Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale [Hématologie, Oncologie et Biothérapies et ED 561]

DSIMB / INSERM UMR S_1134

Dynamics of protein structures and its impact on local structural behaviors

Par Tarun Jairaj NARWANI

Thèse de doctorat de Biotechnologies et Biothérapies

Dirigée par Alexandre G. de Brevern

Présentée et soutenue publiquement à Paris le 27 Juin 2018

Président de jury : Rodriguez-Lima, Fernando / PR / Université Paris Diderot

Rapporteurs : Callebaut, Isabelle / HDR / CNRS, Université Pierre et Marie Curie

Rapporteurs : Leclerc, Fabrice / HDR / CNRS, Université Paris-Sud, Saclay

Examinateurs : André, Isabelle / HDR / CNRS, INSA, Toulouse

Examinateurs : Offmann, Bernard / PR / Université de Nantes

Directeur de thèse : de Brevern, Alexandre G. / HDR / INSERM, Université Paris Diderot

Résumé en français :

Dynamique des structures protéiques et son impact sur les comportements structuraux locaux

Les structures protéiques sont de nature hautement dynamique contrairement à leur représentation dans les structures cristallines. Une composante majeure de la dynamique structurelle est la flexibilité des protéines inhérentes. L'objectif principal de cette thèse est de comprendre le rôle de la dynamique inhérente dans les structures protéiques et leur propagation. La flexibilité des protéines est analysée à différents niveaux de complexité structurelle, du niveau d'organisation primaire au niveau quaternaire. Chacun des cinq premiers chapitres traite un niveau différent d'organisation structurelle locale avec le premier chapitre traitant des structures secondaires classiques tandis que le second analyse la même chose en utilisant un alphabet structurel - les blocs protéiques. Le troisième chapitre se concentre sur l'impact d'événements physiologiques spéciaux comme les modifications post-traductionnelles et le désordre sur les transitions d'ordre sur la flexibilité des protéines. Ces trois chapitres indiquent une mise en œuvre dépendante du contexte de la flexibilité structurelle dans leur environnement local. Dans les chapitres suivants, des structures plus complexes sont prises en compte. Le chapitre 4 traite de l'intégrine $\alpha_{IIb}\beta_3$ impliquée dans des troubles génétiques rares. L'impact des mutations pathologiques sur la flexibilité locale est étudié dans deux domaines rigides de l'intégrine $\alpha_{IIb}\beta_3$ ectodomaine. La flexibilité inhérente dans ces domaines est montrée pour moduler l'impact des mutations vers les boucles. Le chapitre 5 traite de la modélisation structurelle et de la dynamique d'une structure protéique plus complexe du récepteur des chimiokines des antigènes du groupe Duffy incorporé dans un système de membrane mimétique érythrocytaire. Le modèle est soutenu par l'analyse phylogénétique la plus complète sur les récepteurs de chimiokines jusqu'à ce jour, comme expliqué dans le dernier chapitre de la thèse.

Mots clés :

Flexibilité de la structure des protéines, allostérie, Blocs Protéiques, fragments de double personnalité, modification post-translationnelle, Intégrine $\alpha_{IIb}\beta_3$, Thrombasthénie de Glanzmann, thrombocytopénie allo-immune fœtale / néonatale, paludisme à *Plasmodium vivax*, récepteurs des chimiokines des antigènes du groupe Duffy, phylogénie moléculaire.

Abstract in English:

Dynamics of protein structures and its impact on local structural behaviors

Protein structures are highly dynamic in nature contrary to their depiction in crystal structures. A major component of structural dynamics is the inherent protein flexibility. The prime objective of this thesis is to understand the role of the inherent dynamics in protein structures and its propagation. Protein flexibility is analyzed at various levels of structural complexity, from primary to quaternary levels of organization. Each of the first five chapters' deal with a different level of local structural organization with first chapter dealing with classical secondary structures while the second one analysis the same using a structural alphabet - Protein Blocks. The third chapter focuses on the impact of special physiological events like post-translational modifications and disorder to order transitions on protein flexibility. These three chapters indicate towards a context dependent implementation of structural flexibility in their local environment. In subsequent chapters, more complex structures are taken under investigation. Chapter 4 deals with integrin αIIbβ3 that is involved in rare genetic disorders. Impact of the pathological mutations on the local flexibility is studied in two rigid domains of integrin αIIbβ3 ectodomain. Inherent flexibility in these domains is shown to modulate the impact of mutations towards the loops. Chapter 5 deals with the structural modelling and dynamics of a more complex protein structure of Duffy Antigen Chemokine Receptor embedded in an erythrocyte mimic membrane system. The model is supported by the most comprehensive phylogenetic analysis on chemokine receptors till date as explained in the last chapter of the thesis.

Keywords:
Protein structural flexibility, Allostery, Protein Blocks, Dual personality fragments, Post-translational modifications, Integrin $\alpha_{IIb}\beta_3$, Glanzmann Thrombasthenia, Fetal/neonatal Alloimmune Thrombocytopenia, *Plasmodium vivax* malaria, Duffy Antigen Chemokine Receptors, Molecular phylogenetics

ACKNOWLEDGEMENTS:

Gratitude of a graduate

Similar to proteins that are composed of amino acids, a person's character is developed by the blocks of interactions with different people throughout life. There are many such people in my life that I need to thank for I can be eligible to compile a thesis manuscript.

I would like to thank my thesis supervisor, advisor, mentor, and a great friend, Dr. Alexandre G. de Brevern (Alex) for being the backbone of my thesis work. He provided a conductive and motivating environment for research and innovation. I thank him for our discussions in the central room of the lab that were quintessential to our project developments. I also thank him for putting trust in me with varied research themes that might felt like a challenge at first but Alex's constant motivation and energy translated it to an adventurous joy ride. I am indebted towards him and his family, Helene, Snoopy and, aunt Roselyne, for providing the love and affection that acted as an immune system towards my home sickness. I will cherish our times in the lab and outsides, especially our prefecture and CROUS fun times.

I will extend my gratitude towards the INSERM unit head Prof. Yves Collin (Yves) and our team director, Prof. Catherine Etchebest (Cathy) for being extensively student friendly in administrative matters. Recommendation letters from Yves were very instrumental in getting my VISA to France. While, my interactions with Cathy have been a great learning experience. I take this place to thank Cathy for our brainstorming sessions on varied scientific questions that have always enhanced my knowledge. I would like to thank Prof. Olivier Bertand for fruitful discussions on DARC and his wife for a wonderful dinner at their place. These memories will always own a sweet spot in my heart.

I would like to thank my colleagues from Team 1, 3, and 4, Claude, Jean-Phillippe, Dominique, Arnauld, Jean-Lopez, Mariano, Wasim, Sophie, Melanie, Sebastian, Maria, Slm, Saby, Zaineb, Marilou, Megane, Irene, Nelly, for providing a friendly environment that helped me adjust easily in a new workplace. My lab mates for being a great support during my tenure and especially during my thesis writing days. I thank Jean-Christophe (J-C) for being my French culture reference and a mentor in need. I thank Sofiane for our Tacos treats and making me learn football, in theory. Along with him, I thank Akhila and Rajas for taking good care of me during my thesis writing days. They have been of great help to my French speaking disability, alongside master interns, Natacha, Sarah, Madeliene, Katia, and Soubika who handled my prefecture, bank, CAF, doctor, and velib calls. I am indebted towards their open heartedness and affection. I will like to thank Charlotte, Tatiana, and Marja for organizing the evening plans for drinks that were a great stress buster and fun times. I especially thank Tatiana for showing tolerance towards my power naps during the thesis writing days.

It will be unjustified if I claim that my thesis would have been completed without my collaborators and the hard-working master interns. I thank all my scientific collaborators starting with Dr. Vincent Jallu, Dr. Sophie Abby, Prof. N. Srinivasan, Dr. Salwa Karboune, Dr. Pierrick Craveur, Dr. Joseph Rebehmed, and Dr. Agnel Praveen Joseph. Their contributions to my projects and research is quintessential. I would especially thank Dr. Sophie Abby for providing a new perspective to DARC (Chapter 6) by mentoring me on molecular phylogenetics. I will extend my sincere appreciations to four master interns that I got the opportunity to mentor, M. Matthieu Gouget, Madam Sali Anies, Madam Sarah Kaddah, and Madam Soubika Bisoo. Their work and efforts have been largely instrumental in the Integrins project.

My life has been quite a joy-ride and a major portion of the happiness is contributed by my dear friends, Udit, Gauravs, Abhishek, Harjot, Chetan, and Ishwar. I might have missed them a lot during my stay in Paris however, the quotient of happiness and craziness is fulfilled by Aurore, Matthieu (Chicken parents), Natacha (my lady), Hubert (stack.overflow), Yassine (Yes Sin), and Nicolas (Nikola-San). Aurore and Matthieu provided the care and affection that helped me merge easily with a new culture while I thank Hubert for being there professionally and personally to help me being a better person. With lots of affection I thank Natacha for her amazing cookies along with Yassine for the fun and craziness drive that we share in lab and ‘elsewheres’. I will duly cherish the fandom of Christopher nolan’s that I shared with Nicolas and thank him for introducing me to funny Japanese games.

I extend my sincere gratitude towards my mentors, Dr. Nandita Bacchhawat, Prof. N. Srinivasan, Dr. Srikrishna Subramanian, and Prof. P.P Mathur for their encouragements and careful honing of my skills that forms the very foundation of my research. I sincerely thank my parents for being an open-minded couple and letting me pursue research. I especially thank my father for always fighting for me and defeating all my defeats. The constant support, enlightenment, care, and love that I have received from Miss Himani Tandon is much beyond a thank-you. However, I will still thank her along with my parents for being the most instrumental people in my life.

Towards the end, I would like to thank some people with superhuman capabilities. I extend my gratitude to the security personnel and housekeeping staff of INTS, M. Gabriel, M. Kareem, M. Habib, M. Aziz, M. Kareem, M. Jamal, M. Nabeel, and Madame Sali and Gasama for being kind towards me. I especially thank Gabriel, Kareem, and Habib for taking good care of my appetite during their security checks at middle of the night.

My research, writing skills, personality, energy, and character is a blessing of all these people and therefore Dr. Tarun Jairaj Narwani does not exist without his experiences with these people.

Sincerely,

Narwani

Tarun Jairaj Narwani..

Table of Contents

Index for Figures

Index for Tables

INTRODUCTION

I.1 *Proteins*

Protein are the mediators of information content stored in the DNA and perform myriad of functions inside the cell. From being signalling molecules [1] to enzymes [2] to cellular receptors [3] to forming cytoskeleton [4], they are involved in regulating the cellular function at every step. Genetic information stored in the DNA is transcribed by RNA polymerase into an mRNA transcript which is then used by ribosomes to translate a protein biopolymer. Such sequential transfer of information from DNA to RNA to Protein is known as the central dogma of molecular biology [5]. For a protein to mediate its function, it is required to fold correctly and therefore the relation between protein structure and function is quintessential [6].

I.1.1 *Protein sequence*

Proteins are made up of amino-acids attached to one another via peptide bonds [7]. Amino acids are the organic molecules containing amine (-NH2) and carboxyl (-COOH) groups as well as an R (side-chain) group which is specific to each amino acid (Fig I.1a). A presence of the amine and carboxyl group states that between pH 2.2 - 9.4, the amino acids contain a negative carboxylate as well as positive ammonium group, hence existing as Zwitterions. Based on the R group, the amino acids can be classified into polar/nonpolar or aliphatic/aromatic or charge/uncharged residues (Fig I.1b). Generally, the non-polar residues form the core of the protein structures while the polar residues are present on the surface to carry out important functions such as catalysis [8].

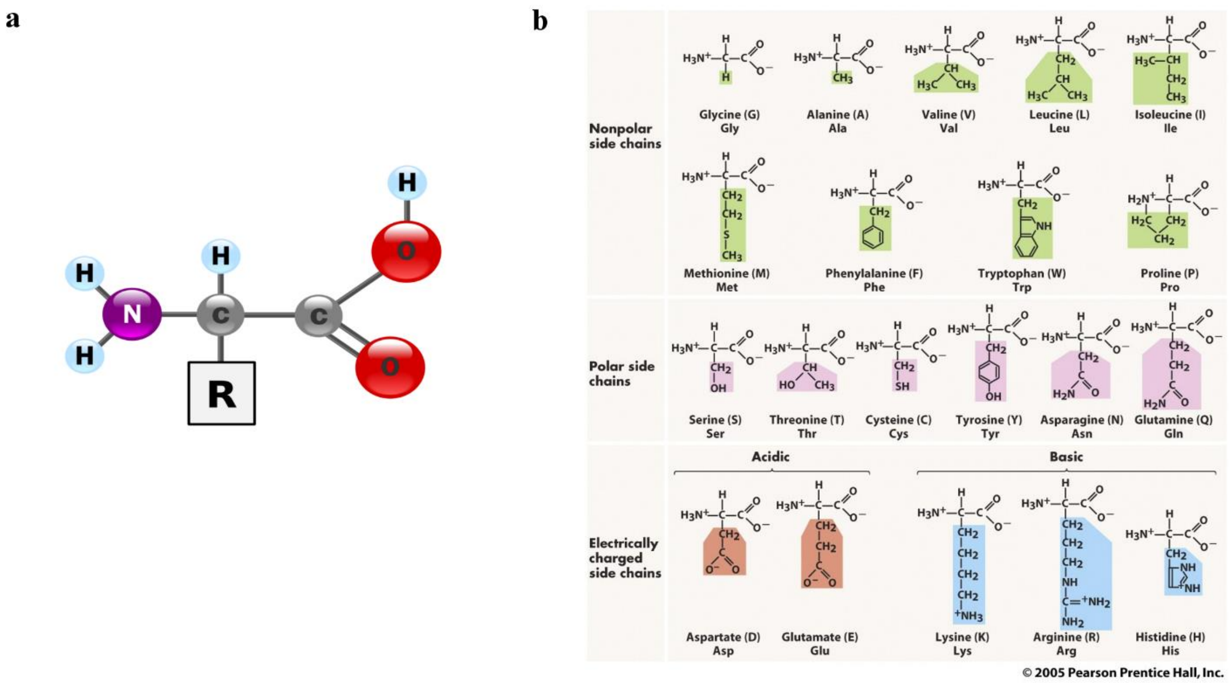


Figure I.1. *Amino acid structure in* ***(a)*** *shows the –NH_2, R and –COOH groups and* ***(b)*** *shows the classification of amino acid based on the nature of R group.*

+Source internet, www.stanford.edu/tutorials/biochem/

I.1.2 *Structure of proteins*

The structure of a protein is dictated by its sequence. Anfinsen, in early 70s conducted a very elegant experiment to prove this characteristic point [9]. The paradigm of sequence-structure-function relationship holds true for majority of the proteins. It states that if two proteins have similar sequences, then it is highly likely that they have similar structures which in turn implies that they should have similar function [10,11]. This underlying principle has been extensively used to predict function for a given protein [12,13]. In order to classify protein structures, it is important to understand the recurring patterns in their 3D arrangement. Ramachandran *et al* used a model system of two-linked peptide units to identify those conformations which displayed no short-contact between any two atoms [14]. The conformations were selected based on the accessible rotations around N-Cα and C-Cα single bonds called phi (Φ) and psi (Ψ) dihedral angles respectively (Fig I.2).

The values obtained for each accessible Φ and Ψ were plotted as what is famously known today as Ramachandran map. It was observed that the polypeptide backbone can take up certain conformations which are allowed according to the Ramachandran map confirming the occurrence of regular structures in proteins.

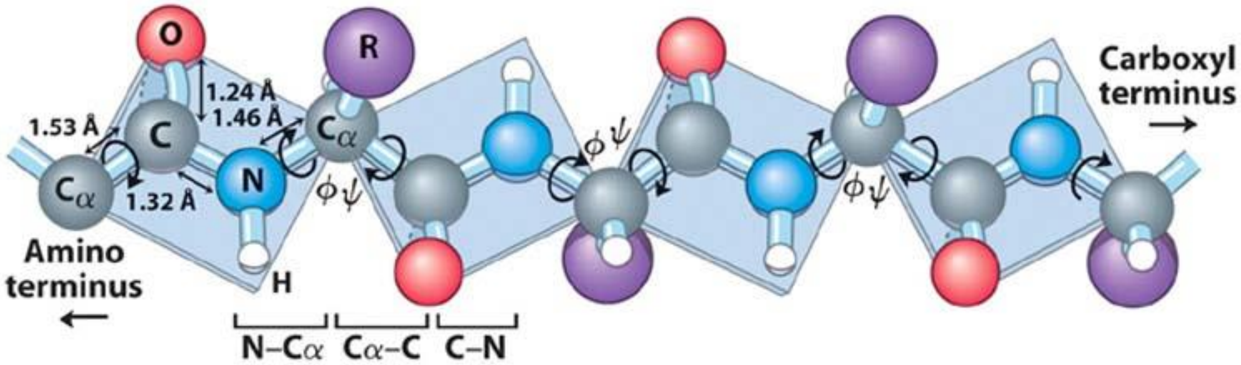


Figure I.2. *Polypeptide chain in fully extended conformation showing the Φ and Ψ dihedral angles. The different bond lengths are also shown.*

++Source internet: Nelson & Cox, Lehninger Principles of Biochemistry, 4th ed. (2004)

I.2 *Protein structure organisation*

Protein structures can be organised into four levels of structural hierarchy viz. primary, secondary, tertiary and quaternary (Fig I.3).

I.2.1 *Primary structure*

An organisation of amino-acids in a linear fashion, next to one another, depicts the primary structure of the protein (Fig I.3). Sequential arrangement of amino acids in a polypeptide chain refers to its primary structure. A protein generally adopts this conformation during the translation process, when the peptide is being polymerized from the 'P' site of the ribosome.

I.2.2 *Secondary structure*

The second level of structural organisation in a protein is defined by secondary structures and is governed by a highly regular, local sub-structural arrangement of polypeptide backbone. Traditionally, these include α-helices and β-sheets (Fig I.3) which are identified by a fixed intra

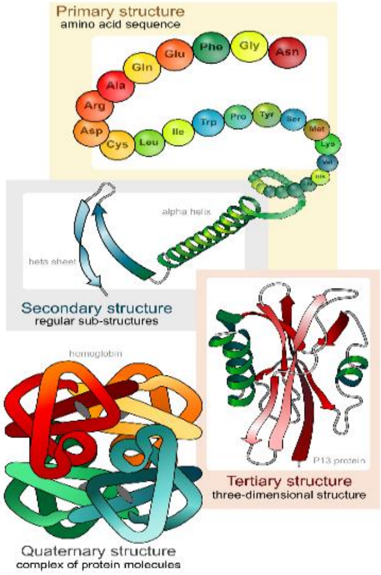


Figure I.3: *Levels of protein structure organisation; Primary, Secondary, Tertiary and Quaternary structures.*

+Source internet, www.ucsf:com/lectures/molecularbiology/proteins

and inter-chain hydrogen-bonding pattern. These secondary structures have a fixed geometry defined by their backbone dihedral angles (Φ and Ψ) and are known to occupy definite location in the Ramachandran map. Example, α-helices occupy the Φ, Ψ position -57, -47 while β-sheets typically occupy position -140, 130. Apart from α-helices, the most prominent helices occurring in the protein, 3_{10} and π-helices are other types of helices that occur in proteins. These types of 3 helices are distinguished by the H-bonding pattern between them. The intra-chain H-bond is formed between *i* and *i+4* residue in the α-helix, between *i* and *i+3* residue in the 3_{10}-helix and *i* and *i+5* residue in the π-helix [15,16]. The Φ and Ψ values for these helices are given in Table I.1.

Table I.1. *Characteristics of different types of helices found in proteins.*

Helix type	Phi (Φ)	Psi (Ψ)	Description
α-helix (R)	-57	-47	Right handed α-helix
3_{10}-helix	-49	-26	Right-handed 3_{10}-helix
π-helix	-57	-80	Right-handed π-helix

Apart from these, proteins can have also kinks in helices which is often introduced due to a proline residue in the middle of α-helix [17]. Separately, polyproline helices and collagen triple helix enjoy status of being special cases of secondary structure elements which occur at different Φ, Ψ values than the regular helices and sheets [18,19]. β-sheets are formed by arrangement of β-strands in register with each other. They are also classified further into antiparallel and parallel β-sheets depending upon the direction of the strands in register which in turn decides the geometry of H-bond. In an antiparallel arrangement, the consecutive β-strands are in opposite direction to each other such that the N-terminus of one strand is adjacent to the C-terminus of the next (Fig I.4a). This results in a planar H-bond arrangement which deems most suitable for the stability of the β-sheets. The backbone dihedral angles Φ and Ψ for antiparallel sheets are –140°, 135°. On the other hand, parallel β-sheets have the two strands running in same direction making the inter-strand H-bond slightly out-of-plane (Fig I.4b), hence lowering the stability of parallel sheets when compared to antiparallel. The dihedral angles Φ and Ψ are –120°, 115° for the parallel sheets. Strands are rarely long, maximum 15 residues in length and most β-sheets contain less than 6 strands. Side chains from adjacent residues of a strand in a β-sheet are found on opposite sides of the sheet and do not interact with one another. Therefore, like α-helices, β-sheets have the potential for amphiphilicity with one face being polar and the other being non-polar. It has also been noted that parallel sheets are generally buried inside while antiparallel sheets have one side exposed to the solution [20].

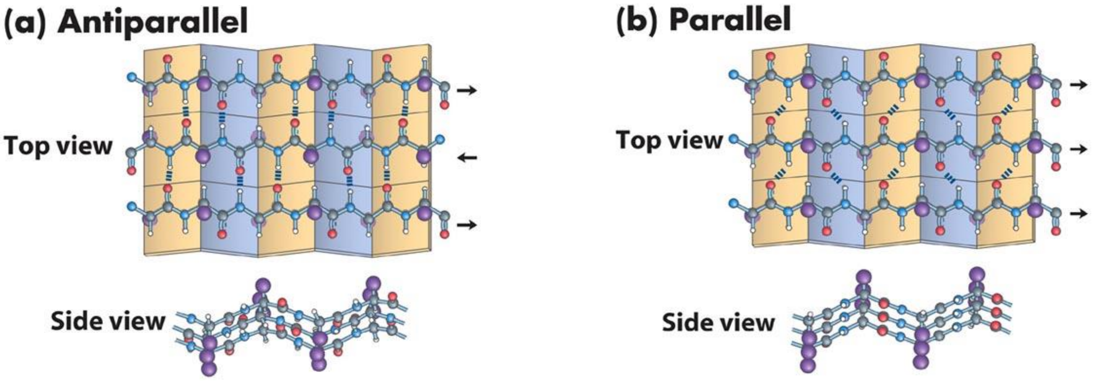


Figure I.4. *The top and side views of **(a)** antiparallel β-sheet **(b)** parallel β-sheet.*

Like the kinks in helices, β-sheets are known to have β-bulges which are caused by interruption in the hydrogen bonding [21]. A β-bulge is a region between two consecutive β-type hydrogen bonds which includes two residues (positions 1 and 2) on one strand opposite a single residue (position x) on the other strand. Two common types of β –bulges are known viz. the "classical" β-bulge and the "G1" β-bulge. The main functional attribute of β-bulges is to accommodate for any single residue insertion or deletion within a β-structure [21].

Besides helices and sheets exist a third kind of structural element called turns which help in reversing the direction of a polypeptide chain. Turns are mainly found on the protein surface and hence contain polar or charged residues and were first identified in protein structures by Venkatachalam [22]. Turns are further classified into α-, β- and γ-turns, out of which β-turns are the most common ones consisting of a sequence of four residues They were defined as linked by a 1-4 (3_{10}-type) hydrogen bond between the -C=O of the first residue and the -NH of the fourth residue [20].

α-helices and β-sheets are composed of repetetive units of particular hydrogen bonding patterns as shown in Table I.1 and Fig. I.4. These repetetive units are classified based on their hydrogen bonding and length. A β-bridge is a singelton hydrogen bond observed in isolation with a length of 3 to 4 residues. When multiple bridges exist together, they form a β-sheet [21]. If such hydrogen bonding is missing but the local curvature around the Cα atoms has an angle of 70, it is classified as a bend. A bend is the only secondary structure element whose principle identification is not done by hydrogen bonding pattern [21]. The structural examples of bend and bridges can be seen in Fig. 1.5

I.2.3 *Super secondary structures*

The secondary structural elements can come together in more than one ways to form some higher order structures. Their structural complexity is smaller than the tertiary structures and denote topological arrangement of helices, sheets and turns. These are called as super-secondary structures and usually occur as small yet functionally important structural motifs. These motifs are generally involved in either biological or structural functions. Some examples include the *helix-turn-helix* motif which is known to bind DNA [23], the EF-hand motif known to bind Ca^{2+} [24] and β-hairpin motif which plays a structural role and connects two antiparallel β-strands [25]. Figure I.3 shows a β-hairpin motif in the secondary structure section.

I.2.4 *Tertiary and quaternary level of protein folding*

Tertiary structure denotes the 3D arrangement of the secondary structure elements to form a well-folded functional polypeptide (Fig I.3). Tertiary structure is generally defined for a single polypeptide chain where the interior of the folded protein is known as core and is formed by hydrophobic amino acids. The concept of protein domains can be defined at this level. Protein domains are the compact globular modules that are capable of folding and functioning independently of the rest of the protein. Within a protein, different domains can be identified e.g. ligand-binding domain, DNA-binding domain etc.

In many cases, two or more tertiary structures join together to constitute the functional state of a protein. Such organisation of two or more tertiary structures is called quaternary structure of a protein (Fig I.3). Each polypeptide chain in the quaternary structure is termed as a subunit. In other words, quaternary structure can be a homomer, formed from the self-assembly of repeated copies of a single subunit. On the other hand, heteromeric complexes are composed of multiple distinct protein subunits, usually encoded by different genes [26]. Classical example of proteins with similar structure, but one fully functional in tertiary state and the other in quaternary is Myoglobin and Haemoglobin where the former is functional as a single chain while the latter requires association of 4 chains to form a functional molecule [27].

I.3 *Structural and functional classification of proteins*

Proteins can be classified into various groups depending upon sequence or structural similarity. The classification of proteins becomes important to propose function for a novel protein. One of the ways to classify proteins is to group them into families and superfamilies. A protein family consists of a set of proteins that are evolutionarily related by virtue of similarities in sequence or structure and function. The families can be arranged into hierarchy where the proteins having a common ancestor are grouped into smaller subgroups, indicating more closely related members, called subfamilies. A superfamily consists of different families of proteins where the members within the superfamily are distantly related [28]. A schema of classifying proteins into families is shown in Figure I.5.

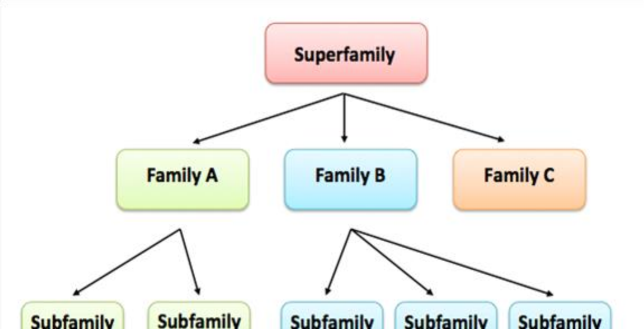


Figure I.5: *A hypothetical protein family classification showing Family, Superfamily and Subfamily level hierarchy.*

+Source internet. www.ebi.ac.uk

Proteins are grouped into families depending upon similarities in their functional regions, commonly termed as domain. Two classification schemes, one based on similarities between domain sequence and the other based on similarities between domain structures are available. These are Pfam (http://pfam.xfam.org) domain definitions and the SCOP (http://scop.berkeley.edu/) domain definitions.

According to SCOP [29], the protein structures can be classified into following categories depending upon similarities between protein domains:

a) *Domain:* A part of a protein. For simple proteins, it can be the entire protein

b) *Species:* The domains in "protein domains" are grouped according to species name

c) *Protein domain:* Grouping together similar sequences having essentially the same functions

d) *Family:* It contain proteins with similar sequences signifying homology but typically distinct functions

e) *Superfamily:* Proteins in a family are grouped together which have at least a distant common ancestor

f) *Fold:* It groups structurally similar superfamilies

g) *Class:* Grouping the protein structures mainly on secondary structure content and organization.

Four classes are defined in SCOP - domains containing all α-helices (a.*), domains containing all β-sheets (b.*), domains containing α/β (c.*) and domains containing α+β (mainly segregated, represented as superfamily d.*). SCOP classification also goes beyond 'c.*' uptil 'g.*' containing different categories of transmembrane proteins. *denotes the sub-family structure after the superfamily a, b, c, d, e, f, and g.

I.4 *Protein types based on cellular environment*

Protein structures have to correctly fold to perform correct function. Depending on the nature of cellular environment and functional requirements proteins can be either globular, fibrous or membrane proteins.

I.4.1 *Globular proteins*

Globular proteins are those polypeptide chains that fold into a compact shape. They are the most common protein types. These proteins have a well-defined hydrophobic core such that the apolar residues face towards protein interior while the polar residues face outwards. Functionally these proteins can be enzymes, regulatory proteins, messengers, and transporters etc. Many different folds are associated with globular proteins.

I.4.2 *Fibrous proteins*

On the other hand, fibrous proteins are generally elongated and are mostly involved in cellular support and structural functions. They are more stable than globular proteins. Some very well characterized fibrous proteins include collagen, actin, myosin, and keratin etc.

