Atreya, H. S. (2016). Rapid NMR Assignments of Proteins by Using Optimized Combinatorial Selective Unlabeling. *Chembiochem, 17*(4), 334-340. doi: 10.1002/cbic.201500513

Dyson, H. J., & Wright, P. E. (1991). Defining solution conformations of small linear peptides. *Annu Rev Biophys Biophys Chem, 20*, 519-538. doi: 10.1146/annurev.bb.20.060191.002511

Edwards, M. S., Sternberg, J. E., & Thornton, J. M. (1987). Structural and sequence patterns in the loops of beta alpha beta units. *Protein Eng, 1*(3), 173-181.

Efimov, A. V. (1991). Structure of coiled beta-beta-hairpins and beta-beta-corners. *FEBS Lett, 284*(2), 288-292.

Eisenberg, D., Wilcox, W., Eshita, S. M., Pryciak, P. M., Ho, S. P., & DeGrado, W. F. (1986). The design, synthesis, and crystallization of an alpha-helical peptide. *Proteins, 1*(1), 16-22. doi: 10.1002/prot.340010105

Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J., & Oliva, B. (2004). ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res, 32*(Database issue), D185-188. doi: 10.1093/nar/gkh002

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J Chem Phys, 103*(19), 8577-8593. doi: 10.1063/1.470117

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., . . . Sali, A. (2006). Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics, Chapter 5*, Unit 5 6. doi: 10.1002/0471250953.bi0506s15

Ewert, S., Honegger, A., & Pluckthun, A. (2004). Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering. *Methods, 34*(2), 184-199. doi: 10.1016/j.ymeth.2004.04.007

Feng, W., Shi, Y., Li, M., & Zhang, M. (2003). Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. *Nat Struct Biol, 10*(11), 972-978. doi: 10.1038/nsb992

Fernandez-Fuentes, N., Hermoso, A., Espadaler, J., Querol, E., Aviles, F. X., & Oliva, B. (2004). Classification of common functional loops of kinase super-families. *Proteins, 56*(3), 539-555. doi: 10.1002/prot.20136

Fetrow, J. S. (1995). Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J, 9*(9), 708-717.

Fezoui, Y., Weaver, D. L., & Osterhout, J. J. (1994). De novo design and structural characterization of an alpha-helical hairpin peptide: a model system for the study of protein folding intermediates. *Proc Natl Acad Sci U S A, 91*(9), 3675-3679.

Finke, J. M., & Onuchic, J. N. (2005). Equilibrium and kinetic folding pathways of a TIM barrel with a funneled energy landscape. *Biophys J, 89*(1), 488-505. doi: 10.1529/biophysj.105.059147
Fiser, A., Do, R. K., & Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci, 9*(9), 1753-1773. doi: 10.1110/ps.9.9.1753

Fiser, A., & Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. *Bioinformatics, 19*(18), 2500-2501.

Frenkel, Z. M., & Trifonov, E. N. (2005). Closed loops of TIM barrel protein fold. *J Biomol Struct Dyn, 22*(6), 643-656. doi: 10.1080/07391102.2005.10507032

Fritz-Wolf, K., Schnyder, T., Wallimann, T., & Kabsch, W. (1996). Structure of mitochondrial creatine kinase. *Nature, 381*(6580), 341-345. doi: 10.1038/381341a0

Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins, 77*(3), 491-498. doi: 10.1002/prot.22509

Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., & Nishikawa, K. (2003). Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol, 327*(2), 347-357.

Fuller-Schaefer, C. A., & Kadner, R. J. (2005). Multiple extracellular loops contribute to substrate binding and transport by the Escherichia coli cobalamin transporter BtuB. *J Bacteriol, 187*(5), 1732-1739. doi: 10.1128/JB.187.5.1732-1739.2005

Gangadhara, B. N., Laine, J. M., Kathuria, S. V., Massi, F., & Matthews, C. R. (2013). Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J Mol Biol, 425*(6), 1065-1081. doi: 10.1016/j.jmb.2013.01.002

Gassner, N. C., Baase, W. A., Mooers, B. H., Busam, R. D., Weaver, L. H., Lindstrom, J. D., . . . Matthews, B. W. (2003). Multiple methionine substitutions are tolerated in T4 lysozyme and have coupled effects on folding and stability. *Biophys Chem, 100*(1-3), 325-340.

Gavrilov, Y., Dagan, S., & Levy, Y. (2015). Shortening a loop can increase protein native state entropy. *Proteins*. doi: 10.1002/prot.24926

Gekko, K., Kunori, Y., Takeuchi, H., Ichihara, S., & Kodama, M. (1994). Point mutations at glycine-121 of Escherichia coli dihydrofolate reductase: important roles of a flexible loop in the stability and function. *J Biochem, 116*(1), 34-41.

Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol, 274*(4), 562-576. doi: 10.1006/jmbi.1997.1412

Goddard, T. D., & Kneller, D. G. (2006). SPARKY 3, University of California, San Francisco.

Gracy, R. W. (1975). Triosephosphate isomerase from human erythrocytes. *Methods Enzymol, 41*, 442-447.

Grishin, N. V. (2001). Fold change in evolution of protein structures. *J Struct Biol, 134*(2-3), 167-185. doi: 10.1006/jsbi.2001.4335

Gromiha, M. M. (2001). Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys Chem, 91*(1), 71-77.

Gualfetti, P. J., Bilsel, O., & Matthews, C. R. (1999). The progressive development of structure and stability during the equilibrium folding of the alpha subunit of tryptophan synthase from Escherichia coli. *Protein Sci, 8*(8), 1623-1635. doi: 10.1110/ps.8.8.1623

Gunasekaran, K., Ramakrishnan, C., & Balaram, P. (1997). Beta-hairpins in proteins revisited: lessons for de novo design. *Protein Eng, 10*(10), 1131-1141.

