**MIAMI UNIVERSITY**
**The Graduate School**

Certificate for Approving the Dissertation

We hereby approve the Dissertation

of

Ruojing Zhang

**Candidate for the Degree**

DOCTOR OF PHILOSOPHY

_____
Michael A. Kennedy, Advisor

_____
David L. Tierney, Committee Chair

_____
Carole Dabney-Smith, Reader

_____
Rick Page, Reader

_____
Paul Urayama, Graduate School Representative

# ABSTRACT

## THE INVESTIGATION OF BIOPHYSICAL AND BIOLOGICAL FUNCTION OF PRPS FROM *NOSTOC* PCC 7120

by

Ruojing Zhang

Nostoc sp. PCC 7120 are filamentous cyanobacteria capable of both oxygenic photosynthesis and nitrogen fixation, with the latter taking place in specialized cells known as heterocysts that terminally differentiate from vegetative cells under conditions of nitrogen starvation. Pentapeptide repeat proteins (PRPs), which occur most abundantly in cyanobacteria, adopt a right-handed quadrilateral β-helical structure, also referred to as a repeat five residue (Rfr) fold, with four-consecutive pentapeptide repeats constituting a single coil in the β-helical structure. Despite their intriguing structure and importance to understanding ancient cyanobacteria, the biochemical function of PRPs in cyanobacteria remains largely unknown. Here we report the crystal structure of Alr5209 and Alr1298, two PRPs from Nostoc sp. PCC 7120 predicted to be involved in oxidative phosphorylation and response to nitrogen starvation and/or heterocyst differentiation. The Alr5209 structure was analyzed in comparison to all other PRPs to determine how type I β turns can be accommodated in Rfr folds and the consequences of type I β turns on the right-handed quadrilateral β-helical structure. Given that Alr5209 represents the first PRP structure containing type I β turns, the PRP consensus sequence was reevaluated and updated. The structure of Alr1298 displays the typical right-handed quadrilateral β-helical structure and includes a four-α-helix cluster capping the N-terminus and a single α helix capping the C-terminus. Furthermore, we provide the preliminary investigation of the function of Alr5209 and Alr1298 by measurement of phenotype and localization. Protein β turn classification remains an area of ongoing development in structural biology research. We recently encountered a specific problem when classifying β-turns in crystal structures of pentapeptide repeat proteins (PRPs) determined in our lab that are largely composed of β-turns that often lie close to, but just outside of, canonical β-turn regions. To address this problem, we devised a new scheme that merges the Klyne-Prelog stereochemistry nomenclature and definitions with the Ramachandran plot. The resulting Klyne-Prelog-modified Ramachandran plot schema defines 1296 distinct potential β-turn classifications that cover all possible protein β-turn space with a nomenclature that indicates the stereochemistry of i+1 and i+2 backbone dihedral angles. The utility of the new classification scheme was illustrated by re-classification of the β turns in all known protein structures in the PRP superfamily and further assessed using a database of 16657 high-resolution protein structures (≤1.5 Å) from which 522776 β turns were identified and classified.

# THE INVESTIGATION OF BIOPHYSICAL AND BIOLOGICAL FUNCTION OF PRPS FROM NOSTOC PCC 7120

**A DISSERTATION**

Presented to the Faculty of

Miami University in partial

fulfillment of the requirements

for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

by

Ruojing Zhang

The Graduate School
Miami University
Oxford, Ohio

2021

Dissertation Director: Michael A. Kennedy

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# DEDICATION

This work is dedicated to my family for their infinite trust and support. To my father, Hong Zhang, and my mother Shiru Yu, thank you for your devotion to me including but not limited to love, education, life, future guidance and so on. To my grandparents, Quanqing Yu and Feng Wang, thank you for your encouragement and nurture in my childhood. I also dedicate this dissertation to all my relatives in both my father's and mother's family, thanks for supporting and guiding me in life. Last but not least, I dedicate this dissertation to all my friends, thanks for encouragement and companionship.

ACKNOWLEDGEMENTS

Firstly and most importantly, I would like to show my deepest respect and appreciation to my advisor, Dr.Michael Kennedy. Thank you for giving me the opportunity to complete my dissertation study. Without your trust and support, it is impossible for me to make any progress in the research life. Knowledge is the best charity. In the process towards my future research career, not only do you provide me with the powerful tools to study the structural biology and the biochemical function of proteins, but also you guide me to learn the method of thinking as well as the scientific and scrupulous intellection. Both are the treasures of my career and life and will assist me to solve more problems in the future. Thank you for all your help and support which benefit the rest of my life, I sincerely show my gratitude and appreciation. I hope all good luck and nice blessings will be with you in the future.

In addition, I would like to show my appreciation to Dr.Shuisong Ni. Thank you for guiding me to learn all experimental skills in the laboratory. Those skills are essential for my future research studies and powerful tools to investigate the evidence to prove my research opinions. Furthermore, thank you for the guidance and support in my future development and life. Your suggestions provide me the solid foundation to move forward to the next step. I sincerely appreciate all you have done for me and encouragement to me. They are the warmest support to inspire me to complete the dissertation. I hope all best wishes follow you in the future.

I would also like to show my appreciation to Dr.Loralyn Cozy in Illinois Wesleyan University and Prof.Michael Stahr in College of Engineering in Miami University. Without your corporation and guidance, I could not complete my projects and dissertation.

Also, I would like to thank all the current and former graduate students in Dr. Kennedy's Lab for helping each other.

Last but not least, I would like to thank all my committee members: Dr. David L.Tierney, Dr. Carole Dabney-Smith, Dr. Rick Page and Dr. Paul Urayama for the help and guidance of my research life in Miami and future including but limiting to the courses, examination and future development.

# Chapter 1: Introduction

The subject of this dissertation is an investigation of the structure and function of pentapeptide repeat proteins (PRPs) from the filamentous cyanobacterium Nostoc sp. st. PCC 7120. Cyanobacteria are one of the most important and abundant prokaryotes found in multiple diverse environments on earth. It is widely believed that cyanobacteria are the first main group of microorganisms with the function of photosynthesis. The current distribution of PRPs in PF00805 Pfam indicates that 38,981 PRP sequences distributed over 3,338 species. PRPs are found most abundantly in cyanobacteria, with 26.9% of all PRP sequences occurring in cyanobacteria, which represents only 3.7% of the species in which PRPs have been discovered, indicating that PRPs likely played an important physiological or structural role in the evolution and lifecycle of ancient cyanobacteria. However, the knowledge of function and structure to PRPs still remains largely unknown. The dissertation includes the structural studies relating two PRPs and is organized into research four chapters, as described below.

Chapter two is a review of the current understanding of the function and structure of PRPs. PRPs represent a large superfamily of proteins with almost 39,000 sequences identified in more than 3,300 species. However, remarkably little is known about the structure and biochemical function of PRPs. In the past 26 years, since the first PRPs were identified in 1995, only 16 PRP structures have been solved by X-ray crystallography and the biochemical function of only two PRPs have been determined. In this review, PRPs were grouped into six categories based on known or putative biochemical functions or based on known structures and their structures and functions were analyzed and discussed.

In chapter three, the structure of Alr5209, a 129 amino acid PRP from Nostoc sp. st. PCC 7120 is presented, described and analyzed. Alr5209 was the first PRP structure to incorporate type I beta turns into its β helix. Due to the unique presence of type I β turns in the Alr5209 structure in comparison to other PRPs, the influence of different combinations of β turn types on the structures of PRPs was analyzed, including the dimensions and helical twist of the β helix. To further investigate the biophysical characteristics of Alr5209, the electrostatic potential surface was analyzed an compared to other PRPs. The stability of Alr5209 was investigated using CD melting experiments. Based on the new Alr5209 structure, the pentapeptide repeat sequence consensus was revised. To investigate the potential function of Alr5209, a gene cluster analysis was performed. Based on the annotated function of other proteins belonging to the same putative operon, it was concluded that Alr5209 might be involved in the process of oxidative phosphorylation.

In chapter four, the structure of Alr1298, a 169 amino acid encoded PRP from Nostoc sp. st. PCC 7120 is presented, described and analyzed. The Alr1298 structure adopted a four-coil Rfr β helix capped by a four-α-helix bundle at its N-terminus. The stability of Alr1298 was investigated using a CD melting analysis. The solution behavior of Alr1298 was investigated by determining the rotational correlation time based on NMR relaxation analysis at room temperature. PISA analysis was performed to investigate the potential flexibility an independent motion of the N-terminal helix-bundle relative to the Rfr domain. The electronic surface potential of Alr1298 was investigated and compared to that of all other PRPs with known structures. The potential biochemical function of

Alr1298 was investigated by performing genetic analysis of the proteins encoded by flanking genes potentially belonging to a common putative operon with Alr1298, indicating that Alr1298 may play a role in the response to nitrogen starvation or in the process of heterocyst differentiation.

In chapter five, a new schema is introduced to classify proteins β turns. The motivation for the new schema was inspired by 1) the fact that PRP structures consist almost exclusively of β-ladders joined by β-turns although the β turns can vary in type and composition, 2) the analysis of the new Alr5209 and Alr1298 PRP structures described in Chapters 3 and 4, respectively, and 3) based on our new analysis and comparison with other PRP structures. Specifically, our investigations revealed the structures of PRPs involved many type IV β-turns, which is a catch-all category for all β-turns that do not belong to type I or type II, and therefore do not specify any particular structure characteristics. To overcome this limitation, a new schema was established based on organic chemistry stereochemistry definitions and conventions to eliminate the intrinsic ambiguity of "border β turns" that inevitably end up being grouped into the type IV β turn category. In the study, 16657 protein structures of all categories with resolution less than 1.5 Å were included in the database to evaluate the new schema. Application of the new schema to the β turn database resulted in identification of 582 new turn types. A hydrogen bond analysis was performed to summarize the occurrence of hydrogen bonds in each new turn type. Since the distance between Cα of the first and the last residue in β turn is an important parameter to identify β turns, the distribution of this distance was analyzed for each new turn type. The amino acid distribution for each turn type was evaluated to determine if an amino acids preference existed for each new turn type. The analysis was also applied to all known PRP structures as well. Collectively, the analysis illustrated the capability of the new schema to resolve common ambiguities present in the analysis, classification and description of β turns in protein structures.

In chapter six, we summarized the results from three research projects (chapter three to chapter five) and concluded the previous and current studies relating to the target proteins. Following the current conclusion, we outline potential future studies that could be performed to investigate the biochemical function of the Alr5209 and Alr1298 PRPs.

# Chapter 2: Current Understanding of the Structure and Function of Pentapeptide Repeat Proteins

Reproduced with permission from:

Ruojing Zhang[1], Michael A. Kennedy*[1]

[1]Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056

*Corresponding Author: Department of Chemistry and Biochemistry, 106 Hughes Laboratories, Miami University, 651 East High Street, Oxford, OH 45056. Email: kennedm4@miamioh.edu. Phone: 513-529-8267. Fax: 513-529-5715.

This paper will be submitted for publication.

Author contributions: RZ contributed to data collection, data analysis, manuscript preparation. MAK contributed to data analysis and manuscript preparation.

## 2.1 Abstract

The pentapeptide repeat protein (PRP) superfamily, identified in 1998 by Bateman et al., has grown to nearly 39,000 sequences from over 3,300 species. PRPs, recognized as having at least eight contiguous pentapeptide repeats (PRs) of a consensus pentapeptide sequence, adopt a remarkable structure, namely, a right-handed quadrilateral β-helix with four consecutive PRs forming a single β-helix coil. Adjacent coils join together to form a β-helix "tower" stabilized by β-ladders on the tower faces and type I, type II or type IV β-turns facilitating an approximately -90° redirection of the polypeptide chain joining one coil face to the next. PRPs have been found in all branches of life, but they are predominantly found in cyanobacteria. Cyanobacteria have existed on earth for more than two billion years and are thought to be responsible for oxygenation of the earth's atmosphere. Filamentous cyanobacteria such as Nostoc sp. strain PCC 7120 may also represent the oldest and simplest multicellular organisms known to undergo cell differentiation on earth. Knowledge of the biochemical function of these PRPs is essential to understanding how ancient cyanobacteria achieved functions critical to early development of life on earth. PRPs are predicted to exist in all cyanobacteria compartments including thylakoid and cell-wall membranes, cytoplasm and thylakoid periplasmic space. Despite their intriguing structure and importance to understanding ancient cyanobacteria, biochemical functions of PRPs in cyanobacteria remain almost completely unknown. The precise biochemical function of only a handful of PRPs is currently known from any organisms, and three-dimensional structures of only 16 PRPs or PRP-containing multidomain proteins from any organism have been reported. In this review, the current knowledge of structures and functions of PRPs is presented and discussed.

## 2.2 Introduction

The first description of a pentapeptide repeat protein (PRP) was reported in 1995, when Haselkorn and coworkers[3] identified a gene from the filamentous cyanobacterium Nostoc (formerly Anabaena) sp. strain PCC 7120, which when mutated, altered the composition of glycolipids encasing the heterocysts, among other alterations (**Figure 2.1**). They named the gene heterocyst-specific glycolipids-directing protein K (hglK) for the role that the gene played in localization of glycolipids to heterocysts. The protein encoded by the hglK gene, HglK, was predicted to contain four trans-membrane spanning regions and an unusual alanine- and leucine-rich pentapeptide repeat (PR) region made up of 36 PRs with the consensus sequence AXLXX.[3] Therefore, HglK was the first PRP to be associated with a putative biochemical function. To date, however, the precise mechanism for the role that HglK plays in regulating glycolipid localization to heterocysts remains unknown.

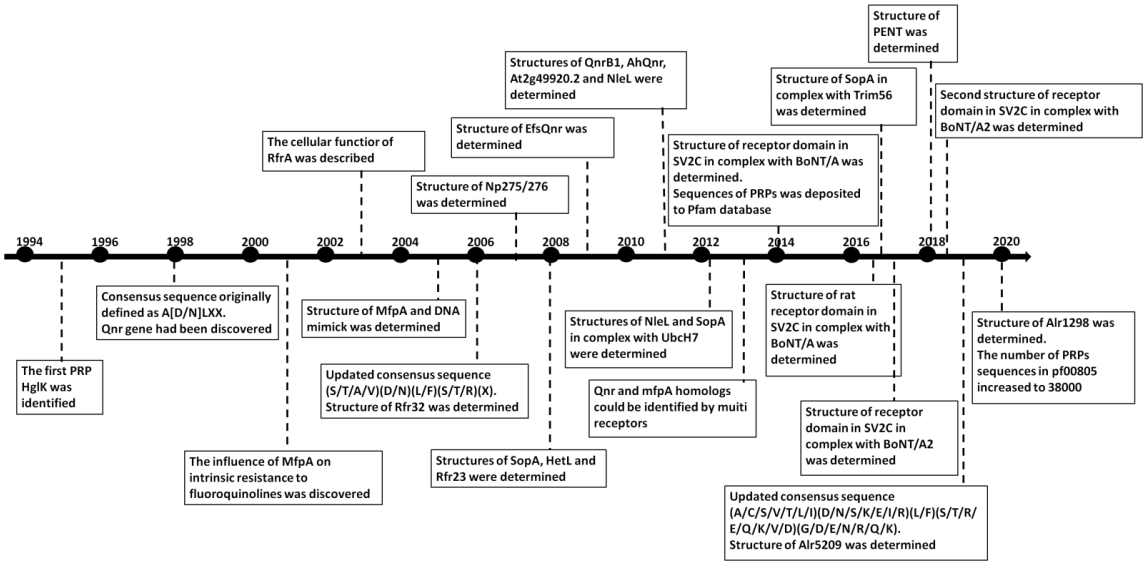



**Figure 2. 1 Milestones in the timeline investigation of the structure and function of PRPs.**

In 1998, Bateman et al.[5] reported the discovery of a novel family of proteins, to which HglK belonged, that contained tandem PRs with the sequence motif A(D/N)LXX, based on the analysis of recently determined complete genomes of several bacteria at the time. They observed that PRPs were most commonly found in cyanobacteria[6]. The authors also proposed a model of PRP structures, rightly predicting that PRPs would adopt a right-handed β helical architecture, however they predicted a triangular-shaped helix, which would prove to be in error once the first three-dimensional structures of PRPs were determined several years later.

Also in 1998, Martínez- Martínez confirmed that quinoline resistance in bacteria could be carried on a multi-resistance plasmid (pMG252), which they discovered in a clinical isolate of Klebsiella pneumonia.[7] We now know that bacterial acquisition of antimicrobial resistance undergoes constant evolution with horizontal gene transfer through plasmids playing a major role.[8] In 2001, Montero et al. discovered that intrinsic resistance to fluoroquinolines was influenced by MfpA, a putative PRP-containing protein encoded by a chromosomal gene in Mycobacterium smegmatis.[9] In 2005, Hegde et al.[10] solved the three-dimensional structure of MfpA from Mycobacterium tuberculosis, a homologue of MfpA from M. smegmatis, representing the first three-dimensional structure of a PRP, revealing that it adopted a right-handed quadrilateral β helical structure. Hegde et al. also reported that the structure and electrostatic charge distribution of MfpA mimicked that of DNA, thereby conferring fluoroquinoline resistance to M. tuberculosis due to its ability to bind to DNA gyrase and inhibit its function.[10] Soon after, many more chromosomal genes encoding homologs of MfpA were discovered in the genomes of a variety of organisms and in 2013, Jacoby and Hooper reported a phylogenetic tree analysis that showed that quinoline resistance genes (qnr) and mfpA homologs could be identified in 58 Gram-negative bacteria, 34 Gram-positive organisms and 14 plasmid-mediated genes.[11]

In 2003, Chandler et al.[12] ascribed a cellular function to RfrA, a PRP from the cyanobacterium Synechocystis 6803, showing that it played a role in regulating a novel manganese uptake system, however, the nature of the system and the precise role that RfrA plays in regulating the manganese uptake system remains unknown.

By 2006, Vetting et al.[13] reported that the PRP family had grown to more than 500 members in the prokaryotic and eukaryotic kingdoms and they updated the PR consensus sequence as [S,T,A,V][D,N][L,F][S,T,R][G], and, in 2009, Buchko reviewed the knowledge of the structure and function of PRPs from cyanobacteria.[14] In 2014, Shah and Heddle reported that a query of the Pfam database (http://pfam.xfam.org) for members of the PR family (PF00805) had expanded to 11,082 sequences from 1,513 species with protein structures having been solved for a number of PRPs from Nostoc. sp. PCC 7120[6, 15], Cyanothece 51142,[13, 16] Arabidopsis thaliana, Enterococcus faecalis,[17] K. pneumoniae[18], Xanthomonas albilineans,[18] Aeromonas hydrophila[19] and M. tuberculosis.[10] In 2019, Zhang et al. updated the PR consensus sequence to (A/C/S/V/T/L/I)/(D/N/S/K/E/I/R)/(L/F)/(S/T/R/E/Q/K/V/D)/(G/D/E/N/R/Q/K) based on the consideration of several newly available PRP crystal structures.[5, 10, 20] By 2020, the number of PRPs in the PF00805 Pfam had increased to 38,000 sequences in nearly 3,500 sequences[21]. The current distribution of PRPs in PF00805 Pfam is depicted in **Figure 2.2**, indicating 38,981 PRP sequences distributed over 3,338 species (https://pfam.xfam.org/family/PF00805#tabview=tab7). In this sunburst plot, 82.2% of the species and 84.7% of the sequences belong to bacteria, 14.1% of species and 13.7% of sequences belong to eukaryota, 0.5% of species and 1.4% of sequences belong to viruses and 1.1% of species and 2.2% of sequences belong to archaea. The plot also indicates that PRPs are found most abundantly in cyanobacteria, with 26.9% of all PRP sequences occurring in cyanobacteria, which represents only 3.7% of the species in which PRPs have been discovered, indicating that PRPs likely played an important physiological or structural role in the evolution and lifecycle of ancient cyanobacteria. Despite the large and growing nature of the PRP superfamily, three-dimensional structures of only sixteen PRP or PRP-containing proteins have been determined, thirteen of which contain a single PRP domain with α helices capping the N and/or C termini and three of which contain two or more domains including the PRP domain.

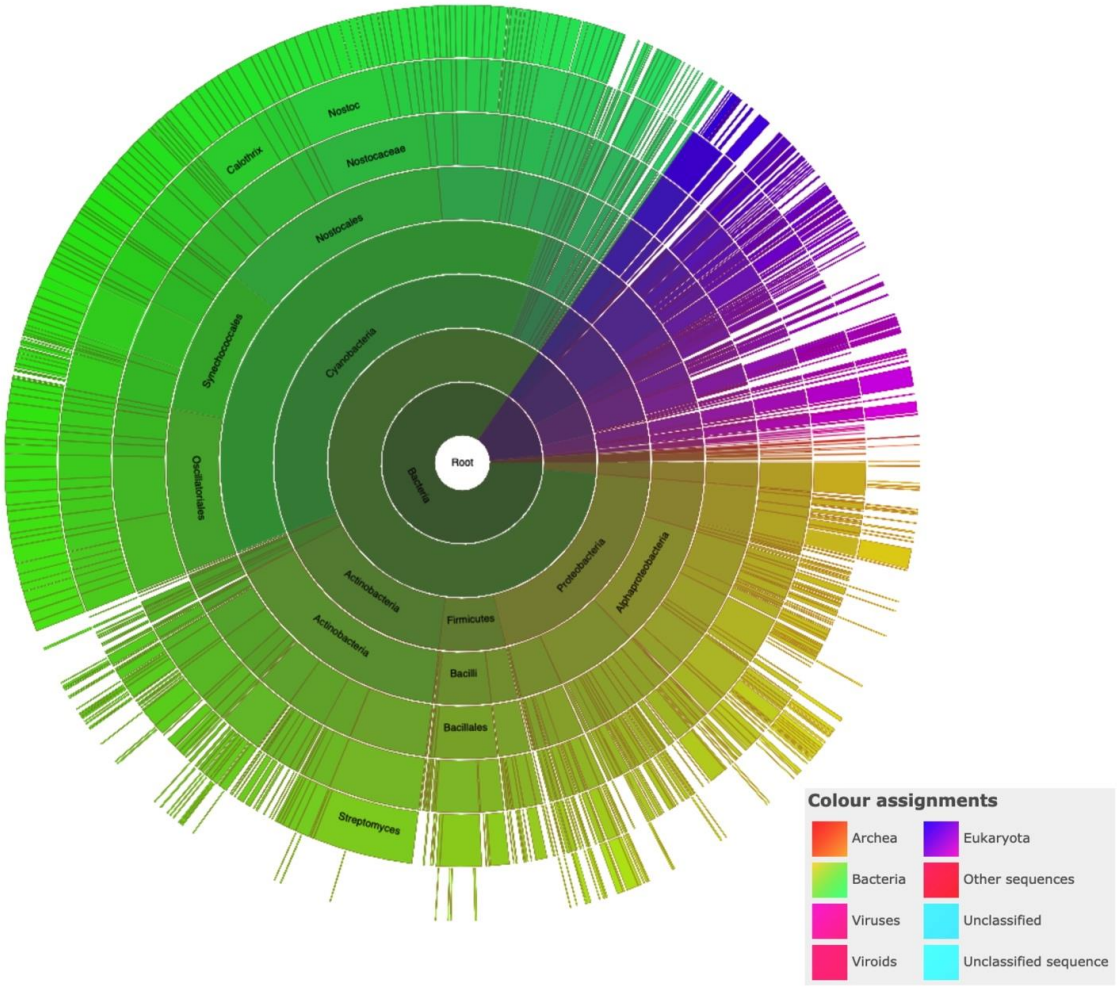



**Figure 2. 2 Distribution of PRP sequences across species.** This sunburst plot of the PF00805 PRP Pfam shows the distribution of 38981 sequences across 3338 species. The color-coding in the sunburst plot is indicated in the legend.

In this review, several PRP categories are discussed, including those that have putative associated biochemical or cellular functions and those that have had structures determined but with unknown putative functions, including those involved in 1) heterocyst glycolipid synthesis, 2) manganese uptake, 3) gyrase inhibition, 4) ubiquitin E3 ligases, 5) synaptic vesicle glycoprotein 2 isoform C (SV2C) receptors and 6) plant and cyanobacteria proteins with three-dimensional structures but no functional characterization (**Figure 2.3**). Although the biological functions of most PRPs remain unknown, three-dimensional structures of PRPs and PRP-containing multidomain proteins continue to be solved and reported with the hope of helping to eventually understand their biological, biochemical or cellular functions.

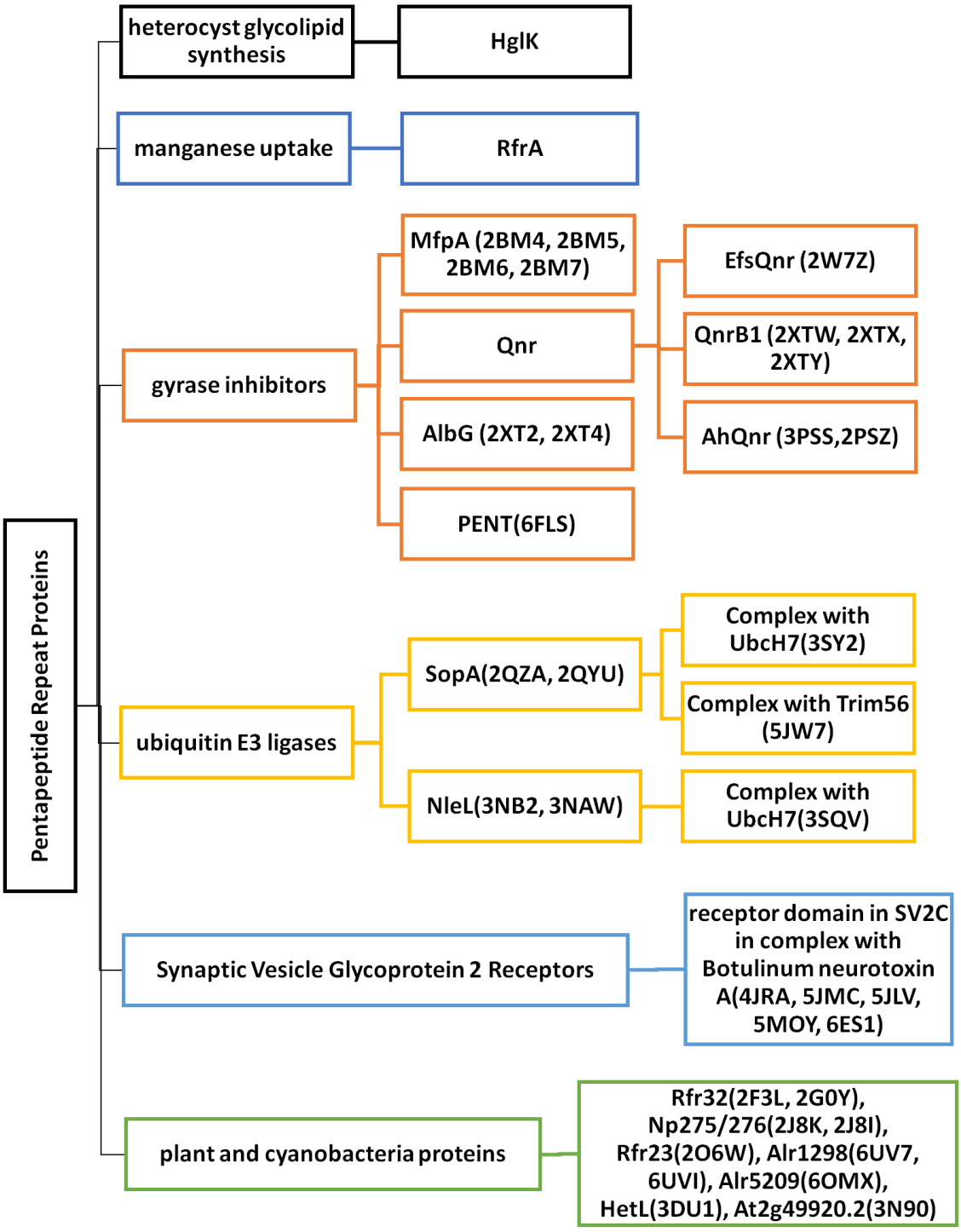



**Figure 2. 3 Summary of the PRPs discussed with and without known three-dimensional structures.** The six category groups are show in the first branch. The second and subsequent branches indicated specific PRPs. PRPs with known structures include the corresponding PDB ID inside parentheses immediately to the right of the PRP name.

## 2.3 Cyanobacterial and Eubacterial PRPs with an Associated Biochemical or Cellular Function

### 2.3.1 heterocyst glycolipid biosynthesis – HglK

In 1995, Haselkorn and coworkers identified the hglK gene in mutant strain 543 of the filamentous Nostoc sp. strain PCC 7120. The mutant strain was isolated as a $Fox^-$ (lack of ability to fix dinitrogen in an oxygen-depleted environment) mutant following chemical mutagenesis.[3] The ultrastructural phenotype of the mutant strain showed that, in nitrogen replete media, i.e. media containing an abundant usable soluble nitrogen source, the vegetative cells in the filaments, i.e. those cells capable of dividing and extending the filament length, were more cylindrical and had thicker septa compared to the wild-type strain whereas in nitrogen depleted media, the mutant heterocysts lacked the glycolipid layer that is normally exterior to the cell wall and isolates the nitrogenase enzyme that is required for fixation of atmospheric nitrogen inside the heterocysts from oxygen that inactivates the nitrogenase enzyme. Hydropathy analysis indicated that the 727 amino acid HglK protein contained four potential trans-membrane regions in its N-terminal region and 36 PRs in its C-terminal region starting at amino acid position 501. Analysis of the mutant strain indicated that it contained a stop codon just upstream of the DNA encoding the PRP domain. Because heterocysts in the mutant strain lacked the glycolipid layer exterior to the cell wall, the authors used thin-layer chromatography to analyze the lipid content of the mutant and wild-type strains and found no difference. The authors therefore concluded that the hglK gene encoded a protein that was necessary for localization of glycolipids to the heterocyst walls and that this function required the PRP domains.

Arévalo and Flores further characterized the function of the hglK gene and discovered that hglK mutants were also defective in heterocyst differentiation, being impaired in the expression of the heterocyst-related genes coxB2A2C2 (a cytochrome c oxidase) and nifHDK (a nitrogenase)[22]. The authors also observed that HglK was predominantly localized at the intercellular septa and was required for biogenesis of long filaments, to produce normal numbers of nanopores, and for normal intercellular molecular transfer activity. The authors concluded that HglK contributed to the architecture of the intercellular septa and impacted the function of septal junctions.[22] The precise biochemical role of HglK remains unknown and no three-dimensional structure of HglK is currently available.

### 2.3.2 Regulator of manganese uptake - RfrA

In 2003, Pakrasi and coworkers discovered that RfrA (gene name sll1350), a PRP from Synechocystis 6803, was a regulator of a novel high-affinity manganese uptake system.[12] RfrA and its function were identified in a suppressor screen in which the mutant strain was deficient in both mntC, a gene encoding a component of an ABC transport system for manganese, and psbO, which encoded an extrinsic manganese stabilizing protein of photosystem II. In a suppressor screen, one looks for additional mutations that reverse the mutant phenotype, in this case, deficiency in manganese transport. The authors discovered that a point mutation in rfrA restored photosynthetic activity of the □mntC, □psbO double deletion mutant. Radioactive manganese uptake experiments indicated that RfrA was a regulator of a high affinity manganese transport system that was different

from the known manganese ABC transport system. The authors named the 398 amino acid RfrA protein for the repeat five-residues (Rfr) domain, which is another name for PRPs, that occurred in the N-terminus of the protein. Genetic analysis indicated that Synechocystis 6803 contained 16 PRPs. RfrA was the first member of the PRP family to be linked to a specific physiological process. RfrA has no sequence or structural similarities to previously described bacterial manganese transcription factors and it does not have any known DNA-binding domains.[23] It has been postulated that RfrA may regulate the second manganese transporter through a mechanism other than transcriptional control, such as by reversible protein modifications at the post-translational level.[23] Despite the link to regulation of manganese uptake, the nature of the hypothetical second high-affinity manganese importer, its regulatory mechanism, and the precise biochemical role that RfrA plays regulating the putative manganese transporter remains unknown.[24] The three-dimensional structure of RfrA also remains unknown.

### 2.3.3 Gyrase inhibitors

#### 2.3.3.1 MfpA (2BM4[10], 2BM5[10], 2BM6[10], 2BM7[10])

In the early 2000s, due to growing antibiotic resistance of M. tuberculosis to two bactericidal compounds, isoniazid and rifampicin,[25] fluoroquinolines had become the most common new antibiotic therapy to treat M. tuberculosis infections.[8, 26] In 2001, Montero et al. identified a gene mfpA that encoded a PRP that conferred a new mechanism of fluoroquinoline resistance to M. smegmatis.[9] The protein encoded by the mfpA gene resulted in a low level of resistance to ciprofloxacin and sparfloxacine.[10] Hegde et al.[10] identified a 183-amino acid MfpA homolog from M. tuberculosis (MtMfpA) encoded by the Rv3361c gene that was 67% identical to the 192-residue M. smegmatis MfpA protein. Hegde et al.[10] reported the three-dimensional structure of MfpA from M. tuberculosis revealing it to be a PRP made up of eight complete PR coils (**Figure 2.4**). MfpA was the first PRP to have its three-dimensional structure solved. Hegde et al. reported that MfpA expression in vivo conferred resistance to the antibiotic fluoroquinolone. Fluoroquinolines are chemotherapeutic bactericidal drugs that interfere with DNA replication in bacteria, leading to bacterial cell death. Fluoroquinolines exert their antibacterial activity by interfering with the normal function of the type II topoisomerases, DNA gyrase and DNA topoisomerase IV. These enzymes normally cut the genomic DNA to allow supercoiling and then ligate the DNA to stabilize the supercoiled DNA. Fluoroquinoline acts by inhibiting the ligase activity of these enzymes and leaving the nuclease activity intact, resulting in accumulation of single- and double-strand breaks that leads to disrupted DNA replication and cell death. Fluoroquinoline acts by binding reversibly to the gyrase-DNA complexes and stabilizing the covalent enzyme tyrosyl-DNA phosphate ester that is normally a transient intermediate in the topoisomerase reaction. Hegde et al. reported that the three-dimensional structure of MfpA exhibited a size, shape and electrostatic surface that was similar to that of B-form DNA, and concluded that its mechanism of action was due to DNA mimicry.[10] Since fluoroquinoline only binds to DNA gyrase-DNA complexes, binding of MfpA to DNA gyrase blocks fluoroquinoline binding to DNA gyrase-DNA complexes, thus interfering with the bactericidal activity of fluoroquinoline.[10]

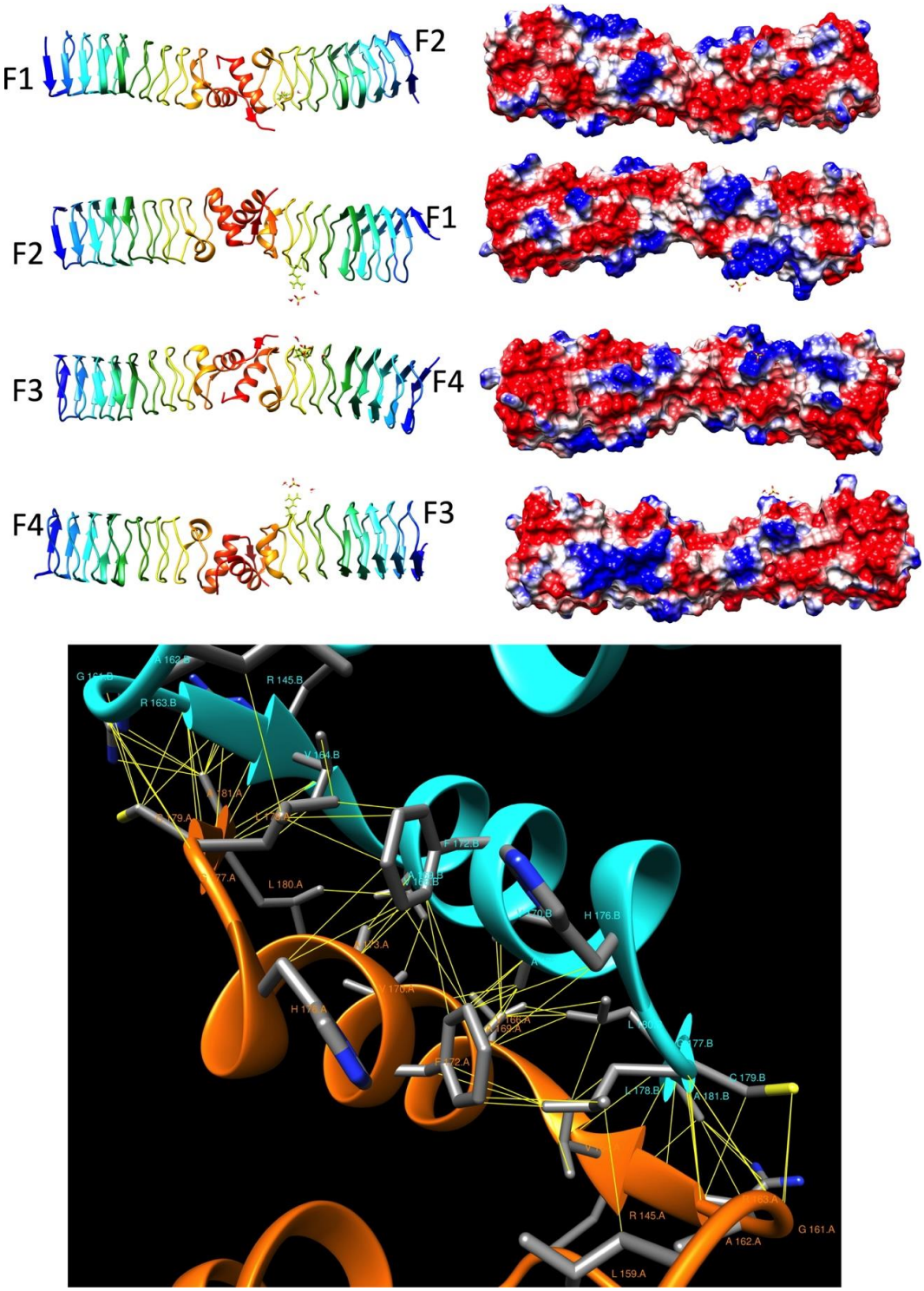



**Figure 2. 4 Structure and analysis of MfpA.** The top panel (at left) shows the ribbon diagrams looking at the four different faces of the Rfr coil the structure and the corresponding electronic potential surface analysis (at right). The bottom panel shows the head-to-head dimer interface with non-bonded interactions depicted by yellow lines as identified using Intersurf[1] and depicted using Chimera[4]. Chain A is colored orange, Chain B is colored cyan, and interacting residues are labeled.

Because its function relies on a mechanism of DNA mimicry, MfpA exists as functional dimer. Two molecules of MfpA undergo a head-to-head interaction mediated by two α helices at the C terminus of each molecule to form an asymmetric rod-like shape with a hydrophobic dimer interface (**Figure 2.4, top panel**). Based on the electrostatic surface potential of the MfpA dimer, Hegde et al. generated a model in which the β helices of the MfpA dimer interacted with N terminal region of gyrase dimer through electrostatic interactions. The secondary structure of MfpA contains a β bulge in its β helix and two α helices at the C terminus and 28 to 29 β turns (**Figure 2.4, top panel**).

The head-to-head interaction two molecules of MfpA involves 81 contacts mediated by 18 residues on one chain and 17 residues on the second chain that interact through hydrogen bonds and non-bonded contacts involving a surface area of about 1700 $Å^2$ (**Figure 2.4, bottom panel**). An interweaved interaction between two MfpA molecules is mediated by a short parallel β sheet involving residues 162-164 (Chain A) on one strand and residues 178-181 on the second strand (Chain B), an orthogonal interaction between α-helices (165-176 on each MfpA molecule), and another short parallel β sheet involving residues 178-181 (Chain A) and 162-164 (Chain B) (**Figure 2.4, bottom panel**). The dimer interaction is stabilized by a hydrophobic core formed by the interacting hydrophobic side chains from each α-helix involving F172 (12 interactions), H176 (2 interactions), L178 (9 interactions), plus the H-bonds stabilizing the parallel β-sheets, and multiple interactions involving R145 (six interactions), R164 (7 interactions) and C179 (7 interactions)(**Figure 2.4, bottom panel**). The energy to form the dimer interaction was -15.8 kcal/mole (2BM4). Considering, that a single hydrogen-bond has a typical energy of 1-3 kcal/mol, this indicates that the head-to-head dimer formation is energetically favored (**Figure 2.4, bottom panel**).

### 2.3.3.2 Qnr Family

#### 2.3.3.2.1 EfsQnr (2W7Z)[17]

In 2007, Arsene and Leclercq discovered a qnr-like gene from E. faecalis, EfsQnr, that conferred fluoroquinoline resistance to E. faecalis indicating that EfsQnr likely functioned as a DNA-gyrase inhibitor.[27] In 2009, Vetting et al.[17] determined the three-dimensional structure of EfsQnr revealing that the 211-residue protein was a PRP made up of eight complete Rfr coils composed of a mixture of type II and type IV β turns and a 12-residue C-terminal α helix capping the β helix. Two molecules of EfsQnr formed a head-to-head dimer mediated by an interaction between the C-terminal α helix of each EfsQnr molecule (**Figure 2.5, top panel**) similar to that observed in the structure of MfpA. Despite their structural similarity, a pairwise sequence alignment using EMBOSS Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) between EfsQnr and MfpA over 234 residues indicated just 19.7% sequence identity, 29.5% similarity and 30.3% gaps. The head-to-head dimer interaction in EfsQnr involved 16 residues on one chain (chain A) and 15 residues on a second chain (chain B) involving 69 contacts creating an interface area encompassing 1568 $Å^2$. The head-to-head interaction was mediated similarly to in MfpA, involving interactions between parallel regions of the polypeptide backbone (residues 190-194 on chain A and 207-210 on chain B), interactions of the sidechains of the orthogonally oriented α-helices from each chain (residues 195-206 on each chain), and backbone and sidechain interactions from another short section of parallel polypeptide chains (residues 207-211 on chain A and residues 191-194 on chain

B) (**Figure 2.5, bottom panel**). The hydrophobic interactions between the α-helices and adjacent strands was established by all hydrophobic, non-aromatic residues, including V94 (5 interactions), P196 (10 interactions), I200 (6 interactions), V209 (7 interactions) and I210 (7 interactions) and T211 (7 interactions) (Figure 5, bottom panel). The binding energy of the head-to-head dimer interaction was -14.9 kcal/mole, which was slightly weaker than in MfpA, that involved more residue interactions and a hydrophobic core involving aromatic side chains.

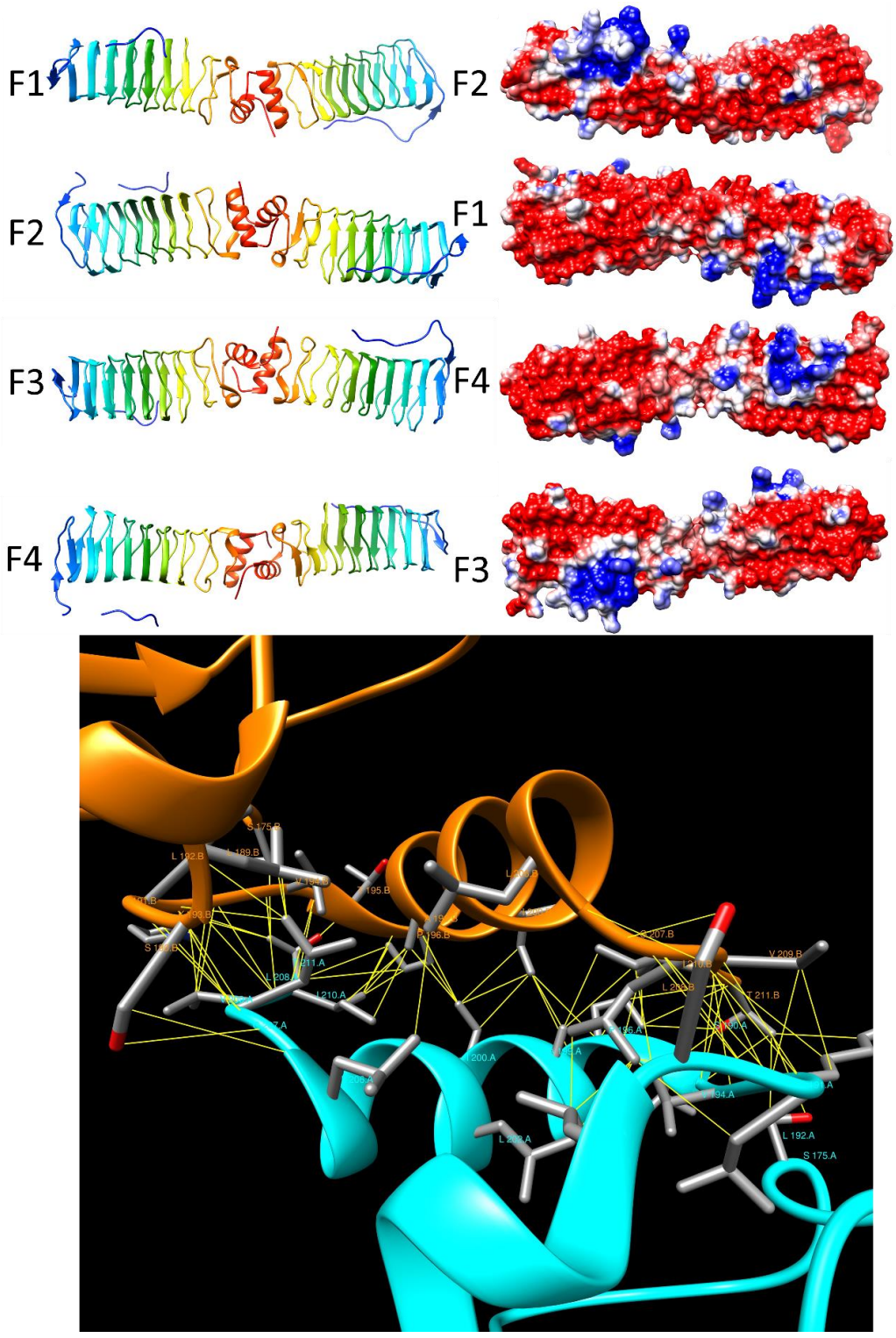



**Figure 2. 5 Structure and analysis of EfsQnr.** The top panel shows the ribbon diagrams of the structure for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel is the interface model calculated by Intersurf and depicted using Chimera.

#### 2.3.3.2.2 QnrB1 (2XTW[18], 2XTX[18], 2XTY[18])

QnrB1, encoded by multi-resistance plasmids in isolates of Enterobacteriaceae from around the world,[18] is a PRP that confers moderate fluoroquinoline resistance. Genes encoding QnrB1 can also be found on bacterial chromosomes.[28] QnrB1 belongs to the qnrB subfamily of Qnr genes that include qnrA, qnrB, qnrC, qnrD, qnrS and qnrVC subfamilies.[29-31]

The three-dimensional structure of QnrB1 from K. pneumoniae was determined by Vetting et al. in 2011.[18] The 226-residue protein contained nine Rfr coils, a 12-residue α helix capping its C terminus (D197-L208) and two loops projecting outward from the Rfr coil (a 8-residue loop (Loop A: Y46-G53) in coil 2 connecting face 2 to face 3 and projecting outward from face 2, and a 12-residue loop (Loop B: S102-S113) that projects outward from the corner between face 4 and face 1 joining coil 4 and col 5) (**Figure 2.6, top panel**). The head to head dimer was established by the interaction of the C-terminal helix of two QnrB1 molecules. The head-to-head dimer interaction was mediated by 40 non-bonded contacts involving 23 residues (chain A: 13 residues and chain D: 10 residues) for the interaction between chain A and chain D and 28 residues between chain B and chain C (chain B: 14 and chain C: 14) for the two dimers observed in the crystallographic asymmetric unit. The hydrophobic core was established by interactions between non-aromatic sidechains from each α helix, including I186 (5 interactions), M205 (8 interactions, and I210 (8 interactions) (**Figure 2.6, bottom panel**). The interface between the A and D chains had an area of 1131 $Å^2$, whereas the interface between the B and C chains had an area of 1432 $Å^2$. Despite the rather large difference between the interface areas, the interaction energies between two dimers composed of different pairs of chains in the asymmetric unit were quite similar with an average of -10.6 kcal/mole. The dimer interface in QnrB1 involved fewer residues (23) and fewer contacts (40) and weaker binding (-10.6 kcal/mole) compared to MfpA (35 residues, 81 contacts and -15.8 kcal/mole) and EfsQnr (31 residues, 69 contacts and -14.9 kcal/mole). The authors noted potentially interestingly positioned residues in the two loops that are also conserved, indicating that they may define a possible contact surface with topoisomerases to assist binding with gyrase in addition to the guiding electrostatic interaction.[18] The authors also pointed out that deletion of the smaller loop affects the inhibition to gyrase while the loss of the larger loop totally makes them lose the ability to inhibit gyrase.[18] In 2019, Li et al. showed that QnrB increases bacterial mutation rates and the selection of quinolone-resistant mutants.[32] Transcriptomic and whole genome sequencing analysis indicated that QnrB upregulates gene expression and increases the number of gene copies near the origin of replication in both E. coli and K. pneumoniae. [32] The authors also reported that Bacterial two-hybrid and in vitro pull-down assays indicated that QnrB interacts with the DNA replication initiator DnaA. [32]

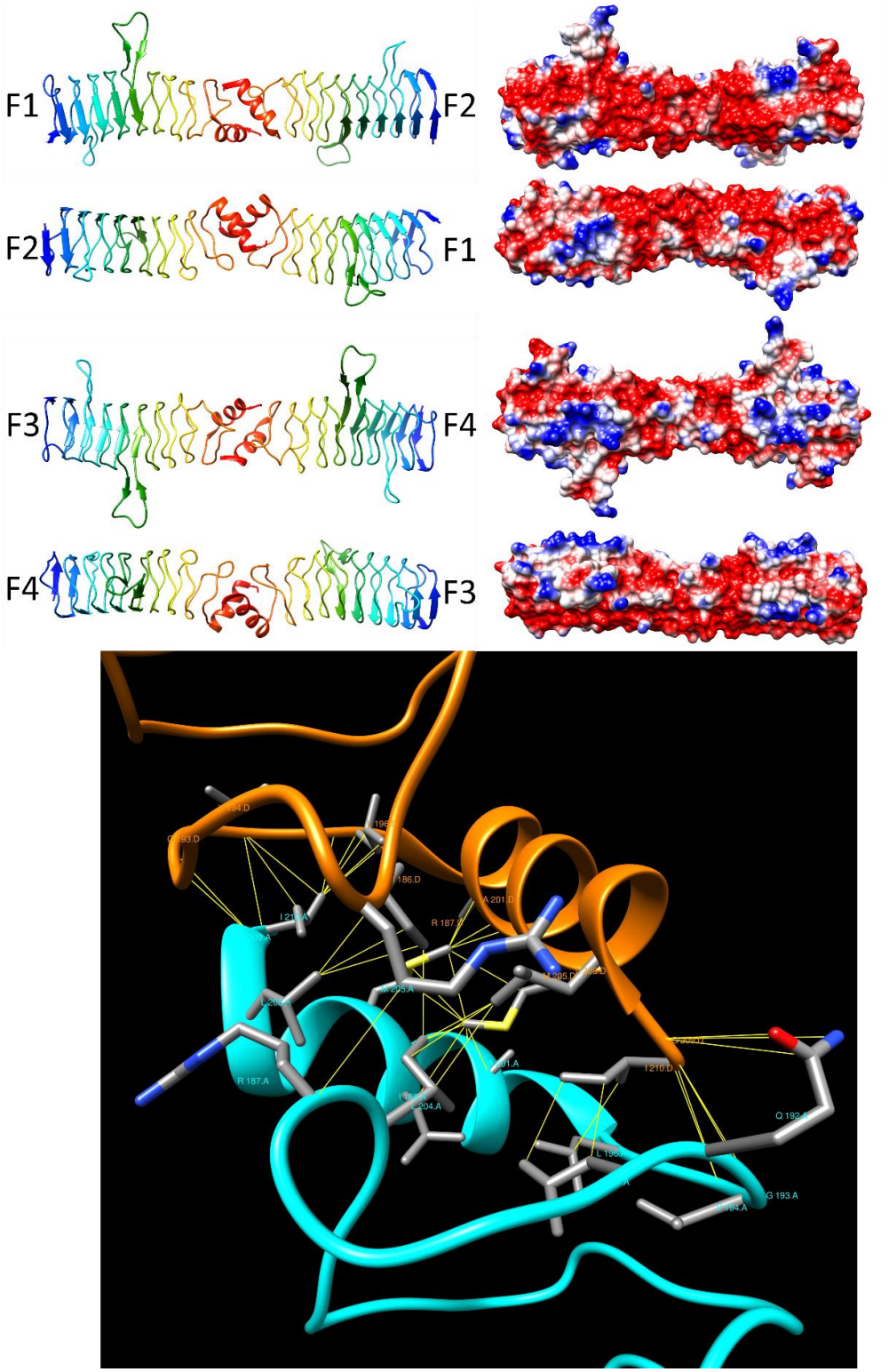



**Figure 2. 6 Structure and analysis of QnrB1.** The top panel shows the ribbon diagrams of the structure for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel is the interface model calculated by Intersurf and depicted using Chimera.

#### 2.3.3.2.3 AhQnr (3PSS[19], 3PSZ[19])

Xiong et al. reported the structure of AhQnr, a Qnr-like protein from A. hydrophila in, 2011.[19] AhQnr adopted a Rfr fold with nine complete coils capped by a 10-residue α-helix (D101-I210) and two conserved loops (Loop A (F47-C56) and Loop B (F103-C114) (**Figure 2.7, top panel**). The overall structure of AhQnr was very similar to that of QnrB1, consistent with the 38.8% sequence identity plus 56% similarity. The dimer interaction between two AhQnr chains involved a common motif with a short parallel β-sheet formed by residues Q197-N199 of chain A with I213-F215 of chain B, sidechain interactions between the two orthogonally-oriented α-helixes from each molecule, and a final short parallel β-sheet formed by L212-F215 of chain A with Q197-N199 of chain B with the interaction mediated by 96 contacts involving 20 residues on chain A and 20 residues on chain B. Interestingly, the dimer interaction was established by an equal mixture of polar residues and eight hydrophobic residues, with Q23 and F215 making the largest numbers of contacts at 11 and 12 interactions, respectively (**Figure 2.7 bottom panel**). The number of residues establishing the dimer interface in AhQnr (40) was greater than in MfpA (35), EfsQnr (31) and QnrB1 (23) and, correspondingly, the interface area of AhQNr (2039 $Å^2$) was significantly larger than that of MfpA (1700 $Å^2$), EfsQnr (1568 $Å^2$) and QnrB1 (1131-1432 $Å^2$). Interestingly, despite having the largest interface area, the energy of the AhQNr dimer interaction was the weakest of those discussed so far at -9.9 kcal/mole, compared to MfpA (-15.8 kcal/mole), EfsQnr (-14.9 kcal/mole) and QnrB1 (-10.6 kcal/mole). It is possible that more interactions were required to compensate for a less stable hydrophobic core, with the aromatic sidechain involved in the largest number of contacts, F215, residing on the surface of the protein and not participating in formation of a hydrophobic core.

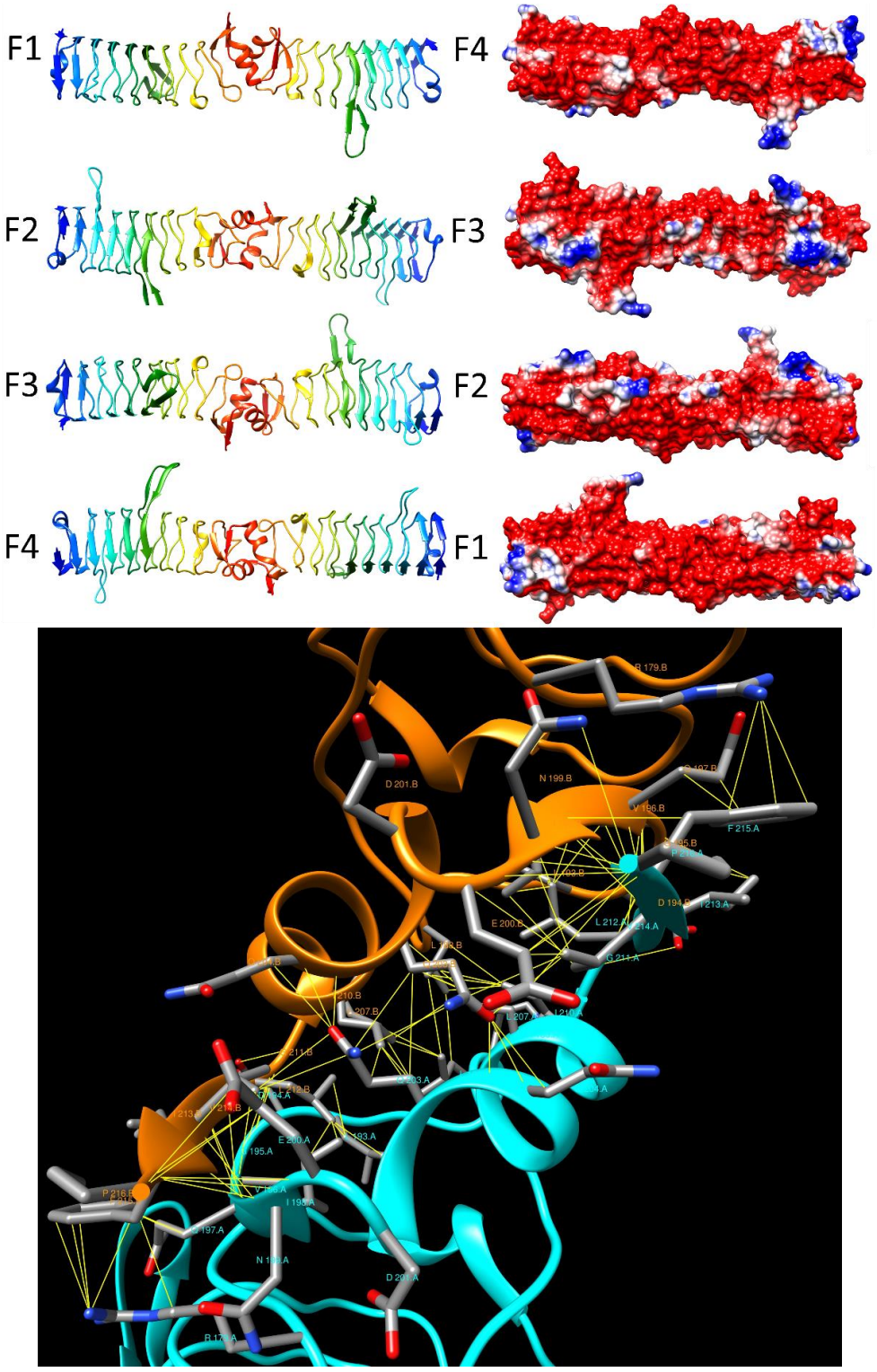



**Figure 2. 7 Structure of AhQnr.** The top panel shows the ribbon diagrams of the structure for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel is the interface model calculated by Intersurf and depicted using Chimera.

#### 2.3.3.2.4 AlbG (2XT2[33], 2XT4[33])

AlbG is a self-resistance factor from X. albilineans against albicidin, a nonribosomally-encoded hybrid polyketide-peptide with antibiotic and phytotoxic properties produced by the pathogenic bacterium X. albilineans.[8-29] As a self-resistance factor, AlbG protects X. albilineans from the antibiotic and cytotoxic effects of albicidin produced by X. albilineans itself. Albicidin shares a common mode of action with fluoroquinolines, also stabilizing the DNA gyrase-cleaved DNA complex and leading to single and double-strand breaks and eventual cell death.[33-34] X. albilineans uses multiple self-protection mechanisms, including the DNA mimicking activity of the AlbG PRP. AlbG increases the resistance of E. coli gyrase to albicidin both in vivo and in vitro, and at higher concentrations it inhibits supercoiling by the E. coli gyrase even in the absence of albicidin.[34]

The three-dimensional structure of AlbG was solved by Vetting et al. in 2011.[31] The structure is similar to other Qnr proteins, having eight complete coils with one half coil as the $0^{th}$ coil and a quarter coil as $9^{th}$ coil, capped at the N-terminus of the Rfr helix by a small N-terminal extension (residues 1-8) and capped at the C terminal end of the Rfr helix by an 11-residue α helix that is involved in the formation of the head to head functional dimer (**Figure 2.8, top panel**). A 13-residue loop insertion (T87-A99) disrupted the β helix structure of AlbG, with the loop and β-helical kink stabilized by several noncanonical PRP residues.[31] The dimer interface established by the head-to-head interaction of the C-terminal helix of two AlbG molecules involves 36 amino acids (chain A: 18 residues and chain B: 18 residues) that involved the formation of a hydrophobic core involving eight aromatic and aliphatic sidechains from each chain and interactions between five charged sidechains from each chain (**Figure 2.8, bottom panel**). The surface area for the dimer interface was 911 $Å^2$ and the binding energy for dimer formation was -13.5 kcal/mol.

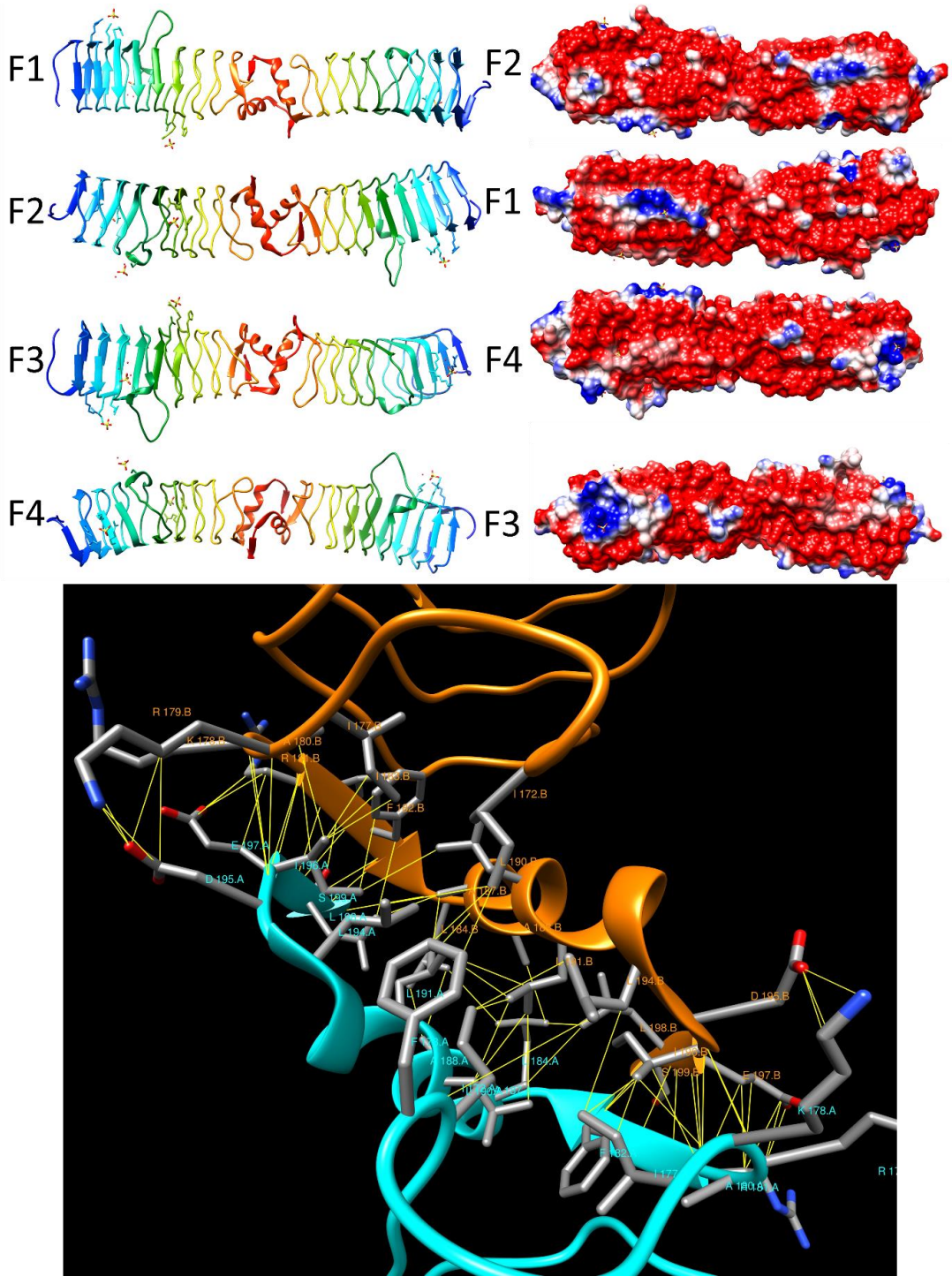



**Figure 2. 8 Structure and analysis of AlbG.** The top panel shows the ribbon diagrams of the structure for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel is the interface model calculated by Intersurf and depicted using Chimera.

**2.3.3.2.5 PENT (6FLS[35])**

PENT, a PRP from the human pathogen Clostridium botulinum, is a homolog of the fluoroquinoline resistance protein, MfpA from M. Typhimurium[35] (20.2% sequence identity and 39.3% sequence similarity over 183 residues). PENT is made up of 217 residues, crystallizes as a dimer and adopts a right-handed quadrilateral β helix with eight coils, with a 12-residue α helix with its axis perpendicular to the β helix axis capping both the N- and C-termini of the β helix (**Figure 2.9**). The head-to-head dimer interface is mediated by the C-terminal α helix cap of each of two PENT molecules (**Figure 2.9, top panel**). The head-to-head dimer interface interaction was observed between chains A and B and between chains C and D in the crystallographic asymmetric unit. The architecture of the dimer interface involved the formation of a short parallel β involving A197-I199 on chain A and I213-V215 on chain B, orthogonally oriented α helices (S200-G212 on each chain), and another short parallel β sheet formed by residues I213-V215 on chain A and residues A197-I199 on chain B (**Figure 2.9, bottom**). The dimer interaction involved 37 residues (chain A: 18, chain B: 19) and 63 non-bonded contacts mediated by a network of aromatic and aliphatic sidechains that formed a hydrophobic core with I199, M201, I213 and I214 all making six or more contacts in chain A and with S189, I199, W211, I213 and I214 making at least six contacts in chain B (**Figure 2.9, bottom panel**). The interface area of those two interactions was the same at ~1035 $Å^2$ and the energy of the interaction was -20 kcal/mole. Although the PENT PRP forms a head-to-head dimer with an overall structure similar to that of known PRP gyrase inhibitors, i.e. the MfpA and EfsQnr, it lacked the conserved extra-helix loops observed in other PRP gyrase inhibitors, such as QnrB1, AhQnr and AlbG. At this time, the function of PENT has not been confirmed experimentally.

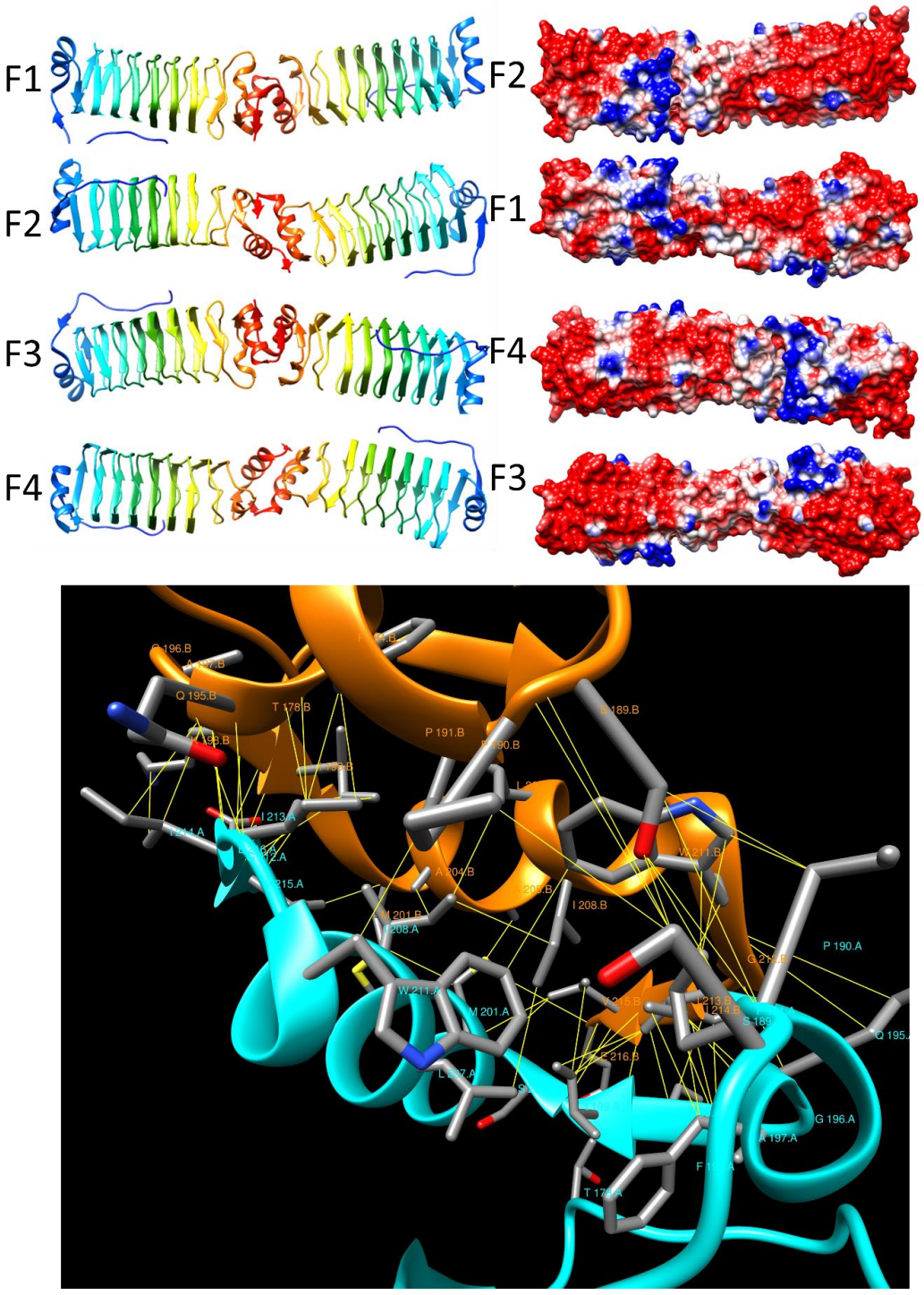



**Figure 2. 9 Structure and analysis of PENT.** The top panel shows the ribbon diagrams of the structure for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel is the interface model calculated by Intersurf and depicted using Chimera.

### 2.3.4 Ubiquitin E3 ligases

#### 2.3.4.1 SopA (2QYU[36], 2QZA[36], 3SY2[37], 5JW7[38])

Salmonella enterica serovar Typhimurium, a rod-headed, flagellate, facultative anaerobic, Gram-negative pathogenic bacterium, stimulates inflammatory responses in the intestinal tract that are required in order for it to replicate in the intestinal tract.[39] Bacterial pathogens can infect host cells and stimulate the inflammatory response by delivering effector proteins by either a type III or type IV secretion system.[38] One of the effector proteins is SopA, which is a Homologous to E6AP C-terminus (HECT)-type E3 ligase[40] that is required for efficient stimulation of inflammation in S. Typhimurium infections.[38] The process of attaching ubiquitin to a targeted proteins requires an enzyme cascade including ubiquitin activation enzyme (E1), conjugating enzymes (E2) and ligase (E3). HECT E3s is one of two types of E3 which has the function of forming a thioester intermediate with ubiquitin to transfer ubiquitin to substrate while the other is called RING E3s. Bacterial infection is normally sensed by host pattern recognition receptor-mediated detection of pathogen-associated molecular patterns (PAMPS), which induces a pro-inflammatory response to fight the infection.[38] The tripartite motif-containing (TRIM) TRIM56 and TRIM65 host really interesting new gene (RING)[41] E3 ubiquitin ligases are normally involved recognizing foreign proteins and stimulating release of interferons to communicate to nearby cells to launch an immune response to combat infection.[38] Kamanova et al. showed that SopA inhibits the host immune responses by targeting the tripartite motif-containing (TRIM) TRIM56 and TRIM65 host really interesting new gene[41] E3 ubiquitin ligases.[39]

In 2008, Diao *et al.* solved the structure of $SopA_{163-782}$, which was a fragment of the full-length SopA that was stable to proteolysis.[36] The structure was described as being organized into a 147-residue N-terminal β-helix domain (residues 163-370), a central domain (residues 371-590), a helical linker (residues 591-611) and a C-terminal domain (residues 612-782). SopA was the first PRP-containing multi-domain protein to have its structure determined. The SopA structure contains an N terminal PRP domain (around 200 amino acids) and a catalytic domain containing N- and C-lobes (**Figure 2.10, top panel**). In 2017, Fiskin *et al.*, solved the structure of SopA bound to the RING domain of TRIM56[38] which revealed the structural basis for selectivity of SopA for TRIM56 and TRIM6.The molecular basis of interaction showed that the TRIM56 domain interacts with the interface of the β-helix and the N-lobe domains of SopA through packing with the first $Zn^{2+}$-binding loop in a cleft of SopA. With a combination mutation experiments, it was shown that this interaction relies on three key residues including T338 of SopA, and L25 and E26 of TRIM56. In TRIM56, E25 interact with R296, H297 and K298 by polar contacts while L25 is inserted between the hydrophobic pocket constructed by P334 and F345. In SopA, T338 has a close hydrophobic contact with central α-helix of TRIM56. From the mutation experiments, T338 also shown to support the interaction between SopA and TRIM65 (**Figure 2.10, bottom panel**).[38]The authors performed structure-based biochemical analyses to show that SopA inhibited the TRIM56 E3 ligase activity by occluding the E2-interacting surface of TRIM56, and further showed that SopA ubiquinates TRIM56 and TRIM65 resulting in their proteasomal degradation during infection, thus disrupting the host immune response to *S. typhimurium* infection.

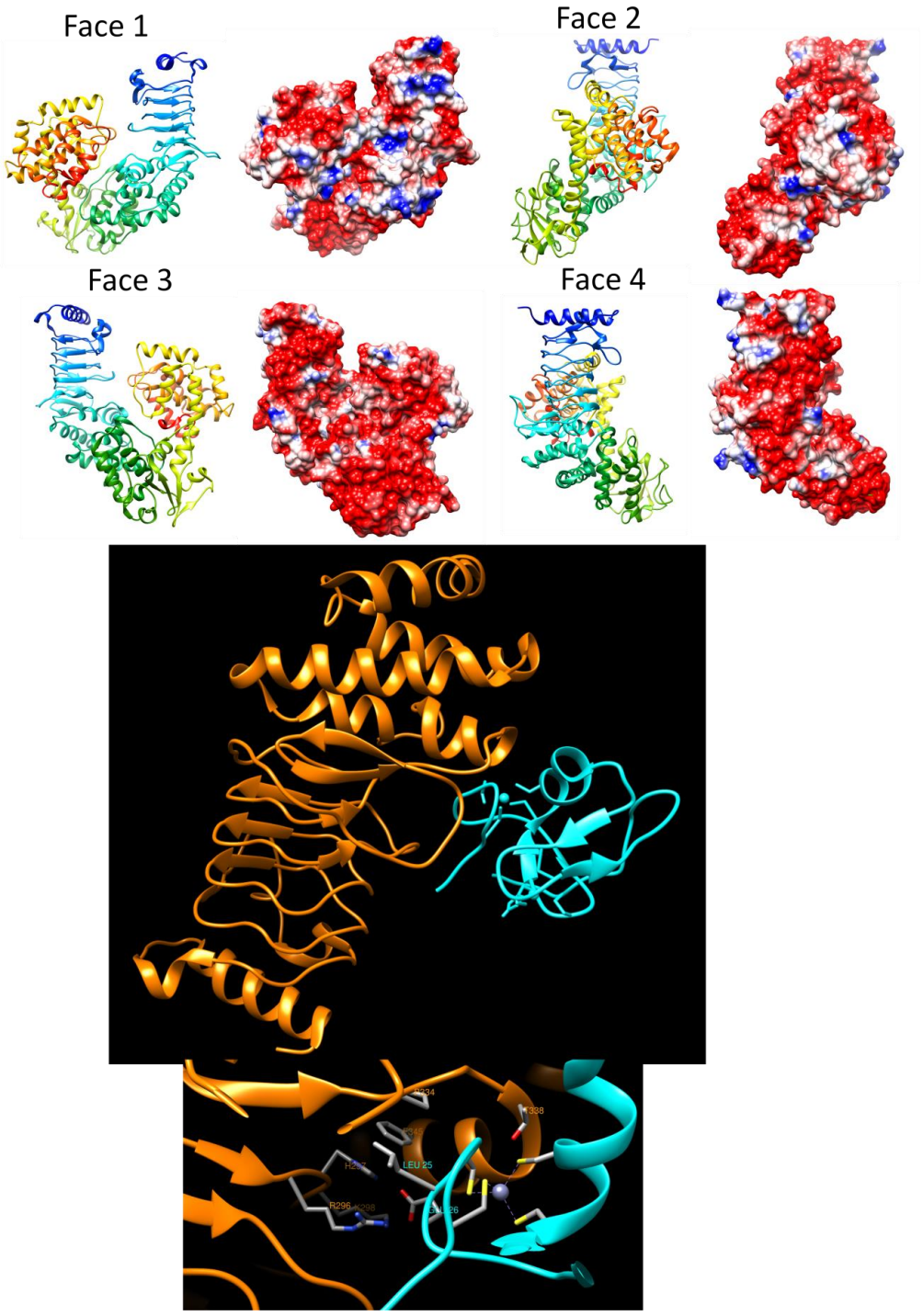



**Figure 2. 10 Structure and analysis of SopA.** The top panel shows ribbon diagrams of the structure (PDB ID: 2QZA) for four different orientations (at left) and the corresponding electronic potential surface plots (at right). The bottom panel (upper figure) shows the overall interaction between SopA and Trim56(5JW7) and the details of the interaction are shown in the lower figure. SopA is colored orange and Trim56 is colored cyan.

### 2.3.4.2 NleL (3NB2[42], 3NAW[42], 3SQV[37])

The non-Lee-encoded effector ligase (NleL) from enterohemorrhagic Escherichia coli (EHEC) 0157:H7 is a homolog of SopA from S. Typhimurium. Lin et al.[42] solved the crystal structure of NleL and showed that NleL functionally and structural mimics eukaryotic E3 ligases and catalyzes formation of unanchored polyubiquitin chains using linkages with residues Lys6 and Lys 48 and with the catalytic cysteine residue forming a thioester intermediate with ubiquitin.[42] In recent studies, it has been shown that NleL uses JNK proteins (stress-activated protein kinase) as the first substrate to promote EHEC-induced attaching and effacing (A/E) lesions[42] and interacts with TRAF2, TRAF5, TRAF6, IKKα and IKKβ to disrupt the host NF-κB pathway.[42]

The structure of NleL shares a common N-terminal PRP domain with SopA, which contains five and a half Rfr coils, one α helix at the N terminus and three α helices at the C terminus flanking the β helix domain. Superposition of the SopA and NleL crystal structures indicated that the PR and N-lobe domains maintain a fixed relative orientation, the C-lobes can adopt different orientations relative to the PR and N-lobe, perhaps but due to their different functional requirements. Based on the structures of complex between NleL and the UbcH7 domain (3SQV), UbcH7 contacts the N-lobes by both hydrogen bond and van der Waals interactions. In SopA, the PRP domain is N-terminal to the catalytic domain and the carbohydrate-modified substrate proteins may be recognized by that. However, the actual functions and substrates still remain unknown. Although the function of pentapeptide repeat domains is still unknown, a cleft at the interface of pentapeptide repeat domain and N-lobe domain provide some clues to identify the potential functions. Two of tripartite-motif-containing (TRIM) E3 ligases, belonging to the family of RING-type E3 ligases, are involved in the regulation of SopA.[39] Complex structure involving SopA and TRIM 56 unmasked that the first $Zn^{2+}$-binding loop of TRIM 56 contacts with the cleft of SopA. In TRIM 56, Leu25 and Glu 26 are the two key amino acids contacting SopA even though they adopt different strategies. Leu25 interacts with Phe345 and Pro334 of SopA by inserting into a hydrophobic pocket while Glu26 contacts Arg296, His297 and Lys298 by their polar groups. (**Figure 2.11**). There are two complexes in the crystallographic asymmetric unit. In the interaction between chain A (NleL) and chain C (E2 UbcH7), one salt bridge, two hydrogen bond and 47 non-bonded contacts form the contact interface involving 15 residues from chain A and 13 residues from chain C. Asn578 from NleL and Phe63 from E2 UbcH7 are responsible for most of the contacts for this interaction. The interface area between two chains was 1640 $Å^2$ with the energy of interaction at -11.5 kcal/mol. Different from interaction between chain A and C, in the interaction between chain B and chain D, a disulfide bond between Cys753 (chain B) and Cys86 (chain D) stabilized the interaction. 20 residues from chain B and 15 residues from chain D formed the interaction and the interface area was 1922 $Å^2$ with an energy of formation of -15.3 kcal/mol. Phe63 from chain D was major contributor to support this interaction.