I.4.3 *Membrane proteins*

As the name suggests, membrane proteins are the ones that interact with phospholipid membrane. They can be integral membrane proteins which can be permanently attached to the membrane or peripheral membrane proteins which are temporarily attached to the lipid bilayer

[30]. The integral membrane proteins are transmembrane proteins which span across the membranes. It can either be single pass (passing the membrane only once) or multi-pass membrane (passing the membrane more than once) protein. They function mostly as membrane receptors, transport channels, Ion channels, and cell-adhesion and aggregation molecules. Membrane proteins form a separate class in SCOP (e.* to g.*).

An example of globular protein, membrane protein and fibrous protein each is shown in Figure I.6.

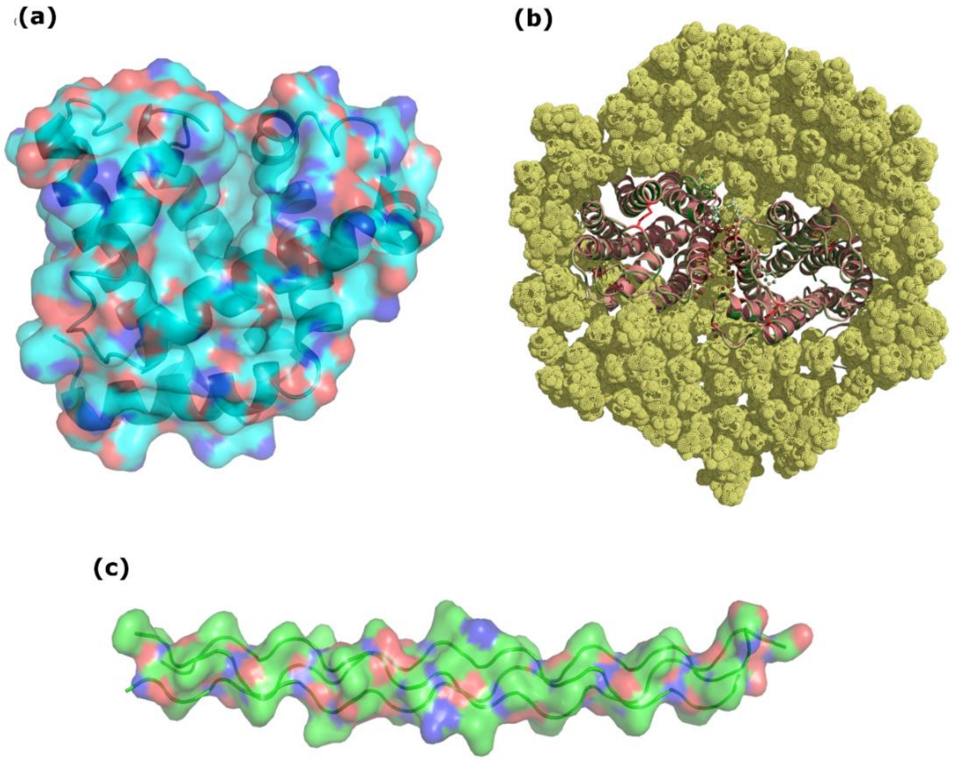


Figure I.6: *Examples of* ***(a)*** *globular protein – myoglobin (PBDid- 1vxc)* ***(b)*** *membrane protein – DARC dimer embedded in membrane (modelled in chapter 5)* ***(c)*** *fibrous protein – collagen triple helix (PBDid- 1bkv)*

+Source internet, www.rcsb.org/

I.5 *Experimental determination of protein structure*

Coordinates for 139717 structures have been deposited in the protein data bank (PDB) as of April 29, 2018. Three main experimental methods exist to determine the structure of a protein. These are X-ray crystallography, nuclear magnetic resonance (NMR) Spectroscopy and cryo-electron microscopy (cryoEM). Out of these, X-ray crystallography has been the method of choice for solving majority of the structures. Though, in the recent years, cryoEM is becoming a popular method to solve the protein structures, especially macromolecular assemblies.

I.5.1 *X-ray Crystallography*

X-ray crystallography is a technique used to determine the atomic structure of a protein. The technique itself is more than a 100 years old but became a popular method of choice for protein structure determination since 1950s when Sir John Kendrew first solved the structure of sperm whale myoglobin [31]. Since then, 123230 structures solved by crystallography have been deposited in PDB so far. Following structure determination of the lysozyme from bacteriophage T4 (T4 lysozyme) [32], it became a prototype for the study of protein folding and thermodynamics [33].

Briefly, crystallography requires protein crystallisation – a process of forming protein crystals. A unit cell is the crystal repeating unit which defines the smallest group of atoms which has the overall symmetry of a crystal, and from which the entire lattice can be built up by repetition in three dimensions. The protein atoms systematically arrange themselves in three dimension in a unit cell. The protein crystal is then exposed to X-rays causing diffraction according to the Bragg's law; which states that a constructive interference happens when the condition ‘$n\lambda=2d\sin\theta$’ is satisfied, where d is the distance between two planes, θ is the angle of incidence and λ is the wavelength of X-ray beam (Fig I.7).

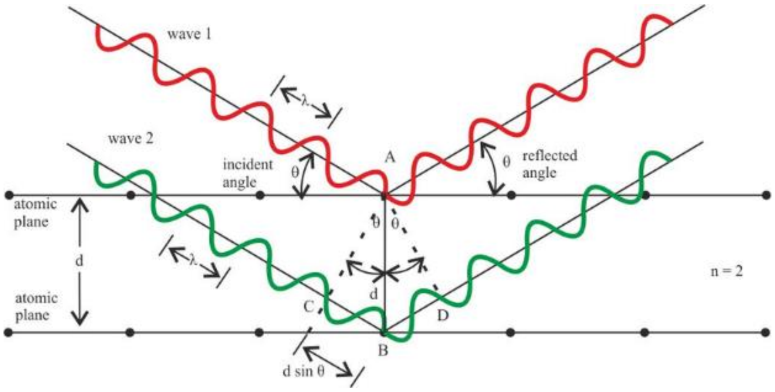


Figure I.7. *Pictorial representation of Bragg's law.*

$^{+}$Source internet, www.wikipedia.org/braggslaw/

A three-dimensional picture of the electron density within the crystal is produced by measuring the angles and intensities of these diffracted beams. The main challenge in generating the electron density from the diffraction pattern is deciphering the phases which are lost while collecting the diffraction data. This is the notorious *phase problem* in crystallography, which is the problem of loss of information concerning the phase that can occur when making a physical measurement [34,35]. The phases in crystallography can be obtained by various methods such as molecular replacement (MR), multi-wavelength anomalous diffraction (MAD), multiple isomorphous replacement (MIR) etc. Once the electron density is obtained, the mean positions of the atoms in the crystal as well as the extent of disorder in the structure can be determined. A flow chart for X-ray crystallography is shown in Figure I.8.

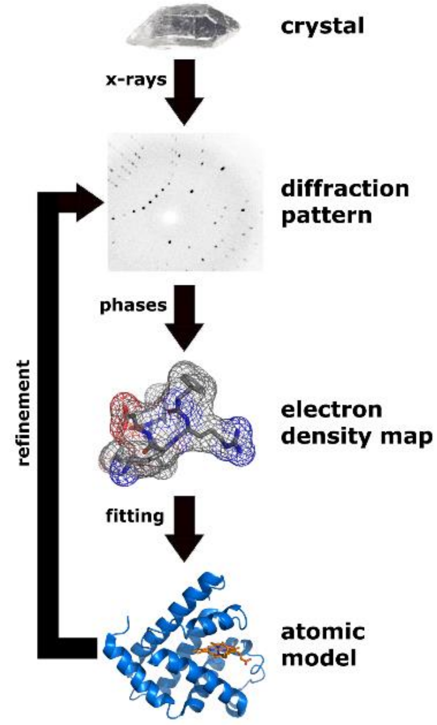


Figure I.8. *Flowchart showing workflow of X-ray crystallography. A crystal is bombarded with X-rays to obtain a diffraction pattern. The pattern is then used to generate an electron density map into which atoms are fitted, sometimes based on best guess.*

+Source internet, www.stanford.edu/tutorials/biophysics/

I.5.1.1 *Crystal packing defects*

Besides, the uncertainty value induced by phase problem there lies another concern with the X-ray technique- *The packing of the structure in the crystal*. [36]. Although X-ray crystallography gives detailed atomic information about the structure, all the interactions observed in the packed crystals may not be biologically relevant. Many of them may be an artifact of crystal packing and may not be observed in solution. Differentiating the true interactions from such non-specific interaction may become a daunting task [37,38]. An analysis of general interface properties has revealed some features to distinguish specific vs non-specific interactions within crystals [39,40]. These properties include interface area, composition of the interface, spatial distribution of the interface residues, secondary structure, core interface conservation and the space group to which they belong. A recent study has shown that many of these properties are indistinguishable for the specific and non-specific interactions [41] and hence one has to be cautious while analysing protein-protein contacts obtained from the crystal structures.

I.5.2 *Nuclear Magnetic Resonance- NMR*

NMR is the second most common method to determine protein structures. Most NMR structures consist of a single type of polypeptide chain and a majority of the structures solved using NMR are monomers [42]. This is because smaller proteins are easily characterized using NMR than the larger proteins and hence the proteins that tend to exist as huge oligomers are not amenable to structure determination using NMR. It has been shown that since 2005, the number of structures determined using NMR in PDB has sharply decreased showing a fall in the popularity of the method [42]. Even though NMR spectroscopy is usually limited to proteins smaller than 35 kDa, it is often the only method to study the conformational heterogeneity and intrinsically disordered nature of proteins.

NMR exploits the quantum mechanical properties of the central core ("nucleus") of the atom. These properties depend on the local environment of the molecules and their measurement provides a map of how the atoms are chemically linked, how close are they in space, and how rapidly they move with respect to each other. Principle behind obtaining NMR spectra is that each distinct nucleus in a protein experiences a distinct electronic environment and thus has a distinct chemical shift by which it can be recognized. A resonance assignment is obtained for the protein to find out the chemical shift corresponding to each atom. To perform structure calculations, a number of experimentally determined restraints are generated like distance restraints and angle restraints. These restraints are used as an input to generate multiple structures satisfying these restraints. Hence, NMR generates an ensemble of structures while X-ray crystallography provides one structure which generally is a space and time-averaged structural snapshot.

I.5.3 *Cryo-electron Microscopy*

Though X-ray crystallography is considered as the gold standard for providing atomic resolution structures, it suffers from the drawback of providing a static snapshot which may be far from the physiological structure. Also, many proteins resist crystal formation and a lot of time and effort has to be invested to solve the structure using X-ray crystallography. NMR on the other hand, though being capable of elucidating dynamics information is limited by the size considerations. Hence, cryoEM can provide solutions to these limitations such that it can be used for bigger protein complexes and can image complexes in their physiological environment [43]. Although the use of cryoEM technique has been limited to medium to low (5-15 Å) resolution range yet structures with resolution better that 3 Å are getting published thus making the technique tractable [44]. CryoEM is becoming the most sought after technique to extract structural information about the macromolecular complexes not amenable to either X-ray or NMR. After procuring a density map, the most important task is to obtain a high confident pseudo-atomic model for the same. The structures to be fitted into the electron density map can either be the crystal structures of a subcomponent or can be a homology modeled structure. If the density map resolution is better than 4 Å, de-novo modeling can be used to calculate the pseudo-atomic model.

CryoEM is a type of Transmission Electron Microscopy (TEM), in which the sample is studied at cryogenic temperatures. The information obtained is invaluable in understanding the macromolecular assembly at physiological conditions. CryoEM techniques can be classified into three types: a) Electron Crystallography, b) Single particle analysis, and c) Cryo-electron tomography. Out of these single-particle analysis or Single particle cryoEM is emerging as a technique of choice to determine 3D structure of proteins with an increasingly advancing electron beams, detectors and ability to analyse isolated complexes (“Single particles”) under native conditions [45]. Its emergence can be judged from the number of EM maps being released every year as well as the improvement in their resolution ranges (Fig I.9).

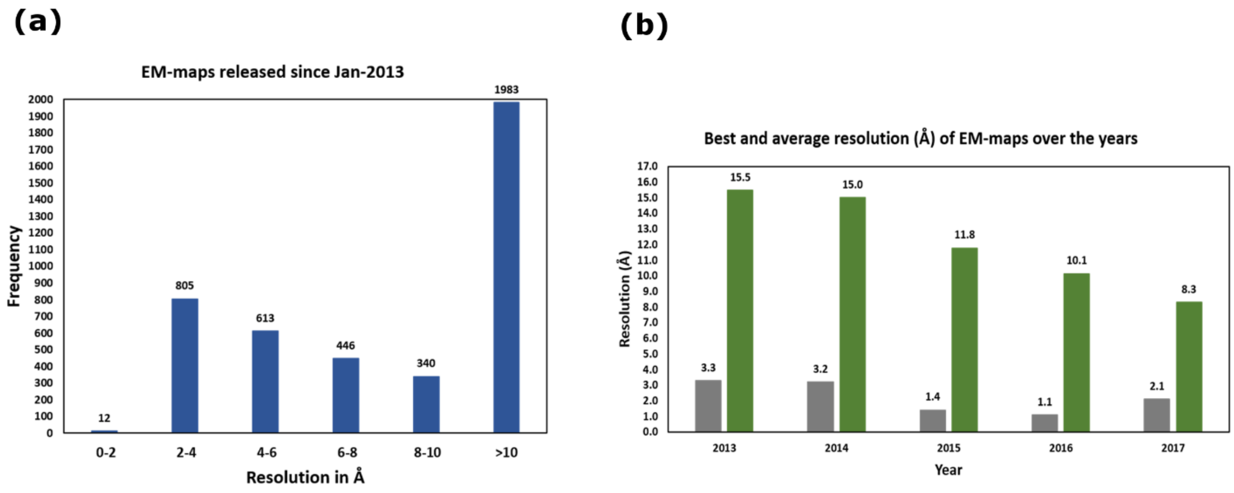


Figure I.9. *Statistics of EM maps and their resolution since 2013.* ***(a)*** *shows the frequency of occurrence of various resolutions of EM maps and* ***(b)*** *shows the best and average resolutions over the years. The worst resolution has been clipped to 30Å for this plot.*

$^+$Source internet, www.emdb.org/

The basic principle behind electron microscopy is the deflection of electrons in an electromagnetic field. An EM consists of an electron source, a series of lenses, and an image detecting system, which currently are high-end digital cameras [46]. As the electrons from the source hit the condenser lens, they are converged and fall on the object as a parallel beam. The aperture at the back focal plane of objective lens filters out the electrons scattered at very high angles, hence preventing them from reaching the image plane. Image magnification is provided by the objective lens and the projector lens. Once a good quality image is obtained, next task is to generate 3D reconstruction of the 2D projections of the objects using the phase information present in the image itself. The next important task after obtaining 3D electron density map is to calculate the pseudo-atomic model for the structure in question. Depending on the resolution, this can be achieved either through de-novo model building or through rigid body fitting or flexible fitting of predicted models / structures from other techniques. A simplified view of the electron microscope is given in Figure I.10.

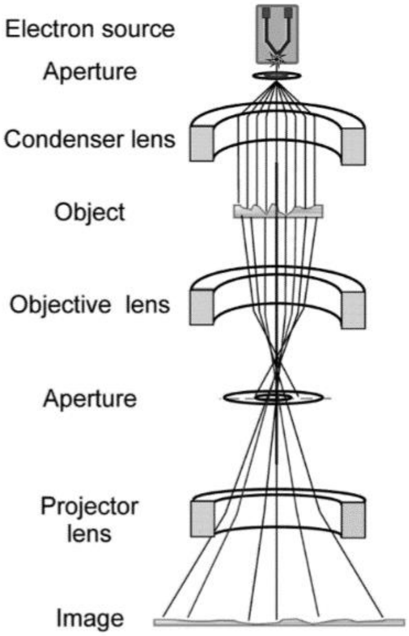


Figure I.10: *A simplified view of the electron microscope.*

$^{+}$Taken from [46]

I.5.4 *Circular Dichroism spectroscopy*

Another popular technique to determine the secondary structure content of the proteins is the Circular Dichroism (CD) spectroscopy which is based on the principle of the differential absorption of the left and right-handed circular polarised light. This technique is a routine to precisely estimate the secondary structure content of the protein and confirm their proper folding under the experimental setup. A typical CD spectral profile for different secondary structure looks like Figure I.11.

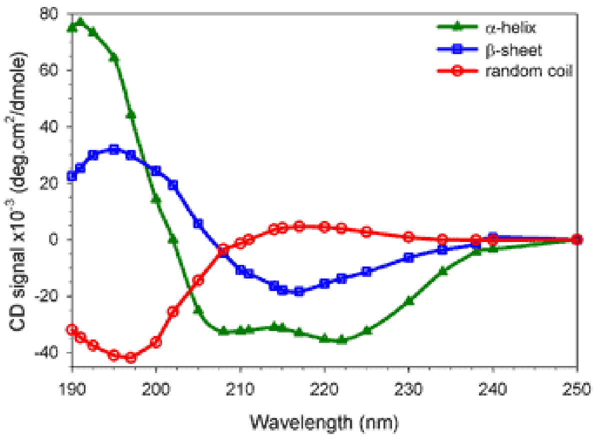


Figure I.11. ***A sample CD-spectra for a protein.*** *The profiles for different secondary structures are shown in different colours.*

+Source internet, www.fbs.leeds.ac.uk/facilities/cd/

I.6 *Computational Structural Biology*: *In silico techniques for protein structures*

It takes a considerable amount of time and effort to experimentally determine the three-dimensional structure of proteins using any of the above mentioned techniques. It sometimes takes from months to years to obtain a protein crystal which can successfully diffract. Though cryoEM circumvents the need of getting crystals, the technique is more amenable to the proteins with higher molecular weight. Moreover, the growth in structural space for proteins does not match up to the speed with which sequence space is growing (Fig I.12).

Hence, it becomes increasingly important to resort to the computational methods to predict three-dimensional structure of a given protein. The history of theoretically predicting the structural elements dates back to 1970s when Chou and Fasman calculated the propensities of amino acids in α-helices, β-sheets and turns [47]. Since then various methods have been developed to predict the secondary structures from the sequence [48–50]. Protein secondary structure prediction refers to the prediction of the conformational state of each amino acid residue of a protein sequence as one of the three possible states, namely, helices, strands, or coils, denoted as H, E, and C, respectively.

Another important application of computational methods is their ability to predict tertiary structure of proteins. Three main approaches are employed in computational 3D prediction are: homology modelling, threading, and ab-initio prediction. The first two are knowledge-based methods; they predict protein structures based on knowledge of existing protein structural information in databases. Homology modelling builds an atomic model based on an experimentally determined structure that is closely related at the sequence level. Threading identifies proteins that are structurally similar, with or without detectable sequence similarities. The ab initio approach requires molecular simulations to predict structures based on physicochemical principles governing protein folding without the use of structural templates. There are meta-servers that combine fold recognition and homology modelling to model a structure based on multiple templates matching different folds.

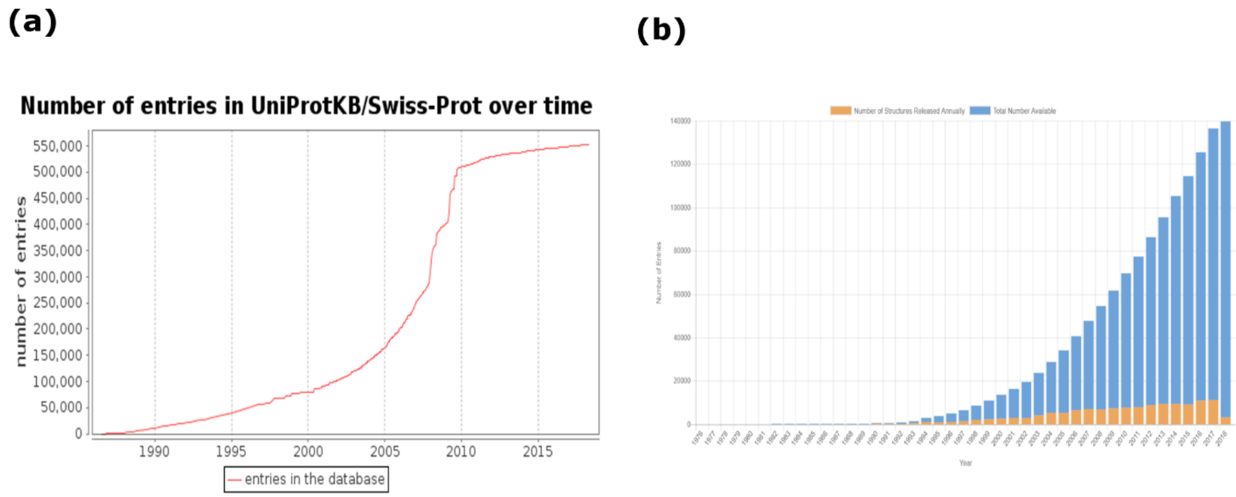


Figure I.12. *Growth of sequence space vis-à-vis the structural space.* ***(a)*** *shows the number of sequences deposited in Uniprot since 1990 (Taken from Uniprot)* ***(b)*** *shows the number of structures deposited in PDB over the years*

$^{+}$Source internet www.rcsb.org

I.6.1 <u>*Secondary structure assignment*</u>

Given their fundamental importance in protein structures, it is important to define and characterize secondary structure elements for a given protein structure. Various standard methods are available for this purpose. Several assignment methods can be used like, DSSP [51], STRIDE [52] and predefined libraries of secondary structure can also be used.

I.6.1.1 *DSSP – Define secondary structure of proteins*

The DSSP algorithm is a standard method for assigning secondary structure to the amino acids of a protein using the coordinates of the structure. The DSSP program was designed by Wolfgang Kabsch and Chris Sander. It identifies the intra-backbone hydrogen bonds of the protein using a purely electrostatic definition. A hydrogen bond is identified if E in the following equation is less than -0.5 kcal/mol. Based on the identified hydrogen bonds, eight types of secondary structure are assigned. These eight types are usually grouped into three larger classes: helix (G, H and I), strand (E and B) and loop (S, T, and C, where C sometimes is represented also as blank space).

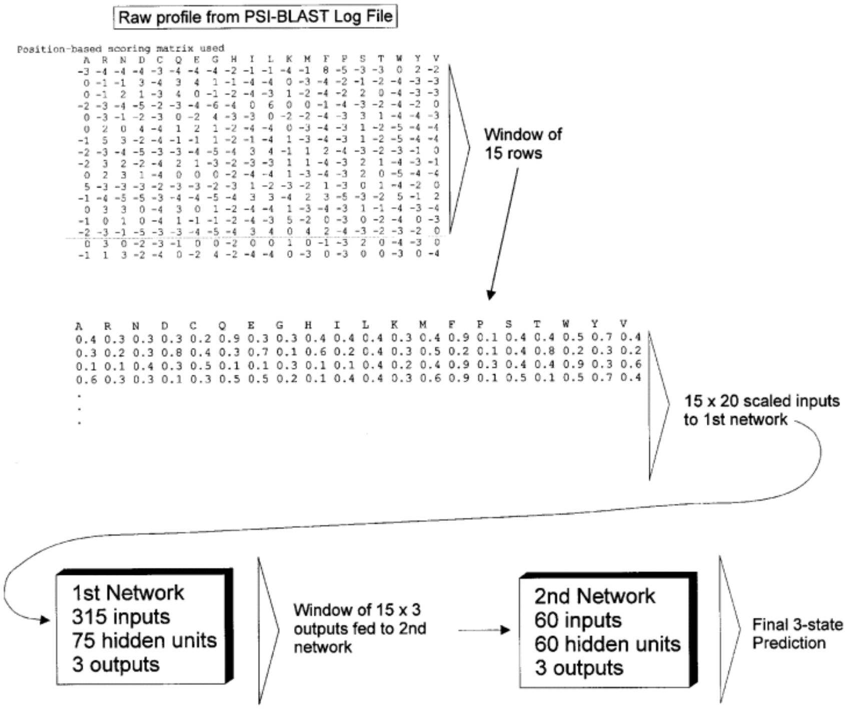
I.6.1.2 *STRIDE* - *STRuctural Identification*

STRIDE is also a secondary structure assignment tool like DSSP but instead of using only the hydrogen bond potential, it also includes dihedral angle potentials to define secondary structures within a protein. Hence, its criteria for defining individual secondary structures are more complex than those of DSSP. The STRIDE energy function contains a hydrogen-bond term containing a Lennard-Jones-like 8-6 distance-dependent potential and two angular dependence factors reflecting the planarity of the optimized hydrogen bond geometry. The criteria for individual secondary structural elements, which are divided into the same groups as those reported by DSSP, also contain statistical probability values derived from empirical examinations of solved structures. There have been comparisons between DSSP and STRIDE since their inception. It has been shown than DSSP and STRIDE agree for 95% of the cases [53]. I should be noted that it has been shown than both DSSP and STRIDE under-represent π-helix [54].

I.6.2 *Secondary structure prediction*

The prediction of secondary structures is based on the regular arrangement of amino acids in the secondary structures which are stabilized by hydrogen bonding patterns. The structural regularity serves the foundation for these prediction algorithms. Protein secondary structure prediction with high accuracy is not a trivial task. It has remained a very difficult problem for decades. Specifically, because protein secondary structure elements are context dependent. The formation of α-helices is determined by short-range interactions, whereas the formation of β-strands is strongly influenced by long-range interactions. Prediction for long-range interactions is theoretically difficult. Albeit, after more than three decades of effort, prediction accuracies have only been improved from about 50% to about 82%. There are many methods available for secondary structure prediction. Out of these PSIPRED is the most popular one [49].

PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) is a web-based program that predicts protein secondary structures using a combination of evolutionary information and neural networks. PSIPRED incorporates two feed-forward neural networks which performs an analysis on output obtained from PSI-BLAST. A profile is extracted from the multiple sequence alignment generated from three rounds of the PSI-BLAST. This profile is then used as input for a neural network prediction. To achieve higher accuracy, a unique filtering algorithm is implemented to filter out unrelated PSI-BLAST hits during profile construction. A schematic of PSIPRED is shown in Figure I.13.



```
Position-based scoring matrix used
        A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
       -3 -4 -4 -4 -3 -4 -4 -4 -2 -1 -1 -4 -1  8 -5 -3 -3  0  2 -2
        0 -1 -1  3 -4  3  4  1 -1 -4 -4  0 -3 -4 -2 -1 -2 -4 -3 -3
        0 -1  2  1 -3  4  0 -1 -2 -4 -3  1 -2 -4 -2  2  0 -4 -3 -3
       -2 -3 -4 -5 -2 -3 -4 -6 -4  0  6  0  0 -1 -4 -3 -2 -4 -2  0
        0 -3 -1 -2 -3  0 -2  4 -3 -3  0 -2 -2 -4 -3  3  1 -4 -4 -3
        0  2  0  4 -4  1  2  1 -2 -4 -4  0 -3 -4 -3  1 -2 -5 -4 -4
       -1  5  3 -2 -4 -1 -1  1 -2 -1 -4  1 -3 -4 -3  1 -2 -5 -4 -4
       -2 -3 -4 -5 -3 -3 -4 -5 -4  3  4 -1  1  2 -4 -3 -2 -3 -1  0
       -2  3  2 -2 -4  2  1 -3 -2 -3 -3  1  1 -4 -3  2  1 -4 -3 -1
        0  2  3  1 -4  0  0  0 -2 -4 -4  1 -3 -4 -3  2  0 -5 -4 -4
        5 -3 -3 -3 -2 -3 -3 -2 -3  1 -2 -3 -2  1 -3  0  1 -4 -2  0
       -1 -4 -5 -5 -3 -4 -4 -5 -4  3  3 -4  2  3 -5 -3 -2  5 -1  2
        0  3  3  0 -4  3  0  1 -2 -4 -4  1 -3 -4 -3  1 -1 -4 -3 -4
       -1  0  1  0 -4  1 -1 -1 -2 -4 -3  5 -2  0 -3  0 -2 -4  0 -3
       -2 -3 -1 -5 -3 -3 -4 -5 -4  3  4  0  4  2 -4 -3 -2 -3 -2  0
        0  3  0 -2 -3 -1  0  0 -2  0  0  1  0 -1 -3  2  0 -4 -3  0
       -1  1  3 -2 -4  0 -2  4 -2 -4 -4  0 -3  0 -3  0  0 -3  0 -4
```

```
A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
0.4 0.3 0.3 0.3 0.2 0.9 0.3 0.3 0.4 0.4 0.4 0.3 0.4 0.9 0.1 0.4 0.4 0.5 0.7 0.4
0.3 0.2 0.3 0.8 0.4 0.3 0.7 0.1 0.6 0.2 0.4 0.3 0.5 0.2 0.1 0.4 0.8 0.2 0.3 0.2
0.1 0.1 0.4 0.3 0.5 0.1 0.1 0.3 0.1 0.1 0.4 0.2 0.4 0.9 0.3 0.4 0.4 0.9 0.3 0.6
0.6 0.3 0.3 0.1 0.3 0.5 0.5 0.2 0.1 0.4 0.4 0.3 0.6 0.9 0.1 0.5 0.1 0.5 0.7 0.4
.
.
.
```

Figure I.13. *Workflow of PSIPRED.*

$^{+}$Taken from [42].

I.6.3 *Protein Blocks: A comprehensive structural alphabet*

A structural alphabet (SA) is a library of *N* structural prototypes (the letters). Each prototype is representative of a backbone local structure of *l*-residues length. The combination of those structural prototypes is assumed to approximate any given protein structure. One of the most developed and comprehensive SA is the Protein Blocks (PBs) [53].