Guruprasad, K., Prasad, M. S., & Kumar, G. R. (2000). Analysis of gammabeta, betagamma, gammagamma, betabeta multiple turns in proteins. *J Pept Res, 56*(4), 250-263.

Guruprasad, K., & Rajkumar, S. (2000). Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *J Biosci, 25*(2), 143-156.
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, . . . Bourne, P. E. (2000). The Protein Data Bank *Nucleic Acids Res, 28*(1), 235-242.

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science, 282*(5393), 1462-1467.

Hardy, F., Vriend, G., Veltman, O. R., van der Vinne, B., Venema, G., & Eijsink, V. G. (1993). Stabilization of Bacillus stearothermophilus neutral protease by introduction of prolines. *FEBS Lett, 317*(1-2), 89-92.

Hess, B., Bekker, H., Berendsen, H. J. C. and Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem, 18*(12), 1463–1472.

Ho, B. K., & Dill, K. A. (2006). Folding very short peptides using molecular dynamics. *PLoS Comput Biol, 2*(4), e27. doi: 10.1371/journal.pcbi.0020027

Hocker, B. (2014). Design of proteins from smaller fragments-learning from evolution. *Curr Opin Struct Biol, 27*, 56-62. doi: 10.1016/j.sbi.2014.04.007

Hoedemaeker, F. J., van Eijsden, R. R., Diaz, C. L., de Pater, B. S., & Kijne, J. W. (1993). Destabilization of pea lectin by substitution of a single amino acid in a surface loop. *Plant Mol Biol, 22*(6), 1039-1046.

Hsu, H. J., Chang, H. J., Peng, H. P., Huang, S. S., Lin, M. Y., & Yang, A. S. (2006). Assessing computational amino acid beta-turn propensities with a phage-displayed combinatorial library and directed evolution. *Structure, 14*(10), 1499-1510. doi: 10.1016/j.str.2006.08.006

Huang, P. S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Hocker, B., & Baker, D. (2016). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol, 12*(1), 29-34. doi: 10.1038/nchembio.1966

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *J Mol Graph, 14*(1), 33-38, 27-38.

Hutchinson, E. G., & Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci, 5*(2), 212-220. doi: 10.1002/pro.5560050204

Iacovache, I., Paumard, P., Scheib, H., Lesieur, C., Sakai, N., Matile, S., . . . van der Goot, F. G. (2006). A rivet model for channel formation by aerolysin-like pore-forming toxins. *EMBO J, 25*(3), 457-466. doi: 10.1038/sj.emboj.7600959

Ihalainen, J. A., Paoli, B., Muff, S., Backus, E. H., Bredenbeck, J., Woolley, G. A., . . . Hamm, P. (2008). Alpha-Helix folding in the presence of structural constraints. *Proc Natl Acad Sci U S A, 105*(28), 9588-9593. doi: 10.1073/pnas.0712099105

Jager, M., Deechongkit, S., Koepf, E. K., Nguyen, H., Gao, J., Powers, E. T., . . . Kelly, J. W. (2008). Understanding the mechanism of beta-sheet folding from a chemical and biological perspective. *Biopolymers, 90*(6), 751-758. doi: 10.1002/bip.21101

Jiang, X., Chen, G., & Wang, L. (2016). Structural and dynamic evolution of the amphipathic N-terminus diversifies enzyme thermostability in the glycoside hydrolase family 12. *Phys Chem Chem Phys, 18*(31), 21340-21350. doi: 10.1039/c6cp02998a

Johnson, L. N., Lowe, E. D., Noble, M. E., & Owen, D. J. (1998). The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett, 430*(1-2), 1-11.

Johnson, T. A., & Holyoak, T. (2012). The Omega-loop lid domain of phosphoenolpyruvate carboxykinase is essential for catalytic function. *Biochemistry, 51*(47), 9547-9559. doi: 10.1021/bi301278t

Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., . . . Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Res, 39*(Database issue), D411-419. doi: 10.1093/nar/gkq1105

Jung, S., & Pluckthun, A. (1997). Improving in vivo folding and stability of a single-chain Fv antibody fragment by loop grafting. *Protein Eng, 10*(8), 959-966.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), 2577-2637. doi: 10.1002/bip.360221211

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

Kawasaki, H., & Kretsinger, R. H. (1995). Calcium-binding proteins 1: EF-hands. *Protein Profile, 2*(4), 297-490.

Keskin, O., Jernigan, R. L., & Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys J, 78*(4), 2093-2106. doi: 10.1016/S0006-3495(00)76756-7

Khan, S., Farooq, U., & Kurnikova, M. (2016). Exploring Protein Stability by Comparative Molecular Dynamics Simulations of Homologous Hyperthermophilic, Mesophilic, and Psychrophilic Proteins. *J Chem Inf Model, 56*(11), 2129-2139. doi: 10.1021/acs.jcim.6b00305

Kim, S. T., Shirai, H., Nakajima, N., Higo, J., & Nakamura, H. (1999). Enhanced conformational diversity search of CDR-H3 in antibodies: role of the first CDR-H3 residue. *Proteins, 37*(4), 683-696.
Kimura, S., Kanaya, S., & Nakamura, H. (1992). Thermostabilization of Escherichia coli ribonuclease HI by replacing left-handed helical Lys95 with Gly or Asn. *J Biol Chem, 267*(31), 22014-22017.

Knudsen, M., & Wiuf, C. (2010). The CATH database. *Hum Genomics, 4*(3), 207-212.