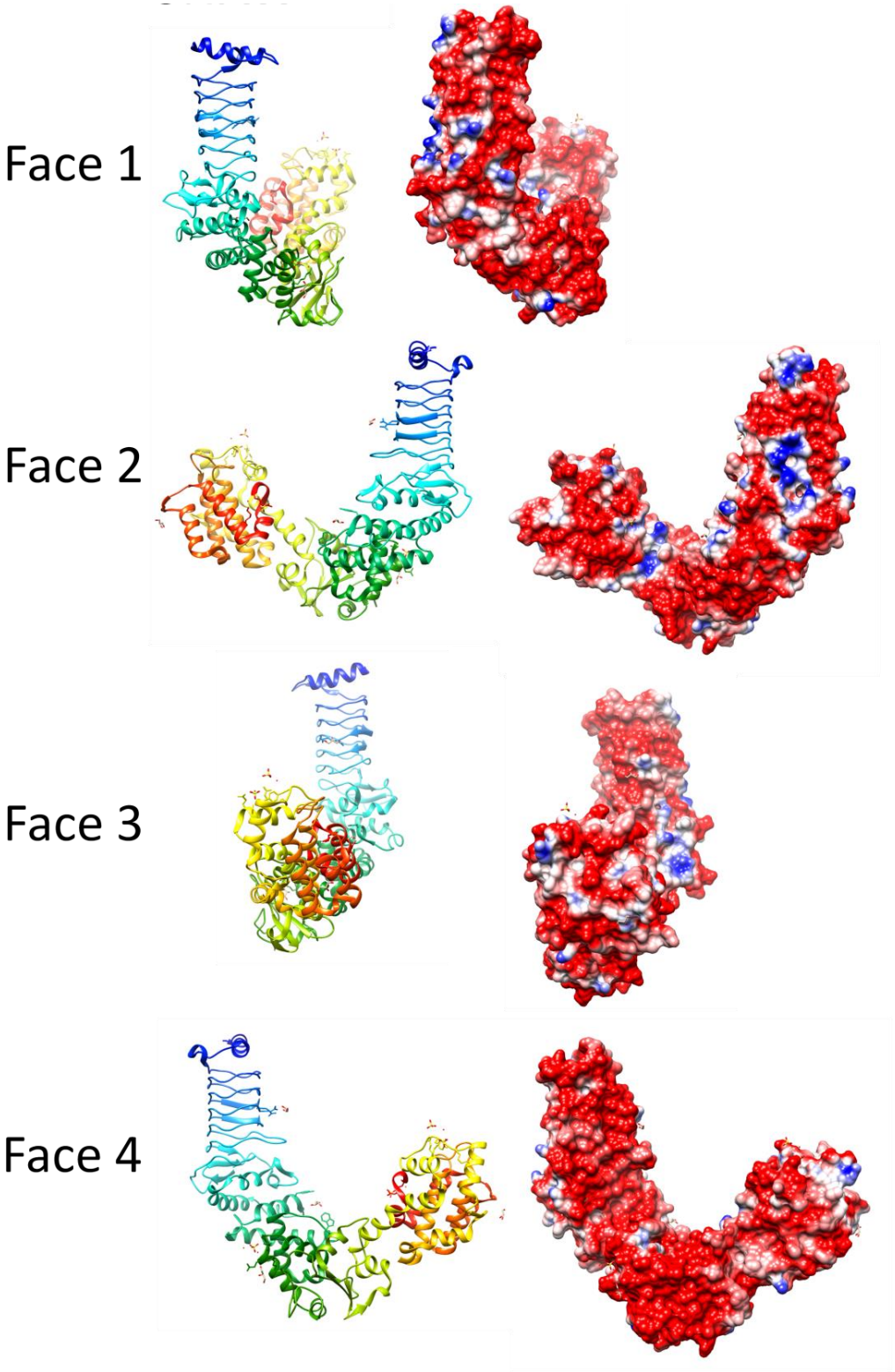



**Figure 2. 11 Structure and analysis of NleL.** Ribbon diagrams of the structure (PDB ID: 3NAW) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.3.5 Synaptic Vesicle Glycoprotein 2 Receptors

#### 2.3.5.1 SV2C-LD (4JRA[43], 5JMC[44], 5JLV[44], 5MOY[45], 6ES1[46])

Synaptic vesicles, also referred to as neurotransmitter vesicles, store various neurotransmitters that are released at the synapse, i.e. the junction, between nerve cells. The synaptic vesicles are essential for propagation of nerve impulses and constantly regenerated in nerve cells. The synaptic lumen refers to the volume contained inside the synaptic vesicles. Synaptic vesicle glycoprotein 2 (SV2) receptors represent a protein family with two essential complementary major isoforms, SV2A and SV2B, and one minor isoform, SV2C, that are putative transport proteins.[47] All three isoforms are composed of a 12-transmembrane domain and a luminal domain (SVC-LD), i.e. the domain sticks into the vesicle lumen, composed of a four and half-coil PRP β helix that acts as a receptor for binding to Botulinum neurotoxin A (BoNT/A) from the bacterium C. botulinum and related bacteria.[46, 48] Botulinum neurotoxins (BoNTs) are the most toxic class of bioweapons and also have a popular and widely used cosmetic application as an anti-wrinkle agent, e.g. Botox. BoNTs exist as seven main serotypes from BoNT/A to BoNT/G.[33] In 2006, Dong et al. showed that the luminal domain of SV2 (SVC-LD) acts as a receptor for BoNT.[49] The SV2C-LD is necessary in the process of translocation of BoNT/A and glycosylation of SV2C-LD and SV2 glycan are also crucial for BoNT/A binding to neurons.[50]

From 2014 to 2018, five structures were reported for complexes between BoNTs and the SV2C-LD (also referred to as SV2C-L4).[43-46] The LD of SVC2 is a five-coil PRP domain (**Figure 2.12**). All five structures were similar with slight variations in the relative orientations between the SV2C-LD and the different subtypes of BoNT/A. The interaction between the BoNT/A and the SV2C-LD receptor occurs at the exposed β-strand of the 5th Rfr coil at C-terminal un-capped edge of the PRP domain by forming an interchain β-sheet mediated by backbone to backbone hydrogen bonds between the open β strand of SV2C-LD and the β-strand edge of the BoNT/A. Even though the structural features of the multiple SV2C-LD/BoNT/A complexes are similar, the orientation of the BoNT/A relative to the SV2C-LD β-helix receptor varied slightly among the structures due to differences in amino acids sidechains mediating the binding interaction and due to flexibility of the interactions. For example, in two structures reported for the same complex SV2C-LD/BoNT/A but crystallized in different space groups and with different resolution (PDB-IDs 5MOY and 6ES1), the orientation of the SV2C-LD β-helix was rotated by 15° relative to BoNT/A to maintain the tight interaction between β-hairpin of BoNT/A and the continuing β-sheet. This rotation caused shifts in both components. The residues in the β-hairpin (T1146 and N1147) moved by 1.8 Å and the residues from C-terminus of SV2C-LD (N480 to Y497, D546 to K566) moved between 0.4–8.1 Å.[46]

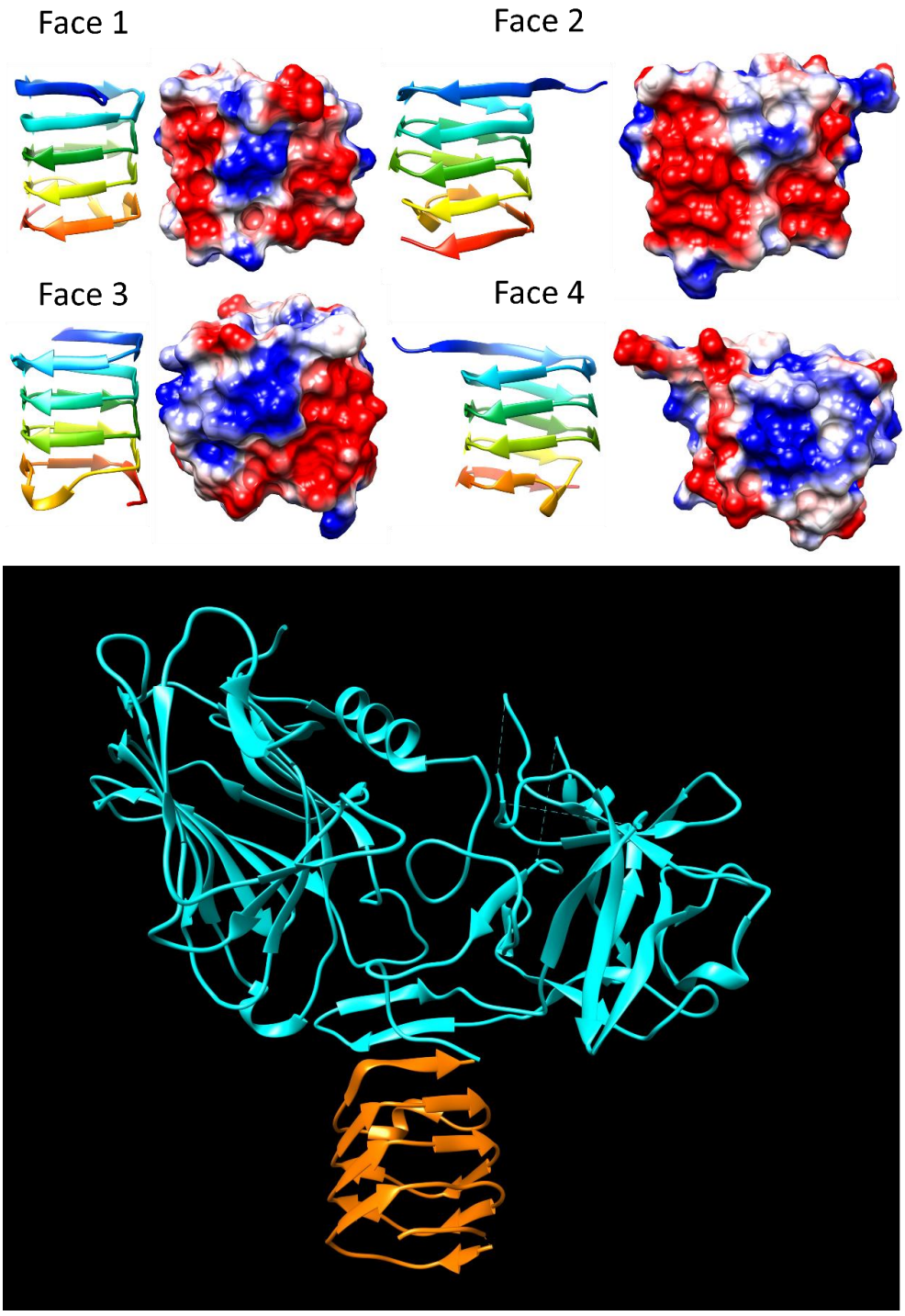



**Figure 2. 12 Structure and analysis of SV2C.** The top panel is the ribbon diagram of the structure and electronic potential surface analysis. The bottom panel is the interface model calculated by Chimera.

## 2.4 PRPs with Three-Dimensional Structures but Unknown Function.

### 2.4.1 HetL (3DU1[6])

HetL (gene all3740) is one of more than 30 PRPs from Nostoc sp. PCC 7120.[6] In 2002, it was shown that HetL overexpression using a heterologous promoter in wild-type Nostoc PCC 7120 induced multiple-contiguous heterocysts in nitrate-containing medium.[51] Addition of a synthetic peptide containing the last five residues of PatS, which was known to suppress heterocyst differentiation in wild type Nostoc PCC 7120, did not suppress heterocyst differentiation in the hetL overexpression strain, indicating that HetL acts downstream of PatS production. Interestingly, a hetL null-mutant showed normal heterocyst development and diazotrophic growth, i.e. the ability to fix atmospheric nitrogen into more usable forms such as ammonia, leading Liu and Golden to conclude that HetL may not normally involved in regulating heterocyst development, many only play a non-essential accessory role, or that its function may be compensated for by cross talk or redundancy with other PRPs.[51] Liu and Golden observed that the predicted HetL protein was composed almost entirely of PRs.

In 2009, the three-dimensional structure of HetL was determined, the first PRP structure from Nostoc sp. PCC 7120 to have its structure determined.[6] The structure revealed that HetL adopted the standard right-handed quadrilateral β helical structure composed of ten complete coils, a ten-residue α helix that caps its N terminus, a two-stranded anti-parallel β sheet that sits on the C-terminus of Face 1 of the β helix, a six-residue loop insertion protruding from the corner adjoining Face 3 and Face 4 near the middle of the helix, and a nine-residue insertion loop protruding from the corner joining Face 3 and Face 4 in the C-terminal half of the helix (**Figure 2.13**). The PRP β helix in HetL is entirely composed of type II β turns. The electrostatic surface potential of HetL contains patches of negative charge but is otherwise unremarkable. Although HetL has been shown to play a role in heterocyst differentiation, its precise biochemical function remains unknown.

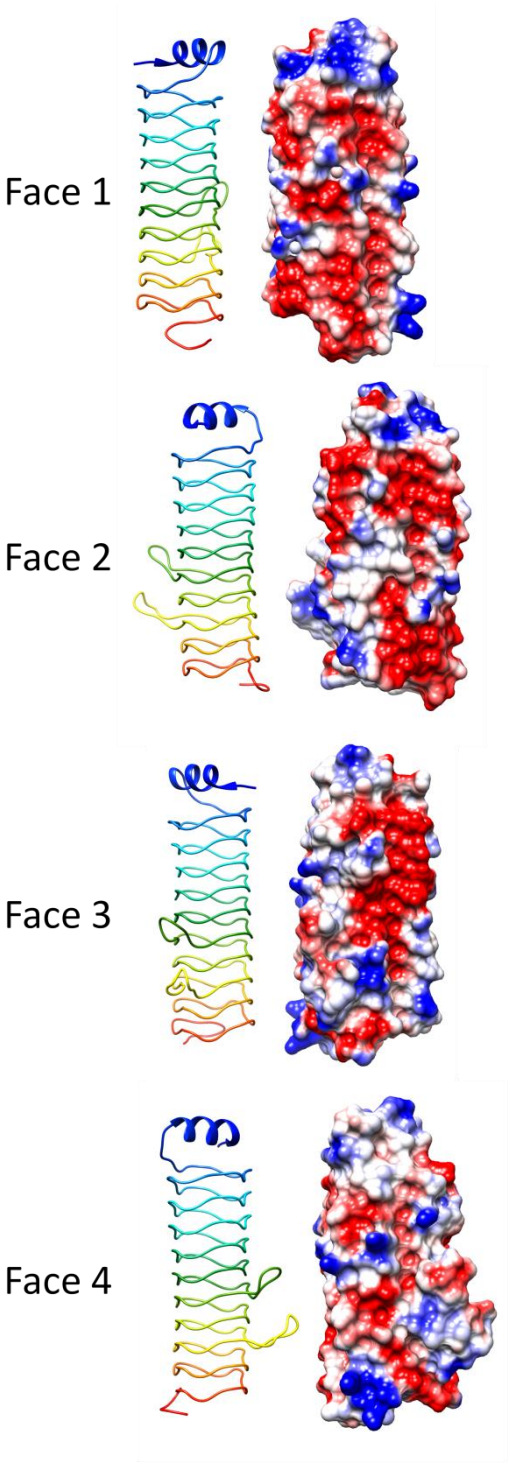



**Figure 2. 13 Structure and analysis of HetL.** Ribbon diagrams of the structure (PDB ID: 3DU1) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.2 Alr1298 (6UV7[52], 6UVI[52])

The full-length Alr1298 protein is predicted to contain 167 amino acids, which includes 15 PRs. The three-dimensional structure of Alr1298 was found to have three and three quarter Rfr coils with the incomplete coil occurring at the C-terminus of the β helix. (**Figure 2.14**). The β helix is capped at the C-terminus by a single α helix and a five α helix bundle at its N-terminus. The β helix was composed of a combination of type II and type IV β turns. The electrostatic surface potential contained large patches of clustered positive and negative charge that are poised to interact with charged binding partners. Potential clues regarding the function of Alr1298 were investigated by analyzing the gene cluster to which the alr1298 gene belonged. Given the gene in the cluster possibly belonged to a common operon that often share related functions,[53-55] the genes flanking alr1298 were examined. The gene cluster contains three genes preceding and three genes following alr1298. Alr1295 was found to be conserved in 14 of 15 aligned genomes and encoded a prohibitin, which generally act as inhibitors to cell proliferation. In cyanobacteria, prohibitins have been linked to thylakoid biogenesis and membrane synthesis. Alr1297 was annotated as an ABC transport system. Alr1299 was predicted to be involved in purine metabolism, metabolic pathways and biosynthesis of secondary metabolites. The other genes were unannotated. It is possible that Alr1298 plays a role in cell proliferation and thylakoid biogenesis. A genome-wide microarray analysis revealed that alr1298 was upregulated following nitrogen starvation,[2] peaking at a 4x increase at eight hours post nitrogen starvation. Since the primary response to nitrogen starvation is patterned differentiation of vegetative cells into heterocysts capable of fixing atmospheric nitrogen, the microarray result supports the observation that alr1298 may either be involved in the response to nitrogen starvation or play a role in heterocyst differentiation.[52]

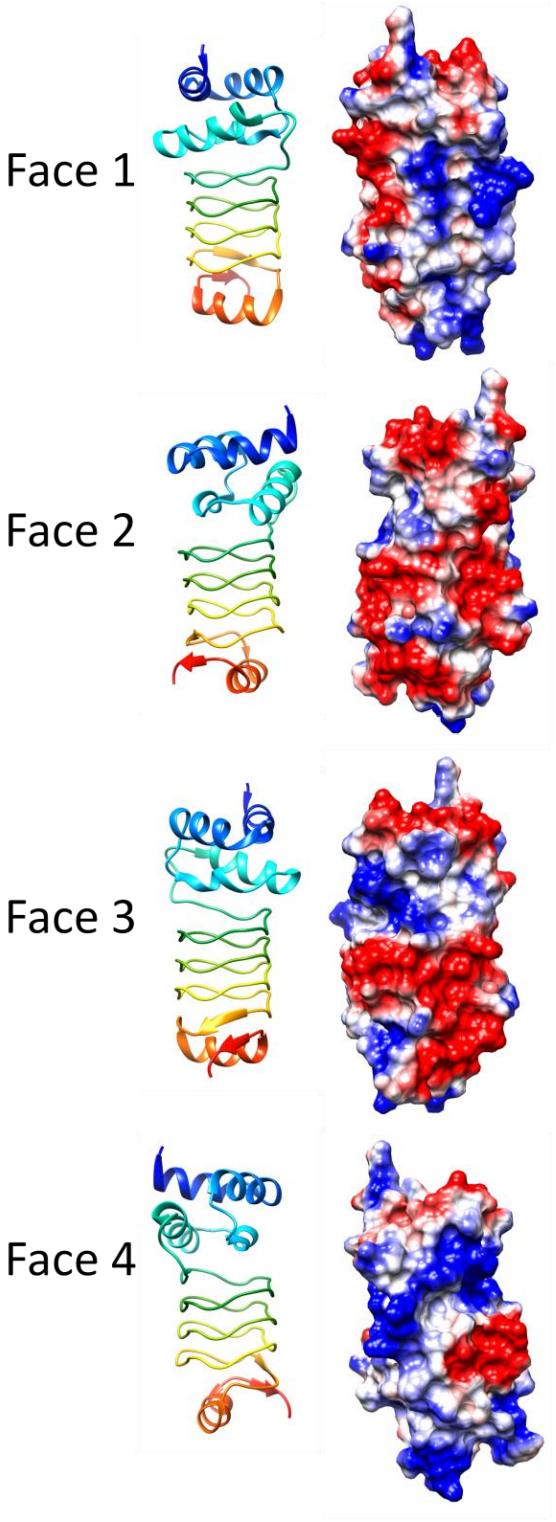



**Figure 2. 14 Structure and analysis of Alr1298.** Ribbon diagrams of the structure (PDB ID: 6UV7) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.3 Alr5209 (6OMX[20])

The alr5209 gene from Nostoc sp. st. PCC7120 encodes a 129 amino acid protein that contains 16 tandem PRs.[20] The three-dimensional structure of Alr5209 was determined in 2019[20] revealing that it was composed of four complete Rfr coils of a β helix that was capped by a nine-residue α helix at the N-terminus and by a four-reside α helix at the C-terminus (**Figure 2.15**). Alr5209 was the first PRP identified to contain type I β turns in its β helix structure, with all four Rfr coils joining Face 2 to Face 3 being type I β turns and all the remaining 12 turns being type II β turns. In comparison with other PRPs, Alr5209 had a more compact structure due the effect on the structure of the combination of type I and type II β turns used to form β coil stack. Analysis of the electrostatic surface potential revealed that two faces of the β helix were predominantly negatively charged with the other two faces being of mixed charge and generally neutral overall. Analysis of the gene cluster that alr5209 belonged to indicated that it may play a role in oxidative phosphorylation.[20]

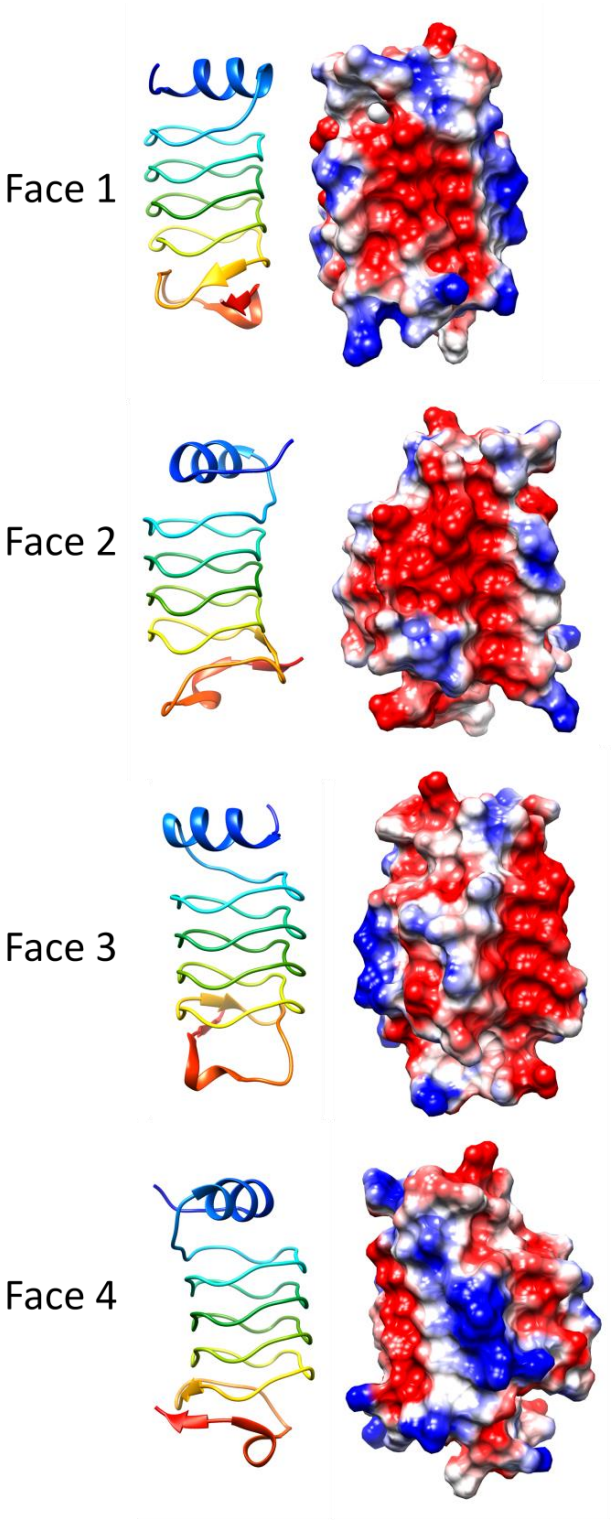



**Figure 2. 15 Structure and analysis of Alr5209.** Ribbon diagrams of the structure (PDB ID: 6UV7) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.4 Np275/Np276 (2J8K[15], 2J8I[15])

The Np275 and Np276 genes in Nostoc punctiforme, which are adjacent to one another and encode proteins of 98 and 75 amino acids, respectively, have sequences that are composed of tandem PRs.[15] The structure of Np275 was solved at 2.1 Å resolution (2J8I). The majority of the Np275 structure adopts a Rfr fold composed of four complete coils will all type II β turns and with the N-terminal end being capped by an α helix and the C-terminal end of the coil being uncapped exposing the hydrophobic core and terminal β strands of the β helix. The intervening sequence between the stop codon of the Np275 gene and the start codon of Np276 gene also encoded an in-frame PR sequence suggesting that Np275 and Np276 previously existed as a single longer protein. This suggestion was supported by the fact that it was possible to solve the structure of a Np275-Np276 fusion protein composed of seven and three-quarters complete Rfr coils with the N-terminal Np275 portion having virtually the same structure as the Np275 monomer structure and the Np276 portion also being uncapped and with the entire Rfr coil composed of type II β turns (**Figure 2.16**). Interestingly, the authors noted that Np275/Np276 has an unoccupied internal molecular surface along the Rfr coil helical axis that is continuous with a volume of 281 $Å^3$, and pointed out that while the function of MfpA is related to glycolipid localization, the cavity in Np275/Np276 would not be large enough to accommodate the hydrophobic tail of a glycolipid without expansion.[15] A putative function of the tunnel in Np275/Np276 remains unknown.

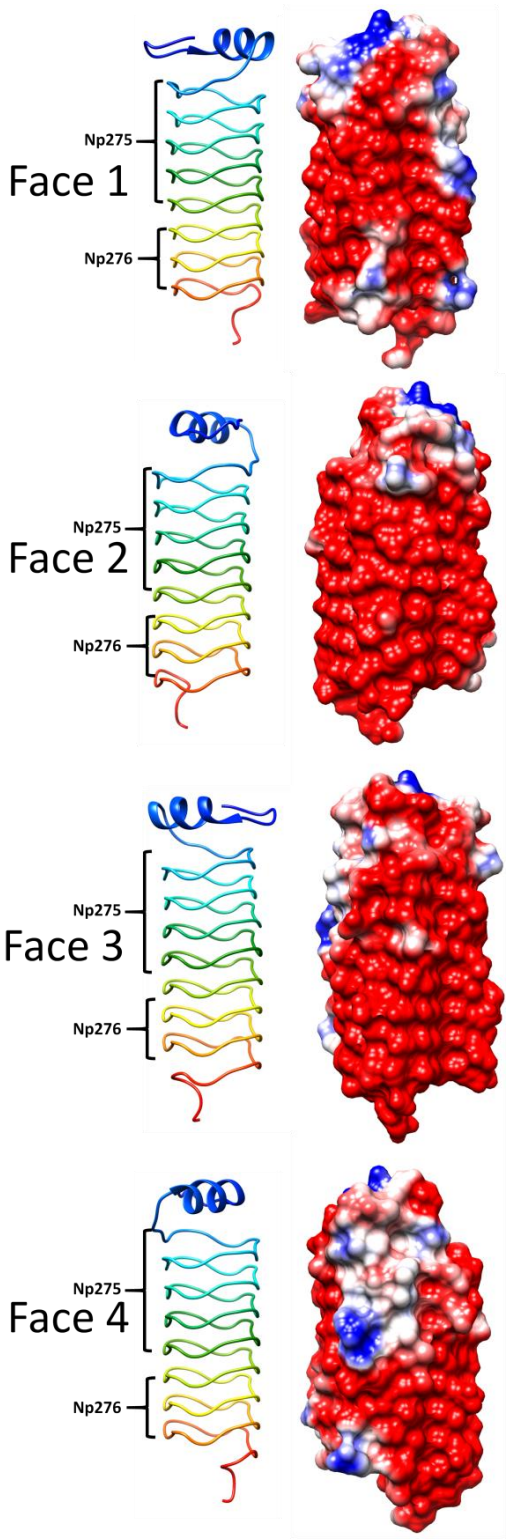



**Figure 2. 16 Structure and analysis of Np275/276.** Ribbon diagrams of the structure (PDB ID: 2J8K) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.5 Rfr32 (2F3L[13], 2G0Y[13])

The Rfr32 gene from Cyanothece sp. 51142 encodes a 167-residue that includes a 29-residue N-terminal signal peptide.[13] The three-dimensional structure of Rfr32 minus the N-terminal 29 residue signal peptide was determined at 2.1 Å revealing a structure dominated by five and one-quarter uninterrupted Rfr coils (**Figure 2.17**). The C-terminus of the Rfr coil is capped by a two-α helix bundle that is stabilized by an internal disulfide bond. The Rfr coil of Rfr32 contains mixture of type II and IV β turns. The electrostatic surface potential of Rfr32 contains contiguous patches of negative charge on Face 3 and in a deep crevasse present on Face 4. Ref32 is predicted to reside in the thylakoid lumen (https://www.uniprot.org/uniprot/B1WVN5#subcellular_location). The function of Rfr32 remains unknown and the UniProt database reports that existence of the protein is predicted based on homology (https://www.uniprot.org/uniprot/B1WVN5).

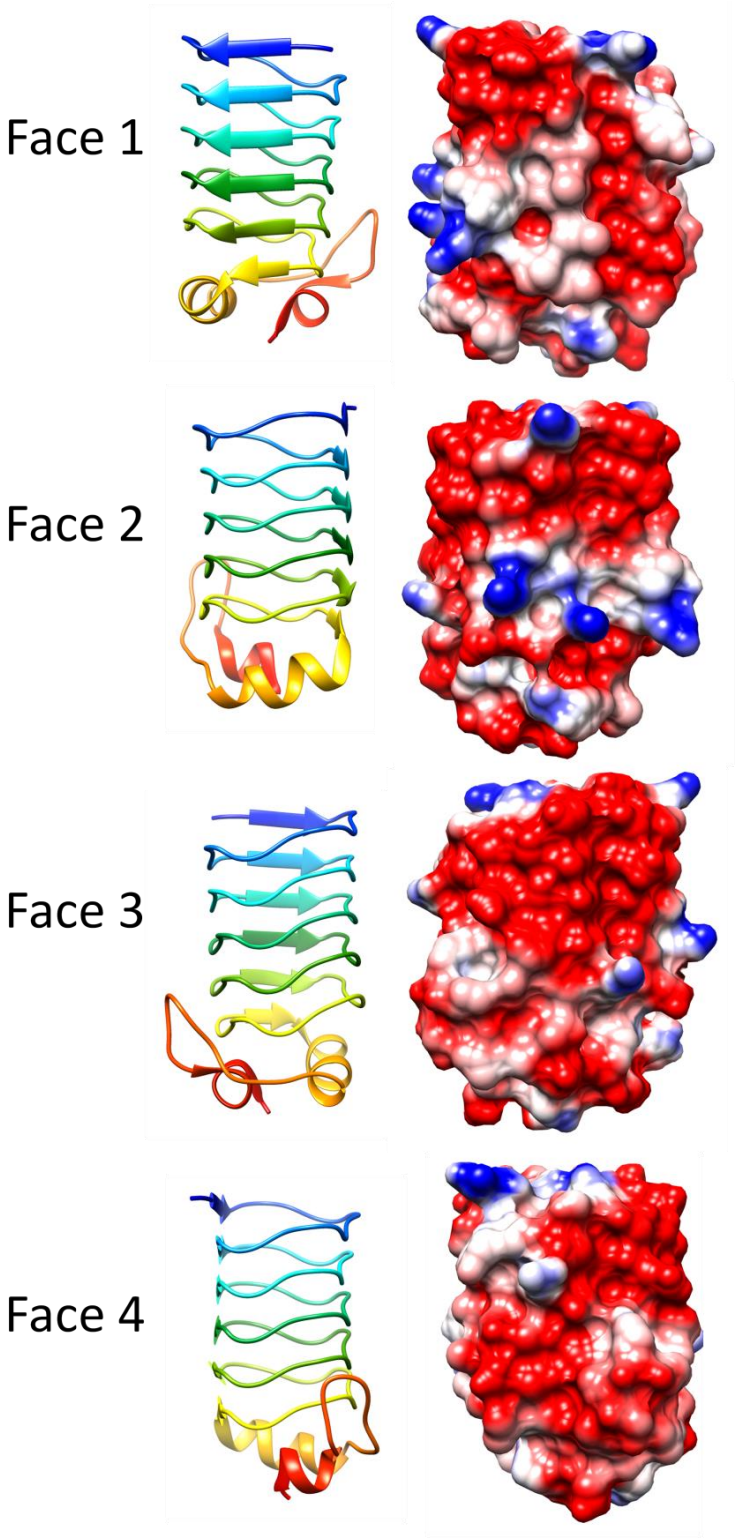



**Figure 2. 17 Structure and analysis of Rfr32.** Ribbon diagrams of the structure (PDB ID: 2G0Y) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.6 Rfr23 (2O6W[16])

Rfr23 is the second PRP with known structure from Cyanothece sp. 51142.[16] The β helix contains five complete coils with one helix in the N terminal (**Figure 2.18**). Different from Rfr32, there are two significant structural specifications in Rfr23, one is the 24-residue insertion, the other is disulfide bracket. The 24-residue insertion happens between the conjunction of the first and secondary coil which causes a break of consensus sequence. Due to the missing electron density, the structure of this insertion still remains unknown. However, according to the analysis of sequence, this insertion has a positive charge. Also, the formation of disulfide bracket between Cys39 and Cys42 make this structure more stable and it is possible to contribute the activity of Rfr23. As for the secondary structure, the composition elements of Rfr23 is simple, it only contains helix and β turn, and Rfr23 is proven as a PRPs with entire type II β turn. While the function of Rfr23 remains unknown, the UniProt database indicates that experimental evidence exists for expression of Rfr23 at the protein level (https://www.uniprot.org/uniprot/D0VWX3).

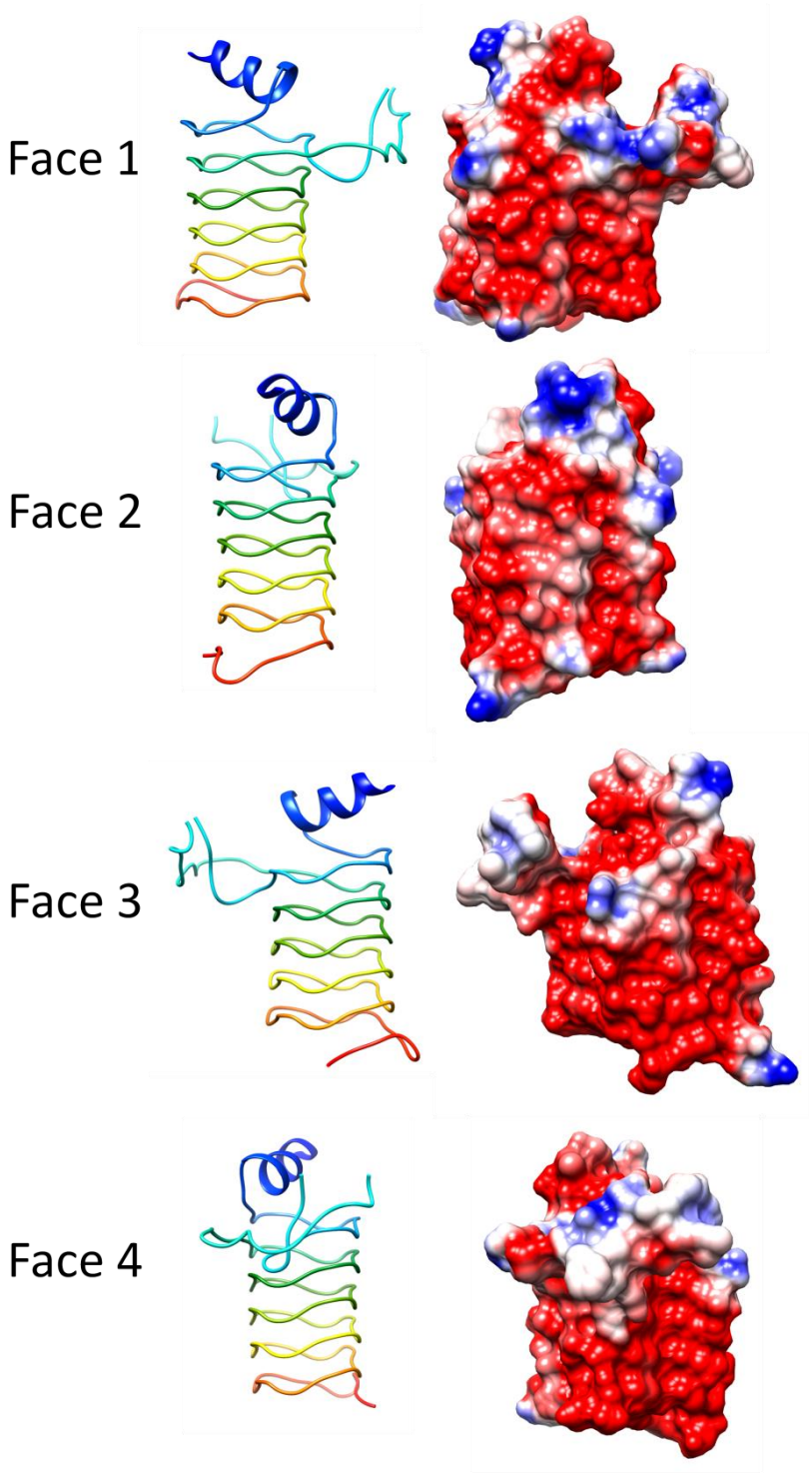



**Figure 2. 18 Structure and analysis of Rfr23.** Ribbon diagrams of the structure (PDB ID: 2O6W) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.7 At2g49920.2 (3N90[56])

The genome of A. thaliana is predicted to contain three genes encoding PRPs (At2g44920, At5g53490 and At1g12250) all of which are predicted to be located in the thylakoid lumen.[56] In 2011, Ni et al. reported the three-dimensional structure of At2g49920.2, one of two isoforms of At2g49920 identified in A. thaliana.[57] At2g49920.2 contained five complete Rfr coils made up of contained 25 uninterrupted PRs with a single-turn α helix capping the N-terminus and two α helices stabilized by a disulphide bond capping the C-terminus of the β helix. At2g49920.2 is made exclusively by type II β turns with one gamma turn. (**Figure 2.19**). Although the function of At2g49920.2 is still unknown, the chloroplast thylakoid lumen, in which At2g49920.2 is predicted to be located, is a compartment where the reactions of oxygenic photosynthesis take place. It has also been shown that At2g49920.2 is primarily expressed in the leaves of A. thaliana.[58]

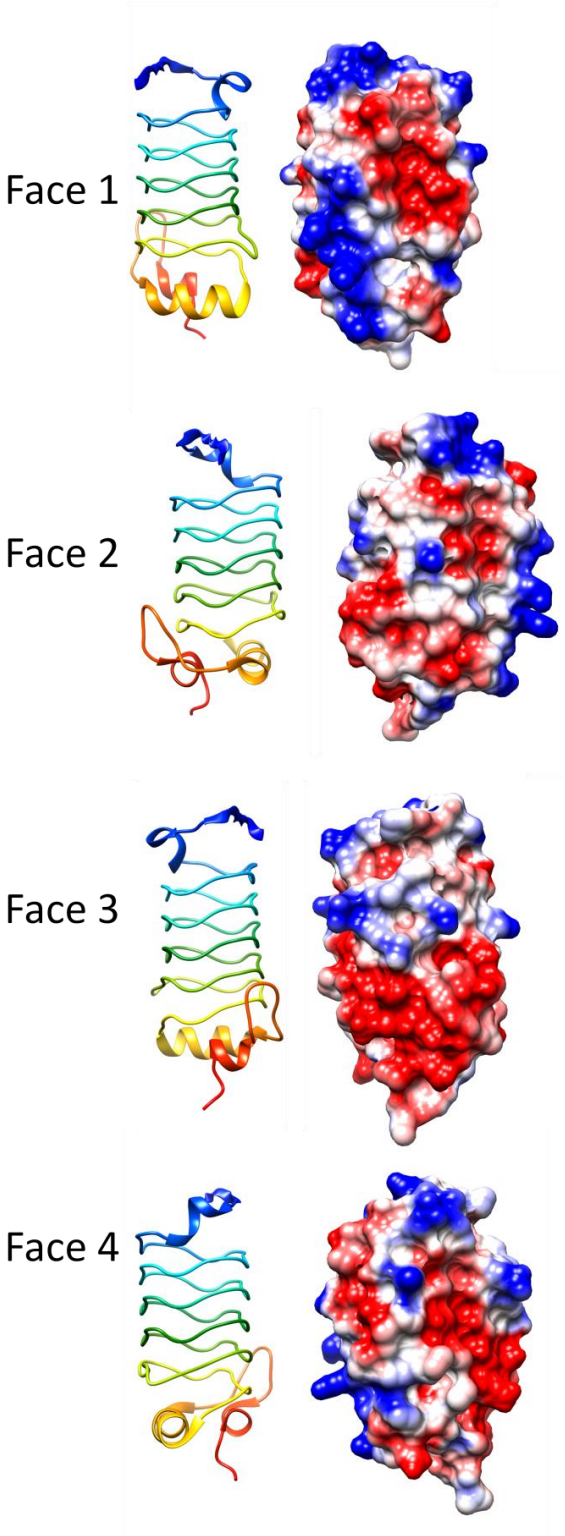



**Figure 2. 19 Structure and analysis of At2g49920.2.** Ribbon diagrams of the structure (PDB ID: 3N90) for four different orientations (at left) and the corresponding electronic potential surface plots (at right).

### 2.4.8 Changes in PRP gene expression levels Nostoc sp. st. PCC 7120 in response to nitrogen deprivation

In 2006, Ehira and Ohmori published a genome-wide gene expression experiment using a Nostoc (Anabaena) microarray containing 5336 probes specific for genes on the chromosome whose expression levels changed in response to nitrogen deprivation at 3, 8 and 24 hours post nitrogen deprivation compared to at the 0-hour timepoint.[2] Of the 26 chromosomally encoded PRP genes, expression of 25 of the genes was detected in the microarray analysis, 15 experienced at least a two-fold increase in expression following nitrogen deprivation for at least one of the time points (**Figure 2.20**). By comparison, expression of only two of the 25 genes decreased by at least 40% (**Figure 2.20**).

| | 3 hrs | 8 hrs | 24 hrs |
|---|---|---|---|
| All0813 | 1.2 | 1.75 | 0.05 |
| All0958 | 1.08 | 0.43 | 1.48 |
| All1812 | 0.46 | 2.3 | 1.4 |
| All2395 | 0.35 | 0.27 | -0.22 |
| All3048 | 0.94 | 1.67 | 0.98 |
| All3114 | 0.81 | 0.28 | 1 |
| All3256 | 1.02 | 0.68 | 0.26 |
| All3305 | 0.39 | -0.5 | -0.53 |
| All3306 | 0.57 | 0.34 | 0.3 |
| All3332 | 0.14 | -0.31 | -0.25 |
| All3740 | | | |
| All3869 | 0.33 | 0.59 | 1.1 |
| All4152 | 1.36 | 1.14 | 0.43 |
| All4220 | 0.81 | *2.09 | 1.12 |
| Alr0433 | | | 1.2 |
| Alr0704 | 1.26 | *1.7 | 1.37 |
| Alr1142 | 0.11 | 0.85 | 0.99 |
| Alr1298 | 0.88 | 2.03 | 1.19 |
| Alr1331 | 0.56 | -0.58 | |
| Alr1579 | 1.1 | *1.72 | 0.67 |
| Alr1746 | 0.32 | 0.1 | 0.05 |
| Alr2741 | 0.42 | 0.5 | 0.66 |
| Alr2768 | 1.19 | 1.45 | 0.4 |
| Alr3268 | 1.31 | 1.12 | 1.31 |
| Alr4610 | 1.13 | 1.05 | -0.01 |
| Alr5209 | | -0.16 | |

**Figure 2. 20 Summary of PRP gene expression in Nostoc sp st PCC 7120 following nitrogen deprivation.** The numbers indicate the log base 2 of the relative change following nitrogen deprivation relative to the 0-hr time point. Positive values indicate increased expression and negative values indicate decreased expression. Values shaded in blue indicate at least a two-fold increase in expression. Values marked by an asterisk indicate statistically significant changes. Values shaded in light pink indicate at least a 40% decrease in expression. Values were obtained from the supplementary tables reported by Ehira and Omori. [2]

## 2.5 Discussion

Despite their intriguing structure, and variations therein, the large size of their Pfam superfamily with nearly 39,000 members, and their relative abundance in one of the most ancient and important organisms on earth, i.e. oxygenic filamentous cyanobacteria, the biochemical function of PRPs remains remarkably elusive. To date, there are only three examples where the explicit biochemical function of a PRP is known, as discussed in this review. The first cellular function being that of conferring antibiotic resistance to fluoroqunionline antibiotics through the biochemical function of acting as DNA gyrase inhibitors, such as MfpA and the Qnr family of proteins, that exert their function by acting as a DNA mimic, binding to the complex of DNA gyrase and DNA, and blocking binding of fluoroquinoline to the DNA gyrase DNA complex, and therefore blocking its antibiotic activity. The second clear function is seen in SV2C that functions as a BoNT/A receptor in the synaptic vesicles of neurons. In this activity, the BoNT/A toxin binds to the SV2C PRP luminal domain of a synaptic vesicle neurotransmitter membrane protein, with the binding interaction mediated by a dovetailing of the β-strands of the BoNT/A neurotoxin with an exposed β-strand edge of the PRP luminal domain, resulting in the formation of an extended β-sheet that spans and crosses over the PRP luminal domain and the BoNT/A neurotoxin molecule. This completes the list of examples for which a PRP or a PRP domain of a larger protein is known to carry out a specific biochemical function that directly involves the PRP structure itself. From there, our understanding of PRP function becomes less clear. We have seen that PRP domains are present in the SopA ubiquitination inhibitor whose biochemical function is to bind to TRIM56 and TRIM65 and block the host immune response of interferon production that would normally stimulate proteasome targeting of the bacterial proteins as part of the host immune response to infection, however, the PRP domain of SopA does not directly interact with the targeted TRIM proteins leaving the precise function of the PRP domain of SopA in question. From here, a biochemical function has been associated with a few other PRPs, e.g. the prototypical HglK protein was associated with localization of glycolipids to the heterocyst outer layer, but the structure of HglK is unknown and the precise biochemical function of HglK remains unknown. The same is true for RfrA, for which a putative role in regulating an uncharacterized manganese uptake system was proposed, but the structure of RfrA and the nature of the putative manganese uptake system remains uncharacterized. Similarly, HetL has been shown to play a role in regulating heterocyst differentiation, and the structure of HetL has been determined, but no connection between the structure and the proposed biochemical function has been elucidated. The remaining structure/function space of the PRP superfamily remains completely uncharacterized.

What we can deduce at this point regarding the structure and function of PRPs, and multi-domain proteins containing PRP domains, is that in some cases the function is a consequence of the shape and electrostatic surface potential of the PRP, as is observed in the case of DNA mimicry in MfpA[10] and the Qnr [17-19, 33, 35]family of proteins. In other cases, the PRP domain simply acts as a scaffold to provide a surface to support binding interactions with another protein, as with the SV2C-LD binding to BoNT/A.[43-46] We also

have seen that variations on the PRP scaffold structure can play functional roles, such as the extra-β-helix loop excursions observed in some Qnr-family proteins, such as QnrB1[18] and AhQnr[19], that appear to play critical roles in guiding interactions with the DNA gyrase. We also observe many subtle variations in Rfr fold structures that may turn out to be important to function, including small bulges that project from the β-helix structure, variations in the compositions of the β-turns, i.e. the mixture of type I, type II and type IV β turns, that cause subtle changes in the β-helix dimensions or β-helix twist,[20] as well as the presence or absence of N-terminal or C-terminal capping α-helices, and whether or not the PRP constitutes a domain in a multidomain protein.

In closing, while the structure-space of PRPs becomes richer, our understanding of the biochemical function of members of the PRP superfamily lags increasingly behind. Targeted and carefully designed studies are required to begin to chip away at expanding our understanding the repertoire of structures and functions of the enigmatic members of the PRP superfamily.

## 2.6 Acknowledgements

The research was conducted with the support of Miami University. MAK acknowledges support of Miami University and the Ohio Board of Regents with funds used to establish the Ohio Eminent Scholar Laboratory where the work was performed. Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

## 2.7 References

1. Ray, N.; Cavin, X.; Paul, J. C.; Maigret, B., Intersurf: dynamic interface between proteins. Journal of molecular graphics & modelling 2005, 23 (4), 347-54.
2. Ehira, S.; Ohmori, M., NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium Anabaena sp. strain PCC 7120. Mol Microbiol 2006, 59 (6), 1692-703.
3. Black, K.; Buikema, W. J.; Haselkorn, R., The hglK gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium Anabaena sp. strain PCC 7120. Journal of Bacteriology 1995, 177 (22), 6440-6448.
4. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 2004, 25 (13), 1605-12.
5. Bateman, A.; Murzin, A. G.; Teichmann, S. A., Structure and distribution of pentapeptide repeats in bacteria. Protein Science 1998, 7 (6), 1477-1480.
6. Ni, S. S.; Sheldrick, G. M.; Benning, M. M.; Kennedy, M. A., The 2 angstrom resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp strain PCC 7120. Journal of Structural Biology 2009, 165 (1), 47-52.
7. Martínez-Martínez, L.; Pascual, A.; Jacoby, G. A., Quinolone resistance from a transferable plasmid. Lancet (London, England) 1998, 351 (9105), 797-9.
8. Rozwandowicz, M.; Brouwer, M. S. M.; Fischer, J.; Wagenaar, J. A.; Gonzalez-Zorn, B.; Guerra, B.; Mevius, D. J.; Hordijk, J., Plasmids carrying antimicrobial

resistance genes in Enterobacteriaceae. Journal of Antimicrobial Chemotherapy 2018, 73 (5), 1121-1137.
9. Montero, C.; Mateu, G.; Rodriguez, R.; Takiff, H., Intrinsic resistance of Mycobacterium smegmatis to fluoroquinolones may be influenced by new pentapeptide protein MfpA. Antimicrob Agents Chemother 2001, 45 (12), 3387-92.
10. Hegde, S. S.; Vetting, M. W.; Roderick, S. L.; Mitchenall, L. A.; Maxwell, A. M.; Takiff, H. E.; Blanchard, J. S., A fluoroquinolone resistance protein from mycobacterium tuberculosis that mimics DNA. Abstracts of Papers of the American Chemical Society 2005, 230, U538-U539.
11. Jacoby, G. A.; Hooper, D. C., Phylogenetic analysis of chromosomally determined qnr and related proteins. Antimicrob Agents Chemother 2013, 57 (4), 1930-4.
12. Chandler, L. E.; Bartsevich, V. V.; Pakrasi, H. B., Regulation of Manganese Uptake in Synechocystis 6803 by RfrA, a Member of a Novel Family of Proteins Containing a Repeated Five-Residues Domain. Biochemistry 2003, 42 (18), 5508-5514.
13. Buchko, G. W.; Ni, S. S.; Robinson, H.; Welsh, E. A.; Pakrasi, H. B.; Kennedy, M. A., Characterization of two potentially universal turn motifs that shape the repeated five-residues fold - Crystal structure of a lumenal pentapeptide repeat protein from Cyanothece 51142. Protein Science 2006, 15 (11), 2579-2595.
14. Buchko, G. W., Pentapeptide Repeat Proteins and Cyanobacteria. HANDBOOK ON CYANOBACTERIA: BIOCHEMISTRY, BIOTECHNOLOGY AND APPLICATIONS 2009, 233.
15. Vetting, M. W.; Hegde, S. S.; Hazleton, K. Z.; Blanchard, J. S., Structural characterization of the fusion of two pentapeptide repeat proteins, Np275 and Np276, from Nostoc punctiforme: Resurrection of an ancestral protein. Protein Science 2007, 16 (4), 755-760.
16. Buchko, G. W.; Robinson, H.; Pakrasi, H. B.; Kennedy, M. A., Insights into the structural variation between pentapeptide repeat proteins - Crystal structure of Rfr23 from Cyanothece 51142. Journal of Structural Biology 2008, 162 (1), 184-192.
17. Vetting, M. W.; Hegde, S. S.; Blanchard, J. S., Crystallization of a pentapeptide-repeat protein by reductive cyclic pentylation of free amines with glutaraldehyde. Acta Crystallographica Section D-Biological Crystallography 2009, 65, 462-469.
18. Vetting, M. W.; Hegde, S. S.; Wang, M. H.; Jacoby, G. A.; Hooper, D. C.; Blanchard, J. S., Structure of QnrB1, a Plasmid-mediated Fluoroquinolone Resistance Factor. Journal of Biological Chemistry 2011, 286 (28), 25265-25273.
19. Xiong, X. L.; Bromley, E. H. C.; Oelschlaeger, P.; Woolfson, D. N.; Spencer, J., Structural insights into quinolone antibiotic resistance mediated by pentapeptide repeat proteins: conserved surface loops direct the activity of a Qnr protein from a Gram-negative bacterium. Nucleic Acids Research 2011, 39 (9), 3917-3927.
20. Zhang, R.; Ni, S.; Kennedy, M. A., Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7 angström resolution. Journal of Structural Biology: X 2019, 3, 100010.
21. Xu, S.; Kennedy, M. A., Structural dynamics of pentapeptide repeat proteins. Proteins: Structure, Function, and Bioinformatics 2020, 88 (11), 1493-1512.
22. Arévalo, S.; Flores, E., Pentapeptide-repeat, cytoplasmic-membrane protein HglK influences the septal junctions in the heterocystous cyanobacterium Anabaena. Mol Microbiol 2020, 113 (4), 794-806.

23. Botello-Morte, L.; González, A.; Bes, M. T.; Peleato, M. L.; Fillat, M. F., Chapter Four - Functional Genomics of Metalloregulators in Cyanobacteria. In Advances in Botanical Research, Chauvat, F.; Cassier-Chauvat, C., Eds. Academic Press: 2013; Vol. 65, pp 107-156.
24. Eisenhut, M., Manganese Homeostasis in Cyanobacteria. Plants (Basel, Switzerland) 2019, 9 (1).
25. Vilchèze, C.; Jacobs, W. R., Jr., Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis: Genes, Mutations, and Causalities. Microbiology spectrum 2014, 2 (4), Mgm2-0014-2013.
26. Berning, S. E., The role of fluoroquinolones in tuberculosis today. Drugs 2001, 61 (1), 9-18.
27. Arsène, S.; Leclercq, R., Role of a qnr-like gene in the intrinsic resistance of Enterococcus faecalis to fluoroquinolones. Antimicrobial agents and chemotherapy 2007, 51 (9), 3254-3258.
28. Strahilevitz, J.; Jacoby, G. A.; Hooper, D. C.; Robicsek, A., Plasmid-mediated quinolone resistance: a multifaceted threat. Clinical microbiology reviews 2009, 22 (4), 664-89.
29. Wang, M. H.; Guo, Q. L.; Xu, X. G.; Wang, X. Y.; Ye, X. Y.; Wu, S.; Hooper, D. C.; Wang, M. G., New Plasmid-Mediated Quinolone Resistance Gene, qnrC, Found in a Clinical Isolate of Proteus mirabilis. Antimicrobial Agents and Chemotherapy 2009, 53 (5), 1892-1897.
30. Cavaco, L. M.; Hasman, H.; Xia, S.; Aarestrup, F. M., qnrD, a Novel Gene Conferring Transferable Quinolone Resistance in Salmonella enterica Serovar Kentucky and Bovismorbificans Strains of Human Origin. Antimicrobial Agents and Chemotherapy 2009, 53 (2), 603-608.
31. Jacoby, G.; Cattoir, V.; Hooper, D.; Martinez-Martinez, L.; Nordmann, P.; Pascual, A.; Poirel, L.; Wang, M. G., qnr gene nomenclature. Antimicrobial Agents and Chemotherapy 2008, 52 (7), 2297-2299.
32. Li, X. J.; Zhang, Y. J.; Zhou, X. T.; Hu, X. L.; Zhou, Y. X.; Liu, D.; Maxwell, A.; Mi, K. X., The plasmid-borne quinolone resistance protein QnrB, a novel DnaA-binding protein, increases the bacterial mutation rate by triggering DNA replication stress. Molecular Microbiology 2019, 111 (6), 1529-1543.
33. Vetting, M. W.; Hegde, S. S.; Zhang, Y.; Blanchard, J. S., Pentapeptide-repeat proteins that act as topoisomerase poison resistance factors have a common dimer interface. Acta Crystallographica Section F-Structural Biology Communications 2011, 67, 296-302.
34. Hashimi, S. M.; Wall, M. K.; Smith, A. B.; Maxwell, A.; Birch, R. G., The phytotoxin albicidin is a novel inhibitor of DNA gyrase. Antimicrobial Agents and Chemotherapy 2007, 51 (1), 181-187.
35. Notari, L.; Martinez-Carranza, M.; Farias-Rico, J. A.; Stenmark, P.; von Heijne, G., Cotranslational Folding of a Pentarepeat beta-Helix Protein. Journal of Molecular Biology 2018, 430 (24), 5196-5206.
36. Diao, J.; Zhang, Y.; Huibregtse, J. M.; Zhou, D.; Chen, J., Crystal structure of SopA, a Salmonella effector protein mimicking a eukaryotic ubiquitin ligase. Nature Structural & Molecular Biology 2008, 15 (1), 65-70.

37. Lin, D. Y. W.; Diao, J. B.; Chen, J., Crystal structures of two bacterial HECT-like E3 ligases in complex with a human E2 reveal atomic details of pathogen-host interactions. Proceedings of the National Academy of Sciences of the United States of America 2012, 109 (6), 1925-1930.
38. Fiskin, E.; Bhogaraju, S.; Herhaus, L.; Kalayil, S.; Hahn, M.; Dikic, I., Structural basis for the recognition and degradation of host TRIM proteins by Salmonella effector SopA. Nature Communications 2017, 8.
39. Kamanova, J.; Sun, H.; Lara-Tejero, M.; Galan, J. E., The Salmonella Effector Protein SopA Modulates Innate Immune Responses by Targeting TRIM E3 Ligase Family Members. Plos Pathogens 2016, 12 (4).
40. Wang, Y.; Argiles-Castillo, D.; Kane, E. I.; Zhou, A.; Spratt, D. E., HECT E3 ubiquitin ligases – emerging insights into their biological roles and disease relevance. Journal of Cell Science 2020, 133 (7), jcs228072.
41. Deshaies, R. J.; Joazeiro, C. A., RING domain E3 ubiquitin ligases. Annual review of biochemistry 2009, 78, 399-434.
42. Lin, D. Y. W.; Diao, J. B.; Zhou, D. G.; Chen, J., Biochemical and Structural Studies of a HECT-like Ubiquitin Ligase from Escherichia coli O157:H7. Journal of Biological Chemistry 2011, 286 (1), 441-449.
43. Benoit, R. M.; Frey, D.; Hilbert, M.; Kevenaar, J. T.; Wieser, M. M.; Stirnimann, C. U.; McMillan, D.; Ceska, T.; Lebon, F.; Jaussi, R.; Steinmetz, M. O.; Schertler, G. F. X.; Hoogenraad, C. C.; Capitani, G.; Kammerer, R. A., Structural basis for recognition of synaptic vesicle protein 2C by botulinum neurotoxin A. Nature 2014, 505 (7481), 108-+.
44. Yao, G. R.; Zhang, S. C.; Mahrhold, S.; Lam, K. H.; Stern, D.; Bagramyan, K.; Perry, K.; Kalkum, M.; Rummel, A.; Dong, M.; Jin, R. S., N-linked glycosylation of SV2 is required for binding and uptake of botulinum neurotoxin A. Nature Structural & Molecular Biology 2016, 23 (7), 656-662.
45. Benoit, R. M.; Scharer, M. A.; Wieser, M. M.; Li, X. D.; Frey, D.; Kammerer, R. A., Crystal structure of the BoNT/A2 receptor-binding domain in complex with the luminal domain of its neuronal receptor SV2C. Scientific Reports 2017, 7.
46. Gustafsson, R.; Zhang, S. C.; Masuyer, G.; Dong, M.; Stenmark, P., Crystal Structure of Botulinum Neurotoxin A2 in Complex with the Human Protein Receptor SV2C Reveals Plasticity in Receptor Binding. Toxins 2018, 10 (4).
47. Janz, R.; Goda, Y.; Geppert, M.; Missler, M.; Südhof, T. C., SV2A and SV2B function as redundant Ca2+ regulators in neurotransmitter release. Neuron 1999, 24 (4), 1003-16.
48. Montal, M., Botulinum Neurotoxin: A Marvel of Protein Design. In Annual Review of Biochemistry, Vol 79, Kornberg, R. D.; Raetz, C. R. H.; Rothman, J. E.; Thorner, J. W., Eds. 2010; Vol. 79, pp 591-617.
49. Dong, M.; Yeh, F.; Tepp, W. H.; Dean, C.; Johnson, E. A.; Janz, R.; Chapman, E. R., SV2 Is the Protein Receptor for Botulinum Neurotoxin A. Science 2006, 312 (5773), 592-596.
50. Li, X. D.; Brunner, C.; Wu, Y. F.; Leka, O.; Schneider, G.; Kammerer, R. A., Structural insights into the interaction of botulinum neurotoxin a with its neuronal receptor SV2C. Toxicon 2020, 175, 36-43.
51. Liu, D.; Golden, J. W., hetL overexpression stimulates heterocyst formation in Anabaena sp. strain PCC 7120. J Bacteriol 2002, 184 (24), 6873-81.

52. Zhang, R.; Ni, S.; Kennedy, M. A., Crystal structure of Alr1298, a pentapeptide repeat protein from the cyanobacterium Nostoc sp. PCC 7120, determined at 2.1 Å resolution. Proteins: Structure, Function, and Bioinformatics n/a (n/a).
53. Jacob, F.; Monod, J. In On the regulation of gene activity, Cold Spring Harbor symposia on quantitative biology, Cold Spring Harbor Laboratory Press: 1961; pp 193-211.
54. Lawrence, J. G., Shared strategies in gene organization among prokaryotes and eukaryotes. Cell 2002, 110 (4), 407-13.
55. Ralston, A., Operons and prokaryotic gene regulation. Nature Education 2008, 1 (1), 216.
56. Ni, S. S.; McGookey, M. E.; Tinch, S. L.; Jones, A. N.; Jayaraman, S.; Tong, L.; Kennedy, M. A., The 1.7 A resolution structure of At2g44920, a pentapeptide-repeat protein in the thylakoid lumen of Arabidopsis thaliana. Acta Crystallographica Section F-Structural Biology Communications 2011, 67, 1480-1484.
57. Agrawal, G. K.; Yonekura, M.; Iwahashi, Y.; Iwahashi, H.; Rakwal, R., System, trends and perspectives of proteomics in dicot plants Part I: Technologies in proteome establishment. Journal of chromatography. B, Analytical technologies in the biomedical and life sciences 2005, 815 (1-2), 109-23.
58. Barbazuk, W. B.; Fu, Y.; McGinnis, K. M., Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome research 2008, 18 (9), 1381-92.

# Chapter 3: Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7 Angström resolution

Reproduced with permission from:

Ruojing Zhang[1], Shuisong Ni[1], Michael A. Kennedy*[1]

[1]Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056

*Corresponding Author: Department of Chemistry and Biochemistry, 106 Hughes Laboratories, Miami University, 651 East High Street, Oxford, OH 45056. Email: kennedm4@miamioh.edu. Phone: 513-529-8267. Fax: 513-529-5715.

This paper has been published in Journal of Structural Biology: X, 2019, 3:100010



Author contributions: RZ contributed to plasmid preparation, protein expression, data collection, and data analysis, manuscript preparation. SN and MAK contributed to data analysis and manuscript preparation.

## 3.1 Abstract

Pentapeptide repeat proteins (PRPs), found abundantly in cyanobacteria, number in the dozens in some genomes, e.g. in Nostoc sp. PCC 7120. PRPs, comprised of a repeating consensus sequence of five amino acids, adopt a distinctive right-handed quadrilateral β-helical structure, also referred to as a repeat five residue (Rfr) fold, made up of stacks of coils formed by four consecutive pentapeptide repeats. The right-handed quadrilateral β-helical PRP structure is constructed by repeating β turns at each of four corners in a given coil, each causing a 90° change in direction of the polypeptide chain. Until now, all PRP structures have consisted either of type II and IV β turns or exclusively of type II β turns. Here, we report the first structure of a PRP comprised of type I and II β turns, Alr5209 from Nostoc sp. PCC 7120. The alr5209 gene encodes 129 amino acids containing 16 tandem pentapeptide repeats. The Alr5209 structure was analyzed in comparison to all other PRPs to determine how type I β turns can be accommodated in Rfr folds and the consequences of type I β turns on the right-handed quadrilateral β-helical structure. Given that Alr5209 represents the first PRP structure containing type I β turns, the PRP consensus sequence was reevaluated and updated. Despite a growing number of PRP structural investigations, their function remains largely unknown. Genome analysis indicated that alr5209 resides in a five-gene operon (alr5208-alr5212) with Alr5211 annotated to be a NADH dehydrogenase indicating Alr5209 may be involved in oxidative phosphorylation.

## 3.2 Introduction

Cyanobacteria, ancient prokaryotic microorganisms capable of both oxygenic photosynthesis and nitrogen fixation, are thought to be the first organisms responsible for oxygenation of the earth's atmosphere more than two billion years ago (1-3). In the filamentous Nostoc sp. Strain PCC 7120 cyanobacterium, the filaments can grow to contain several hundred cells due to division of actively dividing vegetative cells (4). Nitrogen fixation in Nostoc sp. PCC 7120 takes place in specialized cells known as heterocysts (4) that differentiate from vegetative cells under conditions of low available nitrogen. Under such conditions, 5 to 10% of the vegetative cells in the filament in Nostoc sp. PCC 7120 differentiate into heterocysts, with adjacent heterocysts regularly spaced by about ten vegetative cells (5), thus providing a source of nitrogen to the surrounding vegetative cells in the filament. Both the vegetative cells and heterocysts in filaments of Nostoc sp. PCC 7120 are capable of performing multiple functions to adapt to changing conditions in their surroundings. The adaptability of Nostoc sp. PCC 7120 to its environment requires both vegetative and heterocyst cells to carry out many biochemical functions including photosynthesis, nitrogen fixation, signal communication and cell differentiation (6). In 2001, the complete genome of Nostoc sp. PCC 7120, containing a 6.4 Mb chromosome and six plasmids, was sequenced and 6228 proteins were predicted to be encoded by the chromosome. Given the availability of its complete genome sequence and the fact that filamentous cyanobacteria represent among the oldest and simplest living organisms to exhibit cell differentiation (7), Nostoc sp. PCC 7120 has become an important model organism to study biochemical functions found in cyanobacteria (8).

Pentapeptide repeat proteins (PRPs) represent a large superfamily of proteins with 52,787 sequences grouped into four clans in the Pfam database (9). Analysis of the largest PRP clan, represented by the Pentapeptide family (Pfam 00805), that includes 38471 sequences from 3485 species indicates that ~90% of the sequences belong to bacteria and archaea while ~10% of sequences belong to eukaryotes. Further analysis indicates that nearly half of the PRP sequences in bacteria belong to cyanobacteria and that PRPs are most abundant in cyanobacteria in terms of the numbers of PRPs per genome (10). PRPs, defined as proteins containing at least eight tandem repeating sequences of five amino acids with a consensus sequence originally defined as A[D/N]LXX in 1998 (11), also referred to here as PRP domains, adopt a distinctive right-handed β-helical structure composed of stacks of coils composed of four pentapeptide repeats. Thirty PRPs have been identified in Nostoc sp. PCC 7120, including HglK (All0813), a membrane protein reported to be involved to the localization of heterocyst-specific glycolipids (3). In 2009, the structure of HetL, a PRP from Nostoc sp. PCC 7120 containing 40 tandem repeats involved in regulating differentiation of heterocysts, was reported (10). Despite the important role that cyanobacteria played in evolution of the earth's atmosphere and oxygen-based life on earth, and the relative abundance of PRPs in cyanobacteria, the biochemical functions of PRPs remain largely unknown and only sixteen PRP structures have been reported (12-16).

In this study, we determined the structure of Alr5209, a PRP found in Nostoc sp. PCC 7120. The structure adopts a repeat five residue (Rfr) fold composed of 16 tandem PRP domains. The resulting right-handed β helix is composed of four coils held together by β ladders composed of β bridges on each face and a 1:3 mixture of type I and type II β turns. Alr5209 is the first PRP reported to contain type I β-turns in its Rfr fold. The structural consequences of including type I turns in the Rfr fold are examined and discussed. Combined structure and sequence analysis of Alr5209 enabled refinement of the pentapeptide consensus sequences that encode PRPs, which should allow for more sensitive and accurate prediction of PRPs in existing and newly reported genomes. Finally, a gene cluster analysis based on the KEGG database indicated that Alr5209 may be involved in oxidative phosphorylation.

### 3.3 Materials and Method

#### 3.3.1 Cloning, expression and purification

The *alr5209* gene was amplified from the genomic DNA of Nostoc. PCC7120 using standard PCR methods. Based on analysis of the KEGG sequence for *alr5209*, the following two primers were designed containing Ndel and Xhol ligation sites to facilitate construction of the expression plasmid:
cccgcccgcatATGTCTGAAGTCAATTATCAACAG and
gcccgctcgagttaTTGTTCTTTGAGTTGCAAGCC. The PCR product was cloned into the pET28b expression vector (Novagen, Inc.) under the control of the T7 promoter, and the construct contained a N-terminal 6xHis tag to allow purification by nickel affinity chromatography. The constructed plasmid was transformed into JM109 competent cells (Novagen, Inc.), spread on agar plates and resulting colonies collected for sequencing. After sequencing confirmed successful cloning of the *alr5209* gene into the expression plasmid, the plasmid was transformed into the *Escherichia coli* BL21 (DE3) (Novagen,

Inc) host strain for overexpression of Alr5209 protein. Protein was isolated from a one-liter culture grown in M9 minimal medium using N15-labeled ammonium chloride as a nitrogen source to enable isotopic labeling for future nuclear magnetic resonance spectroscopy experiments. Cell growth in the bacterial culture was maintained at 37 ℃ with 250 rpm shaking until the $OD_{600}$ reached to 0.6 - 0.8. At this point, the cell culture was cooled to 15 ℃ and 0.5 mL 1 M Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5mM. The culture was then incubated at 15 ℃ with 250 rpm shaking overnight. The cells were collected using 5000 g centrifugation at 4 ℃ for 20 min. The resulting cell pellet was resuspended in 20 mL B1 buffer (20 mM Tris, 250 mM NaCl, 10% glycerol, PH7.8) and the resuspended cells were lysed by three passes through a French press (Thermo, Inc.). The cell lysate was centrifugated at 17418 g for 30 minutes. The His-tagged protein in the supernatant was purified on a 20 mL Ni-NTA affinity column (Qiagen). Proteins in the supernatant lacking a His-tag were removed during successive washing steps with 60 mL B1 buffer containing 0 and 30 mM imidazole, respectively. The purified His-tagged Alr5209 protein eluted with 300 mM imidazole was then dialyzed three times with 1 L B1 buffer to remove imidazole. Purified Alr5209 protein was confirmed by SDS-PAGE gel and concentrated to final concentration of 35 mg/mL.

### 3.3.2 Crystallization, data collection, phasing and refinement

Crystallization conditions were determined using the Hampton Research kit (HR2-112 and HR2-121) to screen for protein crystallization. Screening was performed by combining 1 μL of protein with 1 μL of each buffer on a 48-well plate using the hanging-drop vapor-diffusion method. Plates were maintained at room temperature. Overlapped spherical crystals were obtained in a buffer containing 0.2 M potassium sodium tartrate tetrahydrate, 0.1 M sodium citrate tribasic dihydrate pH 5.6, 2.0 M ammonium sulfate. These crystals were crushed in 50 μL crystallization buffer using a crystal crusher and by glass beads to make a stock seeding solution. Final cubic crystals were obtained by adding 0.5 μL of a 10,000x diluted seeding solution to 1 μL protein and 1 μL cryo-buffer, consisting 0.15 M potassium sodium tartrate tetrahydrate, 0.075 M sodium citrate tribasic dihydrate pH 5.6, 1.5 M ammonium sulfate, 25% v/v glycerol.

All experiment diffraction data were collected at the Advanced Photon Source (APS) at Argonne National Laboratory using the beamline 31-ID at 100 K. Truncated I to F experimental data analyzed by CCP4 7.0.057 were submitted to the CCP4 online server and a molecular replacement solution was found by BALBES (17) using the PDB ID 2J8I structure as a starting model (18). Manual model building was performed using COOT (19). Phenix 1.13 (20) was used for phasing improvement, automatic amino acid building and refinement. The final structure was submitted to the Protein Data Bank (PDB ID: 6OMX). The electrostatic potential surface was calculated using the PDB2PQR server (21) and depicted using the Chimera software (22).

### 3.3.3 Secondary structure and sequence analysis

Distances in the PRPs and the φ and ψ angles were measured using Chimera (22). All distance measurements were performed on PRPs with known structures and structure-based sequence alignment starting from the first, N-terminal pentapeptide repeat domain. The face of the right-handed β-helices containing the first, N-terminal complete

pentapeptide repeat domain was designated as face 1, except for 3PSS whose first pentapeptide in coil 1 was incomplete. The face of the right-handed β-helices containing the second pentapeptide repeat domain was designated as face 2, and so on. The β turn types and distributions were measured from the PDB coordinates of published structures. The length of each face was measured from the carbonyl carbon of the i-2 amino acid to that of the i+2 amino acid for each face. The face-to-face distances between the 1 and 3 faces were measured from the carbonyl carbon of the i-residue in face 1 to the carbonyl carbon of the i residue in face 3. The face-to-face distances between the 2 and 4 faces were measured from the carbonyl carbon of the i-residue in face 2 to the carbonyl carbon of the i-residue in face 4. The distances across the face 1 to face 2 turns were measured from the carbonyl carbon of the i-residue in face 1 to the carbonyl carbon of the i-residue in face 2. The distances across the face 1 to face 4 turns were measured from the carbonyl carbon of the i-residue in face 1 to the carbonyl carbon of the i-residue in face 4. Any PRP coils interrupted by an inner loop or other secondary structures rather than β helix were not counted in the summary statistics. Consensus sequence distribution plots were completed using the Web Logo server (23,24). Sequences belonging to secondary structures other than the PRP domains were not included in the consensus sequence analysis. For calculation of twist angle among coils, the angle calculation tool in Chimera was used (22). Twist angles were measured as the angle between the two vectors defined by the carbonyl carbons of the i-2 and i+2 amino acids from coil and the carbonyl carbons of the i-2 and i+2 amino acids of the following coil. Once those vectors were defined based on those two pairs of atoms, the twist angles were determined using the angle calculation tool. Due to the influence of an α helix near the N-terminus, the twist angles in 6OMX, 2J8K, 3PSS and 3DU1 were measured between the second coil and subsequent coils.

### 3.3.4 Circular dichroism (CD) spectroscopy and thermal protein denaturation

Purified protein was dialyzed and diluted to final concentration at 20 μM with 20 mM potassium phosphate pH 7.8 and 150 mM NaF buffer. Diluted protein samples were loaded into 1 mL quartz cuvettes. Experiments were performed with AVIV model435 circular dichroism spectrophotometer (Aviv Biomedical, Inc). Far-UV wavelength spectra were recorded from 180 nm to 300 nm to determine a suitable wavelength for temperature melting experiments at 25 ℃. Thermal denaturation curves for 20 μM samples were collected both at 226 nm and 210 nm, separately, from 15 to 85 ℃ using 1 ℃ intervals. Wavelength scans were measured for both samples at 85 and 95 ℃ after the thermal denaturation experiments. Experiments with buffer only were performed under the same conditions as with the protein samples and used as blanks for correction.

## 3.4 Results and Discussion

### 3.4.1 Crystal and data quality of Alr5209

Original crystals were spherical and overlapping. High-quality single crystals were obtained using seeding and addition of glycerol. Crystals used for diffraction data collection were orthorhombic (unit cell dimensions: a=71.001 Å, b=27.835 Å, c=60.837 Å, α=β=γ=90°) and the space group was $P222_1$. Single wavelength data collected at 0.97931 Å was used for molecular replacement. The data was truncated to 1.71 Å with an overall completeness of 98.71% measured for 13598 unique reflections. Xtriage (25,26)

analysis indicated a single molecule in the asymmetric unit with a solvent content of 0.407. Molecular replacement phasing was accomplished using 2J8I as a starting model. The final structure included 121 out of 129 amino acids with six residues missing at the N-terminus and two residues missing at the C-terminal end. The structure quality was checked using MolProbity (27) and the PDB validation server. The report showed no Ramachandran outliers and clash scores and sidechain outliers were 2 and 2.1%, respectively. All data and refinement statistics are listed in **Table 3.1**.

| | |
|---|---|
| Resolution range (Å) | 35.5 - 1.706 (1.767 - 1.706)[1] |
| Space group | $P222_1$ |
| Unit cell (Å, °) | $\alpha$= 71.001 $\beta$= 27.835 $\gamma$= 60.837 $\alpha$= $\beta$= $\gamma$=90 |
| Total reflections | 27104 (2671) |
| Unique reflections | 13598 (1339) |
| Multiplicity | 2.0 (2.0) |
| Completeness (%) | 98.71 (99.26) |
| Mean I/sigma (I) | 21.77 (2.49) |
| Wilson B-factor ($Å^2$) | 28.66 |
| R-merge | 0.00942 (0.2226) |
| R-means | 0.01332 (0.3148) |
| R-pim | 0.00942 (0.2226) |
| CC1/2 | 1 (0.91) |
| CC* | 1 (0.976) |
| Reflections used in refinement | 13584 (1336) |
| Reflections used for R-free | 1359 (133) |
| R-work | 0.2194 (0.3275) |
| R-free | 0.2468 (0.3726) |
| CC (work) | 0.974 (0.785) |
| CC (free) | 0.960 (0.704) |
| Number of non-hydrogen atoms | 938 |
| macromolecules | 925 |
| solvent | 13 |
| Protein residues | 121 |
| RMS (bonds) (Å) | 0.006 |
| RMS (angles) (°) | 1.16 |

| | |
|---|---|
| Ramachandran favored (%) | 99.16 |
| Ramachandran allowed (%) | 0.84 |
| Ramachandran outliers (%) | 0.00 |
| Rotamer outliers (%) | 0.00 |
| Clashscore | 2.18 |
| Molprobity | 0.99 |
| Average B-factor ($Å^2$) | 42.34 |
| Macromolecules ($Å^2$) | 42.22 |
| Solvent ($Å^2$) | 50.75 |

**Table 3. 1 Summary of data collection and structure refinement data for Alr5209.**

### 3.4.2 Structure analysis of Alr5209

Alr5209 contained 16 pentapeptide repeat domains (**Figure 3.1**) that formed a right-handed quadrilateral β helix consisting of a stack of four Rfr coils with α-helices at the N- and C-termini (**Figure 3.2**). The N-terminal α-helix contained nine amino acids (13-VATLIEMYT-21) while the C-terminal α-helix was shorter being comprised of four amino acids (119-LLKA-122). The Rfr folds of PRPs are constructed by four β-turns per coil with the type of β-turn being defined by combinations of φ and ψ angles of the residues involved in making up the turns (28,29). Type I and II β turns are distinguished by differences in the ψ angle in the i+1 position and the φ angle in the i+2 position. In canonical type II β turns, the φ and ψ angles are +80° and +120° (29) (**Table 3.2**). Based on the analysis of φ and ψ angles, Alr5209 is composed of a mixture of type I and type II β turns (**Table 3.2**). The type I β turns in Alr5209 appeared in every coil in the same single position (joining face 2 and face 3) and the rest of the turns were type II β turns. In the i+1 position of face 2, the φ and ψ angles were -61±3° and -35±4° consistent with the canonical definition of type I β turns ( -60°/-30°) (28,29). In the i+2 position of face 2, the φ and ψ angles were -127±4° and 32±2°, whereas the the φ and ψ angles of the i+2 residues in that canonical definition of type I β turns are -90° and 0°, respectively. Therefore, the φ and ψ angles of the i+2 residues are -/+30° from the canonical type I values, respectively, putting them just outside the edge of canonical values used to define type I β turns (**Figure 3.3**) (28). While all other PRPs contain combinations of type II and type IV β turns, Alr5209 is the only known PRP that contains exclusively type I β turns in the same corner of its Rfr solenoid (**Figure 3.3**). Close inspection of the graphs in Figure 3 reveals that three PRPs classified as containing mixtures of type II and type IV β turns contain one (2W7Z and 6FLS) or two (2XTZ) type I β turns, respectively. The remaining PRPs classified as containing mixtures of type II and type IV β turns (2BM5, 2G0Y, 2XT2 and 3PSS) did not contain any β turns that could be classified as type I β turns.

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | MSEVNYQQPISTVATLIEMYTAGR 24 | | | | | |
| | Face 1 | Face 2 | Face 3 | Face 4 | | Coil |
| | -2 -1 i +1 +2 | -2 -1 i +1 +2 | -2 -1 i +1 +2 | -2 -1 i +1 +2 | | |
| 25 | R D F N R | A E L G D | A N L Q N | V D I K G | 44 | C1 |
| 45 | S D L S Y | A D L S T | A N L R G | A N L R G | 64 | C2 |
| 65 | T D L S F | A D L S Q | A D L Q D | A D L R G | 84 | C3 |
| 85 | A L L M S | A N L R Q | A N L Q G | A K L E K | 104 | C4 |
| 105 | ADCDRNTHFPENFDLLKAGLQLKEQ 129 | | | | | |

**Figure 3. 1 Alignment of the PRP domains in Alr5209 based on its structure.** Underlined residues were not visible in the electron density and were not modeled, α-helical residues are highlighted in yellow. Residues 25 to 104 comprised the pentapeptide repeat domains defining the Rfr solenoid.

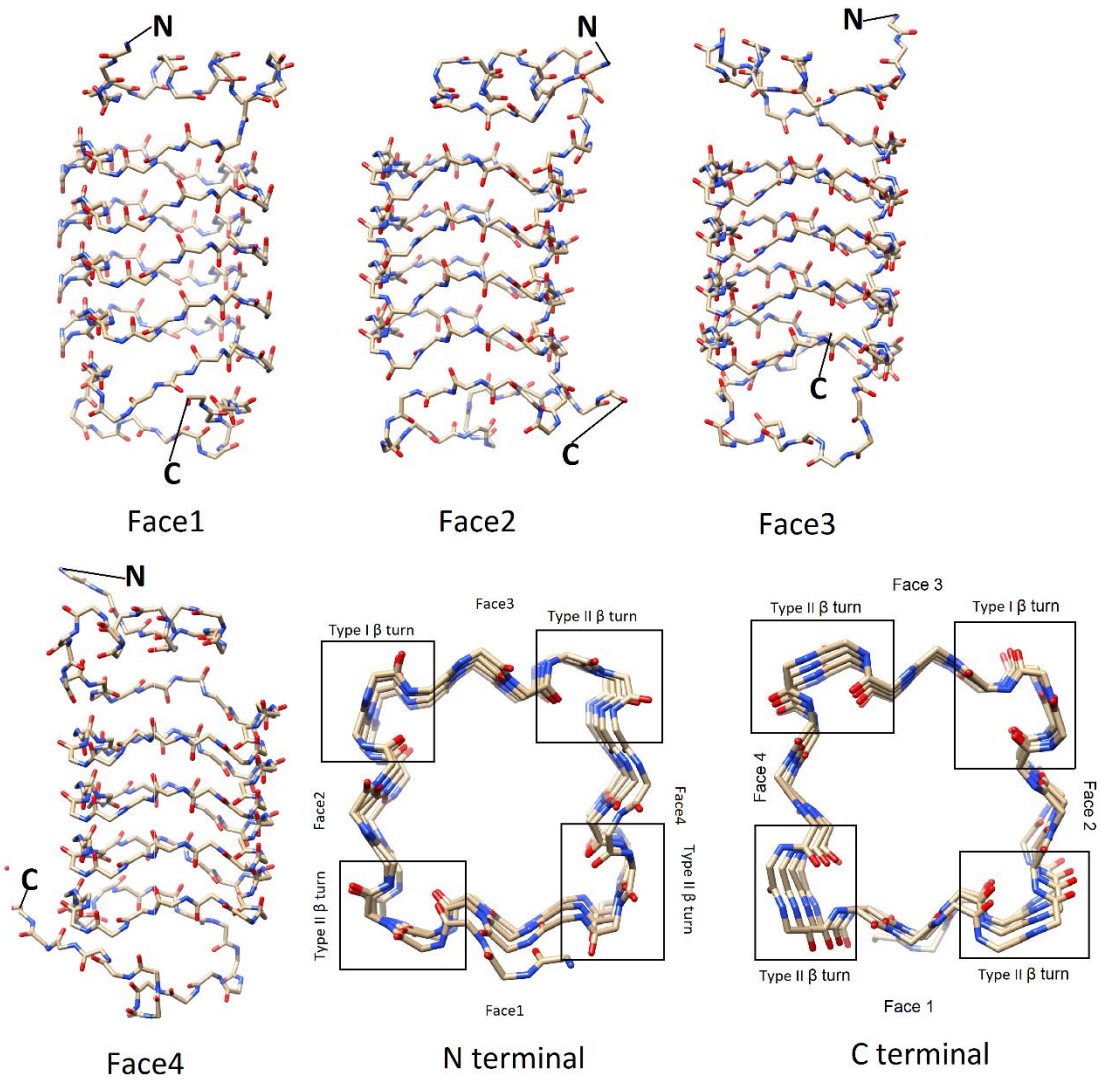



**Figure 3. 2 Overview of the backbone structure in the Rfr fold of Alr5209.** The four faces of the Alr5209 PRP structure are depicted using a stick representation colored by heteroatom type. N- and C-termini are labeled in each representation. The two on-axis views are depicted at the lower right excluding the α-helix facing the viewer for clarity. The type of β-turn type was labeled for the on-axis views.

| | Face 1 | | Face 2 | | Face 3 | | Face 4 | |
|---|---|---|---|---|---|---|---|---|
| | φ (°) | ψ (°) | φ (°) | ψ (°) | φ (°) | ψ (°) | φ (°) | ψ (°) |
| i-2 | -4±75 | 112±73 | -70±9 | 147±7 | -71±3 | 146±2 | -74±3 | 150±5 |
| i-1 | -93±6 | 109±5 | -100±5 | 102±5 | -91±3 | 107±3 | -103±8 | 107±3 |
| i | -117±7 | 23±7 | -123±6 | 34±2 | -115±4 | 25±7 | -118±7 | 23±11 |
| i+1 | -56±7 | 135±5 | -61±3 | -35±4[2] | -61±1 | 128±3 | -60±3 | 135±7 |
| i+2 | 64±2 | 15±4 | -127±4[2] | 32±2 | 68±4 | 10±9 | 72±6 | 8±6 |

**Table 3. 2 Summary of φ and ψ angles for each amino acid position in the PRP domains in Alr5209.**

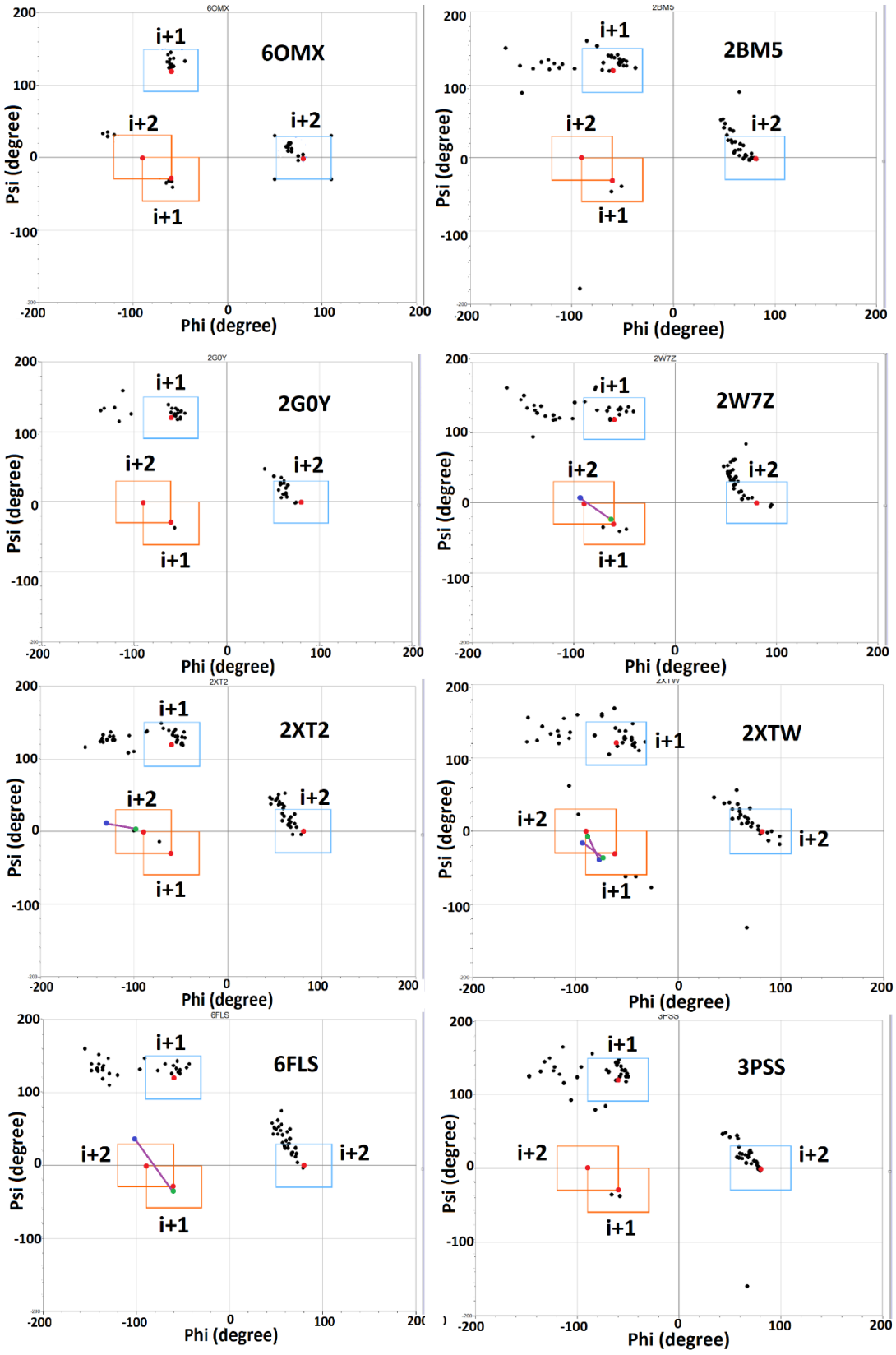

6OMX
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
2BM5
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
2G0Y
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
2W7Z
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
2XT2
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
2XTW
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
6FLS
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)
3PSS
i+1
i+2
i+2
i+1
Psi (degree)
Phi (degree)

**Figure 3. 3 Ramachandran plot of type I and type II β turns in Alr5209 in comparison to other PRPs.** The orange boxes indicate canonical values (red points) +/- 30° for type I β turns. The blue boxes indicate canonical values (red points) +/- 30° for type II β turns. Except 6OMX, all type I/IV β turns locating in or near orange part are linked by purple lines, the blue points stand i+2 and the green points are i+1.