PBs are a structural alphabet composed of a set of 16 local prototypes each of 5 residues length, labeled from *a* to *p* (see Fig I.14 Bottom). They are described as series of eight Φ, Ψ dihedral angles. An unsupervised classifier similar to Kohonen Maps [55,56] and Hidden Markov Models [57] was used to define PBs. Therefore, they approximate all the local regions of a protein structure with an average RMSD of 0.41 Å [58]. The PBs *m* and *d* can be roughly described as prototypes for the central region of α-helix and β-strand, respectively. PBs *a*-*c* primarily represent the N-cap of β-strand while *e* and *f* correspond to C-caps; PBs *g* - *j* are specific to coils, PBs *k* and *l* correspond to N cap of α-helix while PBs *n* - *p* to C-cap.

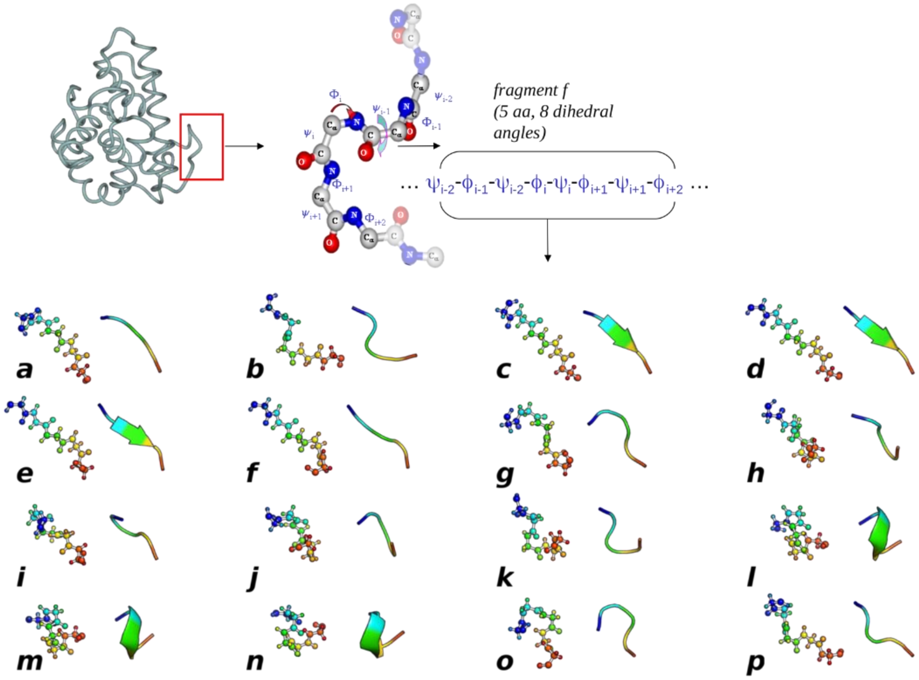

fragment *f*
(5 aa, 8 dihedral angles)
… ψ_{i-2}-ϕ_{i-1}-ψ_{i-2}-ϕ_i-ψ_i-ϕ_{i+1}-ψ_{i+1}-ϕ_{i+2} …
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

Figure I.14: Protein blocks. *Top row depicts the 5 residue long prototype. Bottom row shows the 16 protein blocks along with their respective secondary structure approximations.*

+Adapted from [53] and [58].

<u>*PB Assignment:*</u> For each "*n*th" position of the structure, 8 dihedrals ψ (*n* − 2), φ (*n* − 1), ψ (*n* − 1), φ (*n*), ψ (*n*), φ (*n* + 1), ψ (*n* + 1), φ (*n* + 2) are compared to the dihedrals of each of the 16 PBs. The comparison is made by a least squares approach to match the RMSDA criteria (*Root mean square Deviation on Angular Values*) [59].

$$RMSDA\,(V_1, V_2) \; = \; \sqrt{\frac{1}{2(M-1)} \sum_{i=1}^{M-1} \quad [\psi_i(V_1) - \psi_i(V_2)]^2 + \; [\phi_{i+1}(V_1) - \phi_{i+1}(V_2)]^2}$$

where, V_1 is the 8 dihedrals vector extracted from the 5 residues long window; V_2 is the 8 dihedrals vector corresponding to the compared PBs. PB, which gets lowest RMSDA is chosen as the representing conformation observed in the window.

Applications: PBs have been used to address various problems including, protein superimposition [60,61], general analyses of flexibility [62,63] and prediction of structure and flexibility [64–67] and protein binding sites, and structural analysis of β-bulges [68]. PBs can be assigned to a given structure or an ensemble with valid coordinates using PBxplore [69]. The structural analysis of different structural dataset is assisted by two statistical measures derived from the assigned PBs.

Neq: Quantification of the structural flexibility at a given position *n*, can be obtained by calculating the average number of PBs across a set of conformers at position *n*. This is called the "equivalent number" of PBs or *Neq*. *Neq* is based on a statistical metric similar to Shannon entropy [53]. It is calculated as:

$$Neq \; = \; exp\left(-\sum_{i=1}^{16} \quad f_x . \, ln(f_x) \right)$$

where f_x is the frequency of PB 'x'. The value of x can be any PB from *a* to *p*. An *Neq* value of 1 will indicate that only one type of PB is observed at position *n* while *Neq* value of 16 will denote a random and propotional (1:16) distribution.

I.6.4 *Tertiary structure prediction*

The tertiary structure of the proteins is predicted either using *ab-initio* methods or based on a template identified through homology. The latter is the more common, reliable, less time-consuming method and is based on the paradigm that similar sequences have similar structures and hence similar functions [10], [11]. Homology modelling starts with identification of a suitable template which shares homologous relationship with the sequence of interest. Using elegant computational algorithms, the coordinates of the backbone of the template are copied to the query and the side chains are optimised. Modeller is the most popular program to perform molecular modelling and is described in brief herein [70].

I.6.4.1 *Modeller*

Modeller is a computer program that models three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints (Fig I.15). The initial step before starting the modelling procedure is to identify a suitable template. This forms the foundation for rest of the modelling process. The template selection involves searching for a homologous structure in PDB using either BLAST [71] or any other fold recognition tool such as Phyre2 [72] or HHpred [73]. Generally, the structures with sequence identity greater than 30% are considered safely as homologous to the query protein. Once the structure of suitable confidence is identified as a template, an alignment is performed between the query and the template. This can be achieved either using scripts from Modeller or using a suitable alignment tool. This alignment, in PIR format, is the input to the Modeller program. From its alignment with template 3D structures, Cα- Cα distances, hydrogen bonds and dihedral angle restraints for the target sequence are calculated by Modeller. The form of these restraints has been obtained from a systematic statistical analysis of the relationships between many pairs of homologous structures [74]. The spatial restraints are obtained empirically, from a database of protein structure alignments. These restraints are expressed as probability density functions (pdfs) for the features to be restrained. For example, the probabilities for main-chain conformation of an equivalent residue in a related protein are expressed as a function of the local similarity between the two sequences. A smoothening procedure has been employed in the derivation of these relationships to minimise the problem of sparse database. Next, these spatial restraints and Charmm energy terms enforcing proper stereochemistry are combined into an objective function [75]. The output is a 3D model for the target sequence containing all main-chain and side-chain non-hydrogen atoms which ensures a minimal deviation from the input restraints. The final model is then optimised using variable target function methods employing methods of conjugate gradients and molecular dynamics with simulated annealing. Several slightly different models can be calculated by varying the initial structure. The variability among these models can be used to estimate the errors in the corresponding regions of the fold. Also, the loops are further refined using different protocols for loop modelling. Side chains are further optimised using the rotamer libraries, which are favoured side chain torsion angles extracted from known protein crystal structures.

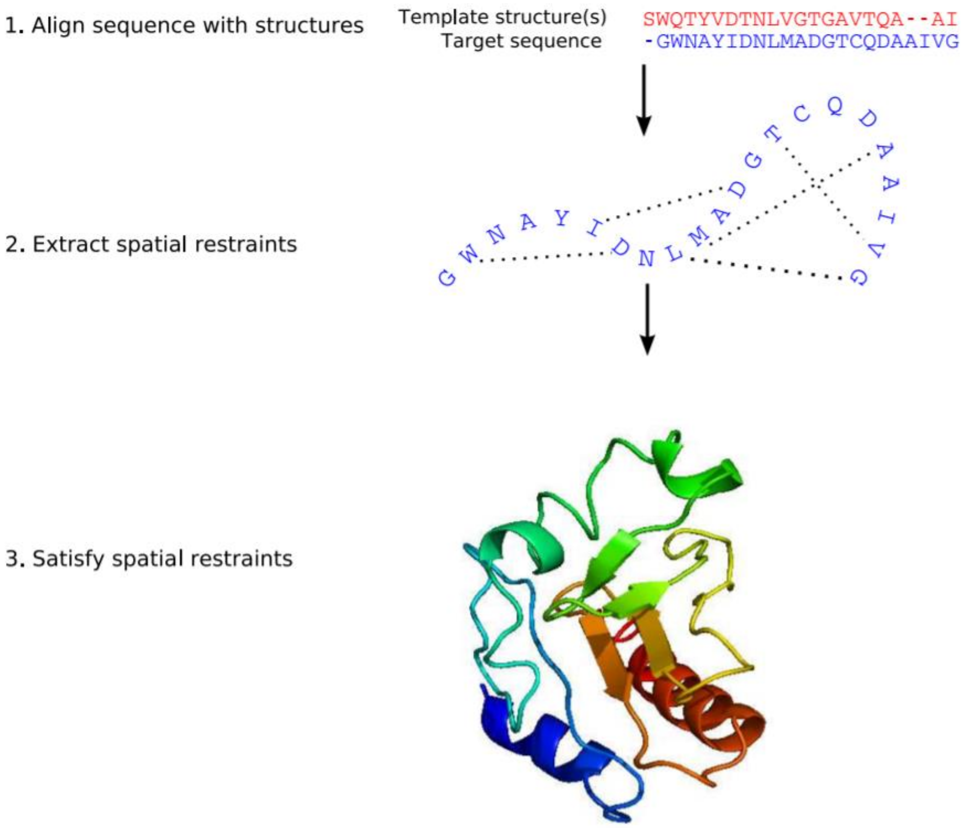

1. Align sequence with structures
Template structure(s)
Target sequence
SWQTYVDTNLVGTGAVTQA--AI
-GWNAYIDNLMADGTCQDAAIVG
2. Extract spatial restraints
G W N A Y I D N L M A D G T C Q D A A I V G
3. Satisfy spatial restraints

Figure I.15. ***Workflow of Modeller***. *The target sequence is aligned to the template, spatial restraints are obtained and then satisfied to obtain a 3D model.*

+Taken from Modeller v9.14 tutorial pages

I.6.5 *Accuracy of predicted models*

The accuracy of comparative models depends on the extent of the sequence identity between the query and the template [76]. Usually, errors are expected to be more in the structurally variable region than in the structurally conserved region. The CASP (Critical Assessment of protein structure prediction) assessments happen every two years to test the ability of different structure prediction methods to accurately model the query proteins. It has generally been seen that many of the models show higher RMSD to the true native structure than the one selected by a structural alignment to be the best available template.

I.6.6 *Dynamic nature of protein structures*

Proteins may exist in multiple conformations and are always in motion under cellular environment. The paradigm of sequence-structure-function also includes dynamics before function these days. Thus redefining the paradigm as; proteins with similar sequences share similar structures which in turn share similar dynamics and hence give rise to similar functions. While experimental techniques such as NMR and cryoEM help understanding the underlying dynamics behind these proteins, computational methods also provide insights into the same. Due to their high efficiency, molecular dynamics simulation and normal mode analysis are the methods of choice in majority of the cases to understand the dynamics associated with a protein in a simulated cellular environment.

I.6.7 *Molecular dynamics*

Molecular dynamics (MD) is a computer simulation method for studying the physical movements of atoms and molecules. In molecular dynamics, successive configurations of the system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary as a function of time. The trajectory is obtained by solving the differential equations in the form of Newton's second law (F = ma):

$$\frac{\delta^2 x_i}{\delta t^2} = \frac{F_x}{m_i}$$

This equation describes the motion of a particle of mass m; along one coordinate (x,) with F_x. being the force on the particle in that direction. MD simulation is based on an assumption that system follows ergodicity which means that all accessible microstates are equally probable over a long period of time.

Briefly, molecular dynamics simulation begins by defining the initial coordinates for the system of interest. A small time-step Δt is chosen such that the next coordinates can be evolved. Next atom positions are predicted and the velocities are updated. The forces are calculated for the new set of positions and the positions are further adjusted. Periodic boundary conditions are employed and then the next iteration follows. This procedure is repeated until the given time. Few concepts important for understanding the MD theory are described below.

Force fields: In order to calculate the potential energy of the system, mathematical functional forms and parameters have been defined, called force fields. Potential energy is further used to calculate the forces on the atoms. The force field is a collection of equations and associated constants designed to reproduce molecular geometry and selected properties of tested structures. The parameters for energy functions have been derived from physical or chemical experiments or from quantum mechanical calculations. The equation for the calculation of potential energy in molecular mechanics include interaction terms from bonded and non-bonded interactions. The specific parameters for the interactions vary between force fields, but a general expression for total potential energy can be written as:

$$E_{total} = E_{bonded} + E_{non-bonded}$$

Where $E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$ and $E_{non-bonded} = E_{electrostatic} + E_{van\ der\ waals}$

The bond and angle terms are modelled as quadratic energy functions and non-bonded terms are modelled as Lennard-Jones and Coulombs potential. The detailed energy function is calculated as:

$$U(R) = \sum_{bonds} k_i^{bond}(r_i - r_0)^2 + \sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2 + \sum_{dihedrals} k_i^{dihedral}[1 + \cos(n_i\phi_i + \delta_i)] + \underbrace{\sum_{i}\sum_{j\neq i} 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \sum_{i}\sum_{j\neq i}\frac{q_i q_j}{\varepsilon r_{ij}}}_{\text{Non-bonded}}$$

Different force-fields are designed for different purposes. e.g. AMBER [77] force-field is used majorly for simulating DNA and proteins. CHARMM [78] can be used for both small molecules and macromolecules. GROMOS is a general purpose force-field for the study of biomolecules [79].

Energy minimisation: Classical MD simulations try to explore all possible conformations of a protein in a given energy well assuming that the protein structure indeed lies at the energy minima. Crystal structures may not always be trapped in their minimum energy conformation while crystallisation. Hence, before proceeding with the simulations, it is mandatory to find minima for the protein structure. Mathematically, minima occur when the when the first derivative of potential is zero and when the second derivative is positive. There are two commonly use methods to perform energy minimization, steepest descent [80] and conjugate gradient [81]. Steepest descent is the simplest method to use for performing energy minimization. It follows the fastest decrease of the potential “U” opposite of the gradient. It is the fastest method from a poor starting geometry but can converge very slowly near energy minima. This is due to the fact that it can oscillate back and forth across a minimum. Conjugate gradient on the other hand, adds history to the steepest descent method to gather second derivative information and guides the search where the derivative determines the pathway.

Solvation: Since the biomolecules have to be simulated in a cellular environment, the protein has to be solvated. Two types of solvation methods are available: Implicit solvation [82] and explicit solvation [83]. In implicit solvation model, the solvent is defined as a continuum of homogenous polarizable medium which possess properties equivalent to the solvent. While in an explicit solvent model, the coordinates for the solvent molecules are explicitly defined. This solvation method is more realistic and can give a true picture of interaction between the solute and solvent. The protein molecule is solvated in a box before performing an energy minimisation step and then the main MD run is performed.

Periodic boundary conditions (PBC): In majority of the simulations, the simulation box should be large enough to circumvent the boundary artefacts. Such scenario can be avoided by employing the periodic boundary conditions, where one side of the simulation comes back from the opposite side, mimicking a bulk phase [84]. For PBCs, particles are enclosed in a box, and the box is replicated to infinity by rigid translation in all the three Cartesian directions, completely filling the space. The basic idea behind the PBC is that if an atom moves in the original simulation box, all its images move in a concerted manner by the same amount and in the same fashion. When PBCs are applied, there is a chance that the number of interacting pairs increase enormously. The reason is that there is an interaction not only interacts with other particles in the simulation box, but also with their images. Such a problem is avoided by choosing a finite range potential within the criteria of *minimum image convention* [84]. The essence of the minimum image criteria is that it allows only the nearest neighbors of particle images to interact.

MD Ensembles: In molecular mechanics, the ensembles are the statistical entities that are used to represent the possible states of a system. Different ensembles utilised in MD simulation are canonical ensemble, isothermal-isobaric ensemble and microcanonical ensemble. Canonical ensemble conserves the number of molecules (N), volume (V) and temperature (T) of the system, hence also called as NVT ensemble. The temperature is maintained through the association of a thermostat. In isothermal-isobaric ensemble, the number of molecules (N), pressure (P) and temperature (T) of a system is conserved, hence popularly known as NPT ensemble. In micro-canonical ensemble, the system's energy (E) is conserved along with number of molecules (N) and volume (V), hence called NVE ensemble.

I.6.8 *Normal mode analysis (NMA)*

Classical molecular dynamics simulations generally provide information on the dynamics happening at the μs time-scales which mostly includes the side chain motions or at most some loop motions if the energy barrier between the different states are within a difference of few K_BT. Bigger conformational changes such as domain motions can be accessed using advance sampling techniques in MD, which requires more computational power and an expert level understanding of the field and parameters. In such a scenario, a simpler, yet powerful, network-based technique, normal mode analysis (NMA) can be used. NMA is purely a geometry-based approach where a protein is modelled as a network of mass and springs. Generally, the Cα atoms of a protein are defined as nodes and a spring is defined for the edges connecting these Cα atoms within a certain cut-off distance. The movement of each node is expressed in terms of squared fluctuations i.e. displacement of nodes from their mean positions. Collective motion of many such nodes in a certain direction defines the global motions which are biologically relevant and correspond to the domain motions [85]. Such a simplified approach to study protein dynamics has been shown to successfully reproduce biologically relevant motions [86].

Two types of NMA can be implemented; Gaussian network model (GNM) or Anisotropic network model (ANM). In GNM, the squared fluctuations are assumed to be isotropic while in ANM the fluctuations are anisotropic. An adjacency matrix is diagonalized in GNM while in ANM a hessian matrix is diagonalized to calculate the Eigen vectors and Eigen values. These values correspond to the direction of motion in ANM. Hessian matrix consists of the double derivative of the hooks potential defined for the system.

I.7 *Biomolecular interactions*

Proteins seldom work in isolation. Multiple interactions within a cell viz. protein-protein, protein-ligand and protein-DNA are key to proper functioning of a cell. Besides the experimental methods to study biomolecular interactions, various computational methods are also available. Two biomolecular entities can be computationally docked to study their binding modes. Protein can be docked with another protein using HADDOCK [87].

HADDOCK (High Ambiguity Driven biomolecular DOCKing) is an information-driven flexible docking approach for the modelling of biomolecular complexes [88]. Docking is defined as the modelling of the structure of a complex based on the known three-dimensional structures of its constituents. HADDOCK incorporates a wide variety of experimental and/or bioinformatics data to drive the modelling. This allows focusing the search to relevant portions of the interaction space using a more sophisticated treatment of conformational flexibility.

AutoDock [89] is used to predict the binding mode of a protein with a ligand. It is a freely available, open-source software which practically is a suit for automatic docking tools. AutoDock has been widely-used and there are many examples of its successful application in the literature [90,91]. It is very fast, provides high quality predictions of ligand conformations, and good correlations between predicted inhibition constants and experimental ones. AutoDock has also been shown to be useful in blind docking, where the location of the binding site is not known.

I.8 *Molecular phylogenetics*

Genetic changes due to mutations or recombination gets accumulated in each generation of an organism or population. In subsequent generations, these accumulations may exhibit phenotypic changes in the organism thus leading to its evolution. The rate of such genetic changes is fundamental to understand the evolution of a given species or taxonomic group. In molecular biology, the rate of change of particular biomolecules like, DNA (nucleotides), RNA (gene) or amino acids (protein) is of interest. The rate of accumulation of changes in these biomolecules studied along with the evolution of a species in tree of life is called Molecular phylogenetics.

I.8.1 *Type of changes*: In molecular biology mutations are caused by substitutions of nucleotide bases or amino acids. However, given the central dogma of molecular biology the substitutions in nucleic acids are fundamental. In DNA the substitutions are of two types:

A) Transitions: a substitution of a purine by purine base or pyrimidine by a pyrimidine base. A ↔ T or C ↔ G substitutions will classify as transitions [Figure I.16].

B) Transversions: a substitution of a purine base by a pyrimidine base is called transversion. (A or G) ↔ (C or T, or U *in RNA*) is termed as transversion. Such substitutions are less frequent than transitions [Figure I.16].

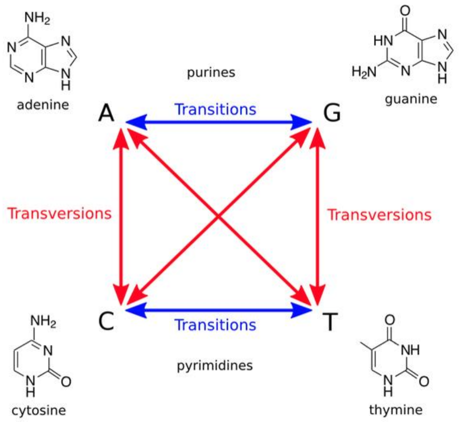


Figure I.16. ***DNA substitutions.*** *The exchanges between purine to purine nucleotide bases is called transitions and are more frequent in nature. The exchange of a purine base with a pyrimidine base is called transversion. Transversions are rare.*

+Source internet, www.humangenomeproject.org

Due to the degeneracy of the genetic code, substitutions in the DNA or RNA may not affect the amino acid sequence and thus the protein function will be unaltered. Such synonymous substitutions are called silent mutations at amino acid level. However, silent mutations can accumulate over generations and put selective pressure on a specific codon for an amino acid, altering the extent of protein expression. Besides, there are missense (non-synonymous) and nonsense mutations (stop codon) that can lead to change in expression and loss of expression or truncated expression, respectively. Due to failure of DNA repair machinery mostly due to external factors there can also be an insertion or deletion of a nucleotide base that can lead to frameshift mutation causing adverse effects in codon reading by ribosome.

I.8.2 *Substitution rates*: Depending on the type of molecule, the rate of selective substitutions differ. For instance, in DNA and RNA, the minimum possibility of a mutation is 1 in 4, i.e 25% while it goes down to 5% (1 in 20) in case of proteins. Therefore, separate substitution models, sensitive to the type of biomolecules is required while studying their evolution. With the advancement of molecular phylogenetics, there are distinguished substitution models available for varied kinds of molecular analysis. More details on practical use of different matrices is provided in chapter 6, section 6.2.4.

I.8.3 *SNP or mutation*: When considered in regards to an individual organism, the nucleotide or amino acid substitution is termed as a point mutation. However, when a population or species is considered, environment factors and genetic events like genetic drift, geographic displacement can lead to more than one kind of substitution at the same position. Therefore, in context of a population, it is termed as a single nucleotide polymorphism (SNP) or amino acid variant [92]. An example is shown in Figure I.17

I.8.4 *Phylogenetic tree generation*: A phylogenetic tree is generated by using the selective substitution rates on a biomolecule and the variations induced by those substitutions in the sequence of the biomolecule in different species. Therefore, quintessential for generating a phylogram is the sequences of all possible homologs (based on local identity) of a given protein/DNA or RNA. All the sequences are then globally aligned to identify the conservation sites as well as highly mutated sites. The multiple sequence alignment (MSA) can be generated by two prominent methods:

Progressive or hierarchical method [93]:

A crude MSA is generated by first aligning the most similar sequences and subsequently adding less related sequences or groups to the alignment. The inclusion of new sequences is carried out until the entire query set has been incorporated into the solution. The initial tree describing the sequence relatedness is based on pairwise comparisons that may include heuristic pairwise alignment methods. Thus, the alignment results are dependent on the choice of "most related" sequences and therefore can be sensitive to inaccuracies in the initial pairwise alignments.

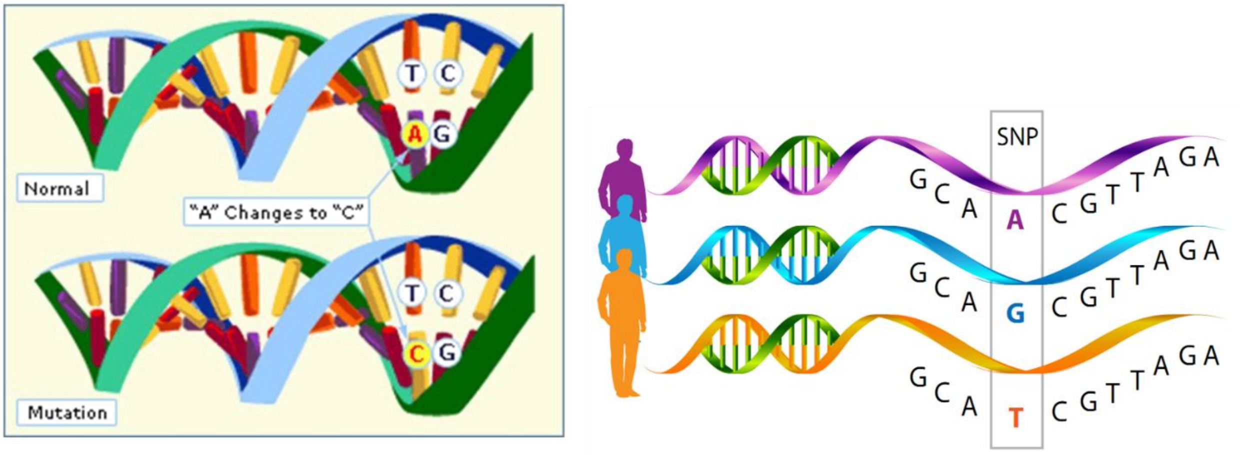


Figure I.17. ***Point mutations and SNP***. *A) A point mutation is any change in the sequence of DNA that may or may not alter the amino acid sequence. B) Single nucleotide polymorphism (SNP) are the co-existing changes in the DNA all of which can affect the amino acid sequence. However, none of the effects of SNP is deleterious. The different polymorphisms (A, G, T) are termed as alleles. If the lower representing allele in a SNP has a fixation value of above 1%, then it can be called as a mutation. However, defining an allele by 1% fixation is still debatable.*

$^+$Source internet, www.humangenomeproject.org

Iterative method [93,94]:

It optimizes an objective function based on a selected alignment scoring method by assigning an initial global alignment and then realigning sequence subsets according to the scoring method. The realigned subsets are then themselves aligned to produce the next iteration of MSA. The iterations are continued each of the query sequence is aligned at least twice. Therefore, an improvement over iterative method removes the bias of an initial alignment. However, since it usually takes multiple iterations to achieve a final MSA, the method is computationally exhaustive.

I.8.5 *Types of phylogenetic trees*: The goal of a molecular phylogenetics is to construct a tree topology that best explains the evolutionary history of the given sequences. There are four approaches to analyze the generated MSA and define a tree topology. The most primitive and basic ones are: distance based methods and parsimony based methods. Distance based methods uses the substitution models to estimate pairwise evolutionary distances among each sequence of MSA. The distance matrix is then analyzed by hierarchical clustering type methods such as neighbor-joining (single linkage clustering) or unweighted pair-group with arithmetic mean (average clustering) [94]. In the parsimony approach, the goal is to identify a topology that requires the fewest necessary changes to explain the differences among the observed sequences. Both of these methods works better in very closely related sequences but often fail to work with datasets comprising of distant homologs. In case of highly diversified homologs, the two character based methods are highly useful as they employ probabilities and thus ignore initial bias.

A) *Maximum likelihood based [95]*:

An initial tree is first built using a fast but suboptimal method such as Neighbor-Joining. A likelihood function is calculated based on the substitution model. The branch lengths of the NJ tree are then adjusted to maximize the likelihood of the data set under the given substitution model. Then variants of the topology are created using the NNI (nearest neighbor Interchange) method to search for topologies that fit the data better. Maximum-Likelihood branch lengths are computed for these variant tree topologies and the greatest likelihood is retained as the best choice. The search continues until no greater likelihoods are found [96].

B) *Bayesian inference based [97]*:

BI based methods can be seen as an extension of the ML based methods with a major difference being the use of a prior probability. BI based methods use the initial MSA and the substitution matrix to generate *a priori* probability that the given topology should belong to the base tree architecture. Therefore, instead of testing the likelihood as in ML, BI used the prior probability to make the decision [96].

I.8.6 *Reliability of the tree topology*: After a tree topology have been generated, statistical measure like bootstrapping, jackknifing are used to test its robustness [95]. Bootstrapping works very similar to e-value calculation. Random sites in the alignment is replaced and is used to build new trees. It is possible that some positions will be repeated in the subsample, while some positions will be left out. Such multiple resamples are run for mostly 100 to 1000 times based on the size of the dataset. A bootstrap value closer to 100 gives higher confidence in the branching.

An introduction to these concepts is important to understand the subsequent chapters in the thesis. All of these concepts would be used at different places in different chapters. For simplification, the thesis is organized into three major sections; A) Chapters 1 to 3 that focuses on secondary structures rather than whole proteins. B) Chapter 4 deals with domain level analysis of a structural assembly while Chapter 5 studies a complete protein structure of a membrane protein. Proteins involved in both the chapters are crucial due to the pathologies they are involved in. C) Chapter 6 is though related to chapter 5 but it does not study the structural behaviors. Rather it is focused on sequence analysis and phylogenetics which adds a fresh perspective towards the end of the thesis.