Ko, J., Lee, D., Park, H., Coutsias, E. A., Lee, J., & Seok, C. (2011). The FALC-Loop web server for protein loop modeling. *Nucleic Acids Res, 39*(Web Server issue), W210-214. doi: 10.1093/nar/gkr352

Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., & Baker, D. (2012). Principles for designing ideal protein structures. *Nature, 491*(7423), 222-227. doi: 10.1038/nature11600

Kundu, S., & Roy, D. (2008). Temperature-induced unfolding pathway of a type III antifreeze protein: insight from molecular dynamics simulation. *J Mol Graph Model, 27*(1), 88-94. doi: 10.1016/j.jmgm.2008.03.002

Kwasigroch, J. M., Chomilier, J., & Mornon, J. P. (1996). A global taxonomy of loops in globular proteins. *J Mol Biol, 259*(4), 855-872. doi: 10.1006/jmbi.1996.0363

Lacroix, E., Viguera, A. R., & Serrano, L. (1998). Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol, 284*(1), 173-191. doi: 10.1006/jmbi.1998.2145

Lamiable, A., Thevenet, P., Rey, J., Vavrusa, M., Derreumaux, P., & Tuffery, P. (2016). PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res, 44*(W1), W449-454. doi: 10.1093/nar/gkw329

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., & Wilmanns, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science, 289*(5484), 1546-1550.

Lazaridis, T., Lee, I., & Karplus, M. (1997). Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci, 6*(12), 2589-2605. doi: 10.1002/pro.5560061211

Levitt, M., & Chothia, C. (1976). Structural patterns in globular proteins. *Nature, 261*(5561), 552-558.

Lewandowska, A., Oldziej, S., Liwo, A., & Scheraga, H. A. (2010). beta-hairpin-forming peptides; models of early stages of protein folding. *Biophys Chem, 151*(1-2), 1-9. doi: 10.1016/j.bpc.2010.05.001

Li, C., Heatwole, J., Soelaiman, S., & Shoham, M. (1999). Crystal structure of a thermophilic alcohol dehydrogenase substrate complex suggests determinants of substrate specificity and thermostability. *Proteins, 37*(4), 619-627.

Li, H., Helling, R., Tang, C., & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science, 273*(5275), 666-669.

Li, J., Chen, Y., Yang, J., & Hua, Z. (2015). Thermal- and urea-induced unfolding processes of glutathione S-transferase by molecular dynamics simulation. *Biopolymers, 103*(5), 247-259. doi: 10.1002/bip.22589

Li, W., Liang, S., Wang, R., Lai, L., & Han, Y. (1999). Exploring the conformational diversity of loops on conserved frameworks. *Protein Eng, 12*(12), 1075-1086.

Li, W., Liu, Z., & Lai, L. (1999). Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers, 49*(6), 481-495.

Liang, H., Chen, H., Fan, K., Wei, P., Guo, X., Jin, C., . . . Lai, L. (2009). De novo design of a beta alpha beta motif. *Angew Chem Int Ed Engl, 48*(18), 3301-3303. doi: 10.1002/anie.200805476

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res, 28*(1), 257-259.

Lupas, A. N., Ponting, C. P., & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol, 134*(2-3), 191-203. doi: 10.1006/jsbi.2001.4393

Madan, B., Seo, S. Y., & Lee, S. G. (2014). Structural and sequence features of two residue turns in beta-hairpins. *Proteins, 82*(9), 1721-1733. doi: 10.1002/prot.24526

Maguid, S., Fernandez-Alberti, S., Parisi, G., & Echave, J. (2006). Evolutionary conservation of protein backbone flexibility. *J Mol Evol, 63*(4), 448-457. doi: 10.1007/s00239-005-0209-x

Magyar, C., Gromiha, M. M., Savoly, Z., & Simon, I. (2016). The role of stabilization centers in protein thermal stability. *Biochem Biophys Res Commun*. doi: 10.1016/j.bbrc.2016.01.181

Manoutcharian, K., Terrazas, L. I., Gevorkian, G., Acero, G., Petrossian, P., Rodriguez, M., & Govezensky, T. (1999). Phage-displayed T-cell epitope grafted into immunoglobulin heavy-chain complementarity-determining regions: an effective vaccine design tested in murine cysticercosis. *Infect Immun, 67*(9), 4764-4770.

Marcelino, A. M., & Gierasch, L. M. (2008). Roles of beta-turns in protein folding: from peptide models to protein engineering. *Biopolymers, 89*(5), 380-391. doi: 10.1002/bip.20960

Marqusee, S., Robbins, V. H., & Baldwin, R. L. (1989). Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci U S A, 86*(14), 5286-5290.

Masumoto, K., Ueda, T., Motoshima, H., & Imoto, T. (2000). Relationship between local structure and stability in hen egg white lysozyme mutant with alanine substituted for glycine. *Protein Eng, 13*(10), 691-695.

Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc Natl Acad Sci U S A, 84*(19), 6663-6667.

Matthews, C. R., & Crisanti, M. M. (1981). Urea-induced unfolding of the .alpha. subunit of tryptophan synthase: evidence for a multistate process. *Biochemistry, 20*(4), 784-792. doi: 10.1021/bi00507a021

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Cryst, 38*, 871-873.
Michalsky, E., Goede, A., & Preissner, R. (2003). Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. *Protein Eng, 16*(12), 979-985. doi: 10.1093/protein/gzg119

Mooers, B. H., Baase, W. A., Wray, J. W., & Matthews, B. W. (2009). Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. *Protein Sci, 18*(5), 871-880. doi: 10.1002/pro.94

Mortazavi, M., & Hosseinkhani, S. (2011). Design of thermostable luciferases through arginine saturation in solvent-exposed loops. *Protein Eng Des Sel, 24*(12), 893-903. doi: 10.1093/protein/gzr051

Munoz, V., & Serrano, L. (1995a). Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J Mol Biol, 245*(3), 275-296.

Munoz, V., & Serrano, L. (1995b). Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J Mol Biol, 245*(3), 297-308. doi: 10.1006/jmbi.1994.0024

Munoz, V., & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers, 41*(5), 495-509.

Munoz, V., Thompson, P. A., Hofrichter, J., & Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature, 390*(6656), 196-199. doi: 10.1038/36626

Nagano, N., Orengo, C. A., & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol, 321*(5), 741-765.