Analysis of the Alr5209 structure showed that the direction of the inter-coil hydrogen bond linkages that establish the β-bridges in type I β turns were different in i+1 and i+2 positions compared to in the type II β turns. In both type I and type II β turns, the i+1 carbonyls always acted as hydrogen bond acceptors and the i+2 amide groups always acted as the hydrogen bond donors. However, in the type II β turns, the i+1 amino acid carbonyl hydrogen bond acceptor is always on the coil C-terminal to the coil containing the i+2 amino acid amide hydrogen bond donor (**Figure 3.4**). In contrast, in type I β turns, the linkage of hydrogen bonds establishing the β-bridges is flipped with the i+2 amino acid amide hydrogen bond donor always in the coil C-terminal to the coil containing the i+1 amino acid containing the carbonyl hydrogen bond acceptors (**Figure 3.4**).

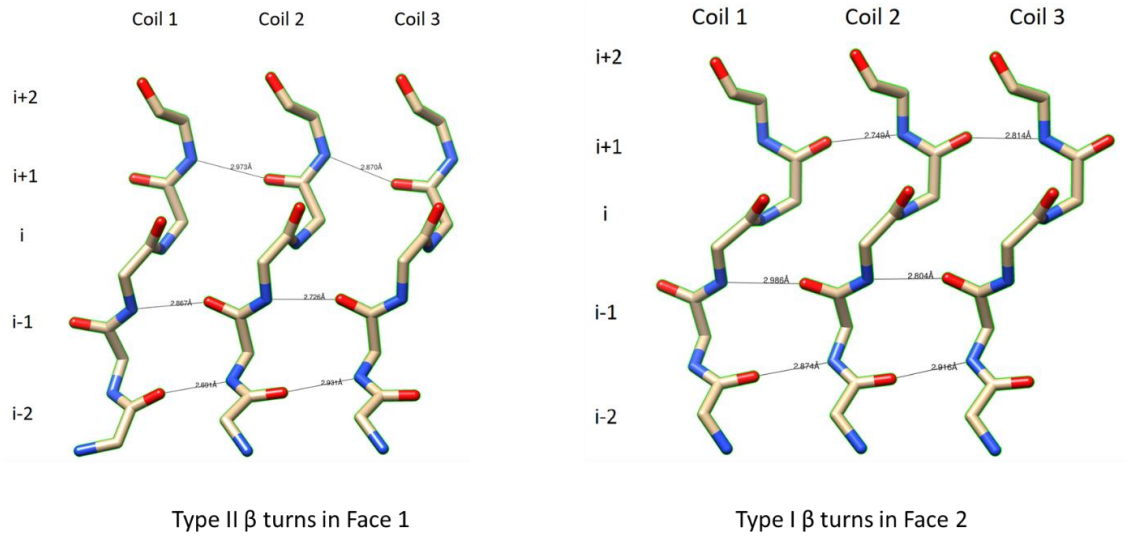



**Figure 3. 4 Details of type I and type II β turns in Alr5209.** The β turns are defined by residues in the i+2, i+1, i and i-1 positions in PRPs. The difference in the combination of the φ and □ angles that distinguish the type I and type II turns results in a change in the direction of the hydrogen bonds formed between i+1 and i+2 residues involved in stabilizing the intercoil structure involving type II turns (left) and type I turns (right).

### 3.4.3 Electrostatic potential surface of Alr5209

The electrostatic potential surface of Alr5209 is shown in **Figure 3.5**. Faces 1 and 2 were dominated by strong negative charge whereas face 3 showed a mostly neutral charge distribution and face 4 showed predominantly positive charge (**Figure 3.5**). The C-

terminal surface was neutral and the N-terminal surface contained a mixture of positive, negative and neutral charge distribution. This charge distribution would be consistent with functioning as a DNA mimic which has been reported for the fluoroquinolone resistance protein from *Mycobacterium tuberculosis* (30).

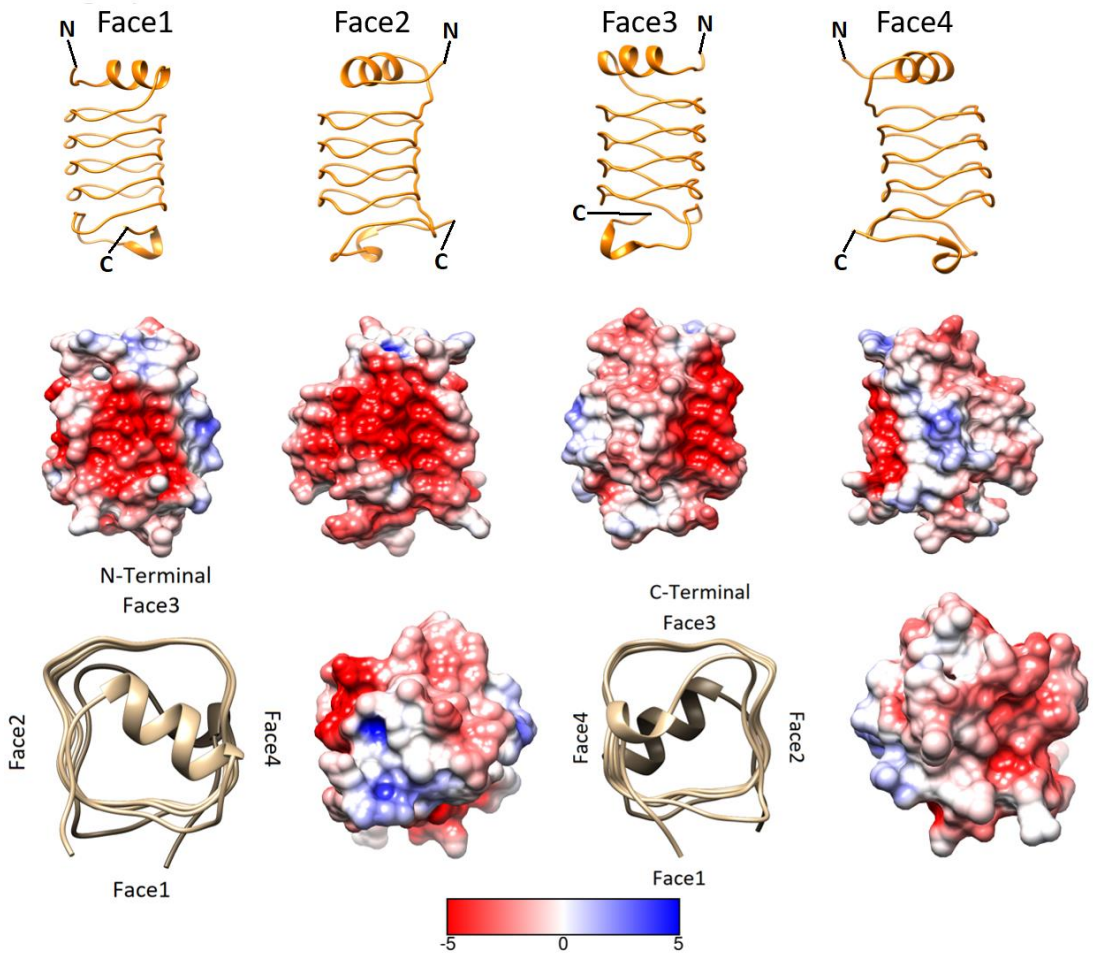



**Figure 3. 5 Electrostatic surface potential of Alr5209 for each face of the right-handed quadrilateral β helix.** The electrostatic surface potential surface is depicted for each of the four faces. The Rfr fold coil structure is depicted above each electrostatic potential plot for reference. Red indicates negative charge and blue indicates positive charge with the relative intensity indicated by the scale bar at the bottom. The electrostatic potential at the N-terminus and C-terminus of the right-handed β helix is depicted at the bottom using two on-axis plots. The scale for the surface potential color gradient has units of kT/e where 1 kT/e = 25.7 mV.

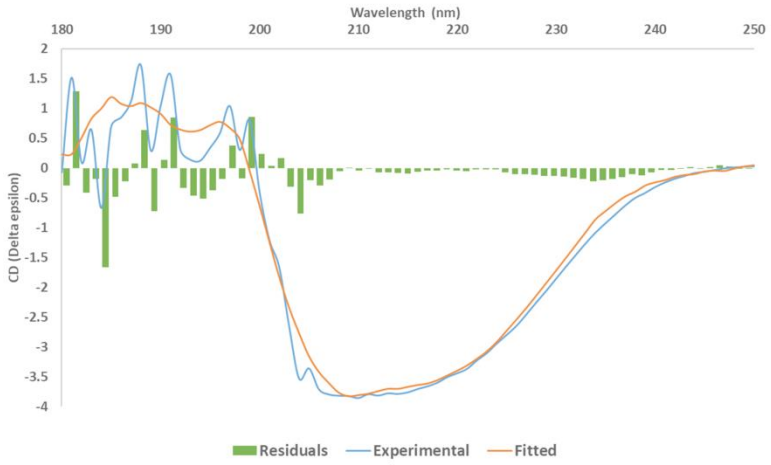
### 3.4.4 Circular dichroism spectroscopy analysis of the Alr5209 structure and thermal stability

The thermodynamic stability of the right-handed quadrilateral β-helical structure of Alr5209 was investigated by CD-monitored thermal melting analysis. The room-temperature CD spectrum was consistent with a structure dominated by type 1 and type II β-turns with short N-terminal and C-terminal α-helices (**Figure 3.6**). At 25 ℃, the strongest ellipticity appeared at 210 nm, which could be fit with a composition of secondary structural components consisting of 21.9% α-helix, 13% turn, antiparallel and parallel β-sheet occupy 13.1% and 4.5%, respectively (**Figure 3.6A, Table 3.3**). Lack of perfect fitting may reflect an incomplete basis set, for example, lack of characteristic CD contributions of type I and type II β turns in PRP structures. The CD melting experiment (**Figure 3.6B**) indicated melting temperature of Alr5209 was 58.5 ± 0.5 ℃. Compared to the average melting temperature of 62.1± 15.0 ℃ reported for a distribution of over 1100 proteins (31), the melting temperature of Alr5209 fell within the average range (32). The reverse melting experiment indicated that denatured alr5209 could be mostly refolded (73.8%) after thermal denaturation (30). Alr5209's melting temperature was ~4 ℃ lower compared to that of At2g44920 (32). The enthalpy of unfolding of Alr5209 was +30.7 kcal/mol, also significantly smaller than that of At2g44920, which was reported to be +120 kcal/mol, which could be correlated with the right-handed quadrilateral β-helix of Alr5209 being comprised of just four Rfr coils compared to the six Rfr coils present in At2g44920. The longer hydrogen bonding network in the extended Rfr coil structure of At2g44920 could require substantially more thermal energy to denature the overall right-handed quadrilateral β-helical structure compared to the Alr5209 structure, which contains only two internal Rfr coils sandwiched by two terminal Rfr coils.

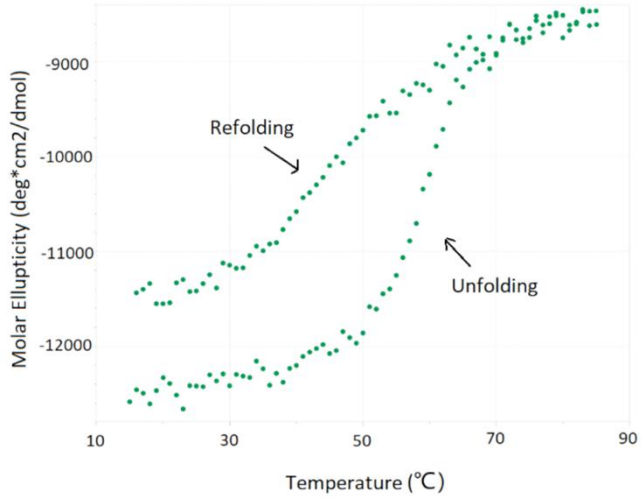
| | | | |
|---|---|---|---|
| Helix | 21.9% | Helix1 (regular) | 6.5% |
| | | Helix2 (distorted) | 15.5% |
| Antiparallel | 13.1% | Anti1 (left-twisted) | 0% |
| | | Anti2 (relaxed) | 1.7% |
| | | Anti3 (right-twisted) | 11.4% |
| Parallel | 4.5% | | |
| Turn | 13% | | |
| Others | 47.5% | | |

**Table 3. 3 Summary of secondary structure contributions used to fit the CD spectrum of Alr5209.**

**A**


Wavelength (nm)
180
190
200
210
220
230
240
250
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
CD (Delta epsilon)
Residuals
Experimental
Fitted

**B**


-9000
-10000
-11000
-12000
Molar Ellupticity (deg*cm2/dmol)
Refolding
Unfolding
10
30
50
70
90
Temperature (℃)

**C**

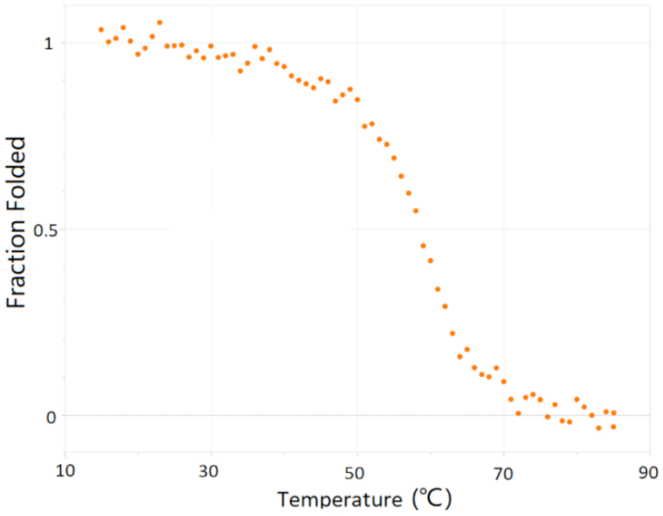

1
0.5
0
Fraction Folded
10
30
50
70
90
Temperature (℃)

**Figure 3. 6 CD spectrum and temperature melting experiments for Alr5209.** A) Wavelength scan for 20 □M protein at 25 ℃ depicted with buffer scan correction and fitted curve. The peak in ellipticity occurred at 210 nm. B) Graph of data points for the temperature melting experiments measured from 15 ℃ to 85 ℃ recorded at 210 nm. Increasing temperature from 15 ℃ to 85 ℃ resulted in unfolding of the protein and subsequent decreasing of the temperature from 85 ℃ to 15 ℃ allowed protein refolding. C) Graph of folded fraction as a function of temperature.

### 3.4.5 Insight into potential function of Alr5209 from gene cluster analysis

A gene cluster analysis based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (34) indicated that five genes, including *alr5209*, belong to the same operon. Of those genes (*alr5208, Alr5209, alr5210, alr5211* and *alr5212*), *alr5208, alr5209,* and *alr5212* were annotated as hypothetical proteins with unknown functions. Alr5210 was annotated as a two-component hybrid sensor and regulator but its function is still unknown and Alr5211 was recognized as a NADH dehydrogenase involved in oxidative phosphorylation based on analogy to the gene cluster with *slr0851, slr1743,* and *sll1484* in cyanobacterium *Synechocystis* sp strain PCC 6803. Therefore, Alr5209 may be involved in oxidative phosphorylation (34).

### 3.4.6 Re-examination of PRP domain consensus sequences

Pentapeptide repeat domains have been reported to have the approximate consensus sequence (S/T/A/V)(D/N)(L/F)(S/T/R)(X) (11,35). However, prior to solving the crystal structure of Alr5209, we were able to predict the location of its pentapeptide repeat domains using this consensus sequence. Once the structure of Alr5209 was determined it was possible to map the pentapeptide repeat domains onto the Alr5209 amino acid sequence (**Figure 3.1**). To facilitate reevaluation of the consensus sequences of pentapeptide repeat domains, a sequence Logo analysis was performed for all known PRP structures (**Figure 3.7**). The sequence logo analysis in **Figure 3.7** is organized into representations for seven type I plus type IV β turn PRPs, four pure type II β turns PRPs and Alr5209, which is a mixture of type I and type II β turns. Based on our structure-based sequence analysis of all currently known PRP structures, we recommend that the consensus sequence of PRPs should be amended to (A/C/S/V/T/L/I)/(D/N/S/K/E/I/R)/(L/F)/(S/T/R/E/Q/K/V/D)/(G/D/E/N/R/Q/K). The complete list of the frequency of occurrence of every amino acid at every position in the pentapeptide repeat domain positions is compiled in **Table 3.4**. The general consensus from this analysis indicates that any uncharged or small hydrophobic amino acid can be accommodated in the i-2 position, any charged or polar amino acid can be found in the i-1 position, the i position is mostly occupied by L or F, followed by I, M, W, but can be occupied by any strongly hydrophobic residue, including A, C, V, the i+1 positions can be occupied by any charged or polar amino acid, and the i+2 positions can be occupied by any charged or polar amino acid. These rules are consistent with the topology of the PRPs in that the side chains of the i-1 and i+1 amino acids always point away from the axis of the right-handed β-helix, which in a water-soluble PRP would position the hydrophilic and charged side chains towards the solvent environment. Likewise, the side chain of the i position amino acid strongly prefers L or F to establish the hydrophobic core of the protein, but can also accommodate the side chain of any other hydrophobic

amino acid. No charged amino acids have ever been observed in the i-2 position, but uncharged, polar, hydrophilic amino acids have been observed in the i-2 position.

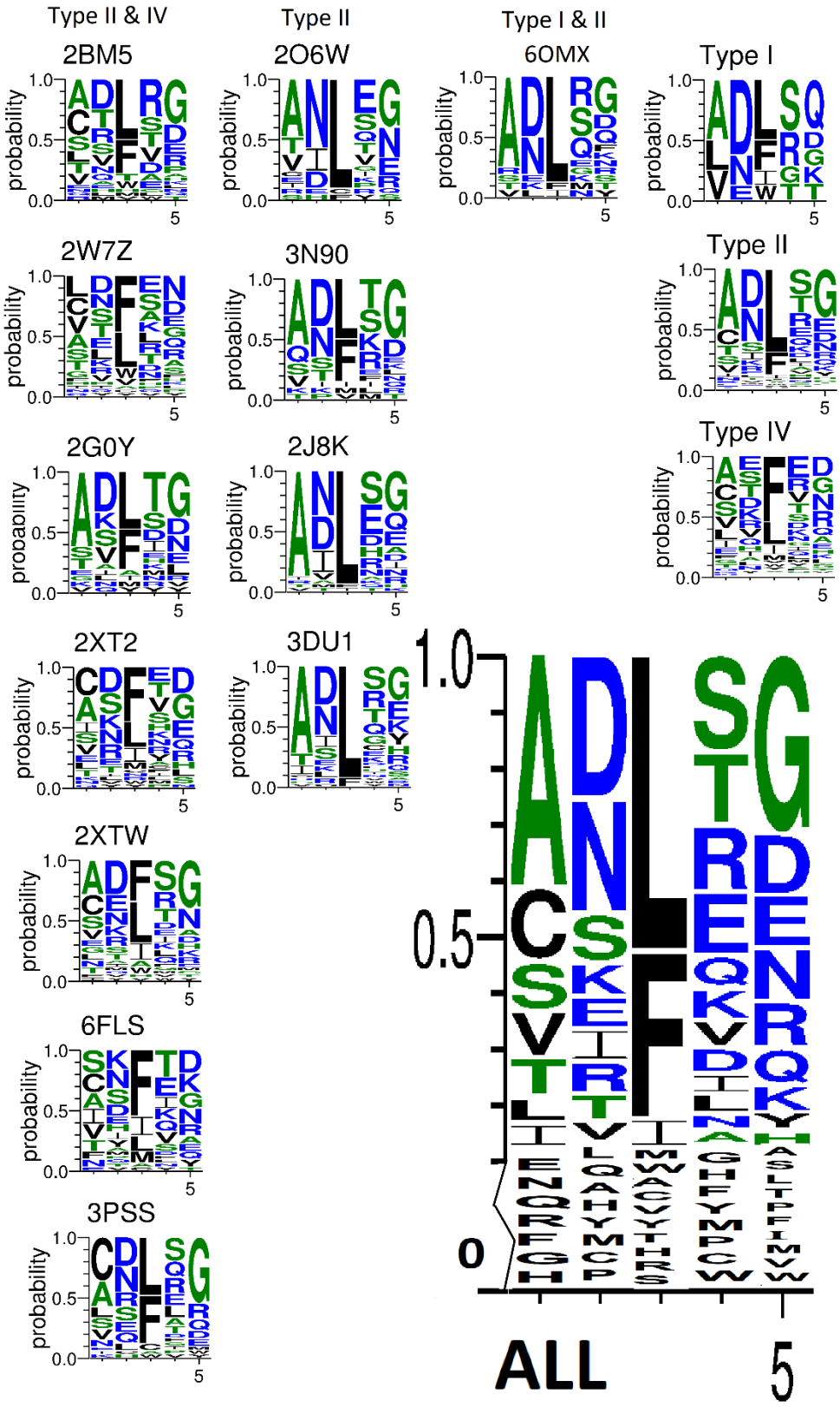

Type II & IV
2BM5
Type II
2O6W
Type I & II
6OMX
Type I
2W7Z
3N90
Type II
2G0Y
2J8K
Type IV
2XT2
3DU1
2XTW
6FLS
3PSS
probability
1.0
0.5
0.0
5
0
ALL

**Figure 3. 7 The sequence logo summary of all PRPs with known structures and alignments.** The codes above each graph are the PDB code for each protein. The large sequence logo plot at the lower right was calculated using the all sequence alignment for all the other individual PRPs included in the figure.

| 2J8K | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
| I-2 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| I-1 | 1 | 0 | 9 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 12 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I+1 | 1 | 0 | 3 | 7 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 3 | 9 | 1 | 0 | 0 | 0 |
| I+2 | 2 | 0 | 2 | 4 | 0 | 10 | 1 | 2 | 1 | 0 | 0 | 2 | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 |
| 2O6W | | | | | | | | | | | | | | | | | | | | |
| I-2 | 10 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 3 | 0 | 0 |
| I-1 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 4 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I+1 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 2 | 1 | 3 | 2 | 2 | 0 | 1 |
| I+2 | 0 | 0 | 0 | 3 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 |
| 2XT2 | | | | | | | | | | | | | | | | | | | | |
| I-2 | 7 | 8 | 0 | 2 | 0 | 0 | 0 | 3 | 1 | 2 | 0 | 1 | 0 | 1 | 0 | 3 | 2 | 3 | 0 | 0 |
| I-1 | 0 | 0 | 7 | 2 | 0 | 0 | 0 | 1 | 4 | 1 | 0 | 4 | 0 | 0 | 4 | 6 | 2 | 1 | 0 | 1 |
| I | 0 | 1 | 0 | 0 | 15 | 0 | 0 | 4 | 0 | 7 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| I+1 | 1 | 1 | 0 | 4 | 1 | 0 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 0 | 2 | 3 | 4 | 4 | 0 | 2 |
| I+2 | 0 | 0 | 7 | 5 | 0 | 7 | 2 | 0 | 0 | 2 | 0 | 1 | 0 | 3 | 3 | 2 | 0 | 0 | 0 | 0 |
| 2G0Y | | | | | | | | | | | | | | | | | | | | |
| I-2 | 13 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 0 |
| I-1 | 1 | 0 | 7 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 0 | 3 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 1 | 0 | 0 | 0 | 7 | 0 | 0 | 1 | 0 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| I+1 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 3 | 7 | 0 | 0 | 1 |
| I+2 | 0 | 0 | 3 | 2 | 0 | 8 | 0 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 3N90 | | | | | | | | | | | | | | | | | | | | |
| I-2 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 1 | 2 | 0 | 0 |
| I-1 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 5 | 1 | 0 | 0 | 2 | 2 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 1 | 0 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| I+1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 3 | 1 | 1 | 0 | 0 | 0 | 3 | 4 | 5 | 0 | 0 | 0 |
| I+2 | 0 | 0 | 3 | 1 | 0 | 10 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2BM5 | | | | | | | | | | | | | | | | | | | | |
| I-2 | 9 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 1 | 1 | 4 | 3 | 2 | 0 | 0 |
| I-1 | 1 | 0 | 8 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 2 | 5 | 3 | 5 | 3 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 9 | 0 | 1 | 0 | 0 | 16 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 0 |
| I+1 | 2 | 0 | 4 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 9 | 4 | 4 | 4 | 1 | 0 |
| I+2 | 1 | 0 | 5 | 3 | 0 | 12 | 1 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 3 | 0 | 0 | 1 | 0 | 0 |
| 3DU1 | | | | | | | | | | | | | | | | | | | | |
| I-2 | 29 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4 | 1 | 0 | 0 |
| I-1 | 0 | 1 | 13 | 2 | 0 | 0 | 0 | 4 | 2 | 2 | 0 | 10 | 0 | 0 | 2 | 4 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I+1 | 0 | 2 | 1 | 2 | 1 | 3 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 4 | 6 | 8 | 5 | 1 | 1 | 1 |
| I+2 | 1 | 0 | 1 | 6 | 0 | 11 | 3 | 0 | 4 | 0 | 0 | 1 | 0 | 2 | 3 | 2 | 0 | 0 | 0 | 4 |
| 2XTW | | | | | | | | | | | | | | | | | | | | |
| I-2 | 11 | 7 | 0 | 2 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 4 | 2 | 3 | 0 | 0 |
| I-1 | 1 | 0 | 10 | 5 | 0 | 0 | 0 | 1 | 3 | 0 | 1 | 4 | 0 | 2 | 3 | 2 | 2 | 1 | 0 | 1 |
| I | 2 | 0 | 0 | 0 | 14 | 0 | 0 | 5 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| I+1 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 2 | 5 | 9 | 4 | 1 | 1 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I+2 | 1 | 0 | 2 | 1 | 0 | 14 | 2 | 0 | 2 | 0 | 1 | 6 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 |
| 2W7Z | | | | | | | | | | | | | | | | | | | | |
| I-2 | 4 | 5 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 6 | 0 | 1 | 0 | 1 | 0 | 3 | 3 | 5 | 0 | 0 |
| I-1 | 0 | 0 | 5 | 3 | 0 | 0 | 1 | 1 | 2 | 3 | 0 | 4 | 0 | 1 | 2 | 4 | 4 | 2 | 0 | 1 |
| I | 1 | 1 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 1 |
| I+1 | 3 | 0 | 2 | 5 | 1 | 0 | 1 | 0 | 3 | 3 | 0 | 2 | 1 | 1 | 3 | 5 | 3 | 1 | 0 | 0 |
| I+2 | 2 | 0 | 4 | 3 | 1 | 3 | 1 | 0 | 1 | 1 | 0 | 7 | 1 | 3 | 3 | 2 | 0 | 0 | 0 | 1 |
| 6OMX | | | | | | | | | | | | | | | | | | | | |
| I-2 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| I-1 | 0 | 0 | 8 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I+1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 3 | 4 | 4 | 0 | 0 | 0 | 0 |
| I+2 | 0 | 0 | 2 | 0 | 1 | 5 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 1 | 1 | 1 | 0 | 0 | 1 |
| 6FLS | | | | | | | | | | | | | | | | | | | | |
| I-2 | 4 | 5 | 0 | 1 | 2 | 0 | 0 | 4 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 6 | 3 | 4 | 0 | 0 |
| I-1 | 1 | 0 | 3 | 2 | 0 | 0 | 2 | 2 | 5 | 0 | 1 | 5 | 1 | 1 | 0 | 4 | 1 | 1 | 0 | 2 |
| I | 1 | 1 | 0 | 0 | 17 | 0 | 0 | 5 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I+1 | 0 | 0 | 1 | 5 | 1 | 0 | 0 | 3 | 3 | 1 | 0 | 1 | 0 | 3 | 0 | 2 | 7 | 3 | 0 | 1 |
| I+2 | 2 | 0 | 6 | 2 | 0 | 4 | 0 | 0 | 5 | 0 | 0 | 4 | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 2 |
| 3PSS | | | | | | | | | | | | | | | | | | | | |
| I-2 | 7 | 12 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 3 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 3 | 0 | 0 |
| I-1 | 0 | 1 | 8 | 3 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 7 | 0 | 3 | 4 | 4 | 0 | 0 | 0 | 0 |
| I | 1 | 2 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| I+1 | 2 | 0 | 1 | 3 | 1 | 1 | 0 | 1 | 0 | 3 | 0 | 0 | 1 | 4 | 4 | 6 | 2 | 1 | 0 | 0 |
| I+2 | 0 | 0 | 2 | 2 | 0 | 17 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 1 | 1 |
| Type I | | | | | | | | | | | | | | | | | | | | |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I-2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| I-1 | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| I+1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 1 | 0 | 0 | 0 |
| I+2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| Type II | | | | | | | | | | | | | | | | | | | | |
| I-2 | 115 | 30 | 0 | 3 | 0 | 0 | 0 | 8 | 2 | 4 | 0 | 7 | 0 | 3 | 5 | 17 | 20 | 16 | 0 | 0 |
| I-1 | 2 | 1 | 74 | 6 | 0 | 0 | 1 | 15 | 11 | 6 | 3 | 65 | 1 | 3 | 11 | 19 | 5 | 5 | 0 | 2 |
| I | 4 | 5 | 0 | 0 | 43 | 0 | 0 | 8 | 0 | 159 | 3 | 0 | 0 | 0 | 1 | 0 | 1 | 3 | 4 | 0 |
| I+1 | 7 | 2 | 10 | 20 | 5 | 3 | 6 | 8 | 13 | 9 | 1 | 5 | 3 | 16 | 27 | 50 | 34 | 7 | 2 | 3 |
| I+2 | 3 | 0 | 19 | 27 | 2 | 94 | 6 | 2 | 11 | 3 | 1 | 18 | 1 | 13 | 14 | 5 | 1 | 0 | 1 | 10 |
| Type IV | | | | | | | | | | | | | | | | | | | | |
| I-2 | 25 | 15 | 0 | 6 | 4 | 4 | 1 | 7 | 2 | 10 | 0 | 2 | 0 | 3 | 0 | 14 | 4 | 11 | 0 | 0 |
| I-1 | 4 | 1 | 10 | 13 | 0 | 0 | 5 | 5 | 10 | 4 | 1 | 4 | 1 | 7 | 8 | 12 | 12 | 8 | 0 | 3 |
| I | 2 | 1 | 0 | 0 | 59 | 0 | 1 | 9 | 0 | 20 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 3 | 3 | 3 |
| I+1 | 2 | 1 | 6 | 19 | 2 | 4 | 2 | 4 | 6 | 3 | 3 | 5 | 1 | 4 | 11 | 7 | 9 | 10 | 1 | 3 |
| I+2 | 6 | 0 | 17 | 5 | 0 | 14 | 4 | 0 | 4 | 4 | 1 | 14 | 1 | 7 | 14 | 4 | 2 | 2 | 0 | 1 |
| All | | | | | | | | | | | | | | | | | | | | |
| I-2 | 144 | 46 | 0 | 9 | 4 | 4 | 1 | 15 | 4 | 16 | 0 | 9 | 0 | 6 | 5 | 31 | 24 | 29 | 0 | 0 |
| I-1 | 6 | 2 | 89 | 20 | 0 | 0 | 6 | 20 | 21 | 10 | 4 | 71 | 2 | 10 | 20 | 31 | 17 | 13 | 0 | 5 |
| I | 6 | 6 | 0 | 0 | 104 | 0 | 1 | 18 | 0 | 184 | 8 | 1 | 0 | 0 | 1 | 0 | 2 | 6 | 8 | 3 |
| I+1 | 9 | 3 | 16 | 39 | 7 | 8 | 8 | 12 | 19 | 12 | 4 | 10 | 4 | 20 | 41 | 60 | 44 | 17 | 3 | 6 |
| I+2 | 9 | 0 | 37 | 32 | 2 | 109 | 10 | 2 | 16 | 7 | 2 | 32 | 3 | 23 | 28 | 9 | 4 | 2 | 1 | 11 |

**Table 3. 1 Summary of amino acids distributions in all PRPs with known structures.**

### 3.4.7 Structural consequences of type I beta turns in PRPs

Based on analysis of all existing PRP structures in Protein Data Bank (18,36-40), Alr5209 is the first PRP that contains type I β turns in its Rfr fold. All other PRPs structures reported to date contain Rfr folds composed exclusively of type II β turns (206W (36), 3DU1 (10) , 2J8K (18), and 3N90 (12) ) or mixture of type II and IV β turns (2G0Y (41), 2W7Z (38), 2BM5 (30), 2XT2 (37), 2XTW (40), 6FLS (42), and 3PSS (39) ) (**Figure 3.7**). In order to determine if there was any visible consequence of including type I β turns in the Rfr fold, we compared all existing PRP structures looking along the right-handed β-helical structure (**Figure 3.8**). One pattern that is apparent is that PRPs composed of combinations of type II and IV β turns experience a significant negative twist in the relative position of the quadrilateral coils along the N-terminal to C-terminal direction (**Figure 3.8**). PRPs comprised exclusively of type II β turns appear to also contain twist, but the magnitude of the negative twist is significantly smaller compared to PRPs containing both type II and type IV β turns (**Figure 3.8**). Finally, Alr5209, composed of type I and type II β turns exhibits the least helical twist among known PRPs (**Figure 8**). The twist angles for all PRPs are summarized in **Table 3.5**. Based on this analysis, increased magnitude of helical twist appears to be correlated with the presence of loops inserted into the pentapeptide repeat domain sequence, this being especially obvious among the combined type II and type IV β turn PRPs. However, when the twist magnitude was averaged on a per coil basis, the type II plus type IV β turn PRPs still had a significantly larger twist per coil magnitude (**Figure 3.9A**), suggesting that a fundamental difference in the turn structure was responsible for introducing twist in the Rfr fold. To better understand the origin of increased negative twist in PRPs containing type IV β turns, the distances across each type of β turn were measured (**Figure 3.9B**). These measurements indicated that the distance across type I β turns was the shortest at ~5.6 Å, compared to ~5.7-5.8 Å for type II β turns, however, the distance across type IV β turns was substantially longer at ~6.4 Å. Consequently, a negative helical twist is required to accommodate the extended β turn distance in comparison to the type I and type II β turn distances. Another consequence of the extended type IV β turns is a general increase in the area spanned by the individual quadrilateral coils (**Table 3.6**). This is evident both in the distance between the opposite faces of the quadrilateral β-helix, which increased by as much as 1 Å in going from type I plus type II β turn PRPs to type II plus type IV β turn PRPs (**Table 3.6**), and in the diagonal distances across the individual coils, which increased by about 1 Å in each direction (**Table 3.6**). Consequently, in this first example of a PRP comprised of both type I plus II β turns, the PRP solenoid is smaller and more compact with less negative helical twist compared to PRP structures made up exclusively of type II β turns and significantly smaller and more compact compared to PRP structures containing both type II and IV β turns, which, in general, have the largest Rfr folds.

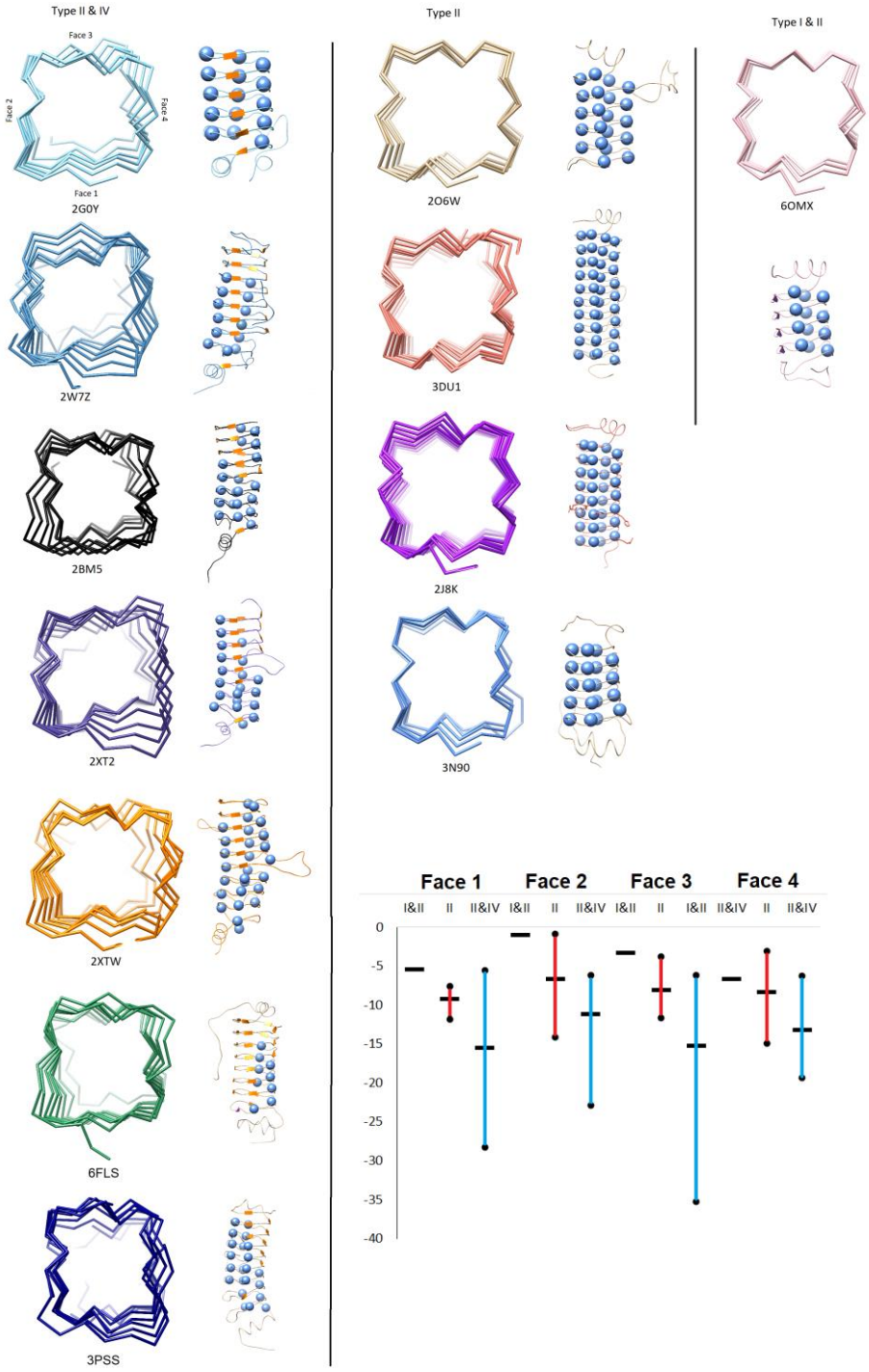



**Figure 3. 8 Backbone traces for all PRPs with known structures and sequence alignments.** The PDB code is indicated below each structure. The first column shows all PRPs containing mixtures of type II and IV β turns with the turn distribution of turns indicated at the right where spheres indicate type II turns and orange sheets indicate type IV turns. The second column shows all PRPs made up exclusively of type II β turns. The last column shows Alr5209 which is the first example of a PRP that contains a mixture of

type I and II β turns with the turn distribution indicated at the right where spheres indicate type II turns and purple sheets indicate type I turns. All structures are depicted with the N terminus facing the reader and the faces oriented the same as with 2G0Y with face 1 at the bottom, face 2 at the left, face3 at the top, and face 4 at the right. The inserted graph shows the average (bar) and range of twist angles of the three classes of PRPs based on their composition of β turns: type I plus type II, pure type II, or type II plus type IV as listed in **Table 3.5**.

| β-Turn | PDB Code | Face 1(°) | Face 2(°) | Face 3(°) | Face 4(°) |
|---|---|---|---|---|---|
| Type I & II | 6OMX | -5.5±1.8[4] | -1.3±1.6 | -3.3±2.1 | -6.7±2.9 |
| Type II | 3N90 | -7.6±2.7 | -0.9±0.5 | -11.7±5.5 | -3.1±3.7 |
| | 2J8K | -8.8±3.5[4] | -6.8±3.3 | -8.9±3.6 | -8.3±3.5 |
| | 2O6W | -8.7±2.1 | -4.7±3.1 | -3.8±1.3 | -7.0±2.8 |
| | 3DU1 | -11.8±6.2[4] | -14.1±6.4 | -7.8±7.3 | -15.0±7.8 |
| Type II & IV | 2XTW | -16.4±6.5 | -12.4±6.9 | -35.3±12.5 | -13.0±8.4 |
| | 2W7Z | -28.2±12.9 | -6.2±5.8 | -6.2±5.8 | -6.3±2.2 |
| | 2XT2 | -26.4±10.0 | -2.8±6.1 | -21.5±11.9 | -16.1±5.2 |
| | 2G0Y | -8.8±2.5 | -10.1±5.7 | -9.3±2.4 | -10.9±4.7 |
| | 2BM5 | -12.2±5.2 | -15.3±6.4 | -11.5±5.1 | -16.8±7.7 |
| | 6FLS | -5.6±3.1 | -22.8±12.2 | -13.1±5.6 | -19.3±6.4 |
| | 3PSS | -10.8±5.6[4] | -8.2±5.0 | -9.9±6.0 | -9.9±3.5 |

**Table 3. 4 Summary of twist angles between coils for all PRPs with known structures.[3]**

A

| Name | Type of turn | Distance (Å) |
|---|---|---|
| Alr5209 | I | 5.65±0.03 |
| | II | 5.82±0.23 |
| Rfr23 | II | 5.80±0.27 |
| Rfr32 | II | 5.71±0.23 |
| | IV | 6.44±0.29 |

B

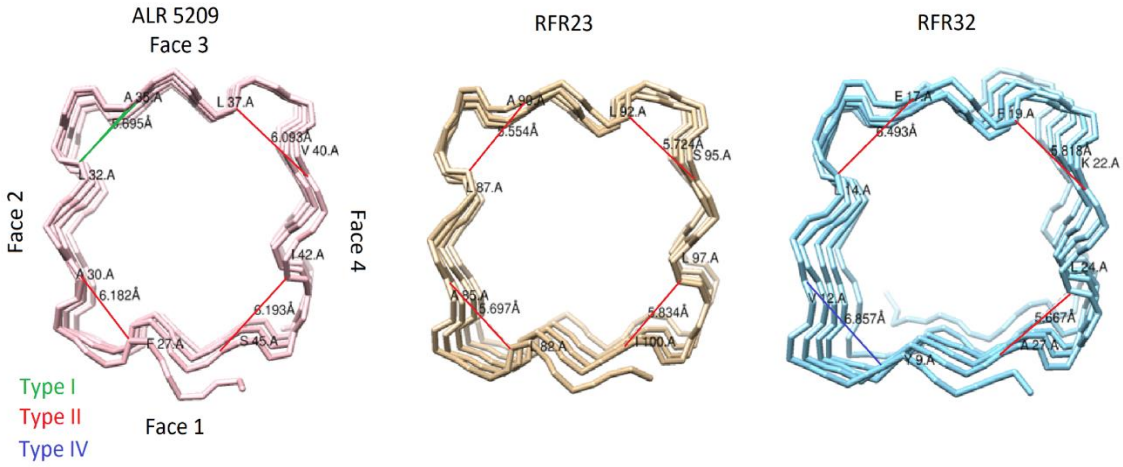



**Figure 3. 9 Graphs showing cross-turn distances for different types of turns and the summary of distance between carbon in i (i) and i-2 (i+1) position based on different types of turn.** A) average distances across type I, type II and type IV β turns measured in three representative PRPs. B) Left) Distances measured across the β turns in Alr5209 (PDB ID 6OMX). Middle) Distances measured across the β turns in Rfr23 (PDB ID 2O6W). Right) Distances measured across the β turns in Rfr32 (PDB ID 2G0Y).

[3] Twist angles are defined between first coil and following coils. Negative values negative twists and positive values indicate positive twists.

[4] Angles compared starting from the second coil rather than the first coil.

| β-Turn | PDB Code | Face 1(Å) | Face 2(Å) | Face 3(Å) | Face 4(Å) | Face 1-3(Å) | Face 2-4(Å) | Face 1-2(Å) | Face 1-4(Å) |
|---|---|---|---|---|---|---|---|---|---|
| Type I & II | 6OMX | 10.88±0.02 | 10.41±0.14 | 10.82±0.14 | 11.09±0.15 | 15.16±0.25 | 14.41±0.25 | 10.47±0.36 | 10.95±0.13 |
| Type II | 3N90 | 11.12±0.18 | 11.02±0.22 | 10.95±0.18 | 11.68±1.11 | 15.43±0.34 | 14.89±0.82 | 10.82±0.21 | 11.00±0.39 |
| | 2J8K | 11.05±0.12 | 10.99±0.12 | 10.92±0.13 | 10.95±0.18 | 14.77±0.51 | 14.99±0.31 | 10.51±0.25 | 11.05±0.13 |
| | 2O6W | 11.28±0.37 | 11.06±0.12 | 11.24±0.18 | 10.88±0.27 | 14.93±0.38 | 15.51±0.43 | 10.52±0.13 | 11.82±0.51 |
| | 3DU1 | 11.10±0.12 | 11.09±0.11 | 11.07±0.22 | 11.05±0.25 | 15.29±0.50 | 15.02±0.48 | 10.52±0.23 | 11.40±0.21 |
| | Average | 11.13±0.22 | 11.04±0.15 | 11.04±0.22 | 11.10±0.56 | 15.10±0.53 | 15.08±0.55 | 10.57±0.24 | 11.33±0.43 |
| Type II & IV | 2XTW | 12.08±0.53 | 11.44±0.55 | 10.57±1.27 | 10.95±0.77 | 15.22±0.54 | 11.08±1.40 | 11.08±0.87 | 12.09±1.46 |
| | 2W7Z | 11.84±0.96 | 11.01±0.39 | 11.40±0.14 | 11.77±0.76 | 16.01±2.38 | 15.86±1.27 | 11.20±0.98 | 13.39±0.88 |
| | 2XT2 | 11.79±0.74 | 11.18±0.29 | 10.92±1.59 | 11.46±0.72 | 15.71±1.22 | 16.30±0.50 | 10.74±0.62 | 12.60±0.64 |
| | 2G0Y | 12.07±0.29 | 11.24±0.26 | 11.05±0.18 | 10.99±0.17 | 15.71±0.23 | 15.45±0.32 | 11.04±0.13 | 12.04±0.49 |
| | 2BM5 | 11.63±0.57 | 11.54±0.50 | 11.26±0.49 | 10.98±0.34 | 16.59±0.91 | 15.14±0.66 | 11.54±0.92 | 11.87±0.57 |
| | 6FLS | 12.05±0.42 | 12.26±0.86 | 11.83±0.47 | 11.37±0.19 | 17.00±1.37 | 16.69±0.51 | 12.68±0.79 | 12.37±0.22 |
| | 3PSS | 11.35±0.57 | 10.70±1.38 | 10.95±0.37 | 11.90±0.60 | 16.01±1.51 | 15.19±0.62 | 11.35±0.57 | 12.14±0.71 |
| | Average | 11.82±0.68 | 11.35±0.85 | 11.14±0.95 | 11.37±0.69 | 16.05±1.49 | 15.83±1.49 | 11.40±0.98 | 12.35±0.95 |

**Table 3. 5 Summary of distances between and across faces of all PRPs with known structures and sequence alignments.[5]**

[5] The distances of face 1,2,3,4 were measured from the carbonyl carbon of the first amino acids to that of the last amino acid. The distances between faces 1 and 3 were measured from the carbonyl carbon in face 1 i position to that in face 3 i position. The distances of face 2-4 are measured from the carbonyl carbon in face 2 i position to that in face 4 i position. The distances of face 1-2 are measured from the carbonyl carbon in face 1 i position to that in face 2 i position. The distances of face 1-4 are measured from the carbonyl carbon in face 1 i position to that in face 4 i position.

## 3.5 Conclusion

Alr5209 from *Nostoc* sp. PCC 7120 represents the first PRP structure that includes type I β turns in its Rfr fold. A combined analysis of its sequence and structure allowed us to investigate how type I β turns, along with type II and type IV β turns can be accommodated into Rfr folds, to characterize the consequences that the occurrence of type I β turns has on the right-handed β-helical coil structure, and to significantly expand our understanding of the consensus sequence observed in pentapeptide repeat protein domains. The thermal titration measurements obtained from CD experiments added to our understanding of how the relative thermal stability PRPs depends on the number of coils comprising the Rfr fold. While an understanding of the biochemical function of Alr5209 remains unknown, genomic analysis indicated that it may play a role in oxidative phosphorylation, however confirmation of such a role will require further examination.

**3.6 Acknowledgements**
This work was supported by Miami University. We acknowledge access to the x-ray beamline provided at the Advanced Photon Source (APS) at Argonne National Laboratory. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. We also acknowledge molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

**3.7 References**

1. (2012) *Ecology of Cyanobacteria II: Their Diversity in Space and Time*, Springer Science & Business Media, London
2. Giovannoni, S. J., Turner, S., Olsen, G. J., Barns, S., Lane, D. J., and Pace, N. R. (1988) Evolutionary Relationships among Cyanobacteria and Green Chloroplasts. *J Bacteriol* **170**, 3584-3592
3. Black, K., Buikema, W. J., and Haselkorn, R. (1995) The Hglk Gene Is Required for Localization of Heterocyst-Specific Glycolipids in the Cyanobacterium *Anabaena* Sp Strain Pcc-7120. *J Bacteriol* **177**, 6440-6448
4. Herrero, A., Stavans, J., and Flores, E. (2016) The multicellular nature of filamentous heterocyst-forming cyanobacteria. *FEMS Microbiol Rev* **40**, 831-854
5. Golden, J. W., and Yoon, H. S. (2003) Heterocyst development in *Anabaena*. *Curr Opin Microbiol* **6**, 557-563
6. Wang, L., Sun Yp Fau - Chen, W.-L., Chen Wl Fau - Li, J.-H., Li Jh Fau - Zhang, C.-C., and Zhang, C. C. Genomic analysis of protein kinases, protein phosphatases and two-component regulatory systems of the cyanobacterium *Anabaena* sp. strain PCC 7120.
7. Hamilton, T. L., Bryant, D. A., and Macalady, J. L. (2016) The role of biology in planetary evolution: cyanobacterial primary production in low-oxygen Proterozoic oceans. *Environ Microbiol* **18**, 325-340
8. Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M., and Tabata, S. (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**, 205-213; 227-253
9. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432
10. Ni, S. S., Sheldrick, G. M., Benning, M. M., and Kennedy, M. A. (2009) The 2 angstrom resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium *Nostoc* sp strain PCC 7120. *J Struct Biol* **165**, 47-52

11. Bateman, A., Murzin, A. G., and Teichmann, S. A. (1998) Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci* **7**, 1477-1480
12. Ni, S., McGookey, M. E., Tinch, S. L., Jones, A. N., Jayaraman, S., Tong, L., and Kennedy, M. A. (2011) The 1.7 A resolution structure of At2g44920, a pentapeptide-repeat protein in the thylakoid lumen of *Arabidopsis thaliana*. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **67**, 1480-1484
13. Diao, J., Zhang, Y., Huibregtse, J. M., Zhou, D., and Chen, J. (2008) Crystal structure of SopA, a *Salmonella* effector protein mimicking a eukaryotic ubiquitin ligase. *Nat Struct Mol Biol* **15**, 65-70
14. Lin, D. Y., Diao, J., Zhou, D., and Chen, J. (2011) Biochemical and structural studies of a HECT-like ubiquitin ligase from *Escherichia coli* O157:H7. *J Biol Chem* **286**, 441-449
15. Yao, G. R., Zhang, S. C., Mahrhold, S., Lam, K. H., Stern, D., Bagramyan, K., Perry, K., Kalkum, M., Rummel, A., Dong, M., and Jin, R. S. (2016) N-linked glycosylation of SV2 is required for binding and uptake of *botulinum* neurotoxin A. *Nature Structural & Molecular Biology* **23**, 656-662
16. Benoit, R. M., Frey, D., Hilbert, M., Kevenaar, J. T., Wieser, M. M., Stirnimann, C. U., McMillan, D., Ceska, T., Lebon, F., Jaussi, R., Steinmetz, M. O., Schertler, G. F. X., Hoogenraad, C. C., Capitani, G., and Kammerer, R. A. (2014) Structural basis for recognition of synaptic vesicle protein 2C by *botulinum* neurotoxin A. *Nature* **505**, 108-+
17. Long, F., Vagin, A. A., Young, P., and Murshudov, G. N. (2008) BALBES: a molecular-replacement pipeline. *Acta Crystallogr D Biol Crystallogr* **64**, 125-132
18. Vetting, M. W., Hegde, S. S., Hazleton, K. Z., and Blanchard, J. S. (2007) Structural characterization of the fusion of two pentapeptide repeat proteins, Np275 and Np276, from *Nostoc punctiforme*: resurrection of an ancestral protein. *Protein Sci* **16**, 755-760
19. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501
20. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221
21. Dolinsky, T. J., Nielsen Je Fau - McCammon, J. A., McCammon Ja Fau - Baker, N. A., and Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations.
22. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612
23. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190
24. Schneider, T. D., and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-6100

25. Zwart, P. H., Grosse-Kunstleve, R. W., and Adams, P. D. (2005) Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsletter* **43**, contribution 7
26. Zwart, P. H., Grosse-Kunstleve, R. W., and Adams, P. D. (2005) Characterization of X-ray data sets *CCP4 Newsletter* **42**, contribution 10
27. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21
28. Shapovalov, M., Vucetic, S., and Dunbrack, R. L., Jr. (2019) A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Comput Biol* **15**, e1006844
29. Hutchinson, E. G., and Thornton, J. M. (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci* **3**, 2207-2216
30. Hegde, S. S., Vetting, M. W., Roderick, S. L., Mitchenall, L. A., Maxwell, A., Takiff, H. E., and Blanchard, J. S. (2005) A fluoroquinolone resistance protein from *Mycobacterium tuberculosis* that mimics DNA. *Science* **308**, 1480-1483
31. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research* **32**, D120-D121
32. Xu, S., Ni, S., and Kennedy, M. A. (2017) NMR Analysis of Amide Hydrogen Exchange Rates in a Pentapeptide-Repeat Protein from *A. thaliana*. *Biophys J* **112**, 2075-2088
33. Greenfield, N. J. (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat Protoc* **1**, 2527-2535
34. Howitt, C. A., Udall, P. K., and Vermaas, W. F. J. (1999) Type 2 NADH dehydrogenases in the cyanobacterium *Synechocystis* sp strain PCC 6803 are involved in regulation rather than respiration. *J Bacteriol* **181**, 3994-4003
35. Vetting, M. W., Hegde, S. S., Fajardo, J. E., Fiser, A., Roderick, S. L., Takiff, H. E., and Blanchard, J. S. (2006) Pentapeptide repeat proteins. *Biochemistry* **45**, 1-10
36. Buchko, G. W., H., R., Pakrasi, H. B., and Kennedy, M. A. (2007) Insights into the structural variation between pentapeptide repeat proteins--crystal structure of Rfr23 from *Cyanothece* 51142. *J. Struct. Biol.* **162**, 184-192
37. Vetting, M. W., Hegde, S. S., Zhang, Y., and Blanchard, J. S. (2011) Pentapeptide-repeat proteins that act as topoisomerase poison resistance factors have a common dimer interface. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **67**, 296-302
38. Hegde, S. S., Vetting, M. W., Mitchenall, L. A., Maxwell, A., and Blanchard, J. S. (2011) Structural and biochemical analysis of the pentapeptide repeat protein EfsQnr, a potent DNA gyrase inhibitor. *Antimicrob Agents Chemother* **55**, 110-117
39. Xiong, X., Bromley Eh Fau - Oelschlaeger, P., Oelschlaeger P Fau - Woolfson, D. N., Woolfson Dn Fau - Spencer, J., and Spencer, J. Structural insights into quinolone antibiotic resistance mediated by pentapeptide repeat proteins:

conserved surface loops direct the activity of a Qnr protein from a gram-negative bacterium.
40. Vetting, M. W., Hegde, S. S., Wang, M., Jacoby, G. A., Hooper, D. C., and Blanchard, J. S. (2011) Structure of QnrB1, a plasmid-mediated fluoroquinolone resistance factor. *J Biol Chem* **286**, 25265-25273
41. Buchko, G. W., S., N., H., R., Welsh, E. A., Pakrasi, H. B., and Kennedy, M. A. (2006) Characterization of two potentially universal turn motifs that shape the repeated five-residues fold--crystal structure of a lumenal pentapeptide repeat protein from *Cyanothece* 51142. *Protein Science* **15**, 2579-2595
42. Notari, L., Martinez-Carranza, M., Farias-Rico, J. A., Stenmark, P., and von Heijne, G. (2018) Cotranslational Folding of a Pentarepeat beta-Helix Protein. *J Mol Biol* **430**, 5196-5206

# Chapter 4: Crystal structure of Alr1298, a pentapeptide repeat protein from the cyanobacterium Nostoc sp. PCC 7120, determined at 2.1 Å resolution

Reproduced with permission from:

Ruojing Zhang[1], Shuisong Ni[1], Michael A. Kennedy*[1]

[1]Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056

*Corresponding Author: Department of Chemistry and Biochemistry, 106 Hughes Laboratories, Miami University, 651 East High Street, Oxford, OH 45056. Email: kennedm4@miamioh.edu. Phone: 513-529-8267. Fax: 513-529-5715.

This paper has been published in Proteins. 2020; 88:1143–1153



Author contributions: RZ contributed to plasmid preparation, protein expression, data collection, and data analysis, manuscript preparation. SN and MAK contributed to data analysis and manuscript preparation.

## 4.1 Abstract

*Nostoc* sp. PCC 7120 are filamentous cyanobacteria capable of both oxygenic photosynthesis and nitrogen fixation, with the latter taking place in specialized cells known as heterocysts that terminally differentiate from vegetative cells under conditions of nitrogen starvation. Cyanobacteria have existed on earth for more than two billion years and are thought to be responsible for oxygenation of the earth's atmosphere. Filamentous cyanobacteria such as *Nostoc* sp. PCC 7120 may also represent the oldest multicellular organisms on earth that undergo cell differentiation. Pentapeptide repeat proteins (PRPs), which occur most abundantly in cyanobacteria, adopt a right-handed quadrilateral β-helical structure, also referred to as a repeat five residue (Rfr) fold, with four-consecutive pentapeptide repeats constituting a single coil in the β-helical structure. PRPs are predicted to exist in all compartments within cyanobacteria including the thylakoid and cell-wall membranes as well as the cytoplasm and thylakoid periplasmic space. Despite their intriguing structure and importance to understanding ancient cyanobacteria, the biochemical function of PRPs in cyanobacteria remains largely unknown. Here we report the crystal structure of Alr1298, a PRP from *Nostoc* sp. PCC 7120 predicted to reside in the cytoplasm. The structure displays the typical right-handed quadrilateral β-helical structure and includes a four-α-helix cluster capping the N-terminus and a single α□helix capping the C-terminus. A gene cluster analysis indicated that Alr1298 may belong to an operon linked to cell proliferation and/or thylakoid biogenesis. Elevated *alr1298* gene expression following nitrogen starvation indicates that Alr1298 may play a role in response to nitrogen starvation and/or heterocyst differentiation.

## 4.2 Introduction

Cyanobacteria are ancient oxygenic photosynthetic prokaryotic microorganisms thought to be responsible for oxygenation of the earth's atmosphere more than 2.3 billion years ago.[1, 2] Cyanobacteria also have the ability to fix atmospheric nitrogen gas during periods of nitrogen starvation in their growth environment into soluble nitrogen-containing compounds to support metabolic requirements.[1-5] The process of nitrogen fixation in cyanobacteria depends on the activity of oxygen-sensitive nitrogenase enzymes,[6, 7] which presents a problem since the process of photosynthesis in cyanobacteria generates oxygen. Unicellular and filamentous cyanobacteria have evolved different strategies to overcome this problem with unicellular cyanobacterial species separating oxygenic photosynthesis and nitrogen fixation temporally during light and dark cycles[7] and filamentous cyanobacterial species, such as *Nostoc* sp. PCC 7120, carrying out nitrogen fixation in specialized heterocysts, which are terminally differentiated from vegetative cells, under conditions of nitrogen starvation.[7-12] In *Nostoc* sp. PCC 7120, heterocysts occur at about every 10 to 20 vegetative cells reaching a final occupancy of around 7% of all cells in the filaments.[11, 13] Heterocyst differentiation is initiated in response the vegetative cells sensing reduced levels of 2-oxoglutarate(2-OG), which serves as a signal of nitrogen limitation for synthesis of ammonia by the sequential action of glutamine synthetase and glutamate synthase.[13-18] The expression of NtcA, a transcription factor responsible for activation and repression of a number of genes involved in nitrogen metabolism in a cyanobacteria, is influenced by the intracellular levels of 2-OG.[16] Elevated expression of NtcA in combination with initiation of expression of HetR initiates heterocyst

differentiation in filamentous cyanobacteria.[13, 19] HetL has also been implicated in regulation of heterocyst differentiation.[4, 20] Although many genes related to nitrogen fixation in cyanobacteria have been identified,[21-24] the genetic control and biological processes that regulate and control heterocyst differentiation remain incompletely understood.

Pentapeptide repeat proteins (PRPs) represent a large superfamily of proteins with about 38,000 PRP sequences identified in nearly 3500 species in the Pfam database.[25-34] PRPs are recognized as containing at least eight tandem pentapeptide repeat sequences, originally defined with the consensus A[D/N]LXX.[35] The PRPs adopt a classic right-handed quadrilateral β-helical structure, also referred to as a repeat five residue (Rfr) fold.[36] Based on expanded availability of both sequence databases and new PRP crystal structures, we recently updated the pentapeptide repeat sequence consensus sequence as (A/C/S/V/T/L/I)/(D/N/S/K/E/I/R)/(L/F)/(S/T/R/E/Q/K/V/D)/(G/D/E/N/R/Q/K).[37] As of 2019, only seventeen crystal structures of proteins containing PRPs have been reported, including Alr1298 in this manuscript. Despite the large size of the PRP superfamily and their abundance in the genomes of ancient photosynthetic bacteria that played a critical role in oxygenation of the earth's atmosphere, the function of PRPs remain largely unknown.

In this study, we report the crystal structure of Alr1298 determined at 2.1 Å resolution. The protein exhibits the classic right-handed quadrilateral β helical structure containing three and three-quarters complete coils. In comparison to all other known structures of PRPs, Alr1298 is unique in that it contains a four-helix cluster at its N-terminus. In contrast, five of the twelve other experimentally-determined PRP structures contain just a single N-terminal α-helix and seven of the 12 experimentally-determined structures contain no secondary structural elements at the N-terminus. The N-terminal four-α-helical cluster played a key role in the structural organization of the two molecules in the crystallographic asymmetric unit in which two Alr1298 molecules pack in a head-to-head interaction between the four-α-helix cluster of each molecule, with the helix axes making an angle of nearly 90° but being slightly obtuse. Despite the head-to-head dimer formation observed in the asymmetric unit, the protein was found to behave as a monomer in solution based on solution-state nuclear magnetic resonance spectroscopy-based measurement of the rotational correlation time. Consideration of the putative gene-cluster operon to which alr1298 belongs, and based on a genome-wide analysis of gene expression patterns in *Nostoc* sp PCC 7120 following nitrogen starvation, Alr1298 may be involved in sensing nitrogen starvation, responding to nitrogen starvation and/or heterocyst differentiation.

## 4.3 Materials and Method

### 4.3.1 Cloning, mutation, expression, purification

The genomic DNA of Nostoc sp. PCC 7120 (ATCC 27893) was used to amplify the native alr1298 gene using the following primer sequences: cccgcccgcatATGATCATGATCAATCCTCATACTC and gcccgctcgagttaATTATCATCTGCCAAATAGTTG. Standard PCR protocols were used to amplify the alr1298 gene. The gene encodes a 167 amino acid protein with a predicted

molecular weight of 18763.36 Da. The PCR product and expression vector pET28b were prepared for insertion of the alr1298 gene by digestion with the NdeI and XhoI restriction enzymes, which produced sticky ends as required for incorporation of the gene into the pET28b plasmid followed by ligation using T4 ligase. The constructed expression plasmid was transformed into JM109 for sequencing, which was confirmed. Next, the expression plasmid was transferred to E. coli BL21(DE3) for expression of the native protein and for expression of the mutant proteins.

Following analysis of the alr1298 sequence, L89M and L124M were selected as mutation sites to introduce methionine residues to facilitate preparation of SeMet labeled proteins required to permit phasing of the crystallographic diffraction data. Two sets of mutation primers were used to sequentially introduce site-directed mutations into the native alr1298 sequence in the expression vector (L89M: ggatgaagtaagtttaattcgaggtaatATGtcagaagcaaatttacaaggaag and cttccttgtaaatttgcttctgacatattacctcgaattaaacttacttcatcc; L124M: gctgatttaagaggtgcaactATGaatggaactgtttggctag and ctagccaaacagttccattcatagttgcacctcttaaatcagc) by using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies). The sequences were confirmed and the plasmid containing the mutated gene was transformed into E. coli BL21(DE3) for expression of SeMet labelled protein.

Alr1298 protein was overexpressed using an E. coli bacterial culture grown at 37 °C in LB medium for the native Alr1298 protein and using M9 medium containing SeMet to allow incorporation of SeMet into the mutated Alr1298 protein. The cell cultures were allowed to grow until the OD600 reached 0.6-0.8, then the cells were cooled to 15 °C followed by addition isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM for overnight overexpression. The bacterial cells were collected by centrifugation with 5000xg at 4°C. Cell pellets were resuspended in 20 mL of buffer (20 mM Tris, 250 mM NaCl, 10% glycerol, pH 7.8) and lysed using a French press. A Ni-NTA affinity column was used to purify the proteins by Ni-affinity chromatography. Imidazole present in the solution of purified protein following elution off of the Ni-NTA affinity column was removed by dialysis in 3 L of buffer (20 mM Tris, 250 mM NaCl, 10% glycerol, pH 7.8). The two proteins were analyzed by SDS-PAGE to confirm purity and then concentrated to around 8 mg/ml for crystallization trials.

The same procedure was applied to prepare the native protein sample including incorporation of $^{15}$N-stable isotope label for NMR correlation time experiments. For $^{15}$N labeling, M9 medium was used for cell culture using $^{15}$N-labeled ammonium chloride to facilitate incorporation of $^{15}$N label into the native protein.

### 4.3.2 Crystallization, phasing and refinement

The hanging-drop vapor-diffusion method was used to screen the crystallization conditions for the native Alr1298 and mutated Alr1298. Buffer number ten from the Hampton Research Kit (HR2-110), which contained 0.2 M ammonium acetate, 0.1 M sodium acetate trihydrate pH 4.6, and 30% w/v polyethylene glycol 4,000, was used to grow crystals used for X-ray diffraction experiments. Crystals started to form in about 48 hours. Cuboidal crystals aggregated in clusters from which single crystals were separated

using the loop used for picking crystals. Beamline 31-ID at the Advanced Photon Source (APS) at Argonne National Laboratory was used to collect both native data sets and for collection of single-wavelength anomalous dispersion (SAD) data on the SeMet-labeled protein as required for SAD phasing of the data at 100 K. The experimental SAD data collected using the SeMet-labeled, mutated Alr1298 collected at resolution 2.3 Å was used to solve the structure using AutoSol in PHENIX 1.13_2998.[38] The structure coordinates obtained for the SeMet-labeled, mutated Alr1298 were used to solve the structure of the native protein using the truncated diffraction data, i.e. structure factors obtained from the reflection intensities, collected from the native Alr1298 protein crystal at resolution 2.1 Å by molecular replacement in PHENIX. All structure refinements and structure validation information were generated within the PHENIX platform. The final structure coordinates and structure factors were submitted to PDB (PDB code of native Alr1298: 6UVI; PDB code of mutated Alr1298: 6UV7). The analysis of both structures, including electrostatic potential surface (combined with PDB2PQR server) and secondary structure calculation, were conducted using Chimera.[39, 40]

### 4.3.3 Circular dichroism (CD) spectroscopy and thermal protein denaturation

The native Alr1298 was diluted to 40 μM in 20 mM potassium phosphate pH 7.8 and 150 mM NaF buffer for CD measurements. The sample was transferred to a 1 mL quartz cuvette and an AVIV model435 CD spectrophotometer (Aviv Biomedical, Inc) was used to collect CD scans and to conduct the CD temperature melting experiments. The far-UV wavelength CD spectra were collected from 180 nm to 300 nm at 25°C and analyzed using the BeStSel (Beta Structure Selection) software package.[41] The temperature melting experiment was performed from 15 °C and 90 °C with step as 1 °C. Relating data was recorded at 221 nm, which had the strongest signals in wavelength scan spectra. Analysis of thermal parameters and fitted curve construction were carried out using the Calfitter 3.1 software package[42] using the natured-state equilibrium with denatured-state (N=D) model.

### 4.3.4 Nuclear magnetic resonance (NMR) correlation time determination

The $^{15}$N-labeled Alr1298 sample was prepared as described above and concentrated to 0.5 mM for all relaxation time measurements. All NMR data was collected at 298 K. All NMR data were collected at 850-MHz using an Avance III NMR spectrometer (Bruker). For the acquisition of $^{15}$N $T_1$ and $T_2$ relaxation data, the hsqct1etf3gpsi3d and hsqct2etf3gpsi3d implemented in TopSpin 3.6.1 were used. Ten inversion recovery delay times used for determination of the $T_1$ relaxation constant were set at 100, 200, 300, 400, 600, 800, 1000, 1500, 1700 and 2200 ms. For determination of the $T_2$ relaxation constant, the CPMG pulse sequence was used with ten loop count values set at 1, 2, 3, 4, 5, 7, 10, 12, 15 and 20 where the length of a single loop was 19.69 ms.

## 4.4 Results and discussion

### 4.4.1 Crystal and structure data quality

Crystals in the crystallization buffer grew as clusters of cuboids for both native and SeMet-labeled Alr1298. Single crystals of suitable size were separated from the clusters using the cryoloop used to pick the crystals. All crystallographic data collection and refinement statistics are included in **Table 4.1**. The native and SeMet data were collected

and truncated at 2.1 and 2.3 Å, respectively, with corresponding completeness of 98.03% and 99.60%, respectively. Diffraction data from both samples belonged to orthorhombic crystal class (native: a = 45.762 Å, b=85.922 Å, c=95.686 Å, $\alpha = \beta = \gamma = 90°$; SeMet: a = 45.842 Å, b = 86.363 Å, c = 96.047, $\alpha = \beta = \gamma = 90°$) and both space groups were $P2_12_12_1$. The SAD data, collected based on the anomalous diffraction from three SeMet sites in the mutated protein (M69, L89M, L124M), was used to solve the phases of the reflection data and after refinement, the R-work and R-free values were 0.20 and 0.25, respectively. The first factor considered when selecting amino acids for mutation to allow Se-Met incorporation was their position within the pentapeptide repeat sequences, with amino acids in i position considered ideal candidates since M residues are the fourth most common amino acid to occupy the i position in PRPs[37] and therefore were assumed to cause minimal perturbation to the native structure. The second factor considered was which i-residue codons could be most conveniently altered to allow mutation to a codon encoding for a M residue. For the native protein reflection data, the final R-work and R-free values were 0.21 and 0.27. The two structures had 3 - 4% Ramachandran outliers and 1 - 2% rotamer outliers (**Figure 4.1**). Both structures contained two chains and a total of 306 refined residues, which have been deposited to the PDB (PDB IDs: native-6UVI, SeMet: 6UV7).

| Name | Mutated Alr1298 | Native Alr1298 |
|---|---|---|
| Resolution range (Å) | 26.25 - 2.275 (2.356 - 2.275) | 29.9 - 2.1 (2.175 - 2.1) |
| Space group | P 21 21 21 | P 21 21 21 |
| Unit cell (Å, °) | a=45.842 b=86.363 c=96.047 $\alpha = \beta = \gamma = 90$ | a=45.762 b=85.922 c=95.686 $\alpha = \beta = \gamma = 90$ |
| Total reflections | 36202 (3530) | 44459 (4349) |
| Unique reflections | 18106 (1765) | 22287 (2178) |
| Multiplicity | 2.0 (2.0) | 2.0 (2.0) |
| Completeness (%) | 99.60 (99.55) | 98.03 (97.84) |
| Mean I/sigma(I) | 17.18 (2.83) | 18.53 (2.35) |
| Wilson B-factor ($Å^2$) | 55.91 | 40.41 |
| R-merge | 0.02771 (0.2676) | 0.01859 (0.3056) |
| R-meas | 0.03918 (0.3784) | 0.02629 (0.4322) |
| R-pim | 0.02771 (0.2676) | 0.01859 (0.3056) |
| CC1/2 | 0.999 (0.832) | 1 (0.862) |
| CC* | 1 (0.953) | 1 (0.962) |
| Reflections used in refinement | 18070 (1758) | 22260 (2176) |
| Reflections used for R-free | 1804 (176) | 1996 (196) |
| R-work | 0.1990 (0.2994) | 0.2085 (0.3201) |

| | | |
|---|---|---|
| R-free | 0.2509 (0.3637) | 0.2743 (0.3734) |
| CC (work) | 0.973 (0.898) | 0.967 (0.841) |
| CC (free) | 0.947 (0.855) | 0.964 (0.807) |
| Number of non-hydrogen atoms | 2430 | 2481 |
| macromolecules | 2413 | 2413 |
| solvent | 17 | 68 |
| Protein residues | 306 | 306 |
| RMS(bonds) (Å) | 0.014 | 0.012 |
| RMS(angles) (°) | 1.48 | 1.41 |
| Ramachandran favored (%) | 95.36 | 94.37 |
| Ramachandran allowed (%) | 3.31 | 4.30 |
| Ramachandran outliers (%) | 1.32 | 1.32 |
| Rotamer outliers (%) | 1.49 | 2.24 |
| Clashscore | 3.73 | 3.72 |
| Average B-factor ($Å^2$) | 65.04 | 54.73 |
| macromolecules ($Å^2$) | 65.10 | 54.87 |
| solvent ($Å^2$) | 56.55 | 49.83 |

**Table 4. 1 Data collection and refinement statistics.** Statistics for the highest-resolution shell are shown in parentheses.

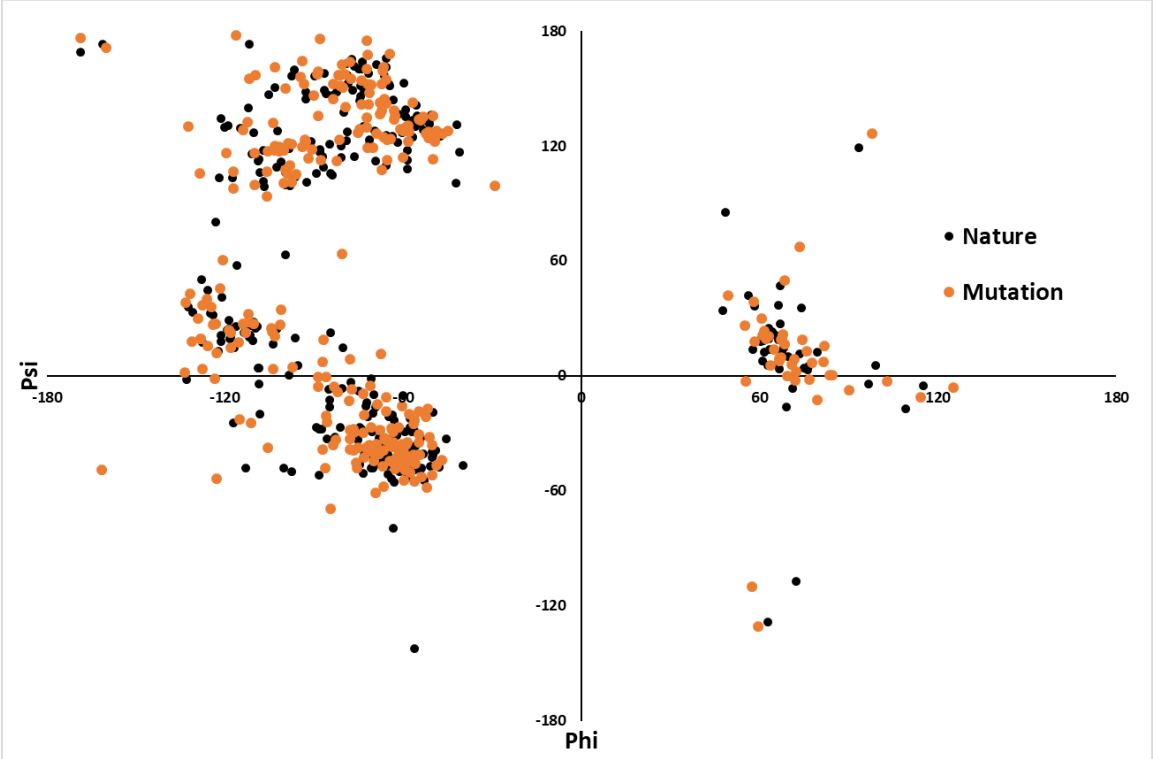


**Figure 4. 1 Ramachandran plot of nature and mutated Alr1298.** The black dots indicate the native Alr1298 Ramachandran distribution. The orange dots indicate the SeMet-mutated Alr1298 Ramachandran distribution.

### 4.4.2 Sequence and structure analysis

The full-length of Alr1298 is predicted to contain 167 amino acids (18.37 kDa) based on the sequence provided by the Kyoto Encyclopedia Genes and Genomes (KEGG) database (https://www.genome.jp/dbget-bin/www_bget?ana:alr1298). A single amino acid difference between the KEGG database sequence and the sequenced clone was discovered corresponding to amino acid 159, which was lysine in the database and glutamic acid in the sequenced clone (**Figure 4.2**). In chain A, nine amino acids at the N-terminus and six amino acids at C-terminus were not modelled in the deposited structures due to missing density in the electron density maps. In chain B, eight amino acids at the N-terminus and five amino acids at the C-terminus were missing. In the 167 amino acids defining the structure of Alr1298, the 75 amino acids from P67 to I141 constituted three and three-quarter Rfr coils with the incomplete coil occurring at the C-terminal end of the β helix (**Figure 4.3**). Of the remaining amino acids, 34 were found in the four α-helix cluster at the N-terminus, and eight constituted the α-helix at the C-terminus (**Figure 4.3**). Most of the remaining ~50 amino acids occurred in sections linking the α-helices to the β-helical structure, or in turns within the four-α-helix cluster at the N-terminus (**Figure 4.3**). PISA calculations[43, 44] indicated that the buried area between the four-helix cluster and the N-terminal base of the Rfr fold was 1051.8 Å$^2$ and 1054.5 Å$^2$ for the native and SeMet proteins, respectively and the free energy of dissociation, $\Delta G_{diss}$, was -2.3 kcal mol$^{-1}$ and -3.1 kcal mol$^{-1}$ , for the native and SeMet proteins, respectively (**Table 4.2**), indicating that the four-helix cluster is likely to undergo independent motion relative to that of the Rfr fold. Based on the comparison between the native and the mutated protein structures, mutation of the two leucine sites to introduce SeMet required for SAD phasing did not appear to change the structure of the main chain, especially in the pentapeptide repeat domain. Analysis of the backbone dihedral angles indicated that the PRP was composed of a combination of type II and IV β turns[37] and the average twist angles between each coil at different faces (native Alr1298: face1: -5.561°, face2: 0.0783°, face3: -9.073°, face 4: 1.708°).[37] Twist angles were defined as being equal to the angle between two vectors linking the carbonyl carbons of the $i-2$ and $i+2$ amino acids between two adjacent coils. In both the native and mutated structures, the two protein molecules in the asymmetric unit packed through an interaction between their N-terminal four α-helix clusters such that the axes of the two Rfr-folds made a nearly 90°, but slightly obtuse, angle (**Figure 4.4**). PISA calculations[43, 44] indicated that the buried area between the four-α-helix clusters from each molecule was around 850 Å$^2$, which was somewhat smaller than expected, e.g. >1000 Å$^2$, for a ~18.3 kDa protein that would form a stable homodimer structure.[45] The PISA-calculated $\Delta G_{diss}$ for the crystallographic dimer in the asymmetric unit was -8.5 kcal/mol and -9.9 kcal/mol for the SeMet and native protein structures, respectively (**Table 4.3**), indicating that the dimer state was energetically unfavorable and that Alr1298 likely exists as a monomer in its physiological active state. The four-α-helix cluster at the N-terminus in Alr1298 (**Figure 4.3**). stands out in comparison to all other known single domain PRP structures contain at most a single α-helix at the N-terminus, occurring in five out of the 12 existing PRP structures

(**Figure 4.5**). The remaining seven structures lack any element of secondary structure at the N-terminus (**Figure 4.5**).

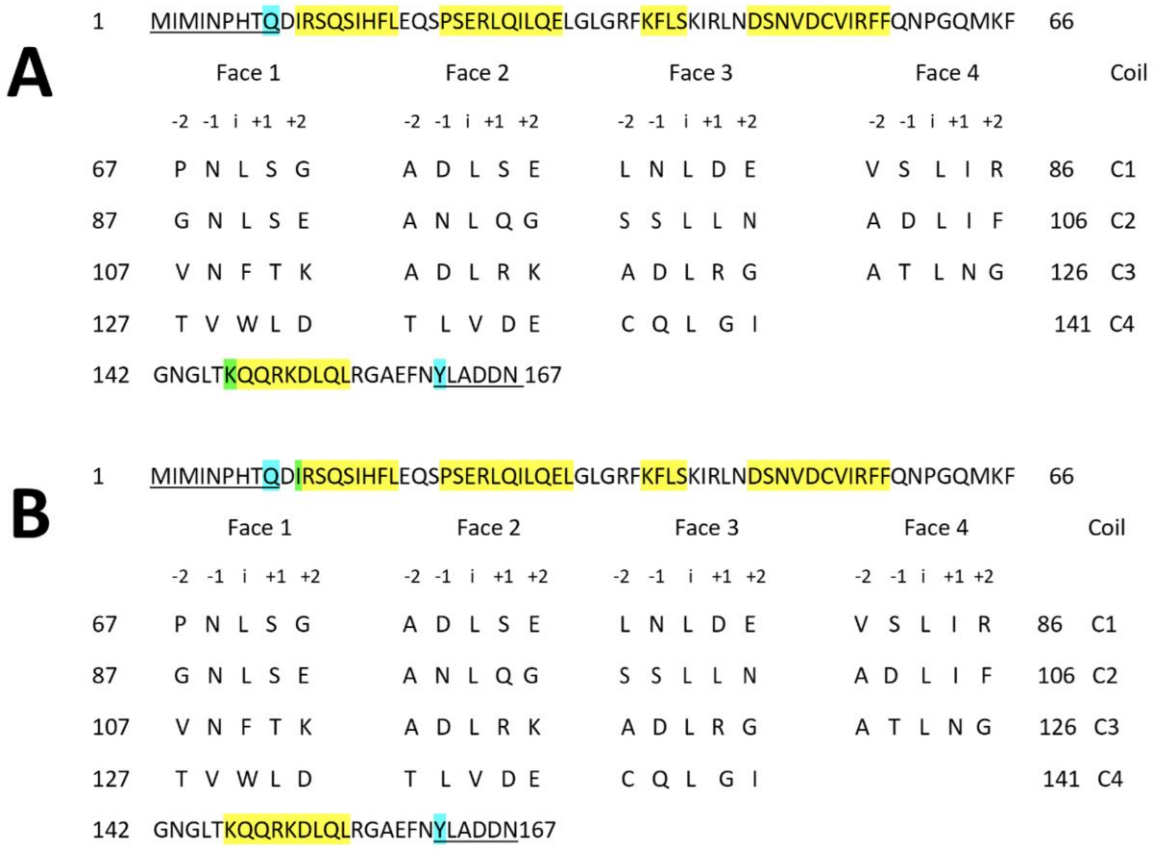



**Figure 4. 2 Identification of the pentapeptide repeat sequences in the native and Se-Met-substituted Alr1298 based on the crystal structure.** A) native Alr1298. B) Se-met substituted Alr1298. Residues 67 to 141 comprised the pentapeptide repeat domains defining three complete and one incomplete coils in the PRP structure. Underlined residues were missing in the electron density and not depicted in the 3-D structure. Residues highlighted in yellow were found in α helices. Residues highlighted in green were observed in the α-helix in chain B but not the chain A. Residues highlighted in blue were observed in the electron density for chain B but absent in the electron density for chain A.