Each chapter deals with a systematic study of an individual idea. There are some supplementary (marked as ‘S’) and sub-chapters (marked as ‘a’, ‘b’, ‘c’ with the chapter number). The supplementary chapter provides information related to the parent chapter while sub-chapters are individual studies dealing with the question in parts. Each chapter including sub-chapters and supplementary chapters contains information regarding, introduction to the topic, methods used, results and discussion and conclusiong and future perspectives. In some chapters, an additional section for acknowledgment is included to thank the team members and collaborators for their support. At last there is a conclusive outline of the thesis that is written from the perspective of personal learning experience during the PhD and the impact of different projects on that learning curve. There are two chapters that should have been included in this thesis document but cannot be included due to restraints of space and time. Each of these two chapters are associated with our collaborators in India and Canada. My responsibility towards the collaboration with Karboune lab, McGill university, Canada is to perform docking analysis of seven different sugars with Levansucrase from five different species. The collaboration with N. Srinivasan lab at Indian Institute of Science involves the investigation of structural dynamics of different protein kinases in their active and inactive state. Recently, a research article derived on molecular dynamics analysis of inactive and active protein kinase A (PKA) have been submitted to Biophysical Journal.

Following are the main objectives on which chapters are constructed:

OBJECTIVES

O.1 *Understanding local dynamics of repetitive secondary structures.*

O.2 *Studying the dynamic behavior of structural alphabets- Protein Blocks.*

O.3 *Systematic study of local structural changes in special events in a protein structure biology, like Post-translational modifications and Disorder to order transitions.*

O.4 *To understand role of inherent flexibility at a more complex structural organization and function; Study of protein domains in a protein structure that undergoes structural transition.*

O.5 *Systematic analysis of structural organization and dynamics in a protein protein complex; Study of a structural oligomer and interactions involved in pathology.*

Chapter 1: Understanding dynamic behaviors of secondary structures- Dynamics and deformability of α-, 3_{10} - and π-helices

1.1 Introduction

Before the first protein 3D structure was solved at atomic resolution [31], Pauling and Corey provided evidence that polypeptide chains can adopt a limited number of repetitive local protein structures stabilized by intramolecular hydrogen bonds [98–100]. The two major local folds are: (i) the α-helix (or 3.6_{13} helix) with hydrogen bond between amino acid residues *i* and *i + 4*, and (ii) the β-sheet composed of extended strands with hydrogen bonds between adjacent strands, running parallel or antiparallel. They roughly represent 1/3rd and 1/5th of the residues found in proteins, respectively. Therefore, protein structures are often represented as seen in crystals as (i) rigid macromolecules (ii) comprising of repetitive units of helices, sheets and coils. However, both the definitions are partial because in physiology proteins are highly dynamic macromolecules and the description of protein structures could be more precise.

Of the current popular secondary structures, helices and β-sheets are the two predominant conformational forms. The hydrogen bonding pattern of the two differs considerably and therefore they can be treated as two separate independent conformations with respect to protein folding. For instance, different types of helices like 3_{10}- and π-helices have been proposed as intermediate conformations in the folding of an α-helix [101–103]. Similarly, β -turns, bends, and strands are commonly observed during the formation of β-amyloid aggregates. Moreover, our team have previously worked with ‘chameleon sequences’ that are short stretch of structured regions that can interchange between helix and strand conformations [104]. Therefore, for simplicity of the analysis, the dynamics of these secondary structures will be studied individually. The first one being different types of helices.

1.1.1 *α-helices*

Since the characterization of helices in 1951 [100], extensive explorations have been conducted to better understand their formation and their role in the kinetics of protein folding. Although a general view of the folding kinetics is too complex to define theoretical folding models for helices. Recent cutting-edge experiments have underlined their significant contribution through different examples [105]. Furthermore, with accumulating high-resolution experimental 3D structure, many studies have been carried out to decipher the sequence features that preferentially drives folding towards a given local fold.

In this context, α-helices have been intensely analyzed [106–108]. It has been emphasized that the length of an α-helix depends on its amino acid composition [109,110] and that its extremities (or caps) have specific signatures [111–113]. These caps can be stabilized by hydrophobic interactions between helical residues and residues outside the repetitive structures [114–116]. The importance of such interactions was highlighted for instance in the case of class α-glutathione transferases where, using computational approaches, it was shown that the highly conserved helix 9 modulates their catalytic and binding function and that a mutation of N-cap residue Asp-209 destabilizes the enzyme's function [117,118].

For structural description of α-helices, please see section I.2.2

1.1.2 *3_{10}-helices*

3_{10}-helices (shown in Fig 1.1) are less frequent than α-helices and represent about 4% of the residues in proteins. The 3_{10}-helix is characterized by intramolecular hydrogen bonds between residues i and i+3, and is usually short, containing three or four residues per turn [119,120]. Nonetheless, two-turn and longer 3_{10}-helices have also been reported [120]. In terms of location, they are preferentially observed at the termini of α-helices and are considered as connectors between two α-helices [20,121,122]. However, the 3_{10}-helix is also often found in the regions connecting strands within β-hairpin or β-β-corner motifs [123]. In terms of sequence, their amino acid content is different from the α-helix [124].

A specific analysis of a 3_{10}-helix adjoining the α-helix and β-strand has shown that the composition of 3_{10}-helices in vicinity of β-strands is much more conserved among family members of homologous structures than those 3_{10}-helices adjacent to two helices [123]. The preferred length of the 3_{10}-helix occurring between an α-helix and β-strand is equal to 3 residues, but extends to 4 residues when located between two α-helices (α-3_{10}-α) [125].

1.1.3 *π-helices*

π-helices (shown in Fig 1.1) which are less frequent than both of α- and 3_{10}-helices. They represent about 0.02% of the residues in proteins. In the π-helix (or 4.4_{16}-helices), hydrogen bonds are formed between amino acid residues *i* and *i+5*. This helix conformation is less stable due to steric constraints, which could also explain why π-helices are rare [16]. In 2000, Weaver found only 14 well-defined π-helices in the available PDB files (i.e. about 13500 structures) [126]. However, the π-helix should occur more frequently in protein structures than has been previously described, and should be conserved within functionally related proteins [54,127]. π-helix show distinct residue preferences that differ from those of α-helices [127]. Interestingly, it was shown that on a limited number of π-helices they were directly linked to the formation or stabilization of a specific binding site [126]. Thus conformationally, π-helices can be of crucial importance in protein-protein or protein-ligand interactions.

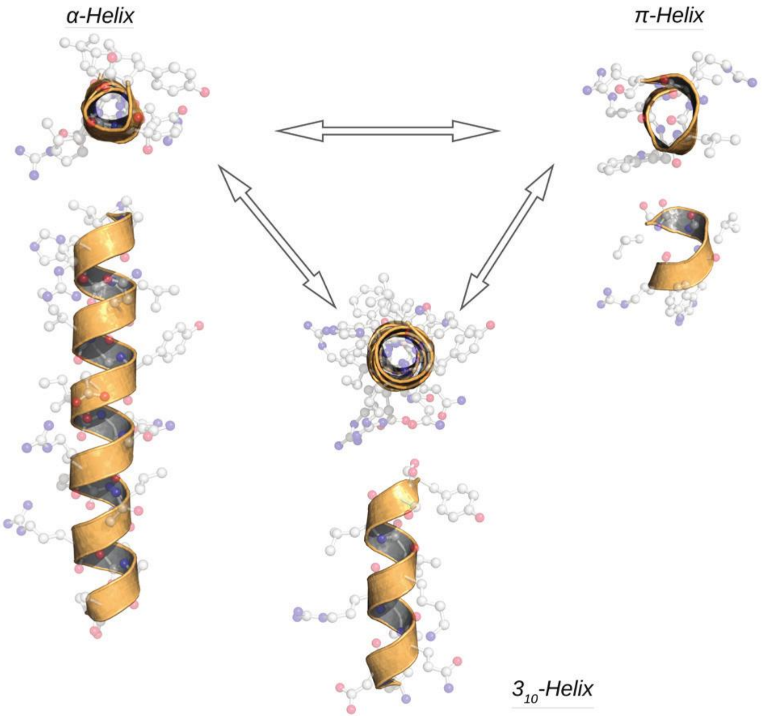


Figure 1.1 ***The different type of helices***. *Two views, lateral and dorsal (bottom) are provided for each helix to appreciate the differences in their helical rise, pitch and the helical core.*

1.1.4 *Prediction of helices from amino acid sequences*

The individual studies on different types of helices point out significant differences in the amino acid composition of various helical motifs, which can be exploited for their prediction. For instance, the secondary structure prediction method SSPRO8 performs reasonably well for 8 different states with a prediction accuracy of ~62 - 63% [127,128]. However, although it aims at separate predictions of α-, 3_{10}- and π-helices; the 3_{10}-helix prediction rate is very low and the π-helix is rarely predicted. The latest approach with RaptorX Property performs slightly better for 3_{10}-helix, but remains unsuccessful for the π-helix [127–129].

The rare occurrences of these motifs largely explain the low rates of predictions. It may also arise from the difficulty to assign 3_{10}- and π-helices. Although the hydrogen bonding pattern and other structural parameters are well characterized for 3_{10}- and π-helices yet assignment methods fail to assign. Their failure might be due to the enhanced flexibility profile of these helical structures [130].

1.1.5 *Dynamic relationship between helices: What is known!*

Indeed, a dynamic relationship would exist between the different kinds of helices, for instance between α- and π-helices as shown in [131]. Importantly, 3_{10}-helices and to a lesser extent π-helices, have been proposed to be intermediates in the folding / unfolding of α-helices [101–103]. However, such studies are often based on model systems like polyalanine peptides and use molecular dynamics (MD) simulations to inspect the effect of chain-lengths and N-terminus residues in α-helix folding [132]. Unfortunately, flexibility profiles and putative interconversion between helical states have never been conducted for a large set of protein structures.

Therefore, it was decided to conduct the first large scale MD simulation study from a large number of structural folds. The underlying motivation being to catalogue the flexibility profile of helices and depict how the helical regions evolve (Fig 1.1). Thus the study provides new insights into the flexibility and deformability of the different helical states, which are an essential component of the structure and function of biological macromolecules. The MD simulations analyse and quantify the stability of helices by considering α-helices as well as 3_{10}- and π-helices.

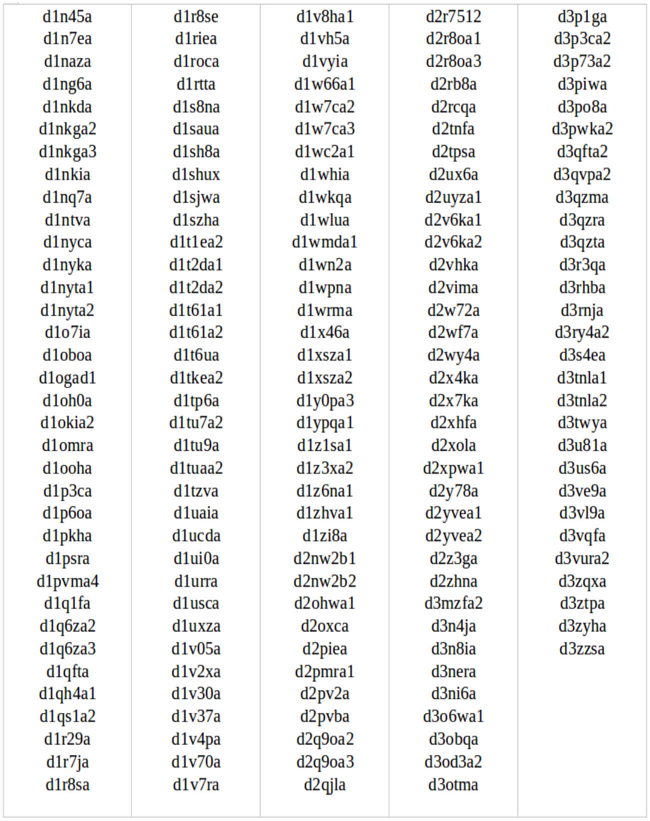
1.2 Methods

1.2.1 *Dataset preparation*

A non redundant dataset at 40% sequence identity was extracted using ASTRAL compendium 2.03 [133–135]. It consists of 5580 protein chains resulting from 4432 PDB files. By filtering on chain lengths between a range of 50-250 residues, resolution better than 1.5 Å and excluding chains with any discontinuity in position numbering, missing residues, modified and/or incomplete residues; only 169 domains were selected. Only globular proteins were used. An in-house parser was used to filter out and fetch the information as implemented in earlier publications from the lab [136]. The selected 169 SCOPids are provided in the Table 1.1.

The 169 domains represent an equilibrated repartition among the different SCOP classes: all-α represent 18.9% of the chains, all-β 29.6%, α/β 24.8% and 26.7% represent α+β class. These SCOP domains belong to 155 X-ray structures in PDB.

Table 1.1 ***SCOP ids of the final selected 169 domains***. *The first four columns contain 35 SCOP ids each while the last one contains 29 entries. The 'd' (first character) signifies that the given structure is a domain. While the following four characters denote the PDB ids of which the domain is a part of. The 6th character in the string is the chain ID of the PDB file. The last or the 7th character if present, signifies the alternate structure form of the domain or isoforms.*



d1n45a	d1r8se	d1v8ha1	d2r7512	d3p1ga
d1n7ea	d1riea	d1vh5a	d2r8oa1	d3p3ca2
d1naza	d1roca	d1vyia	d2r8oa3	d3p73a2
d1ng6a	d1rtta	d1w66a1	d2rb8a	d3piwa
d1nkda	d1s8na	d1w7ca2	d2rcqa	d3po8a
d1nkga2	d1saua	d1w7ca3	d2tnfa	d3pwka2
d1nkga3	d1sh8a	d1wc2a1	d2tpsa	d3qfta2
d1nkia	d1shux	d1whia	d2ux6a	d3qvpa2
d1nq7a	d1sjwa	d1wkqa	d2uyza1	d3qzma
d1ntva	d1szha	d1wlua	d2v6ka1	d3qzra
d1nyca	d1t1ea2	d1wmda1	d2v6ka2	d3qzta
d1nyka	d1t2da1	d1wn2a	d2vhka	d3r3qa
d1nyta1	d1t2da2	d1wpna	d2vima	d3rhba
d1nyta2	d1t61a1	d1wrma	d2w72a	d3rnja
d1o7ia	d1t61a2	d1x46a	d2wf7a	d3ry4a2
d1oboa	d1t6ua	d1xsza1	d2wy4a	d3s4ea
d1ogad1	d1tkea2	d1xsza2	d2x4ka	d3tnla1
d1oh0a	d1tp6a	d1y0pa3	d2x7ka	d3tnla2
d1okia2	d1tu7a2	d1ypqa1	d2xhfa	d3twya
d1omra	d1tu9a	d1z1sa1	d2xola	d3u81a
d1ooha	d1tuaa2	d1z3xa2	d2xpwa1	d3us6a
d1p3ca	d1tzva	d1z6na1	d2y78a	d3ve9a
d1p6oa	d1uaia	d1zhva1	d2yvea1	d3vl9a
d1pkha	d1ucda	d1zi8a	d2yvea2	d3vqfa
d1psra	d1ui0a	d2nw2b1	d2z3ga	d3vura2
d1pvma4	d1urra	d2nw2b2	d2zhna	d3zqxa
d1q1fa	d1usca	d2ohwa1	d3mzfa2	d3ztpa
d1q6za2	d1uxza	d2oxca	d3n4ja	d3zyha
d1q6za3	d1v05a	d2piea	d3n8ia	d3zzsa
d1qfta	d1v2xa	d2pmra1	d3nera	
d1qh4a1	d1v30a	d2pv2a	d3ni6a	
d1qs1a2	d1v37a	d2pvba	d3o6wa1	
d1r29a	d1v4pa	d2q9oa2	d3obqa	
d1r7ja	d1v70a	d2q9oa3	d3od3a2	
d1r8sa	d1v7ra	d2qjla	d3otma	

1.2.2 *Protocol for MD simulations*

Three independent MD simulations of 50 ns each were performed for all protein structures with GROMACS 4.5.7 software [137], using AMBER99sb force field [138]. Thus generating a collective simulation time of 150 (3*50) ns. Each protein structure was immersed in a periodic dodecahedron box using TIP3P water molecules and neutralized with Na^+ or Cl^- counter-ions. The system was then energetically minimized with a steepest-descent algorithm for 2000 steps. The MD simulations were performed in isothermal-isobaric thermodynamics ensemble (NPT) with temperature fixed at 300 K and pressure at 1 bar. A short run of 1 ns was performed to equilibrate the system using the Berendsen algorithm for temperature and pressure control [139]. The coupling time constants were equal to 0.1 ps for each physical parameter. A production step of 50 ns was done using the Parrinello-Rahman algorithm [140]for temperature and pressure control, with coupling constants of T= 0.1 ps and P= 4 ps. All bond lengths were constrained with the LINCS algorithm [141], which allowed an integration step of 2 fs. The PME algorithm [142] was applied for long-range electrostatic interactions using a cut-off of 1 nm for nonbonded interactions.

This protocol was applied to each of the 169 protein domains. From each MD simulation, the conservation of the secondary structures was observed and the structural deviation of each snapshot from the initial structure was measured. Conformations were saved after every ps. For each MD simulation, the secondary structures were analyzed and the structural deviation of each snapshot from the initial structure was measured. Trajectory analyses were performed with the GROMACS v4.6.5, in-house Python and R scripts. Root mean square deviations (RMSD) and root mean square fluctuations (RMSF) were computed on Cα atoms. Normalized RMSFs and normalized B-factors were computed as in Bornot *et al*, 2011 [143].

1.2.3 *Analysis of the local protein conformation*

Secondary structure assignment was performed using DSSP version 2.2.1 with default parameters. DSSP assigns secondary structures as a 7 state model based on intra-hydrogen bonding pattern. The 7 states are represented as ‘H’- α-helix, ‘G’- 3_{10}-helix, ‘I’- π-helix, ‘S’- bend, ‘T’- hydrogen bonded turn, ‘B’- β-bridge, and ‘E’- extended β-strand. In contrast to the previous version of DSSP (cmbi version, 2000) or as termed by the authors as DSSPold, the irregular or coil or loops are marked with a blank in the output. Thus reducing 8-state assignment by DSSPold to 7-state assignment. Since loops or coils are highly flexible structures the blank spaces in the DSSP output were replaced by ‘C’.

It should be noted that the Gromacs v4.5.7 that was used for the molecular dynamics, support the DSSPold developed by Kabsch and Sander in 1983. Since, 25 years, a prominent error in the judgement of hydrogen bonding pattern lead to the misassignment of π-helices as H or T [144]. Therefore, the new version of DSSP (v2.2.1) was used on each frame from the MD trajectory to avoid any errors that may have been induced due to the use of DSSPold.

Protein Blocks (PBs) were also assigned to the same number of frames and *Neq* was used to analyse the behavior of PBs throughout the simulations. Detailed discussion of the methodology and results from PB analysis will be discussed in Chapter 2.

1.2.4 *Clustering approach*

In the initial state, i.e the input structure, each residue is associated to one of the 8 defined secondary states assigned by DSSP [51]. Post simulation, the states for each residue is again assigned using DSSP. Hence, each residue is associated to a vector of size, S=8 representing the 8 defined secondary states and more specifically, the occurrence of each observed state. To define common behaviors between residues, a *k-means* clustering approach was used [145].

At first a subset is created that represents all the residues that were associated to a particular state before MD, e.g. a subset of 3_{10}-helices. Then a fixed number of clusters ‘*k*’ is determined, with *k*=5 (*selected after few tests*). As per the DSSP states, the *k*-clusters are of size S=8. All the data of the subset is then compared to each *k*-cluster and the one with the minimal Euclidean distance is considered the winner. After one cycle (*and after all subsets had been used*), the values of the *k clusters* are modified in order to correspond with the associated observations. The modifications done to a cluster is such that each cluster is the barycenter of the associated observations. After a few cycles, the *k clusters* are stable and can be analyzed for behaviors of different helices.

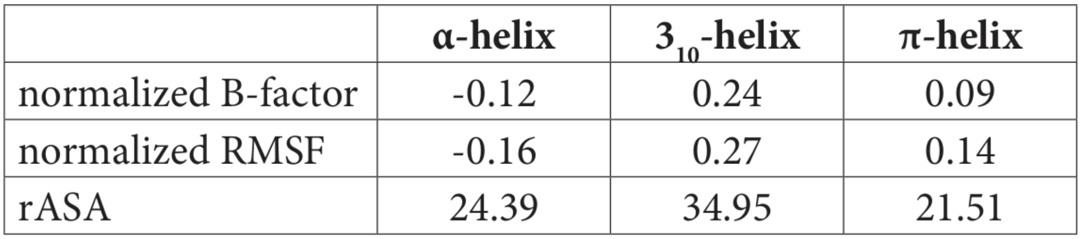
1.3 Results and discussions

1.3.1 Analyses of protein structures

The DSSP assigned 8 states for each frame of all the domains under simulation. In the dataset, the distribution of the helices assigned using DSSP is as follows: 31.5% are assigned as α-helix while 3.99% as 3_{10}-helix and 0.28% as π-helices. This distribution is similar to the distribution observed by Tyagi et al.in 2009 [146]. As shown in Table 1.2, the lowest B-factors are associated to α-helices, an expected feature since α-helices are found most prominently in the ordered state [18].

Interestingly, π-helices are observed in the dataset to be less flexible than 3_{10}-helices with average normalized B-factor values of 0.09 and 0.24, respectively. Both correspond to the flexible region as defined in [18,147,148]. This tendency is correlated with the relative accessibility of the residues computed by DSSP, a higher accessibility being observed for 3_{10}-helices than for π-helices (Table 1.2 and Figure 1.2, row 2).

Table 1.2 ***Behaviors of helices.*** *Average normalized B-factors (from X-ray structures), average normalized RMSF (from the MDs) and the average relative accessibility surface area (for X-ray structures) of α-, 3 10 - and π-helices are presented.*



	α-helix	**3_{10}-helix**	**π-helix**
normalized B-factor	-0.12	0.24	0.09
normalized RMSF	-0.16	0.27	0.14
rASA	24.39	34.95	21.51

1.3.2 Analyses of molecular dynamics

The distribution of normalized B-factors (Fig 1.3A) and normalized RMSF (Fig 1.3B) is highly similar to the distribution observed in a previous studies from our lab, performed with a smaller dataset [143,149]. Figure 1.3C shows the correlation between normalized B-factor values and normalized RMSF (Pearson's coefficient $r = 0.43$). The correct correlation is also very close to the one previously observed by Bornot *et al*, 2011 [143].

Interestingly, 60.2% of the positions do not change at all. Thus no local deformability is observed as is also reflected with an *Neq* value of 1.0. Furthermore, the behavior observed with B-factors is confirmed with RMSF analysis. The most rigid helical structures are α-helices while π-helices appears more flexible but 3_{10}-helices are observed to be the most flexible ones (refer to Table 1.2).

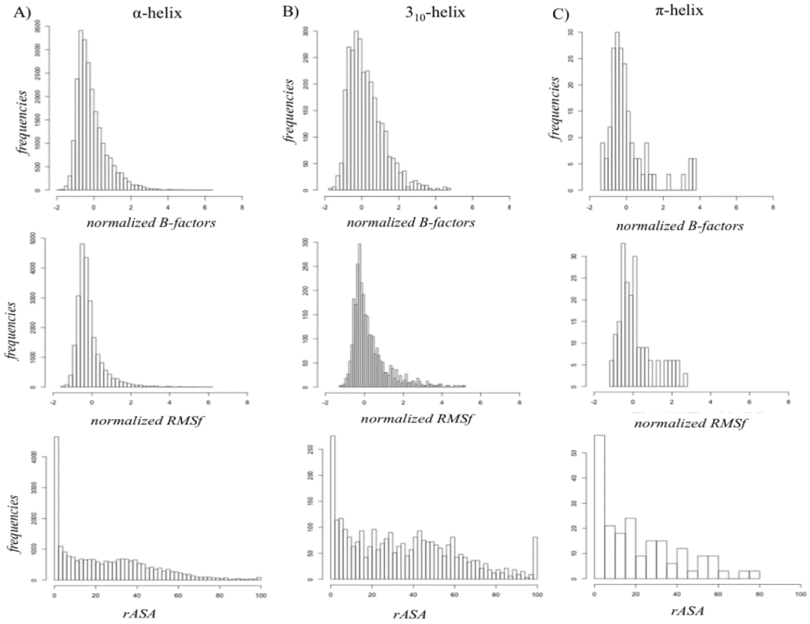


Figure 1.2 ***Normalized B-factors, RMSf, and rASA for α-, 3_{10}- , and π-helix.*** *The 3 x 3 matrix shows individual plots of normalized B-factors (Row1), normalized RMSf (Row2), and relative Solvent Accessibility (row 3) for α-helix (col1), 3_{10}-helix (col2) and π-helices (col3). The values in Table 1.2 are calculated from these plots.*

1.3.3 Helical persistence during simulation

Based on the frequency of the initial DSSP state during the dynamics, perseverance of a state can be estimated. Estimation of perseverance can answer questions such as, how many times an initially assigned 'H' persisted as an α-helix during the simulation of 150 ns and how many times does it changes its conformation? However, it does not provide any details about the changed state.

1.3.3.1 α-helix

DSSP α-helix state represents nearly 30% of the residues in the complete dataset of 169 domains. Of those residues (Fig 1.4A), 31.5% always remain as α-helical during the entire simulation. However, 91.4% maintains an α-helical state for more than 50% of the simulation time, while only 3.9% remain as an α-helix for less than 25% of the time. These statistics illustrate the very stable behavior of the α-helix.

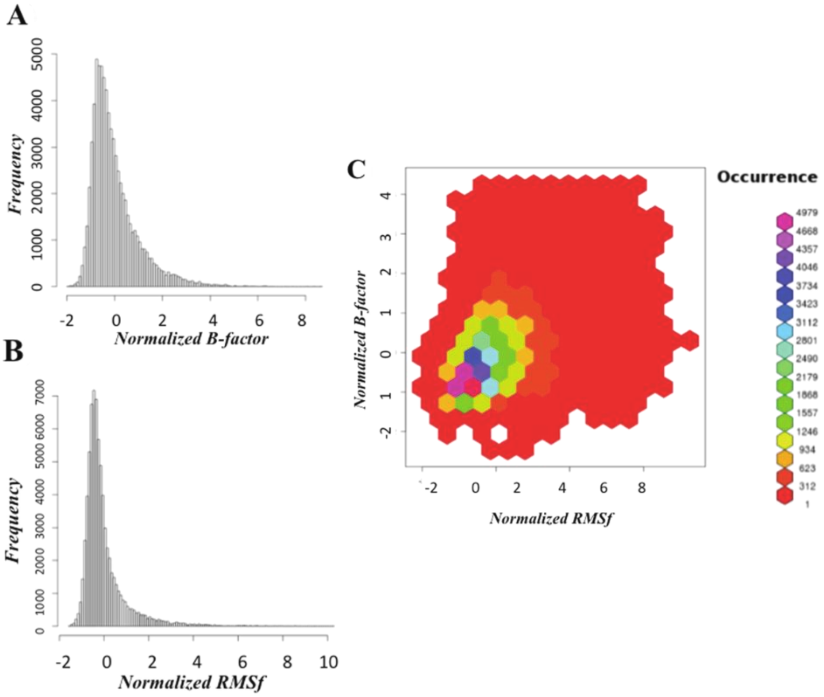


Figure 1.3 ***Normalized B-factors and RMSf behaviors on the whole dataset.*** *A) – Normalized B-factor distribution; B) – Normalized RMSf distribution; C) Correlation between normalized B-factor distribution and normalized RMSf.*

1.3.3.2 *3_{10}-helix*

Despite its relative importance, the 3_{10}-helix is observed to be a less stable local structure during simulations in comparison to the α-helix. As can be seen in Figure 1.4B, the tendency of 3_{10}-helix residues to remain in the 3_{10}- configuration is very limited. Indeed, no residue was found to retain the 3_{10}-helix conformation for the collective simulation time of 150 ns (Table 1.3). The residues adopting a 3_{10}-helix conformation in the initial structure are 3.9% of the total residues in the dataset. Therefore, the representation of 3_{10}-helix is just 13% of the α-helical representation. Among the residues initially observed in the 3_{10}-helix state, only 15.7% retained initial state for more than 90% of the simulation time. However, 54.1% of the residues were observed more than half of the time as 3_{10}-helix.

1.3.3.3 *π-helix*

π-helix was observed to be an extremely rare state in the initial input structures. It was observed 14 times less than the 3_{10}-helices, i.e 0.02% of the total dataset. They are depicted to be slightly less accessible and with regards to their average B-factor and RMSF values, they are supposedly more stable. However, Figure 1.4C shows that this is not the case. Indeed, only 2.4% of the residues remained as a π-helix more than half of the simulation time and the rest is not stable. More than 97.6% were observed as a π-helix for less than 1/4th of the time.

However, rare nature of π-helix is a serious concern given that the selected 169 structures spanned all the SCOP classes. This observation formed the basis of re-assignment of the structures and MD trajectories using DSSP v2.2.1 as the DSSPold underestimates π-helix.