Nagarajan, D., Deka, G., & Rao, M. (2015). Design of symmetric TIM barrel proteins from first principles. *BMC Biochem, 16*, 18. doi: 10.1186/s12858-015-0047-4

Nagi AD, R. L. (1997). An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold Des, 2*(1), 67-75.

Nakashima, H., Fukuchi, S., & Nishikawa, K. (2003). Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J Biochem, 133*(4), 507-513.

Ochoa-Leyva, A., Barona-Gomez, F., Saab-Rincon, G., Verdel-Aranda, K., Sanchez, F., & Soberon, X. (2011). Exploring the Structure-Function Loop Adaptability of a (beta/alpha)(8)-Barrel Enzyme through Loop Swapping and Hinge Variability. *J Mol Biol, 411*(1), 143-157. doi: 10.1016/j.jmb.2011.05.027

Ochoa-Leyva, A., Montero-Moran, G., Saab-Rincon, G., Brieba, L. G., & Soberon, X. (2013). Alternative splice variants in TIM barrel proteins from human genome correlate with the structural and evolutionary modularity of this versatile protein fold. *PLoS One, 8*(8), e70582. doi: 10.1371/journal.pone.0070582

Ochoa-Leyva, A., Soberon, X., Sanchez, F., Arguello, M., Montero-Moran, G., & Saab-Rincon, G. (2009). Protein design through systematic catalytic loop exchange in the (beta/alpha)8 fold. *J Mol Biol, 387*(4), 949-964. doi: 10.1016/j.jmb.2009.02.022

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., & Sternberg, M. J. (1997). An automated classification of the structure of protein loops. *J Mol Biol, 266*(4), 814-830. doi: 10.1006/jmbi.1996.0819

Orengo, C. A., & Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annu Rev Biochem, 74*, 867-900. doi: 10.1146/annurev.biochem.74.082803.133029

Pang, J., & Allemann, R. K. (2007). Molecular dynamics simulation of thermal unfolding of Thermatoga maritima DHFR. *Phys Chem Chem Phys, 9*(6), 711-718. doi: 10.1039/b611210b

Papaleo, E., Riccardi, L., Villa, C., Fantucci, P., & De Gioia, L. (2006). Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family. *Biochim Biophys Acta, 1764*(8), 1397-1406. doi: 10.1016/j.bbapap.2006.06.005

Parge, H. E., Hallewell, R. A., & Tainer, J. A. (1992). Atomic structures of wild-type and thermostable mutant recombinant human Cu,Zn superoxide dismutase. *Proc Natl Acad Sci U S A, 89*(13), 6109-6113.

Park, H. S., Nam, S. H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., & Kim, H. S. (2006). Design and evolution of new catalytic activity with an existing protein scaffold. *Science, 311*(5760), 535-538. doi: 10.1126/science.1118953

Parthasarathy, S., & Murthy, M. R. (2000). Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng, 13*(1), 9-13.

Paul, M., Hazra, M., Barman, A., & Hazra, S. (2014). Comparative molecular dynamics simulation studies for determining factors contributing to the thermostability of chemotaxis protein "CheY". *J Biomol Struct Dyn, 32*(6), 928-949. doi: 10.1080/07391102.2013.799438

Petsko, G. A. (2000). Enzyme evolution. Design by necessity. *Nature, 403*(6770), 606-607. doi: 10.1038/35001176

Petukhov, M., Tatsu, Y., Tamaki, K., Murase, S., Uekawa, H., Yoshikawa, S., . . . Yumoto, N. (2009). Design of stable alpha-helices using global sequence optimization. *J Pept Sci, 15*(5), 359-365. doi: 10.1002/psc.1122

Pompliano, D. L., Peyman, A., & Knowles, J. R. (1990). Stabilization of a reaction intermediate as a catalytic device: definition of the functional role of the flexible loop in triosephosphate isomerase. *Biochemistry, 29*(13), 3186-3194.

Predki, P. F., Agrawal, V., Brunger, A. T., & Regan, L. (1996). Amino-acid substitutions in a surface turn modulate protein stability. *Nat Struct Biol, 3*(1), 54-58.

Ptitsyn, O. B., & Finkelstein, A. V. (1989). Prediction of protein secondary structure based on physical theory. Histones. *Protein Eng, 2*(6), 443-447.

Qi, Y., Huang, Y., Liang, H., Liu, Z., & Lai, L. (2010). Folding simulations of a de novo designed protein with a betaalphabeta fold. *Biophys J, 98*(2), 321-329. doi: 10.1016/j.bpj.2009.10.018

Ramakrishna, V., & Sasidhar, Y. U. (1997). A pentapeptide model for an early folding step in the refolding of staphylococcal nuclease: the role of its turn propensity. *Biopolymers, 41*(2), 181-191. doi: 10.1002/(SICI)1097-0282(199702)41:2<181::AID-BIP5>3.0.CO;2-P

Ramirez-Alvarado, M., Blanco, F. J., Niemann, H., & Serrano, L. (1997). Role of beta-turn residues in beta-hairpin formation and stability in designed peptides. *J Mol Biol, 273*(4), 898-912. doi: 10.1006/jmbi.1997.1347

Regan, L., & DeGrado, W. F. (1988). Characterization of a helical protein designed from first principles. *Science, 241*(4868), 976-978.

Religa, T. L., Johnson, C. M., Vu, D. M., Brewer, S. H., Dyer, R. B., & Fersht, A. R. (2007). The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. *Proc Natl Acad Sci U S A, 104*(22), 9272-9277. doi: 10.1073/pnas.0703434104

Renner, C., Schleicher, M., Moroder, L., & Holak, T. A. (2002). Practical aspects of the 2D 15N-[1h]-NOE experiment. *J Biomol NMR, 23*(1), 23-33.