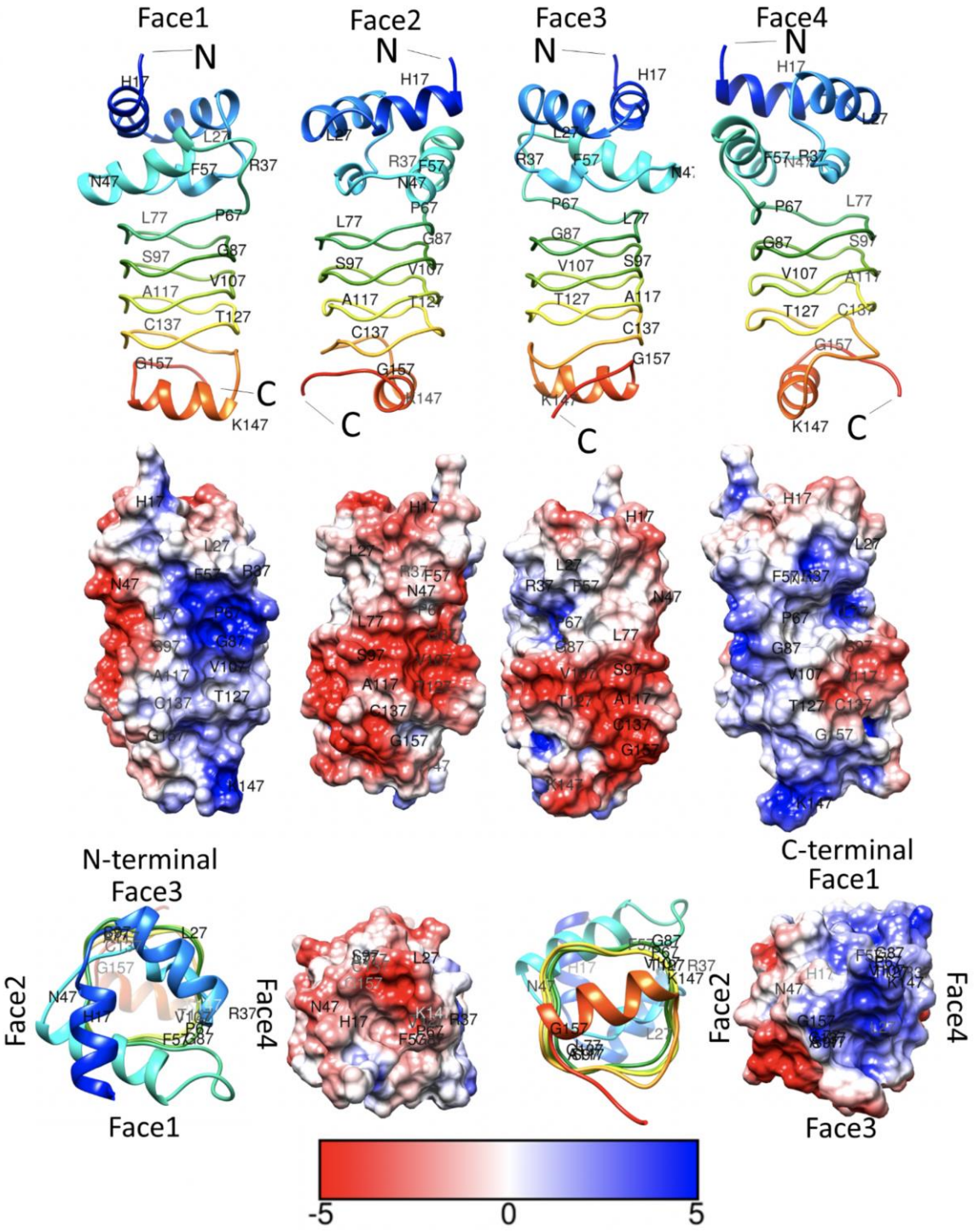



**Figure 4. 3 The structure and electrostatic surface potential of Alr1298.** The structures and the corresponding electrostatic surface potentials are depicted for each of the four faces and facing the two terminal ends of the protein. Red and blue on electrostatic surface potential are negative charge and positive charge, respectively. At

the bottom, in the structure facing the N-terminus, face 1 is pointing downward and faces 2, 3, and 4 can be identified in a clockwise direction relative to face 1. In the structure facing the C-terminus, face1 is at the top and faces 2, 3, and 4 can be identified in an anticlockwise direction relative to face 1.

| **Complex Summary** | | | | | |
|---|---|---|---|---|---|
| | Native Alr1298 | Mutated Alr1298 | | Native Alr1298 | Mutated Alr1298 |
| Multimeric State | 2 | 2 | Surface Area ($Å^2$) | 7570.9 | 7686.7 |
| Copies in unit cell | N/A | N/A | Buried Area ($Å^2$) | 1051.8 | 1054.5 |
| Formula | AB | AB | $\Delta G^{int}$ (kcal/mol) | -4.4 | -4.1 |
| Composition | BA | BA | $\Delta G^{diss}$ (kcal/mol) | -2.3 | -3.1 |
| Dissociation Pattern | B+A | B+A | $T\Delta S^{int}$ (kcal/mol) | 9.7 | 9.8 |
| Symmetry Number | 1 | 1 | | | |

**Table 4. 2 PISA results for the interaction between the four-α-helix cluster and the Rfr fold in Alr1298.**

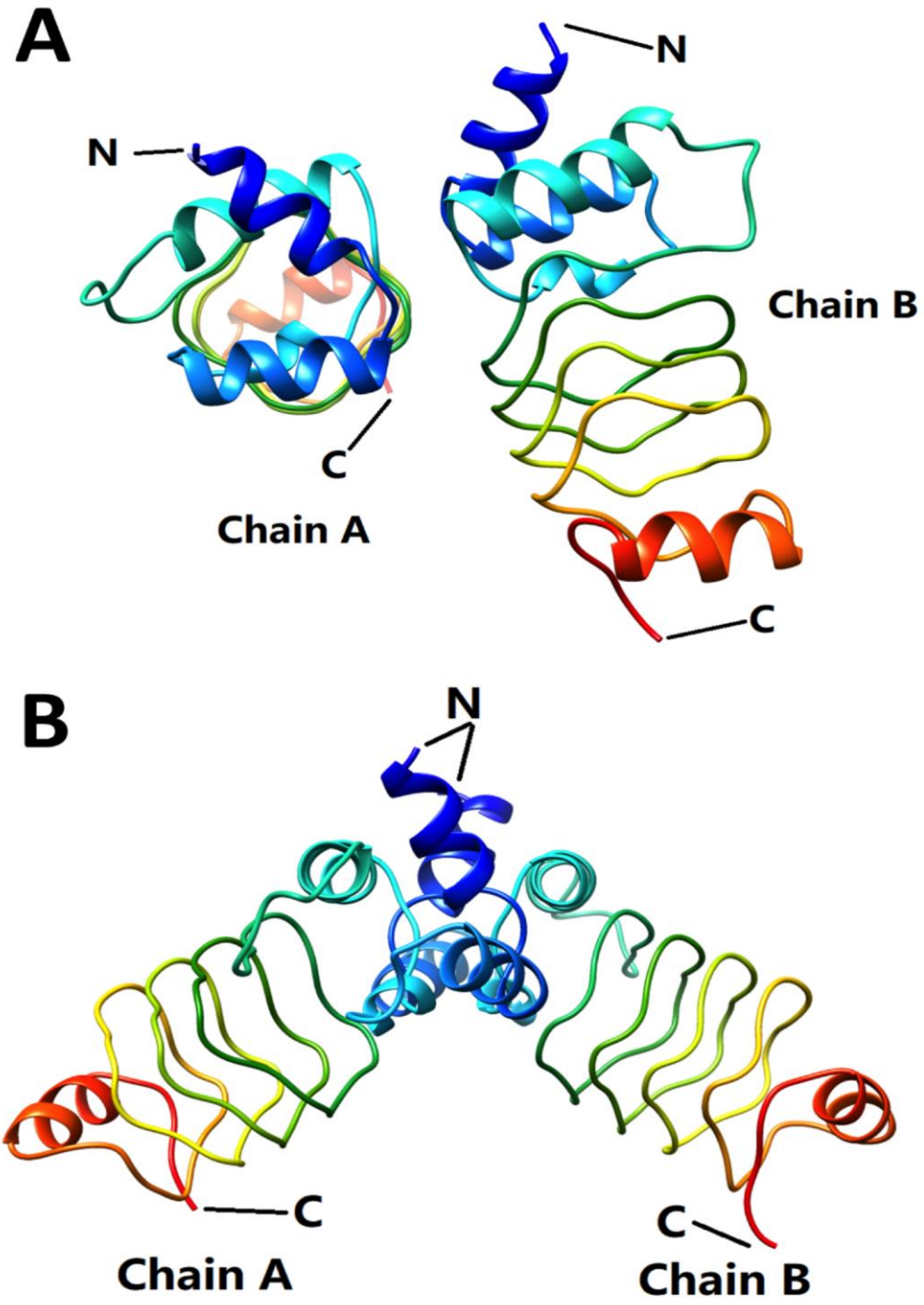



**Figure 4. 4 Depiction of the two molecules of Alr1298 crystal packing in the crystallographic asymmetric unit.** A) Along the Rfr-fold helix axis depiction of the crystal packing of the native Alr1298 in the asymmetric unit. B) In-plane Rfr-fold helix axis depiction of the crystal packing in the asymmetric unit of the native Alr1298 crystal structures.

| Complex Summary | | | | | |
|---|---|---|---|---|---|
| | Native Alr1298 | Mutated Alr1298 | | Native Alr1298 | Mutated Alr1298 |
| Multimeric State | 2 | 2 | Surface Area ($Å^2$) | 14770.2 | 15061.1 |
| Copies in unit cell | N/A | N/A | Buried Area ($Å^2$) | 856.8 | 837.7 |
| Formula | $A_2$ | $A_2$ | $\Delta G^{int}$ (kcal/mol) | -1.2 | -2.2 |
| Composition | AB | AB | $\Delta G^{diss}$ (kcal/mol) | -9.9 | -8.5 |
| Dissociation Pattern | A+B | A+B | $T\Delta S^{int}$ (kcal/mol) | 12.0 | 12.0 |
| Symmetry Number | 2 | 2 | | | |

**Table 4. 3 PISA results for the interaction between the four-α-helix clusters in the two molecules of Alr1298 in the crystallography asymmetric unit.**

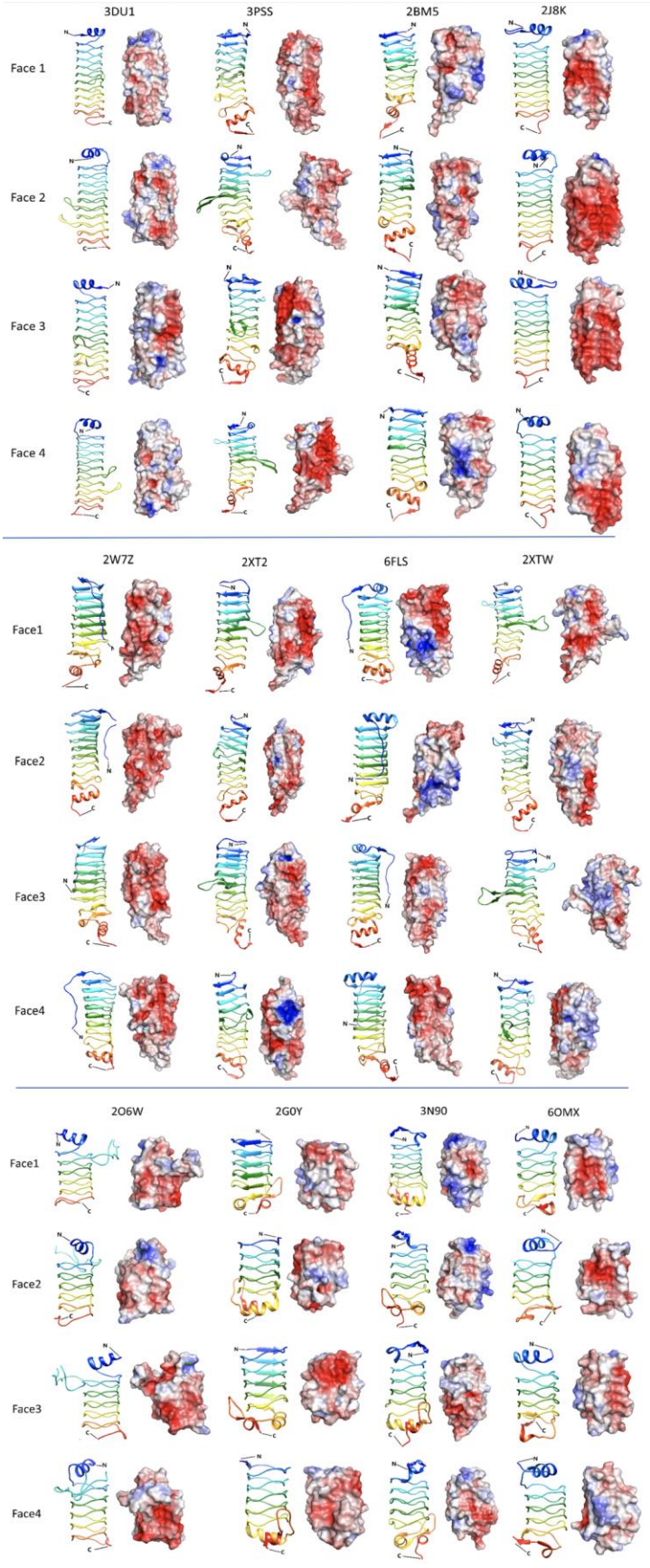

3DU1
3PSS
2BM5
2J8K
Face 1
Face 2
Face 3
Face 4
2W7Z
2XT2
6FLS
2XTW
Face1
Face2
Face3
Face4
2O6W
2G0Y
3N90
6OMX
Face1
Face2
Face3
Face4

**Figure 4. 5 Summary of the structures and electrostatic surface potentials for all other PRPs with known structures.** The structures are organized starting with the PRPs containing the most coils at the top of the figure to the fewest numbers of coils at the bottom of the figure. The PDB ID codes are indicated above each structure and four faces are indicated beneath each PDB ID code. The structures are rendered in ribbons and depicted with a rainbow coloring scheme.

### 4.4.3 Analysis of the electrostatic potential surface

In face 1, the electrostatic surface potential exhibited a dense pocket of positive-charge localized to coils 1 and 2, but otherwise the face was generally neutral in charge (**Figure 4.3**). In contrast, face 2 contained a dense pocket of negative charge localized to coils 3 and 4 with the remainder of the face being largely neutral in charge (**Figure 4.3**). Face 3 had a similar distribution of charge as compared to face 2 (**Figure 4.3**). Face 4 had a single dense pocket of negative charge localized on coils 2 and 3 while the remainder of the face had some pockets of positive charge and an otherwise neutral surface. The N-terminal four α-helix cluster exhibited a generally negative electrostatic surface potential while the C-terminal helix had a generally positive electrostatic surface potential (**Figure 4.3**).

Comparison of the electrostatic surface potential of Alr1298 with that of all other known PRP proteins (PBP IDs: 3DU1[20], 3PSS[46], 2BM5[47], 2J8K [48], 2W7Z[49], 2XT2[50], 6FLS[51], 2XTW[52], 2O6W[53], 2G0Y[54], 3N90[55], 6OMX[56]) indicated that the dense localized pocket of positive charge observed on face 1 of Alr1298 was observed in some other examples, most prominently in face 4 of 2XT2, and less dramatically in face 4 of 2BM5 and in face 1 of 6FLS (**Figure 4.5**). Faces 2 and 3 in Alr1298 had both dense pockets and generally localized regions of negative charge (**Figure 4.3**), of which one or both of these features are common to several existing PRP structures including 3PSS, 2J8K, 2W7Z and 2G0Y (**Figure 4.5**).

### 4.4.4 Analysis of the rotational correlation time ($\tau_c$)

Analysis of the crystal structure indicated that Alr1298 contained two copies in the asymmetric unit in the arrangement depicted in **Figure 4.4**. While there is no fundamental relationship between the number of molecules in observed in the crystallographic asymmetric unit and the number of molecules present in the biologically relevant physiological state, solution-state NMR spectroscopy was used to determine if Alr1298 existed as a monomer or dimer in solution by measuring and analyzing its rotational correlation time, $\tau_c$, in solution.[57] The $T_1$ and $T_2$ relaxation time constants were determined to be 1228 ms and 38.97 ms, respectively. Based on the $\tau_c$ calculated from the $T_1$ and $T_2$ values, the rotational correlation time of Alr1298 was determined to be 12.46 ns. Based on comparison of $\tau_c$ vs molecular weight data generated by the Northeast Structural Genomics (NESG) consortium,[58] it was determined that Alr1298 behaved as monomer in solution at 0.5 mM at 298 K with the estimated molecular weight (MW) based on the relaxation data corresponding to about 20 kDa in comparison to the predicted MW of 18.37 kDa based on the published amino acid sequence, at the concentration used for the NMR relaxation time measurements.

### 4.4.5 Circular dichroism (CD) spectral analysis and thermal melting analysis

The far-UV wavelength CD spectrum of Alr1298 had a minimum at 221 nm (**Figure 4.6**). Analysis of secondary structure content of Alr1298 using the BeStSel algorithm[41] (**Table 4.4**) predicted 7.4% α-helix, 27.9% antiparallel β-sheet and 7.9% parallel β-sheet. However, based on the analysis of the crystal structure, Alr1298 had 25.7% (86/334) α-helix, and 44.9% (150/334) β turns, indicating that the current prediction software could be improved to better account for Rfr fold CD contributions. Of the 30 β turns, 20 were type II, 8 were type IV, and the rest were type II'. In order to characterize the thermal stability of Alr1298, a CD-monitored thermal melt was conducted. Alr1298 was completely denatured by raising the temperature from 15 °C to 90 °C with the apparent melting temperature ($T_m$) determined to be 57.72 ± 0.15 °C. This $T_m$ is not unusual as far as protein stability goes, compared to the average $T_m$ of 62.2 °C reported for >1100 proteins for which thermodynamic parameters are compiled in the ProTherm database.[59-61] The $T_m$ for Alr1298 was significantly lower than the $T_m$ = 62.8 °C determined for At2g44920, which contained four and three quarters Rfr coils.[62] The enthalpy of unfolding of Alr1298 was determined to be +100.38 ± 5.54 kcal/mol (**Figure 4.6**). The denatured Alr1298 protein did not refold upon cooling, in contrast Alr5209, which was found to reversibly refold following thermal denaturation.[37]

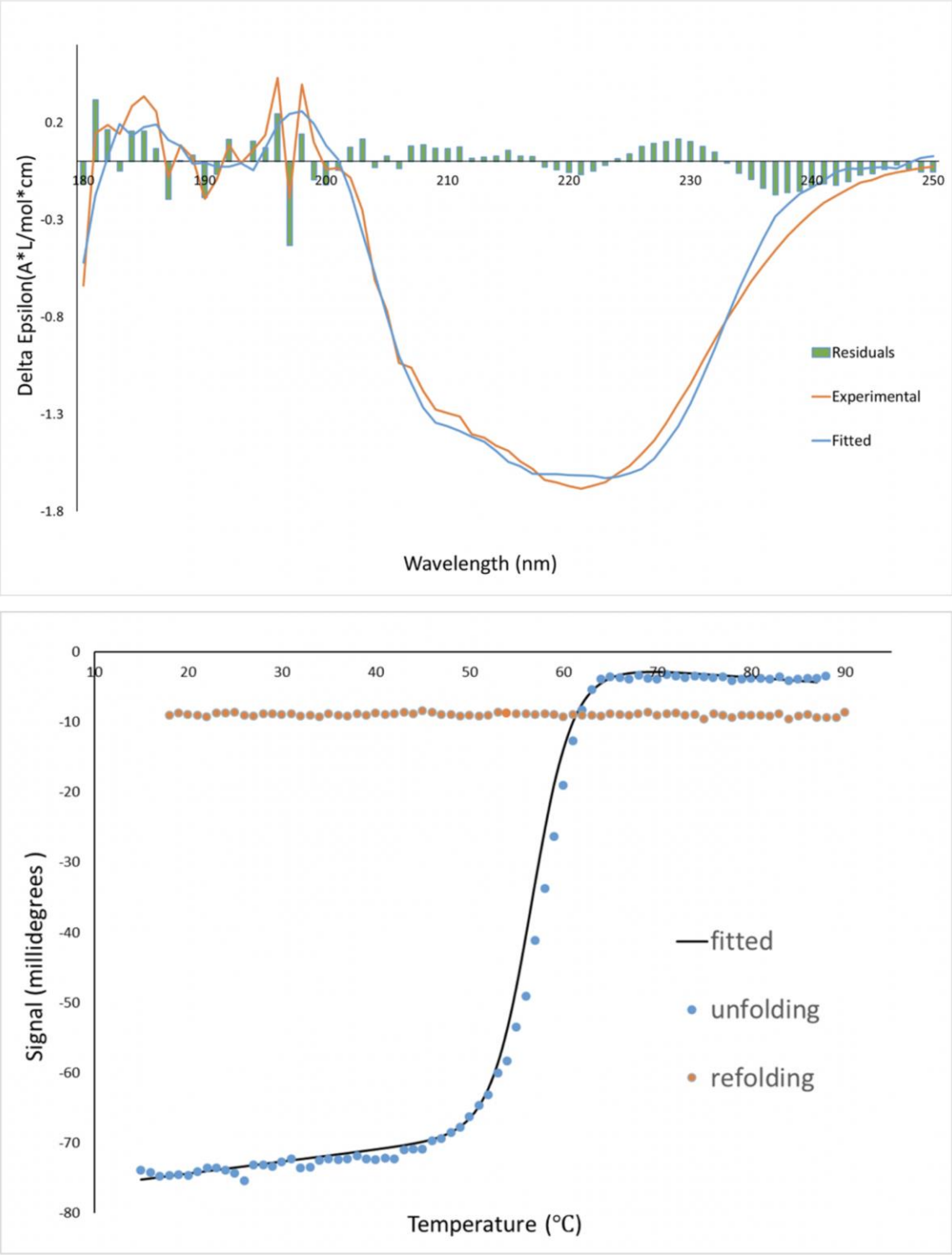



**Figure 4. 6 Circular dichroism data collected on Alr1298.** A) CD wavelength scan of Alr1298 from 180 nm to 250 nm at 25℃. The spectrum was collected using a protein concentration of 40 μM. The orange line indicates the experiment data and the blue line is the fitted curve (RMSD at 0.113). B) Temperature melting spectrum recorded from 15 to

90 ℃ recorded at 221 nm, the wavelength of strongest ellipticity. The blue data indicated the unfolding process and the orange data indicates the refolding process. The black line

| | | | |
|---|---|---|---|
| Helix | 7.4% | Helix1 (regular) | 4.0% |
| | | Helix2 (distorted): | 3.4% |
| Antiparallel | 27.9% | Anti1 (left-twisted): | 4.8% |
| | | Anti2 (relaxed): | 11.6% |
| | | Anti3 (right-twisted): | 11.4% |
| Parallel | 8.9% | Turn | 14.9% |
| Others | | 40.9% | |

**Table 4. 4 Predicted secondary structure content in the native Alr1298.**

indicates the fit of the unfolding process.

### 4.4.6 Analysis of the Alr1298 gene cluster for potential functional analysis

An analysis of the *Nostoc* sp PCC 7120 genome using the KEGG database indicated that Alr1298 belongs to a gene cluster potentially indicating that *alr1298* belongs to an operon. Given that genes that belong to a common operon often times share a related function,[63, 64] we analyzed the genes surrounding the alr1298 gene for clues about a potential function for Alr1298. The gene cluster contained three genes preceding and three genes following the *alr1298* gene. *Alr1295* was conserved in 14 of 15 aligned genomes and encodes a prohibitin homolog. Prohibitins are evolutionarily conserved genes that are generally recognized as inhibitors to cell proliferation and their function is related to tumor suppressors in animals and humans. In cyanobacteria, prohibitins have been linked to thylakoid biogenesis[65] and membrane synthesis.[66] Alr1296 is a 138 amino acid predicted hypothetical protein that is conserved in all 15 aligned genomes in KEGG. Alr1297 is an ABC transporter ATP binding protein annotated as an ABC transport system ATP-binding/permease protein that was also conserved in all 15 aligned genomes in KEGG. Alr1299 is annotated as a phosphoribosylglycinamide formyltransferase 2 annotated to be involved in purine metabolism, metabolic pathways, biosynthesis of secondary metabolites and biosynthesis of antibiotics.[67] Alr1302, annotated as a ribosomal protein alanine acetyltransferase, was also conserved in all 15 aligned genomes in KEGG. Two additional genes in the cluster, alr1300 and alr1301, were annotated as proteins of unknown function and were only conserved in four of the 15 aligned genomes in KEGG. Collectively, based on the function of some of the genes in the cluster, it appears that the operon may play a role in cell proliferation and potentially thylakoid biogenesis. Clearly, with four of the eight genes identified in the cluster having been annotated as proteins with unknown function, additional work is necessary to identify

both the function of Alr1298, the other genes in the operon, and the overall function of the operon.

### 4.4.7 Alr1298 gene expression in response to nitrogen deprivation

A genome-wide microarray analysis of 5336 out of 5338 ORFs in the *Nostoc* sp PCC 7120 genome was conducted to assess the changes in gene expression patterns in response to nitrogen deprivation.[68] The changes in gene expression were evaluated at 3 hours, 8 hours, and 24 hours following nitrogen starvation. Alr1298 gene expression was upregulated at all three points, reaching a peak at 8 hours post nitrogen starvation, with the fold change in expression levels equal to 1.85:1 at 3 hours, 4.10:1 at 8 hours, and 2.29:1 at 24 hours post nitrogen starvation. Since the primary response to nitrogen starvation in the filamentous *Nostoc* sp. PCC 7120 cyanobacterium is the spatially-controlled differentiation of vegetative cells into terminally differentiated heterocysts in a process that takes about 24 hours to complete following deprivation of combined nitrogen, this would suggest that the function of alr1298 is in some way related to either the response to nitrogen starvation or to differentiation of vegetative cells into heterocysts.

## 4.5 Conclusions

PRPs are a large superfamily of proteins found predominantly in ancient cyanobacteria. Despite the critical role that cyanobacteria played in the oxygenation of the earth's atmosphere, and the fact that they potentially represent the earliest organism on earth to undergo cell differentiation, the function of PRPs in cyanobacteria remains largely unknown. In the absence of experimental data establishing the function of PRPs, we are generating structural information for PRPs as a first step towards understanding the structure and function of PRPs in cyanobacteria. In this case, we solved the crystal structure of Alr1298 from *Nostoc* sp. PCC 7120. Alr1298 contains three and three-quarters Rfr coils, which positions it as equal to the PRP with a solved crystal structure containing the fewest number of Rfr coils. Concomitantly, Alr1298 exhibited a significantly reduced thermal stability in comparison to another PRP that contains four and three quarters Rfr coils. A unique structural feature of Alr1298 is the fact that it represents the first PRP structure that contains an elaboration of secondary structural elements at its N-terminus, specifically, a four-α-helix cluster. In contrast, all other existing PRP crystal structures have either a single α-helix or no secondary structural elements at the N-terminus. An analysis of its electrostatic surface potential indicated that it had distinct patches of dense both positive and negative charge, which may play important roles in establishing binding interactions with potential binding partners identified in future functional studies. In an attempt to discover its potential function, the KEGG database was queried and a gene cluster was identified indicating that Alr1298 likely belongs to an operon, and that the genes belonging to this cluster are likely related to a common function. Collectively, based on the annotated function of the genes in the cluster, it appears that the operon, and Alr1298, may play a role in cell proliferation and potentially thylakoid biogenesis. Finally, analysis of a genome-wide analysis of gene expression in *Nostoc* sp. PCC 7120 indicated that expression of Alr1298 is upregulated following initiation of nitrogen starvation, indicating that Alr1298 may be functionally

linked to initiation or biogenesis of heterocyst differentiation. Further investigations of the function of Alr1298 are underway in our laboratory.

**4.6 Acknowledgements**

We acknowledge access to the x-ray beamline 31-ID-D provided at the Advanced Photon Source (APS) at Argonne National Laboratory. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the Lilly Research Laboratories Collaborative Access Team (LRL-CAT) beamline at Sector 31 of the Advanced Photon Source was provided by Eli Lilly Company, which operates the facility. We also acknowledge molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

**4.7 References**

1. Giovannoni, SJ, Turner, S, Olsen, GJ, Barns, S, Lane, DJ, Pace, NR: Evolutionary relationships among cyanobacteria and green chloroplasts. *Journal of Bacteriology.* 1988; 170**:** 3584-3592.
2. Black, K, Buikema, WJ, Haselkorn, R: The Hglk gene Is required for localization of heterocyst-specific glycolipids in the cyanobacterium Anabaena Sp strain PCC-7120. *Journal of Bacteriology.* 1995; 177**:** 6440-6448.
3. Huang, TC, Lin, RF, Chu, MK, Chen, HM: Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium Synechococcus sp. strain RF-1. *Microbiology.* 1999; 145 ( Pt 3)**:** 743-753.
4. Liu, D, Golden, JW: hetL overexpression stimulates heterocyst formation in Anabaena sp. strain PCC 7120. *J Bacteriol.* 2002; 184**:** 6873-6881.
5. Stewart, WDP: Some aspects of structure and function in N2-fixing cyanobacteria. *Annu Rev Microbiol.* 1980; 34**:** 497-536.
6. Fleming, H, Haselkorn, R: Differentiation in Nostoc-muscorum - Nitrogenase is synthesized in heterocysts. *P Natl Acad Sci USA.* 1973; 70**:** 2727-2731.
7. Issa, AA, M.H., A-A, Ohyama, T: Nitrogen fixing cyanobacteria: Future prospect. In: *Advances in Biology and Ecology of Nitrogen Fixation.* edited by OHYAMA, T., Rijeka: In Tech, 2014.
8. Fay, P, Steward, WD, Walsby, AE, Fogg, GE: Is the heterocyst the site of nitrogen fixation in the blue-green algae? . *Nature* 1968; 220**:** 810-812.
9. Flores, E, Herrero, A: Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat Rev Microbiol.* 2010; 8**:** 39-50.
10. Haselkorn, R: Heterocysts. *Annu Rev Plant Phys.* 1978; 29**:** 319-344.
11. Bohme, H: Regulation of nitrogen fixation in heterocyst-forming cyanobacteria. *Trends Plant Sci.* 1998; 3**:** 346-351.
12. Herrero, A, Muro-Pastor, AM, Flores, E: Nitrogen control in cyanobacteria. *Journal of Bacteriology.* 2001; 183**:** 411-425.
13. Zhang, CC, Laurent, S, Sakr, S, Peng, L, Bedu, S: Heterocyst differentiation and pattern formation in cyanobacteria: a chorus of signals. *Mol Microbiol.* 2006; 59**:** 367-375.

14. Stanier, RY, Cohen-Bazire, G: Phototrophic prokaryotes: the cyanobacteria. *Annu Rev Microbiol.* 1977; 31**:** 225-274.
15. Vazquez-Bermudez, MF, Herrero, A, Flores, E: Uptake of 2-oxoglutarate in Synechococcus strains transformed with the Escherichia coli kgtP gene. *J Bacteriol.* 2000; 182**:** 211-215.
16. Muro-Pastor, MI, Reyes, JC, Florencio, FJ: Cyanobacteria perceive nitrogen status by sensing intracellular 2-oxoglutarate levels. *J Biol Chem.* 2001; 276**:** 38320-38328.
17. Li, JH, Laurent, S, Konde, V, Bedu, S, Zhang, CC: An increase in the level of 2-oxoglutarate promotes heterocyst development in the cyanobacterium Anabaena sp. strain PCC 7120. *Microbiology.* 2003; 149**:** 3257-3263.
18. Flores, E, Herrero, A: Nitrogen assimilation and nitrogen control in cyanobacteria. *Biochem Soc Trans.* 2005; 33**:** 164-167.
19. Buikema, WJ, Haselkorn, R: Characterization of a gene controlling heterocyst differentiation in the cyanobacterium Anabaena 7120. *Genes Dev.* 1991; 5**:** 321-330.
20. Ni, S, Sheldrick, GM, Benning, MM, Kennedy, MA: The 2A resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp. strain PCC 7120. *J Struct Biol.* 2009; 165**:** 47-52.
21. Ehira, S, Miyazaki, S: Regulation of genes involved in heterocyst differentiation in the cyanobacterium Anabaena sp. strain PCC 7120 by a group 2 sigma factor SigC. *Life (Basel).* 2015; 5**:** 587-603.
22. Khudyakov, I, Wolk, CP: hetC, a gene coding for a protein similar to bacterial ABC protein exporters, is involved in early regulation of heterocyst differentiation in Anabaena sp. strain PCC 7120. *Journal of Bacteriology.* 1997; 179**:** 6971-6978.
23. Jang, JC, Wang, L, Jeanjean, R, Zhang, CC: PrpJ, a PP2C-type protein phosphatase located on the plasma membrane, is involved in heterocyst maturation in the cyanobacterium Anabaena sp PCC 7120. *Molecular Microbiology.* 2007; 64**:** 347-358.
24. Ehira, S, Ohmori, M, Sato, N: Genome-wide expression analysis of the responses to nitrogen deprivation in the heterocyst-forming cyanobacterium Anabaena sp. strain PCC 7120. *DNA Res.* 2003; 10**:** 97-113.
25. El-Gebali, S, Mistry, J, Bateman, A, Eddy, SR, Luciani, A, Potter, SC, Qureshi, M, Richardson, LJ, Salazar, GA, Smart, A, Sonnhammer, ELL, Hirsh, L, Paladin, L, Piovesan, D, Tosatto, SCE, Finn, RD: The Pfam protein families database in 2019. *Nucleic Acids Res.* 2019; 47**:** D427-D432.
26. Finn, RD, Coggill, P, Eberhardt, RY, Eddy, SR, Mistry, J, Mitchell, AL, Potter, SC, Punta, M, Qureshi, M, Sangrador-Vegas, A, Salazar, GA, Tate, J, Bateman, A: The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016; 44**:** D279-285.
27. Finn, RD, Bateman, A, Clements, J, Coggill, P, Eberhardt, RY, Eddy, SR, Heger, A, Hetherington, K, Holm, L, Mistry, J, Sonnhammer, EL, Tate, J, Punta, M: Pfam: the protein families database. *Nucleic Acids Res.* 2014; 42**:** D222-230.
28. Punta, M, Coggill, PC, Eberhardt, RY, Mistry, J, Tate, J, Boursnell, C, Pang, N, Forslund, K, Ceric, G, Clements, J, Heger, A, Holm, L, Sonnhammer, EL, Eddy,

SR, Bateman, A, Finn, RD: The Pfam protein families database. *Nucleic Acids Res.* 2012; 40**:** D290-301.
29. Finn, RD, Mistry, J, Tate, J, Coggill, P, Heger, A, Pollington, JE, Gavin, OL, Gunasekaran, P, Ceric, G, Forslund, K, Holm, L, Sonnhammer, EL, Eddy, SR, Bateman, A: The Pfam protein families database. *Nucleic Acids Res.* 2010; 38**:** D211-222.
30. Finn, RD, Tate, J, Mistry, J, Coggill, PC, Sammut, SJ, Hotz, HR, Ceric, G, Forslund, K, Eddy, SR, Sonnhammer, EL, Bateman, A: The Pfam protein families database. *Nucleic Acids Res.* 2008; 36**:** D281-288.
31. Bateman, A, Coin, L, Durbin, R, Finn, RD, Hollich, V, Griffiths-Jones, S, Khanna, A, Marshall, M, Moxon, S, Sonnhammer, EL, Studholme, DJ, Yeats, C, Eddy, SR: The Pfam protein families database. *Nucleic Acids Res.* 2004; 32**:** D138-141.
32. Bateman, A, Birney, E, Cerruti, L, Durbin, R, Etwiller, L, Eddy, SR, Griffiths-Jones, S, Howe, KL, Marshall, M, Sonnhammer, EL: The Pfam protein families database. *Nucleic Acids Res.* 2002; 30**:** 276-280.
33. Bateman, A, Birney, E, Durbin, R, Eddy, SR, Howe, KL, Sonnhammer, EL: The Pfam protein families database. *Nucleic Acids Res.* 2000; 28**:** 263-266.
34. Sonnhammer, EL, Eddy, SR, Durbin, R: Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins.* 1997; 28**:** 405-420.
35. Bateman, A, Murzin, AG, Teichmann, SA: Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci.* 1998; 7**:** 1477-1480.
36. Vetting, MW, Hegde, SS, Fajardo, JE, Fiser, A, Roderick, SL, Takiff, HE, Blanchard, JS: Pentapeptide repeat proteins. *Biochemistry.* 2006; 45**:** 1-10.
37. Zhang, R, Ni, S, Kennedy, MA: Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7 angström resolution. *Journal of Structural Biology X.* 2019; 3**:** 100010.
38. Liebschner, D, Afonine, PV, Baker, ML, Bunkoczi, G, Chen, VB, Croll, TI, Hintze, B, Hung, L-W, Jain, S, McCoy, AJ, Moriarty, NW, Oeffner, RD, Poon, BK, Prisant, MG, Read, RJ, Richardson, JS, Richardson, DC, Sammito, MD, Sobolev, OV, Stockwell, DH, Terwilliger, TC, Urzhumtsev, AG, Videau, LL, Williams, CJ, Adams, PD: Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst.* 2019; D75**:** 861-877.
39. Dolinsky, TJ, Nielsen, JE, McCammon, JA, Baker, NA: PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 2004; 32**:** W665-667.
40. Pettersen, EF, Goddard, TD, Huang, CC, Couch, GS, Greenblatt, DM, Meng, EC, Ferrin, TE: UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* 2004; 25**:** 1605-1612.
41. Micsonai, A, Wien, F, Kernya, L, Lee, YH, Goto, Y, Refregiers, M, Kardos, J: Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci U S A.* 2015; 112**:** E3095-3103.
42. Mazurenko, S, Stourac, J, Kunka, A, Nedeljkovic, S, Bednar, D, Prokop, Z, Damborsky, J: CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res.* 2018; 46**:** W344-W349.

43. Krissinel, E, Henrick, K: Detection of protein assemblies in crystals. *Lect Notes Comput Sc.* 2005; 3695**:** 163-174.
44. Krissinel, E, Henrick, K: Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology.* 2007; 372**:** 774-797.
45. Jones, S, Thornton, JM: Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.* 1996; 93**:** 13-20.
46. Xiong, XL, Bromley, EHC, Oelschlaeger, P, Woolfson, DN, Spencer, J: Structural insights into quinolone antibiotic resistance mediated by pentapeptide repeat proteins: conserved surface loops direct the activity of a Qnr protein from a Gram-negative bacterium. *Nucleic Acids Research.* 2011; 39**:** 3917-3927.
47. Hegde, SS, Vetting Mw Fau - Roderick, SL, Roderick Sl Fau - Mitchenall, LA, Mitchenall La Fau - Maxwell, A, Maxwell A Fau - Takiff, HE, Takiff He Fau - Blanchard, JS, Blanchard, JS: A fluoroquinolone resistance protein from Mycobacterium tuberculosis that mimics DNA. *Science.* 2005; 308**:** 1480-1483.
48. Vetting, MW, Hegde, SS, Hazleton, KZ, Blanchard, JS: Structural characterization of the fusion of two pentapeptide repeat proteins, Np275 and Np276, from Nostoc punctiforme: resurrection of an ancestral protein. *Protein Sci.* 2007; 16**:** 755-760.
49. Vetting, MW, Hegde Ss Fau - Blanchard, JS, Blanchard, JS: Crystallization of a pentapeptide-repeat protein by reductive cyclic pentylation of free amines with glutaraldehyde.
50. Vetting, MW, Hegde, SS, Zhang, Y, Blanchard, JS: Pentapeptide-repeat proteins that act as topoisomerase poison resistance factors have a common dimer interface. *Acta Crystallogr Sect F Struct Biol Cryst Commun.* 2011; 67**:** 296-302.
51. Notari, L, Martinez-Carranza, M, Farias-Rico, JA, Stenmark, P, von Heijne, G: Cotranslational Folding of a Pentarepeat beta-Helix Protein. *J Mol Biol.* 2018; 430**:** 5196-5206.
52. Vetting, MW, Hegde, SS, Wang, M, Jacoby, GA, Hooper, DC, Blanchard, JS: Structure of QnrB1, a plasmid-mediated fluoroquinolone resistance factor. *J Biol Chem.* 2011; 286**:** 25265-25273.
53. Buchko, GW, Robinson, H, Pakrasi, HB, Kennedy, MA: Insights into the structural variation between pentapeptide repeat proteins - Crystal structure of Rfr23 from Cyanothece 51142. *Journal of Structural Biology.* 2008; 162**:** 184-192.
54. Buchko, GW, Ni, SS, Robinson, H, Welsh, EA, Pakrasi, HB, Kennedy, MA: Characterization of two potentially universal turn motifs that shape the repeated five-residues fold - Crystal structure of a lumenal pentapeptide repeat protein from Cyanothece 51142. *Protein Sci.* 2006; 15**:** 2579-2595.
55. Ni, S, M.E., M, S.L., T, Jones, AN, S., J, L., T, Kennedy, MA: The 1.7 A resolution structure of At2g44920, a pentapeptide-repeat protein in the thylakoid lumen of Arabidopsis thaliana. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2011; 67**:** 1480-1484.
56. Zhang, R, Ni, S, Kennedy, MA: Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7 angström resolution. *Journal of Structural Biology: X.* 2019; 3**:** 100010.

57. Kay, LE, Torchia, DA, Bax, A: Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry.* 1989; 28**:** 8972-8979.
58. Aramini, JM, Ma, LC, Zhou, L, Schauder, CM, Hamilton, K, Amer, BR, Mack, TR, Lee, HW, Ciccosanti, CT, Zhao, L, Xiao, R, Krug, RM, Montelione, GT: Dimer interface of the effector domain of non-structural protein 1 from influenza A virus: an interface with multiple functions. *J Biol Chem.* 2011; 286**:** 26050-26060.
59. Bava, KA, Gromiha, MM, Uedaira, H, Kitajima, K, Sarai, A: ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 2004; 32**:** D120-121.
60. Gromiha, MM, Uedaira, H, An, J, Selvaraj, S, Prabakaran, P, Sarai, A: ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. *Nucleic Acids Res.* 2002; 30**:** 301-302.
61. Gromiha, MM, An, J, Kono, H, Oobatake, M, Uedaira, H, Prabakaran, P, Sarai, A: ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 2000; 28**:** 283-285.
62. Xu, S, Ni, S, Kennedy, MA: NMR Analysis of Amide Hydrogen Exchange Rates in a Pentapeptide-Repeat Protein from A. thaliana. *Biophys J.* 2017; 112**:** 2075-2088.
63. Ralston, A: Operons and Prokaryotic Gene Regulation. *Nature Education.* 2008; 1**:** 216.
64. Jacob, F, Perrin, D, Sanchez, C, Monod, J, Edelstein, S: [The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729]. *C R Biol.* 2005; 328**:** 514-520.
65. Mares, J, Strunecky, O, Bucinska, L, Wiedermannova, J: Evolutionary patterns of thylakoid architecture in cyanobacteria. *Front Microbiol.* 2019; 10.
66. Boehm, M, Nield, J, Zhang, P, Aro, EM, Komenda, J, Nixon, PJ: Structural and mutational analysis of band 7 proteins in the cyanobacterium Synechocystis sp. strain PCC 6803. *J Bacteriol.* 2009; 191**:** 6425-6435.
67. Lv, Q, Ma, W, Liu, H, Li, J, Wang, H, Lu, F, Zhao, C, Shi, T: Genome-wide protein-protein interactions and protein function exploration in cyanobacteria. *Sci Rep.* 2015; 5**:** 15519.
68. Ehira, S, Ohmori, M: NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium Anabaena sp. strain PCC 7120. *Mol Microbiol.* 2006; 59**:** 1692-1703.

# Chapter 5: Introduction of a new scheme for classifying β turns in protein structures

Reproduced with permission from:

Ruojing Zhang[1], Michael Stahr [2], Michael A. Kennedy*[1]

[1]Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056

[2]Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056

*Corresponding Author: Department of Chemistry and Biochemistry, 106 Hughes Laboratories, Miami University, 651 East High Street, Oxford, OH 45056. Email: kennedm4@miamioh.edu. Phone: 513-529-8267. Fax: 513-529-5715.

This paper has been submitted for publication.

Author contributions: RZ contributed to data collection, data analysis, manuscript preparation. MS contributed to data collection and software development. MAK contributed to data analysis and manuscript preparation.

## 5.1 Abstract

Protein β turn classification remains an area of ongoing development in structural biology research. While the commonly used nomenclature defining type I, type II and type IV β turns was introduced in the 1970s and 1980s, refinements of β-turn type definitions have been introduced as recently as 2019 by Dunbrack, Jr and co-workers who expanded the number of β-turn types to 18. Based on their analysis of 13,030 turns from 1074 ultrahigh resolution (≤ 1.2Å) protein structures, they used a new clustering algorithm to expand the definitions used to classify protein β-turns and introduced a new nomenclature system. We recently encountered a specific problem when classifying β-turns in crystal structures of pentapeptide repeat proteins (PRPs) determined in our lab that are largely composed of β-turns that often lie close to, but just outside of, canonical β-turn regions. To address this problem, we devised a new scheme that merges the Klyne-Prelog stereochemistry nomenclature and definitions with the Ramachandran plot. The resulting Klyne-Prelog-modified Ramachandran plot schema defines 1296 distinct potential β-turn classifications that cover all possible protein β-turn space with a nomenclature that indicates the stereochemistry of i+1 and i+2 backbone dihedral angles. The utility of the new classification scheme was illustrated by re-classification of the β turns in all known protein structures in the PRP superfamily and further assessed using a database of 16657 high-resolution protein structures (≤ 1.5Å) from which 522776 β turns were identified and classified.

## 5.2 Introduction

While repetitive α-helical and β-sheet secondary structural elements in proteins have been studied for many decades and their description and classification is well established [1-3], methods for classification of irregular secondary structural elements in proteins, such as tight turns, are still being developed [1]. Classification of protein tight turns is generally based on the number of residues involved in the turn and the spacing of residues involved in forming a hydrogen bond between the amide group of the last amino acid involved in the turn with the carbonyl group of a preceding amino acid involved in the turn [2]. Currently, six categories of protein turns are recognized [2]. The β-turn has been given special attention due to the fact that it is the most common type of tight turn and because of its involvement in the processes of molecular recognition and ligand binding [4]. The β-turn was first described by Venkatachalam in 1968 [5] in which three types of β-turns and their mirror configurations were classified based on the dihedral angles of i+1 and i+2 residues (type I, II, III and type I', II', III') [5]. In 1973, the definitions of types of β-turns were increased to ten by Scheraga and coworkers [6] (type I, II, III, IV, V, VI, VII and type I', II', III'). The discovery that β-turns could occur without the involvement of hydrogen bonds prompted the definition of "open β-turns", which led to a refinement of the definition of β-turns that specified that the distance between Cα of i and i+3 residues must be less than 7 Å, which allowed for recognition and definition of open β-turns in the absence of a characteristic hydrogen bond between the carbonyl group of the first residue in the turn, i.e. the i residue, and the amide group of the last residue in the turn, i.e. the i+3 residue [6]. In addition to type I and type II β-turns, type IV β-turns were introduced as a "catch all" category to classify miscellaneous β-turns that had combinations of φ and ψ backbone angles that did not fall into the canonical ranges used to defined type I and type II β turns [6]. In 1981, because the ideal values of type III and I were so close and due to

limited examples of type V and VII turns, the number of β-turns was reduced to seven by Richardson (type I, II, IV, VIa, VIb and type I', II') [7]. The current widely used criteria for classification of β-turns were established in 1994 by Hutchinson and Thornton [8]. These criteria stated that β-turns should have four consecutive residues with a distance between Cα of i and i+3 residues being less than 7 Å and the types of β-turns were organized into nine categories (type I, II, VIII, IV, VIa1, VIa2, VIb and type I', II') with ranges of ± 30° for three of the angles and a more liberal range of ± 45° for one of the angles [2, 8].

Despite the existence of classification rules for nine distinct β-turn types, the miscellaneous type IV β-turn category remains one of the largest populated classification categories, representing about 32% of all β-turns [9]. As the number of solved protein structures has increased over the years, more type IV β-turns have been documented [9]. In 1990, Wilmot and Thornton leveraged the Ramachandran plot [10-11] by applying Ramachandran regions to create a new nomenclature. They divided the Ramachandran plot into six regions and combined those regions to create new ranges for possible β-turns [12]. The updated Ramachandran plot and the introduction of fuzzy borders for each of the regions establishes a degree of ambiguity in the nomenclature used to classify β turns [13]. Wilmot and Thornton's nomenclature also did not cover Ramachandran outliers that existed in the natural protein structure space that were corroborated in high quality crystal structures [12]. Koch and Klebe introduced a classification scheme that considers the value of the ω angles due to the rare existence of cis peptide bond which usually involved in proline and extended the distance from 7 Å to 10 Å to reclassify all β-turns in 2009 [4], and four additional new type IV β-turns were introduced by de Brevern in 2016 [9]. The most recently introduced classification scheme was presented by Shapovalov *et al.* in 2019, which utilized the Ramachandran plot and from which 18 refined categories were established [14]. The continued expansion of the definitions of type IV β-turns has increased ambiguity in the precise characteristics that define type IV β-turns. Also, when the combinations of φ and ψ angles deviate by small values from the canonical ranges used to classify β-turns, this leads to the existence of apparent "border β-turns" (i.e. β-turns that nearly satisfy two or more definitions of different types of β-turns). The existence of "border β-turns" can lead to ambiguity of the description and analysis of experimentally determined [15-16] and *ab initio* calculated protein structures [17].

Here, we introduce a new method to describe β-turns that merges the Klyne-Prelog system used to define the stereochemistry about single bonds [18] with the Ramachandran plot used to define protein backbone torsion angles. The resulting Klyne-Prelog-modified Ramachandran plot can be used to specify all β-turns in a way that eliminates ambiguity in the precise conformation of the β-turns and provides enhanced ability to understand the conformation of the β-turn directly from the classification name. We first illustrate the utility of the new classification scheme by re-classifying all the β-turns in all known structures of the superfamily of pentapeptide repeat protein (PRPs) [9, 19-28], which are highly enriched in β-turns. We further assessed the utility of the new classification scheme using 16657 high-resolution crystal structures from the protein data bank (PDB) with resolution < 1.5 Å from which 522776 β turns were identified and classified.

## 5.3 Materials and Methods

### 5.3.1 Database of high-resolution protein crystal structures used for analysis

The β turn database was established from 16657 high-resolution protein crystal structures (resolution < 1.5Å) extracted from the RCSB PDB[15]. Multiple structures of the same protein in the PDB were *not* excluded so as to avoid missing potentially unique new turn types. The β turn database was generated using the open-source software Betaturn18[14]. β-turn recognition was based on DSSP secondary structure classification[29-30] with the distance between the Cα carbon atoms of the i and i+3 residues was less than 7Å. The Betaturn18 software was executed using Python2. The final database contained 522776 β turns used for analysis.

### 5.3.2 Construction of the new β turn classification algorithm

Classification of the β turns was based on paired values of the φ and ψ backbone torsion angles of the i+1 and i+2 residues in the β turn. The Klyne-Prelog rules defining stereochemistry about single bonds were used to define the naming convention for the new classification scheme where dihedral angles between -30º and 30º are called *synperiplanar* and dihedral angles between -150º and 150º are called *antiperiplanar*. Dihedral angles between +30º to +90º are called + *synclinal* and the dihedral angles between -30º to -90º are called - *synclinal*. Finally, dihedral angles between +90º to +150º are called + *anticlinal* and dihedral angles between -90º to -150º are called - *anticlinal*. These stereochemistry definitions are usually depicted in the context of a Newman projection diagram (**Figure 5.1**). To resolve ambiguities that arise regarding the classification of all possible β turns in the complete universe of β turn space, we mapped the Klyne-Prelog stereochemistry definitions onto the two-dimensional Ramachandran space used to define β turns. Because each dihedral angle can assume one of six possible values, i.e. synperiplanar, antiperiplanar, ±synclinal and ±anticlinal (**Figure 5.1**), there are 36 possible combinations of stereochemistry assignments for a single φ/ψ pair. Mapping the stereochemistry definitions onto a conventional Ramachandran plot using different colors to indicate the distinct stereochemistry ranges results in 36 distinct rectangular-shaped regions (**Figure 5.2**) with strictly-defined delimitations that can be used to classify β turns in proteins. When we consider both the i+1 and i+2 together, there are 36 x 36, i.e. 1296 possible stereochemistry combinations for the φ and ψ backbone dihedral angles in two consecutive amino acids in a protein backbone, such as occurs in β turns.

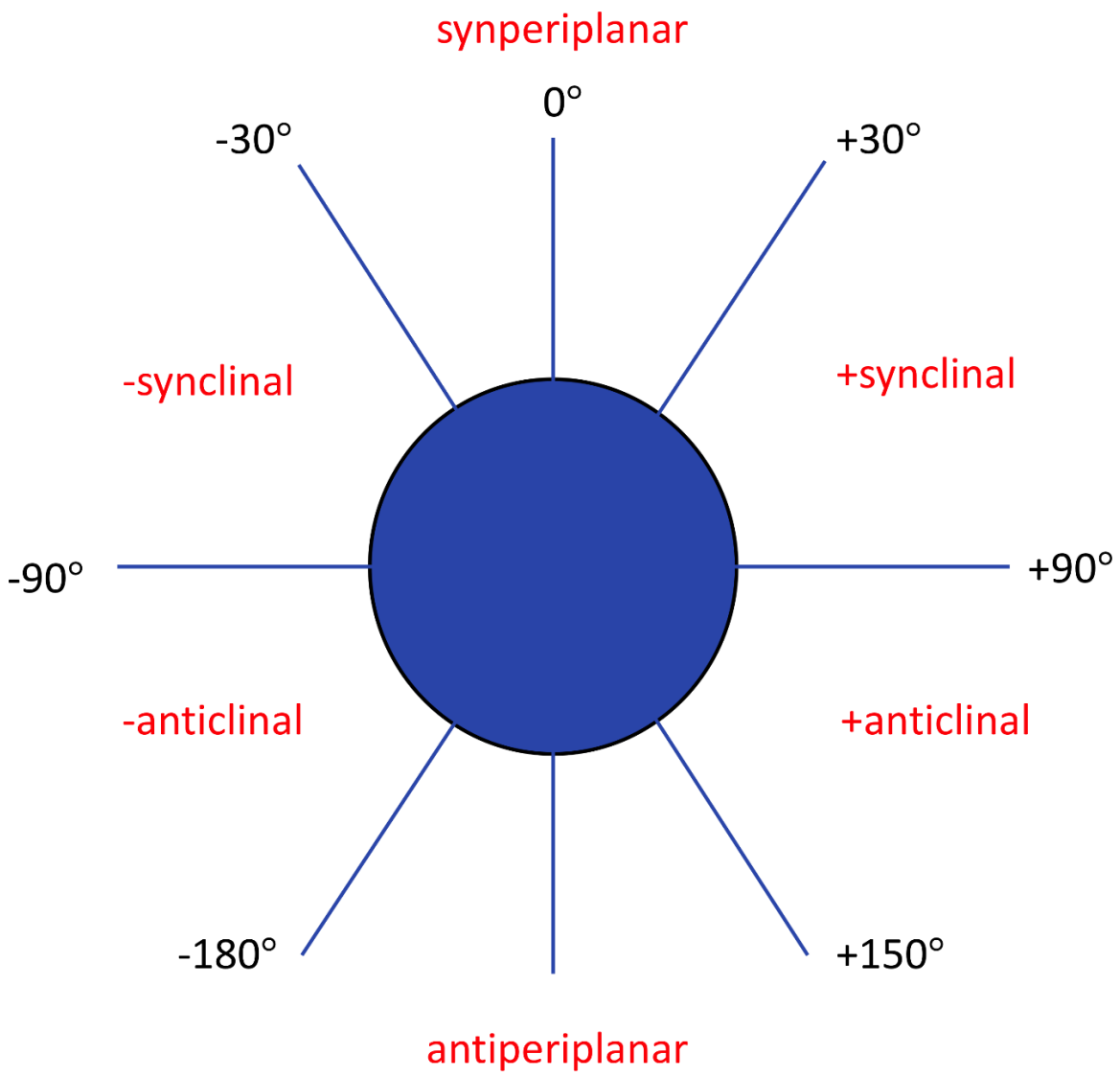



**Figure 5. 1 Newman projection diagram depicting the dihedral angle stereochemistry definitions.**

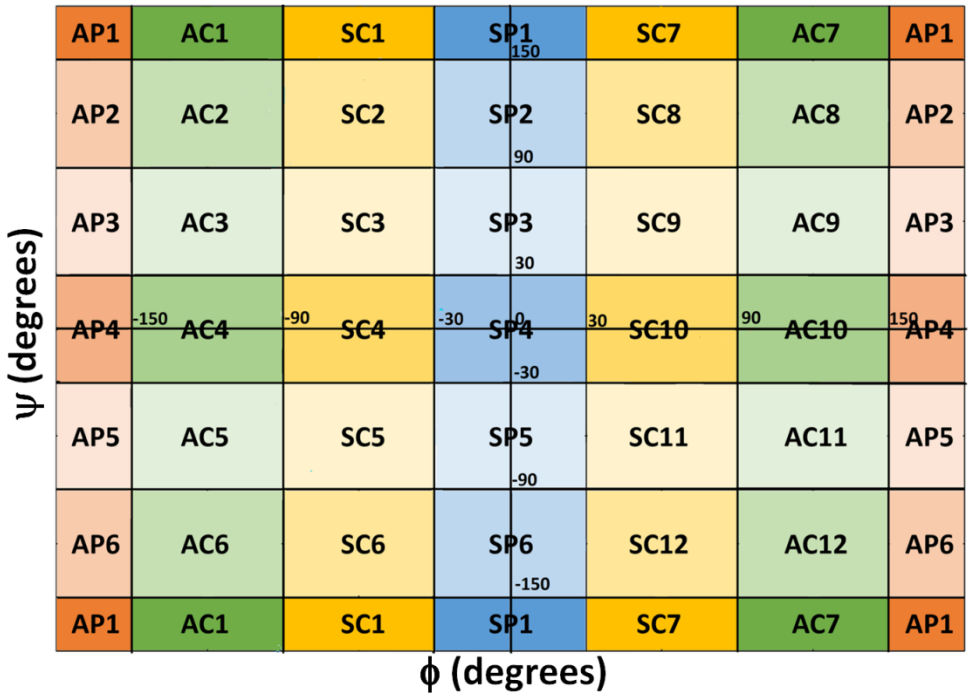



**Figure 5. 2 A Klyne-Prelog-modified Ramachandran plot indicating the φ and ψ dihedral ranges used to specify the stereochemistry values in organic molecules.** The distinct stereochemistry definition ranges for dihedral angles were coded as follows: synclinal (yellow), anticlinal (green), synperiplanar (blue), and antiperiplanar (orange). Since the φ and ψ angles can each take on four different values, this results in 16 possible unique color combinations, which are depicted at the right, with four distinct shadings indicating the four distinct possible combinations of φ and ψ dihedral angle stereochemistries.

### 5.3.3 Data analysis

The database analysis was completed using a SQL Server and the heat maps were generated by an in-house program written in C#. All β turns in the database were classified based on the new schema. Under the new classification, the heat maps were constructed to represent the distribution of φ/ψ angles for the i+2 residues relative to classified i+1 residues. To analyze possible consensus sequences in the new turn types, the occurrence of amino acids in each residue position were depicted using WebLogo[31-32]. Due to the lack of hydrogen atoms in the crystal structures, hydrogen bonds were qualitatively classified based on the following. If the distance between carbonyl oxygen of first residue and nitrogen of the last residue was less than 3.5 Å, the β turn was considered as having hydrogen bond. If the distance of those residues was less than 2.5 Å, the hydrogen bond was determined to be a strong hydrogen bond. If the distance was between 2.5 to 3.2 Å, the hydrogen bond was designated as a moderate hydrogen bond while if the distance was over 3.2 Å, it was designated as a weak hydrogen bond. To evaluate correlations with the ω turns, the trans ω turn, which has an ω angle close to 180°, was recognized as having a positive value, meaning that there was no significant influence on the new schema. The cis ω turns, with values close to 0° was determined as negative, meaning the new turn type has a sub-division of β turn types. Comparisons with the widely accepted classification scheme[6] was based classifications using Betaturn18. The new turn types belonging to two classic turn types were recognized as the existence of overlap.

## 5.4 Results and Discussion

### 5.4.1 Distribution of new β turn types in the superfamily of pentapeptide repeat proteins (PRPs)

We first demonstrate the utility of the new classification scheme by analyzing the β-turn space observed in the superfamily of pentapeptide repeat proteins (PRPs) [19-28]. Out of 1296 possible unique β turn classifications using the Klyne-Prelog-modified Ramachandran plot stereochemistry definitions, only 24 combinations were observed in the PRP structures currently submitted to the PDB, and these are depicted in **Figure 5.3** with each combination given a unique identifier specifying their stereochemistry combinations with the nomenclature based on the Klyne-Prelog stereochemistry definitions. Of the 24 observed combinations, the SC2-SC10 (synclinal range 2, synclinal range 10) combination was by far most commonly observed occurring in 237 out of 394 turns, i.e. 60.2% of the turns, followed by the AC2-SC9 (anticlinal rage 2, synclinal range 9) combination (80 out of 394 or 20.3% of turns) (**Table 5.1**). SC2-SC5 and AC2-SC10 occurred 18 and 17 times, respectively, corresponding to 4.6% and 4.3%, respectively (**Table 5.1**). The remaining 20 unique combinations occurred at frequencies of 2% or less (**Table 5.1**).

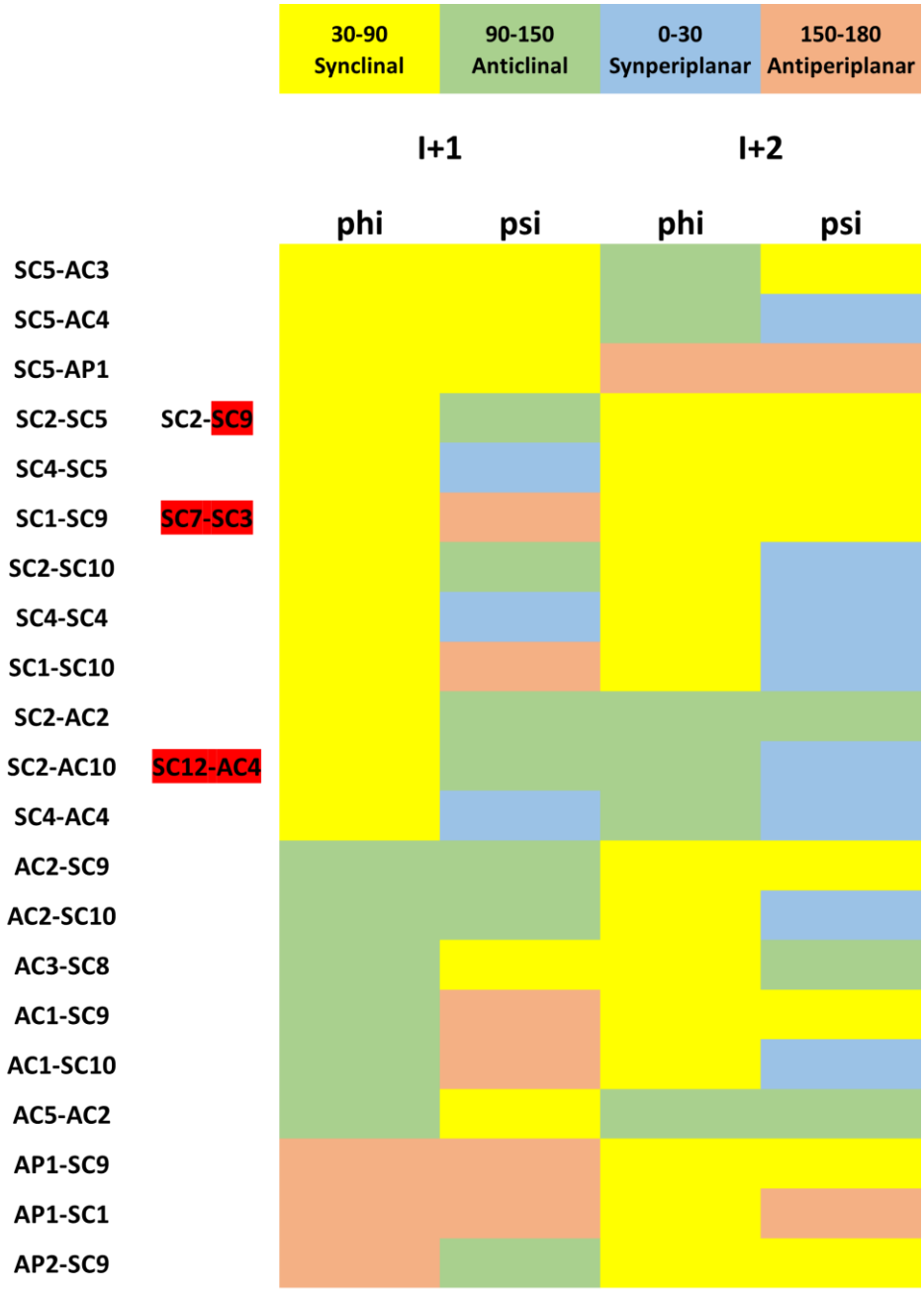



**Figure 5. 3 Graph depicting the 24 unique combinations of φ and ψ backbone dihedral angles observed in PRPs.** The four stereochemistry definitions were given unique colors: synclinal (yellow), anticlinal (green), synperiplanar (blue), and antiperiplanar (orange). The stereochemistry combinations of the four dihedral angles defining a β turn are indicated by four colored blocks, one each depicting the φ or ψ backbone dihedral angle for either the i+1 or i+2 residue in the turn. The one-letter code at the beginning of each combination name is determined by the stereochemistry definition of the i+1 φ dihedral angle. The number following the one-letter code simply indicates the number of the combination in the list of combinations starting with the same

one-letter code. Turns having the same stereochemistry for both the φ and ψ angles are grouped together at the top of the list.

| Combinations | Number | Percentage (%) |
|---|---|---|
| SC5-AC3 | 5 | 1.3 |
| SC5-AC4 | 2 | 0.5 |
| SC5-AP1 | 1 | 0.3 |
| SC2-SC5 | 18 | 4.6 |
| SC4-SC5 | 1 | 0.3 |
| SC2-SC10 | 237 | 60.2 |
| SC1-SC10 | 3 | 0.8 |
| SC2-AC2 | 1 | 0.3 |
| SC2-AC10 | 5 | 1.3 |
| SC4-AC4 | 2 | 0.5 |
| AC2-SC9 | 80 | 20.3 |
| AC2-SC10 | 17 | 4.3 |
| AC3-SC8 | 1 | 0.3 |
| AC1-SC10 | 1 | 0.3 |
| AC5-AC2 | 1 | 0.3 |
| AP1-SC9 | 1 | 0.3 |
| AP1-SC1 | 1 | 0.3 |
| AP2-SC9 | 4 | 1.0 |
| SC2-SC9 | 1 | 0.3 |
| SC1-SC9 | 3 | 0.8 |
| SC12-AC4 | 1 | 0.3 |
| AC1-SC9 | 4 | 1.0 |
| SC7-SC3 | 1 | 0.3 |
| SC4-SC4 | 3 | 0.8 |
| TOTAL | 394 | 100 |

**Table 5. 1 Summary of the number of occurrences of each type of stereochemistry combination of dihedral backbone angles observed in β turns in PRPs.**

All dihedral angle combinations observed for β turns in all known PRPs were plotted on a Klyne-Prelog-modified Ramachandran plot color-coded to indicate the stereochemistry-defined ranges in **Figure 5.4**. The plot also includes an overlay of the canonical type I and type II β turn ranges and the extended ranges for type I and type II β turn ranges [9]. Inspection of **Figure 5.4** reveals that a large majority the turns that belong to the SC2-SC10 group in the new classification scheme belong to the canonical or extended type II β turn categories, however, a large number of the i+1 residues of the turns fall just outside the extended range for type II for i+1 residues, falling into the AC2 region of the modified Ramachandran plot (**Figure 5.4**). By the canonical and extended definitions, these turns would be classified as type IV β turns. Analysis of the plot of the points in

**Figure 5.4** illustrates that the type IV classification is not informative in terms of the absolute stereochemistry of each of the dihedral angle combinations involved in the definition of the β turns. Also, many points lie just outside of a boundary region for either the canonical or extended regions, which could raise the question as to whether such points should be referred to as distorted or borderline type I or type II turns, or whether it is more instructive to use the catch-all type IV classification. The benefit of the new classification scheme based on classic stereochemistry definitions is that a precise, meaningful, unambiguous and definite classifications can be assigned to every point. Because the torsion angles can be calculated to high accuracy from high-resolution crystal structures, points falling exactly on a boundary edge leave their classification ambiguous based on the stereochemistry definitions should be rare. For example, no ambiguous points occurred our classification of β turns for all known PRP structures.

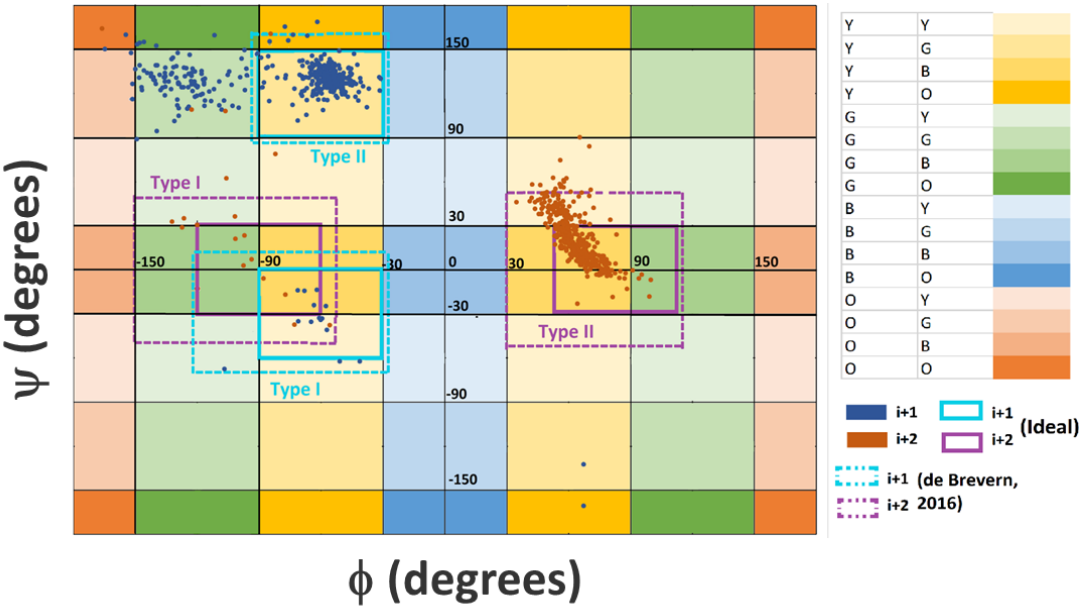


**Figure 5. 4 A Klyne-Prelog-modified Ramachandran plot depicting the distinct ranges defined by the stereochemistry definitions.** All β turns from all existing PRP structures submitted to the PDB are plotted on this Klyne-Prelog- modified Ramachandran plot. Values for i+1 residues are plotted with solid blue circles and those of i+2 residues are plotted with solid orange circles. The canonical dihedral angle ranges for φ and ψ for type I and type II β turns are indicated by solid sky-blue and purple lines for the i+1 and i+2 residues, respectively. The extended ranges defined by de Brevern are indicated by dashed sky-blue and purple lines for the i+1 and i+2 residues, respectively. All measured backbone φ and ψ dihedral angles measured from all PRPs are indicated in blue and orange for the i+1 and i+2 residues, respectively. The color scheme for the plot is explained in the legend to the right of the plot. The first column indicates the Klyne-Prelog region for the i+1 residue and the second column indicates the Klyne-Prelog region for the i+2 residue. The single-letter codes are consistent with those defined in Figure 5.3: Y - synclinal, 30º -90º; G - anticlinal, 90º - 150º; B - synperiplanar, 0º -30º; O - antiperiplanar, 150º -180 º.

To illustrate the utility of this new scheme, we present a more detailed analysis of the β turns in the 2XTW PRP structure that contains the largest number of β turns (**Figure 5.5A**). As can be seen, a majority of the β turns (19 out of 33) fall into the canonical (SC2-SC10 andSC2-AC10) or extended (SC2-SC9) type II β turn category, however, at least a dozen turns are nearby but outside these canonical and extended regions, which are identified at SC1-SC10, SC1-SC9,AC2-SC10, and AC2-SC9, along with several other i+1 or i+2 dihedral angles falling in the AC2 region (**Figure 5.5A**). Another cluster of points are in the area of the Ramachandran space near the type I β turn region, however, closer inspection of these points indicates that in most cases the i+1 and paired i+2 points do not simultaneously fall in the canonical or extended regions, e.g. the points labeled SC12-AC4, SC4-SC4, SC5-AC4, SC5-AC3, and SC4-SC5 (**Figure 5.5A**). This leaves only one turn in this vicinity of the Ramachandran space that conforms to a canonical type I β turn, i.e. SC4-SC4. This exercise illustrates that despite only one out of nine φ/ψ dihedral angle pairs lying in the vicinity of the canonical type l β turn region of the Ramachandran space, this new system allows us to give a unique and informative classification to each φ/ψ dihedral angle pair. A similar analysis has been performed for all of the known PRP crystal structures and the results are included in the supplementary material (**Figure 5.6-4.20**).

To investigate the β-turn space occupied by the 13 different β turn classifications observed in the 2XTW structure, we depicted one representative example of a turn from each category in **Figure 5.5B**. Perhaps of most interest in this collection is analysis of the strongest outliers from the canonical type I and type II β turn ranges. For example, inspection of the SC12-AC4 turn reveals that it adopts a quite regular looking β turn appearance, however, if one considers the symmetry in the Ramachandran space, the SC12 region has 180º rotation symmetry with the SC2 space, i.e. the SC12 φ/ψ angles are +30º to +90º / -90º to -150º whereas the SC2 φ/ψ angles are -30º to -90º / +90º to +150º. Inspection of the representative β turns for SC2 and SC12 illustrates that the two turns are related by a 180º symmetry based on their respective φ and ψ angles.

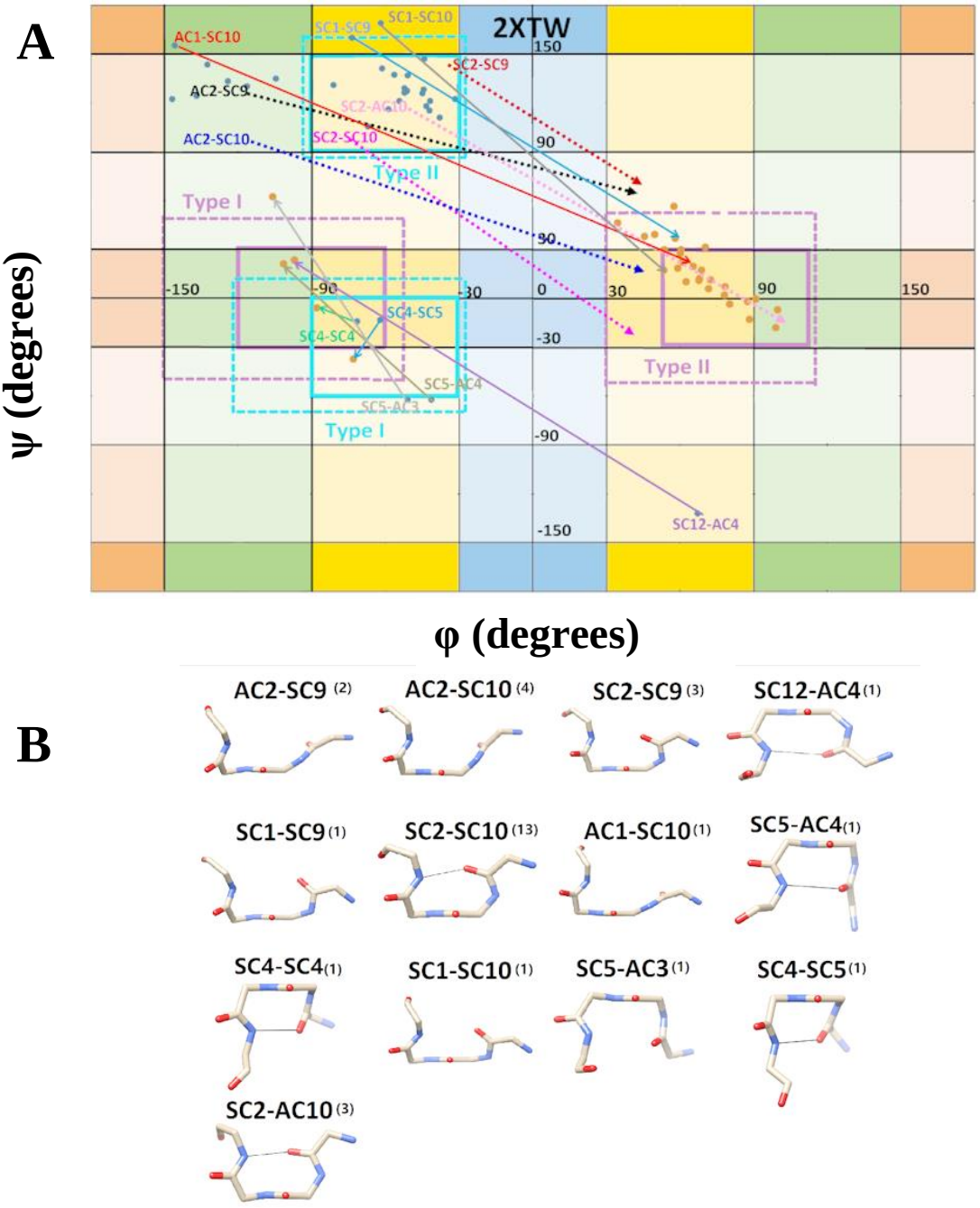



**Figure 5. 5 A Klyne-Prelog-modified Ramachandran plot analysis of the β turns in the 2XTW PRP structure.** A) A modified Ramachandran plot of all φ and ψ backbone dihedral angles for i+1 and i+2 residues in β turns in the 2XTW PRP structure. The base of each arrow joining two points indicates the values of the i+1 residue and the head of each arrow joining two points indicates the values for the i+2 residue. Arrows not joining a pair of points indicates the connection between two general stereochemistry combinations specified by the label definitions given in **Figure 5.2**. B) One representative β turn is depicted for each stereochemistry combination observed in the structure. The number in parentheses indicates the number of occurrences in the 2XTW structure.