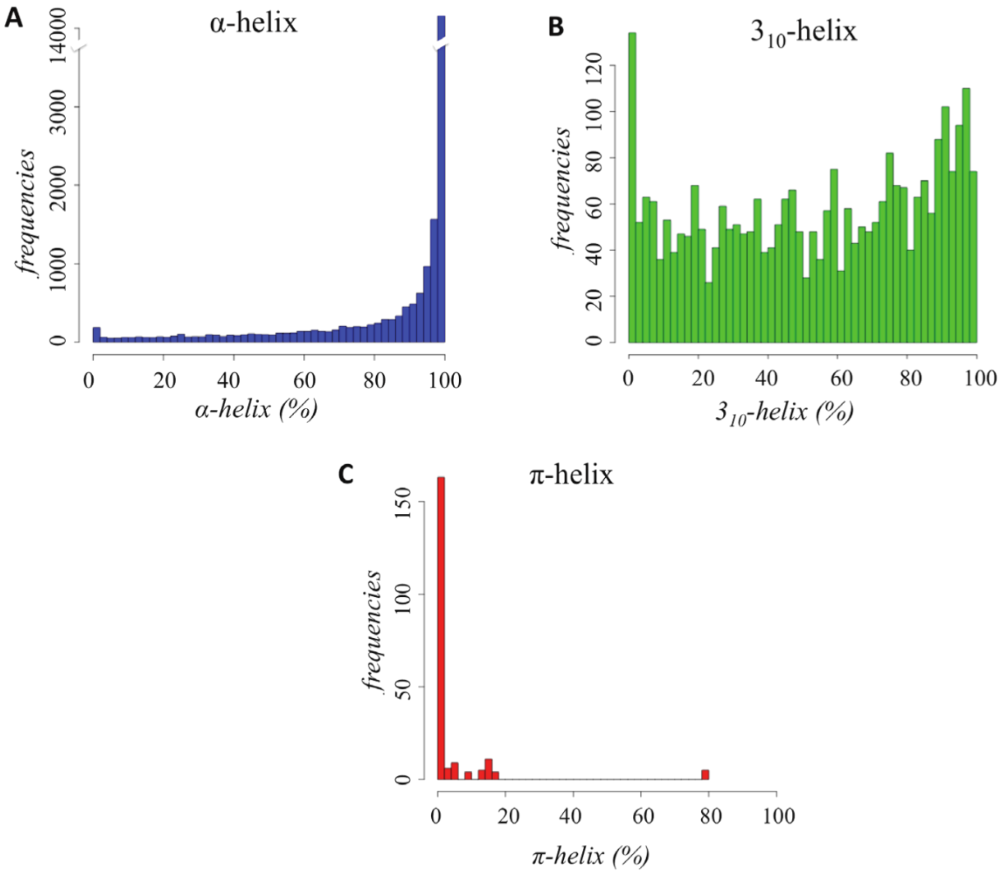


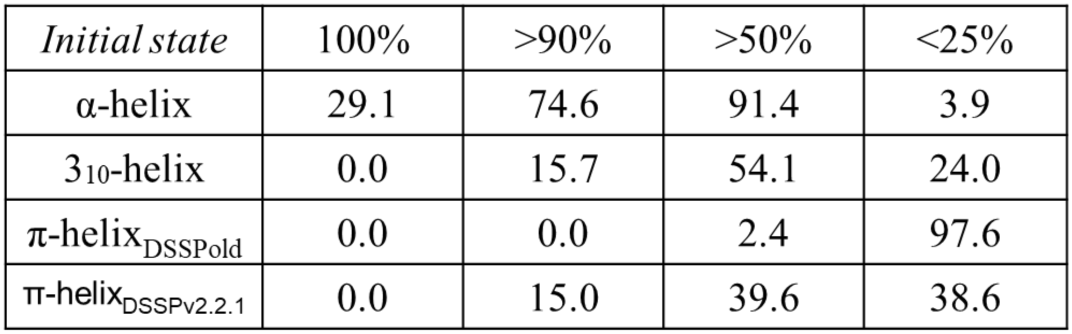
Figure 1.4 ***Persistence of initial helical state.*** *The frequency of residues remaining in the original assigned state of the three types of helices during simulations. A)* α-helix*; B) 3_{10}-helix; C) π-helix.*

1.3.4 *Impact of secondary structure reassignment on the persistence status*

DSSP reassignment had no observable change in 3_{10}-helices (still 3.5%) representation and perseverance frequencies. However, π-helix assignment has a significant 15-fold increase in the initial structures, increasing from 0.02% to 0.32% of the total residues. The increase in π-helix assignment is derived from 2/3rd of the previously assigned α-helices and 1/3rd of turns. As α-helix state has a dominant representation of 30%, the decrease in their representation is non-significant. Thus the DSSP reassignment provides a different view of π-helices. The π-helices are found to be relatively more stable as expected from their B-factor and RMSF analysis.

The updated persistence rates for π-helices are also updated in Table 1.3 as the last row (π-helix$_{DSSPv2.2.1}$). along with their older values for comparison. It is observed that 39.6% of initially assigned π-helices remained as π-helices for more than 50% of the simulation time. Similar to the previous assignments, none of the π-helices remained as π-helices for 100% of the time, however, 15% remained for more than 90% of the time.

Table 1.3 ***Initial state perseverance of each helix.*** *The different columns identify the %age of simulation time, each helix remained in its initial conformation. For e.g. 29.1% of α-helices remain as α-helix during the whole simulation and 91.4% of initial α-helices remains as α-helix for more than 50% of the simulation time. The last column, < 25% defines the persistence of an initial confirmation for less than 1/4th of the simulation time. Therefore, a large value in this column clearly signifies high flexibility or deformability. Please note that the data was reassigned using new version of DSSP and major changes were observed only in π-helices, as depicted by rows 3 and 4 in the table.*



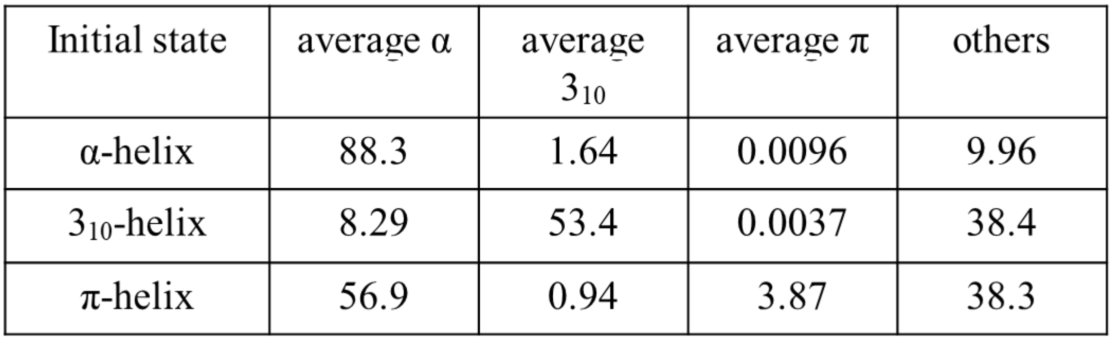
Initial state	100%	>90%	>50%	<25%
α-helix	29.1	74.6	91.4	3.9
3_{10}-helix	0.0	15.7	54.1	24.0
π-helix$_{DSSPold}$	0.0	0.0	2.4	97.6
π-helix$_{DSSPv2.2.1}$	0.0	15.0	39.6	38.6

1.3.5 *Conformational exchanges during simulations*

So far the frequency to retain the initial DSSP state is analysed. It is observed that the initial state is not preserved during the entire simulation, except for α-helix (H). Nonetheless the three helices change their initial state for more than 50% of the simulation time. Therefore, another important question arises: which conformation do they transform into? Do they preferentially explore the conformational space of other helical conformations or the non-helical ones?

Table 1.4 depicts the exchange rates between helical as well as non-helical states. For example, ~10% of times an α-helix adopts a non-helical state. Among the helical states it remains an α-helix for 88.3% of times thus clearly establishing α-helix as the favored conformation. While, an α- to 3_{10}-helix transformation happens at 1.64%, the change from α- to π-helix is negligible. However, 3_{10}-helix transforms to α-helix in 8.29% of cases while retains a 3_{10}-helix conformation for 53.4% of the cases. It shows significant transformations to non-helical states. π-helix in contrast to 3_{10}- prefers α-helical conformation (~57%) than retaining π-helix (3.87% of the times). Apart from α-helix conformation, π-helix to non-helical transformations are significant at 38.3%. Collectively, the helical states transform to non-helical states at 28.8% of times. However, 61.72% of the cases that have 3_{10}-helix as initial conformation adapts a helical conformation while π-helix and α-helix initial conformations stays in the helical fold for 61.7% and 90% of times respectively. Therefore, indicating that positions that have an initial conformation of a helix will tend to remain as a helix.

Table 1.4 ***Exchange rates of helices expressed as percentages.*** *The table quantifies the Helical and non-helical exchange rates for helices. Most of the α-helix (88.3%) tends to remain as α-helix thus denoting the rigidity associated to it. 3_{10}-helix and π-helix changes to non-helical conformations (including coil), the most. Similar results can be interpreted from the table.*



Initial state	average α	average 3_{10}	average π	others
α-helix	88.3	1.64	0.0096	9.96
3_{10}-helix	8.29	53.4	0.0037	38.4
π-helix	56.9	0.94	3.87	38.3

1.3.6 *k-means Cluster analysis: Dynamic behavior of the helices during simulations*

To understand the extent of transitions among helical states, clustering of the ensemble of conformations was done, based on k-means with *k*=5 clusters for comparative purpose. The clusters are named according to their major DSSP state, with subscript indicating a minor DSSP state (T for turn or C for coil). For example, the cluster α^C will indicate a cluster majorly contains the α helix conformation but also has some coil states as well. The detailed composition of each cluster starting from the α-helix, 3_{10}-helix and π-helix initial state is given in Figures. 1.5, 1.6 and 1.7, respectively.

The classification highlights that a residue initiated from an α-helical state (Fig 1.5) tends to transit preferentially towards a β-turn state. Indeed, apart from the most populated cluster α^1 (76.4% of the residues) that is composed of residues remaining as an α-helix, the second cluster α^2 (11.5% of the residues) reveals a decreased content of α-helices in favor of β-turn states. In clusters α^{T1} and α^{T2} (4.2% and 6.6% of the residues, respectively), apart from a small subcluster depicting conformations that switch to the 3_{10}-helix conformation (light blue). This shift causes increase in population of β-turn conformations. The least populated cluster α^C is associated with non repetitive structures (1.1% of the residues). Also, it underlines the correlation between flexibility and the presence of β-turns: the higher the β-turn or coil content, the larger the normalized RMSF (nRMSF). A similar correlation is observed with accessibility and *Neq* values. Clearly, the cluster α^1 represents most of the buried and stable α-helices.

For residues initially assigned as 3_{10}-helices (Fig 1.6), the most populated cluster is 3_{10}. The 3_{10} cluster represents residues that remain in the 3_{10}-helix conformation (40.5% of the residues). It is also the most rigid (both low B-factor and RMSF). From cluster $3_{10}{}^{T1}$ to cluster $3_{10}{}^{T2}$ (25.0 and 17.5% of the residues respectively), the 3_{10}-helix content decreases and the content of β-turns increases. The flexibility increases concomitantly. It also perfectly correlates with the relative accessibility (going from 32.8 to 36.3 and 41.8) and *Neq* values (1.34, 1.49 and 1.59, respectively). The preferred transition to the β-turn was expected as the 3_{10}-helix was shown to overlap by nearly 90% [150]. This is one of the reasons for the disappearance of β-turn type III [151]. The cluster $3_{10}{}^C$ has the highest content of non-repetitive structures and is, as expected, associated with high flexibility. Surprisingly, residues in this cluster are less accessible than those in clusters $3_{10}{}^{T1}$ and $3_{10}{}^{T2}$. Cluster $3_{10}{}^{\alpha}$ (10.5% residues) that represents the transition to α-helical conformation exhibits low accessibility values, the lowest RMSF values compared to clusters $3_{10}{}^C$

, 3_{10}^{T1} and 3_{10}^{T2} , and the lowest *Neq* values,which shows the slightest local conformation change of all the 3_{10} clusters. However, it is associated to high B-factor values.

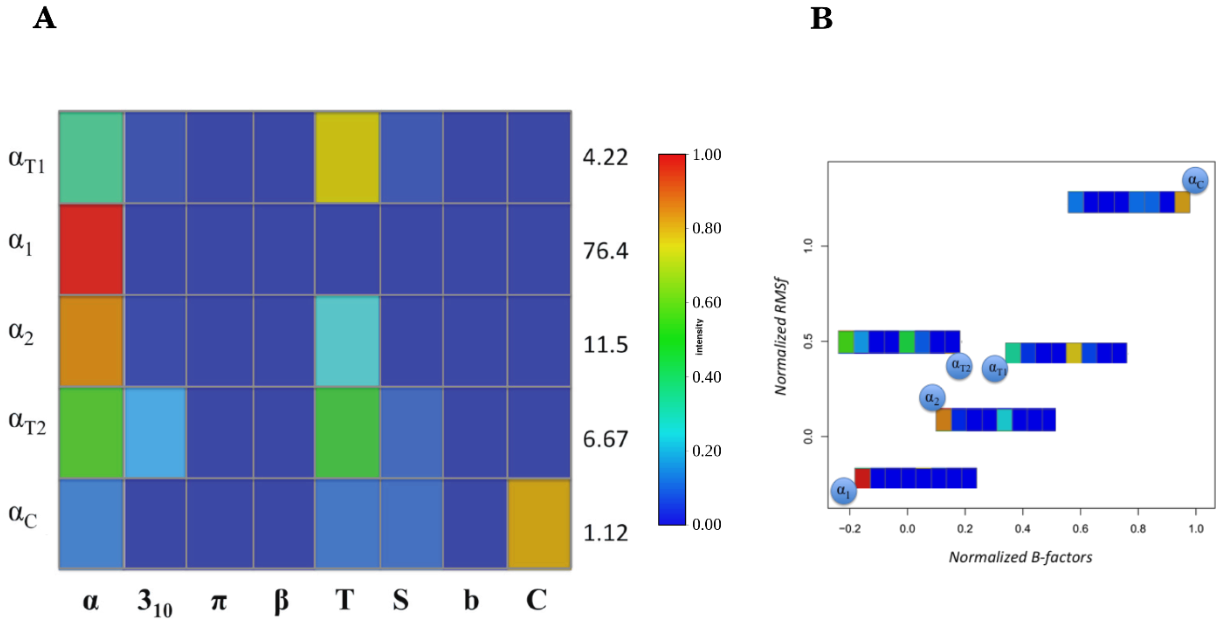


Figure 1.5 ***Different clusters for α-helix.*** *A) Five clusters with a gradient of color are shown (ranging from red 100% to blue 0%). The displayed secondary structures are: α-, 3_{10}- and π-helices, β-strand, turn (T), bend (S), β-bridge (b) and coil (C).*

B) shows the correlation between normalized B-factors and normalized RMSf among different clusters. Extent of flexibility can be estimated from the correlation. For e.g α^{C} and α^{T1} are the most flexible clusters with most of the α-helices transforming to coil and turn conformations, respectively. Similarly, α^{1} is the most rigid cluster with 76.4% of α-helices showing perseverance. (refer to table 1.5)

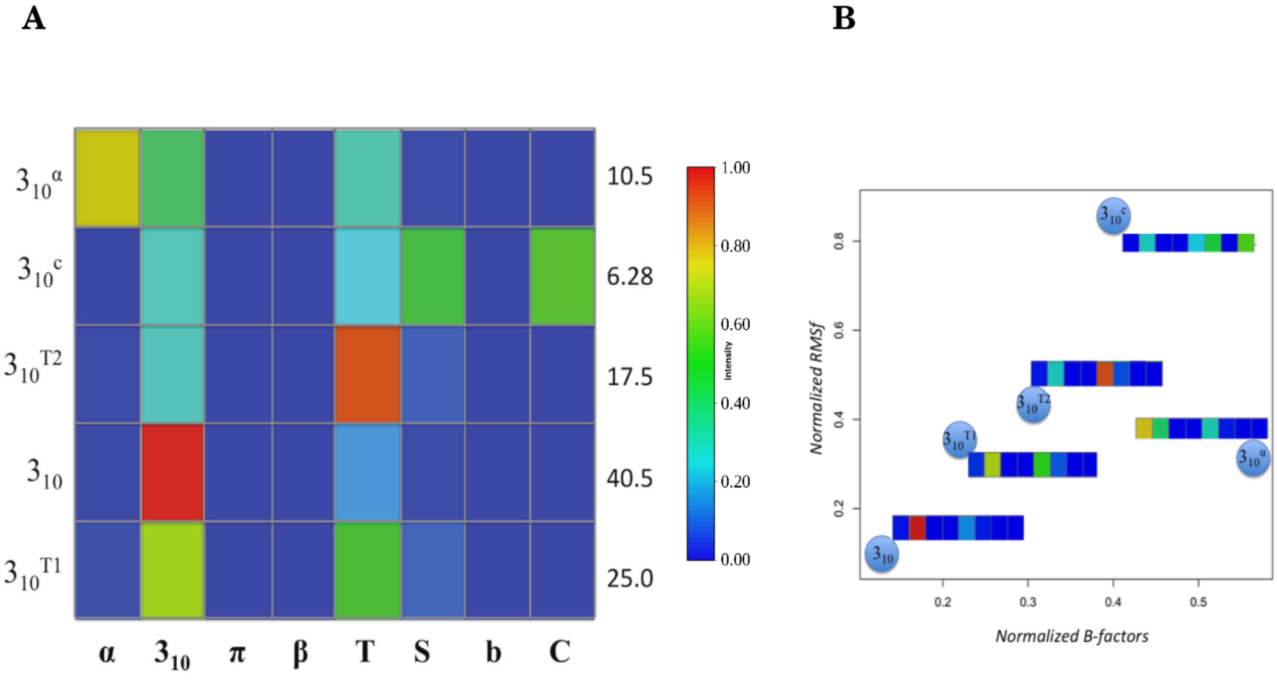


Figure 1.6 ***Different clusters for 3$_{10}$-helix.*** *A) Five clusters with a gradient of color are shown (ranging from red 100% to blue 0%). The displayed secondary structures are: α-, 3$_{10}$- and π-helices, β-strand, turn (T), bend (S), β-bridge (b) and coil (C).*
B) shows the correlation between normalized B-factors and normalized RMSf among different clusters. Extent of flexibility can be estimated from the correlation. For e.g 3$_{10}^{C}$ is the most flexible clusters with most of the 3$_{10}$-helices transforming to coil and bends. Cluster 3$_{10}^{\alpha}$ shows high B-factor value but low RMSf which suggests that transition from 3$_{10}$ to α is not the only one dominating the cluster. As evident from the cluster as well as Table 1.5, that transition to Turns also contribute to the dynamics of this cluster.

Figure 1.7 summarizes the dynamic evolution of the rare π-helices. The cluster named π (10.1% of the residues), which showed the highest frequency of π-helices, was also associated with the β-turn, bends and some coil conformations, but not α- or 3$_{10}$-helices. This is, however, a tip of the iceberg of contradictions in this cluster. Cluster π was found to be associated with the lowest crystallographic B-factors and had the highest relative accessibility. During the MD, it had a very flexible behavior, with the highest RMSF values and also the highest *Neq* values observed. The contradictions could be explained with the following three clusters, $\pi^{\alpha1}$, $\pi^{\alpha2}$ and $\pi^{\alpha T}$ that have a higher α-helical content with no π-helix residues and few β-turns. They showed a slight increase in their *Neq* values. Cluster $\pi^{\alpha1}$ represents π-helix residues (38.7%) that had medium B-factor values associated with a higher stability as an α-helix because they are buried compared to other clusters. Clusters π^{T} and $\pi^{\alpha T}$ have intermediate flexibility behaviors, while the most flexible cluster is cluster $\pi^{\alpha 2}$.

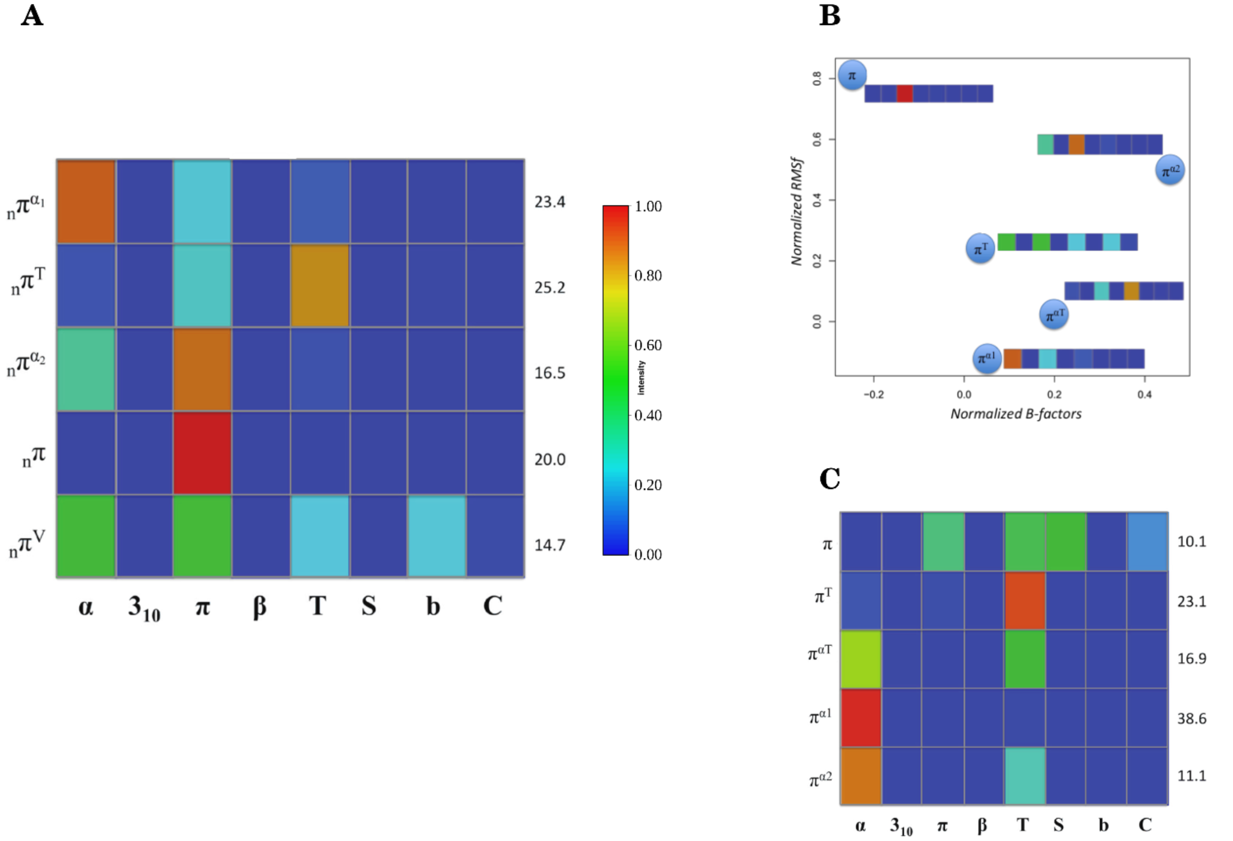
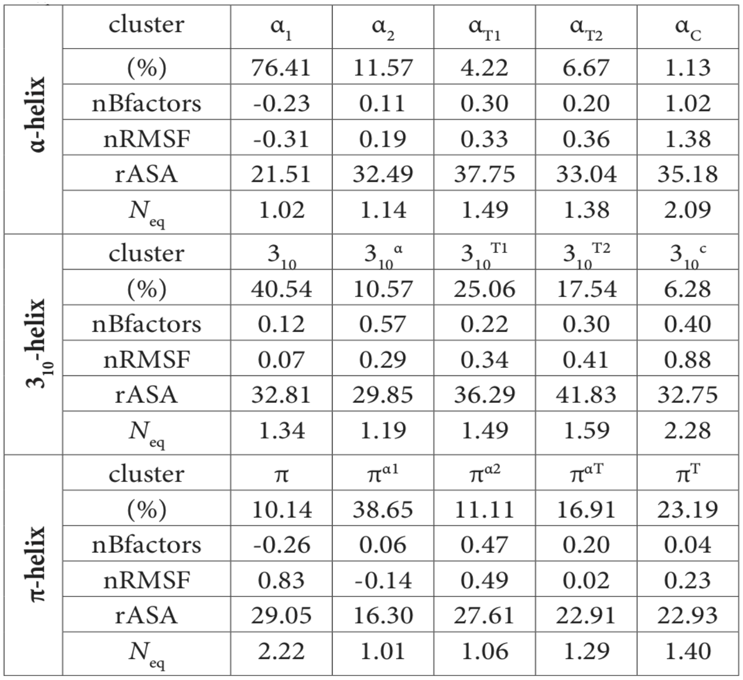


Figure 1.7 ***Different clusters for α-helix.*** *A) Five clusters with a gradient of color are shown (ranging from red 100% to blue 0%). The displayed secondary structures are: α-, 3_{10}- and π-helices, β-strand, turn (T), bend (S), β-bridge (b) and coil (C).*

B) shows the correlation between normalized B-factors and normalized RMSf among different clusters. Extent of flexibility of π-helix can be easily estimated from the correlation, as most of the clusters lie on the right half having higher B-factor and RMSf values. For e.g cluster $_n\pi^T$ has less RMSf but higher B-factor value which suggests that π-helix to Turn transition does not lead to deformability. In contrast, the cluster with π-helices conserved has high RMSf value thus showing the inherent flexibility in π-helices. C) The cluster matrix for the assignments done using $DSSP_{old}$. As can be observed that almost none of the clusters had π-helix representation.

Table 1.5 ***Analysis of different clusters***. *Shown are each cluster, its occurrence, the average normalized B-factors (nBfactors), the average normalized RMSF (nRMSF), the average relative accessibility solvent area (rASA) and the average number of equivalent (Neq) expressed as percentages.*



	cluster	α_1	α_2	α_{T1}	α_{T2}	α_C
α-helix	(%)	76.41	11.57	4.22	6.67	1.13
	nBfactors	-0.23	0.11	0.30	0.20	1.02
	nRMSF	-0.31	0.19	0.33	0.36	1.38
	rASA	21.51	32.49	37.75	33.04	35.18
	N_{eq}	1.02	1.14	1.49	1.38	2.09
3_{10}-helix	cluster	3_{10}	3_{10}^{α}	3_{10}^{T1}	3_{10}^{T2}	3_{10}^{c}
	(%)	40.54	10.57	25.06	17.54	6.28
	nBfactors	0.12	0.57	0.22	0.30	0.40
	nRMSF	0.07	0.29	0.34	0.41	0.88
	rASA	32.81	29.85	36.29	41.83	32.75
	N_{eq}	1.34	1.19	1.49	1.59	2.28
π-helix	cluster	π	$\pi^{\alpha1}$	$\pi^{\alpha2}$	$\pi^{\alpha T}$	π^{T}
	(%)	10.14	38.65	11.11	16.91	23.19
	nBfactors	-0.26	0.06	0.47	0.20	0.04
	nRMSF	0.83	-0.14	0.49	0.02	0.23
	rASA	29.05	16.30	27.61	22.91	22.93
	N_{eq}	2.22	1.01	1.06	1.29	1.40

1.3.7 Impact of secondary structure reassignment on the exchange rates and dynamics

For the reasons described in section 1.3.3 and occurrence of many contradictions in the π-cluster, it was required to reassign structures using DSSP v2.2.1. As expected, the major changes are seen in exchange rates of π-helices and cluster π while clusters α- and 3_{10} remains largely unaffected.

The exchange rates of π-helices are 28.5% to α-helices, 1.43% to 3_{10}-helices, and remains as π-helices 42.2% of the times. Therefore, the exchange rates vary largely from the DSSPold assignments where 56.9% of π-helices transformed to α-helix and only 3.87 remained as π-helices. Also, ~27.7% of π-helices transformed to non-helical conformations which shows a decrease of 10.6% from DSSPold non-helical transformations. This clearly indicates that the 38% of increase in π-helices is contributed by 50% decrease in α-helices conversions.

After reassignment, a pure π-helix cluster was found (named $_{n}\pi$) representing 20% of the residues. Figure 1.7A shows the new clustering for these π-helices. It is totally different from the previous cluster π, which was a mix of β-turn, bends and some coil conformations. Two other clusters ($_{n}\pi^{\alpha1}$ and $_{n}\pi^{\alpha2}$) are also found associated with the transition to α-helices; they represent

23.4% and 16.5%, while cluster $_n\pi^T$ is mainly associated with turns (25.5%). These three share common features with previous clusters $\pi^{\alpha1}$, $\pi^{\alpha2}$ and $\pi^{T,}$ but the proportion of π-helix residues in these clusters is drastically higher. The only fuzzy cluster is cluster $_n\pi^V$, which is a mix of α-helix, π-helix, β-turn, and β-bridge (14.7%). Therefore, the results after reassignment shows that the previous assignment of the π-helix strongly biased our views of this local protein conformation.

1.4 Conclusions and future perspectives

In a previous work from the lab on β-bulges, it was shown that one β-bulge from a 15 β-bulge containing structure disappears after 2/5th of the MD simulation and never returns [68]. Thus depicting that, sometimes non-classical conformations associated with a classical repetitive structure can show some unexpected behaviors. This could be the effect of their inherent flexibility. Therefore, current study was designed to understand the dynamic behaviors of repetitive structures at the basic level of structural complexity, i.e the secondary structures. Herein the focus is specifically on the helical structures while the rest states of secondary structures are under analysis.

The first pertinent result is the quantification of the persistence of helical residues in their original local conformation. More than 3/4th of α-helix residues remain in the helical conformation while it decreases to 40.5% for the 3_{10}-helices. Surprisingly, even if π-helices are mostly buried, they are not observed to be stable. The second interesting result on the flexibility and deformability of helical structures is the huge difference between the three types of helix. The α-helix shows good correlation between stability of the α-helical content and (i) the flexibility as seen through B-factors and RMSf, (ii) accessibility of the residues. The *Neq* analysis of α-helix residues depicts that besides persisting as α-helix, they have a higher tendency to assume β-turn conformations than either the 3_{10}- or π-helices. The 3_{10}-helix shows a similar general behavior in 90% of cases. Indeed, correlation is good in terms of flexibility (both crystallographic and *in silico*), accessibility (with the exception of cluster 3_{10}^C) and *Neq* values.