Richard, J. P., Zhai, X., & Malabanan, M. M. (2014). Reflections on the catalytic power of a TIM-barrel. *Bioorg Chem, 57*, 206-212. doi: 10.1016/j.bioorg.2014.07.001

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem, 34*, 167-339.

Richter, A., Eggenstein, E., & Skerra, A. (2014). Anticalins: exploiting a non-Ig scaffold with hypervariable loops for the engineering of binding proteins. *FEBS Lett, 588*(2), 213-218. doi: 10.1016/j.febslet.2013.11.006

Riechmann, L., & Winter, G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc Natl Acad Sci U S A, 97*(18), 10068-10073. doi: 10.1073/pnas.170145497

Riechmann, L., & Winter, G. (2006). Early protein evolution: building domains from ligand-binding polypeptide segments. *J Mol Biol, 363*(2), 460-468. doi: 10.1016/j.jmb.2006.08.031

Rojsajjakul, T., Wintrode, P., Vadrevu, R., Robert Matthews, C., & Smith, D. L. (2004). Multi-state unfolding of the alpha subunit of tryptophan synthase, a TIM barrel protein: insights into the secondary structure of the stable equilibrium intermediates by hydrogen exchange mass spectrometry. *J Mol Biol, 341*(1), 241-253. doi: 10.1016/j.jmb.2004.05.062

Sadqi, M., de Alba, E., Perez-Jimenez, R., Sanchez-Ruiz, J. M., & Munoz, V. (2009). A designed protein as experimental model of primordial folding. *Proc Natl Acad Sci U S A, 106*(11), 4127-4132. doi: 10.1073/pnas.0812108106

Saerens, D., Pellis, M., Loris, R., Pardon, E., Dumoulin, M., Matagne, A., . . . Conrath, K. (2005). Identification of a universal VHH framework to graft non-canonical antigen-binding loops of camel single-domain antibodies. *J Mol Biol, 352*(3), 597-607. doi: 10.1016/j.jmb.2005.07.038

Sahu, S. C., Bhuyan, A. K., Udgaonkar, J. B., & Hosur, R. V. (2000). Backbone dynamics of free barnase and its complex with barstar determined by 15N NMR relaxation study. *J Biomol NMR, 18*(2), 107-118.

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol, 234*(3), 779-815. doi: 10.1006/jmbi.1993.1626

Saunders, N. F., Thomas, T., Curmi, P. M., Mattick, J. S., Kuczek, E., Slade, R., . . . Cavicchioli, R. (2003). Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. *Genome Res, 13*(7), 1580-1588. doi: 10.1101/gr.1180903

Scheerlinck, J. P., Lasters, I., Claessens, M., De Maeyer, M., Pio, F., Delhaise, P., & Wodak, S. J. (1992). Recurrent alpha beta loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins, 12*(4), 299-313. doi: 10.1002/prot.340120402

Schrodinger, L. (2015). The PyMOL Molecular Graphics System, Version 1.8.

Schuttelkopf, A. W., & van Aalten, D. M. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr, 60*(Pt 8), 1355-1363. doi: 10.1107/S0907444904011679

Searle, M. S., Williams, D. H., & Packman, L. C. (1995). A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. *Nat Struct Biol, 2*(11), 999-1006.

Sham, Y. Y., Ma, B., Tsai, C. J., & Nussinov, R. (2001). Molecular dynamics simulation of Escherichia coli dihydrofolate reductase and its protein fragments: relative stabilities in experiment and simulations. *Protein Sci, 10*(1), 135-148. doi: 10.1110/ps.33301

Shen, Y., Maupetit, J., Derreumaux, P., & Tuffery, P. (2014). Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. *J Chem Theory Comput, 10*(10), 4745-4758. doi: 10.1021/ct500592m

Shin, H. C., Merutka, G., Waltho, J. P., Tennant, L. L., Dyson, H. J., & Wright, P. E. (1993). Peptide models of protein folding initiation sites. 3. The G-H helical hairpin of myoglobin. *Biochemistry, 32*(25), 6356-6364.

Simpson, E. R., Meldrum, J. K., Bofill, R., Crespo, M. D., Holmes, E., & Searle, M. S. (2005). Engineering enhanced protein stability through beta-turn optimization: insights for the design of stable peptide beta-hairpin systems. *Angew Chem Int Ed Engl, 44*(31), 4939-4944. doi: 10.1002/anie.200500577

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol, 213*(4), 859-883.

Smith, J. W., Tachias, K., & Madison, E. L. (1995). Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin alpha IIb beta 3. *J Biol Chem, 270*(51), 30486-30490.

Soding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res, 33*(Web Server issue), W244-248. doi: 10.1093/nar/gki408

Soding, J., & Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays, 25*(9), 837-846. doi: 10.1002/bies.10321

Sormanni, P., Aprile, F. A., & Vendruscolo, M. (2015). Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proc Natl Acad Sci U S A, 112*(32), 9902-9907. doi: 10.1073/pnas.1422401112

Stanger, H. E., Syud, F. A., Espinosa, J. F., Giriat, I., Muir, T., & Gellman, S. H. (2001). Length-dependent stability and strand length limits in antiparallel beta -sheet secondary structure. *Proc Natl Acad Sci U S A, 98*(21), 12015-12020. doi: 10.1073/pnas.211536998

Sterner, R., & Hocker, B. (2005). Catalytic versatility, stability, and evolution of the (betaalpha)8-barrel enzyme fold. *Chem Rev, 105*(11), 4038-4055. doi: 10.1021/cr030191z

Stites, W. E., Meeker, A. K., & Shortle, D. (1994). Evidence for strained interactions between side-chains and the polypeptide backbone. *J Mol Biol, 235*(1), 27-32.

Strub, C., Alies, C., Lougarre, A., Ladurantie, C., Czaplicki, J., & Fournier, D. (2004). Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. *BMC Biochem, 5*, 9. doi: 10.1186/1471-2091-5-9

Struthers, M. D., Cheng, R. P., & Imperiali, B. (1996). Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science, 271*(5247), 342-345.