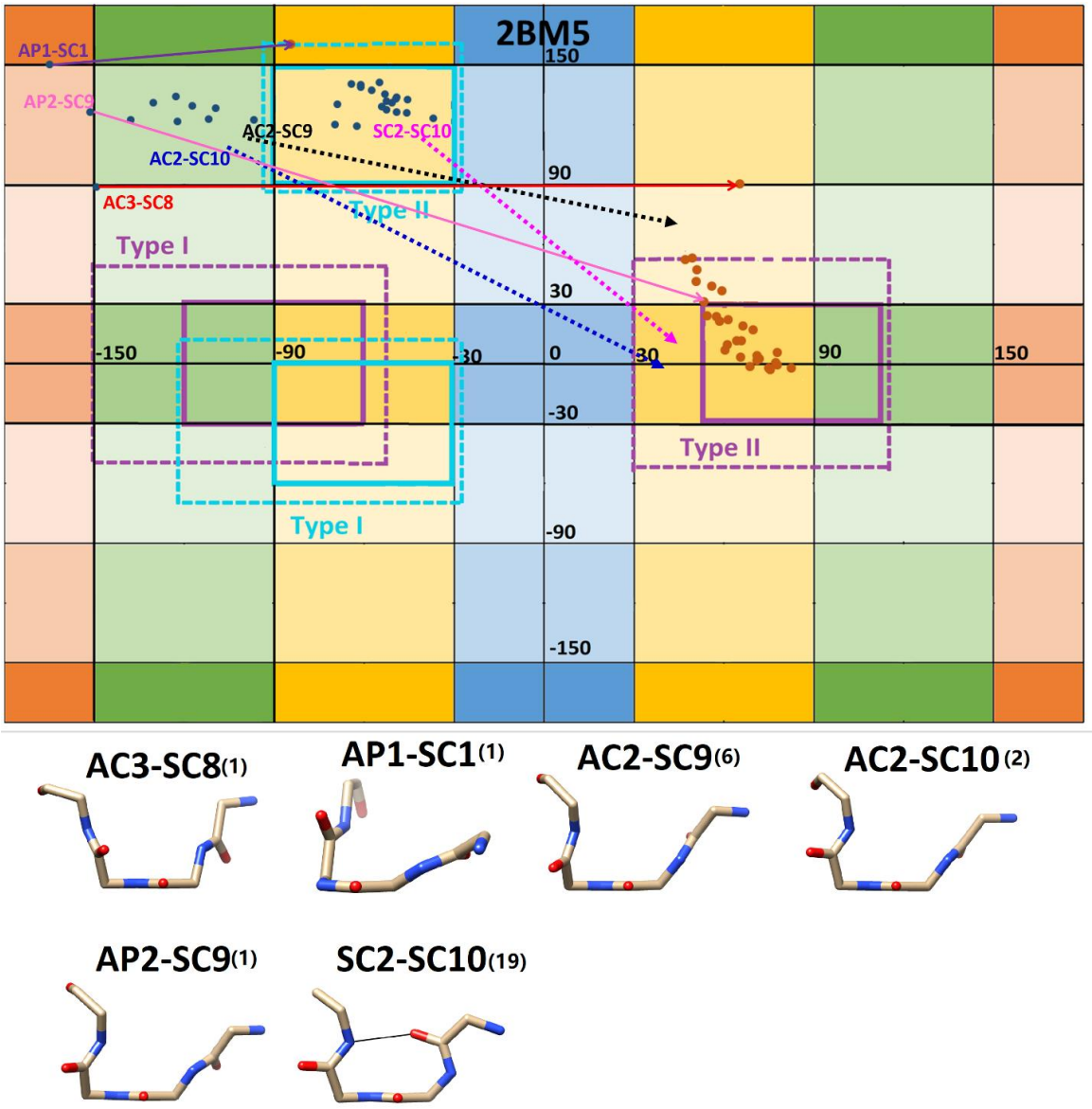



**Figure 5. 6 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2BM5 PRP structure.**

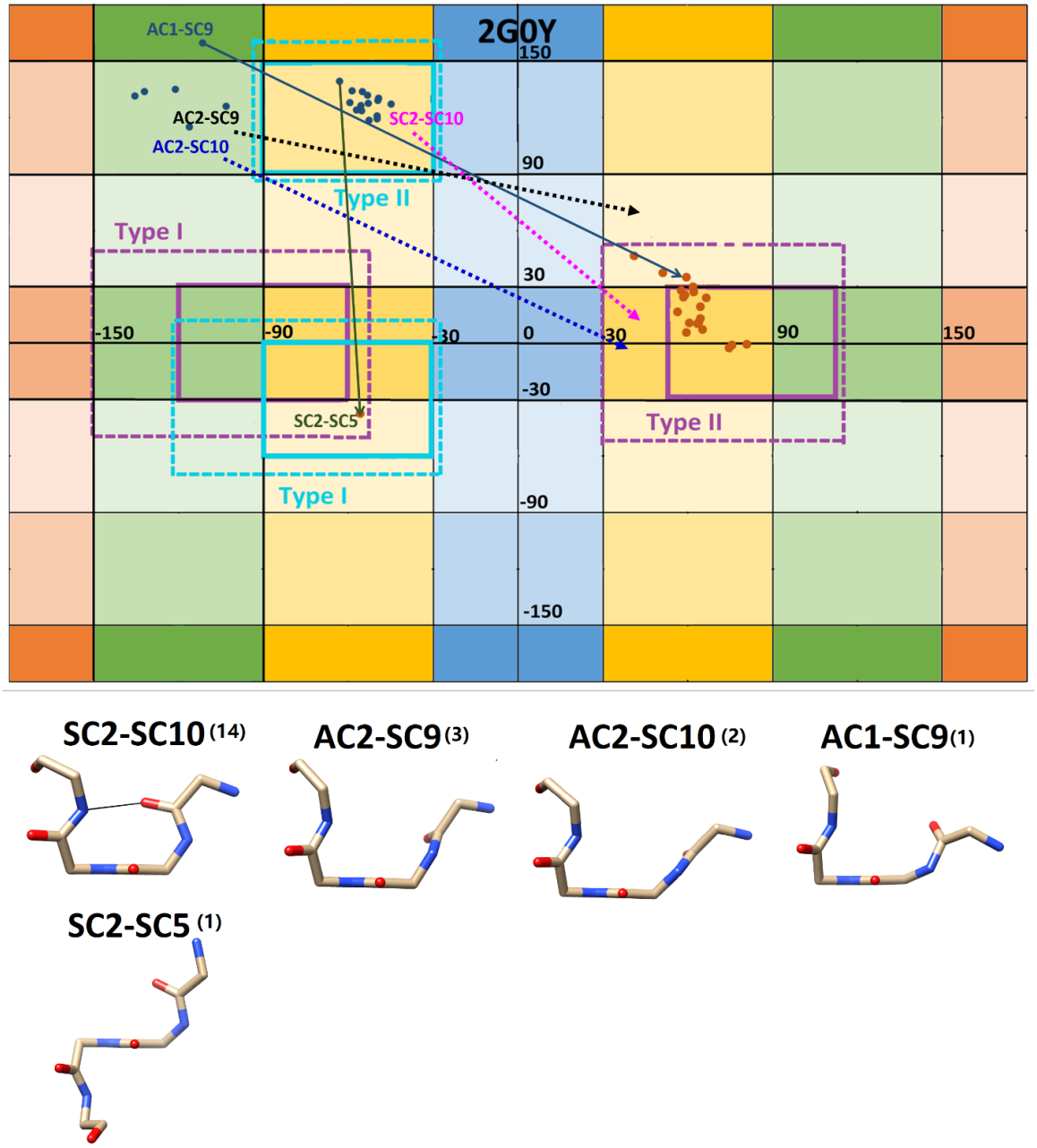



**Figure 5. 7 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2G0Y PRP structure.**

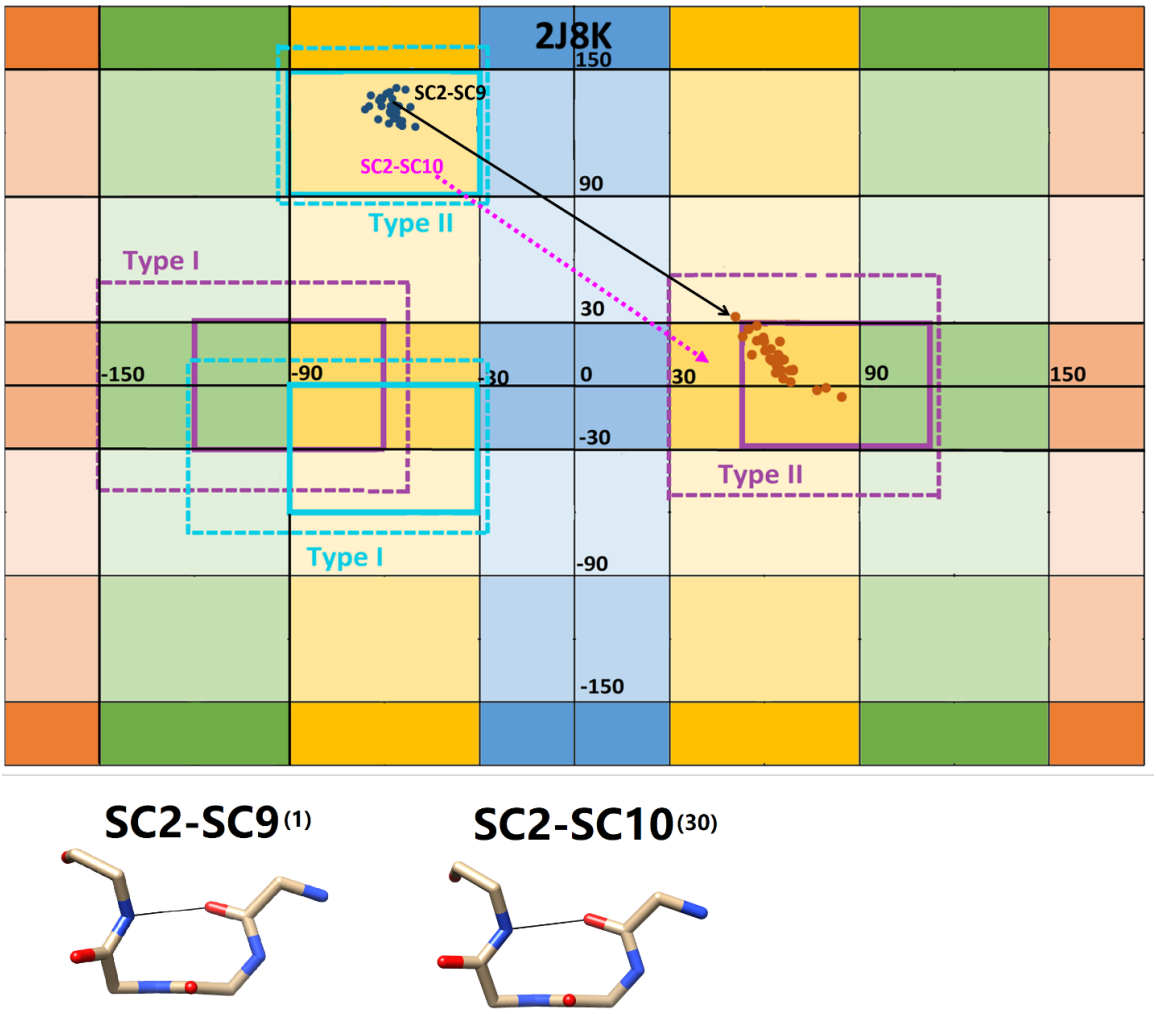



**Figure 5. 8 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2J8K PRP structure.**

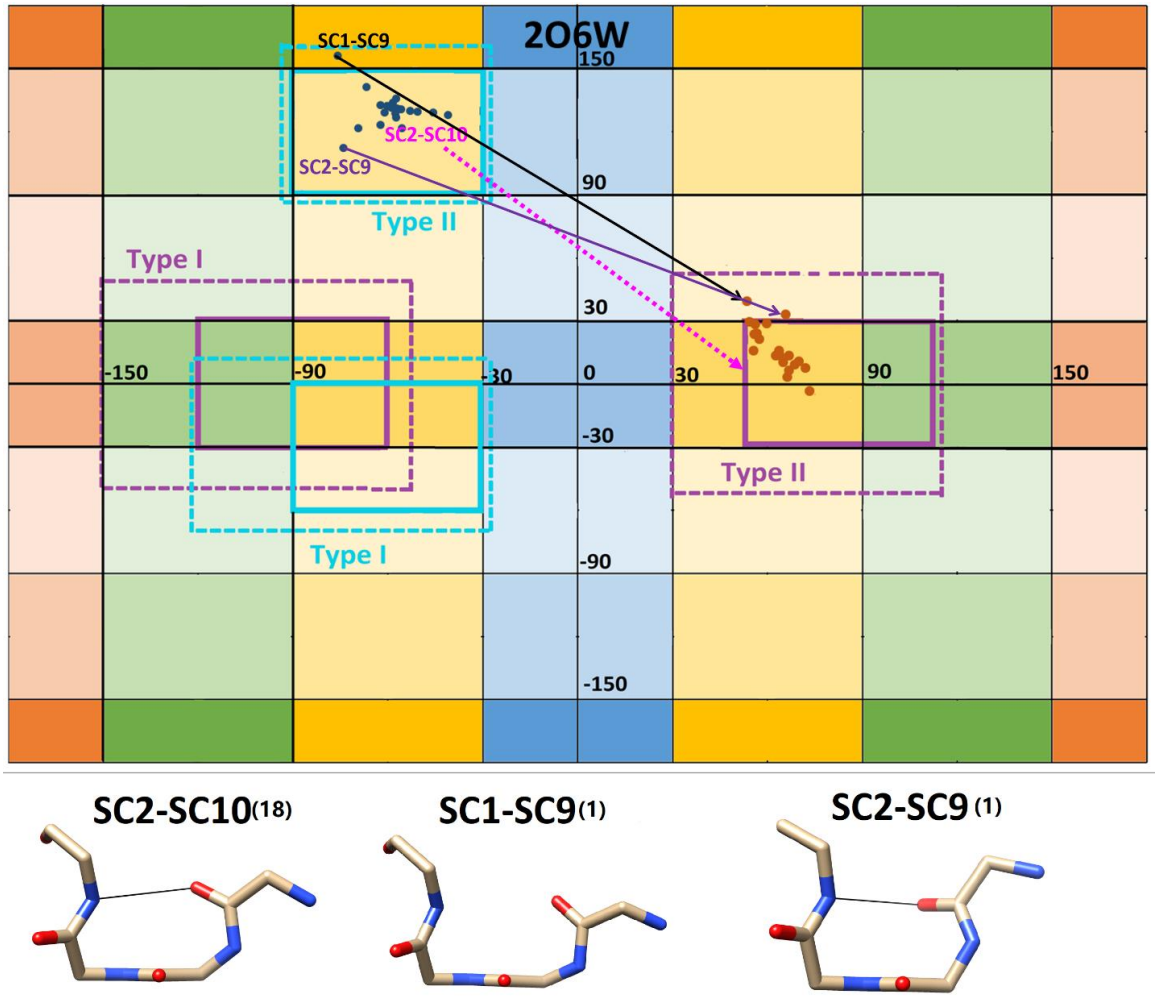



**Figure 5. 9 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2O6W PRP structure.**

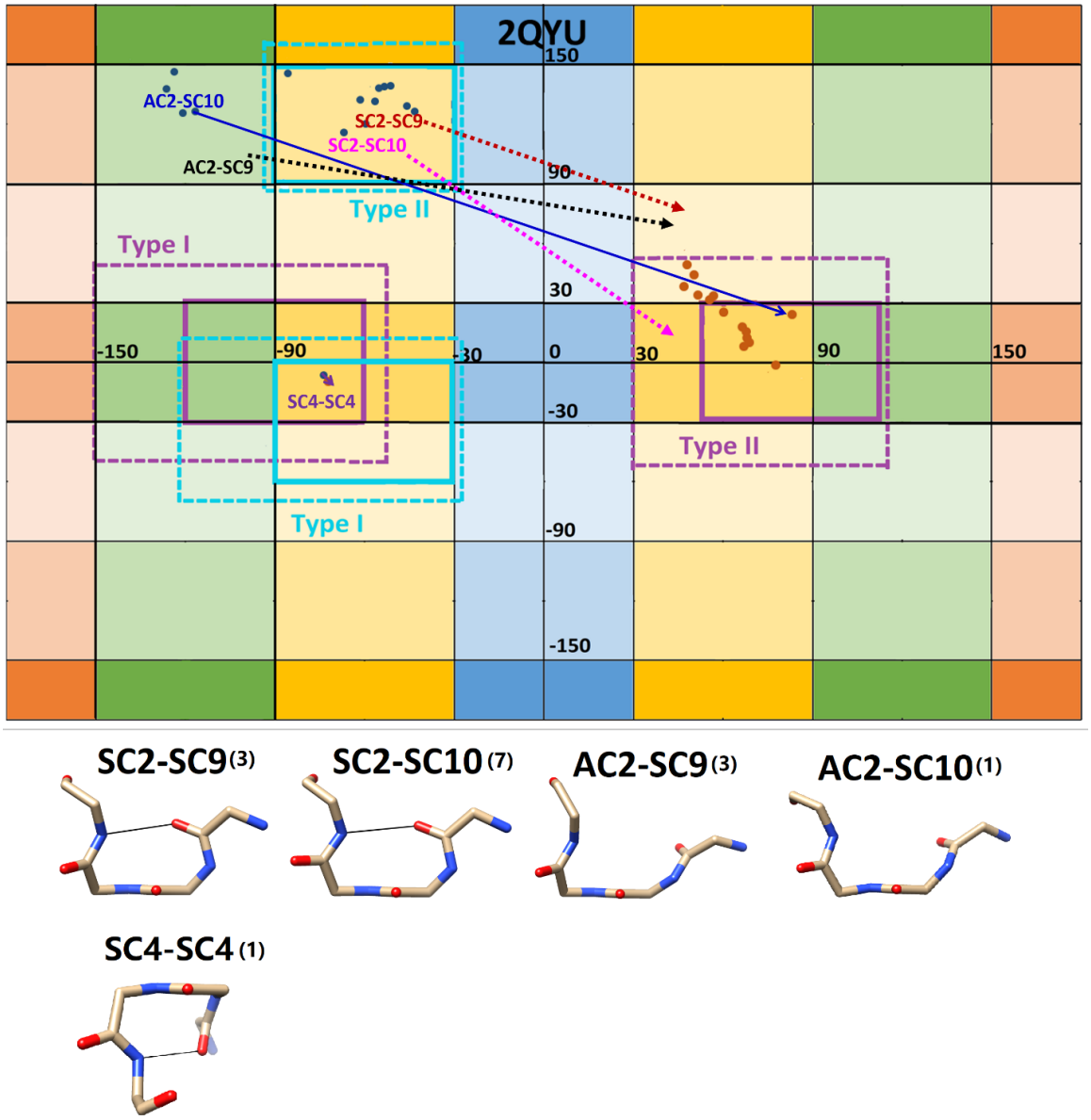



**Figure 5. 10 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2QYU PRP structure.**

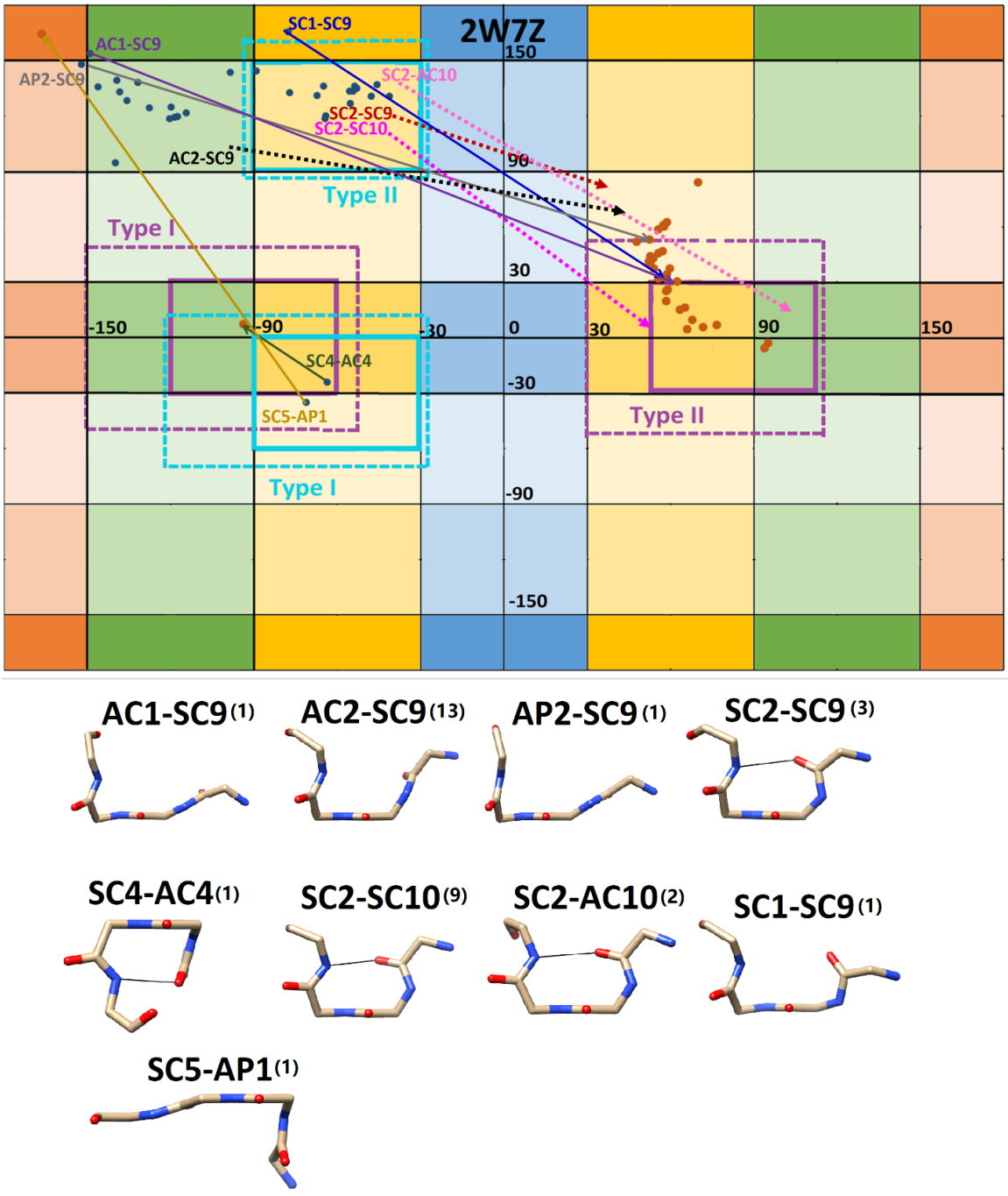



**Figure 5. 11 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2W7Z PRP structure.**

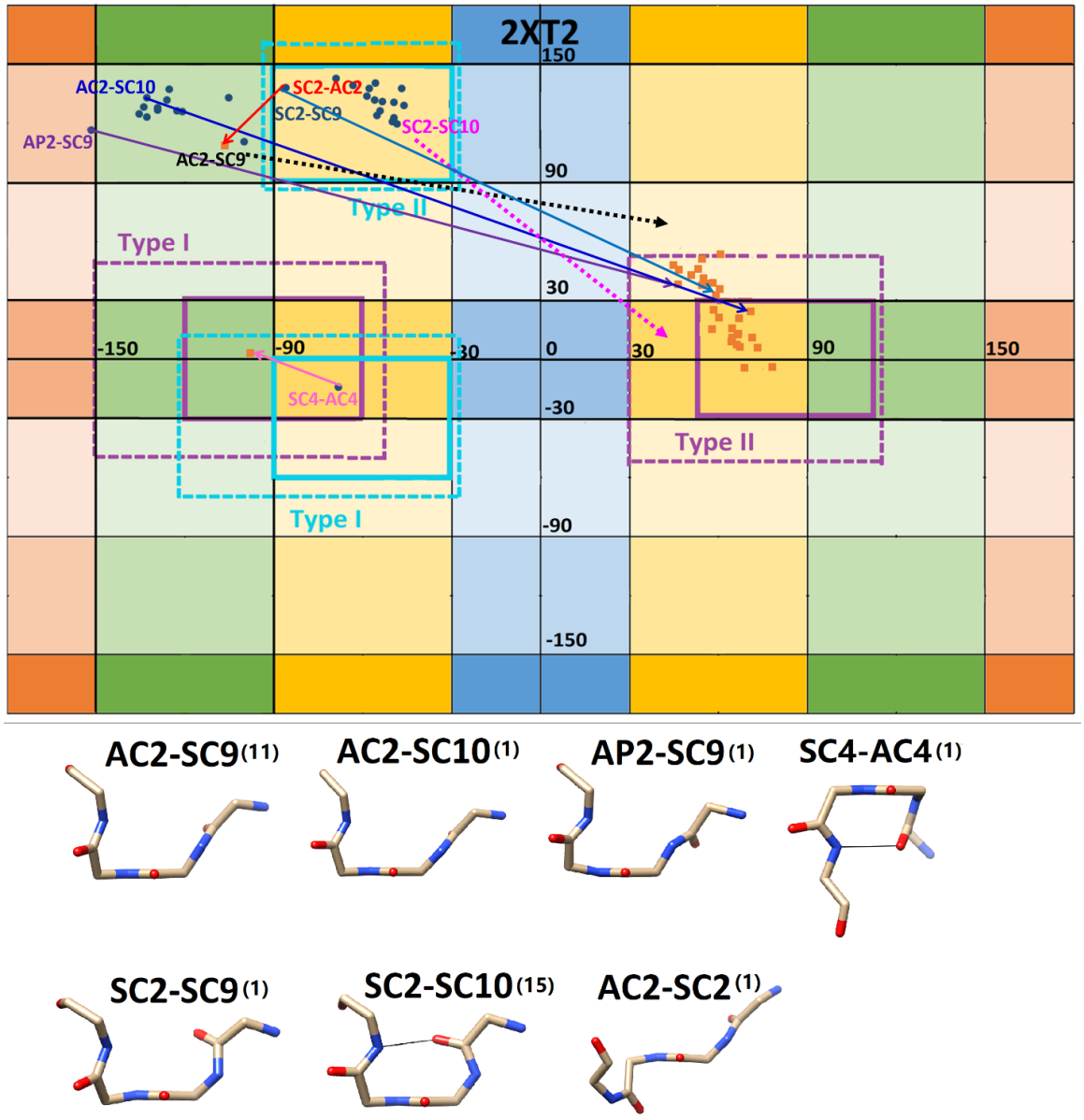



**Figure 5. 12 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 2XT2 PRP structure.**

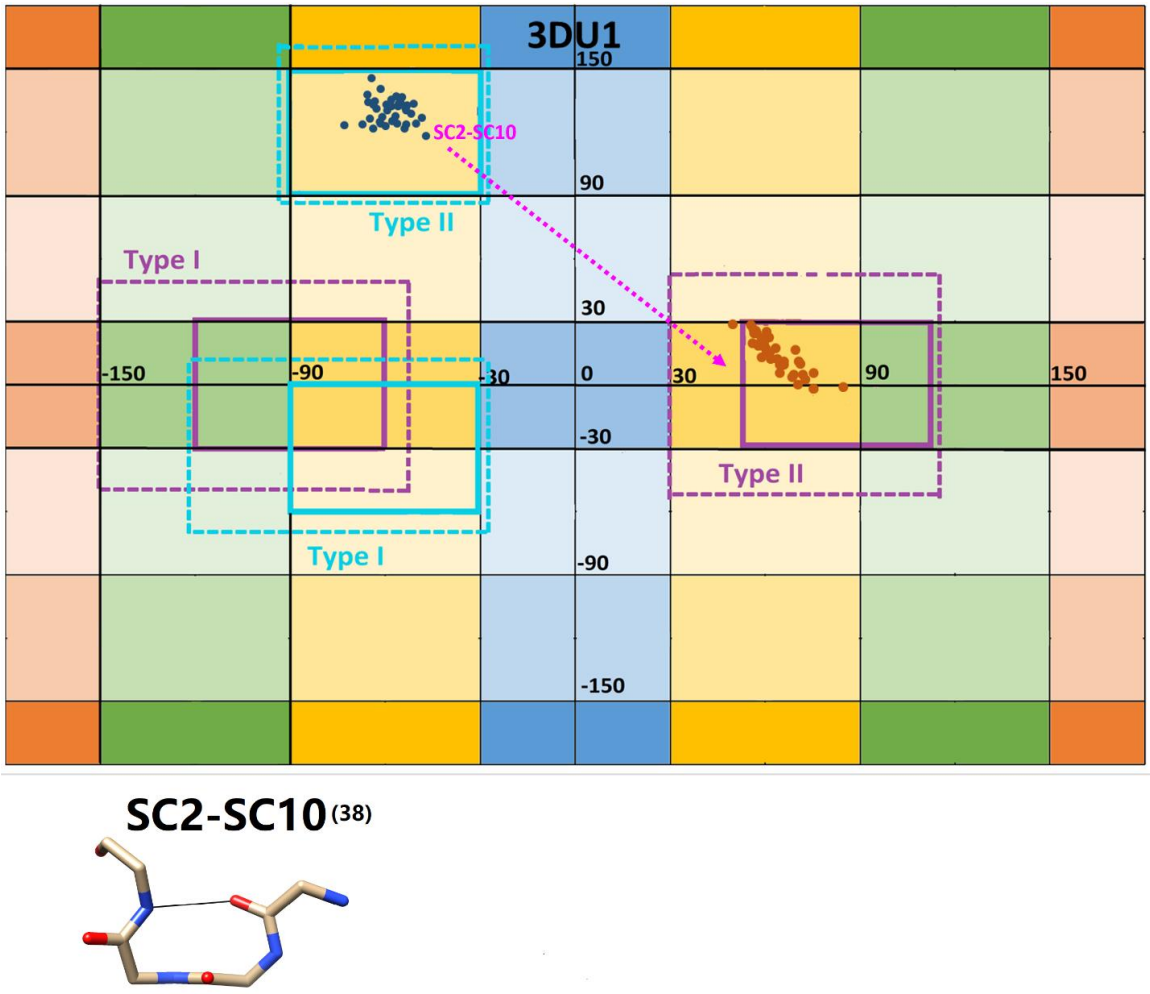



**Figure 5. 13 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 3DU1 PRP structure.**

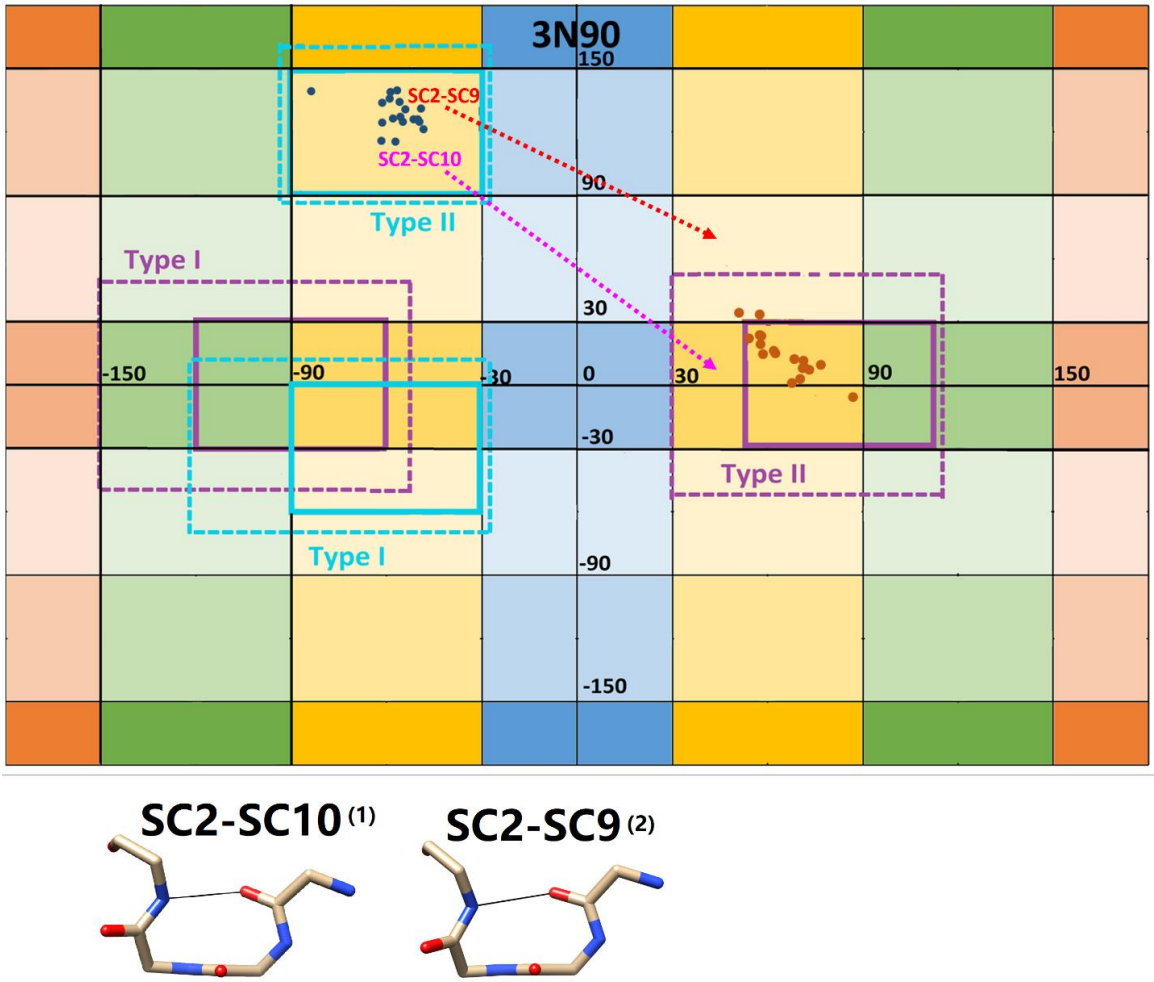



**Figure 5. 14 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 3N90 PRP structure.**

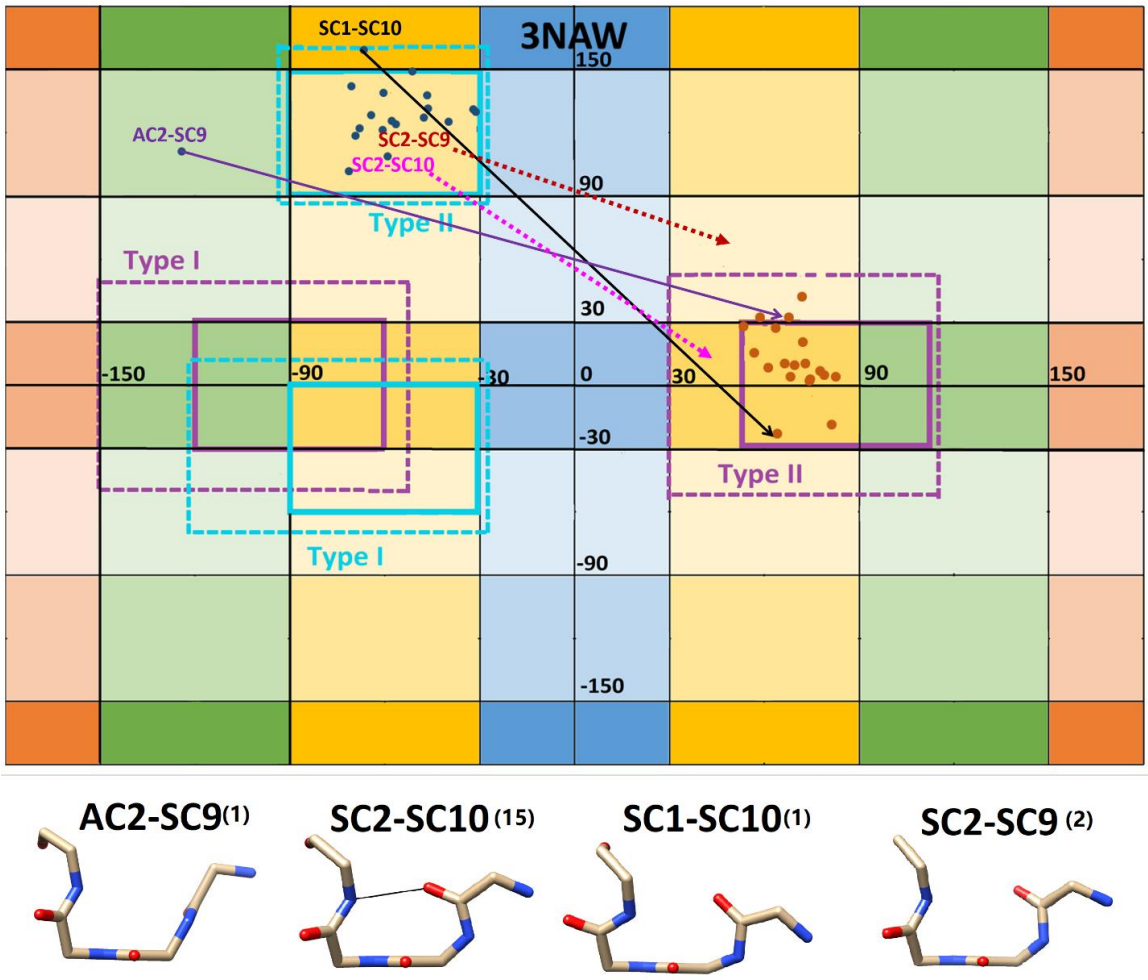



**Figure 5. 15 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 3NAW PRP structure.**

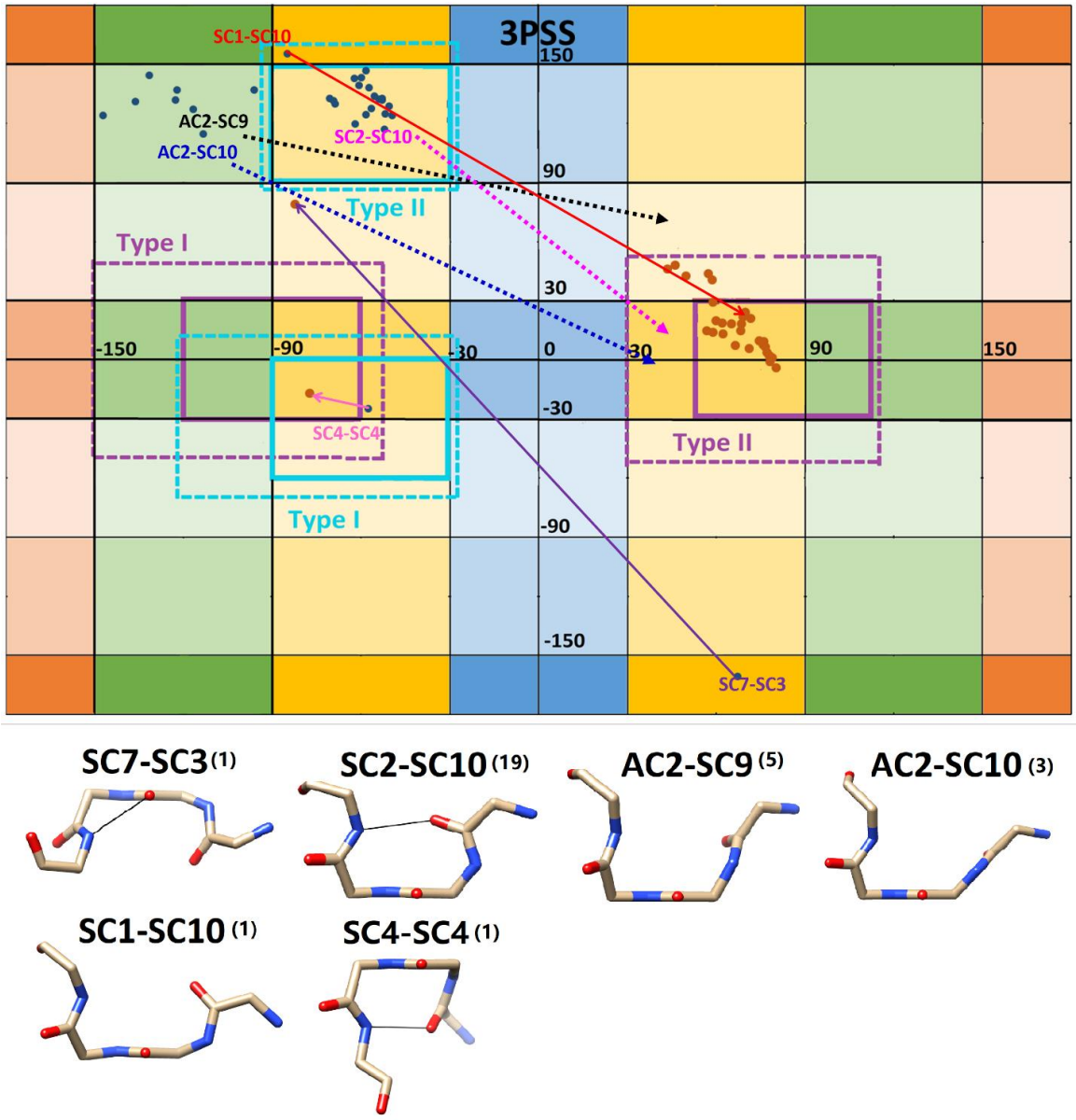



**Figure 5. 16 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 3PSS PRP structure.**

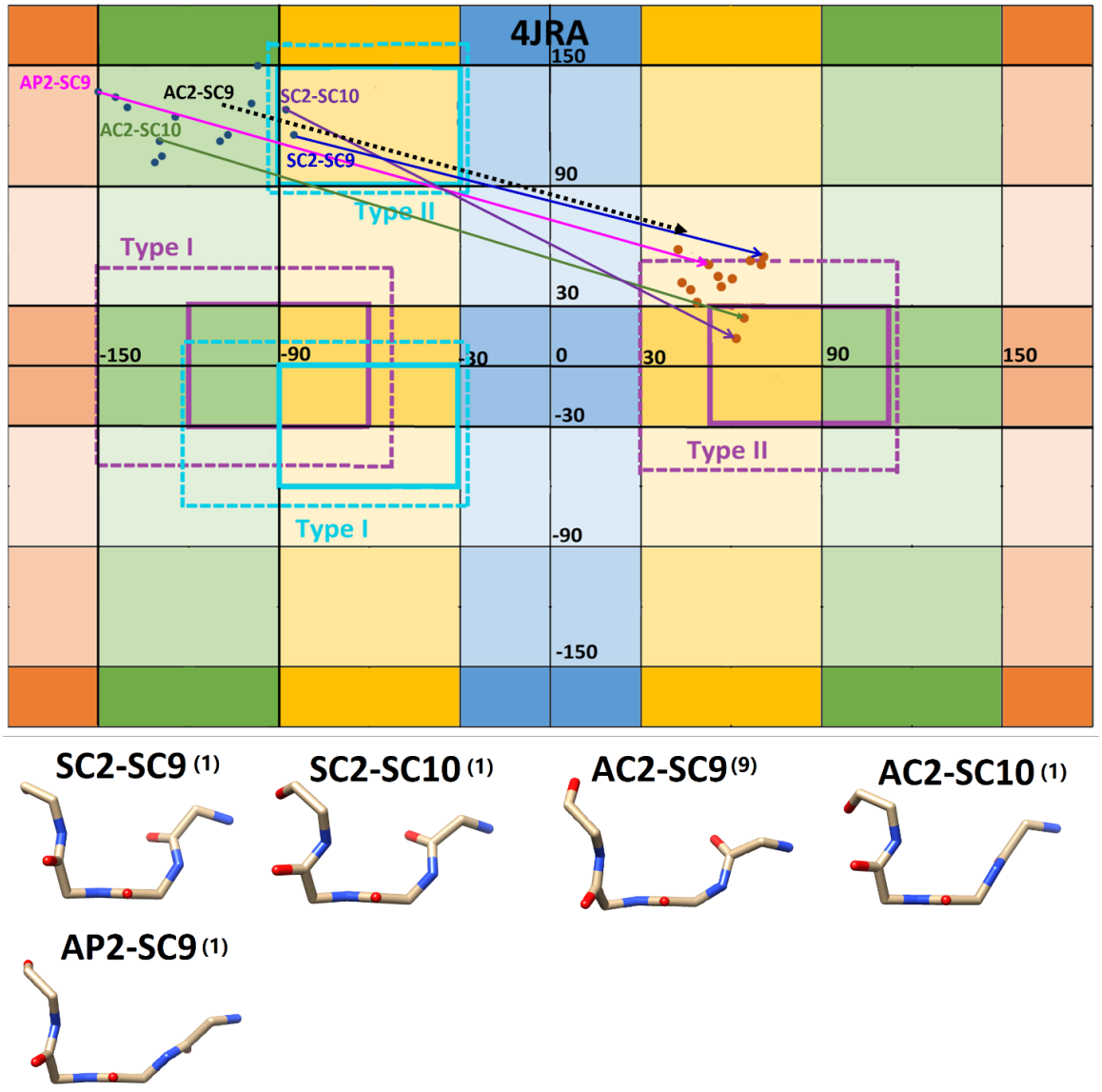



**Figure 5. 17 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 4JRA PRP structure.**

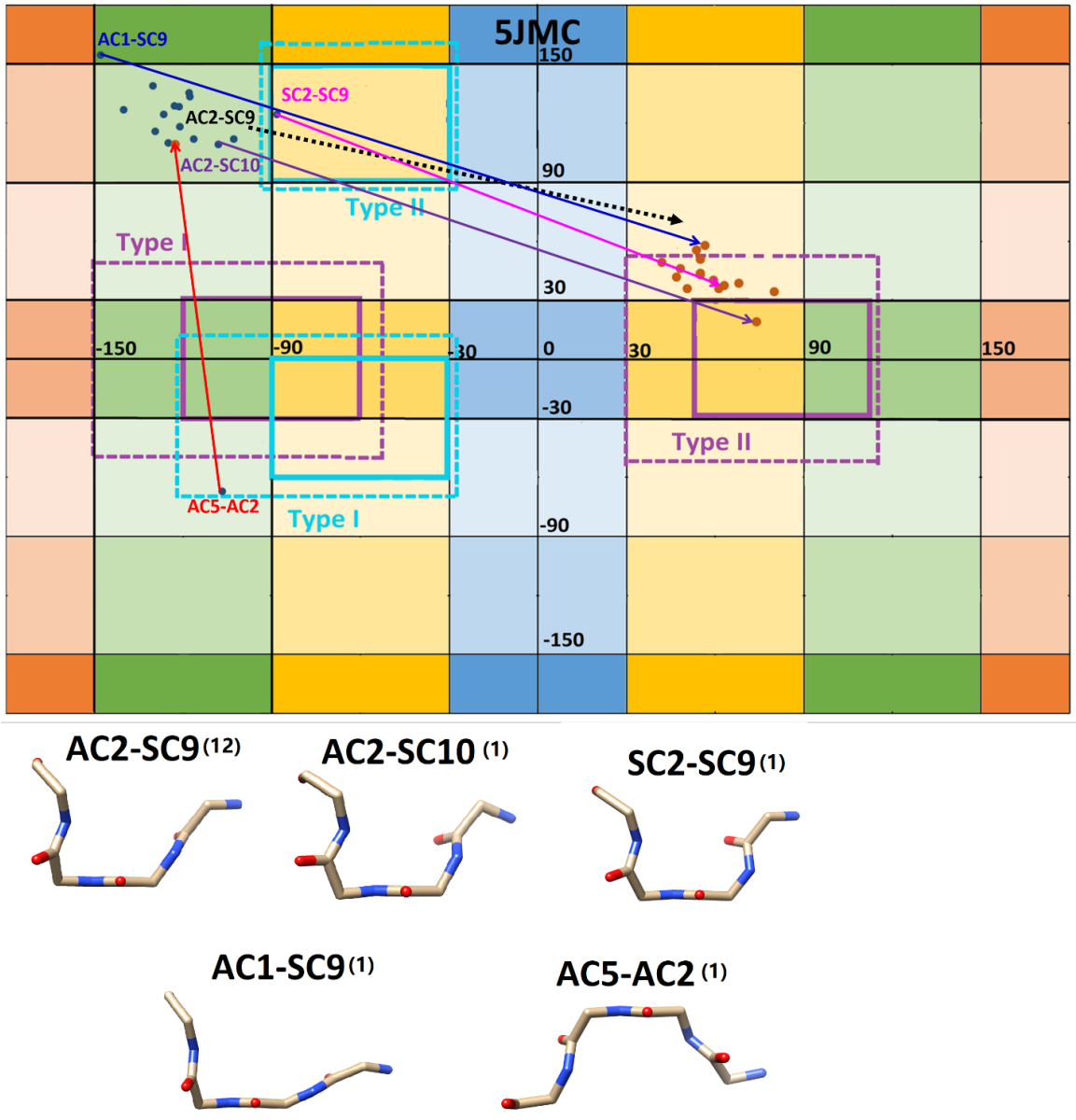



**Figure 5. 18 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 5JMC PRP structure.**

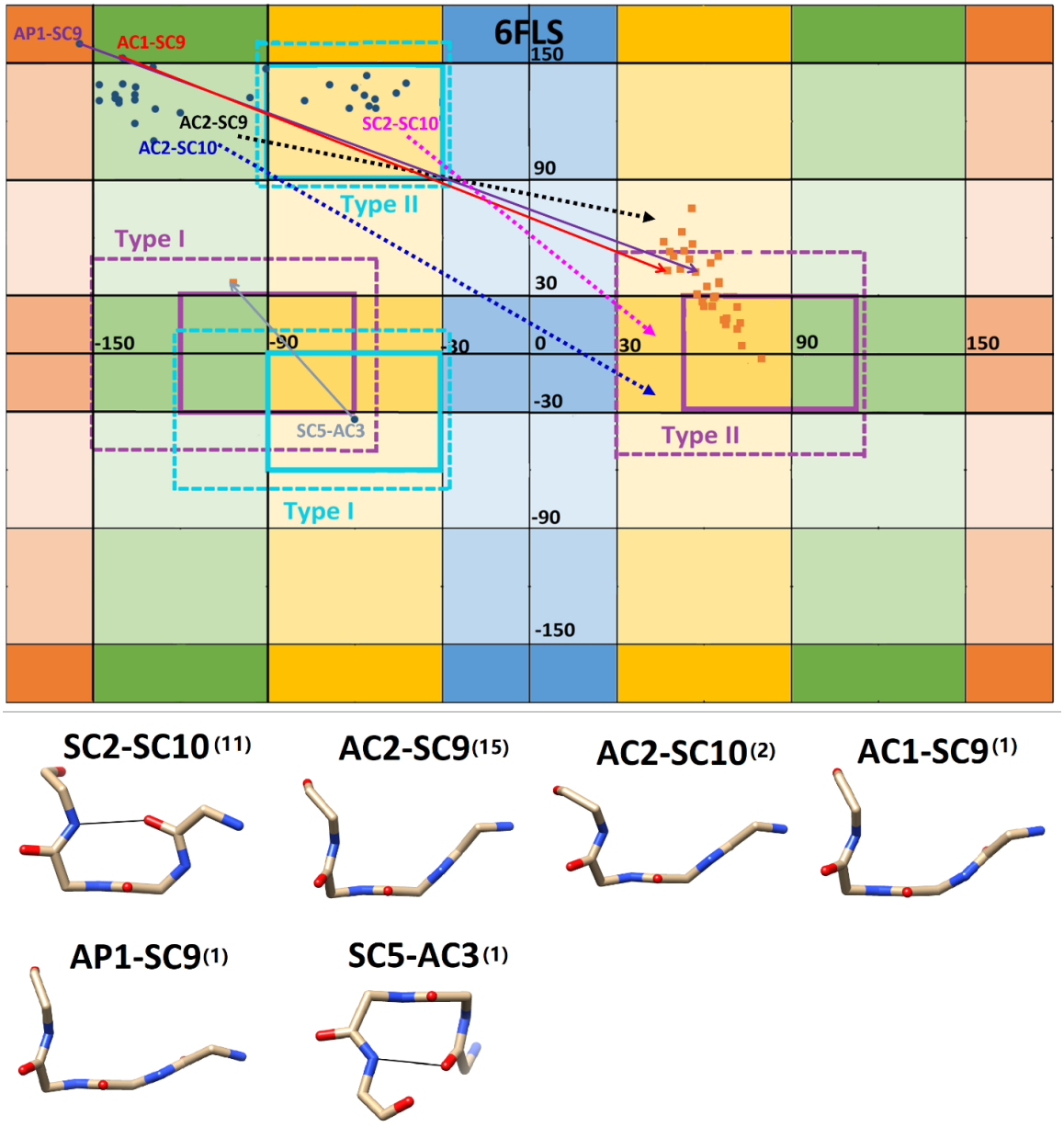



**Figure 5. 19 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 6FLS PRP structure.**

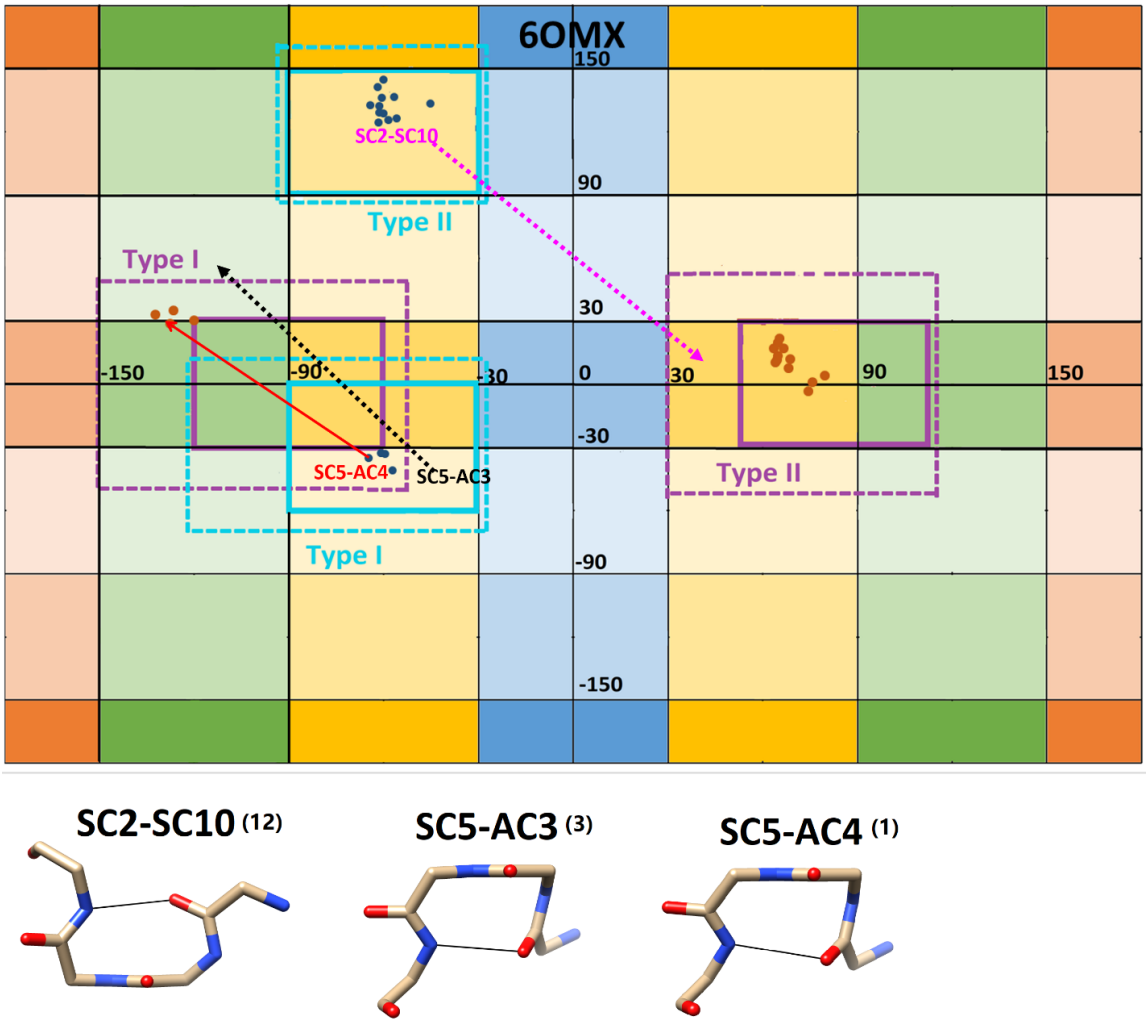



**Figure 5. 20 A modified Klyne-Prelog Ramachandran plot analysis of the β turns and representatives of β turns for each stereochemistry combination observed in the 6OMX PRP structure.**

### 5.4.2 Distribution of the new β turn types in the database of 16657 high-resolution protein structures and 522776 β turns

Our analysis of the large β turn database indicated that 582 new turns were populated out of 1296 possible distinct β turn types (i.e. ~44.9%). Complete information for the distribution of new turn types is available in **Table 5.2**. The top 19 turn types, which contained ~ 79.1% of the turns in the database, are listed in **Table 5.3**. The distributions of turn types were represented by heat maps plotted onto the Ramachandran space for the i+1 and i+2 residues in **Figure 5.21**. Inspection of these heat maps indicated that β turns were found in all 36 Klyne-Prelog sections of the Ramachandran space for both the i+1 and i+2 residues. Some interesting patterns were evident in the heat maps, for example, the AC1-AC6 and SC1-SC6 regions were significantly more populated at the i+1 position in comparison to the i+2 position. Also, the SC9-SC11 and AC9-AC11 regions were much more heavily populated at the i+2 position in comparison to the i+1 position.

Finally, the SC12 and bordering SC7 and AC12 and bordering AC7 were much more heavily populated at the i+2 position in comparison to at the i+1 position.

| Category | Count | Percentage (%) |
|---|---|---|
| [SC4]-[SC4] | 84742 | 16.21000199 |
| [SC4]-[AC4] | 81476 | 15.58526023 |
| [SC5]-[SC4] | 55774 | 10.66881418 |
| [SC2]-[SC10] | 35628 | 6.815156013 |
| [SC5]-[AC4] | 24738 | 4.732045848 |
| [SC5]-[SC5] | 16297 | 3.117396361 |
| [SC9]-[SC10] | 14136 | 2.704026199 |
| [SC2]-[AC10] | 13867 | 2.652570126 |
| [SC4]-[SC5] | 13297 | 2.543536811 |
| [SC5]-[AC2] | 10005 | 1.913821599 |
| [SC4]-[AC3] | 9979 | 1.908848149 |
| [SC12]-[SC4] | 8541 | 1.633778138 |
| [SC4]-[AC2] | 8453 | 1.616944925 |
| [AC4]-[AC4] | 7767 | 1.485722374 |
| [SC4]-[AC5] | 6443 | 1.232459026 |
| [SC2]-[SC9] | 6368 | 1.218112538 |
| [SC5]-[AC3] | 5525 | 1.056858004 |
| [SC12]-[AC4] | 5339 | 1.021278712 |
| [SC4]-[AC1] | 5094 | 0.974413516 |
| [SC5]-[AC1] | 4541 | 0.868632072 |
| [AC2]-[SC9] | 4230 | 0.809141965 |
| [AC4]-[AC3] | 4180 | 0.799577639 |
| [SC10]-[SC10] | 3768 | 0.720767595 |
| [SC1]-[SC9] | 3498 | 0.669120235 |
| [AC2]-[SC1] | 2876 | 0.550140022 |
| [AC5]-[AC4] | 2758 | 0.527568213 |
| [AC4]-[AC5] | 2569 | 0.491415061 |
| [AC5]-[SC5] | 2403 | 0.459661499 |
| [SC9]-[AC10] | 2378 | 0.454879336 |
| [SC5]-[AC5] | 2271 | 0.434411679 |
| [SC4]-[AP1] | 2188 | 0.418534898 |
| [SC1]-[SC10] | 2100 | 0.401701685 |

| | | |
|---|---|---|
| [AC5]-[SC4] | 2007 | 0.383912039 |
| [AC4]-[AC2] | 1987 | 0.380086308 |
| [SC10]-[AC5] | 1987 | 0.380086308 |
| [AC1]-[SC9] | 1959 | 0.374730286 |
| [SC5]-[SC2] | 1861 | 0.355984207 |
| [AC4]-[SC4] | 1847 | 0.353306196 |
| [SC10]-[AC4] | 1843 | 0.35254105 |
| [AC4]-[SC5] | 1795 | 0.343359297 |
| [SC9]-[SC9] | 1720 | 0.329012809 |
| [SC4]-[SC3] | 1650 | 0.315622752 |
| [SC2]-[AC4] | 1522 | 0.291138078 |
| [AC2]-[SC10] | 1516 | 0.289990359 |
| [AC5]-[AC2] | 1380 | 0.263975393 |
| [SC10]-[SC9] | 1321 | 0.252689488 |
| [SC4]-[SC2] | 1207 | 0.230882826 |
| [SC2]-[SC4] | 1173 | 0.224379084 |
| [SC4]-[AP2] | 1054 | 0.201615988 |
| [AC10]-[AC4] | 1045 | 0.19989441 |
| [AC4]-[AC10] | 949 | 0.181530904 |
| [AC3]-[AC10] | 938 | 0.179426753 |
| [SC10]-[AC10] | 884 | 0.169097281 |
| [SC5]-[AP1] | 866 | 0.165654123 |
| [SC5]-[AP2] | 856 | 0.163741258 |
| [SC5]-[SC3] | 848 | 0.162210966 |
| [SC12]-[AC3] | 806 | 0.154176932 |
| [AC5]-[AC3] | 788 | 0.150733775 |
| [AC2]-[SC4] | 749 | 0.143273601 |
| [SC10]-[AC3] | 748 | 0.143082314 |
| [SC7]-[SC4] | 744 | 0.142317168 |
| [AC4]-[AC1] | 743 | 0.142125882 |
| [AC6]-[SC5] | 692 | 0.132370269 |
| [AC1]-[SC10] | 689 | 0.13179641 |
| [SC4]-[AP3] | 672 | 0.128544539 |
| [AC2]-[SC2] | 644 | 0.123188517 |
| [AC6]-[SC4] | 615 | 0.117641208 |

| | | |
|---|---|---|
| [SC12]-[SC2] | 605 | 0.115728343 |
| [SC4]-[AC6] | 578 | 0.110563607 |
| [AC4]-[AP1] | 575 | 0.109989747 |
| [AC2]-[AC4] | 547 | 0.104633725 |
| [AC4]-[AP2] | 546 | 0.104442438 |
| [AC5]-[SC2] | 521 | 0.099660275 |
| [SC2]-[SC2] | 520 | 0.099468989 |
| [AC10]-[AC5] | 519 | 0.099277702 |
| [AP2]-[SC1] | 519 | 0.099277702 |
| [SC1]-[SC4] | 496 | 0.094878112 |
| [SC2]-[AC11] | 464 | 0.088756944 |
| [AC3]-[SC10] | 456 | 0.087226652 |
| [SC2]-[SC11] | 453 | 0.086652792 |
| [AC3]-[SC9] | 431 | 0.082444489 |
| [SC11]-[AC4] | 427 | 0.081679343 |
| [SC12]-[SC5] | 413 | 0.079001331 |
| [AC5]-[SC1] | 393 | 0.075175601 |
| [AC4]-[AP3] | 389 | 0.074410455 |
| [SC7]-[SC5] | 381 | 0.072880163 |
| [SC5]-[AP3] | 379 | 0.07249759 |
| [SC4]-[AC10] | 365 | 0.069819579 |
| [SC5]-[SC1] | 359 | 0.068671859 |
| [AC4]-[AC6] | 357 | 0.068289286 |
| [SC10]-[SC5] | 321 | 0.061402972 |
| [AC11]-[AC4] | 312 | 0.059681393 |
| [AC10]-[SC4] | 305 | 0.058342388 |
| [AC1]-[SC2] | 295 | 0.056429522 |
| [AC5]-[AC5] | 276 | 0.052795079 |
| [SC1]-[SC2] | 273 | 0.052221219 |
| [AC4]-[SC3] | 271 | 0.051838646 |
| [SC7]-[SC3] | 266 | 0.050882213 |
| [AC1]-[SC1] | 254 | 0.048586775 |
| [SC12]-[SC3] | 254 | 0.048586775 |
| [AC5]-[SC3] | 248 | 0.047439056 |
| [SC5]-[AC6] | 227 | 0.043422039 |

| | | |
|---|---|---|
| [SC4]-[SC1] | 225 | 0.043039466 |
| [AP2]-[SC9] | 223 | 0.042656893 |
| [SC1]-[AC10] | 223 | 0.042656893 |
| [SC7]-[SC2] | 223 | 0.042656893 |
| [SC2]-[AC3] | 214 | 0.040935315 |
| [AC12]-[SC5] | 212 | 0.040552742 |
| [SC11]-[SC5] | 200 | 0.038257303 |
| [AP6]-[SC5] | 199 | 0.038066017 |
| [AC10]-[AP1] | 182 | 0.034814146 |
| [AC2]-[SC7] | 178 | 0.034049 |
| [SC9]-[SC11] | 174 | 0.033283854 |
| [AC3]-[AC4] | 171 | 0.032709994 |
| [SC2]-[SC12] | 169 | 0.032327421 |
| [AC7]-[SC4] | 168 | 0.032136135 |
| [SC11]-[SC4] | 168 | 0.032136135 |
| [SC1]-[SC11] | 166 | 0.031753562 |
| [AC10]-[AC3] | 160 | 0.030605843 |
| [SC11]-[AC1] | 159 | 0.030414556 |
| [SC12]-[AC2] | 157 | 0.030031983 |
| [AC4]-[SC10] | 156 | 0.029840697 |
| [AC2]-[SC12] | 152 | 0.029075551 |
| [AC5]-[AC1] | 148 | 0.028310404 |
| [AC8]-[SC10] | 147 | 0.028119118 |
| [AC2]-[SC11] | 146 | 0.027927831 |
| [AC6]-[SC2] | 146 | 0.027927831 |
| [AC6]-[AC4] | 143 | 0.027353972 |
| [SC3]-[SC9] | 142 | 0.027162685 |
| [SC11]-[AC2] | 141 | 0.026971399 |
| [AC7]-[SC5] | 137 | 0.026206253 |
| [AP2]-[SC2] | 135 | 0.02582368 |
| [SC2]-[AC12] | 132 | 0.02524982 |
| [SC2]-[SC7] | 128 | 0.024484674 |
| [SC10]-[AC6] | 124 | 0.023719528 |
| [AP2]-[SC10] | 121 | 0.023145669 |
| [SC10]-[AP3] | 117 | 0.022380522 |

| | | |
|---|---|---|
| [SC3]-[SC10] | 116 | 0.022189236 |
| [AC3]-[SC1] | 110 | 0.021041517 |
| [AC4]-[AP4] | 110 | 0.021041517 |
| [AC3]-[AC1] | 110 | 0.021041517 |
| [AC2]-[AC10] | 109 | 0.02085023 |
| [AC4]-[SC2] | 102 | 0.019511225 |
| [SC4]-[AP4] | 99 | 0.018937365 |
| [SC4]-[AC9] | 97 | 0.018554792 |
| [AC2]-[AC1] | 92 | 0.01759836 |
| [SC3]-[AC10] | 90 | 0.017215786 |
| [AC1]-[SC4] | 90 | 0.017215786 |
| [SC1]-[AC4] | 87 | 0.016641927 |
| [SC4]-[AP5] | 87 | 0.016641927 |
| [AC2]-[AC3] | 85 | 0.016259354 |
| [AP2]-[SC4] | 82 | 0.015685494 |
| [AC11]-[AC5] | 80 | 0.015302921 |
| [AP5]-[SC5] | 79 | 0.015111635 |
| [AP6]-[SC4] | 78 | 0.014920348 |
| [AC11]-[SC5] | 76 | 0.014537775 |
| [AC4]-[AC11] | 75 | 0.014346489 |
| [AC2]-[AC2] | 75 | 0.014346489 |
| [AC4]-[AP5] | 74 | 0.014155202 |
| [SC2]-[SC1] | 74 | 0.014155202 |
| [SC2]-[AC2] | 71 | 0.013581343 |
| [AC10]-[AC10] | 70 | 0.013390056 |
| [AP2]-[SC11] | 69 | 0.01319877 |
| [AC3]-[AC3] | 67 | 0.012816197 |
| [SC9]-[AC7] | 67 | 0.012816197 |
| [SC6]-[SC2] | 64 | 0.012242337 |
| [AC8]-[SC9] | 62 | 0.011859764 |
| [AC12]-[SC4] | 62 | 0.011859764 |
| [AC11]-[SC4] | 61 | 0.011668478 |
| [AC3]-[AC7] | 61 | 0.011668478 |
| [SC5]-[AP4] | 61 | 0.011668478 |
| [AP5]-[SC4] | 60 | 0.011477191 |

| | | |
|---|---|---|
| [AC4]-[SC1] | 60 | 0.011477191 |
| [AC5]-[AP2] | 58 | 0.011094618 |
| [SC6]-[SC5] | 58 | 0.011094618 |
| [SC11]-[AP2] | 57 | 0.010903331 |
| [SC7]-[AC3] | 57 | 0.010903331 |
| [SC7]-[AC4] | 56 | 0.010712045 |
| [AC6]-[AC2] | 56 | 0.010712045 |
| [SC10]-[AP1] | 55 | 0.010520758 |
| [SC9]-[AC1] | 54 | 0.010329472 |
| [SC9]-[SC1] | 54 | 0.010329472 |
| [AC10]-[SC9] | 52 | 0.009946899 |
| [AP3]-[SC9] | 51 | 0.009755612 |
| [SC6]-[SC4] | 51 | 0.009755612 |
| [SC9]-[AC11] | 50 | 0.009564326 |
| [SC1]-[AC11] | 49 | 0.009373039 |
| [SC1]-[SC1] | 47 | 0.008990466 |
| [AC4]-[AC9] | 47 | 0.008990466 |
| [SC9]-[AC5] | 46 | 0.00879918 |
| [SC2]-[SP3] | 46 | 0.00879918 |
| [SC3]-[AC7] | 45 | 0.008607893 |
| [AC10]-[SC5] | 45 | 0.008607893 |
| [AC6]-[SC3] | 45 | 0.008607893 |
| [AP1]-[SC9] | 44 | 0.008416607 |
| [SC3]-[AC4] | 43 | 0.00822532 |
| [SC10]-[SC4] | 42 | 0.008034034 |
| [AC6]-[AC3] | 41 | 0.007842747 |
| [AC3]-[AC11] | 41 | 0.007842747 |
| [AC5]-[AP3] | 41 | 0.007842747 |
| [AC9]-[AC10] | 40 | 0.007651461 |
| [AP5]-[AC4] | 40 | 0.007651461 |
| [AC9]-[SC9] | 39 | 0.007460174 |
| [SC11]-[AC3] | 39 | 0.007460174 |
| [SC3]-[AC8] | 39 | 0.007460174 |
| [AC10]-[SC10] | 38 | 0.007268888 |
| [SP5]-[SC4] | 38 | 0.007268888 |

| | | |
|---|---|---|
| [SC10]-[AP5] | 38 | 0.007268888 |
| [SC9]-[AC8] | 36 | 0.006886315 |
| [SC4]-[SC10] | 35 | 0.006695028 |
| [SC11]-[SC2] | 34 | 0.006503742 |
| [AC12]-[SC3] | 34 | 0.006503742 |
| [SC3]-[SC7] | 34 | 0.006503742 |
| [SC1]-[SP3] | 34 | 0.006503742 |
| [AC3]-[SC7] | 34 | 0.006503742 |
| [SC10]-[AP4] | 32 | 0.006121169 |
| [AC1]-[SC11] | 32 | 0.006121169 |
| [AC4]-[SC9] | 32 | 0.006121169 |
| [AC3]-[AC5] | 32 | 0.006121169 |
| [AC2]-[AC12] | 32 | 0.006121169 |
| [SP5]-[AC4] | 31 | 0.005929882 |
| [SC11]-[AP1] | 30 | 0.005738595 |
| [AC2]-[SC8] | 29 | 0.005547309 |
| [SC5]-[AP5] | 29 | 0.005547309 |
| [AC7]-[SC3] | 28 | 0.005356022 |
| [AC12]-[AC4] | 28 | 0.005356022 |
| [SC8]-[SC10] | 28 | 0.005356022 |
| [SC1]-[SC12] | 28 | 0.005356022 |
| [SC3]-[AP4] | 28 | 0.005356022 |
| [SC2]-[AP5] | 27 | 0.005164736 |
| [SC3]-[AC3] | 27 | 0.005164736 |
| [SC9]-[AC4] | 27 | 0.005164736 |
| [AC4]-[AC8] | 27 | 0.005164736 |
| [AC3]-[AP1] | 27 | 0.005164736 |
| [SC4]-[SC9] | 26 | 0.004973449 |
| [SC3]-[AP1] | 26 | 0.004973449 |
| [SC8]-[AC4] | 26 | 0.004973449 |
| [AC11]-[AC3] | 25 | 0.004782163 |
| [AC5]-[AP4] | 25 | 0.004782163 |
| [AC4]-[AP6] | 25 | 0.004782163 |
| [SP3]-[SC10] | 25 | 0.004782163 |
| [SP2]-[SC10] | 24 | 0.004590876 |

| | | |
|---|---|---|
| [SC10]-[AC7] | 24 | 0.004590876 |
| [SC1]-[SC3] | 24 | 0.004590876 |
| [AC6]-[SC1] | 24 | 0.004590876 |
| [SC9]-[SC2] | 24 | 0.004590876 |
| [AC11]-[AC1] | 24 | 0.004590876 |
| [SC2]-[SC5] | 24 | 0.004590876 |
| [AC3]-[AC9] | 24 | 0.004590876 |
| [SC4]-[SC6] | 23 | 0.00439959 |
| [SC9]-[AP5] | 23 | 0.00439959 |
| [AC10]-[AP3] | 22 | 0.004208303 |
| [SC1]-[SC5] | 22 | 0.004208303 |
| [SC12]-[AC5] | 21 | 0.004017017 |
| [AC1]-[AC2] | 21 | 0.004017017 |
| [AP1]-[SC10] | 21 | 0.004017017 |
| [AC7]-[AC3] | 20 | 0.00382573 |
| [SC3]-[AC9] | 20 | 0.00382573 |
| [AC5]-[AC6] | 20 | 0.00382573 |
| [AC2]-[AC11] | 20 | 0.00382573 |
| [SC2]-[SC8] | 20 | 0.00382573 |
| [AP2]-[AC4] | 20 | 0.00382573 |
| [SC10]-[AC8] | 20 | 0.00382573 |
| [SP5]-[SC5] | 20 | 0.00382573 |
| [AP1]-[SC5] | 20 | 0.00382573 |
| [SC2]-[AC1] | 20 | 0.00382573 |
| [SC4]-[AC11] | 19 | 0.003634444 |
| [AP4]-[AC4] | 19 | 0.003634444 |
| [AC11]-[SC1] | 19 | 0.003634444 |
| [AC3]-[AP5] | 18 | 0.003443157 |
| [SC7]-[AC2] | 18 | 0.003443157 |
| [AP5]-[SC2] | 18 | 0.003443157 |
| [AC9]-[SC10] | 18 | 0.003443157 |
| [SC1]-[AC9] | 17 | 0.003251871 |
| [SC10]-[SC11] | 17 | 0.003251871 |
| [SC10]-[AC9] | 17 | 0.003251871 |
| [AC12]-[SC2] | 17 | 0.003251871 |

| | | |
|---|---|---|
| [SC1]-[AC2] | 16 | 0.003060584 |
| [AC11]-[SC2] | 16 | 0.003060584 |
| [SC9]-[AP3] | 16 | 0.003060584 |
| [AP4]-[AC3] | 16 | 0.003060584 |
| [SP5]-[AC3] | 16 | 0.003060584 |
| [SC3]-[AC11] | 16 | 0.003060584 |
| [SC1]-[AC3] | 15 | 0.002869298 |
| [AC10]-[AC1] | 15 | 0.002869298 |
| [SC1]-[AC12] | 15 | 0.002869298 |
| [SC9]-[SC12] | 14 | 0.002678011 |
| [AC9]-[AC9] | 14 | 0.002678011 |
| [AC1]-[SC3] | 14 | 0.002678011 |
| [SC6]-[SC3] | 14 | 0.002678011 |
| [SC2]-[AC5] | 14 | 0.002678011 |
| [SC6]-[AC4] | 14 | 0.002678011 |
| [SC5]-[SC6] | 14 | 0.002678011 |
| [AC3]-[AP4] | 14 | 0.002678011 |
| [SC10]-[AC11] | 14 | 0.002678011 |
| [SC6]-[AC3] | 13 | 0.002486725 |
| [AP1]-[SC4] | 13 | 0.002486725 |
| [SC8]-[SC9] | 13 | 0.002486725 |
| [SC3]-[AP5] | 13 | 0.002486725 |
| [AC3]-[SC11] | 13 | 0.002486725 |
| [SC3]-[AC5] | 13 | 0.002486725 |
| [AC11]-[AC2] | 13 | 0.002486725 |
| [AC3]-[AC2] | 12 | 0.002295438 |
| [SC9]-[AC9] | 12 | 0.002295438 |
| [AC10]-[AC9] | 12 | 0.002295438 |
| [AP3]-[SC10] | 12 | 0.002295438 |
| [SC2]-[AC9] | 12 | 0.002295438 |
| [AC2]-[AC7] | 12 | 0.002295438 |
| [SC9]-[AP6] | 12 | 0.002295438 |
| [AC3]-[SC4] | 12 | 0.002295438 |
| [SC11]-[SC3] | 12 | 0.002295438 |
| [SP2]-[AC10] | 12 | 0.002295438 |

| | | |
|---|---|---|
| [SC4]-[AC8] | 11 | 0.002104152 |
| [SC11]-[SC1] | 11 | 0.002104152 |
| [SC11]-[AP4] | 11 | 0.002104152 |
| [AP1]-[SC1] | 11 | 0.002104152 |
| [AC8]-[SC1] | 11 | 0.002104152 |
| [AP1]-[SC2] | 11 | 0.002104152 |
| [AC1]-[SC5] | 11 | 0.002104152 |
| [SC10]-[AP2] | 11 | 0.002104152 |
| [SC3]-[SC11] | 10 | 0.001912865 |
| [AC2]-[SP3] | 10 | 0.001912865 |
| [SC7]-[SC11] | 10 | 0.001912865 |
| [SC9]-[AC12] | 10 | 0.001912865 |
| [SC1]-[AP5] | 10 | 0.001912865 |
| [AC12]-[AC2] | 10 | 0.001912865 |
| [SC8]-[SC4] | 10 | 0.001912865 |
| [AC12]-[AC3] | 10 | 0.001912865 |
| [SP6]-[SC4] | 10 | 0.001912865 |
| [AC2]-[AC8] | 10 | 0.001912865 |
| [SC10]-[SC8] | 10 | 0.001912865 |
| [AC6]-[AC5] | 10 | 0.001912865 |
| [AC8]-[AC10] | 10 | 0.001912865 |
| [AP6]-[AC4] | 9 | 0.001721579 |
| [SC4]-[AP6] | 9 | 0.001721579 |
| [AP3]-[SC7] | 9 | 0.001721579 |
| [SC2]-[AC7] | 9 | 0.001721579 |
| [SC3]-[AC12] | 9 | 0.001721579 |
| [SC3]-[SC8] | 9 | 0.001721579 |
| [AC3]-[AP3] | 9 | 0.001721579 |
| [AP5]-[SC1] | 8 | 0.001530292 |
| [AC3]-[AC8] | 8 | 0.001530292 |
| [AC10]-[AC6] | 8 | 0.001530292 |
| [AC2]-[AC5] | 8 | 0.001530292 |
| [SC9]-[SC7] | 8 | 0.001530292 |
| [AC1]-[AC4] | 8 | 0.001530292 |
| [AP5]-[AC3] | 8 | 0.001530292 |

| | | |
|---|---|---|
| [AC5]-[AP1] | 8 | 0.001530292 |
| [AP5]-[AC2] | 8 | 0.001530292 |
| [AP2]-[AC2] | 8 | 0.001530292 |
| [AC2]-[SC5] | 7 | 0.001339006 |
| [AC3]-[SC5] | 7 | 0.001339006 |
| [SC11]-[AC5] | 7 | 0.001339006 |
| [AP4]-[SC5] | 7 | 0.001339006 |
| [SP6]-[AC4] | 7 | 0.001339006 |
| [SC2]-[SP2] | 7 | 0.001339006 |
| [SP6]-[AC3] | 7 | 0.001339006 |
| [AP1]-[SC11] | 7 | 0.001339006 |
| [AP2]-[AC10] | 7 | 0.001339006 |
| [SC9]-[AP1] | 6 | 0.001147719 |
| [AC3]-[SC8] | 6 | 0.001147719 |
| [AC5]-[SC6] | 6 | 0.001147719 |
| [SC9]-[AC3] | 6 | 0.001147719 |
| [SC5]-[AP6] | 6 | 0.001147719 |
| [AC3]-[AP2] | 6 | 0.001147719 |
| [AC10]-[AC12] | 6 | 0.001147719 |
| [AP3]-[AC1] | 6 | 0.001147719 |
| [AC3]-[AC12] | 6 | 0.001147719 |
| [SC11]-[AP3] | 6 | 0.001147719 |
| [SP5]-[AC2] | 6 | 0.001147719 |
| [SC9]-[AP4] | 6 | 0.001147719 |
| [AC3]-[AC6] | 6 | 0.001147719 |
| [AC7]-[SC2] | 6 | 0.001147719 |
| [SP2]-[AC11] | 6 | 0.001147719 |
| [SC4]-[SC12] | 5 | 0.000956433 |
| [SC10]-[AC2] | 5 | 0.000956433 |
| [AC11]-[AP1] | 5 | 0.000956433 |
| [SC9]-[SC8] | 5 | 0.000956433 |
| [AP6]-[SC3] | 5 | 0.000956433 |
| [AC1]-[SC12] | 5 | 0.000956433 |
| [SP2]-[SC9] | 5 | 0.000956433 |
| [SC12]-[AP3] | 5 | 0.000956433 |

| | | |
|---|---|---|
| [SC5]-[SC10] | 5 | 0.000956433 |
| [AC2]-[AC9] | 5 | 0.000956433 |
| [AC1]-[SP3] | 5 | 0.000956433 |
| [SP6]-[SC5] | 4 | 0.000765146 |
| [AC12]-[SC1] | 4 | 0.000765146 |
| [SC12]-[AC11] | 4 | 0.000765146 |
| [SC6]-[AC2] | 4 | 0.000765146 |
| [SC3]-[AP6] | 4 | 0.000765146 |
| [AC7]-[AC4] | 4 | 0.000765146 |
| [SC4]-[SP5] | 4 | 0.000765146 |
| [AP3]-[SC1] | 4 | 0.000765146 |
| [AP4]-[SC4] | 4 | 0.000765146 |
| [SC7]-[SC9] | 4 | 0.000765146 |
| [AP5]-[SC3] | 4 | 0.000765146 |
| [AP3]-[AC4] | 4 | 0.000765146 |
| [AC1]-[SP2] | 4 | 0.000765146 |
| [AC3]-[SC12] | 4 | 0.000765146 |
| [AP1]-[SC3] | 4 | 0.000765146 |
| [SC5]-[SP5] | 4 | 0.000765146 |
| [SC3]-[SC12] | 4 | 0.000765146 |
| [SC1]-[SC7] | 4 | 0.000765146 |
| [AP2]-[AC1] | 4 | 0.000765146 |
| [AP3]-[AC10] | 4 | 0.000765146 |
| [AP6]-[AC3] | 3 | 0.00057386 |
| [SC5]-[AC12] | 3 | 0.00057386 |
| [SC1]-[AC7] | 3 | 0.00057386 |
| [AP1]-[AC3] | 3 | 0.00057386 |
| [SP2]-[SC11] | 3 | 0.00057386 |
| [AC1]-[AC10] | 3 | 0.00057386 |
| [SC11]-[AC6] | 3 | 0.00057386 |
| [SC5]-[AC11] | 3 | 0.00057386 |
| [SP5]-[SC3] | 3 | 0.00057386 |
| [SC5]-[AC10] | 3 | 0.00057386 |
| [SC7]-[SC10] | 3 | 0.00057386 |
| [SC2]-[AP3] | 3 | 0.00057386 |

| | | |
|---|---|---|
| [SC6]-[SC10] | 3 | 0.00057386 |
| [SP5]-[SC2] | 3 | 0.00057386 |
| [SC3]-[AC2] | 3 | 0.00057386 |
| [SP3]-[SC9] | 3 | 0.00057386 |
| [AC2]-[SP2] | 3 | 0.00057386 |
| [SC12]-[SC1] | 3 | 0.00057386 |
| [AP6]-[SC7] | 3 | 0.00057386 |
| [SC12]-[SP5] | 3 | 0.00057386 |
| [AP3]-[AC2] | 3 | 0.00057386 |
| [SP2]-[AP5] | 2 | 0.000382573 |
| [AC5]-[SP5] | 2 | 0.000382573 |
| [AC2]-[SC6] | 2 | 0.000382573 |
| [AP5]-[AC5] | 2 | 0.000382573 |
| [AC1]-[AC11] | 2 | 0.000382573 |
| [AC11]-[AP4] | 2 | 0.000382573 |
| [SC7]-[AC11] | 2 | 0.000382573 |
| [AP4]-[AP3] | 2 | 0.000382573 |
| [SC1]-[SC8] | 2 | 0.000382573 |
| [SC11]-[AC10] | 2 | 0.000382573 |
| [SC5]-[SC8] | 2 | 0.000382573 |
| [SC6]-[SP5] | 2 | 0.000382573 |
| [SP6]-[SP5] | 2 | 0.000382573 |
| [SC2]-[AP4] | 2 | 0.000382573 |
| [SC3]-[AP3] | 2 | 0.000382573 |
| [SP5]-[SC1] | 2 | 0.000382573 |
| [AP4]-[AP4] | 2 | 0.000382573 |
| [AP3]-[SC4] | 2 | 0.000382573 |
| [AC10]-[SC2] | 2 | 0.000382573 |
| [AP3]-[AC11] | 2 | 0.000382573 |
| [SC10]-[AC12] | 2 | 0.000382573 |
| [SC6]-[SC11] | 2 | 0.000382573 |
| [AC1]-[AC12] | 2 | 0.000382573 |
| [SC10]-[SC3] | 2 | 0.000382573 |
| [AC2]-[AP1] | 2 | 0.000382573 |
| [SC8]-[SC11] | 2 | 0.000382573 |

| | | |
|---|---|---|
| [AC9]-[SC2] | 2 | 0.000382573 |
| [SC2]-[SC3] | 2 | 0.000382573 |
| [AP3]-[AC7] | 2 | 0.000382573 |
| [SP2]-[SP3] | 2 | 0.000382573 |
| [SC10]-[AP6] | 2 | 0.000382573 |
| [AC10]-[AP2] | 2 | 0.000382573 |
| [SC1]-[SP2] | 2 | 0.000382573 |
| [SC8]-[SP3] | 2 | 0.000382573 |
| [SP5]-[AC5] | 2 | 0.000382573 |
| [AC3]-[SC3] | 2 | 0.000382573 |
| [AC10]-[AC2] | 2 | 0.000382573 |
| [AC4]-[SC6] | 2 | 0.000382573 |
| [SC5]-[SC7] | 2 | 0.000382573 |
| [AC12]-[AC5] | 2 | 0.000382573 |
| [AP3]-[SC11] | 2 | 0.000382573 |
| [AC4]-[SC8] | 2 | 0.000382573 |
| [SC7]-[SP5] | 2 | 0.000382573 |
| [AC11]-[AP2] | 2 | 0.000382573 |
| [SP5]-[AP4] | 2 | 0.000382573 |
| [SP6]-[AC11] | 1 | 0.000191287 |
| [SP3]-[SC5] | 1 | 0.000191287 |
| [SP3]-[SP3] | 1 | 0.000191287 |
| [SC2]-[AC8] | 1 | 0.000191287 |
| [AP5]-[AC1] | 1 | 0.000191287 |
| [AP1]-[SC12] | 1 | 0.000191287 |
| [AC5]-[AC12] | 1 | 0.000191287 |
| [SP6]-[AC1] | 1 | 0.000191287 |
| [SC12]-[AC9] | 1 | 0.000191287 |
| [AC2]-[AP2] | 1 | 0.000191287 |
| [AC8]-[AC4] | 1 | 0.000191287 |
| [SP2]-[AP6] | 1 | 0.000191287 |
| [SP2]-[AC3] | 1 | 0.000191287 |
| [SC12]-[AP1] | 1 | 0.000191287 |
| [SC5]-[SP6] | 1 | 0.000191287 |
| [SC6]-[SP2] | 1 | 0.000191287 |

| | | |
|---|---|---|
| [SC3]-[SP3] | 1 | 0.000191287 |
| [AP4]-[AC11] | 1 | 0.000191287 |
| [AC4]-[SP6] | 1 | 0.000191287 |
| [AC11]-[AP3] | 1 | 0.000191287 |
| [SC12]-[AC12] | 1 | 0.000191287 |
| [SP5]-[SP4] | 1 | 0.000191287 |
| [SC7]-[SC1] | 1 | 0.000191287 |
| [AC3]-[SC2] | 1 | 0.000191287 |
| [SC6]-[SP6] | 1 | 0.000191287 |
| [AP4]-[SC9] | 1 | 0.000191287 |
| [AC3]-[AP6] | 1 | 0.000191287 |
| [AC5]-[AC11] | 1 | 0.000191287 |
| [AC9]-[AC3] | 1 | 0.000191287 |
| [SC11]-[AP6] | 1 | 0.000191287 |
| [AP2]-[SC5] | 1 | 0.000191287 |
| [SP5]-[SC6] | 1 | 0.000191287 |
| [SC1]-[AP1] | 1 | 0.000191287 |
| [AP1]-[AC2] | 1 | 0.000191287 |
| [AC9]-[AC4] | 1 | 0.000191287 |
| [AP2]-[AP3] | 1 | 0.000191287 |
| [SC8]-[AC3] | 1 | 0.000191287 |
| [SC4]-[SP6] | 1 | 0.000191287 |
| [SC10]-[SC2] | 1 | 0.000191287 |
| [AC11]-[AP6] | 1 | 0.000191287 |
| [SP3]-[SC2] | 1 | 0.000191287 |
| [AC5]-[SC12] | 1 | 0.000191287 |
| [AP2]-[SC8] | 1 | 0.000191287 |
| [SC9]-[AC6] | 1 | 0.000191287 |
| [SC9]-[SP2] | 1 | 0.000191287 |
| [AC10]-[AC8] | 1 | 0.000191287 |
| [AC4]-[AC12] | 1 | 0.000191287 |
| [AP4]-[AC2] | 1 | 0.000191287 |
| [SP3]-[AC3] | 1 | 0.000191287 |
| [AC6]-[AP3] | 1 | 0.000191287 |
| [SP3]-[AP6] | 1 | 0.000191287 |

| | | |
|---|---|---|
| [AC8]-[SC8] | 1 | 0.000191287 |
| [SP4]-[SC2] | 1 | 0.000191287 |
| [AP5]-[SC6] | 1 | 0.000191287 |
| [SC1]-[AC5] | 1 | 0.000191287 |
| [AC12]-[SP5] | 1 | 0.000191287 |
| [SC1]-[SP1] | 1 | 0.000191287 |
| [AP2]-[AC7] | 1 | 0.000191287 |
| [SC2]-[AP2] | 1 | 0.000191287 |
| [SP1]-[SC9] | 1 | 0.000191287 |
| [AP2]-[AC3] | 1 | 0.000191287 |
| [AC8]-[SC2] | 1 | 0.000191287 |
| [SC8]-[AC11] | 1 | 0.000191287 |
| [AP2]-[AC5] | 1 | 0.000191287 |
| [AC9]-[SC5] | 1 | 0.000191287 |
| [AP3]-[AC5] | 1 | 0.000191287 |
| [SP3]-[SC4] | 1 | 0.000191287 |
| [AC10]-[AP4] | 1 | 0.000191287 |
| [SP6]-[SC2] | 1 | 0.000191287 |
| [SC9]-[SC5] | 1 | 0.000191287 |
| [AC11]-[SC3] | 1 | 0.000191287 |
| [AC2]-[AP5] | 1 | 0.000191287 |
| [SC6]-[SC1] | 1 | 0.000191287 |
| [SP2]-[SC1] | 1 | 0.000191287 |
| [SP3]-[AC4] | 1 | 0.000191287 |
| [SP3]-[AP4] | 1 | 0.000191287 |
| [SP5]-[AC1] | 1 | 0.000191287 |
| [SC3]-[AP2] | 1 | 0.000191287 |
| [AP4]-[SC3] | 1 | 0.000191287 |
| [SP4]-[AC10] | 1 | 0.000191287 |
| [SP4]-[AC4] | 1 | 0.000191287 |
| [AC12]-[AC1] | 1 | 0.000191287 |
| [SP3]-[AC11] | 1 | 0.000191287 |
| [AC6]-[SC10] | 1 | 0.000191287 |
| [AC12]-[SC6] | 1 | 0.000191287 |
| [SC8]-[AP2] | 1 | 0.000191287 |

| | | |
|---|---|---|
| [SC4]-[AC7] | 1 | 0.000191287 |
| [SC9]-[SP3] | 1 | 0.000191287 |
| [SC12]-[SC11] | 1 | 0.000191287 |
| [SC6]-[AC5] | 1 | 0.000191287 |
| [SC2]-[AP1] | 1 | 0.000191287 |
| [AC9]-[AP5] | 1 | 0.000191287 |
| [AP3]-[SC3] | 1 | 0.000191287 |
| [SC1]-[AP6] | 1 | 0.000191287 |
| [SC1]-[AP3] | 1 | 0.000191287 |
| [SC10]-[AC1] | 1 | 0.000191287 |
| [SC9]-[AC2] | 1 | 0.000191287 |
| [SP3]-[AC10] | 1 | 0.000191287 |
| [AP5]-[SP5] | 1 | 0.000191287 |
| [SP6]-[AC2] | 1 | 0.000191287 |
| [SP4]-[SC4] | 1 | 0.000191287 |
| [SC8]-[SC8] | 1 | 0.000191287 |
| [AP2]-[SP3] | 1 | 0.000191287 |
| [AP2]-[AC6] | 1 | 0.000191287 |
| [AP4]-[AC10] | 1 | 0.000191287 |
| [SP5]-[AP2] | 1 | 0.000191287 |
| [AC5]-[SC9] | 1 | 0.000191287 |
| [SC11]-[SC6] | 1 | 0.000191287 |
| [AC8]-[SC11] | 1 | 0.000191287 |
| [SC8]-[SC1] | 1 | 0.000191287 |
| [SC12]-[AP4] | 1 | 0.000191287 |
| [SC12]-[SC10] | 1 | 0.000191287 |
| Total | 522776 | |