Nonetheless, the 3_{10}-helix that transforms to the α-helix conformation shows different characteristics. It retains higher B-factor and RMSF values than the average of cluster 3_{10} but is associated with lower accessibility and lower *Neq*. The cluster 3_{10}, seems to have the dynamic characteristics of a local protein conformation that can adopt an α-helical conformation [152].

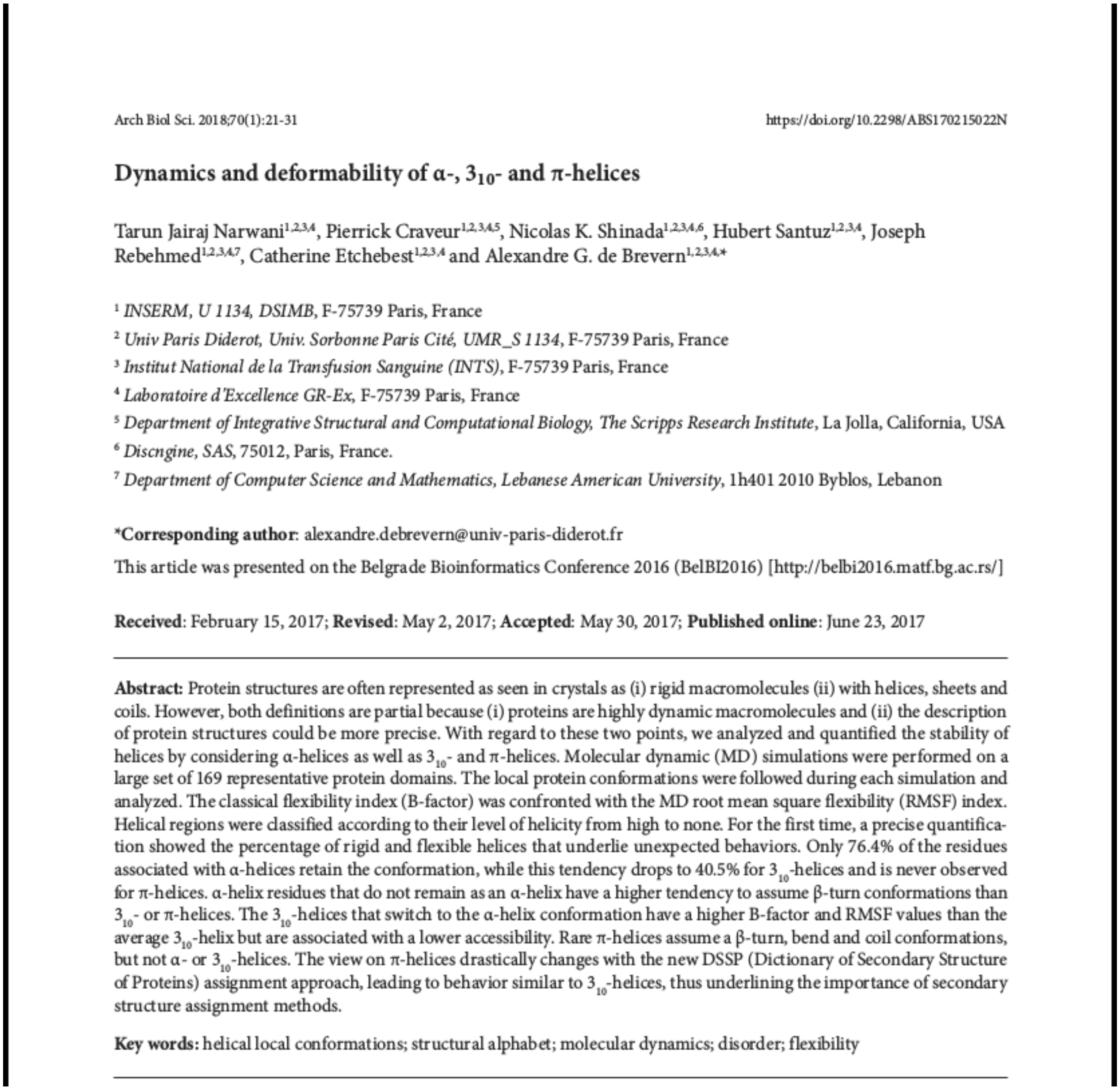
Using classical DSSP (cmbi version), the π-helices cannot be described as stable and therefore new DSSP version 2.2.1 is used to reassign all the initial structures and trajectory frames. The residues that stayed mostly associated with the π-helix conformation are also associated with β-turn, bend and coil conformations, but never α- or 3_{10}-helices. A counterintuitive finding is that they are also associated with low B-factors but due to the high accessibility, they are very flexible/deformable thus showing the highest RMSF and *Neq* values. The other residues lose their initial π-helix conformation and mainly assume an α-helical conformation or to a lesser extent a β-turn conformation. Such dynamic behavior of π-helices with low B-factors and high accessibility may be characteristics of post nucleation, cooperative protein folding effect. Also, it was recently shown that π-helices help the protein chain to fold properly and also in helix packing. They facilitate favorable non bonded interactions by positioning the functionally important helical residues in the correct orientation [127]. Therefore, it indeed becomes fitting for π-helices to exhibit such dramatic flexibility in their dynamics.

For this analysis, 169 protein chains were selected with a limited redundancy from SCOP. They are of high quality. However, one must not downplay the fact that crystallization also produces some crystal contact packing effects, and these were found in a limited number of cases [40,41]. The crystal packing might have an effect on the initial assignment but during molecular simulations, it did not have a significant effect. It is important to properly define the properties of these helical conformations that can have implications in both experimental and computational studies, i.e. analyses of flexibility of protein local conformations, force field parameterization and disorder, etc.

The dynamics of β-strands and related DSSP states are analysed along with more precise local structure estimations using protein blocks. Chapter 2 will contain the details and discussion about their dynamics. Before transitioning to beta-strands, helical component of the protein secondary structures needs to be complete. Therefore, it is highly fitting to discuss about Polyproline-II helices (PPII) that are conventionally ignored by popular assignment softwares like DSSP, Stride, etc. A short discussion dedicated to PPII is provided as a supplementary to chapter1.

Dissemination of results

The results from chapter 1 were published as: Narwani T.J., Craveur P., Shinada N.K., Santuz H., Rebehmed J., Etchebest C., de Brevern A.G. Dynamics and Deformability of α-, 310- and π-Helices. Archives of Biological Sciences (2018) 70(1):21-31.



Arch Biol Sci. 2018;70(1):21-31 https://doi.org/10.2298/ABS170215022N

Dynamics and deformability of α-, 3_{10}- and π-helices

Tarun Jairaj Narwani[1,2,3,4], Pierrick Craveur[1,2,3,4,5], Nicolas K. Shinada[1,2,3,4,6], Hubert Santuz[1,2,3,4], Joseph Rebehmed[1,2,3,4,7], Catherine Etchebest[1,2,3,4] and Alexandre G. de Brevern[1,2,3,4,*]

[1] *INSERM, U 1134, DSIMB*, F-75739 Paris, France
[2] *Univ Paris Diderot, Univ. Sorbonne Paris Cité, UMR_S 1134*, F-75739 Paris, France
[3] *Institut National de la Transfusion Sanguine (INTS)*, F-75739 Paris, France
[4] *Laboratoire d'Excellence GR-Ex*, F-75739 Paris, France
[5] *Department of Integrative Structural and Computational Biology, The Scripps Research Institute*, La Jolla, California, USA
[6] *Discngine, SAS*, 75012, Paris, France.
[7] *Department of Computer Science and Mathematics, Lebanese American University*, 1h401 2010 Byblos, Lebanon

***Corresponding author**: alexandre.debrevern@univ-paris-diderot.fr

This article was presented on the Belgrade Bioinformatics Conference 2016 (BelBI2016) [http://belbi2016.matf.bg.ac.rs/]

Received: February 15, 2017; **Revised**: May 2, 2017; **Accepted**: May 30, 2017; **Published online**: June 23, 2017

Abstract: Protein structures are often represented as seen in crystals as (i) rigid macromolecules (ii) with helices, sheets and coils. However, both definitions are partial because (i) proteins are highly dynamic macromolecules and (ii) the description of protein structures could be more precise. With regard to these two points, we analyzed and quantified the stability of helices by considering α-helices as well as 3_{10}- and π-helices. Molecular dynamic (MD) simulations were performed on a large set of 169 representative protein domains. The local protein conformations were followed during each simulation and analyzed. The classical flexibility index (B-factor) was confronted with the MD root mean square flexibility (RMSF) index. Helical regions were classified according to their level of helicity from high to none. For the first time, a precise quantification showed the percentage of rigid and flexible helices that underlie unexpected behaviors. Only 76.4% of the residues associated with α-helices retain the conformation, while this tendency drops to 40.5% for 3_{10}-helices and is never observed for π-helices. α-helix residues that do not remain as an α-helix have a higher tendency to assume β-turn conformations than 3_{10}- or π-helices. The 3_{10}-helices that switch to the α-helix conformation have a higher B-factor and RMSF values than the average 3_{10}-helix but are associated with a lower accessibility. Rare π-helices assume a β-turn, bend and coil conformations, but not α- or 3_{10}-helices. The view on π-helices drastically changes with the new DSSP (Dictionary of Secondary Structure of Proteins) assignment approach, leading to behavior similar to 3_{10}-helices, thus underlining the importance of secondary structure assignment methods.

Key words: helical local conformations; structural alphabet; molecular dynamics; disorder; flexibility

Chapter S1: Recent advances on polyproline II

About half of the globular proteins are composed of regular secondary structures like α-helices, and β-sheets, while the rest are constituted of irregular secondary structures, such as turns or coil conformations. Other regular secondary structures are often ignored, despite their importance in biological processes. Therefore, three-dimensional structure information is usually described as a simple succession of these repetitive structures (see Figure S1.1), connected by “random” coil [98,153]. Helical structures are locally stabilized by hydrogen bond patterns of backbone atoms (between residues i and i+4) [99], while extended structures are also maintained by hydrogen bonds but at longer distances [100]. As shown in previous chapters, the two forms are highly abundant as they represent 1/3rd (helices) and 1/5th (sheets) of the total residues. A third defined state, called β-turns, is characterized by the reversal of polypeptide chain and is stabilized by a hydrogen bond between the first and last residues [22,151,154]. 25% of the total residues are associated with such structures [155]. However, another common repetitive conformation that was characterized before the β-turns in the 1950s, but often forgotten. Such conformations are called Poly-l-proline-II helices II (PPII) [156,157] (see Figure S1.1B).

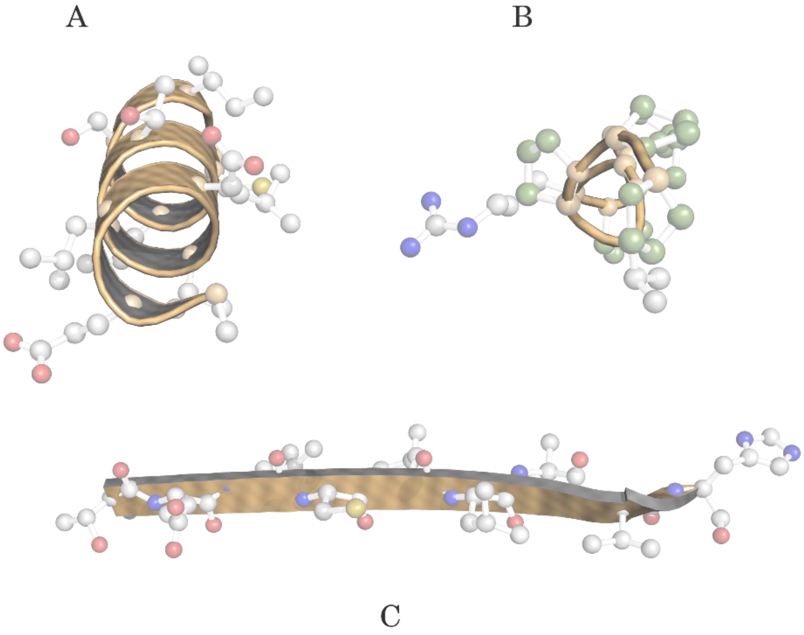


Figure S1.1 ***Structural characteristic of three secondary structures***. *A) Right-handed α-helix, B) left-handed PPII, and C) A β-strand. The cartoon representation highlights the structural geometry, while ball and stick represent the atomic arrangements of the three secondary*

structures. The proline rings can be observed in (B), and the comparison of oxygen (red) and nitrogen (blue) clearly indicates the absence of intra H-bonding in PPII.
In A and C, the close proximity of oxygen and nitrogen atoms makes it favourable for intra H-bonding. High helical rise of the PPII and lack of intra H-bonding make its backbone highly solvent accessible. Images are generated with the PyMOL software.

S1.1 *Introduction to polyproline II helices*

PPII is characterized as a left-handed helical structure with dihedral angles characteristic to that of β-strands and with an overall shape resembling a triangular prism [158,159]. Figure S1.2 shows a comparison of PPII with other local structure helices. The PPII helix has distinct trans-isomers of peptide bonds with dihedral angles of [−75°, +150°]. The rise per residue of PPII helix is 3.1 Å with three residues per turn. Thus, this distinct helical structure rises at 9.3 Å per turn compared to 6.0 Å pitch of a 3_{10}-helix. The primary reason for such open and relatively elongated geometry of PPII is the absence of H-donor atoms due to the cyclic side chain of proline residues. Therefore, the PPII conformation is highly acceptable of H-donor atoms from its environment or third party moieties enhancing its solvation energy. PPII (containing hydroxyproline) is observed commonly in the collagen triple helix and hence was deemed confined to fibrous proteins.

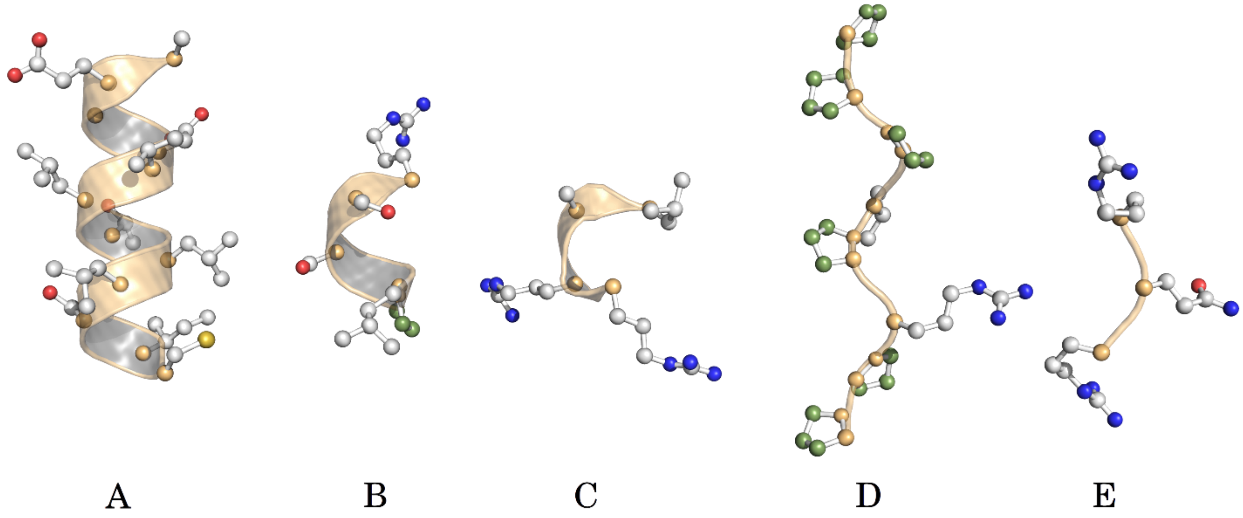


Figure S1.2 ***Orientation and structural organization of the different helices***. *A) α-helix: right handed with a spherical coiling. B) 3_{10}-helix, C) π-helix, and D) Poly-II-proline helix: left handed with a triangular prism coiling. Proline residues are marked in yellow. E) PPII helix with minimum residues possible. Only three residues can adopt a PPII conformation. In this example, none of the*

residue is proline. The proline rings can be observed in (D). High helical rise of the PPII can be clearly seen. Images are generated with the PyMOL software.

Interestingly, PPII stabilizes the collagen triple helix, a conformation that motivated the academic arguments between G.N Ramachandran and F.H Crick eventually leading to the development of the popular Ramachandran map [160–162]. It was found through circular dichroïsm studies that PPII is present in folded proteins and in other structural folding contexts as well [163–165]. Later, Creamer and his team in 2005 demonstrated the existence of PPII in denatured proteins [166], while NMR studies established PPII as a favoured local structure over α-helices in denatured states [167]. Interestingly, the presence of proline residues is not a strict requirement for a PPII and that indeed establishes PPII as a distinct class in secondary structures. Rather, it has been advocated since 1993 to include PPII in mainstream secondary structures, such as α-helices and β-sheets [168]. A striking fact is that residues associated with PPII conformations represent nearly 5% of the total residues in a structure [169], but the lack of popular PPII assignment approaches hinders their systematic analysis.

S1.2 ***Amino acid compositions in PPII helices***

A review article by Adzhubei and Sternberg in 2013 have refreshed the interest in PPII as mainstream secondary structures, such as α-helices and β-sheets. However, it also underlined the non-obligation of the presence of proline residues in PPII [170]. Numerous mutational studies, e.g., SH3 domain—PPII peptide binding analysis provided a desired assertion that PPII conformations are favourable in denatured space [171,172]. Impact of residue level mutations on PPII concludes that PPII conformation is retained even after successive changes of proline with alanine or glycine residues, implying that PPII are not constituted by a succession of proline residues alone. Therefore, PPII should rather be understood as a structural conformation found with different residue propensities in folded and unfolded states. Other experiments further establish PPII as a separate structural class [170,173].

Apart from these studies, restricted coiled library analysis performed by Jha *et al.* in 2005 explores the influence of neighbors on the residues having favourable PPII propensities [174]. Examination of the bias-free coiled library sets reveals dominant PPII conformation for ten of amino acid residues:

Pro, Ala, Met, Glu, Leu, Asn, Cys, Gln, Lys, Gly, and Tyr

Another proposal of similar propensities comes from Cubellis and co-workers who analysed position specific propensities in 5700 PPII helices and classified data with peptide lengths [175]. Thus, residues, such as Ala, Met, Lys, Thr, and Leu, favour PPII conformation in longer peptides, while Asp, Ile, and Glu adopts the conformation in shorter peptides (<3 res). Trp, Phe, and Gly do not favour PPII; however interestingly, Gly is present in a repetitive motif in collagen triple helix, while Trp and Phe have been crystallized in interaction with PPII–hydrophobic motif interactions. Thus, supposedly, these residues could stabilize and mark the terminus of a PPII helix [175]. In the most recent survey, Kumar and Bansal show that 40% of PPII contain no Pro residues at all. Besides, aromatic amino acids are avoided within the helix, while Gly, Asn, and Asp residues are preferred in the proximal flanking regions [176].

Based on hard-sphere Monte Carlo simulations, the propagation of the PPII helix is logically explained by the interaction between the prolyl ring and the backbone (Cβ) of the previous residue. However, this logic breaks when a poly-Alanine adopts a PPII conformation, and therefore, a better explanation could be the neighboring environment and the presence of polar residues. PPII does not have characteristic main chain H-bonding pattern; thus arguably, Ser, Thr, Gln, and other polar residues can stabilize the PPII helix by non-local hydrogen bonding with the backbone [171,175]

Therefore, the overall survey of amino acid propensities reveals that propensities of amino acids in PPII are highly context based. The composition of amino acids seems to deviate according to the presence of PPII in fibrous or globular protein context.

S1.3 *Assignment methods for PPII*

PPII dihedral angles are quite particular. The most classical way to analyse them is to use Ramachandran map (1963), as shown in Figure S1.3. As briefly discussed in thesis introduction section I.1**.**, the map is based on calculations of dihedral angles between the two adjacent planes of protein backbone, hinged at Cα atoms. The dihedral rotation of the planes is restricted by the steric clashes that define the disallowed regions on the map. Therefore, the map is a very powerful tool to assess the stability of a structure based on the local analysis of degrees of freedom for dihedral planes. Further evolution of the map leads to the marking of areas for specific secondary structures, namely, α-helix, β-strands, and later β-turns (see Figure S1.3A). Lately, allowed region for PPII was assigned in the north-western quadrant of the map, allowed for β-strands (see Figure S1.3B). A recent review catalogues the evolution of Ramachandran map very efficiently [177]. It is, however, very distinctive observation that Prof. Ramachandran incepted the idea based on the collagen hydrogen bonding argument [161,178], which arose due to the presence of hydroxyproline [160].

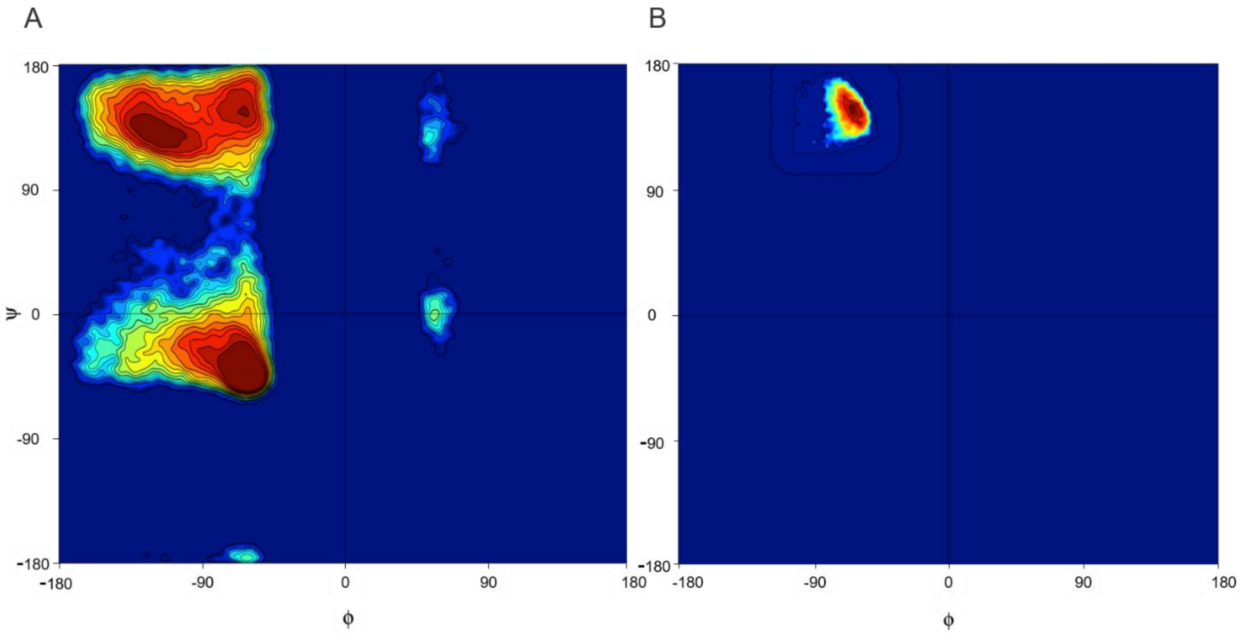


Figure S1.3: Ramachandran Map. *A) From a non-redundant dataset of the Protein Data Bank. B) highlights the allowed region for PPII helix assigned using modified DSSP approach [169]. Visualisation is done with the R software (R CoreTeam 2013).*

More than 20 secondary structure assignment methods (SSAM) had been published in 30 years [51,54,179–198]. They have been defined with various criteria [199]. The most popular SSAM uses backbone hydrogen bonding pattern-based methods. Nonetheless, very few SSAM assigns PPII to the protein coordinates. Only five SSAMs, to be more precise, include the assignment of PPII conformations.

XTLSSTR [187]: The first available approach was XTLSSTR , where a structure is assigned based on a simple approach similar to the visual inspection of secondary structures. It calculates three distances and two angles based on the backbone geometry and then searches for amide–amide interactions. It successfully assigns α-helix, 3_{10}-helix, extended β-strand, hydrogen bonded and non-hydrogen bonded turns, and polyproline (type-II) helices.

SEGNO [182]: makes assignment based on distance and torsion angle calculation. For assigning PPII, it uses dihedral angles between the two-peptide planes separated by one and two residues, respectively, named diheco and diheco2. An important observation is that PPII is assigned when a residue is not defined as β-strand by SEGNO and lies within predefined values of Φ and Ψ angles. Later, taking into account the range of the four diheco angles (220–270 and 100–140), the PPII helical conformation is assigned to the residue. These thresholds are relaxed for the termini of PPII with a minimum length of the helix to be three residues and the overall shape of PPII is deemed to be like a triangular prism.

PROSS [200]: uses the concept of mesostates from a torsional grid for the assignments. The grid is described as the unit squares covering all areas in a Ramachandran map. The grids are of two kinds based on their unit area: smaller unit square: *fine grid* and broader unit square: *coarse grid*. Based on the type, each unit grid is referred to as a coarse/fine mesostate. Therefore, in principle, the Ramachandran plot is converted into a Φ/Ψ grid with marked regions (allowed, favourable, and disallowed) covering more than one mesostates. In a very similar approach related to SEGNO, PROSS also does not directly assign PPII conformation rather resolve it out after β-strand leftovers.

DSSP-PPII [169]: is an extension of DSSP with included dihedral angle parameters for PPII assignment, thus isolating PPII from coils. DSSP-PPII uses dihedral space (Φ and Ψ, −75° and +145°) to define the core of PPII while increasing the space by a metric value, ε, radiating out at 1°. The value of ε is chosen as an equilibrium between the number of amino acids assigned as PPII by the three previous approaches plus, an extra constraint that two consecutive dihedral angles should be assigned as PPII. One of the major features of this method is to use DSSP that is already an established and trusted method for other secondary structure elements (SSE). Therefore, the code can be adapted to apparently any other assignment method, if and when required. A specific database had been proposed to the scientific community [169]

A major caveat to use DSSP-PPII is that the DSSP at its foundation is the old DSSP (cbmi version) and not the corrected version from 2011 [144]. As discussed in section 1.2.3, the old version of DSSP is prejudiced towards π-helices and often assign them into α-helices or β-turns. However, since the script is portable, current efforts are taken into account to adapt the script with new version of DSSP and include PPII helices too in chapter-1 analysis of α-, 3_{10}-, and π-helices.

ASSP [127]: is an extension of helical geometry calculation program, HELANAL-plus (Bansal et al. 2000) that is used to calculate the local helical structure parameters: twist, rise, virtual torsion, and radii. ASSP uses the difference between these parameters calculated over two or more adjacent C_α windows of four residues. Later, in the protocol, the overlaps are resolved based on the established minimum lengths of helices: α- (4), 3_{10}- (3), π (5), and PPII (3). Therefore, PPII conformations are assigned based on the helical geometry of the local region. Since it uses HELANAL, which further is based on Sugeta and Miyazawa, and Shakarji methods for helical geometry, ASSP tends to assign β-sheets with less efficiency [201,202]. Kumar et al, applied ASSP to analyse in detail the PPII [176] and found that near 3/4 of PPIIs occur in conjunction with α-helices and β-strands, and serve as linkers as well. They also underline a large number of CH···OH-bonds.

All these methods are well designed for PPII assignments. However, most of them tend to assign PPII by indirect approaches due to the different bonding patterns of PPII. Unfortunately, the number of PPII assignment approaches is still limited compared to SSAM for other secondary structure elements, and remains a limitation for the use by scientific community.

S1.4 ***Physiological importance of PPII***

A distinct feature of polyproline helices is that unlike other SSE, they do not have intra-hydrogen bonding thus making the backbone as well as the side chains highly solvent accessible. Such conformations would be hankering for finding partners for hydrogen bonding and stabilization. Therefore, the sequence and structural characteristics of PPII make it worth to be probed for partnered interactions. One of the important tools to study the PPII role in protein–protein and DNA–protein interactions is the SH3 domain models. SH3 (Src homology 3) domains are small yet important structural domains involved in cell signalling and regulation, e.g., Tyrosine kinases.

SH3 domains are also well known to interact with PPII conformations [203]. Hence, host-pathogen models designed with SH3 domains are critical to understand the interaction space of PPII conformation with respect to proteins and/or nucleic acids.

Many such studies focusing on signal transduction and cell–cell recognition have been explored for potential PPII–protein and PPII–nucleic acid interactions [204,205]. For instance, C-terminus of Synapsin-I, a protein regulating synaptic vesicle transport in neurons, is a proline-rich region. Synapsin-I interacts with the cytoplasmic polyproline region of membrane protein, vesicle-associated membrane protein 1 (VAMP-I) [205]. Phosphorylation of a serine residue upstream of C-terminus PPII helix regulates the secretion of a synaptic vesicle, while VAMP-I helps in recognition. Similarly, in Ras-GTP signalling pathway, the SH3 domains of the adaptor protein bind to the polyproline region of SoS protein (*xPxxPPPψxPx*) leading to the exchange of GTP. Another set of interactions [206] is in vacuolar sorting, where SH3 domain of phosphatidylinositol-3 kinase binds to the GTP-binding protein.

Structurally, it is acknowledged that the PPII helix-binding region of SH3 domain is a smooth hydrophobic surface flanked by conserved charged residues [206]. The PPII interactions also have a significant structural–functional role in transcription, as many transcription factors have proline-rich terminals [207]. This could also indicate the role of PPII interactions in multimeric complex formation during transcription. A well-characterized case of PPII–protein interaction is the RNA polymerase II (RNApolII). C-terminus of RNApolII has multiple copies of conserved motif *YSPTSPS*, which further is a two-fold *SPXX* motif. *SPXX* is a DNA binding motif found in DNA binding domains [208,209]. Furthermore, Hicks and Hsu (2004) investigated the structural aspects of PPII in DNA binding and recognition [204]. Exemplifying with three DNA interacting proteins viz.; the third K homology domain of NOVA-2 see Figure S1.4 [210], the Epstein–Barr nuclear antigen-1, and the Drosophila paired protein homeodomain, they quantify the binding of PPII to the nucleotides' minor groove and underline the specificity and non-specificity of recognition. The optimal size and specific recognition offered by PPII backbone residues strongly suggest to identify PPII as a nucleic acid binding motif [204].