Suhre, K., & Claverie, J. M. (2003). Genomic correlates of hyperthermostability, an update. *J Biol Chem, 278*(19), 17198-17202. doi: 10.1074/jbc.M301327200

Szilagyi, A., & Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure, 8*(5), 493-504.

Takano, K., Yamagata, Y., & Yutani, K. (2001). Role of amino acid residues in left-handed helical conformation for the conformational stability of a protein. *Proteins, 45*(3), 274-280.

Tanaka Y, T. K., Yasutake Y, Umetsu M, Yao M, Fukada H, Tanaka I, Kumagai , & I. (2004). How oligomerization contributes to the thermostability of an archaeon protein. *J Biol Chem, 279*(31), 32957-32967.

Tawfik, D. S. (2006). Biochemistry. Loop grafting and the origins of enzyme species. *Science, 311*(5760), 475-476. doi: 10.1126/science.1123883

Thevenet, P., Shen, Y., Maupetit, J., Guyon, F., Derreumaux, P., & Tuffery, P. (2012). PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res, 40*(Web Server issue), W288-293. doi: 10.1093/nar/gks419

Thornton, J. M., Sibanda, B. L., Edwards, M. S., & Barlow, D. J. (1988). Analysis, design and modification of loop regions in proteins. *Bioessays, 8*(2), 63-69. doi: 10.1002/bies.950080205

Tompa, D. R., Gromiha, M. M., & Saraboji, K. (2016). Contribution of main chain and side chain atoms and their locations to the stability of thermophilic proteins. *J Mol Graph Model, 64*, 85-93. doi: 10.1016/j.jmgm.2016.01.001

Trevino, S. R., Schaefer, S., Scholtz, J. M., & Pace, C. N. (2007). Increasing protein conformational stability by optimizing beta-turn sequence. *J Mol Biol, 373*(1), 211-218. doi: 10.1016/j.jmb.2007.07.061

Urfer, R., & Kirschner, K. (1992). The importance of surface loops for stabilizing an eightfold beta alpha barrel protein. *Protein Sci, 1*(1), 31-45. doi: 10.1002/pro.5560010105

Vadrevu, R., Falzone, C. J., & Matthews, C. R. (2003). Partial NMR assignments and secondary structure mapping of the isolated alpha subunit of Escherichia coli tryptophan synthase, a 29-kD TIM barrel protein. *Protein Sci, 12*(1), 185-191. doi: 10.1110/ps.0221103

Vadrevu, R., Wu, Y., & Matthews, C. R. (2008). NMR analysis of partially folded states and persistent structure in the alpha subunit of tryptophan synthase: implications for the equilibrium folding mechanism of a 29-kDa TIM barrel protein. *J Mol Biol, 377*(1), 294-306. doi: 10.1016/j.jmb.2007.11.010

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J Comput Chem, 26*(16), 1701-1718. doi: 10.1002/jcc.20291

van Gunsteren, W. F., Billeter, S. R., Eking, A. A., Hiinenberger, P. H., Kriiger, P., Mark, A. E., Scott, W. R. P. and Tironi, I. G. (1996). Biomolecular Simulation, The GROMOS96 Manual and User Guide. *vdf Hochschulverlag AG an der ETH Ziirich and BIOMOS b.v.*

Vanhee, P., Verschueren, E., Baeten, L., Stricher, F., Serrano, L., Rousseau, F., & Schymkowitz, J. (2011). BriX: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Res, 39*(Database issue), D435-442. doi: 10.1093/nar/gkq972

Vemparala, S., Mehrotra, S., & Balaram, H. (2011). Role of loop dynamics in thermal stability of mesophilic and thermophilic adenylosuccinate synthetase: a molecular dynamics and normal mode analysis study. *Biochim Biophys Acta, 1814*(5), 630-637. doi: 10.1016/j.bbapap.2011.03.012

Vieille, C., Burdette, D. S., & Zeikus, J. G. (1996). Thermozymes. *Biotechnol Annu Rev, 2*, 1-83.

Vieille, C., & Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev, 65*(1), 1-43. doi: 10.1128/MMBR.65.1.1-43.2001

Walser, R., Kleinschmidt, J. H., Skerra, A., & Zerbe, O. (2012). beta-Barrel scaffolds for the grafting of extracellular loops from G-protein-coupled receptors. *Biol Chem, 393*(11), 1341-1355. doi: 10.1515/hsz-2012-0234

Wang, J., & Feng, J. A. (2003). Exploring the sequence patterns in the alpha-helices of proteins. *Protein Eng, 16*(11), 799-807.

Watanabe, K., Hata, Y., Kizaki, H., Katsube, Y., & Suzuki, Y. (1997). The refined crystal structure of Bacillus cereus oligo-1,6-glucosidase at 2.0 A resolution: structural characterization of proline-substitution sites for protein thermostabilization. *J Mol Biol, 269*(1), 142-153. doi: 10.1006/jmbi.1997.1018

Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett, 492*(3), 193-198.

Wierenga, R. K., Terpstra, P., & Hol, W. G. (1986). Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. *J Mol Biol, 187*(1), 101-107.

Wijma, H. J., Floor, R. J., & Janssen, D. B. (2013). Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr Opin Struct Biol, 23*(4), 588-594. doi: 10.1016/j.sbi.2013.04.008

Wolfgang, K. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica, A34*, 827-828.

Wright, P. E., Dyson, H. J., & Lerner, R. A. (1988). Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. *Biochemistry, 27*(19), 7167-7175.