**Table 5. 2 The distribution of new turn types**

| Category | Number of turns | Percentage |
|---|---|---|
| [SC4]-[SC4] | 84742 | 16.21 |
| [SC4]-[AC4] | 81476 | 15.59 |
| [SC5]-[SC4] | 55774 | 10.67 |
| [SC2]-[SC10] | 35628 | 6.82 |
| [SC5]-[AC4] | 24738 | 4.73 |
| [SC5]-[SC5] | 16297 | 3.12 |

| | | |
|---|---|---|
| [SC9]-[SC10] | 14136 | 2.70 |
| [SC2]-[AC10] | 13867 | 2.65 |
| [SC4]-[SC5] | 13297 | 2.54 |
| [SC5]-[AC2] | 10005 | 1.91 |
| [SC4]-[AC3] | 9979 | 1.91 |
| [SC12]-[SC4] | 8541 | 1.63 |
| [SC4]-[AC2] | 8453 | 1.62 |
| [AC4]-[AC4] | 7767 | 1.49 |
| [SC4]-[AC5] | 6443 | 1.23 |
| [SC2]-[SC9] | 6368 | 1.22 |
| [SC5]-[AC3] | 5525 | 1.06 |
| [SC12]-[AC4] | 5339 | 1.02 |
| [SC4]-[AC1] | 5094 | 0.97 |
| Top 19 | 413469 | 79.09 |
| … | … | … |
| Total | 522776 | 100.00 |

**Table 5. 3 The distribution of the top 19 β turn types in the large protein database.**

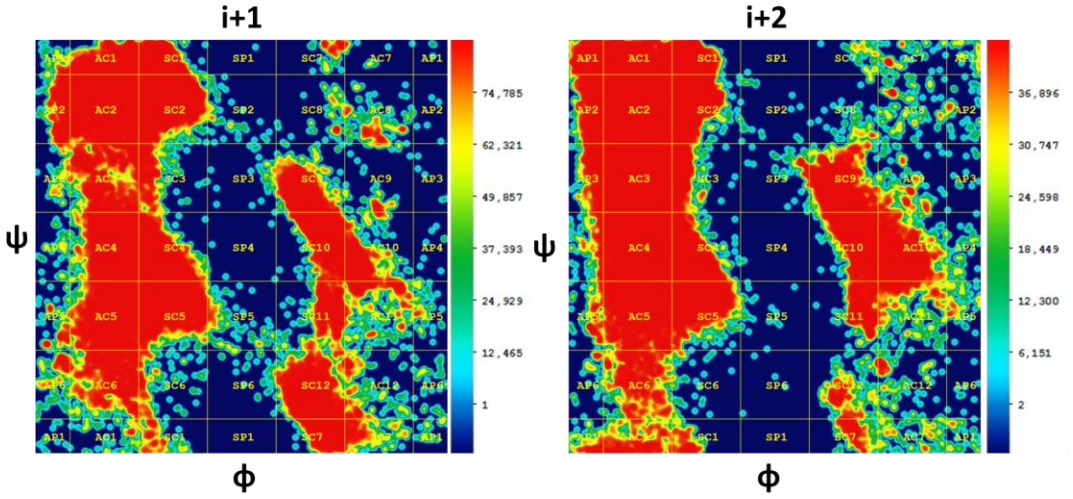
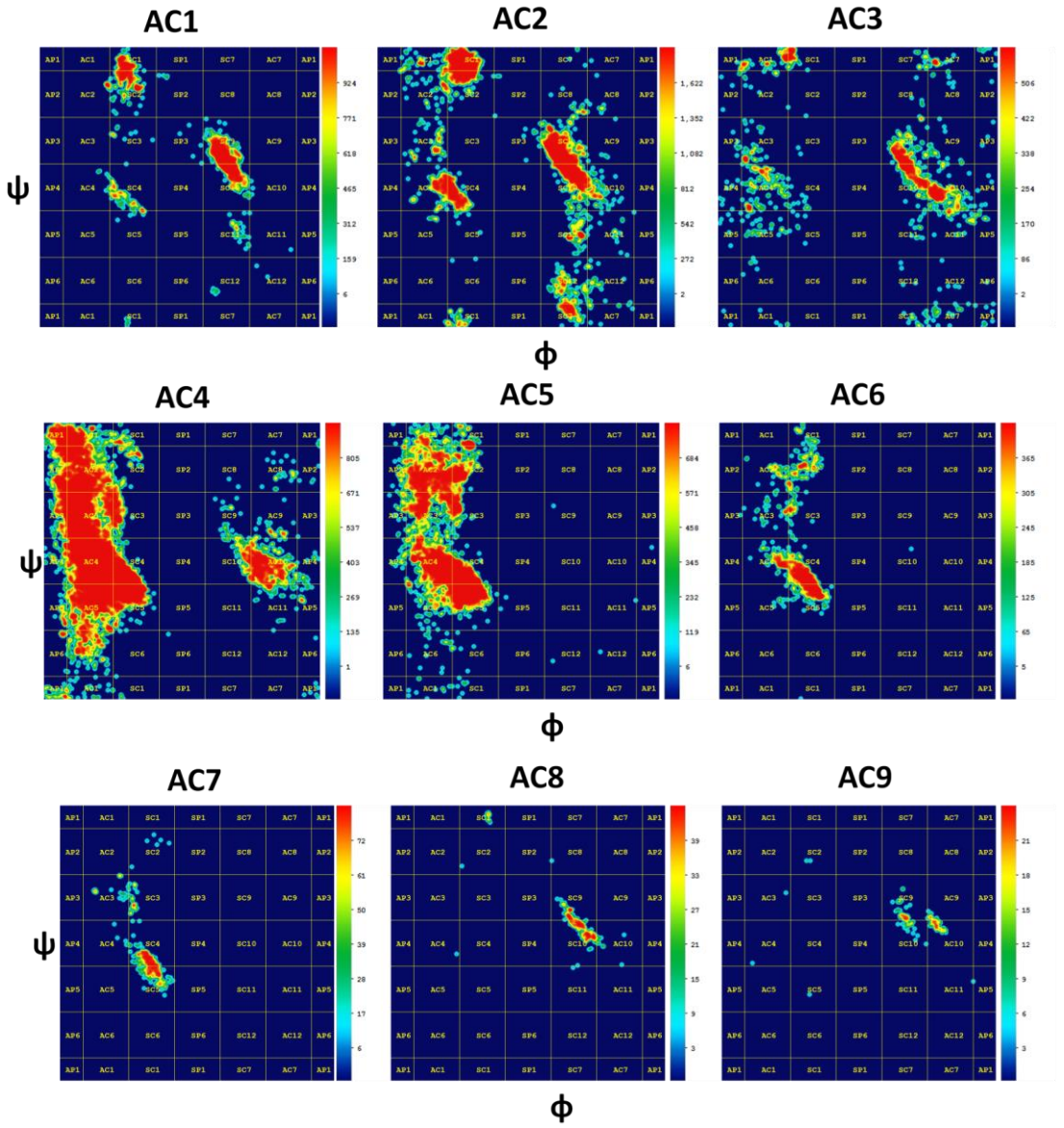
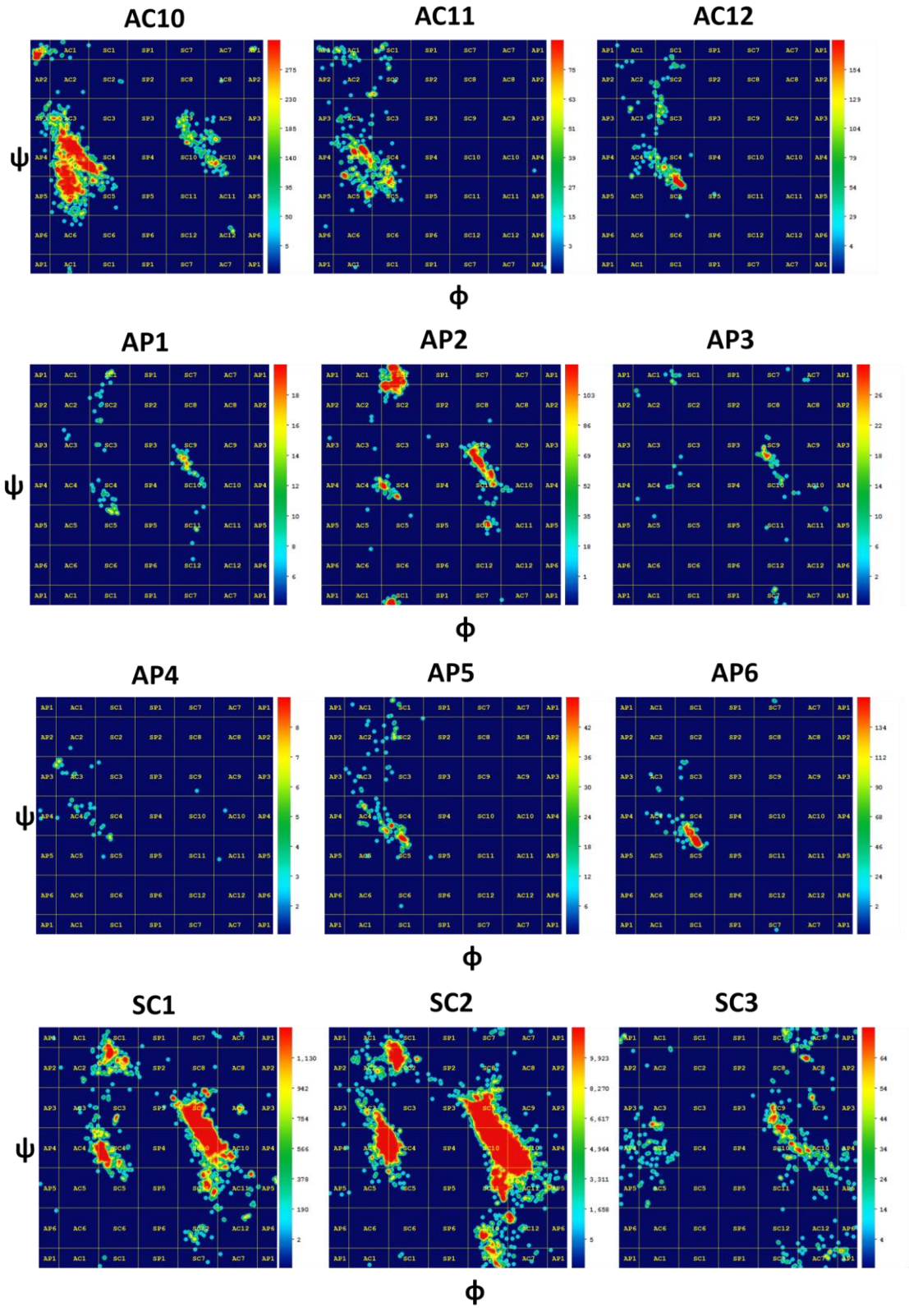


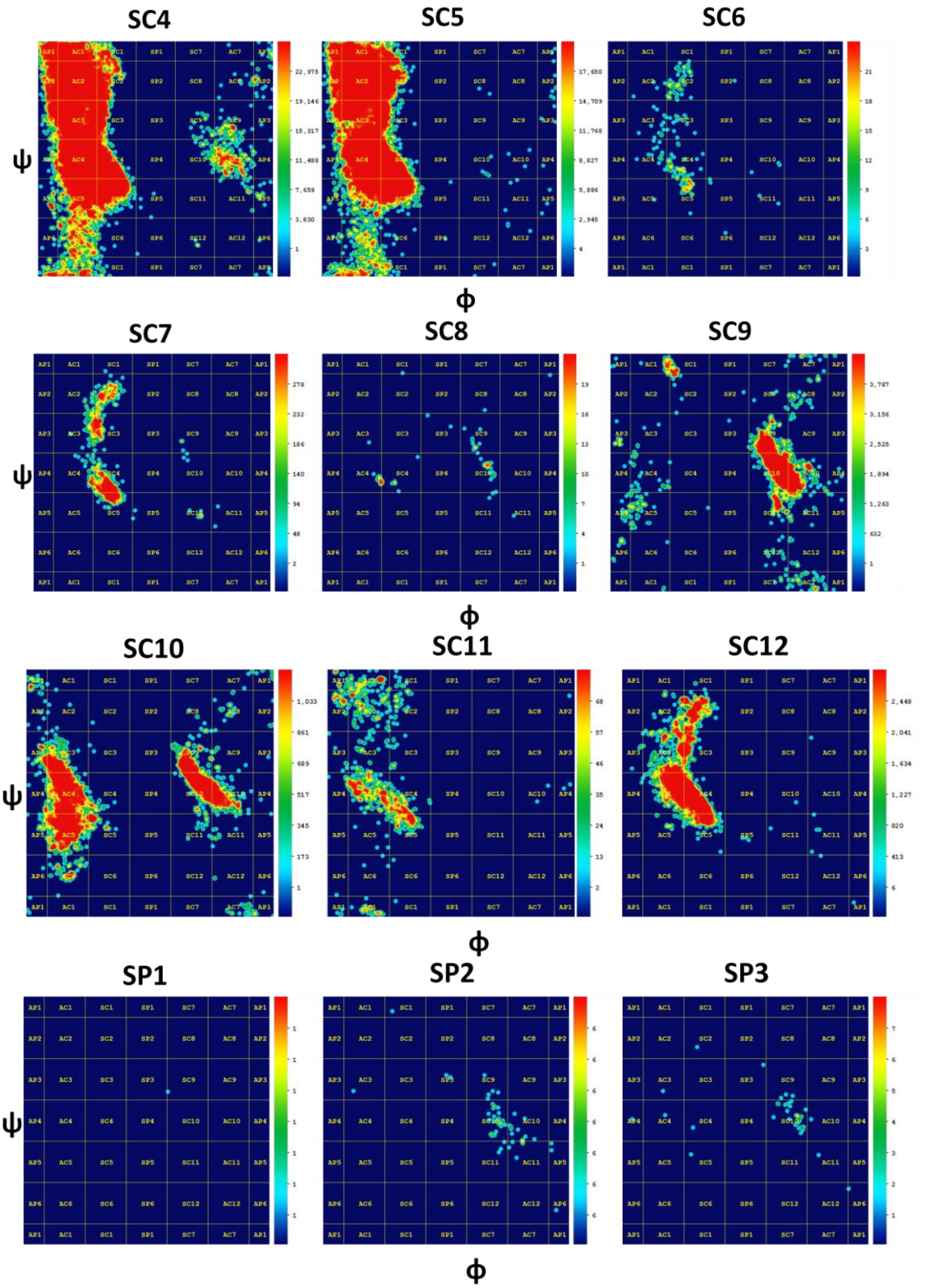

**Figure 5. 21 The heat map of i+1 and i+2 φ and ψ angle distributions for the large β turn database.** This figure represents the distribution of all data points from database. The left panel represents the i+1 residue and right panel represents the i+2 residue. The angles shown on the graphs are φ and ψ dihedral angles. The legend indicating the number of turns corresponding to each color category for the heat map is indicated to the right of each plot.

In order to determine if the conformation of the i+1 residue influenced the conformation of the i+2 residue, heat maps were generated for the distribution of turn types for the i+2

residue for a specific turn type for the i+1 residue, which are depicted for all 36 possible i+1 residue conformations in **Figure 5.22**. As an example, representative heat map plots for three i+1 conformations, AC4, AC5 and AC8, are depicted in **Figure 5.23**. Inspection of these heatmaps indicates that the conformation of the i+2 residue can depend strongly on the conformation of the i+1 residues. For example, when the i+1 residue adopts an AC4 conformation, the conformation of the i+2 residue abundantly populates the AC2-AC5 conformation as well as the AC10 region (**Figure 5.23**). This is in strong contrast to what is observed for the i+2 residue when it follows and i+1 residue with an AC5 conformation, in which the AC10 region is virtually unpopulated (**Figure 5.23**). Also in strong contrast, when the i+2 residue follows an i+1 residue with an AC8 conformation, the AC2-AC5 conformation was virtually unpopulated and the SC9 and SC10 conformations occurred with the highest frequency.




AC10
AC11
AC12
AP1
AP2
AP3
AP4
AP5
AP6
SC1
SC2
SC3
ψ
ϕ


SC4
SC5
SC6
SC7
SC8
SC9
SC10
SC11
SC12
SP1
SP2
SP3
ψ
ϕ

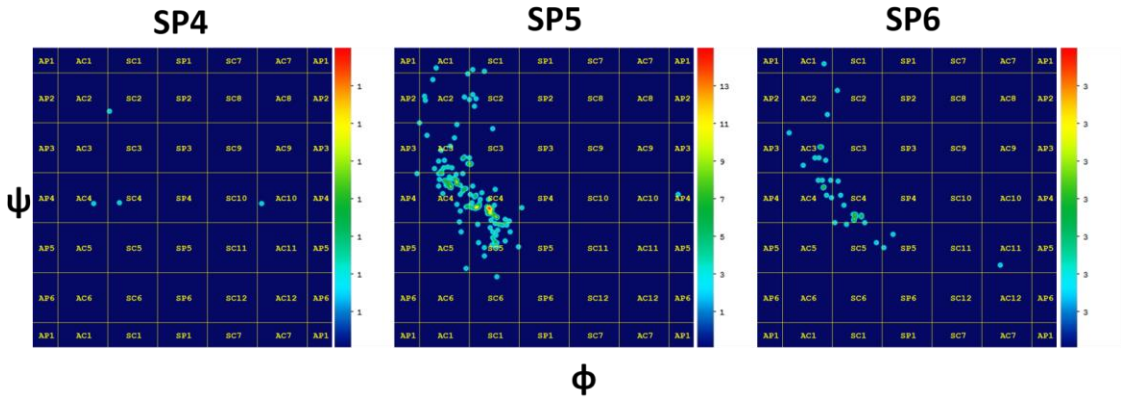



**Figure 5. 22 The heat maps of data points distribution for all categories.** The labels above each figure represent the category of the second residue and the data points in the graph represent the distribution of third residue under each corresponding category.

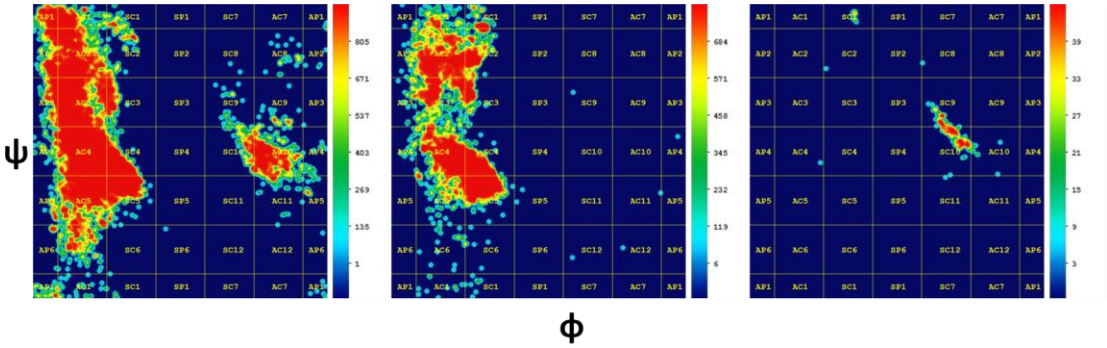



**Figure 5. 23 The heat map of i+2 angle distributions for three representative i+1 categories.** The labels above each figure represent the category of the i+1 residue and the data points in the graph represent the distribution of i+2 residue under each corresponding category. The legend indicating the number of turns corresponding to each color category for the heat map is indicated to the right of each plot.

Finally, we investigated how β turns categorized as type IV were distributed in the new classification scheme. Inspection of the heat maps illustrates that distinct distributions were observed for the i+1 and i+2 residues (**Figure 5.24**). For example, whereas that AC2-AC3 and SC2-SC3 regions were strongly populated at the i+1 residue they were notably less populated at the i+2 position. Similarly, the AC5-AC6 sections are much more densely populated at the i+1 residue in comparison to at the i+2 residue. On the other hand, the SC12 and adjacent SC7 sections are much more densely populated at the i+2 residue in comparison to the i+1 residue.

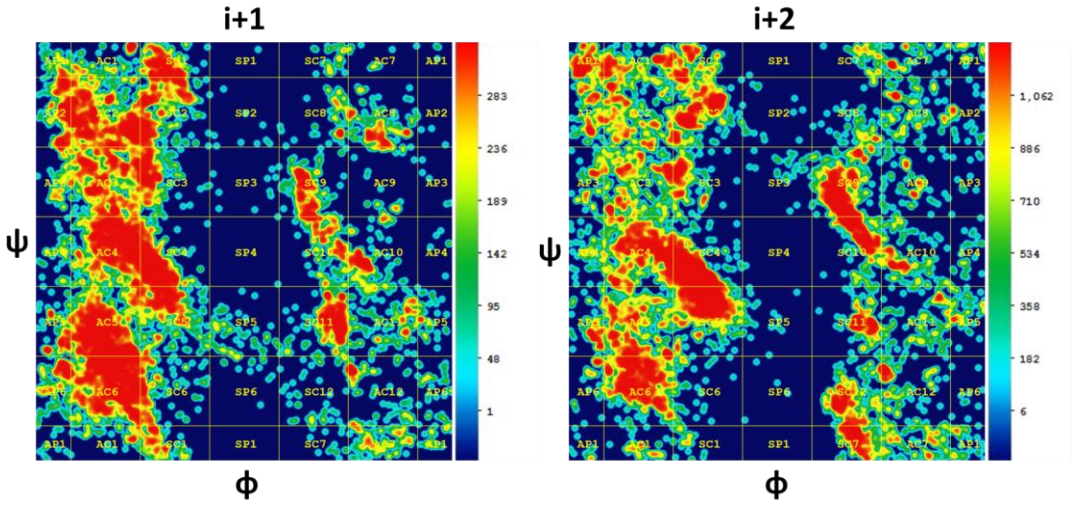



**Figure 5. 24 The heat map of the new β turn distribution for previously designated type IV β turns.** The legend indicating the number of turns corresponding to each color category for the heat map is indicated to the right of each plot.

### 5.4.3 Hydrogen bond occurrences in β turns

Since hydrogen bonds are not an essential element of β turns due to the introduction of open β turns lacking hydrogen bonds, we evaluated how hydrogen bond occurrences were distributed in the new β turn classification scheme (**Table 5.4, Table 5.5, Table 5.6**). Approximately 65.9% of new turn types never contained hydrogen bonds and around 6.7% of new turn types always contained hydrogen bonds (**Table 5.7**). In the turn types that always contained hydrogen bonds, almost all contained moderate rather than weak or strong bonds. Interestingly, around 27.4% of new turn types sometimes contained hydrogen bonds with ~68.1% of the turn types forming no hydrogen bonds (the non-hydrogen bond β turns occupied over 50 % of total β turns in each turn type). The remaining 31.9% of turn types that sometimes included hydrogen bonds tended to have either moderate or weak hydrogen bonds.

| Composition of H-bond | Partial H-Bond | No H-Bond | ALL H-Bond |
|---|---|---|---|
| # of each composition | 160 | 384 | 39 |
| Percentage of each composition | 27.44% | 65.87% | 6.69% |
| Category | [AC1]-[AC4] | [AC1]-[AC10] | [AC10]-[SC2] |
| | [AC1]-[SC1] | [AC1]-[AC11] | [AC3]-[SC2] |
| | [AC1]-[SC2] | [AC1]-[AC12] | [AC4]-[AC12] |
| | [AC1]-[SC4] | [AC1]-[AC2] | [AC9]-[AC3] |
| | [AC11]-[AP4] | [AC1]-[SC10] | [AP1]-[AC2] |
| | [AC11]-[SC1] | [AC1]-[SC11] | [SC1]-[AP6] |
| | [AC11]-[SC2] | [AC1]-[SC12] | [SC12]-[AC11] |
| | [AC12]-[AC2] | [AC1]-[SC3] | [SC12]-[AC12] |
| | [AC2]-[AC1] | [AC1]-[SC5] | [SC12]-[AP4] |
| | [AC2]-[AC10] | [AC1]-[SC9] | [SC12]-[SC10] |
| | [AC2]-[AC12] | [AC1]-[SP2] | [SC2]-[AC3] |
| | [AC2]-[AC2] | [AC1]-[SP3] | [SC2]-[SC3] |
| | [AC2]-[AC3] | [AC10]-[AC1] | [SP2]-[AC11] |
| | [AC2]-[AC4] | [AC10]-[AC10] | [SP2]-[AC3] |
| | [AC2]-[SC1] | [AC10]-[AC12] | [SP2]-[SC10] |
| | [AC2]-[SC10] | [AC10]-[AC2] | [SP2]-[SC11] |
| | [AC2]-[SC11] | [AC10]-[AC3] | [SP2]-[SP3] |
| | [AC2]-[SC2] | [AC10]-[AC4] | [SP3]-[AC10] |
| | [AC2]-[SC4] | [AC10]-[AC5] | [SP3]-[AC11] |
| | [AC2]-[SC5] | [AC10]-[AC6] | [SP3]-[AC3] |
| | [AC2]-[SC9] | [AC10]-[AC8] | [SP3]-[AP4] |

| | | | |
|---|---|---|---|
| | [AC2]-[SP2] | [AC10]-[AC9] | [SP3]-[AP6] |
| | [AC3]-[AC1] | [AC10]-[AP1] | [SP3]-[SC10] |
| | [AC3]-[AC10] | [AC10]-[AP2] | [SP3]-[SC2] |
| | [AC3]-[AC2] | [AC10]-[AP3] | [SP3]-[SC9] |
| | [AC3]-[AC4] | [AC10]-[AP4] | [SP3]-[SP3] |
| | [AC3]-[AP4] | [AC10]-[SC10] | [SP4]-[AC10] |
| | [AC3]-[AP5] | [AC10]-[SC4] | [SP4]-[AC4] |
| | [AC3]-[SC1] | [AC10]-[SC5] | [SP4]-[SC4] |
| | [AC3]-[SC4] | [AC10]-[SC9] | [SP5]-[AP2] |
| | [AC4]-[AC3] | [AC11]-[AC1] | [SP5]-[SC3] |
| | [AC4]-[AC4] | [AC11]-[AC2] | [SP5]-[SC4] |
| | [AC4]-[SC4] | [AC11]-[AC3] | [SP5]-[SC6] |
| | [AC4]-[SC5] | [AC11]-[AC4] | [SP5]-[SP4] |
| | [AP2]-[SC1] | [AC11]-[AC5] | [SP6]-[AC11] |
| | [AP2]-[SC2] | [AC11]-[AP1] | [SP6]-[AC3] |
| | [AP5]-[SC2] | [AC11]-[AP2] | [SP6]-[AC4] |
| | [SC1]-[AC10] | [AC11]-[AP3] | [SP6]-[SC4] |
| | [SC1]-[AC11] | [AC11]-[AP6] | [SP6]-[SC5] |
| | [SC1]-[AC2] | [AC11]-[SC3] | |
| | [SC1]-[AC3] | [AC11]-[SC4] | |
| | [SC1]-[AC4] | [AC11]-[SC5] | |
| | [SC1]-[SC10] | [AC12]-[AC1] | |
| | [SC1]-[SC11] | [AC12]-[AC3] | |
| | [SC1]-[SC2] | [AC12]-[AC4] | |
| | [SC1]-[SC4] | [AC12]-[AC5] | |
| | [SC1]-[SC5] | [AC12]-[SC1] | |
| | [SC1]-[SC9] | [AC12]-[SC2] | |
| | [SC1]-[SP3] | [AC12]-[SC3] | |
| | [SC10]-[AC10] | [AC12]-[SC4] | |
| | [SC10]-[AC11] | [AC12]-[SC5] | |
| | [SC10]-[AC4] | [AC12]-[SC6] | |
| | [SC10]-[AC9] | [AC12]-[SP5] | |
| | [SC10]-[AP4] | [AC2]-[AC11] | |
| | [SC10]-[SC10] | [AC2]-[AC5] | |
| | [SC10]-[SC11] | [AC2]-[AC7] | |

| | | | |
|---|---|---|---|
| | [SC10]-[SC9] | [AC2]-[AC8] | |
| | [SC11]-[AC10] | [AC2]-[AC9] | |
| | [SC11]-[AC3] | [AC2]-[AP1] | |
| | [SC11]-[AC4] | [AC2]-[AP2] | |
| | [SC11]-[AC5] | [AC2]-[AP5] | |
| | [SC11]-[AP3] | [AC2]-[SC12] | |
| | [SC11]-[AP4] | [AC2]-[SC6] | |
| | [SC11]-[SC4] | [AC2]-[SC7] | |
| | [SC11]-[SC5] | [AC2]-[SC8] | |
| | [SC12]-[AC2] | [AC2]-[SP3] | |
| | [SC12]-[AC3] | [AC3]-[AC11] | |
| | [SC12]-[AC4] | [AC3]-[AC12] | |
| | [SC12]-[AC5] | [AC3]-[AC3] | |
| | [SC12]-[SC2] | [AC3]-[AC5] | |
| | [SC12]-[SC3] | [AC3]-[AC6] | |
| | [SC12]-[SC4] | [AC3]-[AC7] | |
| | [SC12]-[SC5] | [AC3]-[AC8] | |
| | [SC2]-[AC1] | [AC3]-[AC9] | |
| | [SC2]-[AC10] | [AC3]-[AP1] | |
| | [SC2]-[AC11] | [AC3]-[AP2] | |
| | [SC2]-[AC12] | [AC3]-[AP3] | |
| | [SC2]-[AC2] | [AC3]-[AP6] | |
| | [SC2]-[AC4] | [AC3]-[SC10] | |
| | [SC2]-[AC9] | [AC3]-[SC11] | |
| | [SC2]-[AP3] | [AC3]-[SC12] | |
| | [SC2]-[AP5] | [AC3]-[SC3] | |
| | [SC2]-[SC1] | [AC3]-[SC5] | |
| | [SC2]-[SC10] | [AC3]-[SC7] | |
| | [SC2]-[SC11] | [AC3]-[SC8] | |
| | [SC2]-[SC2] | [AC3]-[SC9] | |
| | [SC2]-[SC4] | [AC4]-[AC1] | |
| | [SC2]-[SC5] | [AC4]-[AC10] | |
| | [SC2]-[SC8] | [AC4]-[AC11] | |
| | [SC2]-[SC9] | [AC4]-[AC2] | |
| | [SC2]-[SP3] | [AC4]-[AC5] | |

| | | | |
|---|---|---|---|
| | [SC3]-[AC10] | [AC4]-[AC6] | |
| | [SC3]-[AC11] | [AC4]-[AC8] | |
| | [SC3]-[AC3] | [AC4]-[AC9] | |
| | [SC3]-[AC4] | [AC4]-[AP1] | |
| | [SC3]-[AC5] | [AC4]-[AP2] | |
| | [SC3]-[AC9] | [AC4]-[AP3] | |
| | [SC3]-[AP4] | [AC4]-[AP4] | |
| | [SC3]-[AP5] | [AC4]-[AP5] | |
| | [SC3]-[SC10] | [AC4]-[AP6] | |
| | [SC4]-[AC2] | [AC4]-[SC1] | |
| | [SC4]-[AC3] | [AC4]-[SC10] | |
| | [SC4]-[AC4] | [AC4]-[SC2] | |
| | [SC4]-[AC5] | [AC4]-[SC3] | |
| | [SC4]-[AC6] | [AC4]-[SC6] | |
| | [SC4]-[AP2] | [AC4]-[SC8] | |
| | [SC4]-[AP3] | [AC4]-[SC9] | |
| | [SC4]-[AP4] | [AC4]-[SP6] | |
| | [SC4]-[SC2] | [AC5]-[AC1] | |
| | [SC4]-[SC3] | [AC5]-[AC11] | |
| | [SC4]-[SC4] | [AC5]-[AC12] | |
| | [SC4]-[SC5] | [AC5]-[AC2] | |
| | [SC4]-[SC6] | [AC5]-[AC3] | |
| | [SC4]-[SP5] | [AC5]-[AC4] | |
| | [SC5]-[AC11] | [AC5]-[AC5] | |
| | [SC5]-[AC12] | [AC5]-[AC6] | |
| | [SC5]-[AC2] | [AC5]-[AP1] | |
| | [SC5]-[AC3] | [AC5]-[AP2] | |
| | [SC5]-[AC4] | [AC5]-[AP3] | |
| | [SC5]-[AC5] | [AC5]-[AP4] | |
| | [SC5]-[AP4] | [AC5]-[SC1] | |
| | [SC5]-[SC10] | [AC5]-[SC12] | |
| | [SC5]-[SC2] | [AC5]-[SC2] | |
| | [SC5]-[SC3] | [AC5]-[SC3] | |
| | [SC5]-[SC4] | [AC5]-[SC4] | |
| | [SC5]-[SC5] | [AC5]-[SC5] | |

| | | | |
|---|---|---|---|
| | [SC5]-[SC6] | [AC5]-[SC6] | |
| | [SC5]-[SP5] | [AC5]-[SC9] | |
| | [SC6]-[AC3] | [AC5]-[SP5] | |
| | [SC6]-[SC3] | [AC6]-[AC2] | |
| | [SC6]-[SC4] | [AC6]-[AC3] | |
| | [SC6]-[SP5] | [AC6]-[AC4] | |
| | [SC7]-[AC11] | [AC6]-[AC5] | |
| | [SC7]-[AC3] | [AC6]-[AP3] | |
| | [SC7]-[AC4] | [AC6]-[SC1] | |
| | [SC7]-[SC11] | [AC6]-[SC10] | |
| | [SC7]-[SC3] | [AC6]-[SC2] | |
| | [SC7]-[SC4] | [AC6]-[SC3] | |
| | [SC7]-[SC5] | [AC6]-[SC4] | |
| | [SC8]-[SC11] | [AC6]-[SC5] | |
| | [SC9]-[AC1] | [AC7]-[AC3] | |
| | [SC9]-[AC10] | [AC7]-[AC4] | |
| | [SC9]-[AC11] | [AC7]-[SC2] | |
| | [SC9]-[AC3] | [AC7]-[SC3] | |
| | [SC9]-[AC4] | [AC7]-[SC4] | |
| | [SC9]-[AC9] | [AC7]-[SC5] | |
| | [SC9]-[SC10] | [AC8]-[AC10] | |
| | [SC9]-[SC11] | [AC8]-[AC4] | |
| | [SC9]-[SC2] | [AC8]-[SC1] | |
| | [SC9]-[SC9] | [AC8]-[SC10] | |
| | [SP2]-[AC10] | [AC8]-[SC11] | |
| | [SP2]-[AP5] | [AC8]-[SC2] | |
| | [SP2]-[SC9] | [AC8]-[SC8] | |
| | [SP5]-[AC2] | [AC8]-[SC9] | |
| | [SP5]-[AC3] | [AC9]-[AC10] | |
| | [SP5]-[AC4] | [AC9]-[AC4] | |
| | [SP5]-[AP4] | [AC9]-[AC9] | |
| | [SP5]-[SC2] | [AC9]-[AP5] | |
| | [SP5]-[SC5] | [AC9]-[SC10] | |
| | [SP6]-[SP5] | [AC9]-[SC2] | |
| | | [AC9]-[SC5] | |

| | | | |
|---|---|---|---|
| | | [AC9]-[SC9] | |
| | | [AP1]-[AC3] | |
| | | [AP1]-[SC1] | |
| | | [AP1]-[SC10] | |
| | | [AP1]-[SC11] | |
| | | [AP1]-[SC12] | |
| | | [AP1]-[SC2] | |
| | | [AP1]-[SC3] | |
| | | [AP1]-[SC4] | |
| | | [AP1]-[SC5] | |
| | | [AP1]-[SC9] | |
| | | [AP2]-[AC1] | |
| | | [AP2]-[AC10] | |
| | | [AP2]-[AC2] | |
| | | [AP2]-[AC3] | |
| | | [AP2]-[AC4] | |
| | | [AP2]-[AC5] | |
| | | [AP2]-[AC6] | |
| | | [AP2]-[AC7] | |
| | | [AP2]-[AP3] | |
| | | [AP2]-[SC10] | |
| | | [AP2]-[SC11] | |
| | | [AP2]-[SC4] | |
| | | [AP2]-[SC5] | |
| | | [AP2]-[SC8] | |
| | | [AP2]-[SC9] | |
| | | [AP2]-[SP3] | |
| | | [AP3]-[AC1] | |
| | | [AP3]-[AC10] | |
| | | [AP3]-[AC11] | |
| | | [AP3]-[AC2] | |
| | | [AP3]-[AC4] | |
| | | [AP3]-[AC5] | |
| | | [AP3]-[AC7] | |
| | | [AP3]-[SC1] | |

| | | | |
|---|---|---|---|
| | | [AP3]-[SC10] | |
| | | [AP3]-[SC11] | |
| | | [AP3]-[SC3] | |
| | | [AP3]-[SC4] | |
| | | [AP3]-[SC7] | |
| | | [AP3]-[SC9] | |
| | | [AP4]-[AC10] | |
| | | [AP4]-[AC11] | |
| | | [AP4]-[AC2] | |
| | | [AP4]-[AC3] | |
| | | [AP4]-[AC4] | |
| | | [AP4]-[AP3] | |
| | | [AP4]-[AP4] | |
| | | [AP4]-[SC3] | |
| | | [AP4]-[SC4] | |
| | | [AP4]-[SC5] | |
| | | [AP4]-[SC9] | |
| | | [AP5]-[AC1] | |
| | | [AP5]-[AC2] | |
| | | [AP5]-[AC3] | |
| | | [AP5]-[AC4] | |
| | | [AP5]-[AC5] | |
| | | [AP5]-[SC1] | |
| | | [AP5]-[SC3] | |
| | | [AP5]-[SC4] | |
| | | [AP5]-[SC5] | |
| | | [AP5]-[SC6] | |
| | | [AP5]-[SP5] | |
| | | [AP6]-[AC3] | |
| | | [AP6]-[AC4] | |
| | | [AP6]-[SC3] | |
| | | [AP6]-[SC4] | |
| | | [AP6]-[SC5] | |
| | | [AP6]-[SC7] | |
| | | [SC1]-[AC12] | |

| | | | |
|---|---|---|---|
| | | [SC1]-[AC5] | |
| | | [SC1]-[AC7] | |
| | | [SC1]-[AC9] | |
| | | [SC1]-[AP1] | |
| | | [SC1]-[AP3] | |
| | | [SC1]-[AP5] | |
| | | [SC1]-[SC1] | |
| | | [SC1]-[SC12] | |
| | | [SC1]-[SC3] | |
| | | [SC1]-[SC7] | |
| | | [SC1]-[SC8] | |
| | | [SC1]-[SP1] | |
| | | [SC1]-[SP2] | |
| | | [SC10]-[AC1] | |
| | | [SC10]-[AC12] | |
| | | [SC10]-[AC2] | |
| | | [SC10]-[AC3] | |
| | | [SC10]-[AC5] | |
| | | [SC10]-[AC6] | |
| | | [SC10]-[AC7] | |
| | | [SC10]-[AC8] | |
| | | [SC10]-[AP1] | |
| | | [SC10]-[AP2] | |
| | | [SC10]-[AP3] | |
| | | [SC10]-[AP5] | |
| | | [SC10]-[AP6] | |
| | | [SC10]-[SC2] | |
| | | [SC10]-[SC3] | |
| | | [SC10]-[SC4] | |
| | | [SC10]-[SC5] | |
| | | [SC10]-[SC8] | |
| | | [SC11]-[AC1] | |
| | | [SC11]-[AC2] | |
| | | [SC11]-[AC6] | |
| | | [SC11]-[AP1] | |

| | | | |
|---|---|---|---|
| | | [SC11]-[AP2] | |
| | | [SC11]-[AP6] | |
| | | [SC11]-[SC1] | |
| | | [SC11]-[SC2] | |
| | | [SC11]-[SC3] | |
| | | [SC11]-[SC6] | |
| | | [SC12]-[AC9] | |
| | | [SC12]-[AP1] | |
| | | [SC12]-[AP3] | |
| | | [SC12]-[SC1] | |
| | | [SC12]-[SC11] | |
| | | [SC12]-[SP5] | |
| | | [SC2]-[AC5] | |
| | | [SC2]-[AC7] | |
| | | [SC2]-[AC8] | |
| | | [SC2]-[AP1] | |
| | | [SC2]-[AP2] | |
| | | [SC2]-[AP4] | |
| | | [SC2]-[SC12] | |
| | | [SC2]-[SC7] | |
| | | [SC2]-[SP2] | |
| | | [SC3]-[AC12] | |
| | | [SC3]-[AC2] | |
| | | [SC3]-[AC7] | |
| | | [SC3]-[AC8] | |
| | | [SC3]-[AP1] | |
| | | [SC3]-[AP2] | |
| | | [SC3]-[AP3] | |
| | | [SC3]-[AP6] | |
| | | [SC3]-[SC11] | |
| | | [SC3]-[SC12] | |
| | | [SC3]-[SC7] | |
| | | [SC3]-[SC8] | |
| | | [SC3]-[SC9] | |
| | | [SC3]-[SP3] | |

| | | | |
|---|---|---|---|
| | | [SC4]-[AC1] | |
| | | [SC4]-[AC10] | |
| | | [SC4]-[AC11] | |
| | | [SC4]-[AC7] | |
| | | [SC4]-[AC8] | |
| | | [SC4]-[AC9] | |
| | | [SC4]-[AP1] | |
| | | [SC4]-[AP5] | |
| | | [SC4]-[AP6] | |
| | | [SC4]-[SC1] | |
| | | [SC4]-[SC10] | |
| | | [SC4]-[SC12] | |
| | | [SC4]-[SC9] | |
| | | [SC4]-[SP6] | |
| | | [SC5]-[AC1] | |
| | | [SC5]-[AC10] | |
| | | [SC5]-[AC6] | |
| | | [SC5]-[AP1] | |
| | | [SC5]-[AP2] | |
| | | [SC5]-[AP3] | |
| | | [SC5]-[AP5] | |
| | | [SC5]-[AP6] | |
| | | [SC5]-[SC1] | |
| | | [SC5]-[SC7] | |
| | | [SC5]-[SC8] | |
| | | [SC5]-[SP6] | |
| | | [SC6]-[AC2] | |
| | | [SC6]-[AC4] | |
| | | [SC6]-[AC5] | |
| | | [SC6]-[SC1] | |
| | | [SC6]-[SC10] | |
| | | [SC6]-[SC11] | |
| | | [SC6]-[SC2] | |
| | | [SC6]-[SC5] | |
| | | [SC6]-[SP2] | |

| | | | |
|---|---|---|---|
| | | [SC6]-[SP6] | |
| | | [SC7]-[AC2] | |
| | | [SC7]-[SC1] | |
| | | [SC7]-[SC10] | |
| | | [SC7]-[SC2] | |
| | | [SC7]-[SC9] | |
| | | [SC7]-[SP5] | |
| | | [SC8]-[AC11] | |
| | | [SC8]-[AC3] | |
| | | [SC8]-[AC4] | |
| | | [SC8]-[AP2] | |
| | | [SC8]-[SC1] | |
| | | [SC8]-[SC10] | |
| | | [SC8]-[SC4] | |
| | | [SC8]-[SC8] | |
| | | [SC8]-[SC9] | |
| | | [SC8]-[SP3] | |
| | | [SC9]-[AC12] | |
| | | [SC9]-[AC2] | |
| | | [SC9]-[AC5] | |
| | | [SC9]-[AC6] | |
| | | [SC9]-[AC7] | |
| | | [SC9]-[AC8] | |
| | | [SC9]-[AP1] | |
| | | [SC9]-[AP3] | |
| | | [SC9]-[AP4] | |
| | | [SC9]-[AP5] | |
| | | [SC9]-[AP6] | |
| | | [SC9]-[SC1] | |
| | | [SC9]-[SC12] | |
| | | [SC9]-[SC5] | |
| | | [SC9]-[SC7] | |
| | | [SC9]-[SC8] | |
| | | [SC9]-[SP2] | |
| | | [SC9]-[SP3] | |

| | | | |
|---|---|---|---|
| | | [SP1]-[SC9] | |
| | | [SP2]-[AP6] | |
| | | [SP2]-[SC1] | |
| | | [SP3]-[AC4] | |
| | | [SP3]-[SC4] | |
| | | [SP3]-[SC5] | |
| | | [SP4]-[SC2] | |
| | | [SP5]-[AC1] | |
| | | [SP5]-[AC5] | |
| | | [SP5]-[SC1] | |
| | | [SP6]-[AC1] | |
| | | [SP6]-[AC2] | |
| | | [SP6]-[SC2] | |

**Table 5. 4 The list of categories in three types of H-bond composition.**

| Partial H-Bond | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Category | No H-bond | Moderate H-bond | Strong H-bond | Weak H-bond | Total | % of No H-bond | % of Moderate H-bond | % of Strong H-bond | % of Weak H-bond |
| [AC1]-[AC4] | 7 | 0 | 0 | 1 | 8 | 87.5 | 0.00 | 0.00 | 12.5 |
| [AC1]-[SC1] | 253 | 0 | 0 | 1 | 254 | 99.6 | 0.00 | 0.00 | 0.39 |
| [AC1]-[SC2] | 287 | 3 | 0 | 5 | 295 | 97.2 | 1.02 | 0.00 | 1.69 |
| [AC1]-[SC4] | 72 | 0 | 0 | 18 | 90 | 80.0 | 0.00 | 0.00 | 20.0 |
| [AC11]-[AP4] | 1 | 0 | 0 | 1 | 2 | 50.0 | 0.00 | 0.00 | 50.0 |
| [AC11]-[SC1] | 18 | 0 | 0 | 1 | 19 | 94.7 | 0.00 | 0.00 | 5.26 |
| [AC11]-[SC2] | 11 | 0 | 0 | 5 | 16 | 68.7 | 0.00 | 0.00 | 31.2 |
| [AC12]-[AC2] | 8 | 2 | 0 | 0 | 10 | 80.0 | 20.0 | 0.00 | 0.00 |
| [AC2]-[AC1] | 62 | 15 | 0 | 15 | 92 | 67.3 | 16.3 | 0.00 | 16.3 |
| [AC2]-[AC10] | 108 | 0 | 0 | 1 | 109 | 99.0 | 0.00 | 0.00 | 0.92 |
| [AC2]-[AC12] | 31 | 1 | 0 | 0 | 32 | 96.8 | 3.13 | 0.00 | 0.00 |
| [AC2]-[AC2] | 47 | 20 | 0 | 8 | 75 | 62.6 | 26.6 | 0.00 | 10.6 |
| [AC2]-[AC3] | 19 | 66 | 0 | 0 | 85 | 22.3 | 77.6 | 0.00 | 0.00 |
| [AC2]-[AC4] | 421 | 23 | 0 | 103 | 547 | 76.9 | 4.20 | 0.00 | 18.8 |
| [AC2]-[SC1] | 1433 | 246 | 0 | 1197 | 2876 | 49.8 | 8.55 | 0.00 | 41.6 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [AC2]-[SC10] | 1515 | 1 | 0 | 0 | 1516 | 99.9 | 0.07 | 0.00 | 0.00 |
| [AC2]-[SC11] | 145 | 0 | 0 | 1 | 146 | 99.3 | 0.00 | 0.00 | 0.68 |
| [AC2]-[SC2] | 444 | 40 | 0 | 160 | 644 | 68.9 | 6.21 | 0.00 | 24.8 |
| [AC2]-[SC4] | 593 | 33 | 0 | 123 | 749 | 79.1 | 4.41 | 0.00 | 16.4 |
| [AC2]-[SC5] | 4 | 3 | 0 | 0 | 7 | 57.1 | 42.8 | 0.00 | 0.00 |
| [AC2]-[SC9] | 4229 | 0 | 0 | 1 | 4230 | 99.9 | 0.00 | 0.00 | 0.02 |
| [AC2]-[SP2] | 1 | 0 | 0 | 2 | 3 | 33.3 | 0.00 | 0.00 | 66.6 |
| [AC3]-[AC1] | 90 | 1 | 0 | 19 | 110 | 81.8 | 0.91 | 0.00 | 17.2 |
| [AC3]-[AC10] | 936 | 1 | 0 | 1 | 938 | 99.7 | 0.11 | 0.00 | 0.11 |
| [AC3]-[AC2] | 7 | 0 | 0 | 5 | 12 | 58.3 | 0.00 | 0.00 | 41.6 |
| [AC3]-[AC4] | 165 | 1 | 0 | 5 | 171 | 96.4 | 0.58 | 0.00 | 2.92 |
| [AC3]-[AP4] | 11 | 1 | 0 | 2 | 14 | 78.5 | 7.14 | 0.00 | 14.2 |
| [AC3]-[AP5] | 17 | 1 | 0 | 0 | 18 | 94.4 | 5.56 | 0.00 | 0.00 |
| [AC3]-[SC1] | 32 | 6 | 0 | 72 | 110 | 29.0 | 5.45 | 0.00 | 65.4 |
| [AC3]-[SC4] | 11 | 1 | 0 | 0 | 12 | 91.6 | 8.33 | 0.00 | 0.00 |
| [AC4]-[AC3] | 4179 | 1 | 0 | 0 | 4180 | 99.9 | 0.02 | 0.00 | 0.00 |
| [AC4]-[AC4] | 7737 | 8 | 0 | 22 | 7767 | 99.6 | 0.10 | 0.00 | 0.28 |
| [AC4]-[SC4] | 1818 | 6 | 0 | 23 | 1847 | 98.4 | 0.32 | 0.00 | 1.25 |
| [AC4]-[SC5] | 1789 | 0 | 0 | 6 | 1795 | 99.6 | 0.00 | 0.00 | 0.33 |
| [AP2]-[SC1] | 517 | 0 | 0 | 2 | 519 | 99.6 | 0.00 | 0.00 | 0.39 |
| [AP2]-[SC2] | 124 | 0 | 0 | 11 | 135 | 91.8 | 0.00 | 0.00 | 8.15 |
| [AP5]-[SC2] | 11 | 0 | 0 | 7 | 18 | 61.1 | 0.00 | 0.00 | 38.8 |
| [SC1]-[AC10] | 206 | 4 | 0 | 13 | 223 | 92.3 | 1.79 | 0.00 | 5.83 |
| [SC1]-[AC11] | 40 | 2 | 0 | 7 | 49 | 81.6 | 4.08 | 0.00 | 14.2 |
| [SC1]-[AC2] | 12 | 0 | 0 | 4 | 16 | 75.0 | 0.00 | 0.00 | 25.0 |
| [SC1]-[AC3] | 4 | 5 | 0 | 6 | 15 | 26.6 | 33.3 | 0.00 | 40.0 |
| [SC1]-[AC4] | 38 | 20 | 0 | 29 | 87 | 43.6 | 22.9 | 0.00 | 33.3 |
| [SC1]-[SC10] | 1861 | 50 | 0 | 189 | 2100 | 88.6 | 2.38 | 0.00 | 9.00 |
| [SC1]-[SC11] | 141 | 1 | 0 | 24 | 166 | 84.9 | 0.60 | 0.00 | 14.4 |
| [SC1]-[SC2] | 267 | 0 | 0 | 6 | 273 | 97.8 | 0.00 | 0.00 | 2.20 |
| [SC1]-[SC4] | 182 | 93 | 0 | 221 | 496 | 36.6 | 18.7 | 0.00 | 44.5 |
| [SC1]-[SC5] | 21 | 0 | 0 | 1 | 22 | 95.4 | 0.00 | 0.00 | 4.55 |
| [SC1]-[SC9] | 3409 | 2 | 0 | 87 | 3498 | 97.4 | 0.06 | 0.00 | 2.49 |
| [SC1]-[SP3] | 32 | 0 | 0 | 2 | 34 | 94.1 | 0.00 | 0.00 | 5.88 |
| [SC10]-[AC10] | 216 | 520 | 1 | 147 | 884 | 24.4 | 58.8 | 0.11 | 16.6 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [SC10]-[AC11] | 12 | 1 | 0 | 1 | 14 | 85.7 | 7.14 | 0.00 | 7.14 |
| [SC10]-[AC4] | 1841 | 1 | 0 | 1 | 1843 | 99.8 | 0.05 | 0.00 | 0.05 |
| [SC10]-[AC9] | 16 | 0 | 0 | 1 | 17 | 94.1 | 0.00 | 0.00 | 5.88 |
| [SC10]-[AP4] | 31 | 0 | 0 | 1 | 32 | 96.8 | 0.00 | 0.00 | 3.13 |
| [SC10]-[SC10] | 268 | 2665 | 0 | 835 | 3768 | 7.11 | 70.7 | 0.00 | 22.1 |
| [SC10]-[SC11] | 8 | 1 | 0 | 8 | 17 | 47.0 | 5.88 | 0.00 | 47.0 |
| [SC10]-[SC9] | 416 | 581 | 0 | 324 | 1321 | 31.4 | 43.9 | 0.00 | 24.5 |
| [SC11]-[AC10] | 1 | 1 | 0 | 0 | 2 | 50.0 | 50.0 | 0.00 | 0.00 |
| [SC11]-[AC3] | 37 | 2 | 0 | 0 | 39 | 94.8 | 5.13 | 0.00 | 0.00 |
| [SC11]-[AC4] | 380 | 28 | 2 | 17 | 427 | 88.9 | 6.56 | 0.47 | 3.98 |
| [SC11]-[AC5] | 5 | 0 | 0 | 2 | 7 | 71.4 | 0.00 | 0.00 | 28.5 |
| [SC11]-[AP3] | 5 | 1 | 0 | 0 | 6 | 83.3 | 16.6 | 0.00 | 0.00 |
| [SC11]-[AP4] | 7 | 3 | 0 | 1 | 11 | 63.6 | 27.2 | 0.00 | 9.09 |
| [SC11]-[SC4] | 163 | 5 | 0 | 0 | 168 | 97.0 | 2.98 | 0.00 | 0.00 |
| [SC11]-[SC5] | 192 | 7 | 0 | 1 | 200 | 96.0 | 3.50 | 0.00 | 0.50 |
| [SC12]-[AC2] | 156 | 0 | 0 | 1 | 157 | 99.3 | 0.00 | 0.00 | 0.64 |
| [SC12]-[AC3] | 320 | 307 | 2 | 177 | 806 | 39.7 | 38.0 | 0.25 | 21.9 |
| [SC12]-[AC4] | 202 | 4717 | 2 | 418 | 5339 | 3.78 | 88.3 | 0.04 | 7.83 |
| [SC12]-[AC5] | 20 | 1 | 0 | 0 | 21 | 95.2 | 4.76 | 0.00 | 0.00 |
| [SC12]-[SC2] | 601 | 0 | 0 | 4 | 605 | 99.3 | 0.00 | 0.00 | 0.66 |
| [SC12]-[SC3] | 181 | 8 | 0 | 65 | 254 | 71.2 | 3.15 | 0.00 | 25.5 |
| [SC12]-[SC4] | 244 | 7482 | 2 | 813 | 8541 | 2.86 | 87.6 | 0.02 | 9.52 |
| [SC12]-[SC5] | 219 | 135 | 0 | 59 | 413 | 53.0 | 32.6 | 0.00 | 14.2 |
| [SC2]-[AC1] | 17 | 1 | 0 | 2 | 20 | 85.0 | 5.00 | 0.00 | 10.0 |
| [SC2]-[AC10] | 547 | 1152 | 0 | 1795 | 1386 | 3.94 | 83.1 | 0.00 | 12.9 |
| [SC2]-[AC11] | 71 | 294 | 1 | 98 | 464 | 15.3 | 63.3 | 0.22 | 21.1 |
| [SC2]-[AC12] | 131 | 0 | 0 | 1 | 132 | 99.2 | 0.00 | 0.00 | 0.76 |
| [SC2]-[AC2] | 30 | 19 | 0 | 22 | 71 | 42.2 | 26.7 | 0.00 | 30.9 |
| [SC2]-[AC4] | 30 | 1259 | 0 | 233 | 1522 | 1.97 | 82.7 | 0.00 | 15.3 |
| [SC2]-[AC9] | 9 | 2 | 0 | 1 | 12 | 75.0 | 16.6 | 0.00 | 8.33 |
| [SC2]-[AP3] | 1 | 2 | 0 | 0 | 3 | 33.3 | 66.6 | 0.00 | 0.00 |
| [SC2]-[AP5] | 25 | 1 | 0 | 1 | 27 | 92.5 | 3.70 | 0.00 | 3.70 |
| [SC2]-[SC1] | 71 | 0 | 0 | 3 | 74 | 95.9 | 0.00 | 0.00 | 4.05 |
| [SC2]-[SC10] | 2075 | 2748 | 4 | 6068 | 3562 | 5.82 | 77.1 | 0.01 | 17.0 |
| [SC2]-[SC11] | 296 | 45 | 0 | 112 | 453 | 65.3 | 9.93 | 0.00 | 24.7 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [SC2]-[SC2] | 354 | 46 | 0 | 120 | 520 | 68.0 | 8.85 | 0.00 | 23.0 |
| [SC2]-[SC4] | 8 | 1091 | 1 | 73 | 1173 | 0.68 | 93.0 | 0.09 | 6.22 |
| [SC2]-[SC5] | 5 | 3 | 0 | 16 | 24 | 20.8 | 12.5 | 0.00 | 66.6 |
| [SC2]-[SC8] | 18 | 0 | 0 | 2 | 20 | 90.0 | 0.00 | 0.00 | 10.0 |
| [SC2]-[SC9] | 3815 | 1111 | 0 | 1442 | 6368 | 59.9 | 17.4 | 0.00 | 22.6 |
| [SC2]-[SP3] | 43 | 3 | 0 | 0 | 46 | 93.4 | 6.52 | 0.00 | 0.00 |
| [SC3]-[AC10] | 83 | 5 | 0 | 2 | 90 | 92.2 | 5.56 | 0.00 | 2.22 |
| [SC3]-[AC11] | 15 | 0 | 0 | 1 | 16 | 93.7 | 0.00 | 0.00 | 6.25 |
| [SC3]-[AC3] | 19 | 0 | 0 | 8 | 27 | 70.3 | 0.00 | 0.00 | 29.6 |
| [SC3]-[AC4] | 18 | 11 | 1 | 13 | 43 | 41.8 | 25.5 | 2.33 | 30.2 |
| [SC3]-[AC5] | 10 | 1 | 0 | 2 | 13 | 76.9 | 7.69 | 0.00 | 15.3 |
| [SC3]-[AC9] | 19 | 1 | 0 | 0 | 20 | 95.0 | 5.00 | 0.00 | 0.00 |
| [SC3]-[AP4] | 7 | 15 | 0 | 6 | 28 | 25.0 | 53.5 | 0.00 | 21.4 |
| [SC3]-[AP5] | 9 | 2 | 0 | 2 | 13 | 69.2 | 15.3 | 0.00 | 15.3 |
| [SC3]-[SC10] | 113 | 0 | 0 | 3 | 116 | 97.4 | 0.00 | 0.00 | 2.59 |
| [SC4]-[AC2] | 8439 | 0 | 0 | 14 | 8453 | 99.8 | 0.00 | 0.00 | 0.17 |
| [SC4]-[AC3] | 8077 | 958 | 0 | 944 | 9979 | 80.9 | 9.60 | 0.00 | 9.46 |
| [SC4]-[AC4] | 1571 | 4727 | 4 | 1848 | 8147 | 19.2 | 58.0 | 0.00 | 22.6 |
| [SC4]-[AC5] | 6078 | 20 | 0 | 345 | 6443 | 94.3 | 0.31 | 0.00 | 5.35 |
| [SC4]-[AC6] | 577 | 1 | 0 | 0 | 578 | 99.8 | 0.17 | 0.00 | 0.00 |
| [SC4]-[AP2] | 1053 | 1 | 0 | 0 | 1054 | 99.9 | 0.09 | 0.00 | 0.00 |
| [SC4]-[AP3] | 669 | 1 | 0 | 2 | 672 | 99.5 | 0.15 | 0.00 | 0.30 |
| [SC4]-[AP4] | 95 | 3 | 0 | 1 | 99 | 95.9 | 3.03 | 0.00 | 1.01 |
| [SC4]-[SC2] | 1205 | 0 | 0 | 2 | 1207 | 99.8 | 0.00 | 0.00 | 0.17 |
| [SC4]-[SC3] | 1598 | 5 | 0 | 47 | 1650 | 96.8 | 0.30 | 0.00 | 2.85 |
| [SC4]-[SC4] | 4008 | 6913 | 6 | 1159 | 8474 | 4.73 | 81.5 | 0.01 | 13.6 |
| [SC4]-[SC5] | 5863 | 3530 | 0 | 3904 | 1329 | 44.0 | 26.5 | 0.00 | 29.3 |
| [SC4]-[SC6] | 22 | 0 | 0 | 1 | 23 | 95.6 | 0.00 | 0.00 | 4.35 |
| [SC4]-[SP5] | 2 | 1 | 0 | 1 | 4 | 50.0 | 25.0 | 0.00 | 25.0 |
| [SC5]-[AC11] | 2 | 1 | 0 | 0 | 3 | 66.6 | 33.3 | 0.00 | 0.00 |
| [SC5]-[AC12] | 2 | 1 | 0 | 0 | 3 | 66.6 | 33.3 | 0.00 | 0.00 |
| [SC5]-[AC2] | 9366 | 6 | 0 | 633 | 1000 | 93.6 | 0.06 | 0.00 | 6.33 |
| [SC5]-[AC3] | 3922 | 960 | 0 | 643 | 5525 | 70.9 | 17.3 | 0.00 | 11.6 |
| [SC5]-[AC4] | 9008 | 1166 | 0 | 4065 | 2473 | 36.4 | 47.1 | 0.00 | 16.4 |
| [SC5]-[AC5] | 2110 | 4 | 0 | 157 | 2271 | 92.9 | 0.18 | 0.00 | 6.91 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [SC5]-[AP4] | 60 | 1 | 0 | 0 | 61 | 98.3 | 1.64 | 0.00 | 0.00 |
| [SC5]-[SC10] | 1 | 1 | 0 | 3 | 5 | 20.0 | 20.0 | 0.00 | 60.0 |
| [SC5]-[SC2] | 1806 | 1 | 0 | 54 | 1861 | 97.0 | 0.05 | 0.00 | 2.90 |
| [SC5]-[SC3] | 741 | 28 | 0 | 79 | 848 | 87.3 | 3.30 | 0.00 | 9.32 |
| [SC5]-[SC4] | 3728 | 4721 | 14 | 4822 | 5577 | 6.68 | 84.6 | 0.03 | 8.65 |
| [SC5]-[SC5] | 5775 | 5377 | 2 | 5143 | 1629 | 35.4 | 32.9 | 0.01 | 31.5 |
| [SC5]-[SC6] | 13 | 0 | 0 | 1 | 14 | 92.8 | 0.00 | 0.00 | 7.14 |
| [SC5]-[SP5] | 1 | 3 | 0 | 0 | 4 | 25.0 | 75.0 | 0.00 | 0.00 |
| [SC6]-[AC3] | 12 | 0 | 0 | 1 | 13 | 92.3 | 0.00 | 0.00 | 7.69 |
| [SC6]-[SC3] | 12 | 1 | 0 | 1 | 14 | 85.7 | 7.14 | 0.00 | 7.14 |
| [SC6]-[SC4] | 50 | 0 | 0 | 1 | 51 | 98.0 | 0.00 | 0.00 | 1.96 |
| [SC6]-[SP5] | 1 | 0 | 0 | 1 | 2 | 50.0 | 0.00 | 0.00 | 50.0 |
| [SC7]-[AC11] | 1 | 1 | 0 | 0 | 2 | 50.0 | 50.0 | 0.00 | 0.00 |
| [SC7]-[AC3] | 55 | 0 | 0 | 2 | 57 | 96.4 | 0.00 | 0.00 | 3.51 |
| [SC7]-[AC4] | 50 | 1 | 0 | 5 | 56 | 89.2 | 1.79 | 0.00 | 8.93 |
| [SC7]-[SC11] | 9 | 1 | 0 | 0 | 10 | 90.0 | 10.0 | 0.00 | 0.00 |
| [SC7]-[SC3] | 262 | 0 | 0 | 4 | 266 | 98.5 | 0.00 | 0.00 | 1.50 |
| [SC7]-[SC4] | 690 | 15 | 0 | 39 | 744 | 92.7 | 2.02 | 0.00 | 5.24 |
| [SC7]-[SC5] | 375 | 1 | 0 | 5 | 381 | 98.4 | 0.26 | 0.00 | 1.31 |
| [SC8]-[SC11] | 1 | 0 | 0 | 1 | 2 | 50.0 | 0.00 | 0.00 | 50.0 |
| [SC9]-[AC1] | 53 | 1 | 0 | 0 | 54 | 98.1 | 1.85 | 0.00 | 0.00 |
| [SC9]-[AC10] | 186 | 1992 | 0 | 200 | 2378 | 7.82 | 83.7 | 0.00 | 8.41 |
| [SC9]-[AC11] | 16 | 28 | 0 | 6 | 50 | 32.0 | 56.0 | 0.00 | 12.0 |
| [SC9]-[AC3] | 4 | 2 | 0 | 0 | 6 | 66.6 | 33.3 | 0.00 | 0.00 |
| [SC9]-[AC4] | 25 | 0 | 0 | 2 | 27 | 92.5 | 0.00 | 0.00 | 7.41 |
| [SC9]-[AC9] | 6 | 0 | 0 | 6 | 12 | 50.0 | 0.00 | 0.00 | 50.0 |
| [SC9]-[SC10] | 171 | 1304 | 5 | 913 | 1413 | 1.21 | 92.3 | 0.04 | 6.46 |
| [SC9]-[SC11] | 82 | 9 | 0 | 83 | 174 | 47.1 | 5.17 | 0.00 | 47.7 |
| [SC9]-[SC2] | 22 | 2 | 0 | 0 | 24 | 91.6 | 8.33 | 0.00 | 0.00 |
| [SC9]-[SC9] | 343 | 885 | 0 | 492 | 1720 | 19.9 | 51.4 | 0.00 | 28.6 |
| [SP2]-[AC10] | 1 | 9 | 0 | 2 | 12 | 8.33 | 75.0 | 0.00 | 16.6 |
| [SP2]-[AP5] | 1 | 1 | 0 | 0 | 2 | 50.0 | 50.0 | 0.00 | 0.00 |
| [SP2]-[SC9] | 3 | 2 | 0 | 0 | 5 | 60.0 | 40.0 | 0.00 | 0.00 |
| [SP5]-[AC2] | 5 | 0 | 0 | 1 | 6 | 83.3 | 0.00 | 0.00 | 16.6 |
| [SP5]-[AC3] | 1 | 13 | 0 | 2 | 16 | 6.25 | 81.2 | 0.00 | 12.5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [SP5]-[AC4] | 1 | 26 | 0 | 4 | 31 | 3.23 | 83.8 | 0.00 | 12.9 |
| [SP5]-[AP4] | 1 | 0 | 0 | 1 | 2 | 50.0 | 0.00 | 0.00 | 50.0 |
| [SP5]-[SC2] | 2 | 0 | 0 | 1 | 3 | 66.6 | 0.00 | 0.00 | 33.3 |
| [SP5]-[SC5] | 1 | 16 | 0 | 3 | 20 | 5.00 | 80.0 | 0.00 | 15.0 |
| [SP6]-[SP5] | 1 | 1 | 0 | 0 | 2 | 50.0 | 50.0 | 0.00 | 0.00 |

**Table 5. 5 The strength of H-bond in each category for the type of partial H-Bond.**

| All H-Bond | | | | | | | |
|---|---|---|---|---|---|---|---|
| Category | Moderate H-bond | Strong H-bond | Weak H-bond | Total | % of Moderate H-bond | % of Strong H-bond | % of Weak H-bond |
| [AC3]-[SC2] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [AC4]-[AC12] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [AC9]-[AC3] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [AC10]-[SC2] | 0 | 0 | 2 | 2 | 0.00 | 0.00 | 1.00 |
| [AP1]-[AC2] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SC1]-[AP6] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SC12]-[AC11] | 4 | 0 | 0 | 4 | 100.00 | 0.00 | 0.00 |
| [SC12]-[AC12] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SC12]-[AP4] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SC12]-[SC10] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SC2]-[AC3] | 209 | 1 | 4 | 214 | 97.66 | 0.00 | 0.02 |
| [SC2]-[SC3] | 2 | 0 | 0 | 2 | 100.00 | 0.00 | 0.00 |
| [SP2]-[AC11] | 6 | 0 | 0 | 6 | 100.00 | 0.00 | 0.00 |
| [SP2]-[AC3] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP2]-[SC10] | 22 | 0 | 0 | 22 | 100.00 | 0.00 | 0.00 |
| [SP2]-[SC11] | 3 | 0 | 0 | 3 | 100.00 | 0.00 | 0.00 |
| [SP2]-[SP3] | 2 | 0 | 0 | 2 | 100.00 | 0.00 | 0.00 |
| [SP2]-[SC10] | 0 | 0 | 2 | 2 | 0.00 | 0.00 | 1.00 |
| [SP3]-[AC10] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[AC11] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[AC3] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[AP4] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[AP6] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[SC10] | 25 | 0 | 0 | 25 | 100.00 | 0.00 | 0.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [SP3]-[SC2] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP3]-[SC9] | 3 | 0 | 0 | 3 | 100.00 | 0.00 | 0.00 |
| [SP3]-[SP3] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP4]-[AC10] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP4]-[AC4] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP4]-[SC4] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP5]-[SC3] | 2 | 0 | 1 | 3 | 66.67 | 0.00 | 0.33 |
| [SP5]-[SC4] | 34 | 1 | 3 | 38 | 89.47 | 0.03 | 0.08 |
| [SP5]-[SC6] | 0 | 1 | 0 | 1 | 0.00 | 1.00 | 0.00 |
| [SP5]-[SP4] | 0 | 0 | 1 | 1 | 0.00 | 0.00 | 1.00 |
| [SP5]-[AP2] | 0 | 0 | 1 | 1 | 0.00 | 0.00 | 1.00 |
| [SP6]-[AC11] | 1 | 0 | 0 | 1 | 100.00 | 0.00 | 0.00 |
| [SP6]-[AC3] | 6 | 0 | 1 | 7 | 85.71 | 0.00 | 0.14 |
| [SP6]-[AC4] | 4 | 0 | 3 | 7 | 57.14 | 0.00 | 0.43 |
| [SP6]-[SC4] | 10 | 0 | 0 | 10 | 100.00 | 0.00 | 0.00 |
| [SP6]-[SC5] | 3 | 0 | 1 | 4 | 75.00 | 0.00 | 0.25 |

**Table 5. 6 The strength of H-bond in each category for the type of all H-Bond.**

| H-bond status | # of beta turns | Percentage of each |
|---|---|---|
| **Some H-Bonds** | **160** | **27.444%** |
| **No H-Bond** | **384** | **65.866%** |
| **All H-Bond** | **39** | **6.690%** |

**Table 5. 7 Summary of hydrogen bond status in different turn types in the large protein database.**

### 5.4.4 Distances between Cα atoms of the i and i+3 residues in β turns

The distance between the Cα atom of the i and i+3 residues is an important criterion for identifying β turns. The frequency of the number of new turn types as a function of the distance of separation between the Cα atom of the i and i+3 residues is summarized in **Table 5.8**. The data indicated that the majority of new turn types, i.e. 515 out of 583, i.e. 88.3%, had distances between the Cα atom of the i and i+3 residues between 6 and 7 Å. Ten turn types or 1.72% of turns had a distance > 7 Å and ~2.5% of turn types had a distance < 5 Å. The mean and standard deviation of the distances between the Cα atom of the i and i+3 residues for each turn type is included in **Table 5.9**. Plots of the distributions of the distances between the Cα atom of the i and i+3 residues revealed that some turn types had a strongly preferred separation distance which resulted in a single peak in the distribution plot, e.g. in the SC4-AC4 turn type shown which has a single peak in the distance distribution plot centered at 5.4 Å (**Figure 5.25)**. Eleven of the top 19 new turn types exhibited a single peak in the distance distribution, including [SC4]-[SC4], [SC4]-[AC4], [SC12]-[AC4], [SC12]-[SC4], [SC2]-[AC10], [SC2]-[SC10],

[SC5]-[AC4], [SC5]-[SC4], [SC5]-[SC5], [SC9]-[SC10], [SC4]-[SC5], which can be inspected in **Figure 5.26**. In contrast, nine of the top 19 new turn types exhibited a broad distribution of distance as opposed to a single preferred distance **(Figure 5.26)**, e.g. SC5-AC3 which had a broad distribution of distances that spanned 4.5 - 6 Å with high occurrences **(Figure 5.25).** The details regarding the distance distributions for the top 19 new turn types are available in **Figure 5.26**.

| Distance(Å) | Number of turn types | Percentage |
|---|---|---|
| <4.0 | 1 | 0.17 |
| <5.0 | 14 | 2.40 |
| <5.5 | 43 | 7.38 |
| <6.0 | 143 | 24.53 |
| <6.5 | 216 | 37.05 |
| <7.0 | 156 | 26.76 |
| =7.0 | 10 | 1.72 |
| Total | 583 | 100 |