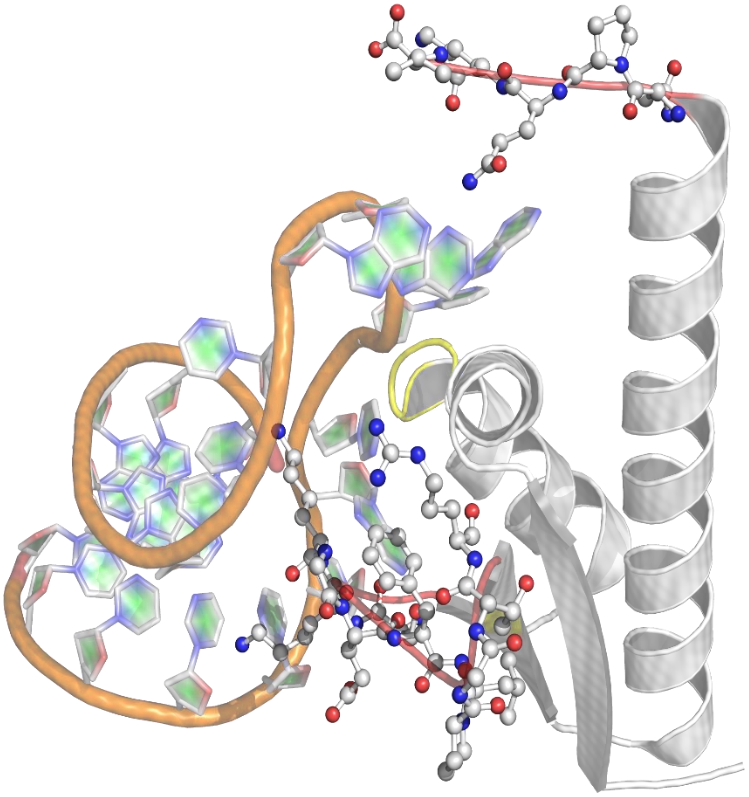


Figure S1.4 ***Interaction of Nova Protein K homology domain with RNA hairpin (PDBid: 1EC6_A [210]).*** *The conserved motif of the variable loop is coloured in yellow. The two PPII helices are coloured in magenta. The occurrence of C-term PPII helix is reported to be the difference between RNA bound and unbound form. Image is generated with the PyMOL software.*

S1.5 ***Pathological roles of PPII***

Role of PPII in protein–protein and DNA–protein interactions, and its role in sorting and transport mechanisms have been investigated. KISS-1 Receptor (KISS1R) in its intracellular domain has three triplets of Proline–Arginine–Arginine (PRR). The addition of a fourth triplet induces the formation of a PPII, and inhibits KISS1R presentation on cell membrane. The retention of KISS1R in cytoplasm ceases the interaction with KISS peptin and thus abolishes the secretion of GnRH leading to Hypogonadotropic hypogonadism [211]. Besides, several studies using ROA (Raman optical activity) and VBD (vibrational circular dichroism) structural visualization techniques confirm the presence of PPII conformation in pathological fibrillar aggregates [163,170,212]. Conversion of PPII to β-sheet conformation in amyloidogenic precursor of human lysozyme may indicate a highly potential role of PPII in numerous amyloid-based conformational disorders [212].

For instance, phosphorylation of a threonine flanked by a PPII in Tau protein leads to the misfolding and aggregation of microtubular proteins in Alzheimer's disease [213]. A similar role of PPII has been found in α-synuclein, responsible for aggregation in Alzheimer's and Parkinson's pathologies [170]. Taken together, this emphasizes a deeper understanding of PPII structural features [214].

S1.6 ***Recent advances in PPII research***

The growing interest in physico-chemical and structural properties of PPII, especially their short extended-helical structure has attracted the attention of pharmaceutical companies. Very recently, cell-penetrating vector approaches are designed based on PPII scaffold [54,215–219]. As discussed in section S1.4, PPII backbone has a high solvent accessibility and thus becomes highly hydrated in solvents. Therefore, use of PPII for cell penetration poses a challenge for hydrating the PPII-based moiety and their convenient uptake in hydrophobic membranes [216]. Chmielewski's group [220,221] addressed this by designing and introducing cationic and hydrophobic moieties on the PPII backbone and observed no structural change. The compactness and inherent flexibility of the PPII conformation is the key to their adaptability and accompanied by cationic and hydrophobic moieties; they become highly suitable for a cell-penetrating vector [222]. A tremendous increase in PPII-based Cell-Penetrating Peptide (CPP) uptake compared to the traditional ones has been reported.

Another important difference is the claimed reduction in toxicity. This is based on the observations that PPII scaffold-based CPP: Sweet ArrowPeptides—SAP(E)—obtain a net negative charge unlike the traditional CPP which are positively charged [216,217].

S1.7 **Summary**: ***To consider PPII as a regular secondary structure***

Polyproline II helix is arguably a distinct member in secondary structure elements, based on its geometry, sequence, and structure. PPII has a left-handed geometry unlike the right-handedness of popular protein helices (see Fig S1.2). Its sequence composition varies based on the presence in a globular or fibrous protein environment. It is quite an interesting observation that proline, a major α-helix breaker or kink inducer, when in succession adapts a distinct helical form itself. Moreover, it dominates the α-helical form in denatured space. Such examples can be appreciated in light of the expanse of the secondary structural space. Although PPII conformation represents only 5% of the conformational space, it is highly recommended for it to be considered in the regular secondary structures. Besides, its representation is equivalent to, if not more than, the 3_{10}-helices. The involvement of PPII–protein and PPII–nucleic acid interactions in different pathologies, structural applications, and drug carriers makes it even more viable candidate to be included in the main regular secondary structures. Its potential role in Alzheimer's and Parkinson's could not be ignored, given recent publications on the subject. The presence of PPII in regular, ordered, and disordered regions while establishes that its distinctiveness is not sufficient to seize the complete structural space of PPII conformations. Therefore, more assignment approaches and coiled library experiments are needed to explore such conformations. Figure S1.5 shows the number of publications about PPII since 1968. The increase is clear, but remains limited. The number of papers had never been higher than 100 papers/per year. Therefore, we can hope for a better representation of PPII among regular secondary structures.

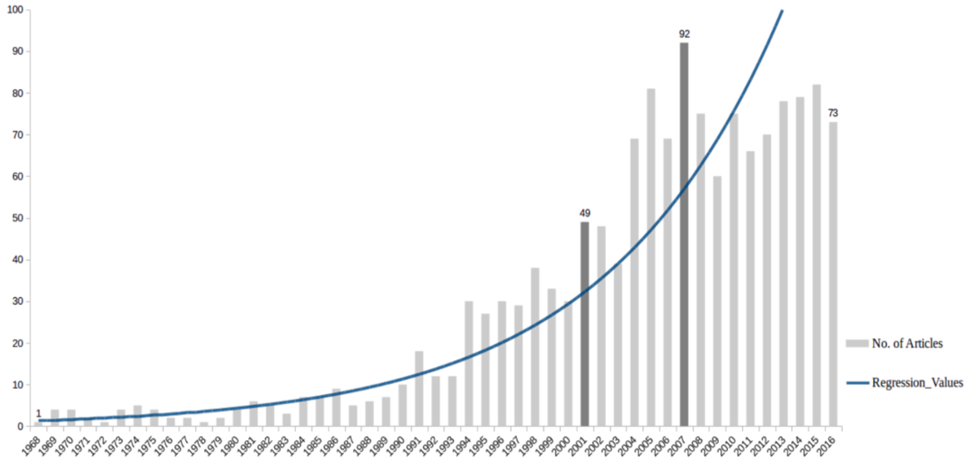
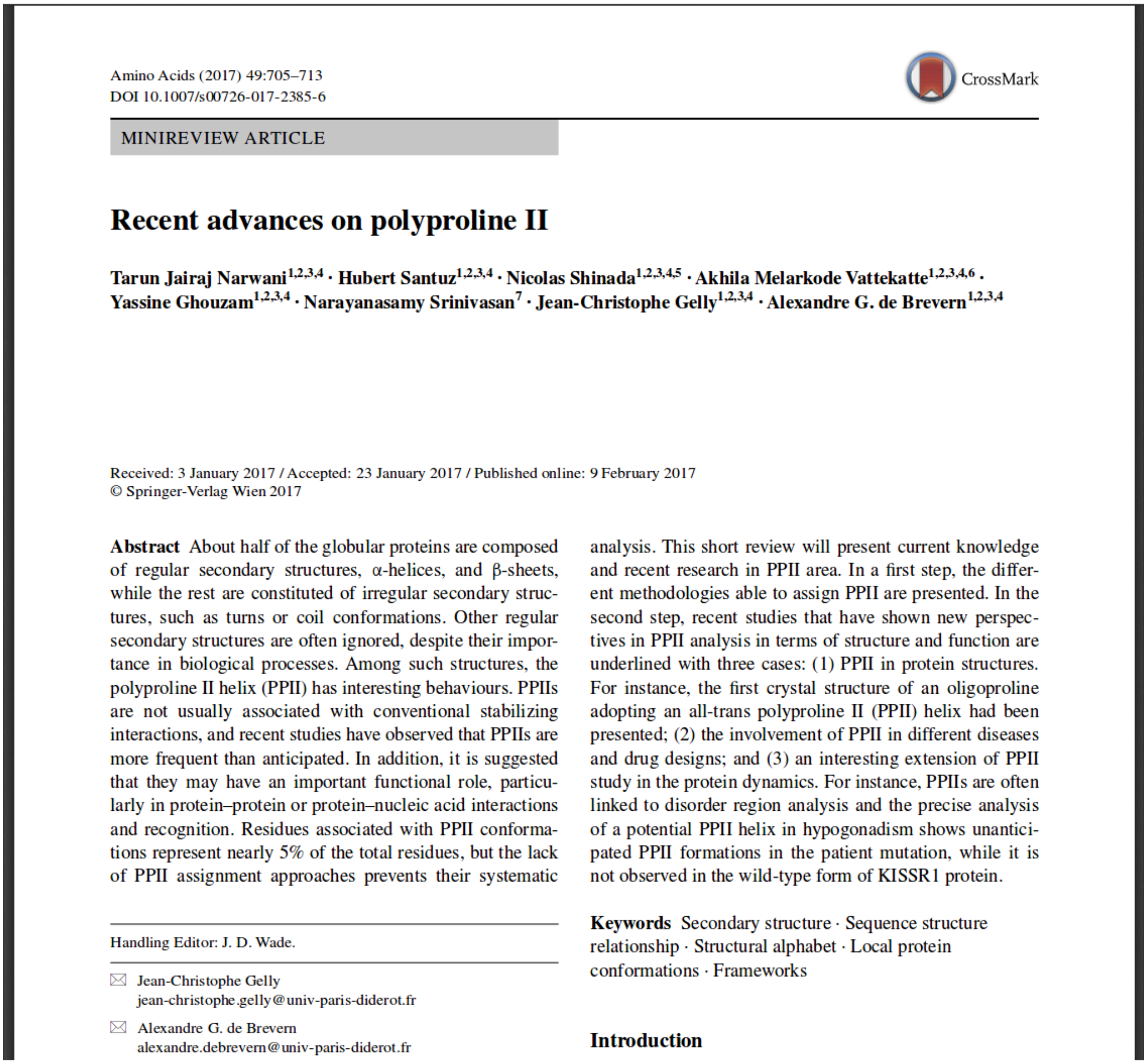


***Figure S1.5. Year wise publications trends on Polyproline II helices.** The bars depict the number of publications corresponding to the year on x-axis. An exponential function is represented in blue curve. Dark towers show the sudden surge in publications compared to the previous year. Visualisation is done with the R software (R Core Team 2013)*

Dissemination of results

The information from chapter S1 was archived as a mini review article: Narwani T.J., Santuz H., Shinada N.K., Melarkode Vattekatte A., Ghouzam Y., Srinivasan N., Gelly J.-C., de Brevern A.G. Recent advances on polyproline II. Amino Acids (2017) 49(4):705-713.



Amino Acids (2017) 49:705–713
DOI 10.1007/s00726-017-2385-6

CrossMark

MINIREVIEW ARTICLE

Recent advances on polyproline II

Tarun Jairaj Narwani[1,2,3,4] · Hubert Santuz[1,2,3,4] · Nicolas Shinada[1,2,3,4,5] · Akhila Melarkode Vattekatte[1,2,3,4,6] · Yassine Ghouzam[1,2,3,4] · Narayanasamy Srinivasan[7] · Jean-Christophe Gelly[1,2,3,4] · Alexandre G. de Brevern[1,2,3,4]

Received: 3 January 2017 / Accepted: 23 January 2017 / Published online: 9 February 2017

Abstract About half of the globular proteins are composed of regular secondary structures, α-helices, and β-sheets, while the rest are constituted of irregular secondary structures, such as turns or coil conformations. Other regular secondary structures are often ignored, despite their importance in biological processes. Among such structures, the polyproline II helix (PPII) has interesting behaviours. PPIIs are not usually associated with conventional stabilizing interactions, and recent studies have observed that PPIIs are more frequent than anticipated. In addition, it is suggested that they may have an important functional role, particularly in protein–protein or protein–nucleic acid interactions and recognition. Residues associated with PPII conformations represent nearly 5% of the total residues, but the lack of PPII assignment approaches prevents their systematic analysis. This short review will present current knowledge and recent research in PPII area. In a first step, the different methodologies able to assign PPII are presented. In the second step, recent studies that have shown new perspectives in PPII analysis in terms of structure and function are underlined with three cases: (1) PPII in protein structures. For instance, the first crystal structure of an oligoproline adopting an all-trans polyproline II (PPII) helix had been presented; (2) the involvement of PPII in different diseases and drug designs; and (3) an interesting extension of PPII study in the protein dynamics. For instance, PPIIs are often linked to disorder region analysis and the precise analysis of a potential PPII helix in hypogonadism shows unanticipated PPII formations in the patient mutation, while it is not observed in the wild-type form of KISSR1 protein.

Keywords Secondary structure · Sequence structure relationship · Structural alphabet · Local protein conformations · Frameworks

Handling Editor: J. D. Wade.

✉ Jean-Christophe Gelly
jean-christophe.gelly@univ-paris-diderot.fr

✉ Alexandre G. de Brevern
alexandre.debrevern@univ-paris-diderot.fr

Introduction

Chapter 2: Understanding local structure behaviors using Protein Blocks

2.1 Introduction

In Chapter 1, a systematic analysis of dynamic behavior of helical structures was performed. Therefore, Chapter 2 extends the similar question to β-strands or rather precisely, non-helical secondary structures which were not focused upon in Chapter 1. Besides using DSSP v2.2.1, Protein Blocks (PB) will also be used to have maximum coverage of the local structural space. Since PB are 5 residue length based abstraction of the protein structure backbone, PBs are expected to provide better insights into the dynamics of local protein structure.

2.1.1 *A brief recap*

The α-helix (or 3.6_{13} helix) and the β-sheet have been extensively analysed since their discovery by Pauling and Corey, 65 years ago [98–100,151]. The secondary structure description of protein structures had led to the development of more than 20 secondary structure assignment methodologies [51,54,179–198], with DSSP being the most popular one [51]. Such descriptions have often been updated from time to time to include other types of secondary structures such as β-turns [22,150,155,186], PolyProline II [168,171,223,224] and loops categorization [225,226]. However, no single secondary structure assignment method (SSAM), have been designed for all types of known secondary structures. Therefore, specialists often have to use different SSAM based on its expertise and then normalize individual results to obtain a meaningful ss assignment. Such an example can be seen in sub-chapter 3b of this thesis.

2.1.2 *Limitations of SSAMs*

Analysis using different SSAM can still have some limitations such as the non-definition or ambiguity of the coil state, some known problems with short repetitive structures and ofcourse, the discrepancies between different algorithms [198,227–229]. Hence, alternative views have been proposed using systematic analysis of all local protein conformations. It has motivated the development of local protein structure libraries that (i) are able to approximate all (or nearly all) local protein structures and (ii) do not take into account the classical secondary structures. Such libraries brought about the categorization of 3D structures, without any *a priori,* into small prototypes that are specific to local folds found in proteins. The complete set of such local structure prototypes defines a structural alphabet [230].

2.1.3 *Why Protein Blocks?*

The precursor research in defining a structural alphabet was carried out by Unger and co-workers [231]. This led to numerous applications, from the analysis of sequence-structure relationship [232] to the prediction of short loops [229], etc. In this context, Protein Blocks have been the most successful structural alphabet [53,233] and following are few studies where PBs are extensively used:

(i) 3D protein backbone description [233], (ii) Local structure prediction [53,58,234], (iii) Description and the prediction of long fragments [64,235–239], (iv) Prediction of short loops [228,229], (v) Analysis of protein contacts [240], (vi) Structural modelling of transmembrane proteins [241,242], (vii) Definition of a reduced amino acid alphabet dedicated to mutation design [243,244], (viii) Protein structure superimposition and comparison [245–247]; (ix) Reconstruction of globular protein structures [248], Peptide designing [249], and (x) Definition of binding site signatures [62]. A recent impressive development concerns the inclusion of PBs in threading approaches [250,251] and especially to an efficient one called ORION [252,253].

Interestingly, PBs show great use for analysis of protein flexibility [18] using molecular dynamics in the specific cases of Integrins [254,255], transmembrane proteins like KISS1R [211] and NMDA Receptor Channel Gate [256]. Therefore, PB was selected as the choice for structural alphabet to analyse the dataset of 169 protein domains (see section 1.2.1).

2.1.4 *Previous studies on local flexibility of protein structures*

The number of large-scale analyses of protein dynamics remain slightly limited. A prominent one is the Dynameomics project [257]. It simulated a representative sample of all globular protein meta folds and focuses on the unfolding process. Although a robust study to probe protein folding, it does not correspond to the analyses of protein flexibility. A database is available and provides visual results [258]. A more related study can be the work of Grubmüller’s group namely Dynasome [259]. They showed that in the 34 different descriptors defined to characterize the simulation of the 112 proteins, only a few Collective Dynasome descriptors describes most of the movement. However, that cannot be defined as a local structural analysis as the Dynasome descriptors define the structures globally.

Thus, it would be worth to see the dynamics of proteins from the local structure perspective that can provide insights on the inherent protein flexibility.

2.2 Methods

All the methods used were same as Chapter 1, given that the same data set was operated upon. However, the clustering method as well as use of PB is different.

2.2.1 *Data set preparation* - Same as section 1.2.1

2.2.2 *Molecular dynamics simulations* - Same as section 1.2.2

2.2.3 *Analysis of local protein conformation*

For the analysis of non-helical states, the secondary structures were assigned using DSSP ver 2.2.1 for the same reasons as cited in section 1.2.3. The DSSP states analysed under the current chapter are:

Turn (T), Bend (S), β-bridge (B), Extended β-strand (E), and coils (C)

Protein Blocks were assigned using PBxplore [69] toolkit from GitHub. To follow the evolution of a local protein conformation in regards to its original PB assignment, a simple constituency PB ratio, named C^{PB} was calculated. C^{PB} is the percentage of times the PB x is found associated at this position where x is the initially assigned PB.

2.2.4 *k-means Clustering*

A residue is initially associated to one of the 8 defined secondary states assigned by DSSP [51] and one of the 16 PBs [53]. During MDs, DSSP and PBxplore are used again to assign the protein chain secondary structures. Hence, each residue is associated to a vector of size $S = 8$, representing the 8 defined secondary states and more specifically the occurrence of each observed state. For PBs the vector size is $S = 16$. To define common behaviors between residues, a *k-means* clustering approach was used [145].

At first, a subset is created. It represents all the residues that were associated to this state before MD, e.g. PB *d*. Then a fixed number of clusters *k* is determined. The *k* clusters are of size

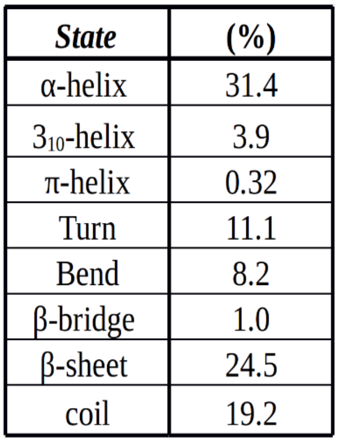
S=16 for PBs and $S = 8$ for secondary structures analysis. All the data of the subset is compared to each of the *k* cluster, and the one with the minimal Euclidean distance is considered as the winner. After one cycle (after all the subset have been used), the values of the *k* clusters are modified to correspond with the associated observations. That is each cluster is the barycentre of the associated observations. After few cycles, the *k* clusters are stable and can be analysed. This will provide answers to questions like, how the residues associated to PB *d* assignment have evolved.

2.3 Results and discussions

2.3.1 *Analyses of the data-set and simulations*

In the secondary structure dataset, α-helix has the most assignments at a frequency of 31.4% succeeded by the frequency of extended β-strand (E) at 24.5%. Least assigned secondary structures by DSSP v2.2.1 were π-helix (0.32) and β-bridge (1%). Table 2.1 summaries the frequencies for different DSSP states.

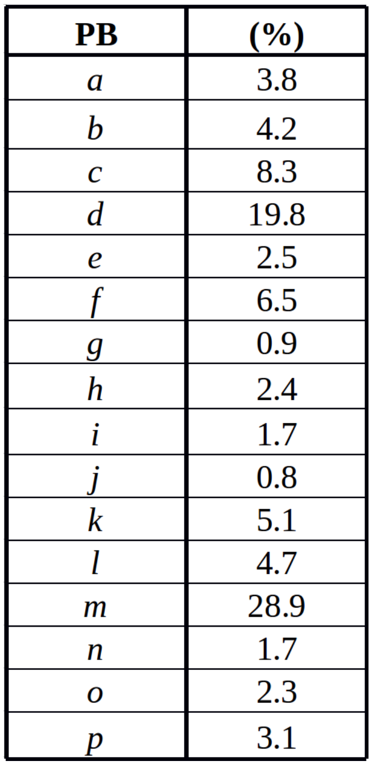
Table 2.1 ***Frequencies of occurrences of DSSP states during dynamics.*** *The table summarizes the frequency of occurrences of all the DSSP states during MD simulations of 169 domains.*



State	**(%)**
α-helix	31.4
3_{10}-helix	3.9
π-helix	0.32
Turn	11.1
Bend	8.2
β-bridge	1.0
β-sheet	24.5
coil	19.2

Protein blocks also showed similar statistics with PB *m* (approximately the core of an α-helix) being assigned the most (28.9%). PB *d* that can be assumed closer to the core of a β-strand was assigned 19.8% of time. If a complete correspondence is to drawn between the frequecies DSSP states and PBs, then α-helix and β-strands are represented at 31.4% and 24.5% of times respectively by DSSP while ~35% and 30% of time respectively in PBs. For PBs, the values are calculated by adding the frequencies of PB *l, m,* and *n,* and PBs *c, d, and e*, since PBs *l* and *n* approximate terminus of an α-helix while PBs *c* and *e* approximates β-strand N and C caps. Table 2.2 summarises the observed frequencies of all the 16 PBs.

Table 2.2 ***Frequencies of occurrences of all PBs during dynamics.*** *The table summarizes the frequency of occurrences of all the 16 PBs during MD simulations of 169 domains. PB m and d are the most commonly observed PBs which is an indicator of order in the structure and less flexibility during dynamics.*



PB	(%)
a	3.8
b	4.2
c	8.3
d	19.8
e	2.5
f	6.5
g	0.9
h	2.4
i	1.7
j	0.8
k	5.1
l	4.7
m	28.9
n	1.7
o	2.3
p	3.1

Distributions of normalized B-factor, normalized RMSf and relative solvent accessibility follow classical distributions as observed in chapter 1 (see Fig 2.1A to 2.1C). The correlation between normalized B-factor and normalized RMSf is 0.43, while their correlation with *Neq* is low, i.e. 0.24 and 0.14, respectively. Both B-factors and RMSf are global properties as they take into account the overall structure for calculations. *Neq* on the other hand is a local and precise value for the given position and can be influenced by ±2 neighbors since PB is a 5 residue long prototype. Therefor, low correlation between B-factors, *Neq* and RMSf, *Neq* may indicate that the local structures in the dataset proteins behave differently than the overall protein dynamics. Indeed, more than 60% of the residues have an *Neq* of 1.0, i.e. does not change during all simulations, while only 0.8% have a *Neq* higher than 4 (see Fig 2.1D). This indicates that overall the 169 chains have high mobility but their local structural regions do not move intrinsically.

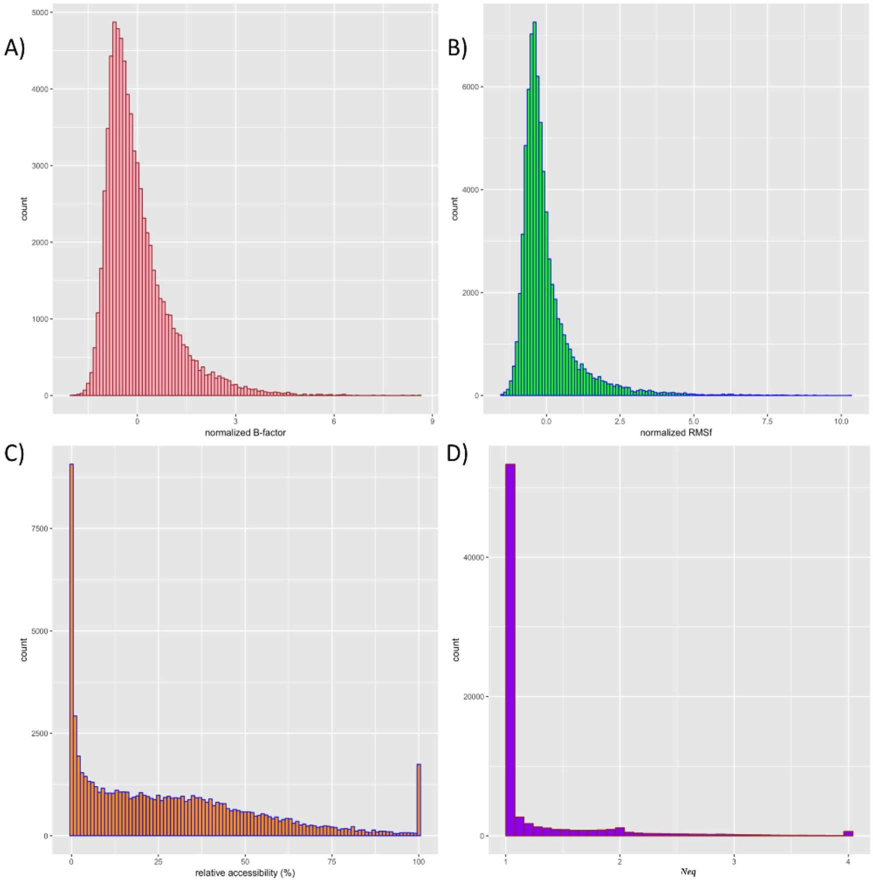


Figure 2.1 ***Distributions of different structural properties.*** *A) Normalized B-factor, B) Normalized Root mean square fluctuations (RMSf), C) Relative solvent accessibility, D) Equivalent number of PBs, Neq. The image can be divided in to two panels. Left panel includes distributions for structural properties of static structures while right panel displays the structural properties of protein dynamics.*

2.3.2 <u>*Dynamics of the non-helical secondary structures*</u>

Figure 2.2 shows a summary of the dynamic evolution of all non-helical states. Near 95% of the residues assigned to β-strand remains as β-strand during the dynamics while 4.2% goes to coil state. Interestingly, the rare β-bridge that tends to remains associated to β-bridge, also goes to coil state 12.8% and β-sheet with 10.3% of times. Unlike helical states, where interconvertibility was seen to some extent, non-helical states tend to have more perseverance. However, slight exchanges can be seen between bends (12.07%) and turns (7.23%).

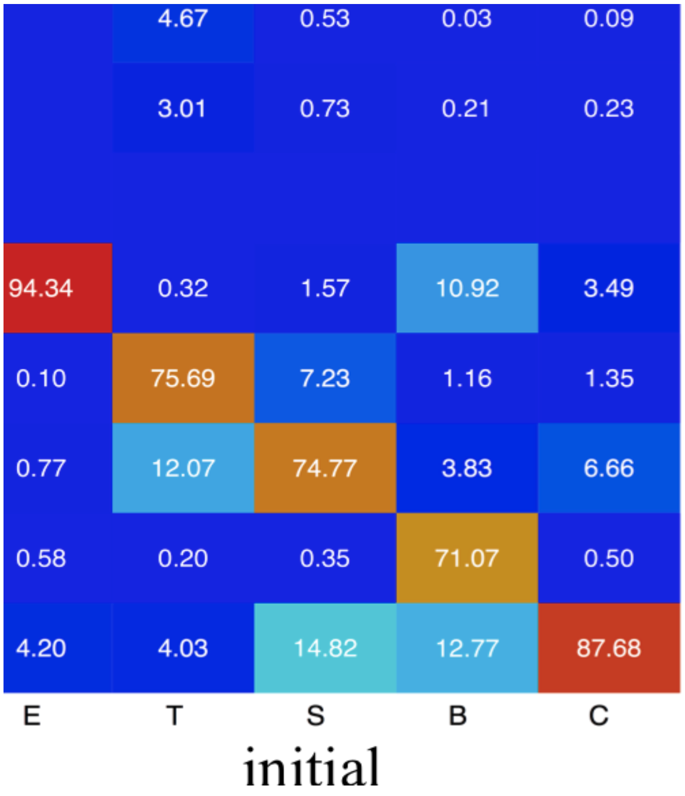


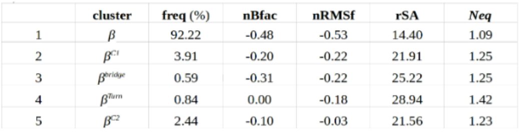
Figure 2.2 ***Dynamic exchanges of non-helical DSSP states.*** *All the initial DSSP non-helical states are shown on x-axis (a vector S=5). On y-axis are all the DSSP states to which changes during dynamics are measured. The color scheme varies from blue to red, with red being the maximum and blue being minimum. The β-strand can be seen to remain as β-strand for 94.34% for times. This suggests the rigidity associated with β-strands. Turns (T) and Bends (S) remain as T and S for 75.69% and 74.77% of times but also interchange with: T to S – 12.07% of times while S to T – 7.23% of times.*

2.3.3 *Cluster analysis of non-helical states*

Clusters were named with the most characteristic state other than initial state. For instance, β^{bridge} would translate to that β-strand dominant cluster have β-bridges as the second most dominant state in the cluster.