Wu, Y., Vadrevu, R., Kathuria, S., Yang, X., & Matthews, C. R. (2007). A tightly packed hydrophobic cluster directs the formation of an off-pathway sub-millisecond folding intermediate in the alpha subunit of tryptophan synthase, a TIM barrel protein. *J Mol Biol, 366*(5), 1624-1638. doi: 10.1016/j.jmb.2006.12.005

Xiao, L., & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J Mol Biol, 289*(5), 1435-1444. doi: 10.1006/jmbi.1999.2810

Xiong, F., Xia, L., Wang, J., Wu, B., Wang, D., Yuan, L., . . . Wang, Y. (2014). A high-affinity CDR-grafted antibody against influenza A H5N1 viruses recognizes a conserved epitope of H5 hemagglutinin. *PLoS One, 9*(2), e88777. doi: 10.1371/journal.pone.0088777

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins, 80*(7), 1715-1735. doi: 10.1002/prot.24065

Xu, J., Baase, W. A., Baldwin, E., & Matthews, B. W. (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci, 7*(1), 158-177. doi: 10.1002/pro.5560070117

Yakimov, A., Rychkov, G., & Petukhov, M. (2014). De novo design of stable alpha-helices. *Methods Mol Biol, 1216*, 1-14. doi: 10.1007/978-1-4939-1486-9_1

Yang, A. S., Hitz, B., & Honig, B. (1996). Free energy determinants of secondary structure formation: III. beta-turns and their role in protein folding. *J Mol Biol, 259*(4), 873-882. doi: 10.1006/jmbi.1996.0364

Yang, X., Kathuria, S. V., Vadrevu, R., & Matthews, C. R. (2009). Betaalpha-hairpin clamps brace betaalphabeta modules and can make substantive contributions to the stability of TIM barrel proteins. *PLoS One, 4*(9), e7179. doi: 10.1371/journal.pone.0007179

Yang, X., Vadrevu, R., Wu, Y., & Matthews, C. R. (2007). Long-range side-chain-main-chain interactions play crucial roles in stabilizing the (betaalpha)8 barrel motif of the alpha subunit of tryptophan synthase. *Protein Sci, 16*(7), 1398-1409. doi: 10.1110/ps.062704507

Zeng, J., Jiang, F., & Wu, Y. D. (2016). Folding Simulations of an alpha-Helical Hairpin Motif alphatalpha with Residue-Specific Force Fields. *J Phys Chem B, 120*(1), 33-41. doi: 10.1021/acs.jpcb.5b09027

Zhang, Y., & Skolnick, J. (2004a). Scoring function for automated assessment of protein structure template quality. *Proteins, 57*(4), 702-710. doi: 10.1002/prot.20264

Zhang, Y., & Skolnick, J. (2004b). SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem, 25*(6), 865-871. doi: 10.1002/jcc.20011

Zhou, H. X. (2004). Loops, linkages, rings, catenanes, cages, and crowders: entropy-based strategies for stabilizing proteins. *Acc Chem Res, 37*(2), 123-130. doi: 10.1021/ar0302282

Zhou, H. X., Hoess, R. H., & DeGrado, W. F. (1996). In vitro evolution of thermodynamically stable turns. *Nat Struct Biol, 3*(5), 446-451.

Zhou, R. (2007). Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol Biol, 350*, 205-223.

Zhou, X. X., Wang, Y. B., Pan, Y. J., & Li, W. F. (2008). Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids, 34*(1), 25-33. doi: 10.1007/s00726-007-0589-x

Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A., & Matthews, C. R. (1999). Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a beta/alpha barrel protein. *Protein Sci, 8*(6), 1200-1209. doi: 10.1110/ps.8.6.1200

Zomot, E., & Kanner, B. I. (2003). The interaction of the gamma-aminobutyric acid transporter GAT-1 with the neurotransmitter is selectively impaired by sulfhydryl modification of a conformationally sensitive cysteine residue engineered into extracellular loop IV. *J Biol Chem, 278*(44), 42950-42958. doi: 10.1074/jbc.M209307200

Biography of Mr. K. Rajashekar Varma

Mr. K. Rajashekar Varma is a Ph. D student in the Department of Biological Sciences. He obtained M. Sc. in Biotechnology from Bangalore University in the year 2010 and joined Metaome Science Pvt Ltd in January 2011 as BCIL trainee, in the interest of learning computational methods for Data mining, Database development and data analysis. Mr. Rajashekar Varma was selected for the Ph.D. program in January 2012 at BITS, Pilani-Hyderabad campus at the Department of Biological Sciences under the Supervision of Dr. Ramakrishna Vadrevu.

He is well trained in experimental lab work and computational studies. During his work he gained proficiency in programming languages like PYTHON and Perl and also developed a web GUI based database with flexible query tools for assisting protein engineering studies. He also gained expertise in expression and purification of ^{15}N and ^{13}C labelled protein for NMR data acquisition and NMR data analysis software's like SPARKY, BRUKER etc. for his thesis. Currently his research interests focus on analysis of $\alpha\beta$ and $\beta\alpha$ loops from TIM barrel proteins.

Biography of Prof. Ramakrishna Vadrevu

Prof. Ramakrishna Vadrevu is currently working as an Associate Professor, Department of Biological Sciences. After obtaining Ph.D. from Indian Institute of Technology, Bombay, he joined as a Post Doc Fellow at Pennsylvania State University and later as a Faculty member in University of Massachusetts Medical School. He joined BITS Pilani-Hyderabad Campus in the year 2008.

His research interests include Protein Design & Engineering: Self-assembly and Bio nano-materials; NMR spectroscopy / Biophysical Approaches to understand protein Folding / mis-Folding and Dynamics in vitro & in vivo and Drug discovery. He is a member of Protein Society and authored more than 20 publications in peer reviewed international journals and chaired in some international conferences. He has served as a reviewer for research proposals submitted to funding agencies, manuscript submitted to journals and theses.

His administrative contributions include serving as the Head, Department of Biology (2012-2014), Chairperson DRC-Departmental Research Committee (2012-2014), General Secretary, Technology Business Incubator Society (2012 onwards).