**Table 5. 8 Distribution of the number of new turn types according to the distance between the of Cα atoms of the i and i+3 residues.**

| Category | Mean of distance | Std |
|---|---|---|
| [AC5]-[SC9] | 3.80 | 0.000 |
| [AC3]-[SC1] | 4.20 | 0.332 |
| [SC9]-[AC2] | 4.20 | 0.000 |
| [SC9]-[SC1] | 4.23 | 0.256 |
| [SP3]-[AC3] | 4.40 | 0.000 |
| [SP6]-[AC1] | 4.40 | 0.000 |
| [AP3]-[AC1] | 4.50 | 0.183 |
| [SP3]-[AP6] | 4.50 | 0.000 |
| [SC9]-[SC2] | 4.59 | 0.559 |
| [SC1]-[AP6] | 4.60 | 0.000 |
| [AP3]-[SC1] | 4.63 | 0.493 |
| [SC9]-[AC1] | 4.72 | 0.767 |
| [AP2]-[AC7] | 4.80 | 0.000 |
| [SC5]-[AC11] | 4.87 | 0.208 |
| [SP5]-[SC6] | 4.90 | 0.000 |
| [AC4]-[AC12] | 5.00 | 0.000 |
| [SC1]-[AP3] | 5.00 | 0.000 |

| | | |
|---|---|---|
| [SC12]-[AC9] | 5.00 | 0.000 |
| [SC5]-[AC12] | 5.07 | 0.551 |
| [SP3]-[AC11] | 5.10 | 0.000 |
| [AC9]-[SC2] | 5.15 | 0.071 |
| [AP3]-[AC4] | 5.17 | 1.159 |
| [AC3]-[SC2] | 5.20 | 0.000 |
| [AC5]-[AC12] | 5.20 | 0.000 |
| [SP5]-[AC3] | 5.20 | 0.429 |
| [AP2]-[SC2] | 5.23 | 0.841 |
| [AC2]-[SC4] | 5.25 | 0.791 |
| [AC3]-[AC1] | 5.25 | 0.996 |
| [SP5]-[SC3] | 5.27 | 0.551 |
| [AC2]-[SC5] | 5.28 | 0.471 |
| [AC11]-[AP4] | 5.30 | 0.566 |
| [AC6]-[AP3] | 5.30 | 0.000 |
| [AP1]-[AC2] | 5.30 | 0.000 |
| [SC9]-[AC9] | 5.30 | 0.158 |
| [SP5]-[AC4] | 5.30 | 0.460 |
| [SP5]-[AP2] | 5.30 | 0.000 |
| [SC12]-[AC11] | 5.33 | 0.263 |
| [SC3]-[AC11] | 5.36 | 0.435 |
| [AC2]-[SP2] | 5.37 | 0.723 |
| [AP2]-[AC1] | 5.40 | 0.678 |
| [AP3]-[SC3] | 5.40 | 0.000 |
| [AP5]-[SC6] | 5.40 | 0.000 |
| [SC2]-[AP2] | 5.40 | 0.000 |
| [SC5]-[SC7] | 5.40 | 0.283 |
| [SC6]-[SP5] | 5.40 | 0.849 |
| [SC9]-[AC6] | 5.40 | 0.000 |
| [SP2]-[SC11] | 5.40 | 0.400 |
| [SP3]-[AC10] | 5.40 | 0.000 |
| [SP3]-[SC2] | 5.40 | 0.000 |
| [SC4]-[AC3] | 5.43 | 0.971 |
| [SC4]-[SC5] | 5.43 | 0.966 |
| [SC4]-[AC4] | 5.44 | 0.955 |

| | | |
|---|---|---|
| [AP2]-[SC4] | 5.45 | 0.650 |
| [SC2]-[SC3] | 5.45 | 0.071 |
| [SC5]-[AC3] | 5.45 | 0.938 |
| [SC9]-[AC10] | 5.46 | 0.683 |
| [SP2]-[AC11] | 5.47 | 0.306 |
| [AC3]-[AC2] | 5.48 | 0.987 |
| [AC2]-[SC1] | 5.50 | 0.909 |
| [SC2]-[SC9] | 5.50 | 0.909 |
| [SC5]-[AC2] | 5.50 | 0.909 |
| [SC5]-[AC4] | 5.50 | 0.851 |
| [SC5]-[SC4] | 5.50 | 0.915 |
| [SC8]-[AC3] | 5.50 | 0.000 |
| [SC8]-[SC11] | 5.50 | 0.707 |
| [SP2]-[AC3] | 5.50 | 0.000 |
| [SP3]-[SC10] | 5.50 | 0.274 |
| [SP6]-[AC11] | 5.50 | 0.000 |
| [SC4]-[AC5] | 5.53 | 0.920 |
| [SP2]-[AC10] | 5.54 | 0.665 |
| [AC2]-[SC2] | 5.55 | 0.880 |
| [AC3]-[AC4] | 5.55 | 0.984 |
| [SC5]-[AC5] | 5.55 | 0.841 |
| [SC5]-[SC5] | 5.55 | 0.886 |
| [SC9]-[AC11] | 5.55 | 0.532 |
| [SC11]-[SC5] | 5.56 | 0.722 |
| [SC3]-[AC10] | 5.56 | 0.624 |
| [SP6]-[AC4] | 5.57 | 0.280 |
| [AC8]-[AC4] | 5.60 | 0.000 |
| [AC9]-[AC3] | 5.60 | 0.000 |
| [AP4]-[AP4] | 5.60 | 0.141 |
| [SC11]-[AP3] | 5.60 | 0.265 |
| [SC12]-[SC10] | 5.60 | 0.000 |
| [SC2]-[SC10] | 5.60 | 0.851 |
| [SC4]-[AP3] | 5.60 | 0.806 |
| [SC5]-[SC3] | 5.60 | 0.851 |
| [SC8]-[AC4] | 5.60 | 0.100 |

| | | |
|---|---|---|
| [SC8]-[SC4] | 5.60 | 0.200 |
| [SP2]-[SC10] | 5.60 | 0.424 |
| [SP2]-[SP3] | 5.60 | 0.000 |
| [SP3]-[AP4] | 5.60 | 0.000 |
| [SP4]-[AC10] | 5.60 | 0.000 |
| [SP6]-[SC4] | 5.60 | 0.100 |
| [SC3]-[AC9] | 5.63 | 0.411 |
| [SC2]-[AP4] | 5.65 | 0.919 |
| [SC9]-[SC10] | 5.65 | 0.649 |
| [SP5]-[SC5] | 5.65 | 0.598 |
| [AC4]-[AC4] | 5.66 | 0.858 |
| [AP2]-[SC1] | 5.66 | 0.864 |
| [SC5]-[AP3] | 5.67 | 0.758 |
| [SC5]-[SC10] | 5.67 | 0.830 |
| [AC2]-[AC4] | 5.68 | 0.785 |
| [SC11]-[AP4] | 5.68 | 0.337 |
| [AC10]-[SC2] | 5.70 | 0.000 |
| [AC4]-[SC4] | 5.70 | 0.794 |
| [AC4]-[SC5] | 5.70 | 0.794 |
| [SC12]-[AC12] | 5.70 | 0.000 |
| [SC12]-[SC4] | 5.70 | 0.794 |
| [SC4]-[SC4] | 5.70 | 0.800 |
| [SC8]-[AP2] | 5.70 | 0.000 |
| [SP3]-[SP3] | 5.70 | 0.000 |
| [SP4]-[SC4] | 5.70 | 0.000 |
| [AC11]-[SC5] | 5.71 | 0.829 |
| [SC3]-[AP5] | 5.71 | 0.746 |
| [SC6]-[AC3] | 5.71 | 0.729 |
| [AC2]-[AC3] | 5.72 | 0.577 |
| [SC1]-[SC9] | 5.72 | 0.815 |
| [SC5]-[SP5] | 5.72 | 0.727 |
| [SP6]-[AC3] | 5.72 | 0.340 |
| [SP3]-[SC9] | 5.73 | 0.208 |
| [AC6]-[SC4] | 5.74 | 0.837 |
| [AC2]-[SC9] | 5.75 | 0.765 |

| | | |
|---|---|---|
| [AC3]-[AC11] | 5.75 | 0.642 |
| [AC3]-[SC4] | 5.75 | 1.062 |
| [AC5]-[AC4] | 5.75 | 0.765 |
| [AC5]-[SC4] | 5.75 | 0.765 |
| [SC11]-[AC4] | 5.75 | 0.762 |
| [SC12]-[AC4] | 5.75 | 0.765 |
| [SC12]-[SC5] | 5.75 | 0.765 |
| [SC4]-[AP4] | 5.75 | 0.553 |
| [SC9]-[SC9] | 5.75 | 0.707 |
| [SC2]-[AC11] | 5.76 | 0.748 |
| [SC3]-[SC9] | 5.76 | 0.756 |
| [SC4]-[SC3] | 5.76 | 0.805 |
| [AC4]-[AC3] | 5.78 | 0.773 |
| [SC2]-[AC10] | 5.78 | 0.767 |
| [SP5]-[SC4] | 5.78 | 0.456 |
| [SC9]-[AC4] | 5.79 | 1.004 |
| [AC10]-[AP4] | 5.80 | 0.000 |
| [AC2]-[SP3] | 5.80 | 0.718 |
| [AC8]-[SC1] | 5.80 | 0.100 |
| [AC8]-[SC2] | 5.80 | 0.000 |
| [AP2]-[AC3] | 5.80 | 0.000 |
| [AP3]-[SC11] | 5.80 | 0.141 |
| [SC12]-[AP1] | 5.80 | 0.000 |
| [SC12]-[AP4] | 5.80 | 0.000 |
| [SC12]-[SC3] | 5.80 | 0.736 |
| [SC2]-[AC3] | 5.80 | 0.389 |
| [SC2]-[SC4] | 5.80 | 0.563 |
| [SC4]-[AC2] | 5.80 | 0.736 |
| [SC5]-[SC2] | 5.80 | 0.743 |
| [SP3]-[SC5] | 5.80 | 0.000 |
| [SP6]-[SC5] | 5.80 | 0.200 |
| [SC5]-[AP4] | 5.81 | 0.625 |
| [AC3]-[AP4] | 5.82 | 0.793 |
| [AC4]-[AC5] | 5.82 | 0.758 |
| [AC2]-[AC1] | 5.83 | 0.662 |

| | | |
|---|---|---|
| [AC2]-[SC10] | 5.83 | 0.746 |
| [SC2]-[SC2] | 5.83 | 0.737 |
| [SC6]-[SC3] | 5.83 | 0.566 |
| [AC5]-[SC5] | 5.84 | 0.796 |
| [AC3]-[AC9] | 5.85 | 0.493 |
| [SC3]-[AP4] | 5.85 | 0.564 |
| [SC11]-[AC3] | 5.86 | 0.699 |
| [SP2]-[SC9] | 5.86 | 0.691 |
| [AC5]-[AC5] | 5.87 | 0.733 |
| [AC6]-[SC5] | 5.87 | 0.743 |
| [SC2]-[SC11] | 5.87 | 0.692 |
| [SC3]-[SC11] | 5.88 | 0.781 |
| [AC3]-[SC10] | 5.89 | 0.694 |
| [SC11]-[SC4] | 5.89 | 0.699 |
| [AC3]-[AC10] | 5.90 | 0.678 |
| [SC1]-[SC4] | 5.90 | 0.846 |
| [SC10]-[SC10] | 5.90 | 0.678 |
| [SC12]-[AC3] | 5.90 | 0.678 |
| [SC3]-[SC10] | 5.90 | 0.728 |
| [SC5]-[AP2] | 5.90 | 0.686 |
| [SC6]-[SP2] | 5.90 | 0.000 |
| [SP2]-[AP5] | 5.90 | 0.283 |
| [SP5]-[AC5] | 5.90 | 0.141 |
| [SP6]-[SC2] | 5.90 | 0.000 |
| [AC3]-[AP5] | 5.91 | 0.697 |
| [SC9]-[SC11] | 5.91 | 0.577 |
| [AC2]-[AC2] | 5.92 | 0.587 |
| [SC11]-[SC3] | 5.92 | 0.856 |
| [SC2]-[SC8] | 5.93 | 0.725 |
| [AC11]-[AC3] | 5.94 | 0.661 |
| [AC11]-[AC4] | 5.94 | 0.666 |
| [SC5]-[AC1] | 5.94 | 0.675 |
| [AC5]-[SC3] | 5.95 | 0.649 |
| [AP3]-[AC11] | 5.95 | 0.071 |
| [AP3]-[AC2] | 5.95 | 0.071 |

| | | |
|---|---|---|
| [SC10]-[AC10] | 5.95 | 0.649 |
| [SC2]-[AC4] | 5.95 | 0.592 |
| [SP5]-[SC1] | 5.95 | 0.212 |
| [SC10]-[AC9] | 5.97 | 0.550 |
| [SC2]-[AP3] | 5.97 | 0.404 |
| [AC11]-[AC5] | 5.98 | 0.483 |
| [AC12]-[SC5] | 5.98 | 0.593 |
| [AC4]-[AP4] | 5.98 | 0.695 |
| [SC1]-[SP3] | 5.99 | 0.565 |
| [AC10]-[AP3] | 6.00 | 0.442 |
| [AC4]-[AC10] | 6.00 | 0.686 |
| [AC4]-[AP3] | 6.00 | 0.620 |
| [AC5]-[AC3] | 6.00 | 0.620 |
| [AC9]-[AP5] | 6.00 | 0.000 |
| [AP2]-[SP3] | 6.00 | 0.000 |
| [SC1]-[AC10] | 6.00 | 0.629 |
| [SC1]-[SC10] | 6.00 | 0.620 |
| [SC11]-[AC5] | 6.00 | 0.432 |
| [SC4]-[SP5] | 6.00 | 0.361 |
| [SC5]-[SC8] | 6.00 | 0.707 |
| [SC6]-[SC11] | 6.00 | 0.990 |
| [SP3]-[SC4] | 6.00 | 0.000 |
| [SP4]-[AC4] | 6.00 | 0.000 |
| [AC2]-[SC11] | 6.01 | 0.633 |
| [AC11]-[SC4] | 6.02 | 0.593 |
| [SC10]-[AC4] | 6.02 | 0.640 |
| [SC2]-[AC9] | 6.02 | 0.680 |
| [AC12]-[SC4] | 6.03 | 0.659 |
| [SC11]-[AC6] | 6.03 | 0.493 |
| [AP2]-[AC4] | 6.04 | 0.581 |
| [SC2]-[SC5] | 6.04 | 0.498 |
| [SC4]-[AP2] | 6.04 | 0.620 |
| [AC3]-[SC9] | 6.05 | 0.592 |
| [AP2]-[SC9] | 6.05 | 0.592 |
| [SC1]-[SC11] | 6.05 | 0.592 |

| | | |
|---|---|---|
| [SC10]-[SC9] | 6.05 | 0.592 |
| [SC6]-[AC4] | 6.05 | 0.481 |
| [SP6]-[SP5] | 6.05 | 0.354 |
| [SC1]-[SC2] | 6.06 | 0.715 |
| [SC4]-[AC6] | 6.06 | 0.616 |
| [AC2]-[AC10] | 6.07 | 0.612 |
| [AP6]-[AC3] | 6.07 | 0.611 |
| [SC4]-[SC12] | 6.07 | 0.153 |
| [SC5]-[AC10] | 6.07 | 1.069 |
| [SC10]-[AP5] | 6.08 | 0.623 |
| [SC7]-[SC4] | 6.08 | 0.599 |
| [AC10]-[AC4] | 6.09 | 0.572 |
| [AC3]-[SC11] | 6.09 | 0.711 |
| [SC1]-[AC11] | 6.09 | 0.557 |
| [SC9]-[SC12] | 6.09 | 0.699 |
| [AC3]-[AC5] | 6.10 | 0.569 |
| [AP3]-[AC5] | 6.10 | 0.000 |
| [SC1]-[AC3] | 6.10 | 0.506 |
| [SC10]-[AC11] | 6.10 | 0.578 |
| [SC8]-[AC11] | 6.10 | 0.000 |
| [SC9]-[SC8] | 6.10 | 0.557 |
| [SC2]-[AC2] | 6.11 | 0.591 |
| [SC3]-[AC4] | 6.11 | 0.606 |
| [SC6]-[SC4] | 6.11 | 0.578 |
| [AC12]-[AC3] | 6.12 | 0.746 |
| [AC3]-[AC3] | 6.12 | 0.557 |
| [AC3]-[AP3] | 6.12 | 0.555 |
| [AC9]-[SC10] | 6.12 | 0.585 |
| [SC2]-[AP5] | 6.12 | 0.455 |
| [SC7]-[SC5] | 6.12 | 0.644 |
| [AC2]-[AC5] | 6.13 | 0.263 |
| [AC4]-[AC2] | 6.13 | 0.559 |
| [AP2]-[AC2] | 6.13 | 0.557 |
| [AP5]-[SC2] | 6.13 | 0.477 |
| [AP5]-[SC5] | 6.13 | 0.512 |

| | | |
|---|---|---|
| [SC10]-[SC5] | 6.13 | 0.611 |
| [SC2]-[SP3] | 6.13 | 0.504 |
| [SP5]-[SC2] | 6.13 | 0.586 |
| [AC5]-[AC2] | 6.14 | 0.544 |
| [AC5]-[SC2] | 6.14 | 0.544 |
| [SC10]-[AP3] | 6.14 | 0.510 |
| [SC9]-[AC12] | 6.14 | 0.535 |
| [AC1]-[SC2] | 6.15 | 0.534 |
| [AC1]-[SC9] | 6.15 | 0.534 |
| [AC10]-[AC5] | 6.15 | 0.534 |
| [AC4]-[AC11] | 6.15 | 0.602 |
| [AP5]-[AC5] | 6.15 | 0.636 |
| [AP6]-[AC4] | 6.15 | 0.517 |
| [SC10]-[AC5] | 6.15 | 0.590 |
| [SC12]-[SP5] | 6.15 | 0.071 |
| [SC6]-[SC5] | 6.15 | 0.541 |
| [SC7]-[SC10] | 6.15 | 1.061 |
| [SC8]-[SC9] | 6.15 | 0.645 |
| [AC1]-[SC1] | 6.16 | 0.775 |
| [SC5]-[SC1] | 6.16 | 0.579 |
| [SC5]-[SC6] | 6.16 | 0.548 |
| [AC2]-[SC12] | 6.17 | 0.634 |
| [AP5]-[AC4] | 6.17 | 0.588 |
| [SC3]-[SC8] | 6.17 | 0.637 |
| [SC9]-[AP5] | 6.17 | 0.701 |
| [AC3]-[SC5] | 6.18 | 0.968 |
| [AC4]-[AP2] | 6.18 | 0.539 |
| [AC4]-[SC3] | 6.18 | 0.571 |
| [AP6]-[SC5] | 6.18 | 0.411 |
| [SC10]-[AP4] | 6.18 | 0.580 |
| [SC3]-[AC5] | 6.18 | 0.648 |
| [SC4]-[AC1] | 6.18 | 0.535 |
| [SC4]-[SC2] | 6.18 | 0.538 |
| [SC9]-[AC3] | 6.18 | 0.998 |
| [AC12]-[AC4] | 6.19 | 0.546 |

| | | |
|---|---|---|
| [SC1]-[SC1] | 6.19 | 0.620 |
| [SC5]-[AC6] | 6.19 | 0.524 |
| [SC9]-[AC8] | 6.19 | 0.482 |
| [AC10]-[AC3] | 6.20 | 0.505 |
| [AC11]-[AP3] | 6.20 | 0.000 |
| [AC2]-[AP5] | 6.20 | 0.000 |
| [AC6]-[SC10] | 6.20 | 0.000 |
| [AC8]-[SC11] | 6.20 | 0.000 |
| [AC9]-[AC4] | 6.20 | 0.000 |
| [AP1]-[SC12] | 6.20 | 0.000 |
| [AP5]-[SP5] | 6.20 | 0.000 |
| [SC11]-[SC6] | 6.20 | 0.000 |
| [SC5]-[SP6] | 6.20 | 0.000 |
| [SC6]-[SC10] | 6.20 | 0.624 |
| [SP5]-[AP4] | 6.20 | 0.566 |
| [AP5]-[AC3] | 6.21 | 0.807 |
| [SC9]-[AP4] | 6.22 | 0.545 |
| [AC2]-[AC11] | 6.23 | 0.476 |
| [AC4]-[SC10] | 6.23 | 0.570 |
| [AC6]-[AC4] | 6.23 | 0.512 |
| [AP4]-[SC4] | 6.23 | 0.586 |
| [SC11]-[AC2] | 6.23 | 0.512 |
| [SC7]-[SC3] | 6.23 | 0.508 |
| [AC9]-[SC9] | 6.24 | 0.512 |
| [SC10]-[SC8] | 6.24 | 0.230 |
| [SC3]-[AC12] | 6.24 | 0.513 |
| [AC5]-[AP3] | 6.25 | 0.463 |
| [AP3]-[SC9] | 6.25 | 0.560 |
| [AP5]-[SC4] | 6.25 | 0.476 |
| [SC1]-[AC4] | 6.25 | 0.550 |
| [SC10]-[AC3] | 6.25 | 0.476 |
| [SC10]-[SC11] | 6.25 | 0.469 |
| [SC11]-[AC10] | 6.25 | 0.778 |
| [SC7]-[AC11] | 6.25 | 0.212 |
| [AC6]-[SC3] | 6.26 | 0.478 |

| | | |
|---|---|---|
| [AC8]-[SC9] | 6.26 | 0.517 |
| [SC10]-[SC4] | 6.26 | 0.540 |
| [SC4]-[AP5] | 6.26 | 0.498 |
| [AC10]-[AC1] | 6.27 | 1.071 |
| [AC4]-[AC6] | 6.27 | 0.516 |
| [AP2]-[SC11] | 6.27 | 0.550 |
| [AC5]-[AC1] | 6.28 | 0.765 |
| [AP2]-[SC10] | 6.28 | 0.570 |
| [AP4]-[SC5] | 6.28 | 0.685 |
| [AP6]-[SC4] | 6.28 | 0.518 |
| [SC5]-[AP5] | 6.28 | 0.410 |
| [AC11]-[SC2] | 6.29 | 0.546 |
| [SC12]-[AC2] | 6.29 | 0.459 |
| [AC1]-[SC10] | 6.30 | 0.447 |
| [AC12]-[AC2] | 6.30 | 0.590 |
| [AC2]-[AC9] | 6.30 | 0.787 |
| [AC3]-[SC3] | 6.30 | 0.141 |
| [AC4]-[AP5] | 6.30 | 0.457 |
| [AC7]-[AC4] | 6.30 | 0.872 |
| [AP3]-[AC10] | 6.30 | 0.100 |
| [AP4]-[SC3] | 6.30 | 0.000 |
| [SC1]-[SP1] | 6.30 | 0.000 |
| [SC12]-[AP3] | 6.30 | 0.408 |
| [SC12]-[SC2] | 6.30 | 0.447 |
| [SC3]-[AP2] | 6.30 | 0.000 |
| [SC4]-[AP1] | 6.30 | 0.447 |
| [SC7]-[AC4] | 6.30 | 0.525 |
| [SC8]-[SP3] | 6.30 | 0.566 |
| [AC10]-[SC10] | 6.31 | 0.463 |
| [AC6]-[AC3] | 6.31 | 0.492 |
| [SC5]-[AP1] | 6.32 | 0.456 |
| [SC6]-[SC2] | 6.32 | 0.528 |
| [AC10]-[AC9] | 6.33 | 0.356 |
| [AC8]-[SC10] | 6.33 | 0.450 |
| [SC2]-[AC7] | 6.33 | 0.314 |

| | | |
|---|---|---|
| [SC3]-[AC8] | 6.33 | 0.393 |
| [SC4]-[SC6] | 6.33 | 0.515 |
| [SC4]-[AC8] | 6.34 | 0.728 |
| [AC10]-[SC4] | 6.35 | 0.418 |
| [AC5]-[SC1] | 6.35 | 0.528 |
| [AP1]-[SC11] | 6.35 | 0.404 |
| [SC2]-[SC1] | 6.35 | 0.418 |
| [SP5]-[AC2] | 6.35 | 0.532 |
| [AC7]-[SC5] | 6.36 | 0.566 |
| [AP4]-[AC3] | 6.36 | 0.670 |
| [SC7]-[AC2] | 6.36 | 0.793 |
| [SC7]-[AC3] | 6.36 | 0.448 |
| [SC8]-[SC10] | 6.37 | 0.550 |
| [AC12]-[SC3] | 6.38 | 0.473 |
| [SC2]-[AC5] | 6.38 | 1.030 |
| [AC4]-[SC1] | 6.39 | 0.478 |
| [SC2]-[AC12] | 6.39 | 0.486 |
| [AC1]-[AC12] | 6.40 | 0.849 |
| [AC1]-[SC4] | 6.40 | 0.389 |
| [AC2]-[SC8] | 6.40 | 0.378 |
| [AP3]-[SC10] | 6.40 | 0.576 |
| [SC1]-[SC8] | 6.40 | 0.566 |
| [SC11]-[AC1] | 6.40 | 0.389 |
| [SC3]-[SC12] | 6.40 | 0.688 |
| [SC4]-[SP6] | 6.40 | 0.000 |
| [AC11]-[SC1] | 6.41 | 0.383 |
| [SC1]-[AC2] | 6.41 | 0.535 |
| [AC4]-[AC9] | 6.42 | 0.411 |
| [AC5]-[AP4] | 6.42 | 0.421 |
| [AP5]-[AC2] | 6.42 | 0.415 |
| [SC4]-[AC10] | 6.42 | 0.428 |
| [AC5]-[SC6] | 6.43 | 0.377 |
| [AP6]-[SC3] | 6.43 | 0.403 |
| [SC2]-[SC12] | 6.43 | 0.409 |
| [SC2]-[SC7] | 6.43 | 0.420 |

| | | |
|---|---|---|
| [SC9]-[SC7] | 6.43 | 0.443 |
| [AC10]-[AC12] | 6.45 | 0.071 |
| [AC10]-[SC5] | 6.45 | 0.497 |
| [AP5]-[SC1] | 6.45 | 0.238 |
| [SC11]-[AP2] | 6.45 | 0.378 |
| [SC2]-[SP2] | 6.45 | 0.351 |
| [SC4]-[SC1] | 6.45 | 0.361 |
| [AC2]-[AC12] | 6.46 | 0.395 |
| [AC4]-[AC1] | 6.46 | 0.396 |
| [SC2]-[AC1] | 6.47 | 0.453 |
| [AC1]-[SC11] | 6.48 | 0.494 |
| [AC10]-[AC10] | 6.48 | 0.366 |
| [SC1]-[AC12] | 6.48 | 0.376 |
| [SC1]-[AP5] | 6.48 | 0.578 |
| [AC7]-[SC4] | 6.49 | 0.433 |
| [AC5]-[AC11] | 6.50 | 0.000 |
| [AC7]-[SC3] | 6.50 | 0.293 |
| [AP1]-[SC2] | 6.50 | 0.742 |
| [AP2]-[AC6] | 6.50 | 0.000 |
| [AP3]-[SC4] | 6.50 | 0.000 |
| [AP4]-[AC11] | 6.50 | 0.000 |
| [AP4]-[AC2] | 6.50 | 0.000 |
| [SC12]-[SC11] | 6.50 | 0.000 |
| [SC3]-[AC7] | 6.50 | 0.374 |
| [SC3]-[SP3] | 6.50 | 0.000 |
| [SC4]-[AC7] | 6.50 | 0.000 |
| [SC8]-[SC8] | 6.50 | 0.000 |
| [SC9]-[AC7] | 6.50 | 0.332 |
| [SP4]-[SC2] | 6.50 | 0.000 |
| [SC1]-[SC5] | 6.51 | 0.302 |
| [SC3]-[SC7] | 6.51 | 0.326 |
| [SC3]-[AP6] | 6.52 | 0.171 |
| [AC6]-[AC5] | 6.53 | 0.427 |
| [AC6]-[SC1] | 6.53 | 0.292 |
| [SC4]-[AC11] | 6.53 | 0.437 |

| | | |
|---|---|---|
| [SC4]-[AC9] | 6.53 | 0.387 |
| [AC4]-[AP1] | 6.54 | 0.364 |
| [AC6]-[AC2] | 6.54 | 0.320 |
| [SC11]-[SC2] | 6.54 | 0.387 |
| [AC4]-[SC2] | 6.55 | 0.303 |
| [AC4]-[SC6] | 6.55 | 0.212 |
| [AC6]-[SC2] | 6.55 | 0.303 |
| [AP1]-[SC9] | 6.55 | 0.303 |
| [AP6]-[SC7] | 6.55 | 0.212 |
| [SC3]-[AC2] | 6.55 | 0.071 |
| [SC7]-[SC11] | 6.55 | 0.592 |
| [SC7]-[SC2] | 6.55 | 0.303 |
| [SC1]-[SC12] | 6.56 | 0.321 |
| [SC12]-[AC5] | 6.56 | 0.611 |
| [SC4]-[SC10] | 6.56 | 0.416 |
| [AC1]-[AC4] | 6.57 | 0.308 |
| [AC3]-[SC12] | 6.57 | 0.306 |
| [AC5]-[AC6] | 6.57 | 0.335 |
| [SC10]-[AC6] | 6.57 | 0.339 |
| [AC5]-[AP2] | 6.58 | 0.307 |
| [SC11]-[SC1] | 6.58 | 0.264 |
| [SC9]-[AP6] | 6.58 | 0.259 |
| [SC1]-[SC3] | 6.59 | 0.376 |
| [SC4]-[SC9] | 6.59 | 0.467 |
| [AC10]-[SC9] | 6.60 | 0.274 |
| [AC11]-[AC1] | 6.60 | 0.383 |
| [AC12]-[AC5] | 6.60 | 0.000 |
| [AC5]-[SC12] | 6.60 | 0.000 |
| [AC9]-[AC9] | 6.60 | 0.100 |
| [AC9]-[SC5] | 6.60 | 0.000 |
| [SC10]-[AC12] | 6.60 | 0.000 |
| [SC10]-[AC7] | 6.60 | 0.274 |
| [SC10]-[AC8] | 6.60 | 0.383 |
| [SC2]-[AC8] | 6.60 | 0.000 |
| [SC3]-[AP3] | 6.60 | 0.283 |

| | | |
|---|---|---|
| [SC9]-[AP3] | 6.60 | 0.294 |
| [SC9]-[SP2] | 6.60 | 0.000 |
| [SP2]-[AP6] | 6.60 | 0.000 |
| [SP2]-[SC1] | 6.60 | 0.000 |
| [SP6]-[AC2] | 6.60 | 0.000 |
| [AC3]-[SC8] | 6.62 | 0.356 |
| [AC4]-[AC8] | 6.62 | 0.232 |
| [AP2]-[AC10] | 6.62 | 0.286 |
| [AC1]-[SC3] | 6.63 | 0.325 |
| [AC1]-[SP3] | 6.63 | 0.250 |
| [AC11]-[AC2] | 6.63 | 0.216 |
| [AC3]-[AC12] | 6.63 | 0.350 |
| [AP1]-[SC1] | 6.63 | 0.222 |
| [AP4]-[AC4] | 6.63 | 0.325 |
| [AC12]-[SC2] | 6.64 | 0.299 |
| [AC2]-[SC7] | 6.64 | 0.267 |
| [AP1]-[SC4] | 6.64 | 0.288 |
| [SC4]-[AP6] | 6.64 | 0.299 |
| [AC2]-[AC8] | 6.65 | 0.311 |
| [AC7]-[AC3] | 6.65 | 0.394 |
| [AC1]-[SC5] | 6.66 | 0.439 |
| [AC3]-[AP1] | 6.67 | 0.356 |
| [AC5]-[AP1] | 6.67 | 0.216 |
| [SC11]-[AP1] | 6.67 | 0.269 |
| [AC10]-[AC6] | 6.68 | 0.279 |
| [AC3]-[SC7] | 6.69 | 0.241 |
| [AC4]-[SC9] | 6.69 | 0.241 |
| [AC1]-[SP2] | 6.70 | 0.000 |
| [AC11]-[AP6] | 6.70 | 0.000 |
| [AC11]-[SC3] | 6.70 | 0.000 |
| [AC12]-[SP5] | 6.70 | 0.000 |
| [AC2]-[AC7] | 6.70 | 0.261 |
| [AC2]-[AP2] | 6.70 | 0.000 |
| [AP2]-[AC5] | 6.70 | 0.000 |
| [AP2]-[SC8] | 6.70 | 0.000 |

| | | |
|---|---|---|
| [AP4]-[SC9] | 6.70 | 0.000 |
| [AP5]-[SC3] | 6.70 | 0.294 |
| [SC1]-[AP1] | 6.70 | 0.000 |
| [SC10]-[AP6] | 6.70 | 0.141 |
| [SC2]-[AP1] | 6.70 | 0.000 |
| [SC6]-[SC1] | 6.70 | 0.000 |
| [SC9]-[AC5] | 6.70 | 0.216 |
| [AP1]-[SC10] | 6.72 | 0.277 |
| [SC1]-[SC7] | 6.72 | 0.171 |
| [SC10]-[AP2] | 6.73 | 0.216 |
| [SC7]-[SC9] | 6.73 | 0.379 |
| [SC5]-[AP6] | 6.74 | 0.207 |
| [AC1]-[AC11] | 6.75 | 0.354 |
| [AC10]-[AP1] | 6.75 | 0.187 |
| [AC10]-[AP2] | 6.75 | 0.071 |
| [AC2]-[AP1] | 6.75 | 0.071 |
| [AC3]-[AP2] | 6.75 | 0.129 |
| [AC4]-[AP6] | 6.75 | 0.071 |
| [AC4]-[SC8] | 6.75 | 0.071 |
| [AC9]-[AC10] | 6.75 | 0.187 |
| [SC10]-[AP1] | 6.75 | 0.187 |
| [SC10]-[SC3] | 6.75 | 0.212 |
| [SC3]-[AC3] | 6.75 | 0.187 |
| [AC8]-[AC10] | 6.76 | 0.230 |
| [SC1]-[AC7] | 6.77 | 0.153 |
| [SC6]-[AC2] | 6.77 | 0.321 |
| [AC3]-[AC7] | 6.78 | 0.192 |
| [AC10]-[AC8] | 6.80 | 0.000 |
| [AC11]-[AP1] | 6.80 | 0.216 |
| [AC11]-[AP2] | 6.80 | 0.141 |
| [AC2]-[SC6] | 6.80 | 0.000 |
| [AC3]-[AC8] | 6.80 | 0.183 |
| [AC3]-[AP6] | 6.80 | 0.000 |
| [AC4]-[SP6] | 6.80 | 0.000 |
| [AC8]-[SC8] | 6.80 | 0.000 |

| | | |
|---|---|---|
| [AP1]-[AC3] | 6.80 | 0.141 |
| [AP3]-[AC7] | 6.80 | 0.283 |
| [AP4]-[AC10] | 6.80 | 0.000 |
| [SC10]-[AC2] | 6.80 | 0.216 |
| [SC3]-[AP1] | 6.80 | 0.158 |
| [SC9]-[AP1] | 6.80 | 0.100 |
| [SC9]-[SC5] | 6.80 | 0.000 |
| [SP1]-[SC9] | 6.80 | 0.000 |
| [AC3]-[AC6] | 6.82 | 0.171 |
| [SC1]-[AC9] | 6.82 | 0.171 |
| [AC1]-[AC2] | 6.83 | 0.153 |
| [SC12]-[SC1] | 6.83 | 0.153 |
| [AC1]-[AC10] | 6.85 | 0.071 |
| [AC10]-[AC2] | 6.85 | 0.212 |
| [AP1]-[SC5] | 6.85 | 0.129 |
| [AP3]-[SC7] | 6.85 | 0.129 |
| [SC1]-[SP2] | 6.85 | 0.071 |
| [AC7]-[SC2] | 6.87 | 0.153 |
| [AC1]-[SC12] | 6.90 | 0.100 |
| [AC12]-[SC1] | 6.90 | 0.000 |
| [AC12]-[SC6] | 6.90 | 0.000 |
| [AP1]-[SC3] | 6.90 | 0.000 |
| [AP2]-[AP3] | 6.90 | 0.000 |
| [AP2]-[SC5] | 6.90 | 0.000 |
| [AP5]-[AC1] | 6.90 | 0.000 |
| [SC10]-[AC1] | 6.90 | 0.000 |
| [SC11]-[AP6] | 6.90 | 0.000 |
| [SC7]-[SC1] | 6.90 | 0.000 |
| [SC7]-[SP5] | 6.90 | 0.141 |
| [AC5]-[SP5] | 6.95 | 0.071 |
| [AP4]-[AP3] | 6.95 | 0.071 |
| [AC12]-[AC1] | 7.00 | 0.000 |
| [SC1]-[AC5] | 7.00 | 0.000 |
| [SC10]-[SC2] | 7.00 | 0.000 |
| [SC6]-[AC5] | 7.00 | 0.000 |

| | | |
|---|---|---|
| [SC6]-[SP6] | 7.00 | 0.000 |
| [SC8]-[SC1] | 7.00 | 0.000 |
| [SC9]-[SP3] | 7.00 | 0.000 |
| [SP3]-[AC4] | 7.00 | 0.000 |
| [SP5]-[AC1] | 7.00 | 0.000 |
| [SP5]-[SP4] | 7.00 | 0.000 |

**Table 5. 9 The mean and standard deviation of the distances between the Cα atom of the i and i+3 residues for each turn type.**

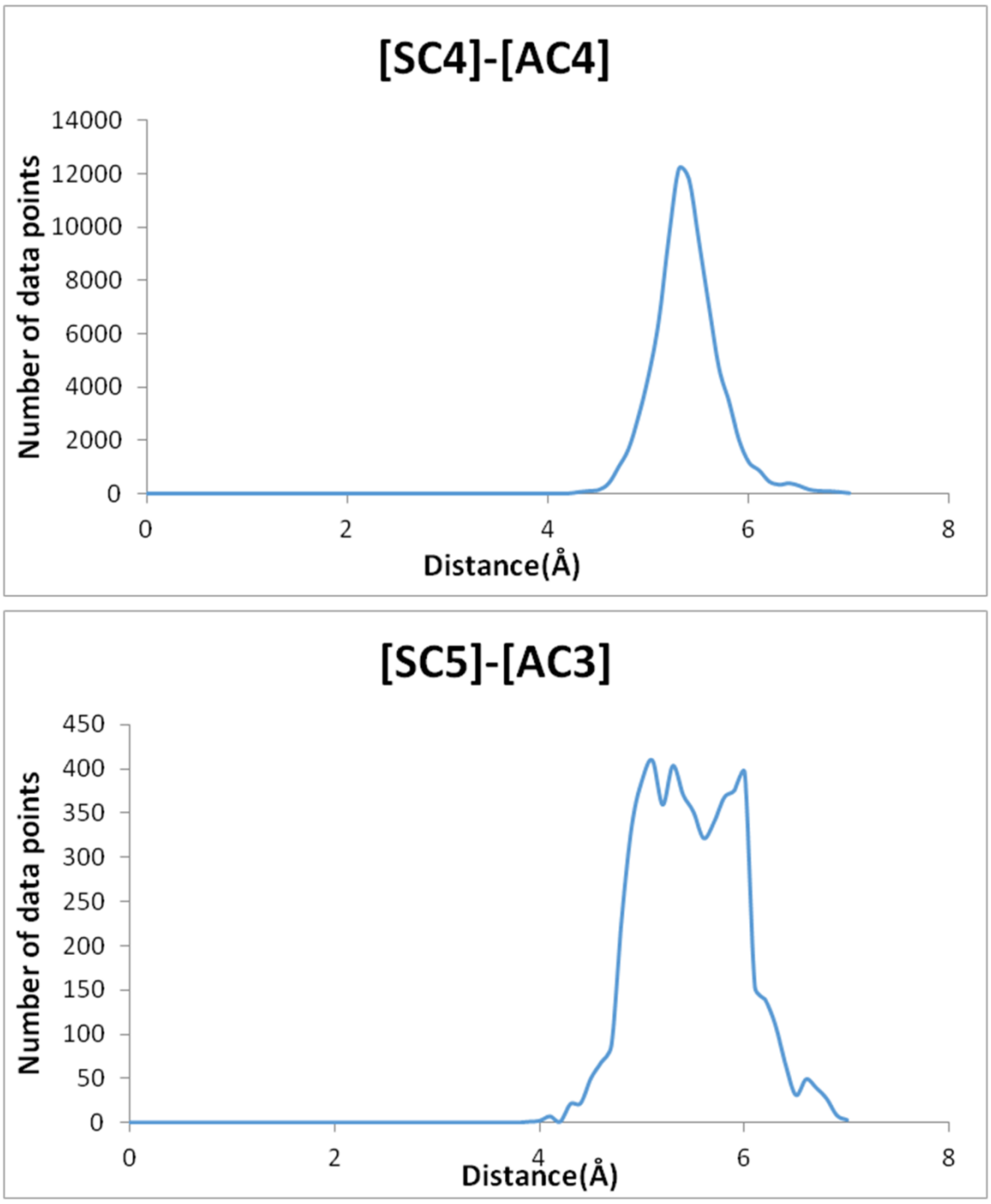
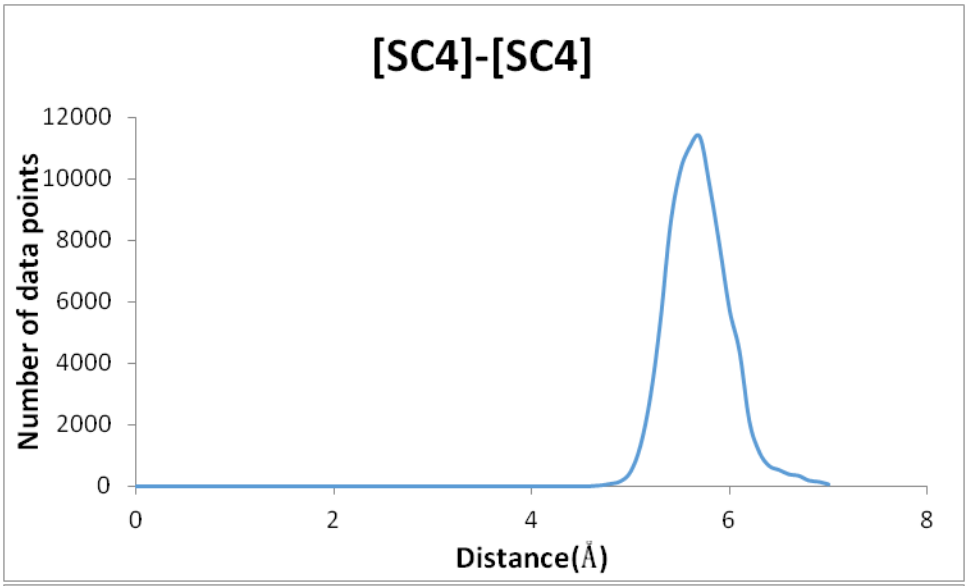
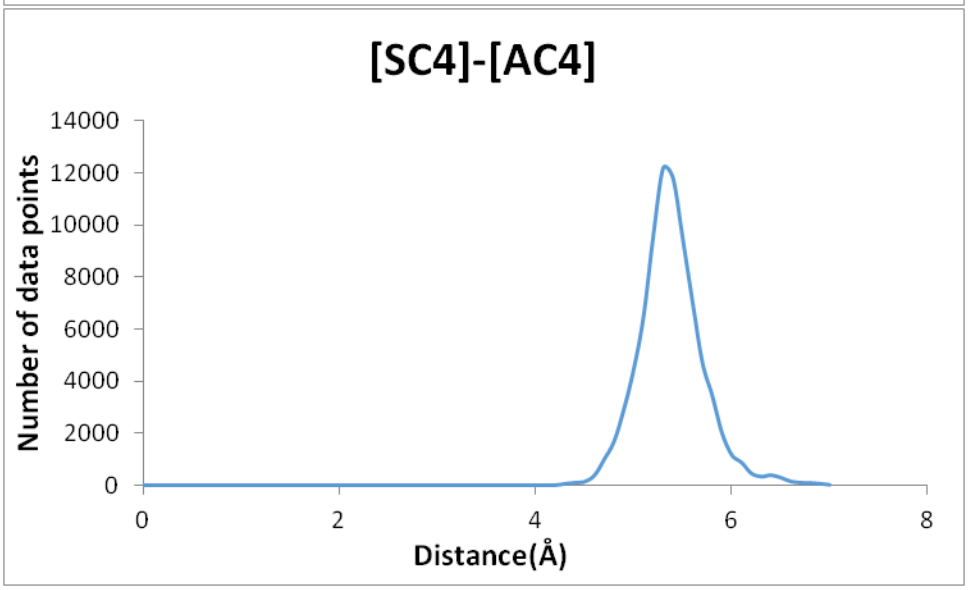
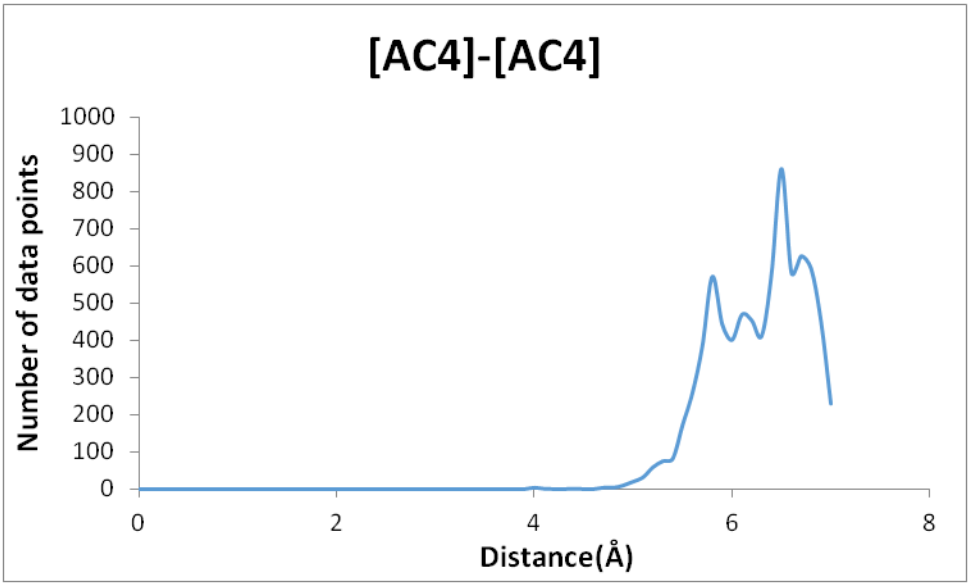
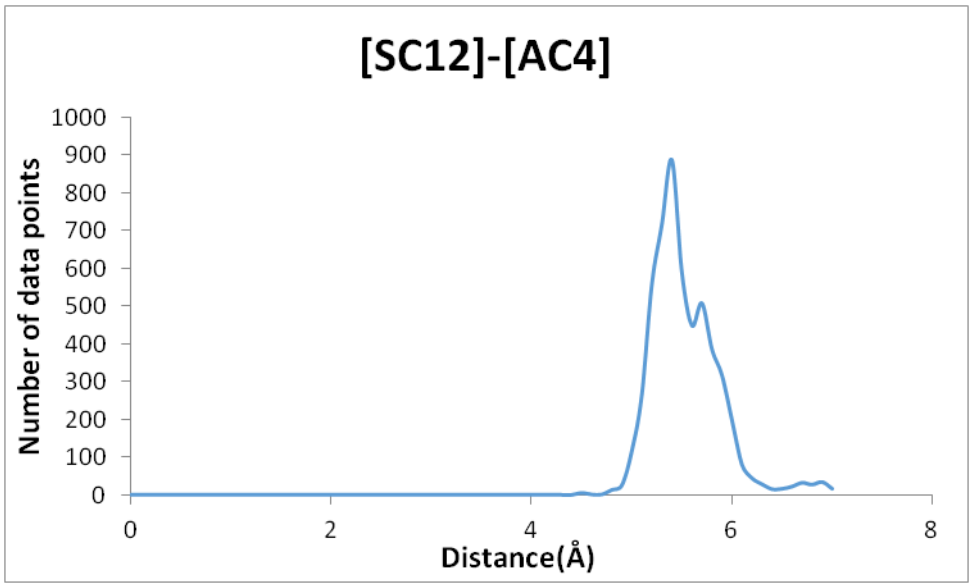


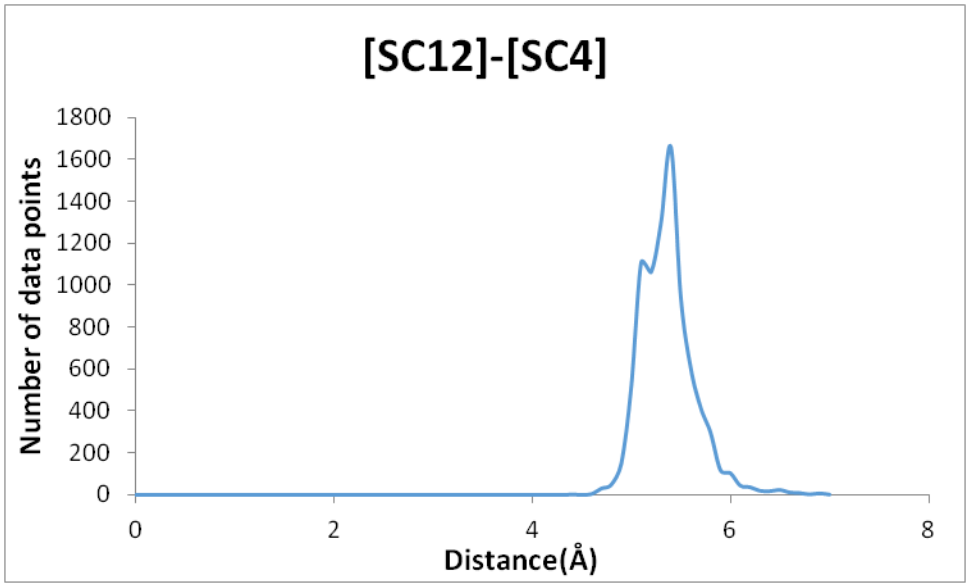

**Figure 5. 25 The Ca distribution for each turn type.** The figure above is the representative graph for the distribution with only one peak. The figure below is the representative graph for the distribution with wide range and no specific peak is found in this category.

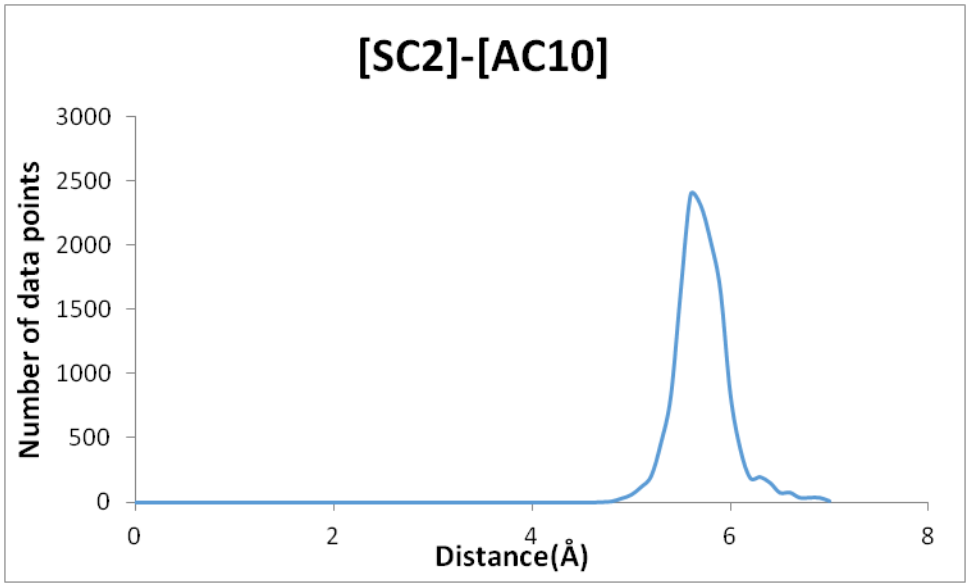

[SC4]-[SC4]
Number of data points
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
Distance(Å)

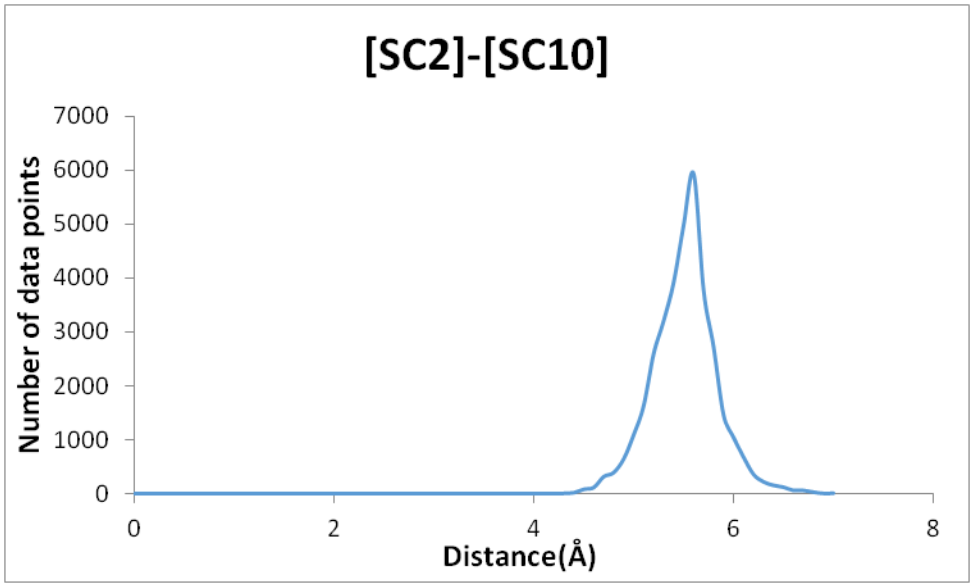

[SC4]-[AC4]
Number of data points
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
Distance(Å)

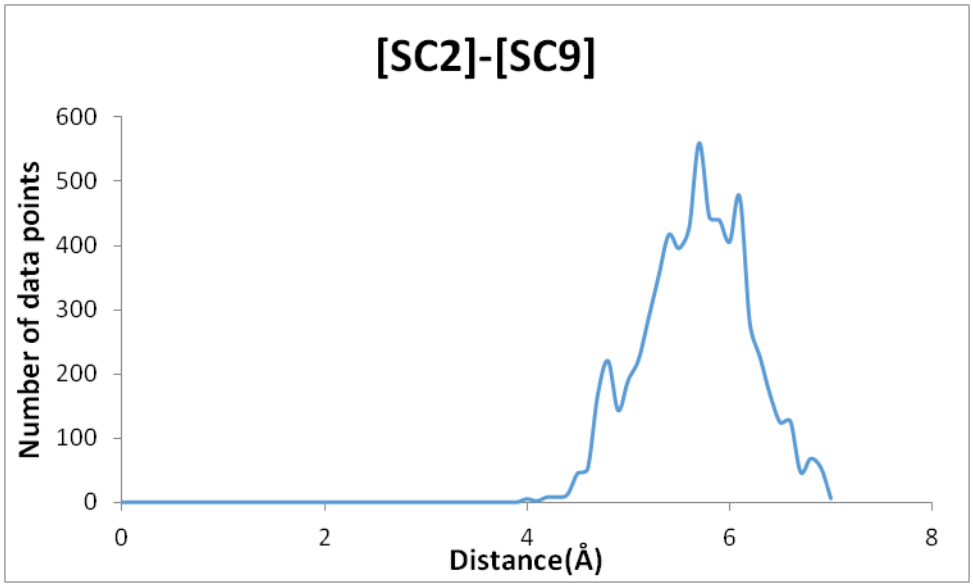

[AC4]-[AC4]
Number of data points
1000
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)

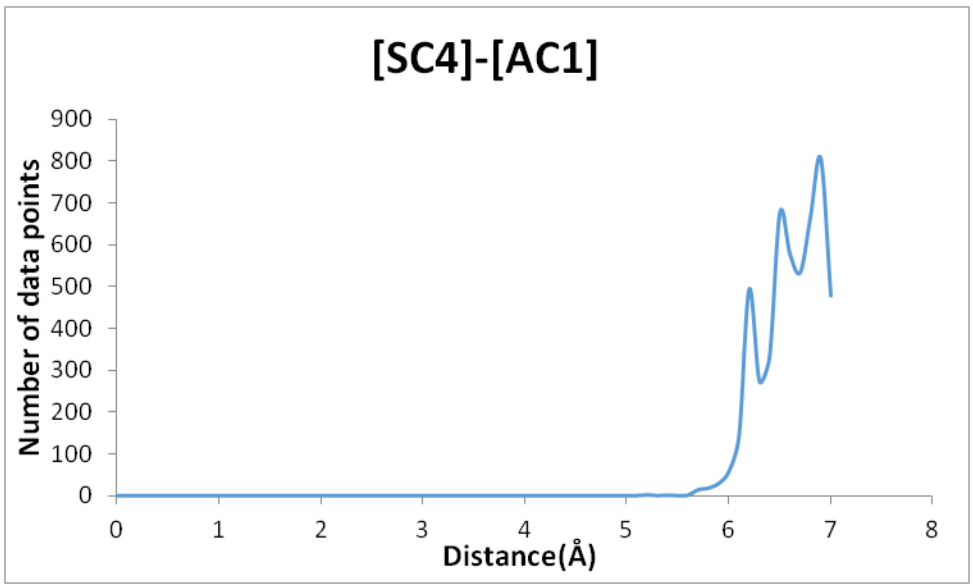

[SC12]-[AC4]
Number of data points
1000
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)

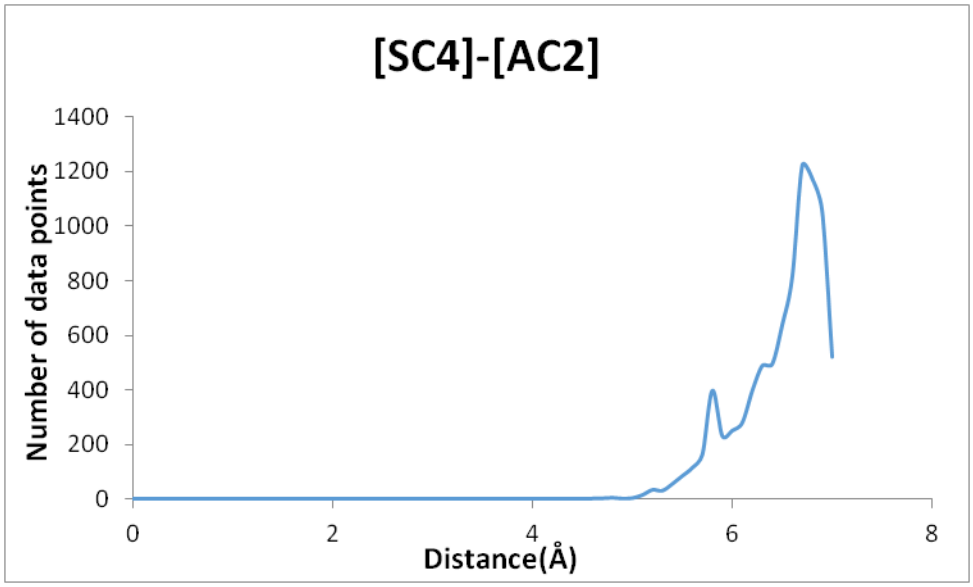

[SC12]-[SC4]
Number of data points
1800
1600
1400
1200
1000
800
600
400
200
0
0
2
4
6
8
Distance(Å)

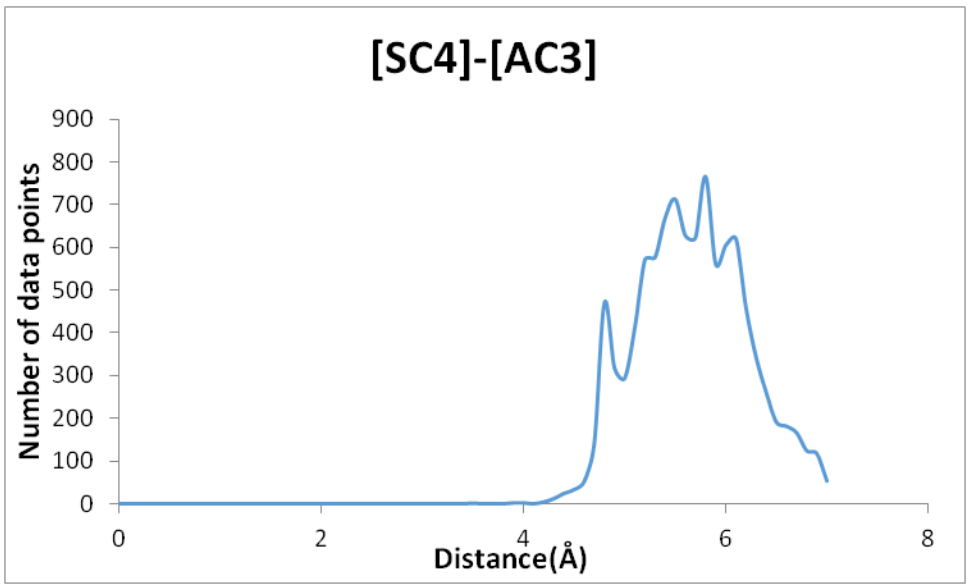

[SC2]-[AC10]
Number of data points
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
Distance(Å)

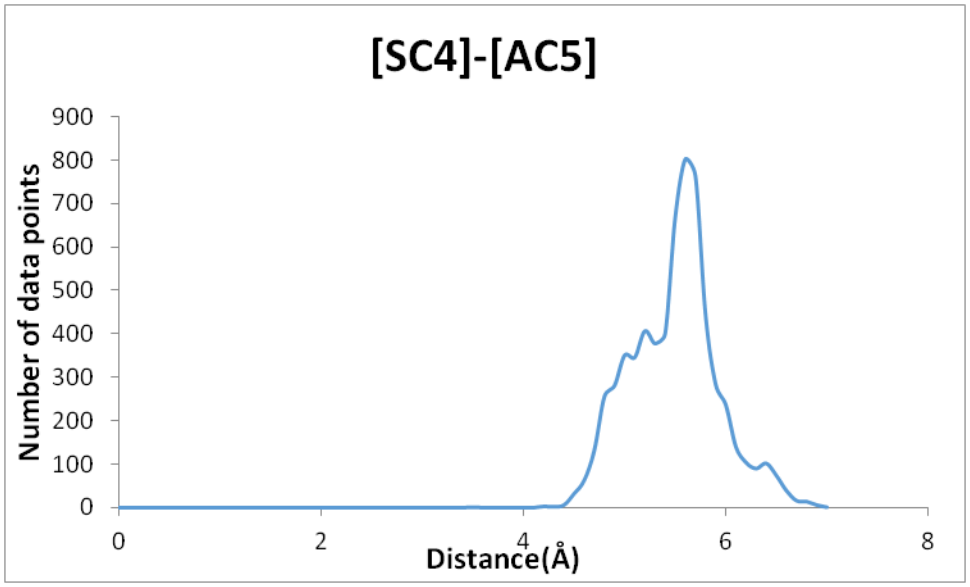

[SC2]-[SC10]
Number of data points
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
Distance(Å)

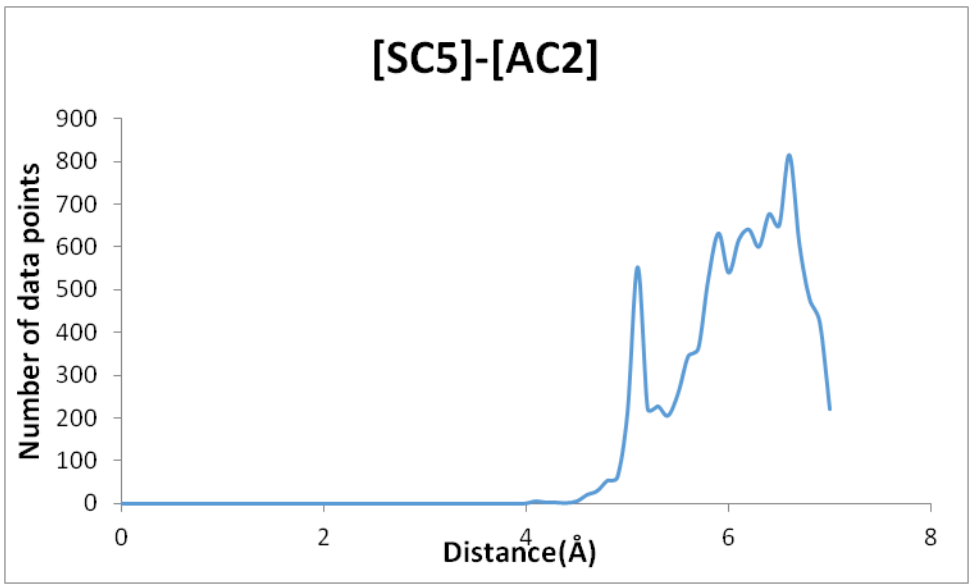

[SC2]-[SC9]
Number of data points
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)

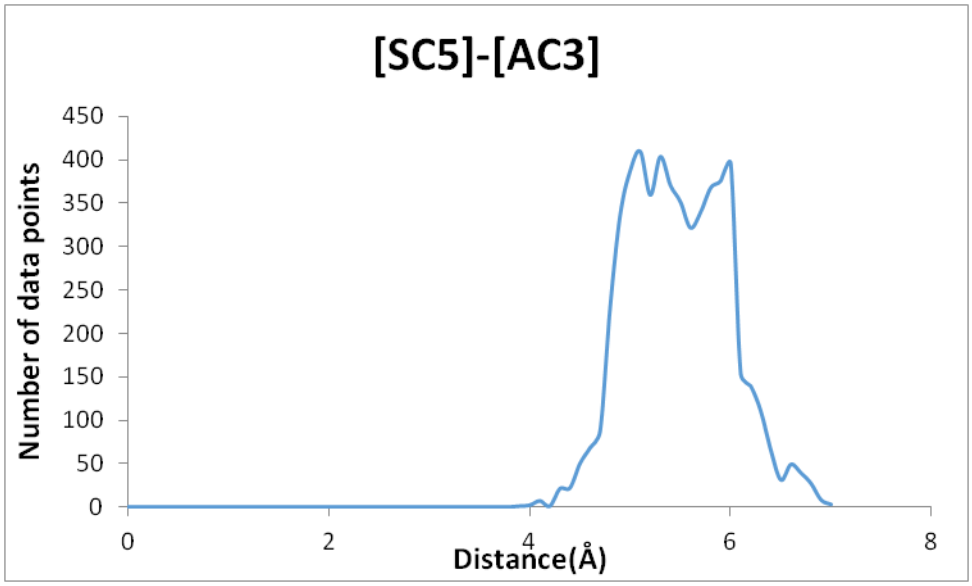

[SC4]-[AC1]
Number of data points
900
800
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
Distance(Å)

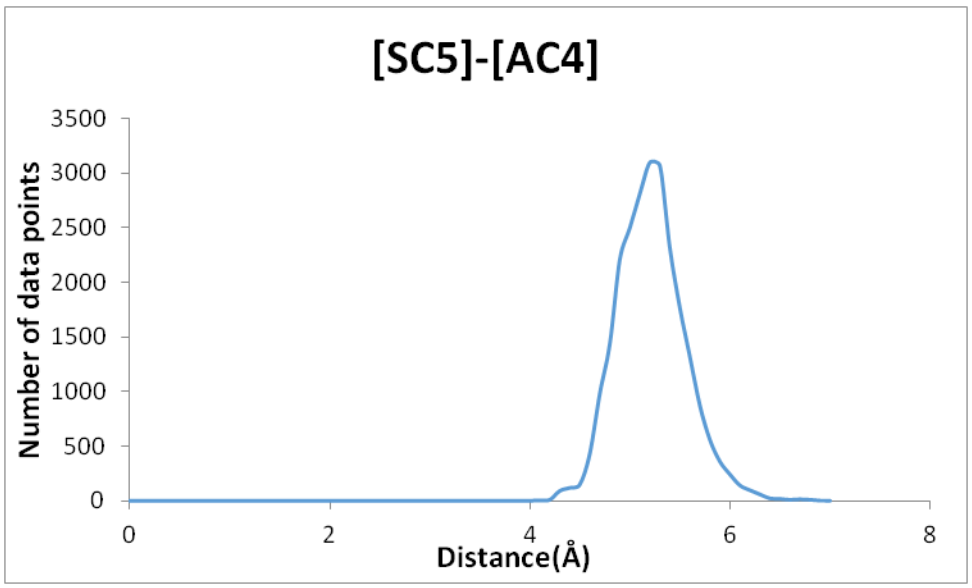

[SC4]-[AC2]
Number of data points
1400
1200
1000
800
600
400
200
0
0
2
4
6
8
Distance(Å)

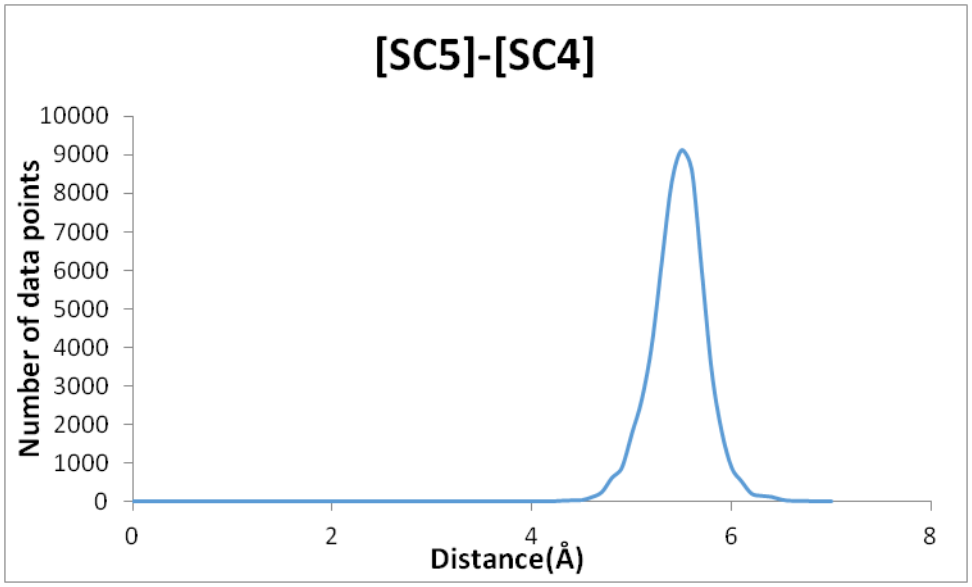

[SC4]-[AC3]
Number of data points
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)

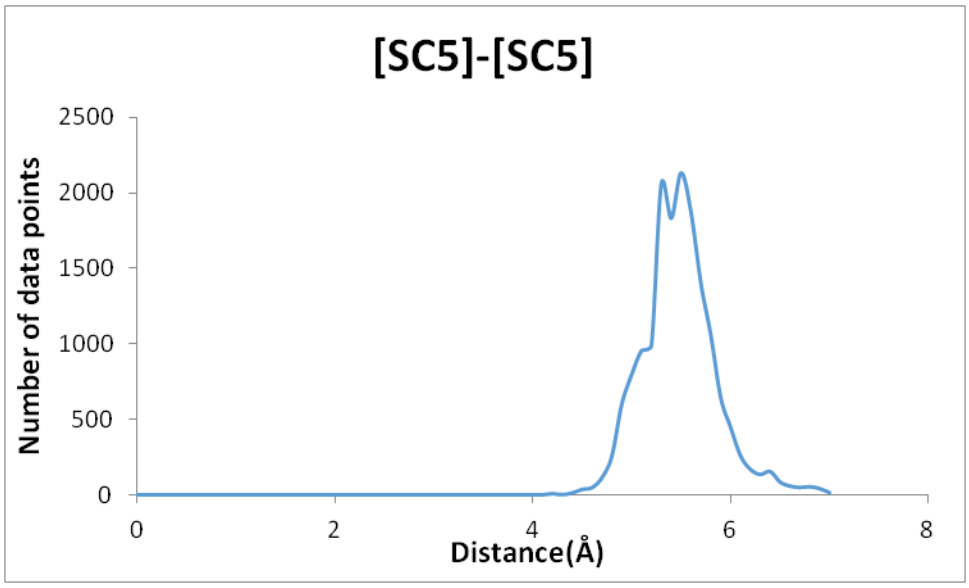

[SC4]-[AC5]
Number of data points
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)

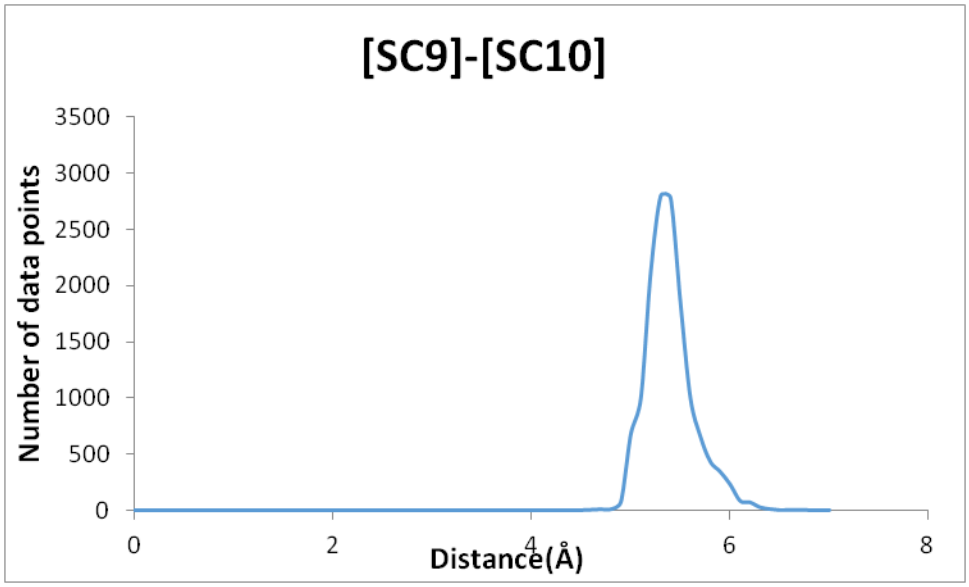

[SC5]-[AC2]
Number of data points
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
Distance(Å)


[SC5]-[AC3]
Number of data points
450
400
350
300
250
200
150
100
50
0
0
2
4
6
8
Distance(Å)


[SC5]-[AC4]
Number of data points
3500
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
Distance(Å)


[SC5]-[SC4]
Number of data points
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
Distance(Å)


[SC5]-[SC5]
Number of data points
2500
2000
1500
1000
500
0
0
2
4
6
8
Distance(Å)


[SC9]-[SC10]
Number of data points
3500
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
Distance(Å)

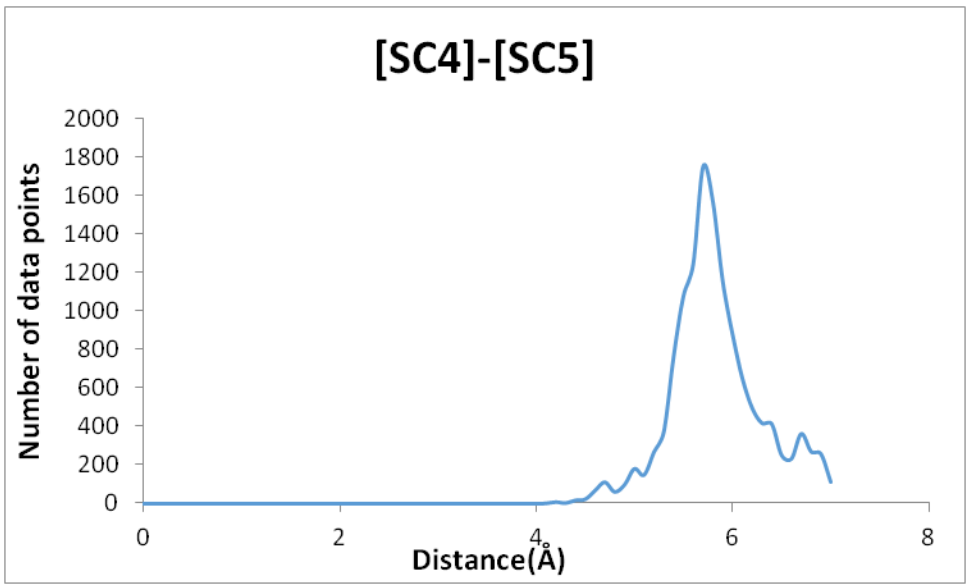



**Figure 5. 26 The Ca distribution for top 19 turn types**

### 5.4.5 Amino acid preferences at distinct residue positions in β turns

To determine if an amino acid preference existed for each residue position in β turns, the amino acid distribution at each position in the entire β turn database was evaluated. From **Figure 5.27**, it can be seen that the top five most frequent amino acids in the i position in decreasing order were D/G/P/S/A, for the i+1 position P/A/G/S/E, for the i+2 position G/D/N/S/E and for the i+3 residue G/A/L/S/T. Although most of the remaining amino acids were found in β turns, their occurrence was relatively low compared to the other amino acids already mentioned. The analysis indicated that the β turns most frequently involved glycine, aspartate, serine, alanine and prolines residues whereas cysteine, methionine and tryptophan were least commonly observed. These conclusions were further confirmed when the fractional occurrences at each residue position in the β turns were plotted for each amino acid type (**Figure 5.28**).

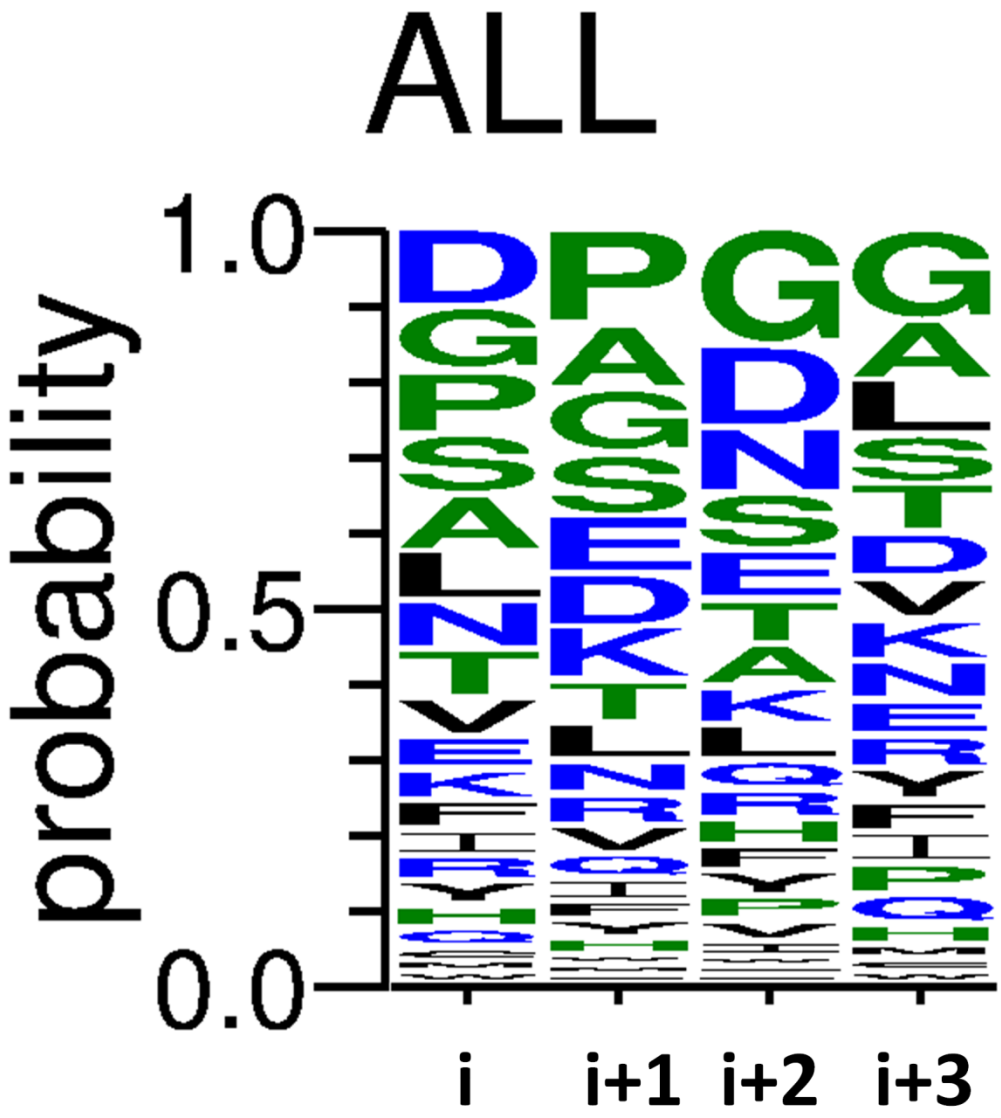



**Figure 5. 27 The summary of amino acid distribution based on all database in each residue.** From the top to bottom is from the most to the least. The size of each letter represents the percentage of each amino acid.

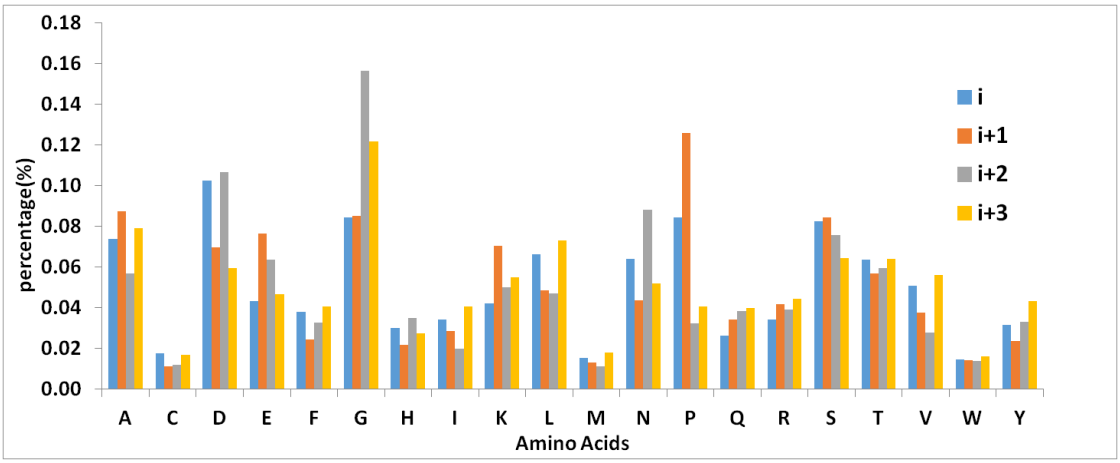
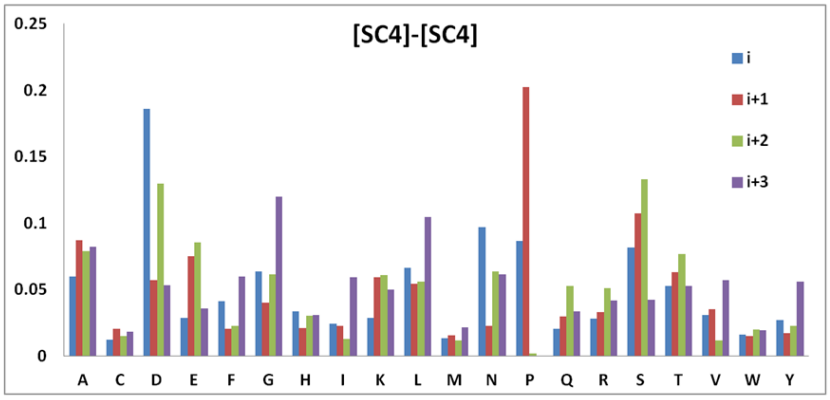


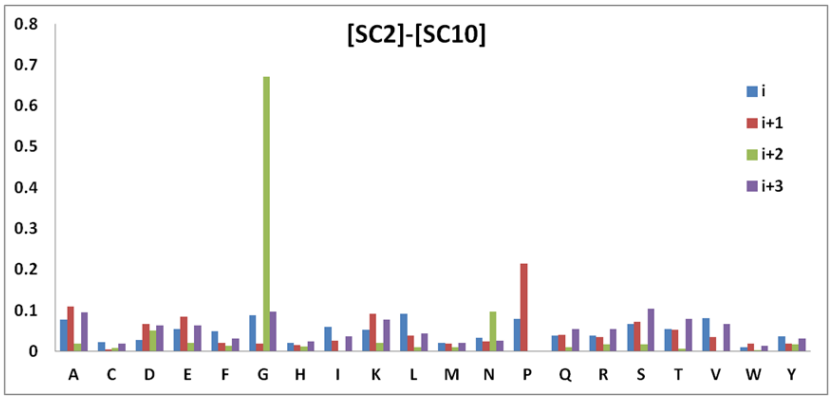

**Figure 5. 28 The amino acid distribution based on all database in each residue.** The bar graph shows the distribution of amino acids in all four residues.

To determine if specific β turn types defined in the new classification scheme exhibited amino acid preferences with regard to the residue position within the different β turn types, we analyzed the amino acid occurrences as a function of the residue position within the β turn. Inspection of the results indicated that some turn types, like the most common SC4-SC4 turn, which overlaps with the type I and III turn types in the conventional scheme, exhibited no strong preference for any specific amino acid type at any of the residue positions with no amino acid preference reaching greater than 20% occurrence at any residue position, although glutamate most commonly occurred at the i position ( ~18% of turns) and proline was most commonly observed in the i+1 position (~20% of turns) (**Figure 5.29A**). In contrast, other turn types showed very strong amino acid preferences at certain positions, e.g. in the SC2-SC10 turn, which with the type II turn region in the conventional scheme, glycine occurred in ~68% of the turns at the i+2 position (**Figure 5.29B**), in the SC12-AC4 turn, which overlaps with the type IV turn region in the conventional scheme, glycine occurred at nearly 70% of the turns in the i+1 position (**Figure 5.29C**), and glycine occurred in nearly 80% of the turns in the i+1 position in the SC12-SC4 turns, which overlapped with the type IV turn region in the conventional scheme (**Figure 5.29D**). The residue amino acid preferences for the top 19 β turn types defined in the new scheme are tabulated in **Table 5.10** to allow for further inspection and analysis.