Structure and Sequence Level Analysis of βαβ Motifs and Loops in TIM Barrel Proteins: Implications for Protein Folding, Design and Engineering

THESIS

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

K.RAJA SHEKAR VARMA

ID. No: 2011PHXF404H

Under the supervision of

Prof. Ramakrishna Vadrevu

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

HYDERABAD CAMPUS

2017

Chapter 6

CONCLUSIONS AND FUTURE PERSPECTIVES

CONCLUSIONS

In summary, this thesis presented (i) a comprehensive analysis of loops from TIM barrel proteins (ii) compiled a detailed library of loops from protein structures in the form of a database, which is powered by a graphical user interface, facilitating users with a diverse set of query-based search options to extract and compare the sequence and structural features of loops (iii) detailed structural and dynamic properties of loops and their role in stability and function of TIM barrel proteins employing both structural and molecular dynamics simulations (iv) identification, analysis and assessment of plausible independently folding/stable βαβ structural motifs from TIM barrel proteins.

The thesis presents a comprehensive compilation and analysis of αβ and βα loops which indicate that size, conformation, and sequence features and preferences differ between the αβ and βα loops of TIM barrel proteins. The αβ loops are dominated by smaller loops in contrast to the longer loops in βα loops. Constrained, short αβ loops can assist establishing efficient hydrophobic packing between the flanking helices and strands. It is expected that the longer βα loops can have a higher degree of freedom to adopt random conformations assisted by multiple turns, required for geometrical flexibility leading to function.

Also, the presence of ordered conformations, in general, and type I turns, in particular, could be the contributing factors for the rigidity of αβ loops. The other likely factor that may also endow rigidity to the αβ loops is the higher proportion of long range side chain hydrogen bonding interactions involving the two positively charged residues, arginine, and lysine.

Glycine residues appear to play an important role in adopting conformations that will result in tight turns leading to optimal packing between the helices and strands of the αβ hairpins. Moreover, lacking side chain, it can maximize the interactions between the helix and strand. In summary, the distinctive role of αβ and βα loops in stability and function, respectively, perhaps is reflecting in their size, sequence profiles, and conformations.

In summary, the present study will help to facilitate short-listing of the candidate loops to be exchanged in place of a target loop. Apart from assisting engineering TIM barrel proteins, the compilation of the αβ and βα loops shed light on the distribution of turn types and preferred template sequences, particularly, in short loops connecting the α and β elements in αβ/βα hairpins. This may encourage experimentalists in designing and or identifying sequences that can adopt α/β structures in solution. As per the findings in this study, four residue loops adopting type I turn conformation may be best suited for connecting the α helix to β strand in αβ hairpins. Analyses from experimental and bioinformatics approaches clearly indicate that the βαβ modules, serve as the minimal unit of stability in β/α class of proteins. Our observations and reasoning will in addition to steering protein engineering efforts on TIM barrel design and stabilization can provide the basis for identifying and or designing independently folding stable βαβ modules.

The analysis from structures and molecular dynamics simulations further provided more insights on the flexibility and dynamic nature of αβ and βα loops in tryptophan synthase alpha subunit, a TIM barrel protein, in particular. Detailed residue level analysis, indicates that αβ loops are more rigid compared to βα loops and alpha helices and may thus play a significant contributing role in the overall stability of the fold. The analysis further indicates that the hydrophobic and long range side chain main chain interactions can play a crucial role in restricting the flexibility of the loops, in αTS, in particular, and TIM barrel proteins,

in general. These observations and findings from the current study will greatly assist in designing independently folding αβ and βαβ units, in addition, to assist engineering stability in TIM barrel proteins by manipulating and or introducing αβ loops at appropriate positions.

The LoopX database presented in this work provides comprehensive information on sequence, conformation, hydrogen bonding interactions of protein loops. Equipped with efficient search tools/algorithms and a visualizer, the database extracts compatible loop candidates for a chosen target loop, thus providing an opportunity for comparing their sequence and structural level information. To the best of our knowledge, LoopX is a comprehensive web-based archive of protein loops which also provides the facility to examine loops and analyze them for their structural similarity, hydrogen bonding interactions, the backbone ϕ, ψ plots. In summary, LoopX serves as a comprehensive graphical user interface driven database for both analysis and comparative evaluation of protein loops, thus adding and complementing to other active databases on protein loops.

The analysis and study of independently folding βαβ motifs from TIM barrels provide opportunities to test the shortlisted βαβ sequences to adopt ordered and stable structures. Given the fact that complex protein structures have evolved from small independently folding super secondary structures, such as βαβ motifs, the exhaustive sequence and structural features based prediction can lead to the identification likely βαβ candidates from TIM barrel proteins that can fold independently. The search methodology and the proposed strategy employed in this process are based on the analysis of loops and the contribution of stabilizing interactions arising from within the loops in addition to basic principles/theories of protein folding and stability. Apart from the identification of likely βαβ candidates that can fold independently this work will pave way for development of algorithms that can

predict independently folding βαβ motifs which are the basic building blocks of TIM barrel proteins.

FUTURE PERSPECTIVES

The design of novel proteins or engineering existing proteins for desired features has been the researchers' efforts. Insights from loops analysis and dynamics work will greatly assist in future protein folding/design/engineering studies.

Selection of compatible loop candidates for successful loop grafting in proteins has been taxing. The challenge of identifying compatible loops for grafting will be greatly aided of LoopX. The comparative analysis of loops at the level of sequence and structure will in addition, to guiding most promising candidates, will help in minimizing the number of candidates to be tested.

The prediction of promising βαβ sequences from TIM barrels will open up opportunities for expression and purification of the sequences and further designing super-secondary βαβ motifs. This also provides insights into exploring nature's secrets that assist the motifs to fold independently combining these motifs to build/engineer new proteins.

Limitations:

The efficiency of LoopX needs to be evaluated experimentally. The foldability and stability of the proposed βαβ candidates need experimental verification.