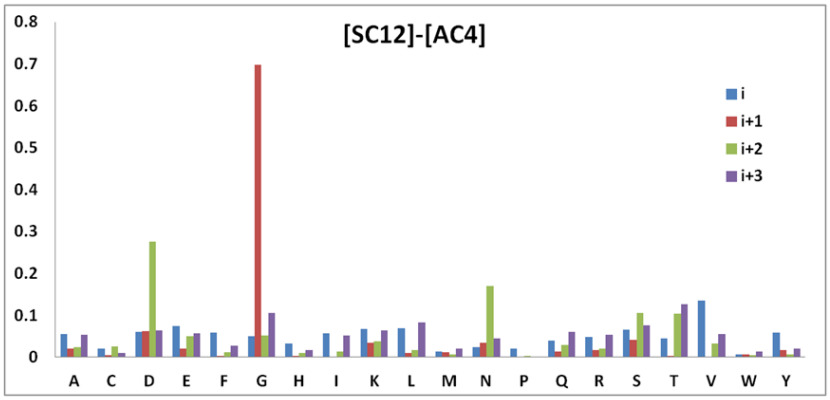
A


[SC4]-[SC4]
0.25
0.2
0.15
0.1
0.05
0
A C D E F G H I K L M N P Q R S T V W Y
i
i+1
i+2
i+3

B


[SC2]-[SC10]
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A C D E F G H I K L M N P Q R S T V W Y
i
i+1
i+2
i+3

C


[SC12]-[AC4]
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A C D E F G H I K L M N P Q R S T V W Y
i
i+1
i+2
i+3

D

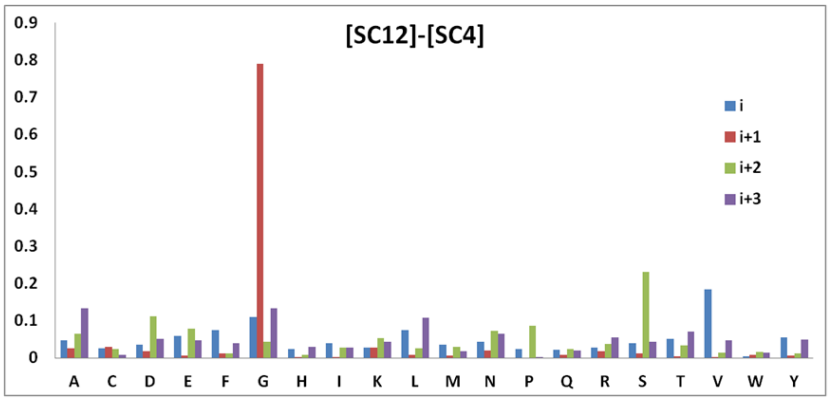

[SC12]-[SC4]
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A C D E F G H I K L M N P Q R S T V W Y
i
i+1
i+2
i+3

**Figure 5. 29 The amino acid distribution represented by different new turn type in each residue.** A) The bar graph shows the distribution of amino acids of four residues for turn type [SC4]-[SC4]. B) The bar graph shows the distribution of amino acids of four residues for turn type [SC2]-[SC10]. C) The bar graph shows the distribution of amino acids of four residues for turn type [SC12]-[AC4]. D) The bar graph shows the distribution of amino acids of four residues for turn type [SC12]-[SC4].

| Aminoaicd | Number of turns | Percentage | Number of turns | Percentage | Number of turns | Percentage | Number of turns | Percentage |
|---|---|---|---|---|---|---|---|---|
| Position | 1 | | 2 | | 3 | | 4 | |
| [SC4]-[SC4] | | | | | | | | |
| A | 5077 | 0.059911 | 7392 | 0.087229 | 6674 | 0.078757 | 6968 | 0.082226 |
| C | 1065 | 0.012568 | 1720 | 0.020297 | 1301 | 0.015352 | 1540 | 0.018173 |
| D | 15774 | 0.186141 | 4824 | 0.056926 | 11002 | 0.129829 | 4510 | 0.05322 |
| E | 2450 | 0.028911 | 6387 | 0.07537 | 7251 | 0.085566 | 3037 | 0.035838 |
| F | 3494 | 0.041231 | 1759 | 0.020757 | 1932 | 0.022799 | 5078 | 0.059923 |
| G | 5386 | 0.063558 | 3396 | 0.040075 | 5219 | 0.061587 | 10138 | 0.119634 |
| H | 2851 | 0.033643 | 1785 | 0.021064 | 2571 | 0.030339 | 2610 | 0.030799 |
| I | 2066 | 0.02438 | 1921 | 0.022669 | 1108 | 0.013075 | 5010 | 0.059121 |
| K | 2449 | 0.028899 | 5048 | 0.059569 | 5185 | 0.061186 | 4261 | 0.050282 |
| L | 5610 | 0.066201 | 4622 | 0.054542 | 4763 | 0.056206 | 8859 | 0.104541 |
| M | 1152 | 0.013594 | 1307 | 0.015423 | 993 | 0.011718 | 1813 | 0.021394 |
| N | 8229 | 0.097107 | 1934 | 0.022822 | 5387 | 0.063569 | 5193 | 0.06128 |
| P | 7339 | 0.086604 | 17165 | 0.202556 | 176 | 0.002077 | 18 | 0.000212 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 1724 | 0.020344 | 2516 | 0.02969 | 4458 | 0.052607 | 2863 | 0.033785 |
| R | 2378 | 0.028062 | 2801 | 0.033053 | 4315 | 0.050919 | 3527 | 0.04162 |
| S | 6942 | 0.081919 | 9105 | 0.107444 | 11258 | 0.13285 | 3611 | 0.042612 |
| T | 4476 | 0.052819 | 5330 | 0.062897 | 6505 | 0.076762 | 4462 | 0.052654 |
| V | 2609 | 0.030788 | 2983 | 0.035201 | 1015 | 0.011978 | 4838 | 0.057091 |
| W | 1369 | 0.016155 | 1279 | 0.015093 | 1714 | 0.020226 | 1664 | 0.019636 |
| Y | 2302 | 0.027165 | 1468 | 0.017323 | 1915 | 0.022598 | 4742 | 0.055958 |
| Total | 84742 | 1 | 84742 | 1 | 84742 | 1 | 84742 | 1 |
| [SC4]-[AC4] | | | | | | | | |
| A | 6588 | 0.080858 | 7106 | 0.087216 | 3196 | 0.039226 | 7927 | 0.097292 |
| C | 1275 | 0.015649 | 848 | 0.010408 | 866 | 0.010629 | 1406 | 0.017257 |
| D | 11250 | 0.138077 | 6672 | 0.081889 | 9737 | 0.119508 | 4152 | 0.05096 |
| E | 3084 | 0.037852 | 7159 | 0.087866 | 5598 | 0.068707 | 2880 | 0.035348 |
| F | 2663 | 0.032684 | 1080 | 0.013255 | 4303 | 0.052813 | 2747 | 0.033715 |
| G | 5732 | 0.070352 | 3775 | 0.046333 | 2863 | 0.035139 | 13387 | 0.164306 |
| H | 2555 | 0.031359 | 1817 | 0.022301 | 3604 | 0.044234 | 2054 | 0.02521 |
| I | 1882 | 0.023099 | 2114 | 0.025946 | 1254 | 0.015391 | 4377 | 0.053721 |
| K | 2605 | 0.031973 | 5654 | 0.069395 | 5994 | 0.073568 | 2918 | 0.035814 |
| L | 4780 | 0.058668 | 3055 | 0.037496 | 5661 | 0.069481 | 6111 | 0.075004 |
| M | 1017 | 0.012482 | 818 | 0.01004 | 1131 | 0.013881 | 1874 | 0.023001 |
| N | 6375 | 0.078244 | 3157 | 0.038748 | 9562 | 0.11736 | 3972 | 0.048751 |
| P | 7319 | 0.08983 | 11063 | 0.135782 | 5 | 6.14E-05 | 23 | 0.000282 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 1865 | 0.02289 | 3449 | 0.042331 | 4672 | 0.057342 | 3164 | 0.038834 |
| R | 2676 | 0.032844 | 3365 | 0.041301 | 3845 | 0.047192 | 3682 | 0.045191 |
| S | 8349 | 0.102472 | 11142 | 0.136752 | 4667 | 0.057281 | 6015 | 0.073825 |
| T | 5666 | 0.069542 | 5364 | 0.065835 | 6535 | 0.080208 | 4323 | 0.053059 |
| V | 2139 | 0.026253 | 1533 | 0.018815 | 1852 | 0.022731 | 5945 | 0.072966 |
| W | 1403 | 0.01722 | 943 | 0.011574 | 1317 | 0.016164 | 1593 | 0.019552 |
| Y | 2253 | 0.027652 | 1362 | 0.016717 | 4814 | 0.059085 | 2926 | 0.035912 |
| Total | 81476 | 1 | 81476 | 1 | 81476 | 1 | 81476 | 1 |
| [SC5]-[SC4] | | | | | | | | |
| A | 4567 | 0.081884 | 7679 | 0.137681 | 5543 | 0.099383 | 3562 | 0.063865 |
| C | 686 | 0.0123 | 485 | 0.008696 | 608 | 0.010901 | 994 | 0.017822 |
| D | 3804 | 0.068204 | 2765 | 0.049575 | 7007 | 0.125632 | 5041 | 0.090383 |
| E | 2675 | 0.047961 | 4966 | 0.089038 | 6929 | 0.124234 | 3833 | 0.068724 |
| F | 1610 | 0.028866 | 1875 | 0.033618 | 834 | 0.014953 | 2640 | 0.047334 |
| G | 4159 | 0.074569 | 2212 | 0.03966 | 3540 | 0.06347 | 6399 | 0.114731 |
| H | 2281 | 0.040897 | 971 | 0.01741 | 2205 | 0.039535 | 2132 | 0.038226 |
| I | 1882 | 0.033743 | 2038 | 0.03654 | 756 | 0.013555 | 1592 | 0.028544 |
| K | 3278 | 0.058773 | 4380 | 0.078531 | 2373 | 0.042547 | 3676 | 0.065909 |
| L | 4078 | 0.073117 | 4008 | 0.071861 | 2133 | 0.038244 | 4559 | 0.081741 |
| M | 807 | 0.014469 | 859 | 0.015401 | 757 | 0.013573 | 1183 | 0.021211 |
| N | 3095 | 0.055492 | 933 | 0.016728 | 2770 | 0.049665 | 3443 | 0.061731 |
| P | 7165 | 0.128465 | 8424 | 0.151038 | 4079 | 0.073134 | 15 | 0.000269 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 1718 | 0.030803 | 1626 | 0.029153 | 2672 | 0.047908 | 2486 | 0.044573 |
| R | 1984 | 0.035572 | 2501 | 0.044842 | 2066 | 0.037042 | 2558 | 0.045864 |
| S | 3494 | 0.062646 | 2700 | 0.04841 | 6703 | 0.120181 | 3780 | 0.067774 |
| T | 4042 | 0.072471 | 2091 | 0.037491 | 2083 | 0.037347 | 2403 | 0.043085 |
| V | 2312 | 0.041453 | 2692 | 0.048266 | 801 | 0.014362 | 1494 | 0.026787 |
| W | 610 | 0.010937 | 1271 | 0.022788 | 825 | 0.014792 | 924 | 0.016567 |
| Y | 1527 | 0.027378 | 1298 | 0.023272 | 1090 | 0.019543 | 3060 | 0.054864 |
| Total | 55774 | 1 | 55774 | 1 | 55774 | 1 | 55774 | 1 |
| [SC2]-[SC10] | | | | | | | | |
| A | 2757 | 0.077383 | 3887 | 0.1091 | 651 | 0.018272 | 3360 | 0.094308 |
| C | 784 | 0.022005 | 166 | 0.004659 | 257 | 0.007213 | 643 | 0.018048 |
| D | 1010 | 0.028348 | 2370 | 0.066521 | 1777 | 0.049877 | 2270 | 0.063714 |
| E | 1941 | 0.05448 | 3029 | 0.085017 | 708 | 0.019872 | 2270 | 0.063714 |
| F | 1758 | 0.049343 | 742 | 0.020826 | 465 | 0.013052 | 1133 | 0.031801 |
| G | 3154 | 0.088526 | 663 | 0.018609 | 23913 | 0.671186 | 3457 | 0.09703 |
| H | 718 | 0.020153 | 515 | 0.014455 | 431 | 0.012097 | 878 | 0.024644 |
| I | 2104 | 0.059055 | 893 | 0.025065 | 8 | 0.000225 | 1288 | 0.036151 |
| K | 1873 | 0.052571 | 3231 | 0.090687 | 755 | 0.021191 | 2732 | 0.076681 |
| L | 3285 | 0.092203 | 1342 | 0.037667 | 379 | 0.010638 | 1570 | 0.044066 |
| M | 714 | 0.02004 | 667 | 0.018721 | 350 | 0.009824 | 750 | 0.021051 |
| N | 1157 | 0.032474 | 863 | 0.024223 | 3447 | 0.09675 | 891 | 0.025008 |
| P | 2846 | 0.079881 | 7655 | 0.214859 | 0 | 0 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 1384 | 0.038846 | 1408 | 0.039519 | 331 | 0.00929 | 1947 | 0.054648 |
| R | 1343 | 0.037695 | 1212 | 0.034018 | 629 | 0.017655 | 1925 | 0.054031 |
| S | 2396 | 0.06725 | 2539 | 0.071264 | 607 | 0.017037 | 3708 | 0.104075 |
| T | 1955 | 0.054873 | 1874 | 0.052599 | 192 | 0.005389 | 2848 | 0.079937 |
| V | 2859 | 0.080246 | 1243 | 0.034888 | 17 | 0.000477 | 2344 | 0.065791 |
| W | 317 | 0.008897 | 635 | 0.017823 | 102 | 0.002863 | 498 | 0.013978 |
| Y | 1273 | 0.03573 | 694 | 0.019479 | 609 | 0.017093 | 1116 | 0.031324 |
| Total | 35628 | 1 | 35628 | 1 | 35628 | 1 | 35628 | 1 |
| [SC5]-[AC4] | | | | | | | | |
| A | 2124 | 0.08586 | 3003 | 0.121392 | 639 | 0.025831 | 2399 | 0.096976 |
| C | 339 | 0.013704 | 93 | 0.003759 | 389 | 0.015725 | 315 | 0.012733 |
| D | 1628 | 0.06581 | 1761 | 0.071186 | 4400 | 0.177864 | 1348 | 0.054491 |
| E | 1113 | 0.044992 | 2605 | 0.105304 | 993 | 0.040141 | 860 | 0.034764 |
| F | 878 | 0.035492 | 490 | 0.019808 | 1236 | 0.049964 | 837 | 0.033835 |
| G | 2693 | 0.108861 | 929 | 0.037554 | 973 | 0.039332 | 6227 | 0.251718 |
| H | 877 | 0.035452 | 576 | 0.023284 | 1407 | 0.056876 | 712 | 0.028782 |
| I | 639 | 0.025831 | 754 | 0.030479 | 386 | 0.015604 | 510 | 0.020616 |
| K | 1258 | 0.050853 | 3260 | 0.131781 | 940 | 0.037998 | 1191 | 0.048145 |
| L | 1461 | 0.059059 | 1114 | 0.045032 | 1507 | 0.060918 | 1159 | 0.046851 |
| M | 430 | 0.017382 | 336 | 0.013582 | 183 | 0.007398 | 380 | 0.015361 |
| N | 987 | 0.039898 | 637 | 0.02575 | 3080 | 0.124505 | 2020 | 0.081656 |
| P | 1989 | 0.080403 | 2081 | 0.084122 | 54 | 0.002183 | 8 | 0.000323 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 502 | 0.020293 | 1198 | 0.048428 | 1134 | 0.04584 | 815 | 0.032945 |
| R | 935 | 0.037796 | 1316 | 0.053198 | 875 | 0.035371 | 1108 | 0.044789 |
| S | 2251 | 0.090994 | 1477 | 0.059706 | 1595 | 0.064476 | 1780 | 0.071954 |
| T | 2362 | 0.095481 | 1181 | 0.04774 | 2936 | 0.118684 | 903 | 0.036503 |
| V | 928 | 0.037513 | 1284 | 0.051904 | 269 | 0.010874 | 930 | 0.037594 |
| W | 369 | 0.014916 | 199 | 0.008044 | 333 | 0.013461 | 369 | 0.014916 |
| Y | 975 | 0.039413 | 444 | 0.017948 | 1409 | 0.056957 | 867 | 0.035047 |
| Total | 24738 | 1 | 24738 | 1 | 24738 | 1 | 24738 | 1 |
| [AC4]-[AC4] | | | | | | | | |
| A | 684 | 0.088065 | 293 | 0.037724 | 578 | 0.074417 | 377 | 0.048539 |
| C | 130 | 0.016737 | 146 | 0.018797 | 189 | 0.024334 | 141 | 0.018154 |
| D | 553 | 0.071199 | 637 | 0.082014 | 992 | 0.12772 | 242 | 0.031157 |
| E | 259 | 0.033346 | 322 | 0.041457 | 316 | 0.040685 | 185 | 0.023819 |
| F | 267 | 0.034376 | 417 | 0.053689 | 263 | 0.033861 | 312 | 0.04017 |
| G | 577 | 0.074289 | 380 | 0.048925 | 180 | 0.023175 | 2120 | 0.27295 |
| H | 129 | 0.016609 | 592 | 0.07622 | 389 | 0.050084 | 131 | 0.016866 |
| I | 264 | 0.03399 | 144 | 0.01854 | 88 | 0.01133 | 313 | 0.040299 |
| K | 566 | 0.072872 | 375 | 0.048281 | 273 | 0.035149 | 428 | 0.055105 |
| L | 399 | 0.051371 | 477 | 0.061414 | 384 | 0.04944 | 551 | 0.070941 |
| M | 64 | 0.00824 | 91 | 0.011716 | 66 | 0.008497 | 76 | 0.009785 |
| N | 711 | 0.091541 | 596 | 0.076735 | 1035 | 0.133256 | 259 | 0.033346 |
| P | 422 | 0.054332 | 214 | 0.027552 | 19 | 0.002446 | 44 | 0.005665 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 140 | 0.018025 | 460 | 0.059225 | 377 | 0.048539 | 181 | 0.023304 |
| R | 362 | 0.046607 | 303 | 0.039011 | 323 | 0.041586 | 228 | 0.029355 |
| S | 757 | 0.097464 | 597 | 0.076864 | 561 | 0.072229 | 360 | 0.04635 |
| T | 784 | 0.10094 | 1247 | 0.160551 | 1139 | 0.146646 | 352 | 0.04532 |
| V | 331 | 0.042616 | 172 | 0.022145 | 154 | 0.019827 | 646 | 0.083172 |
| W | 165 | 0.021244 | 63 | 0.008111 | 67 | 0.008626 | 91 | 0.011716 |
| Y | 203 | 0.026136 | 241 | 0.031029 | 374 | 0.048152 | 730 | 0.093987 |
| Total | 7767 | 1 | 7767 | 1 | 7767 | 1 | 7767 | 1 |
| [SC12]-[AC4] | | | | | | | | |
| A | 294 | 0.055066 | 106 | 0.019854 | 126 | 0.0236 | 285 | 0.053381 |
| C | 110 | 0.020603 | 21 | 0.003933 | 138 | 0.025848 | 53 | 0.009927 |
| D | 320 | 0.059936 | 328 | 0.061435 | 1473 | 0.275894 | 341 | 0.06387 |
| E | 399 | 0.074733 | 111 | 0.02079 | 271 | 0.050759 | 301 | 0.056378 |
| F | 315 | 0.059 | 20 | 0.003746 | 62 | 0.011613 | 143 | 0.026784 |
| G | 263 | 0.04926 | 3721 | 0.696947 | 279 | 0.052257 | 563 | 0.10545 |
| H | 175 | 0.032778 | 19 | 0.003559 | 53 | 0.009927 | 88 | 0.016482 |
| I | 305 | 0.057127 | 10 | 0.001873 | 70 | 0.013111 | 276 | 0.051695 |
| K | 360 | 0.067428 | 181 | 0.033901 | 204 | 0.038209 | 337 | 0.06312 |
| L | 373 | 0.069863 | 50 | 0.009365 | 88 | 0.016482 | 446 | 0.083536 |
| M | 68 | 0.012736 | 61 | 0.011425 | 38 | 0.007117 | 113 | 0.021165 |
| N | 132 | 0.024724 | 187 | 0.035025 | 902 | 0.168945 | 243 | 0.045514 |
| P | 108 | 0.020229 | 0 | 0 | 18 | 0.003371 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 207 | 0.038771 | 70 | 0.013111 | 151 | 0.028282 | 321 | 0.060124 |
| R | 253 | 0.047387 | 93 | 0.017419 | 112 | 0.020978 | 282 | 0.052819 |
| S | 352 | 0.06593 | 217 | 0.040644 | 567 | 0.1062 | 405 | 0.075857 |
| T | 236 | 0.044203 | 14 | 0.002622 | 555 | 0.103952 | 670 | 0.125492 |
| V | 724 | 0.135606 | 6 | 0.001124 | 171 | 0.032028 | 292 | 0.054692 |
| W | 35 | 0.006556 | 33 | 0.006181 | 24 | 0.004495 | 71 | 0.013298 |
| Y | 310 | 0.058063 | 91 | 0.017044 | 37 | 0.00693 | 109 | 0.020416 |
| Total | 5339 | 1 | 5339 | 1 | 5339 | 1 | 5339 | 1 |
| [SC12]-[SC4] | | | | | | | | |
| A | 409 | 0.047887 | 225 | 0.026344 | 557 | 0.065215 | 1136 | 0.133006 |
| C | 219 | 0.025641 | 257 | 0.03009 | 203 | 0.023768 | 69 | 0.008079 |
| D | 307 | 0.035944 | 148 | 0.017328 | 946 | 0.11076 | 440 | 0.051516 |
| E | 498 | 0.058307 | 56 | 0.006557 | 666 | 0.077977 | 405 | 0.047418 |
| F | 639 | 0.074816 | 94 | 0.011006 | 104 | 0.012177 | 332 | 0.038871 |
| G | 929 | 0.108769 | 6742 | 0.789369 | 363 | 0.042501 | 1132 | 0.132537 |
| H | 199 | 0.023299 | 20 | 0.002342 | 72 | 0.00843 | 248 | 0.029036 |
| I | 332 | 0.038871 | 5 | 0.000585 | 228 | 0.026695 | 238 | 0.027866 |
| K | 231 | 0.027046 | 230 | 0.026929 | 451 | 0.052804 | 366 | 0.042852 |
| L | 640 | 0.074933 | 67 | 0.007845 | 217 | 0.025407 | 923 | 0.108067 |
| M | 306 | 0.035827 | 53 | 0.006205 | 250 | 0.029271 | 151 | 0.017679 |
| N | 375 | 0.043906 | 164 | 0.019201 | 626 | 0.073294 | 547 | 0.064044 |
| P | 196 | 0.022948 | 0 | 0 | 729 | 0.085353 | 2 | 0.000234 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 180 | 0.021075 | 74 | 0.008664 | 198 | 0.023182 | 164 | 0.019201 |
| R | 235 | 0.027514 | 146 | 0.017094 | 326 | 0.038169 | 473 | 0.05538 |
| S | 340 | 0.039808 | 107 | 0.012528 | 1963 | 0.229833 | 373 | 0.043672 |
| T | 433 | 0.050697 | 38 | 0.004449 | 287 | 0.033603 | 599 | 0.070132 |
| V | 1571 | 0.183936 | 3 | 0.000351 | 126 | 0.014752 | 401 | 0.04695 |
| W | 42 | 0.004917 | 60 | 0.007025 | 127 | 0.014869 | 121 | 0.014167 |
| Y | 460 | 0.053858 | 52 | 0.006088 | 102 | 0.011942 | 421 | 0.049292 |
| Total | 8541 | 1 | 8541 | 1 | 8541 | 1 | 8541 | 1 |
| [SC2]-[AC10] | | | | | | | | |
| A | 873 | 0.062955 | 1863 | 0.134348 | 2 | 0.000144 | 941 | 0.067859 |
| C | 441 | 0.031802 | 119 | 0.008582 | 2 | 0.000144 | 249 | 0.017956 |
| D | 268 | 0.019326 | 580 | 0.041826 | 11 | 0.000793 | 2141 | 0.154395 |
| E | 955 | 0.068869 | 1324 | 0.095478 | 6 | 0.000433 | 1191 | 0.085887 |
| F | 441 | 0.031802 | 354 | 0.025528 | 6 | 0.000433 | 483 | 0.034831 |
| G | 827 | 0.059638 | 192 | 0.013846 | 13762 | 0.992428 | 1033 | 0.074493 |
| H | 292 | 0.021057 | 125 | 0.009014 | 1 | 7.21E-05 | 267 | 0.019254 |
| I | 539 | 0.038869 | 542 | 0.039086 | 0 | 0 | 585 | 0.042186 |
| K | 1184 | 0.085383 | 1392 | 0.100382 | 2 | 0.000144 | 835 | 0.060215 |
| L | 925 | 0.066705 | 411 | 0.029639 | 3 | 0.000216 | 754 | 0.054374 |
| M | 223 | 0.016081 | 149 | 0.010745 | 1 | 7.21E-05 | 221 | 0.015937 |
| N | 572 | 0.041249 | 472 | 0.034038 | 24 | 0.001731 | 292 | 0.021057 |
| P | 1436 | 0.103555 | 2738 | 0.197447 | 0 | 0 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 667 | 0.0481 | 667 | 0.0481 | 2 | 0.000144 | 701 | 0.050552 |
| R | 869 | 0.062667 | 462 | 0.033317 | 7 | 0.000505 | 347 | 0.025023 |
| S | 707 | 0.050984 | 660 | 0.047595 | 29 | 0.002091 | 898 | 0.064758 |
| T | 909 | 0.065551 | 397 | 0.028629 | 2 | 0.000144 | 1304 | 0.094036 |
| V | 1020 | 0.073556 | 1154 | 0.083219 | 0 | 0 | 860 | 0.062018 |
| W | 224 | 0.016153 | 78 | 0.005625 | 4 | 0.000288 | 205 | 0.014783 |
| Y | 495 | 0.035696 | 188 | 0.013557 | 3 | 0.000216 | 560 | 0.040384 |
| Total | 13867 | 1 | 13867 | 1 | 13867 | 1 | 13867 | 1 |
| [SC2]-[SC9] | | | | | | | | |
| A | 601 | 0.094378 | 384 | 0.060302 | 229 | 0.035961 | 694 | 0.108982 |
| C | 40 | 0.006281 | 19 | 0.002984 | 73 | 0.011464 | 127 | 0.019943 |
| D | 147 | 0.023084 | 435 | 0.06831 | 1042 | 0.163631 | 573 | 0.089981 |
| E | 198 | 0.031093 | 555 | 0.087155 | 260 | 0.040829 | 173 | 0.027167 |
| F | 270 | 0.042399 | 202 | 0.031721 | 143 | 0.022456 | 232 | 0.036432 |
| G | 769 | 0.12076 | 143 | 0.022456 | 515 | 0.080873 | 608 | 0.095477 |
| H | 157 | 0.024655 | 150 | 0.023555 | 370 | 0.058103 | 88 | 0.013819 |
| I | 396 | 0.062186 | 160 | 0.025126 | 10 | 0.00157 | 141 | 0.022142 |
| K | 356 | 0.055905 | 345 | 0.054177 | 165 | 0.025911 | 190 | 0.029837 |
| L | 738 | 0.115892 | 200 | 0.031407 | 214 | 0.033606 | 162 | 0.02544 |
| M | 120 | 0.018844 | 126 | 0.019786 | 94 | 0.014761 | 99 | 0.015546 |
| N | 260 | 0.040829 | 464 | 0.072864 | 2031 | 0.318938 | 220 | 0.034548 |
| P | 709 | 0.111338 | 1012 | 0.15892 | 0 | 0 | 771 | 0.121074 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 141 | 0.022142 | 357 | 0.056062 | 158 | 0.024812 | 179 | 0.028109 |
| R | 187 | 0.029366 | 521 | 0.081815 | 394 | 0.061872 | 208 | 0.032663 |
| S | 192 | 0.030151 | 329 | 0.051665 | 243 | 0.03816 | 609 | 0.095634 |
| T | 328 | 0.051508 | 247 | 0.038788 | 74 | 0.011621 | 896 | 0.140704 |
| V | 461 | 0.072393 | 177 | 0.027795 | 30 | 0.004711 | 220 | 0.034548 |
| W | 74 | 0.011621 | 78 | 0.012249 | 69 | 0.010835 | 36 | 0.005653 |
| Y | 224 | 0.035176 | 464 | 0.072864 | 254 | 0.039887 | 142 | 0.022299 |
| Total | 6368 | 1 | 6368 | 1 | 6368 | 1 | 6368 | 1 |
| [SC4]-[AC1] | | | | | | | | |
| A | 334 | 0.065567 | 328 | 0.064389 | 232 | 0.045544 | 234 | 0.045936 |
| C | 50 | 0.009815 | 44 | 0.008638 | 25 | 0.004908 | 18 | 0.003534 |
| D | 364 | 0.071457 | 386 | 0.075775 | 125 | 0.024539 | 962 | 0.18885 |
| E | 185 | 0.036317 | 226 | 0.044366 | 193 | 0.037888 | 136 | 0.026698 |
| F | 181 | 0.035532 | 47 | 0.009227 | 198 | 0.038869 | 99 | 0.019435 |
| G | 1161 | 0.227915 | 85 | 0.016686 | 104 | 0.020416 | 313 | 0.061445 |
| H | 114 | 0.022379 | 40 | 0.007852 | 127 | 0.024931 | 39 | 0.007656 |
| I | 48 | 0.009423 | 510 | 0.100118 | 217 | 0.042599 | 121 | 0.023753 |
| K | 149 | 0.02925 | 409 | 0.080291 | 421 | 0.082646 | 197 | 0.038673 |
| L | 516 | 0.101296 | 262 | 0.051433 | 593 | 0.116411 | 156 | 0.030624 |
| M | 43 | 0.008441 | 29 | 0.005693 | 110 | 0.021594 | 26 | 0.005104 |
| N | 141 | 0.02768 | 171 | 0.033569 | 81 | 0.015901 | 219 | 0.042992 |
| P | 728 | 0.142913 | 445 | 0.087358 | 0 | 0 | 784 | 0.153907 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 55 | 0.010797 | 287 | 0.056341 | 403 | 0.079113 | 117 | 0.022968 |
| R | 102 | 0.020024 | 226 | 0.044366 | 193 | 0.037888 | 415 | 0.081468 |
| S | 383 | 0.075186 | 1000 | 0.196309 | 972 | 0.190813 | 252 | 0.04947 |
| T | 255 | 0.050059 | 183 | 0.035925 | 528 | 0.103651 | 724 | 0.142128 |
| V | 103 | 0.02022 | 57 | 0.01119 | 370 | 0.072634 | 154 | 0.030232 |
| W | 59 | 0.011582 | 274 | 0.053789 | 43 | 0.008441 | 37 | 0.007263 |
| Y | 123 | 0.024146 | 85 | 0.016686 | 159 | 0.031213 | 91 | 0.017864 |
| Total | 5094 | 1 | 5094 | 1 | 5094 | 1 | 5094 | 1 |
| [SC4]-[AC2] | | | | | | | | |
| A | 539 | 0.063764 | 754 | 0.089199 | 231 | 0.027328 | 487 | 0.057613 |
| C | 169 | 0.019993 | 61 | 0.007216 | 87 | 0.010292 | 32 | 0.003786 |
| D | 438 | 0.051816 | 570 | 0.067432 | 1332 | 0.157577 | 439 | 0.051934 |
| E | 313 | 0.037028 | 872 | 0.103159 | 430 | 0.05087 | 261 | 0.030877 |
| F | 268 | 0.031705 | 185 | 0.021886 | 449 | 0.053117 | 167 | 0.019756 |
| G | 1003 | 0.118656 | 341 | 0.040341 | 57 | 0.006743 | 508 | 0.060097 |
| H | 169 | 0.019993 | 98 | 0.011594 | 274 | 0.032415 | 125 | 0.014788 |
| I | 143 | 0.016917 | 293 | 0.034662 | 721 | 0.085295 | 299 | 0.035372 |
| K | 398 | 0.047084 | 480 | 0.056785 | 370 | 0.043771 | 738 | 0.087306 |
| L | 557 | 0.065894 | 449 | 0.053117 | 569 | 0.067313 | 472 | 0.055838 |
| M | 94 | 0.01112 | 50 | 0.005915 | 77 | 0.009109 | 95 | 0.011239 |
| N | 366 | 0.043298 | 356 | 0.042115 | 639 | 0.075594 | 422 | 0.049923 |
| P | 913 | 0.108009 | 1395 | 0.16503 | 0 | 0 | 1530 | 0.181001 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 150 | 0.017745 | 261 | 0.030877 | 272 | 0.032178 | 554 | 0.065539 |
| R | 223 | 0.026381 | 499 | 0.059032 | 353 | 0.04176 | 301 | 0.035609 |
| S | 1073 | 0.126937 | 868 | 0.102685 | 495 | 0.058559 | 790 | 0.093458 |
| T | 1006 | 0.119011 | 458 | 0.054182 | 770 | 0.091092 | 500 | 0.059151 |
| V | 339 | 0.040104 | 192 | 0.022714 | 1011 | 0.119603 | 446 | 0.052762 |
| W | 89 | 0.010529 | 122 | 0.014433 | 65 | 0.00769 | 57 | 0.006743 |
| Y | 203 | 0.024015 | 149 | 0.017627 | 251 | 0.029694 | 230 | 0.027209 |
| Total | 8453 | 1 | 8453 | 1 | 8453 | 1 | 8453 | 1 |
| [SC4]-[AC3] | | | | | | | | |
| A | 680 | 0.068143 | 785 | 0.078665 | 847 | 0.084878 | 527 | 0.052811 |
| C | 282 | 0.028259 | 89 | 0.008919 | 98 | 0.009821 | 145 | 0.014531 |
| D | 949 | 0.0951 | 981 | 0.098306 | 1458 | 0.146107 | 242 | 0.024251 |
| E | 518 | 0.051909 | 746 | 0.074757 | 384 | 0.038481 | 265 | 0.026556 |
| F | 327 | 0.032769 | 172 | 0.017236 | 666 | 0.06674 | 360 | 0.036076 |
| G | 843 | 0.084477 | 558 | 0.055917 | 81 | 0.008117 | 978 | 0.098006 |
| H | 239 | 0.02395 | 249 | 0.024952 | 773 | 0.077463 | 96 | 0.00962 |
| I | 375 | 0.037579 | 184 | 0.018439 | 202 | 0.020243 | 401 | 0.040184 |
| K | 435 | 0.043592 | 855 | 0.08568 | 407 | 0.040786 | 227 | 0.022748 |
| L | 528 | 0.052911 | 395 | 0.039583 | 479 | 0.048001 | 372 | 0.037278 |
| M | 84 | 0.008418 | 110 | 0.011023 | 129 | 0.012927 | 133 | 0.013328 |
| N | 724 | 0.072552 | 432 | 0.043291 | 1645 | 0.164846 | 199 | 0.019942 |
| P | 831 | 0.083275 | 1380 | 0.13829 | 0 | 0 | 4223 | 0.423189 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 553 | 0.055416 | 373 | 0.037378 | 355 | 0.035575 | 176 | 0.017637 |
| R | 286 | 0.02866 | 388 | 0.038882 | 555 | 0.055617 | 239 | 0.02395 |
| S | 968 | 0.097004 | 1017 | 0.101914 | 537 | 0.053813 | 402 | 0.040285 |
| T | 524 | 0.05251 | 776 | 0.077763 | 426 | 0.04269 | 354 | 0.035474 |
| V | 293 | 0.029362 | 168 | 0.016835 | 401 | 0.040184 | 427 | 0.04279 |
| W | 104 | 0.010422 | 190 | 0.01904 | 111 | 0.011123 | 78 | 0.007816 |
| Y | 436 | 0.043692 | 131 | 0.013128 | 425 | 0.042589 | 135 | 0.013528 |
| Total | 9979 | 1 | 9979 | 1 | 9979 | 1 | 9979 | 1 |
| [SC4]-[AC5] | | | | | | | | |
| A | 433 | 0.067205 | 555 | 0.08614 | 296 | 0.045941 | 482 | 0.07481 |
| C | 281 | 0.043613 | 53 | 0.008226 | 70 | 0.010865 | 264 | 0.040975 |
| D | 1316 | 0.204253 | 550 | 0.085364 | 355 | 0.055099 | 381 | 0.059134 |
| E | 219 | 0.03399 | 635 | 0.098557 | 372 | 0.057737 | 187 | 0.029024 |
| F | 129 | 0.020022 | 204 | 0.031662 | 511 | 0.079311 | 344 | 0.053391 |
| G | 398 | 0.061772 | 405 | 0.062859 | 59 | 0.009157 | 335 | 0.051994 |
| H | 273 | 0.042372 | 135 | 0.020953 | 202 | 0.031352 | 144 | 0.02235 |
| I | 69 | 0.010709 | 69 | 0.010709 | 591 | 0.091727 | 344 | 0.053391 |
| K | 152 | 0.023591 | 406 | 0.063014 | 486 | 0.075431 | 230 | 0.035698 |
| L | 338 | 0.05246 | 263 | 0.040819 | 504 | 0.078224 | 463 | 0.071861 |
| M | 61 | 0.009468 | 49 | 0.007605 | 76 | 0.011796 | 54 | 0.008381 |
| N | 749 | 0.11625 | 434 | 0.06736 | 303 | 0.047028 | 325 | 0.050442 |
| P | 451 | 0.069998 | 829 | 0.128667 | 0 | 0 | 59 | 0.009157 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 153 | 0.023747 | 191 | 0.029645 | 217 | 0.03368 | 183 | 0.028403 |
| R | 203 | 0.031507 | 549 | 0.085209 | 305 | 0.047338 | 379 | 0.058824 |
| S | 634 | 0.098401 | 615 | 0.095452 | 261 | 0.040509 | 366 | 0.056806 |
| T | 302 | 0.046873 | 297 | 0.046097 | 513 | 0.079621 | 841 | 0.130529 |
| V | 138 | 0.021419 | 64 | 0.009933 | 678 | 0.10523 | 497 | 0.077138 |
| W | 59 | 0.009157 | 52 | 0.008071 | 122 | 0.018935 | 165 | 0.025609 |
| Y | 85 | 0.013193 | 88 | 0.013658 | 522 | 0.081018 | 400 | 0.062083 |
| Total | 6443 | 1 | 6443 | 1 | 6443 | 1 | 6443 | 1 |
| [SC4]-[SC5] | | | | | | | | |
| A | 525 | 0.039483 | 1573 | 0.118297 | 1098 | 0.082575 | 735 | 0.055276 |
| C | 468 | 0.035196 | 231 | 0.017372 | 83 | 0.006242 | 385 | 0.028954 |
| D | 3326 | 0.250132 | 620 | 0.046627 | 918 | 0.069038 | 1255 | 0.094382 |
| E | 536 | 0.04031 | 777 | 0.058434 | 1466 | 0.11025 | 671 | 0.050463 |
| F | 231 | 0.017372 | 458 | 0.034444 | 403 | 0.030308 | 539 | 0.040535 |
| G | 980 | 0.073701 | 476 | 0.035798 | 452 | 0.033993 | 886 | 0.066632 |
| H | 421 | 0.031661 | 344 | 0.02587 | 489 | 0.036775 | 386 | 0.029029 |
| I | 132 | 0.009927 | 282 | 0.021208 | 702 | 0.052794 | 505 | 0.037978 |
| K | 200 | 0.015041 | 582 | 0.043769 | 1271 | 0.095585 | 722 | 0.054298 |
| L | 427 | 0.032113 | 920 | 0.069189 | 960 | 0.072197 | 995 | 0.074829 |
| M | 177 | 0.013311 | 208 | 0.015643 | 125 | 0.009401 | 158 | 0.011882 |
| N | 2025 | 0.15229 | 488 | 0.0367 | 736 | 0.055351 | 844 | 0.063473 |
| P | 552 | 0.041513 | 2557 | 0.192299 | 15 | 0.001128 | 332 | 0.024968 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 189 | 0.014214 | 228 | 0.017147 | 479 | 0.036023 | 245 | 0.018425 |
| R | 258 | 0.019403 | 402 | 0.030232 | 870 | 0.065428 | 562 | 0.042265 |
| S | 1535 | 0.11544 | 1387 | 0.104309 | 754 | 0.056705 | 903 | 0.06791 |
| T | 592 | 0.044521 | 972 | 0.073099 | 780 | 0.05866 | 1597 | 0.120102 |
| V | 226 | 0.016996 | 380 | 0.028578 | 1130 | 0.084982 | 422 | 0.031736 |
| W | 210 | 0.015793 | 198 | 0.014891 | 147 | 0.011055 | 457 | 0.034369 |
| Y | 287 | 0.021584 | 214 | 0.016094 | 419 | 0.031511 | 698 | 0.052493 |
| Total | 13297 | 1 | 13297 | 1 | 13297 | 1 | 13297 | 1 |
| [SC5]-[AC2] | | | | | | | | |
| A | 1436 | 0.143528 | 908 | 0.090755 | 251 | 0.025087 | 696 | 0.069565 |
| C | 120 | 0.011994 | 80 | 0.007996 | 74 | 0.007396 | 92 | 0.009195 |
| D | 374 | 0.037381 | 810 | 0.08096 | 1256 | 0.125537 | 398 | 0.03978 |
| E | 314 | 0.031384 | 1013 | 0.101249 | 608 | 0.06077 | 340 | 0.033983 |
| F | 535 | 0.053473 | 328 | 0.032784 | 404 | 0.04038 | 465 | 0.046477 |
| G | 1015 | 0.101449 | 155 | 0.015492 | 42 | 0.004198 | 639 | 0.063868 |
| H | 177 | 0.017691 | 148 | 0.014793 | 322 | 0.032184 | 230 | 0.022989 |
| I | 373 | 0.037281 | 421 | 0.042079 | 875 | 0.087456 | 329 | 0.032884 |
| K | 305 | 0.030485 | 790 | 0.078961 | 502 | 0.050175 | 630 | 0.062969 |
| L | 919 | 0.091854 | 790 | 0.078961 | 562 | 0.056172 | 516 | 0.051574 |
| M | 154 | 0.015392 | 108 | 0.010795 | 94 | 0.009395 | 171 | 0.017091 |
| N | 428 | 0.042779 | 442 | 0.044178 | 852 | 0.085157 | 342 | 0.034183 |
| P | 1524 | 0.152324 | 510 | 0.050975 | 3 | 0.0003 | 2417 | 0.241579 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 271 | 0.027086 | 372 | 0.037181 | 361 | 0.036082 | 268 | 0.026787 |
| R | 335 | 0.033483 | 563 | 0.056272 | 441 | 0.044078 | 395 | 0.03948 |
| S | 422 | 0.042179 | 471 | 0.047076 | 687 | 0.068666 | 599 | 0.05987 |
| T | 389 | 0.038881 | 1142 | 0.114143 | 1138 | 0.113743 | 486 | 0.048576 |
| V | 500 | 0.049975 | 580 | 0.057971 | 984 | 0.098351 | 698 | 0.069765 |
| W | 136 | 0.013593 | 75 | 0.007496 | 91 | 0.009095 | 144 | 0.014393 |
| Y | 278 | 0.027786 | 299 | 0.029885 | 458 | 0.045777 | 150 | 0.014993 |
| Total | 10005 | 1 | 10005 | 1 | 10005 | 1 | 10005 | 1 |
| [SC5]-[AC3] | | | | | | | | |
| A | 633 | 0.11457 | 971 | 0.175747 | 359 | 0.064977 | 364 | 0.065882 |
| C | 155 | 0.028054 | 35 | 0.006335 | 155 | 0.028054 | 181 | 0.03276 |
| D | 164 | 0.029683 | 392 | 0.07095 | 803 | 0.145339 | 168 | 0.030407 |
| E | 268 | 0.048507 | 575 | 0.104072 | 276 | 0.049955 | 86 | 0.015566 |
| F | 155 | 0.028054 | 96 | 0.017376 | 445 | 0.080543 | 82 | 0.014842 |
| G | 800 | 0.144796 | 82 | 0.014842 | 45 | 0.008145 | 407 | 0.073665 |
| H | 82 | 0.014842 | 97 | 0.017557 | 227 | 0.041086 | 92 | 0.016652 |
| I | 439 | 0.079457 | 132 | 0.023891 | 69 | 0.012489 | 121 | 0.0219 |
| K | 295 | 0.053394 | 417 | 0.075475 | 226 | 0.040905 | 171 | 0.03095 |
| L | 436 | 0.078914 | 324 | 0.058643 | 183 | 0.033122 | 170 | 0.030769 |
| M | 126 | 0.022805 | 134 | 0.024253 | 86 | 0.015566 | 71 | 0.012851 |
| N | 245 | 0.044344 | 149 | 0.026968 | 822 | 0.148778 | 121 | 0.0219 |
| P | 299 | 0.054118 | 375 | 0.067873 | 0 | 0 | 2428 | 0.439457 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 163 | 0.029502 | 222 | 0.040181 | 144 | 0.026063 | 76 | 0.013756 |
| R | 143 | 0.025882 | 220 | 0.039819 | 311 | 0.05629 | 114 | 0.020633 |
| S | 260 | 0.047059 | 482 | 0.08724 | 229 | 0.041448 | 299 | 0.054118 |
| T | 333 | 0.060271 | 366 | 0.066244 | 356 | 0.064434 | 268 | 0.048507 |
| V | 322 | 0.058281 | 298 | 0.053937 | 116 | 0.020995 | 172 | 0.031131 |
| W | 64 | 0.011584 | 26 | 0.004706 | 152 | 0.027511 | 55 | 0.009955 |
| Y | 143 | 0.025882 | 132 | 0.023891 | 521 | 0.094299 | 79 | 0.014299 |
| Total | 5525 | 1 | 5525 | 1 | 5525 | 1 | 5525 | 1 |
| [SC5]-[SC5] | | | | | | | | |
| A | 733 | 0.044978 | 1489 | 0.091367 | 2280 | 0.139903 | 1165 | 0.071486 |
| C | 297 | 0.018224 | 72 | 0.004418 | 107 | 0.006566 | 200 | 0.012272 |
| D | 2413 | 0.148064 | 819 | 0.050255 | 1102 | 0.06762 | 1461 | 0.089648 |
| E | 487 | 0.029883 | 1540 | 0.094496 | 2275 | 0.139596 | 1418 | 0.08701 |
| F | 450 | 0.027612 | 553 | 0.033933 | 415 | 0.025465 | 634 | 0.038903 |
| G | 955 | 0.0586 | 601 | 0.036878 | 524 | 0.032153 | 1404 | 0.086151 |
| H | 455 | 0.027919 | 294 | 0.01804 | 270 | 0.016567 | 288 | 0.017672 |
| I | 237 | 0.014543 | 921 | 0.056513 | 405 | 0.024851 | 324 | 0.019881 |
| K | 476 | 0.029208 | 1469 | 0.090139 | 1659 | 0.101798 | 885 | 0.054304 |
| L | 930 | 0.057066 | 1391 | 0.085353 | 809 | 0.049641 | 1184 | 0.072651 |
| M | 106 | 0.006504 | 293 | 0.017979 | 177 | 0.010861 | 372 | 0.022826 |
| N | 1104 | 0.067743 | 319 | 0.019574 | 352 | 0.021599 | 799 | 0.049027 |
| P | 910 | 0.055838 | 1276 | 0.078297 | 553 | 0.033933 | 356 | 0.021845 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 256 | 0.015708 | 428 | 0.026263 | 843 | 0.051727 | 665 | 0.040805 |
| R | 581 | 0.035651 | 685 | 0.042032 | 764 | 0.04688 | 556 | 0.034117 |
| S | 3110 | 0.190833 | 1703 | 0.104498 | 1516 | 0.093023 | 1364 | 0.083696 |
| T | 1821 | 0.111738 | 704 | 0.043198 | 560 | 0.034362 | 1230 | 0.075474 |
| V | 545 | 0.033442 | 1150 | 0.070565 | 1134 | 0.069583 | 862 | 0.052893 |
| W | 139 | 0.008529 | 214 | 0.013131 | 209 | 0.012824 | 189 | 0.011597 |
| Y | 292 | 0.017917 | 376 | 0.023072 | 343 | 0.021047 | 941 | 0.057741 |
| Total | 16297 | 1 | 16297 | 1 | 16297 | 1 | 16297 | 1 |
| [SC9]-[SC10] | | | | | | | | |
| A | 789 | 0.055815 | 793 | 0.056098 | 105 | 0.007428 | 762 | 0.053905 |
| C | 532 | 0.037634 | 167 | 0.011814 | 69 | 0.004881 | 104 | 0.007357 |
| D | 344 | 0.024335 | 3054 | 0.216044 | 597 | 0.042233 | 236 | 0.016695 |
| E | 733 | 0.051853 | 683 | 0.048316 | 212 | 0.014997 | 1181 | 0.083546 |
| F | 672 | 0.047538 | 255 | 0.018039 | 373 | 0.026387 | 167 | 0.011814 |
| G | 424 | 0.029994 | 1160 | 0.08206 | 10045 | 0.710597 | 585 | 0.041384 |
| H | 367 | 0.025962 | 345 | 0.024406 | 90 | 0.006367 | 837 | 0.059211 |
| I | 1761 | 0.124576 | 19 | 0.001344 | 1 | 7.07E-05 | 417 | 0.029499 |
| K | 758 | 0.053622 | 765 | 0.054117 | 549 | 0.038837 | 3045 | 0.215407 |
| L | 991 | 0.070105 | 297 | 0.02101 | 32 | 0.002264 | 555 | 0.039261 |
| M | 203 | 0.01436 | 72 | 0.005093 | 21 | 0.001486 | 139 | 0.009833 |
| N | 432 | 0.03056 | 4190 | 0.296406 | 1154 | 0.081636 | 653 | 0.046194 |
| P | 142 | 0.010045 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Q | 311 | 0.022001 | 474 | 0.033531 | 87 | 0.006154 | 1256 | 0.088851 |
| R | 526 | 0.03721 | 610 | 0.043152 | 203 | 0.01436 | 1284 | 0.090832 |
| S | 853 | 0.060342 | 655 | 0.046336 | 274 | 0.019383 | 597 | 0.042233 |
| T | 569 | 0.040252 | 135 | 0.00955 | 35 | 0.002476 | 592 | 0.041879 |
| V | 2403 | 0.169992 | 38 | 0.002688 | 17 | 0.001203 | 1330 | 0.094086 |
| W | 413 | 0.029216 | 69 | 0.004881 | 49 | 0.003466 | 63 | 0.004457 |
| Y | 913 | 0.064587 | 355 | 0.025113 | 223 | 0.015775 | 333 | 0.023557 |
| Total | 14136 | 1 | 14136 | 1 | 14136 | 1 | 14136 | 1 |

**Table 5. 10 amino acid preferences for the top 19 β turn types defined in the new scheme**

### 5.4.6 The impact of ω turns in the new classification scheme

While the ω turn is not involved in the traditional classification of β turns and, we also did not use the ω turn in constructing the new classification scheme. However, since both cis and trans ω angles can exist in β turns, we evaluated the occurrence and distribution of cis ω angles in the new classification scheme, and our analysis is summarized in **Table 5.11**. Based on our analysis, 201 of new turn types contained cis ω angles. Cis ω angles were not detected in the remaining 381 new turn types. Of the 201 turn types containing cis ω angles, only 33 had more cis ω angles than turns with all trans ω angles. Furthermore, the 33 new turn types containing cis ω angles only contained cis ω angles, i.e. trans ω angles were not observed in those 33 new turn types. Overall, only around 11.3% of new turn types contained cis ω angles. In the case that a researcher encounters a β turn that contains a cis ω angle, our nomenclature scheme can be modified in the following way to indicate the presence of the cis ω angle. If we indicate a cis ω angle in a β turn with a negative sign and a trans ω angle in a β turn with a positive sign, any given β turn in the new schema would have four possible combinations of cis and trans ω angles. For example, the four possible combinations in a [SC4]-[AC2] β turn would be [SC4]-[AC2], [-SC4]-[AC2], [SC4]-[-AC2], [-SC4]-[-AC2]. This nomenclature system is utilized to indicate the distribution of cis ω angles listed in **Table 5.11**.

| Original | With application of omega angle | Number of data points |
|---|---|---|
| [AC1]-[AC11] | [-AC1]-[AC11] | 1 |
| | [-AC1]-[AC11] | 1 |
| | [AC1]-[AC11] | 1 |
| [AC1]-[AC4] | [AC1]-[-AC4] | 5 |

| | | |
|---|---|---|
| | [AC1]-[AC4] | 3 |
| [AC1]-[SC1] | [AC1]-[-SC1] | 41 |
| | [AC1]-[SC1] | 213 |
| [AC1]-[SC2] | [AC1]-[-SC2] | 38 |
| | [AC1]-[SC2] | 257 |
| [AC1]-[SC4] | [AC1]-[-SC4] | 39 |
| | [AC1]-[SC4] | 51 |
| [AC1]-[SC9] | [-AC1]-[SC9] | 2 |
| | [AC1]-[SC9] | 1957 |
| [AC10]-[AC1] | [AC10]-[-AC1] | 2 |
| | [AC10]-[AC1] | 13 |
| [AC10]-[AC4] | [AC10]-[-AC4] | 1 |
| | [AC10]-[AC4] | 1044 |
| [AC11]-[AC4] | [-AC11]-[AC4] | 3 |
| | [AC11]-[AC4] | 309 |
| [AC11]-[AP3] | [-AC11]-[AP3] | 1 |
| [AC11]-[SC1] | [AC11]-[-SC1] | 1 |
| | [AC11]-[SC1] | 18 |
| [AC11]-[SC4] | [-AC11]-[SC4] | 1 |
| | [AC11]-[SC4] | 60 |
| [AC12]-[AC1] | [-AC12]-[AC1] | 1 |
| [AC12]-[AC2] | [-AC12]-[AC2] | 1 |
| | [AC12]-[AC2] | 9 |
| [AC12]-[AC4] | [-AC12]-[AC4] | 1 |
| | [AC12]-[AC4] | 27 |
| [AC12]-[SC2] | [AC12]-[-SC2] | 2 |
| | [AC12]-[SC2] | 15 |
| [AC2]-[AC1] | [AC2]-[-AC1] | 24 |
| | [AC2]-[AC1] | 68 |
| [AC2]-[AC2] | [AC2]-[-AC2] | 17 |
| | [AC2]-[AC2] | 58 |
| [AC2]-[AC3] | [AC2]-[-AC3] | 28 |
| | [AC2]-[AC3] | 57 |
| [AC2]-[AC4] | [-AC2]-[AC4] | 1 |
| | [AC2]-[-AC4] | 518 |

| | | |
|---|---|---|
| | [AC2]-[AC4] | 28 |
| [AC2]-[AC5] | [AC2]-[-AC5] | 8 |
| [AC2]-[AP1] | [AC2]-[-AP1] | 2 |
| [AC2]-[AP2] | [AC2]-[-AP2] | 1 |
| [AC2]-[SC1] | [AC2]-[-SC1] | 785 |
| | [AC2]-[SC1] | 2091 |
| [AC2]-[SC2] | [AC2]-[-SC2] | 315 |
| | [AC2]-[SC2] | 329 |
| [AC2]-[SC4] | [AC2]-[-SC4] | 652 |
| | [AC2]-[SC4] | 97 |
| [AC2]-[SC5] | [AC2]-[-SC5] | 2 |
| | [AC2]-[SC5] | 5 |
| [AC2]-[SC6] | [AC2]-[-SC6] | 1 |
| | [AC2]-[SC6] | 1 |
| [AC2]-[SC9] | [-AC2]-[SC9] | 2 |
| | [AC2]-[SC9] | 4228 |
| [AC2]-[SP2] | [AC2]-[-SP2] | 1 |
| | [AC2]-[SP2] | 2 |
| [AC3]-[AC1] | [-AC3]-[AC1] | 1 |
| | [AC3]-[-AC1] | 49 |
| | [AC3]-[AC1] | 60 |
| [AC3]-[AC2] | [AC3]-[-AC2] | 10 |
| | [AC3]-[AC2] | 2 |
| [AC3]-[AC3] | [-AC3]-[AC3] | 3 |
| | [AC3]-[-AC3] | 41 |
| | [AC3]-[AC3] | 23 |
| [AC3]-[AC4] | [-AC3]-[AC4] | 21 |
| | [AC3]-[-AC4] | 9 |
| | [AC3]-[AC4] | 141 |
| [AC3]-[AC5] | [-AC3]-[AC5] | 4 |
| | [AC3]-[AC5] | 28 |
| [AC3]-[AP1] | [AC3]-[-AP1] | 10 |
| | [AC3]-[AP1] | 17 |
| [AC3]-[AP2] | [AC3]-[-AP2] | 2 |
| | [AC3]-[AP2] | 4 |

| | | |
|---|---|---|
| [AC3]-[AP3] | [AC3]-[-AP3] | 3 |
| | [AC3]-[AP3] | 6 |
| [AC3]-[AP5] | [-AC3]-[AP5] | 1 |
| | [AC3]-[AP5] | 17 |
| [AC3]-[SC1] | [AC3]-[-SC1] | 30 |
| | [AC3]-[SC1] | 80 |
| [AC3]-[SC10] | [-AC3]-[SC10] | 2 |
| | [AC3]-[SC10] | 454 |
| [AC3]-[SC12] | [AC3]-[-SC12] | 1 |
| | [AC3]-[SC12] | 3 |
| [AC3]-[SC2] | [AC3]-[-SC2] | 1 |
| [AC3]-[SC3] | [-AC3]-[SC3] | 1 |
| | [AC3]-[SC3] | 1 |
| [AC3]-[SC4] | [-AC3]-[SC4] | 6 |
| | [AC3]-[-SC4] | 2 |
| | [AC3]-[SC4] | 4 |
| [AC3]-[SC5] | [-AC3]-[SC5] | 1 |
| | [AC3]-[SC5] | 6 |
| [AC4]-[AC1] | [-AC4]-[AC1] | 22 |
| | [AC4]-[AC1] | 721 |
| [AC4]-[AC11] | [-AC4]-[AC11] | 1 |
| | [AC4]-[AC11] | 74 |
| [AC4]-[AC2] | [-AC4]-[AC2] | 262 |
| | [AC4]-[AC2] | 1725 |
| [AC4]-[AC3] | [-AC4]-[AC3] | 34 |
| | [AC4]-[AC3] | 4146 |
| [AC4]-[AC4] | [-AC4]-[AC4] | 134 |
| | [AC4]-[-AC4] | 20 |
| | [AC4]-[AC4] | 7613 |
| [AC4]-[AC5] | [-AC4]-[AC5] | 16 |
| | [AC4]-[AC5] | 2553 |
| [AC4]-[AC6] | [-AC4]-[AC6] | 3 |
| | [AC4]-[AC6] | 354 |
| [AC4]-[AC8] | [-AC4]-[AC8] | 1 |
| | [AC4]-[AC8] | 26 |

| | | |
|---|---|---|
| [AC4]-[AP1] | [-AC4]-[AP1] | 8 |
| | [AC4]-[AP1] | 567 |
| [AC4]-[AP2] | [-AC4]-[AP2] | 1 |
| | [AC4]-[AP2] | 545 |
| [AC4]-[AP3] | [AC4]-[-AP3] | 1 |
| | [AC4]-[AP3] | 388 |
| [AC4]-[SC1] | [-AC4]-[SC1] | 51 |
| | [AC4]-[SC1] | 9 |
| [AC4]-[SC2] | [-AC4]-[SC2] | 63 |
| | [AC4]-[SC2] | 39 |
| [AC4]-[SC3] | [-AC4]-[SC3] | 19 |
| | [AC4]-[SC3] | 252 |
| [AC4]-[SC4] | [-AC4]-[SC4] | 472 |
| | [AC4]-[-SC4] | 14 |
| | [AC4]-[SC4] | 1361 |
| [AC4]-[SC5] | [-AC4]-[SC5] | 395 |
| | [AC4]-[SC5] | 1400 |
| [AC5]-[AC12] | [AC5]-[-AC12] | 1 |
| [AC5]-[AC2] | [AC5]-[-AC2] | 1 |
| | [AC5]-[AC2] | 1379 |
| [AC5]-[AC4] | [-AC5]-[AC4] | 5 |
| | [AC5]-[AC4] | 2753 |
| [AC5]-[SC2] | [AC5]-[-SC2] | 1 |
| | [AC5]-[SC2] | 520 |
| [AC5]-[SC3] | [-AC5]-[SC3] | 1 |
| | [AC5]-[SC3] | 247 |
| [AC5]-[SC4] | [-AC5]-[SC4] | 5 |
| | [AC5]-[SC4] | 2002 |
| [AC5]-[SC5] | [-AC5]-[SC5] | 1 |
| | [AC5]-[SC5] | 2402 |
| [AC5]-[SC9] | [AC5]-[-SC9] | 1 |
| [AC6]-[AC3] | [-AC6]-[-AC3] | 1 |
| [AC7]-[SC2] | [AC7]-[-SC2] | 1 |
| | [AC7]-[SC2] | 5 |
| [AC8]-[AC4] | [AC8]-[-AC4] | 1 |

| | | |
|---|---|---|
| [AC8]-[SC8] | [AC8]-[-SC8] | 1 |
| [AC9]-[AC3] | [-AC9]-[AC3] | 1 |
| [AC9]-[SC2] | [AC9]-[-SC2] | 2 |
| [AP1]-[SC2] | [AP1]-[-SC2] | 2 |
| | [AP1]-[SC2] | 9 |
| [AP1]-[SC4] | [AP1]-[-SC4] | 2 |
| | [AP1]-[SC4] | 11 |
| [AP1]-[SC9] | [AP1]-[-SC9] | 1 |
| | [AP1]-[SC9] | 43 |
| [AP2]-[AC1] | [AP2]-[-AC1] | 3 |
| | [AP2]-[AC1] | 1 |
| [AP2]-[AC2] | [AP2]-[-AC2] | 5 |
| | [AP2]-[AC2] | 3 |
| [AP2]-[AC4] | [AP2]-[-AC4] | 19 |
| | [AP2]-[AC4] | 1 |
| [AP2]-[AC6] | [AP2]-[-AC6] | 1 |
| [AP2]-[SC1] | [AP2]-[-SC1] | 240 |
| | [AP2]-[SC1] | 279 |
| [AP2]-[SC2] | [AP2]-[-SC2] | 87 |
| | [AP2]-[SC2] | 48 |
| [AP2]-[SC4] | [AP2]-[-SC4] | 58 |
| | [AP2]-[SC4] | 24 |
| [AP3]-[AC1] | [AP3]-[-AC1] | 5 |
| | [AP3]-[AC1] | 1 |
| [AP3]-[AC2] | [AP3]-[-AC2] | 3 |
| [AP3]-[AC4] | [AP3]-[-AC4] | 2 |
| | [AP3]-[AC4] | 2 |
| [AP3]-[AC5] | [AP3]-[-AC5] | 1 |
| [AP3]-[SC1] | [AP3]-[-SC1] | 1 |
| | [AP3]-[SC1] | 3 |
| [AP4]-[AC4] | [-AP4]-[AC4] | 1 |
| | [AP4]-[AC4] | 18 |
| [AP5]-[SC2] | [AP5]-[-SC2] | 4 |
| | [AP5]-[SC2] | 14 |
| [AP5]-[SC6] | [AP5]-[-SC6] | 1 |

| | | |
|---|---|---|
| [AP6]-[AC3] | [-AP6]-[AC3] | 1 |
| | [AP6]-[AC3] | 2 |
| [AP6]-[SC7] | [AP6]-[-SC7] | 2 |
| | [AP6]-[SC7] | 1 |
| [SC1]-[AC10] | [-SC1]-[AC10] | 49 |
| | [SC1]-[AC10] | 174 |
| [SC1]-[AC11] | [-SC1]-[AC11] | 5 |
| | [SC1]-[AC11] | 44 |
| [SC1]-[AC12] | [-SC1]-[AC12] | 9 |
| | [SC1]-[AC12] | 6 |
| [SC1]-[AC2] | [-SC1]-[AC2] | 1 |
| | [SC1]-[-AC2] | 1 |
| | [SC1]-[AC2] | 14 |
| [SC1]-[AC3] | [SC1]-[-AC3] | 4 |
| | [SC1]-[AC3] | 11 |
| [SC1]-[AC4] | [SC1]-[-AC4] | 64 |
| | [SC1]-[AC4] | 23 |
| [SC1]-[AC7] | [-SC1]-[AC7] | 3 |
| [SC1]-[AC9] | [-SC1]-[AC9] | 8 |
| | [SC1]-[AC9] | 9 |
| [SC1]-[AP3] | [SC1]-[-AP3] | 1 |
| [SC1]-[AP5] | [-SC1]-[AP5] | 5 |
| | [SC1]-[AP5] | 5 |
| [SC1]-[SC1] | [-SC1]-[-SC1] | 2 |
| | [SC1]-[-SC1] | 4 |
| | [SC1]-[SC1] | 41 |
| [SC1]-[SC10] | [-SC1]-[SC10] | 3 |
| | [SC1]-[SC10] | 2097 |
| [SC1]-[SC11] | [-SC1]-[SC11] | 1 |
| | [SC1]-[SC11] | 165 |
| [SC1]-[SC12] | [-SC1]-[SC12] | 1 |
| | [SC1]-[-SC12] | 1 |
| | [SC1]-[SC12] | 26 |
| [SC1]-[SC2] | [-SC1]-[SC2] | 3 |
| | [SC1]-[-SC2] | 56 |

| | | |
|---|---|---|
| | [SC1]-[SC2] | 214 |
| [SC1]-[SC3] | [SC1]-[-SC3] | 1 |
| | [SC1]-[SC3] | 23 |
| [SC1]-[SC4] | [SC1]-[-SC4] | 390 |
| | [SC1]-[SC4] | 106 |
| [SC1]-[SC5] | [SC1]-[-SC5] | 7 |
| | [SC1]-[SC5] | 15 |
| [SC1]-[SC7] | [-SC1]-[SC7] | 1 |
| | [SC1]-[SC7] | 3 |
| [SC1]-[SC9] | [-SC1]-[SC9] | 50 |
| | [SC1]-[SC9] | 3448 |
| [SC1]-[SP3] | [-SC1]-[SP3] | 1 |
| | [SC1]-[SP3] | 33 |
| [SC10]-[SC10] | [-SC10]-[SC10] | 1 |
| | [SC10]-[SC10] | 3767 |
| [SC11]-[SC5] | [-SC11]-[SC5] | 1 |
| | [SC11]-[SC5] | 199 |
| [SC12]-[AC11] | [SC12]-[-AC11] | 2 |
| | [SC12]-[AC11] | 2 |
| [SC12]-[AC12] | [SC12]-[-AC12] | 1 |
| [SC12]-[AC2] | [-SC12]-[AC2] | 1 |
| | [SC12]-[AC2] | 156 |
| [SC12]-[AC3] | [-SC12]-[AC3] | 1 |
| | [SC12]-[AC3] | 805 |
| [SC12]-[AC4] | [-SC12]-[AC4] | 1 |
| | [SC12]-[AC4] | 5338 |
| [SC12]-[AC9] | [-SC12]-[AC9] | 1 |
| [SC12]-[AP1] | [SC12]-[-AP1] | 1 |
| [SC12]-[SC1] | [SC12]-[-SC1] | 1 |
| | [SC12]-[SC1] | 2 |
| [SC12]-[SC10] | [SC12]-[-SC10] | 1 |
| [SC2]-[AC11] | [-SC2]-[AC11] | 1 |
| | [SC2]-[AC11] | 463 |
| [SC2]-[AC12] | [-SC2]-[AC12] | 2 |
| | [SC2]-[AC12] | 130 |

| | | |
|---|---|---|
| [SC2]-[AC2] | [SC2]-[-AC2] | 9 |
| | [SC2]-[AC2] | 62 |
| [SC2]-[AC3] | [SC2]-[-AC3] | 148 |
| | [SC2]-[AC3] | 66 |
| [SC2]-[AC4] | [SC2]-[-AC4] | 1327 |
| | [SC2]-[AC4] | 195 |
| [SC2]-[AC5] | [-SC2]-[-AC5] | 1 |
| | [SC2]-[-AC5] | 9 |
| | [SC2]-[AC5] | 4 |
| [SC2]-[AC7] | [-SC2]-[AC7] | 2 |
| | [SC2]-[AC7] | 7 |
| [SC2]-[AP2] | [SC2]-[-AP2] | 1 |
| [SC2]-[AP3] | [SC2]-[-AP3] | 2 |
| | [SC2]-[AP3] | 1 |
| [SC2]-[AP4] | [-SC2]-[AP4] | 1 |
| | [SC2]-[-AP4] | 1 |
| [SC2]-[AP5] | [-SC2]-[AP5] | 5 |
| | [SC2]-[AP5] | 22 |
| [SC2]-[SC1] | [SC2]-[-SC1] | 10 |
| | [SC2]-[SC1] | 64 |
| [SC2]-[SC2] | [-SC2]-[SC2] | 1 |
| | [SC2]-[-SC2] | 138 |
| | [SC2]-[SC2] | 381 |
| [SC2]-[SC3] | [SC2]-[-SC3] | 2 |
| [SC2]-[SC4] | [SC2]-[-SC4] | 1117 |
| | [SC2]-[SC4] | 56 |
| [SC2]-[SC5] | [SC2]-[-SC5] | 9 |
| | [SC2]-[SC5] | 15 |
| [SC2]-[SC7] | [-SC2]-[SC7] | 3 |
| | [SC2]-[SC7] | 125 |
| [SC2]-[SC9] | [-SC2]-[SC9] | 1 |
| | [SC2]-[SC9] | 6367 |
| [SC3]-[AP4] | [SC3]-[-AP4] | 1 |
| | [SC3]-[AP4] | 27 |
| [SC4]-[AC1] | [-SC4]-[AC1] | 61 |

| | | |
|---|---|---|
| | [SC4]-[AC1] | 5033 |
| [SC4]-[AC2] | [-SC4]-[AC2] | 413 |
| | [SC4]-[AC2] | 8040 |
| [SC4]-[AC3] | [-SC4]-[AC3] | 65 |
| | [SC4]-[AC3] | 9914 |
| [SC4]-[AC4] | [-SC4]-[AC4] | 74 |
| | [SC4]-[-AC4] | 2 |
| | [SC4]-[AC4] | 81400 |
| [SC4]-[AC5] | [-SC4]-[AC5] | 10 |
| | [SC4]-[AC5] | 6433 |
| [SC4]-[AC6] | [-SC4]-[AC6] | 7 |
| | [SC4]-[-AC6] | 1 |
| | [SC4]-[AC6] | 570 |
| [SC4]-[AP1] | [-SC4]-[AP1] | 32 |
| | [SC4]-[AP1] | 2156 |
| [SC4]-[AP2] | [-SC4]-[AP2] | 3 |
| | [SC4]-[AP2] | 1051 |
| [SC4]-[AP4] | [-SC4]-[AP4] | 1 |
| | [SC4]-[AP4] | 98 |
| [SC4]-[SC1] | [-SC4]-[SC1] | 174 |
| | [SC4]-[SC1] | 51 |
| [SC4]-[SC12] | [-SC4]-[SC12] | 3 |
| | [SC4]-[SC12] | 2 |
| [SC4]-[SC2] | [-SC4]-[SC2] | 160 |
| | [SC4]-[SC2] | 1047 |
| [SC4]-[SC3] | [-SC4]-[SC3] | 31 |
| | [SC4]-[SC3] | 1619 |
| [SC4]-[SC4] | [-SC4]-[SC4] | 106 |
| | [SC4]-[SC4] | 84636 |
| [SC4]-[SC5] | [-SC4]-[SC5] | 539 |
| | [SC4]-[SC5] | 12758 |
| [SC5]-[AC1] | [-SC5]-[AC1] | 5 |
| | [SC5]-[-AC1] | 2 |
| | [SC5]-[AC1] | 4534 |
| [SC5]-[AC10] | [-SC5]-[AC10] | 1 |

| | | |
|---|---|---|
| | [SC5]-[AC10] | 2 |
| [SC5]-[AC11] | [SC5]-[-AC11] | 3 |
| [SC5]-[AC12] | [SC5]-[-AC12] | 3 |
| [SC5]-[AC2] | [-SC5]-[AC2] | 1 |
| | [SC5]-[AC2] | 10004 |
| [SC5]-[AC3] | [-SC5]-[AC3] | 4 |
| | [SC5]-[AC3] | 5521 |
| [SC5]-[AC4] | [-SC5]-[AC4] | 7 |
| | [SC5]-[-AC4] | 2 |
| | [SC5]-[AC4] | 24729 |
| [SC5]-[AC5] | [-SC5]-[AC5] | 1 |
| | [SC5]-[AC5] | 2270 |
| [SC5]-[AP3] | [-SC5]-[AP3] | 1 |
| | [SC5]-[AP3] | 378 |
| [SC5]-[AP4] | [-SC5]-[AP4] | 1 |
| | [SC5]-[AP4] | 60 |
| [SC5]-[AP5] | [SC5]-[-AP5] | 1 |
| | [SC5]-[AP5] | 28 |
| [SC5]-[SC1] | [SC5]-[-SC1] | 1 |
| | [SC5]-[SC1] | 358 |
| [SC5]-[SC10] | [SC5]-[-SC10] | 3 |
| | [SC5]-[SC10] | 2 |
| [SC5]-[SC2] | [-SC5]-[SC2] | 6 |
| | [SC5]-[-SC2] | 11 |
| | [SC5]-[SC2] | 1844 |
| [SC5]-[SC3] | [-SC5]-[SC3] | 2 |
| | [SC5]-[SC3] | 846 |
| [SC5]-[SC4] | [-SC5]-[SC4] | 1 |
| | [SC5]-[SC4] | 55773 |
| [SC5]-[SC5] | [-SC5]-[SC5] | 4 |
| | [SC5]-[SC5] | 16293 |
| [SC5]-[SC7] | [SC5]-[-SC7] | 1 |
| | [SC5]-[SC7] | 1 |
| [SC5]-[SP5] | [SC5]-[-SP5] | 1 |
| | [SC5]-[SP5] | 3 |

| | | |
|---|---|---|
| [SC6]-[AC4] | [SC6]-[-AC4] | 1 |
| | [SC6]-[AC4] | 13 |
| [SC6]-[SC10] | [-SC6]-[-SC10] | 1 |
| | [SC6]-[SC10] | 2 |
| [SC7]-[AC11] | [SC7]-[-AC11] | 2 |
| [SC7]-[AC2] | [-SC7]-[AC2] | 2 |
| | [SC7]-[AC2] | 16 |
| [SC7]-[SC11] | [SC7]-[-SC11] | 1 |
| | [SC7]-[SC11] | 9 |
| [SC7]-[SC5] | [-SC7]-[SC5] | 5 |
| | [SC7]-[SC5] | 376 |
| [SC8]-[AC4] | [SC8]-[-AC4] | 26 |
| [SC8]-[SC1] | [-SC8]-[SC1] | 1 |
| [SC8]-[SC4] | [SC8]-[-SC4] | 9 |
| | [SC8]-[SC4] | 1 |
| [SC9]-[AC1] | [-SC9]-[AC1] | 1 |
| | [SC9]-[-AC1] | 41 |
| | [SC9]-[AC1] | 12 |
| [SC9]-[AC12] | [-SC9]-[AC12] | 1 |
| | [SC9]-[AC12] | 9 |
| [SC9]-[AC2] | [SC9]-[-AC2] | 1 |
| [SC9]-[AC4] | [SC9]-[-AC4] | 6 |
| | [SC9]-[AC4] | 21 |
| [SC9]-[AC6] | [SC9]-[-AC6] | 1 |
| [SC9]-[SC1] | [SC9]-[-SC1] | 46 |
| | [SC9]-[SC1] | 8 |
| [SC9]-[SC12] | [SC9]-[-SC12] | 1 |
| | [SC9]-[SC12] | 13 |
| [SC9]-[SC2] | [SC9]-[-SC2] | 20 |
| | [SC9]-[SC2] | 4 |
| [SC9]-[SC5] | [-SC9]-[SC5] | 1 |
| [SP5]-[AC2] | [-SP5]-[AC2] | 1 |
| | [SP5]-[AC2] | 5 |
| [SP6]-[AC11] | [SP6]-[-AC11] | 1 |

**Table 5. 11 Turn types distribution with application of Omega turn align with the category of turn types.**

### 4.4.7 Overlap between the new and traditional classification schemes

Comparison of the new classification scheme with the traditional schema indicated that 256 new turn types mapped to at least two traditional categories (**Table 5.12**). The majority of the new turns, 234 of 257, overlapped with original type IV β turn type. An additional 18 new turn types overlapped with type I and type VIII turns and four new turn types overlapped with type I and type II' or type I' and type II turns. It is perhaps not surprising that 234 new precise turn type definitions overlap with the original type IV classification, which is generally a "catch all" category for turns that did not satisfy the definition of the common type I and type II turn types in the conventional classification scheme.

| | I | I' | II | II' | IV | VIa1 | VIb | VIII |
|---|---|---|---|---|---|---|---|---|
| I | \ | 0 | 0 | 2 | 38 | 0 | 0 | 18 |
| I' | 0 | \ | 2 | 0 | 23 | 0 | 0 | 0 |
| II | 0 | 2 | \ | 0 | 42 | 0 | 0 | 0 |
| II' | 2 | 0 | 0 | \ | 30 | 0 | 0 | 0 |
| IV | 38 | 23 | 42 | 30 | \ | 15 | 21 | 65 |
| VIa1 | 0 | 0 | 0 | 0 | 15 | \ | 1 | 0 |
| VIb | 0 | 0 | 0 | 0 | 21 | 1 | \ | 0 |
| VIII | 18 | 0 | 0 | 0 | 65 | 0 | 0 | \ |

**Table 5. 12 The overlap summary of new turn type in comparison with tradition classification.** The numbers in the boxes represent the number of new turn types which belong to two classic turn types. The same background color represents those number belong to the same overlap region.

## 5.5 Conclusion

Existing systems for classifying β turns in protein structures are based on delimiting the ranges of φ and ψ backbone dihedral angles of the i+1 and i+2 residues involve in the β turn. Whether these ranges represent canonical limiting values or values optimized based on advanced clustering algorithms, the ranges are subject to potentially subject to revision and evolution as the diversity of protein structure space becomes more populated over time. A notable feature of the new β-turn classification scheme introduced here is that by design the modified Klyne-Prelog Ramachandran plot used for classification covers all possible combinations of φ and ψ backbone dihedral angles of the i+1 and i+2 residues that can occur in β-turns, and therefore the new classification scheme covers the complete potential β-turn space. As a consequence, this scheme should not require future revision or modification. Another advantage of the new classification scheme is that the nomenclature directly provides the specific Klyne-Prelog stereochemistry information for the φ and ψ backbone dihedral angles of the i+1 and i+2 residues involved in the β turns.

The new scheme also avoids problems inherent to canonical or cluster-optimized range-based classification schemes that completely change the β turn designation, for example, when a β turn with φ and ψ backbone dihedral angles of the i+1 and i+2 residues falling just 1º outside the limits defining a type II β turn by definition requires re-classification as a type IV β turn. This type of re-designation blurs the clarity of the nature of the β turn, raising questions as to whether the turn should simply be referred to as type II using expanded ranges, or be designated as a "borderline" or distorted type II β turn, or if it should simply be designated as a type IV β turn. The new classification scheme, by its nature, avoids all such ambiguity by providing clear definitions based on the Klyne-Prelog stereochemistry definitions mapped on to the well-established Ramachandran plot. Of course, the new modified Klyne-Prelog Ramachandran plot-based classification scheme can be used together with any of the other existing classification schemes used to designate β turn types, for example, in our analysis of the 2XTW PRP structure, the β turn that lies at the border of the extended type II range has the classification of a SC1-SC9 turn in the new classification scheme. Therefore, by combining the two classification schemes, the β turn could be referred to as a borderline type II β turn with a classification of SC1-SC9 in the new classification scheme.

In conclusion, the new system for defining β turns introduced here combines the Klyne-Prelog nomenclature used to specify the stereochemistry about single bonds with the conventional Ramachandran plot, yielding a modified Klyne-Prelog Ramachandran plot that can be used to specify definite stereochemistries of the φ and ψ dihedral angles of the i+1 and i+2 residues that distinguish β turns types. The new system eliminates ambiguity in classifying “border β-turns” and provides a precise system to classify β-turns in protein structures. Because the new system provides complete coverage of the dihedral torsion angle space in the Ramachandra plot, all β-turns can be assigned an unambiguous and distinct classification, even in cases where no clear β-turn type definitions exist using the conventional classification schemes. Finally, we illustrated that the new schema can be simply and easily extended to indicate whether the ω angles are cis or trans in the i+1 and i+2 residue positions in a β-turn.

### 5.6 Acknowledgements

The research was conducted with the support of Miami University. MAK acknowledges support of Miami University and the Ohio Board of Regents for funds used to establish the Ohio Eminent Scholar Laboratory where the work was performed.

### 5.7 References

1. Chou, K. C., Prediction of tight turns and their types in proteins. *Anal Biochem* **2000,** *286* (1), 1-16.
2. Chou, K. C.; Blinn, J. R., Classification and prediction of beta-turn types. *J Protein Chem* **1997,** *16* (6), 575-595.
3. Chou, P. Y.; Fasman, G. D., Conservation of Chain Reversal Regions in Proteins. *Biophysical Journal* **1979,** *26* (3), 385-399.
4. Koch, O.; Klebe, G., Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **2009,** *74* (2), 353-367.

5. Venkatachalam, C. M., Stereochemical Criteria for Polypeptides and Proteins .V. Conformation of a System of 3 Linked Peptide Units. *Biopolymers* **1968,** *6* (10), 1425-+.
6. Lewis, P. N.; Momany, F. A.; Scheraga, H. A., Chain Reversals in Proteins. *Biochim Biophys Acta* **1973,** *303* (2), 211-229.
7. Richardson, J. S., The anatomy and taxonomy of protein structure. *Adv Protein Chem* **1981,** *34*, 167-339.
8. Hutchinson, E. G.; Thornton, J. M., A Revised Set of Potentials for Beta-Turn Formation in Proteins. *Protein Sci* **1994,** *3* (12), 2207-2216.
9. de Brevern, A. G., Extension of the classical classification of beta-turns. *Sci Rep* **2016,** *6*, 33191.
10. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V., Stereochemistry of Polypeptide-Chain Configurations. *Curr Sci India* **1990,** *59* (17-18), 813-817.
11. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V., Stereochemistry of Polypeptide Chain Configurations. *J Mol Biol* **1963,** *7* (1), 95-&.
12. Wilmot, C. M.; Thornton, J. M., Beta-Turns and Their Distortions - a Proposed New Nomenclature. *Protein Engineering* **1990,** *3* (6), 479-493.
13. Porter, L. L.; Rose, G. D., Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc Natl Acad Sci U S A* **2011,** *108* (1), 109-13.
14. Shapovalov, M.; Vucetic, S.; Dunbrack, R. L., A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *Plos Comput Biol* **2019,** *15* (3).
15. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.
16. Berman, H., The protein data bank: A retrospective and prospective. *Biophysical Journal* **2000,** *78* (1), 267a-267a.
17. Bonneau, R.; Baker, D., Ab initio protein structure prediction: Progress and prospects. *Annu Rev Bioph Biom* **2001,** *30*, 173-189.
18. Klyne, W.; Prelog, V., Description of Steric Relationships across Single Bonds. *Experientia* **1960,** *16* (12), 521-523.
19. Bateman, A.; Murzin, A. G.; Teichmann, S. A., Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci* **1998,** *7* (6), 1477-80.
20. Buchko, G. W.; Ni, S. S.; Robinson, H.; Welsh, E. A.; Pakrasi, H. B.; Kennedy, M. A., Characterization of two potentially universal turn motifs that shape the repeated five-residues fold - Crystal structure of a lumenal pentapeptide repeat protein from Cyanothece 51142. *Protein Science* **2006,** *15* (11), 2579-2595.
21. Buchko, G. W.; Robinson, H.; Ni, S.; Pakrasi, H. B.; Kennedy, M. A., Cloning, expression, crystallization and preliminary crystallographic analysis of a pentapeptide-repeat protein (Rfr23) from the bacterium Cyanothece 51142. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **2006,** *62* (Pt 12), 1251-1254.
22. Buchko, G. W.; Robinson, H.; Pakrasi, H. B.; Kennedy, M. A., Insights into the structural variation between pentapeptide repeat proteins - Crystal structure of Rfr23 from Cyanothece 51142. *Journal of Structural Biology* **2008,** *162* (1), 184-192.
23. Ni, S. S.; McGookey, M. E.; Tinch, S. L.; Jones, A. N.; Jayaraman, S.; Tong, L.; Kennedy, M. A., The 1.7 A resolution structure of At2g44920, a pentapeptide-repeat

protein in the thylakoid lumen of Arabidopsis thaliana. *Acta Crystallographica Section F-Structural Biology Communications* **2011,** *67*, 1480-1484.
24. Ni, S. S.; Sheldrick, G. M.; Benning, M. M.; Kennedy, M. A., The 2 angstrom resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp strain PCC 7120. *Journal of Structural Biology* **2009,** *165* (1), 47-52.
25. Vetting, M. W.; Hegde, S. S.; Fajardo, J. E.; Fiser, A.; Roderick, S. L.; Takiff, H. E.; Blanchard, J. S., Pentapeptide repeat proteins. *Biochemistry* **2006,** *45* (1), 1-10.
26. Vetting, M. W.; Hegde, S. S.; Hazleton, K. Z.; Blanchard, J. S., Structural characterization of the fusion of two pentapeptide repeat proteins, Np275 and Np276, from Nostoc punctiforme: resurrection of an ancestral protein. *Protein Sci* **2007,** *16* (4), 755-60.
27. Vetting, M. W.; Hegde, S. S.; Zhang, Y.; Blanchard, J. S., Pentapeptide-repeat proteins that act as topoisomerase poison resistance factors have a common dimer interface. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **2011,** *67* (Pt 3), 296-302.
28. Xu, S.; Ni, S.; Kennedy, M. A., NMR Analysis of Amide Hydrogen Exchange Rates in a Pentapeptide-Repeat Protein from A. thaliana. *Biophys J* **2017,** *112* (10), 2075-2088.
29. Joosten, R. P.; te Beek, T. A. H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G., A series of PDB related databases for everyday needs. *Nucleic acids research* **2011,** *39* (Database issue), D411-D419.
30. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-637.
31. Schneider, T. D.; Stephens, R. M., Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **1990,** *18* (20), 6097-100.
32. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome research* **2004,** *14* (6), 1188-90.

# Chapter 6: Conclusion

## 6.1 Biophysical characterization of Alr5209 and Al1298 PRPs from Nostoc sp. PCC 7120

*Nostoc* sp. PCC 7120, one of the most popular model systems for studying many biochemical functions including photosynthesis, nitrogen fixation and so on, is composed of vegetative cells and specialized heterocysts that carry out fixation of atmospheric nitrogen, which differentiate from vegetative cells under the condition of nitrogen starvation.[1] PRPs represent a large superfamily of proteins containing at least eight tandem repeating sequences of five amino acids.[2] In *Nostoc* sp. PCC 7120, 32 PRPs are chromosomally encoded, however their functions still remain unknown.[3] At present, only three PRP structures from *Nostoc* sp. PCC 7120 are solved. One of them is HetL, a PRP which can interfere with the action of PatS in the process of heterocyst formation.[4] Alr5209 and Alr1298 are two PRPs in *Nostoc* sp. PCC 7120 whose biochemical functions remain unknown. Based on analyses of their genetic clusters, we predicted that Alr5209 may be involved in the process of oxidative phosphorylation and Alr1298 may play a role in response to nitrogen starvation and/or heterocyst differentiation.[5-6]

As an initial step towards investigating the function of those two proteins, we solved their structures by x-ray crystallography. According to the structure analysis described in chapter 2, Alr5209 represented the first PRP structure that includes type I β turns in its Rfr fold. Since the influence of type I β turns on Rfr folds had not been analyzed before, we performed a comprehensive structural analysis in comparison with other known PRPs. All structures were grouped based on their composition of β turn types, and the results indicated that PRPs with type I β turns are generally more compact compared to PRPs made up exclusively of type II β turns or PRPs composed of a mixture of type II and type IV β turns. Furthermore, the consensus sequence of PRPs was updated based on the new PRP structures reported in this dissertation, which should be useful for expanding our ability to identify and predict the presence of Rfr folds in new proteins with unknown structures. To determine if proteins containing Rfr folds exhibited any unusual properties of stability, thermal denaturation measurements were performed using CD melting experiments. In chapter 2, it was shown that Alr5209 was slightly more thermally stable compared to the average thermal stability of proteins from a large protein database. In addition, the structural analysis and biophysical characterization described in chapter 3 showed that Alr1298 was the first single-domain PRP structure that contained an elaboration of secondary structural elements at its N-terminus, specifically, a four-helix cluster. Analysis of the electrostatic surface potential of Alr1298 indicated that it had distinct patches of positive and negative charge, which may be important for establishing binding interactions with potential binding partners identified in future functional studies. In chapter 3, due to an interesting packing interaction observed in the crystalline lattice that could have indicated a functional dimer, the solution behavior of Alr1298 was measured by NMR rotational correlation time measurements and computational PISA analysis. Based on the results shown in chapter 3, Alr1298 behaves as a monomer in solution.

## 6.2 New scheme for classifying β turns in protein structures

Different from the regular secondary structure elements of α helices and β sheets, the classification of irregular secondary structural elements, such as tight β turns, are less

well developed.[7-9] Since the first β turn was described by Venkatachalam in 1968, the categories of turn type definitions have been added to and deleted in the next 50 years.[10-13] A widely adopted scheme for classification of β turns was established in 1994 by Hutchinson and Thornton.[13] However, until now, owing to rapid advances in structural biology, schemes for classification of irregular structural elements, such as tight β turns, is undergoing continued development.[14-15] Since classification of β turns remains unsettled and because of continuing discovery of new turn types has led to strong overlap of β turn types as well as existence of "border" β turn types between classical and new turn types,[16] we introduced a new method to eliminate the ambiguous classification of β turns. Based on the description of chapter 4, because classification of β turns depends on the backbone dihedral torsion angles, therefore, we applied well-established organic chemistry stereochemistry conventions in designing the new classification system. All protein structures deposited in PDB before June 2020 with resolution less than 1.5 Å were used for the evaluation of our new schema. To further understand how our new schema can improve the secondary structure analysis of protein structures, we analyzed how our new schema organized β turns based on the following important factors: hydrogen bond occurrence, distances between C atoms of the i and i+3 residues, amino acid preferences and the influence of □ turn. Based on the analysis, we determined that each new turn type may have distinct characteristic distances between C atoms of the i and i+3 residues as well as the preference for specific amino acids at distinct residue positions. Based on the comparison with classical schema, it was clear that the classical schema results in significant ambiguity in classification of protein β turns, and that application of our new classification schema eliminated much of this ambiguity.

### 6.3 Biochemical investigation of PRPs in Nostoc sp. st. PCC 7120

Cyanobacteria, considered to be the first group of microorganisms contributing to the great oxidation event of the earth, have many positive and negative influences on agriculture and the environment.[17-18] As a popular model system of cyanobacteria, *Nostoc* sp. st. PCC 7120 has been studied for decades because it represents the oldest organism to undergo cell differentiation, as the vegetative cell of the filaments undergo patterned cell differentiation to enable the specialized function of fixation of atmospheric nitrogen. At present, a complete understanding of the mechanisms involved in heterocyst differentiation still remains unclear. As detailed the chapter 1, PRPs are abundant in the cyanobacteria, and there are 32 PRPs recognized in the *Nostoc* sp. st. PCC 7120. To date, the function of those PRPs still remains unknown, but based on the genetic analysis, some of them may relate to the process of cell differentiation including Hglk, HetL and Alr1298.[5]

While PRPs are abundant in *Nostoc* sp. PCC 7120, and although the structures and biophysical characteristics of Alr5209 and Alr1298 were determined by our work, the biochemical functions of Alr5209 and Alr1298, as well as the structures, biophysical properties and biochemical functions of the other PRPs in *Nostoc* sp. PCC 7120 still need to be investigated. In carrying out this dissertation research, the other 29 PRPs were also cloned into expression vectors and overexpression in an *E. coli* host was screened for production of soluble protein. These preliminary screening experiments indicated that it should be possible to solve the crystal structures of several additional PRPs from *Nostoc*

sp. PCC 7120 with a little more work. Even though one class PRPs have been shown to have a clear function related to antibiotics resistance, the biochemical function of PRPs in cyanobacteria, especially in the process of nitrogen fixation and cell differentiation, remains unknown. To unravel the structure of those proteins, we can establish the basic idea how those PRPs perform in cyanobacteria and systematically summarize their functional trend in other species.

Our studies indicated that Alr5209 may be involved in the process of oxidative phosphorylation, and Alr1298 may be involved in response to nitrogen starvation as described in chapter 2 and chapter 3, respectively. To further investigate the function of Alr5209 and Alr1298 proteins, studies have been initiated to determine the phenotypes of overexpression mutants of each protein and preliminary experiments have been performed in an attempt to localize each protein in the filament structures both in the presence of abundant nitrogen in the growth media, and in the presence of heterocysts produced during conditions of nitrogen starvation. Experiments are also being planned to characterize the phenotypes of Alr5209 and Alr1298 knock-out strains. We expect that the results of these planned experiments will enhance our understanding of the biological function of those two proteins. Pull-down assays for each protein using constructs to express each protein as a fusion protein containing different affinity tags are planned to attempt to identify potential protein binding partners. Preliminary investigations failed to identify protein binding partners, and therefore it is possible that both proteins bind to small molecules to carry out their function, or that other pull-down affinity tags are necessary for successful pull-down assays. In the future, we also plan to apply proton NMR-based saturation transfer difference (STD) experiments using *Nostoc* sp. PCC 7120 cell extracts for the STD experiments, which is a powerful technology for detecting binding with small molecules. Finally, functional studies will also being planned to characterize phenotypes of Alr5209 and Alr1298 overexpression and knockout strains using a combination of scanning electron microscopy and transmission electron microscopy to observe possible extracellular and intracellular changes that depend on the expression or lack of expression of both proteins that are not visible using simple visible light microscopy or fluorescence microscopy measurements.

### 6.4 Reference

1. Kaneko, T.; Nakamura, Y.; Wolk, C. P.; Kuritz, T.; Sasamoto, S.; Watanabe, A.; Iriguchi, M.; Ishikawa, A.; Kawashima, K.; Kimura, T.; Kishida, Y.; Kohara, M.; Matsumoto, M.; Matsuno, A.; Muraki, A.; Nakazaki, N.; Shimpo, S.; Sugimoto, M.; Takazawa, M.; Yamada, M.; Yasuda, M.; Tabata, S., Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. *DNA research : an international journal for rapid publication of reports on genes and genomes* **2001,** *8* (5), 205-13; 227-53.
2. Bateman, A.; Murzin, A. G.; Teichmann, S. A., Structure and distribution of pentapeptide repeats in bacteria. *Protein Science* **1998,** *7* (6), 1477-1480.
3. Ni, S. S.; Sheldrick, G. M.; Benning, M. M.; Kennedy, M. A., The 2 angstrom resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp strain PCC 7120. *Journal of Structural Biology* **2009,** *165* (1), 47-52.

4. Liu, D.; Golden, J. W., hetL overexpression stimulates heterocyst formation in Anabaena sp. strain PCC 7120. *J Bacteriol* **2002,** *184* (24), 6873-81.
5. Zhang, R.; Ni, S.; Kennedy, M. A., Crystal structure of Alr1298, a pentapeptide repeat protein from the cyanobacterium Nostoc sp. PCC 7120, determined at 2.1 Å resolution. *Proteins: Structure, Function, and Bioinformatics n/a* (n/a).
6. Zhang, R.; Ni, S.; Kennedy, M. A., Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7 angström resolution. *Journal of Structural Biology: X* **2019,** *3*, 100010.
7. Chou, K.-C., Prediction of tight turns and their types in proteins. *Analytical biochemistry* **2000,** *286* (1), 1-16.
8. Chou, K. C.; Blinn, J. R., Classification and prediction of beta-turn types. *Journal of protein chemistry* **1997,** *16* (6), 575-95.
9. Chou, P. Y.; Fasman, G. D., Conservation of chain reversal regions in proteins. *Biophysical journal* **1979,** *26* (3), 385-99.
10. Venkatachalam, C. M., Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **1968,** *6* (10), 1425-36.
11. Lewis, P. N.; Momany, F. A.; Scheraga, H. A., Chain reversals in proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1973,** *303* (2), 211-229.
12. Richardson, J. S., The Anatomy and Taxonomy of Protein Structure. In *Advances in Protein Chemistry*, Anfinsen, C. B.; Edsall, J. T.; Richards, F. M., Eds. Academic Press: 1981; Vol. 34, pp 167-339.
13. Hutchinson, E. G.; Thornton, J. M., A revised set of potentials for beta-turn formation in proteins. *Protein science : a publication of the Protein Society* **1994,** *3* (12), 2207-16.
14. de Brevern, A. G., Extension of the classical classification of β-turns. *Scientific Reports* **2016,** *6* (1), 33191.
15. Shapovalov, M.; Vucetic, S.; Dunbrack, R. L., Jr., A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLOS Computational Biology* **2019,** *15* (3), e1006844.
16. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.
17. Whitton, B. A., *Ecology of cyanobacteria II: their diversity in space and time*. Springer Science & Business Media: 2012.
18. Kerbrat, A. S.; Amzil, Z.; Pawlowiez, R.; Golubic, S.; Sibat, M.; Darius, H. T.; Chinain, M.; Laurent, D., First evidence of palytoxin and 42-hydroxy-palytoxin in the marine cyanobacterium Trichodesmium. *Mar Drugs* **2011,** *9* (4), 543-560.
19. Nicolaisen, K.; Hahn, A.; Schleiff, E., The cell wall in heterocyst formation by Anabaena sp. PCC 7120. *Journal of basic microbiology* **2009,** *49* (1), 5-24.