2.3.3.1 *Clusters of β-strand*

The first cluster is cluster β (>99% of 'E' assignments). is represents 92.2% of the occurrences with extremely low normalized B-factor (-0.48) and normalized RMSf (-0.53), it corresponds to the most buried part of the dataset (mean relative accessibility of 14.4). Figure 2.3 shows the clusters for β-strands (E) in detail. The last four clusters have higher relative accessibility ranging between 21.6 and 28.9. They are named β^{C1}, β^{bridge}, β^{Turn}, and β^{C2}. As expected β^{C1} and β^{C2} are the most occurring with a frequency of 3.9% and 2.4%, respectively. The β^{C1} cluster is relatively more rigid with nBfac value of -0.20 and nRMSf of -0.22 compared to -0.03 of β^{C2} cluster. β^{C1} also have the most β-strand content with near 2/3rd of the cluster consisting of β-strand vs only 14% in case of β^{C2}.



	cluster	freq (%)	nBfac	nRMSf	rSA	*Neq*
1	β	92.22	-0.48	-0.53	14.40	1.09
2	β^{C1}	3.91	-0.20	-0.22	21.91	1.25
3	β^{bridge}	0.59	-0.31	-0.22	25.22	1.25
4	β^{Turn}	0.84	0.00	-0.18	28.94	1.42
5	β^{C2}	2.44	-0.10	-0.03	21.56	1.23

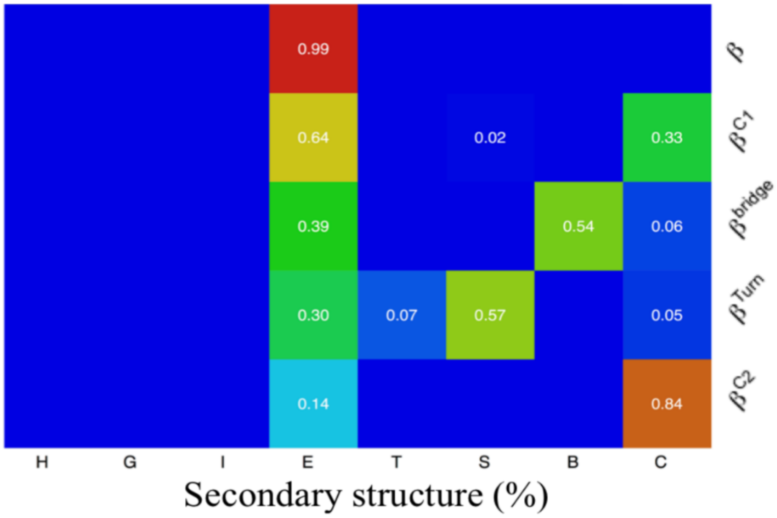
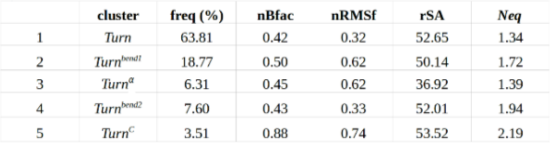


Figure 2.3 ***Evolution of β clusters.*** *The image comprises of a table that provides details about the individual frequency of the cluster and structural properties along with the behavior of DSSP states in different clusters*

A β^{Turn} cluster having a small frequency of 0.84% also appeared. It is quite flexible with regards to other clusters and display the highest accessibility among the 5 β-strand clusters. However, β^{bridge} forms a rigid cluster second to pure β cluster with nBfac of -0.31 and a low RMSf of -0.22. Surprisingly, it is the cluster with least frequency (0.6%) and represents 54% β-bridges, 6% coil and 39% β-sheet.



	cluster	freq (%)	nBfac	nRMSf	rSA	*Neq*
1	*Turn*	63.81	0.42	0.32	52.65	1.34
2	*Turn*^{bend1}	18.77	0.50	0.62	50.14	1.72
3	*Turn*$^{\alpha}$	6.31	0.45	0.62	36.92	1.39
4	*Turn*^{bend2}	7.60	0.43	0.33	52.01	1.94
5	*Turn*^{C}	3.51	0.88	0.74	53.52	2.19

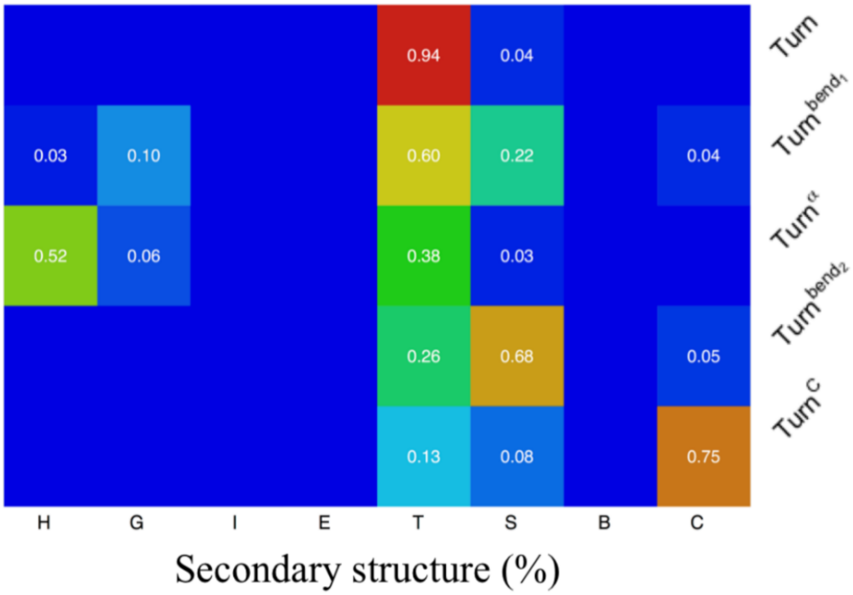


Figure 2.4 ***Evolution of β-turn clusters.*** *The image comprises of a table that provides details about the individual frequency of the cluster and structural properties along with the behavior of DSSP states in different clusters. A major share of the cluster Turn$^{\alpha}$ is contributed by α-helices (52%). This depicts the interconversion among helices and turns.*

2.3.3.2 *Clusters of β-turns*

Venkatachalam first described and classified the β-turns as hydrogen bond turns [22]. Later, the definition of turns evolved from an energetic to a distance criterion between Cα [151]. DSSP differentiates between hydrogen bond turns (namely turns, T) and non-hydrogen bond turns (namely bends, S) [51]. A turn has a perseverance with a rate of 75.7% (see Fig 2.2) while it can transform to bends at 12.1% or to an α-helix (4.7%), coil (4.0%), 3_{10}-helix (3.0%) but rarely it is seen transforming to bridges or strands. The percentage of bends and helical state is expected, as 3_{10}-helix was often confused with type III β-turn (obsolete).

The clustering reflects these results with 5 cluster belonging to Turns (Fig 2.4). Pure turn cluster, represents 63.8% of the initial turns while the rest four in decreasing order of their turn representation are: $\text{Turn}^{\text{bend1}}$, $\text{Turn}^{\text{bend2}}$, Turn^{α}, and Turn^{C}. The clusters are characterized by higher normalized B-factor (0.42 to 0.88) and normalized RMSf (0.32 to 0.74). All clusters show high relative solvent accessibility above 50% with an exception of Turn^{α} having 36.9% accessibility. In conclusion, (hydrogen bond) turns are not so rigid with the extreme being Turn^{C} having an nBfact of 0.88 and nRMSf of 0.74.

2.3.3.3 *Clusters of bends*

Slightly less frequent than turns, they also transform less to helical states relative to turns. Consequently, the perseverance analysis reveal that 71% of bends remain as bends and 14.8% goes to coils, 7.2% to turns and 1.6% to β-strands (Fig 2.2). These results are similarly reflected in the five clusters for bends.

The pure bend cluster occurs 63.8% of the time, followed by bend^{C1}, bend^{C2}, $\text{bend}^{\text{turn}}$, bend^{β} in decreasing order of their occurences of bends. Figure 2.5 show the details about the 5 bend clusters. The clusters, bend, bend^{C1}, and bend^{C2} are more rigid than Turns clusters with lower nBfac, nRMSf and rASA values. The $\text{bend}^{\text{turn}}$ is an equivalent of the two $\text{Turn}^{\text{bend}}$ ($\text{Turn}^{\text{bend1}}$, $\text{Turn}^{\text{bend2}}$). The unique cluster among 5 bend clusters is the bend^{β} that transforms at a rate of 63% to β-sheet, 25% to bends and 11% to coil. It has the lowest nBfac value, lowest rSA yet accessible at 30.1% but surprisingly, it has the highest nRMSF of 0.98 that is the highest mobility observed in any of the non-helical clusters.

	cluster	freq (%)	nBfact	nRMSf	rSA	*Neq*
1	*bend*	62.78	0.26	0.19	40.21	1.44
2	*bend*^{C1}	16.11	0.31	0.45	37.13	1.71
3	*bend*^{turn}	9.20	0.76	0.45	51.61	1.95
4	*bend*$^{\beta}$	1.64	0.21	0.98	30.15	1.43
5	*bend*^{C2}	10.27	0.32	0.18	36.56	1.76

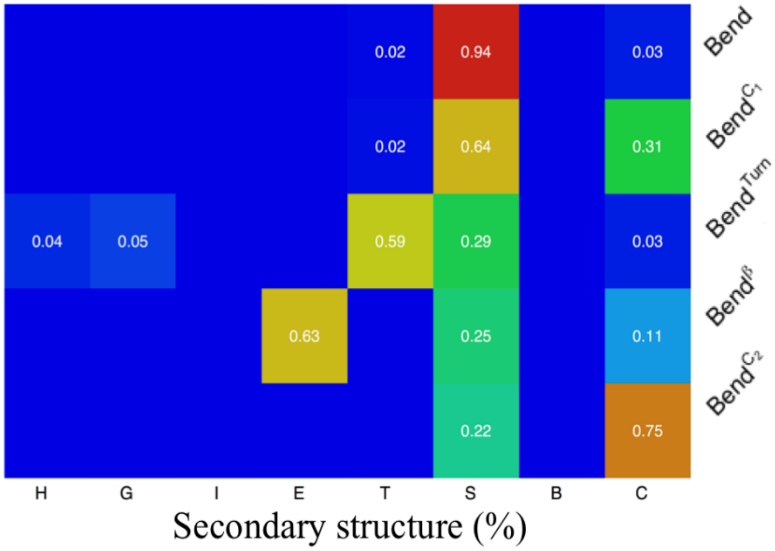


Figure 2.5 ***Evolution of clusters for bend.*** *The image comprises of a table that provides details about the individual frequency of the cluster and structural properties along with the behavior of non-helical DSSP states in different clusters. Although bends and turns are similar in structure yet, for Bends unlike turns, an α cluster does not appear. Although in Bend*^{Turn} *cluster slight tendencies to change to helices can be seen. This can also arise if there is a misidentification of turns (59%) by DSSP.*

Hence, even if turns and bends are highly comparable, they have unexpected specificities. The lack of hydrogen bonds at short range allows for a limited number of bends to participate dynamically in β-sheet and forms a specific recurrent cluster. For the turns, a specific cluster exchanges with α-helix state, but not specifically w ith 3_{10}-helix.

2.3.4 Overall PB analysis with respect to the initial assignment

As seen in Figure 2.1D, more than 60% of the residues have an *Neq* of 1.0, i.e. no change of PB assignment during the whole simulation. This rigid-constituency is highly dependent of the type of PBs. PBs geometrically related to core of repetitive structures like PB *m* (for α-helix) and PB *d* (for β-sheet) remained preserved at 100% with respective frequency of 86.2% and 75.4% (the values pertain to C^{PB} as shown in Table 2.3).

Table 2.3 C^{PB} - Frequency of PB staying in the initial PB assignment. *Analysis done using 8 different thresholds with 100%, the residues that are only seen during their dynamics associated to their initial PB assignment. Followed by 99-90%, 89-75%, 74-50%, 49-25%, 24-10%, 9-1% and finally less than 1%.*

PB	**100%**	**99-90 %**	**89-75 %**	**74-50 %**	**49-25 %**	**24-10 %**	**9-1 %**	**>1 %**
a	59.99	16.19	6.74	5.40	4.76	3.37	2.80	0.75
b	46.50	24.05	9.57	8.16	5.68	3.18	2.22	0.64
c	50.11	27.26	6.68	4.81	4.63	2.94	2.86	0.72
d	**75.35**	13.68	3.24	2.96	1.88	1.43	1.23	0.23
e	43.10	24.66	10.47	10.04	6.71	2.73	1.69	0.60
f	59.93	19.67	6.72	5.20	3.81	2.32	1.85	0.51
g	**16.17**	16.46	15.75	11.25	13.64	10.69	14.49	1.55
h	53.65	21.12	7.99	6.74	4.39	3.20	2.23	0.68
i	60.72	14.24	6.10	7.04	4.54	3.76	2.82	0.78
j	**19.14**	28.05	12.87	16.50	10.23	6.27	5.45	1.49
k	56.52	24.12	7.14	4.46	3.48	2.23	1.49	0.56
l	63.15	11.71	5.97	6.11	6.11	4.12	2.45	0.37
m	86.16	5.60	1.92	2.42	2.04	1.05	0.79	0.01
n	66.59	10.18	3.98	5.17	5.17	3.50	4.06	1.35
o	51.23	14.63	7.97	9.87	7.06	5.22	3.44	0.57
p	39.92	11.41	13.97	17.68	10.57	3.49	2.30	0.66

It decreases very rapidly in a strong gradient with PBs *n* (66.6%), *l* (63.2%), *i* (60.7%), *a* (60.0%), *f* (59.9%), *k* (56.5%), *h* (53.7%), *o* (51.2%), *c* (50.1%), *b* (46.5%), *e* (43.1%), *p* (39.9%), *j* (19.1%) and *g* (16.2%). These C^{PB} tendencies hold valid at every level of perseverance of the initial PB as well as inversely. Therefore, if an assigned PB does not have high conservation for 100% of the time, its occurrences at shorter intervals, like 49-25% or 24-10% will add up to the C^{PB}. Thus the PB at that position remains preserved. For instance, the lowest C^{PB} in 100% threshold is for *g* (16.2%) and j (19.1%). At less than 50% of the times, the highest C^{PB} are found associated to PBs *g* (40.3%) and *j* (16.2%) and not with PBs *d* (4.8%) and *m* (3.9%). Hence a strong correlation exists between the original assignment and the conservation of the local protein conformations. This leads to the occurence of the same PB and therefore low *Neq.*

However, a simple question may arise that if such C^{PB} observations are not due to the accessibility of the residues? If a local protein conformation is accessible, it's probability to change increases. Such tendency does exist but is not a binary case for every PB (see Fig 2.6).

For PBs *m* and *d*, the percentage of rigid position (C^{PB} >75%) is largely higher than in the deformable class (C^{PB} <25%) and is directly linked to their solvent accessibility. However, the PB *n* does not show this simple (and expected) tendency, the difference between rigid and deformable position is not significant. Depending on the type of PBs, it goes from a slight tendency to no tendency at all. For instance, PB *j* that is one of the two less constraint PBs, is more exposed than the others. It has the same distribution of relative accessibility in the deformable classes and the rigid ones. Hence, no specific rules can be observed here.

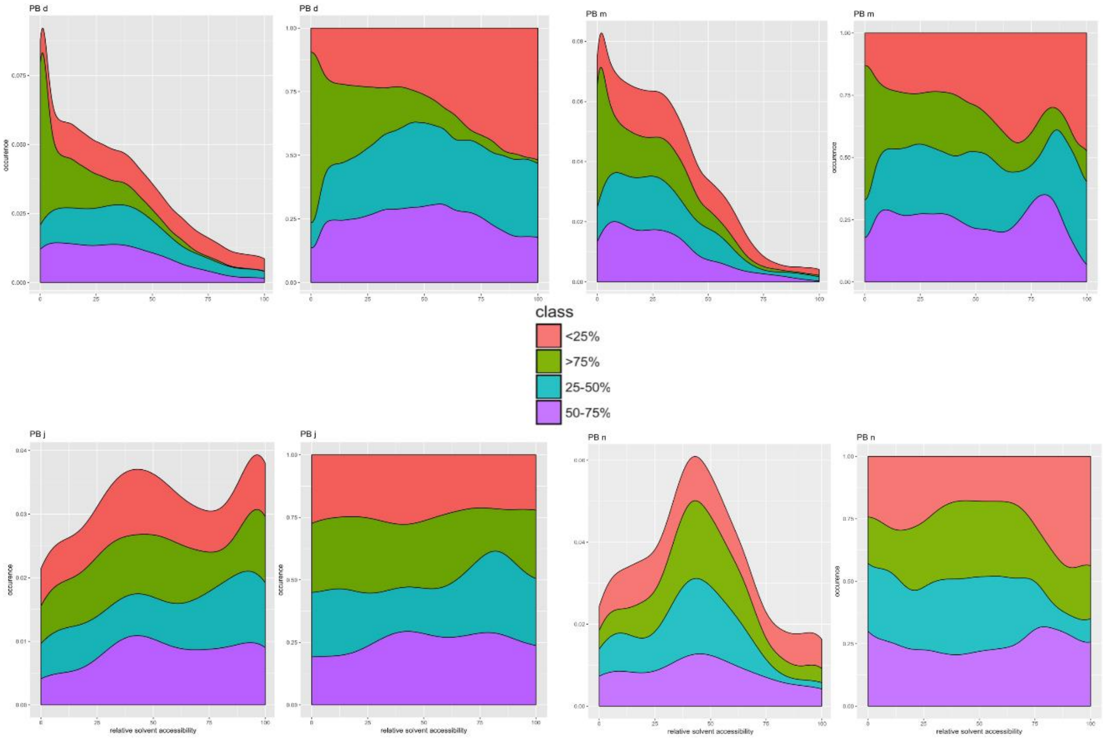


Figure 2.6 *C^{PB} in regards to relative solvent accessibility.* *Following the analysis of C^{PB} in table 2.3, the percentage a residue stay associated to its initial PB assignment was linked to the relative solvent accessibility (rSA). For each PB, four classes were defined to consider the association to initial PB, ranging from high (>75%, green), medium (75-50%, purple and 50%-25%, blue) to low (>25%, pink). In each panel, on the left is shown the occurrences, while on the right are the normalized values to 100% for every rSA value. The four PBs shown here are PB d (top left), PB m (top right), PB j (bottom left), and PB n (bottom right). The difference between rSA tendencies of the lower panel do not agree with those of top panel ones. The rSA profiles of PB d and m differs in different classes while for PB n these is no significant difference and for PB j there is no difference at all.*

Figure 2.7 shows the distribution of PBs accordingly to the initial assigned PB. It reflects the previous results underlying the important frequencies of PBs *m* and *d* (96% and 94%), and the low frequencies of PBs *g* and *j* (59% and 72%) to stay as assigned. It also shows the interconvergency amongst PBs, if any. Hence seven PBs transforms to PB *m* with a frequency higher than 2% (*a threshold used in all representation*) and eight to PB *d*. 22% of these transitions are observed, the highest one being PB *g* to PB *p* (9%), PB *g* to PB c (8%) and to PB *e* (6%), and from PB *p* to PB *m* (6%). While most of the transitions are logical in a way that geometrically they stay in similar neighborhood. Another way to analyse the evolution of PBs is the computation of *Neq* (see Fig 2.7).

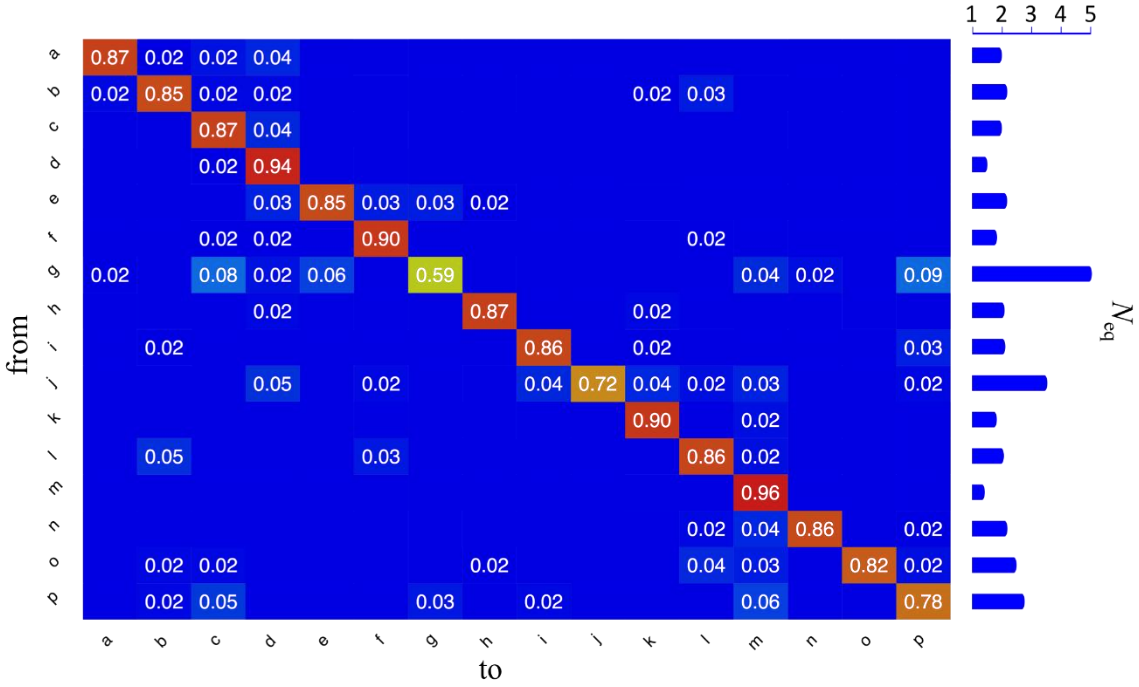


Figure 2.7 ***Exchange rates among different PBs.*** *The plot shows the different PBs a residue adopts during dynamics. x-axis is labelled as 'from' i.e. the initial PB assignments and y-axis are the 16 PBs it has a possibility to change to. The color scheme for the plot is same as used before, blue shows minimum values while red shows maximum establishing a range between them. PB d, f, k, and m shows strong tendencies towards perseverance. PB g however, is the relatively less conserved and can transform to other PBs like, PB p, c, e, m, a, d, and n. This is reflected in the Neq plot on the alternate y-axis.*

2.3.5 *Cluster analysis of Protein blocks- PB evolution*

Clusters were generated using k-means approach to study the behavior of individual PBs. Thus, 80 (16 * 5) clusters were generated and analyzed. Figures 2.8 through 2.11 shows a summary of their recurrent behavior highlighting clusters of PBs *a*, *b*, *g* and *f*. The clusters are named based on the dominant PB in the cluster.

2.3.5.1 *Clusters of PB a*

The five clusters of PB *a* shows interesting results. Figure 2.8 shows the details about clusters of PB *a*. Cluster a_1, is the cluster with >98% of PB *a* and represents 4/5th of positions initially associated to PB *a*. Surprisingly, it does not have the least normalized B-factor values. Although it is associated to the lowest normalized RMSf and one of the lowest relative solvent accessibility values. It also represents a first example of that the more stable cluster (i.e. PB x that stay PB x) is not always associated to lowest nBfac and rSA.

Cluster a_2 represents another behaviors, namely, the cluster that is still highly controlled by the original PB, but also goes to a large number of other local conformations. Therefore, in Cluster a_2, PB *a* still represents 67% of the occurrences but with 6 PBs at an interconversion rate of more than 2%. The PBs to which initial assignment of PB *a* transforms to are: PBs *b, c, d, f, l* and *m.*

Cluster a_3 introduces more fuzziness in PB a cluster. Only, 22% of PBs that were initially assigned, PB *a* remained as PB *a*. Most of the changes are shown to be attributed to PB *c* (65% of PB *c*).

Cluster a_4 represents the fuzzy cluster, it represents only 5.3% of the original PB and therefore, must be the more deformable. The average *Neq* of 2.36 (overall), and cluster *Neq* of 8.43. it is also associated to highest accessibility, highest nBfac and highest nRMSf values.

	frq (%)	PB fr(%)	nBfac	nRMSf	rSA	*Neq*	cl. *Neq*
a₁	80.16	98.62	0.10	-0.06	20.44	1.06	1.10
a₂	9.05	65.95	0.09	0.42	27.93	2.21	4.04
a₃	1.77	22.12	-0.03	0.20	20.36	2.20	2.80
a₄	5.32	20.20	0.60	1.00	36.63	2.36	8.43
a₅	3.69	17.13	-0.02	0.24	30.41	1.93	2.24

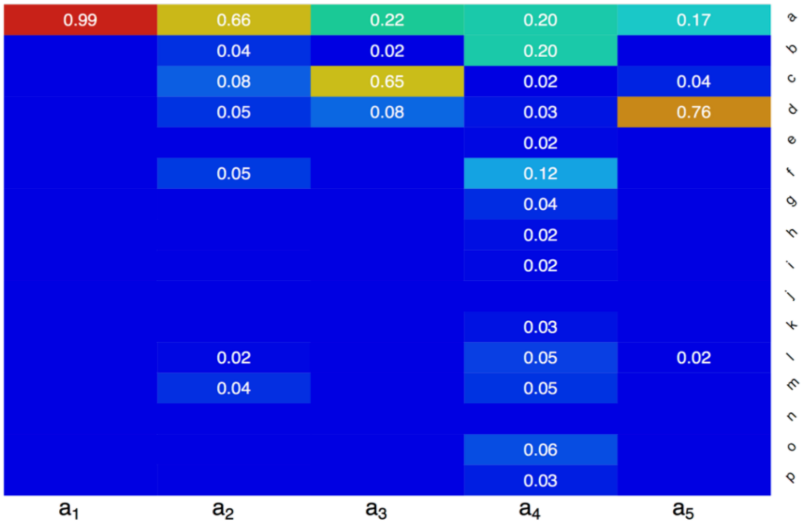
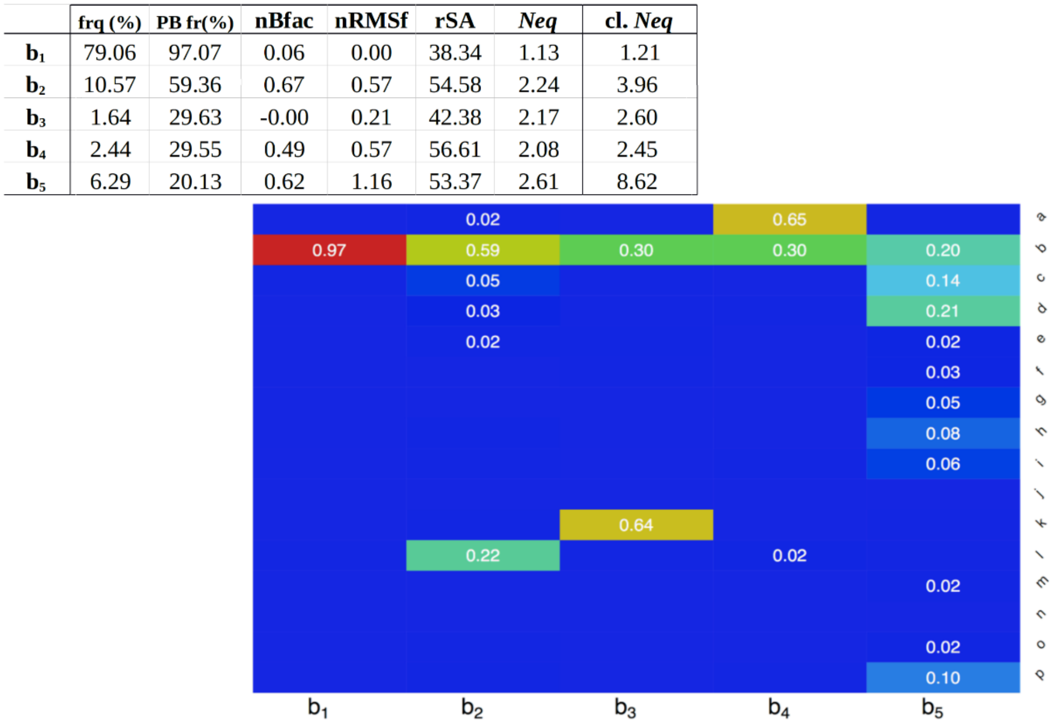


Figure 2.8 ***Evolution of clusters of PB a.*** *The figure comprises of a table that provides details about the individual frequency of the cluster as well as static and dynamic structural properties. Values for both Neq are given; the average Neq (Neq) and Neq of the cluster (cl. Neq). The spread of each cluster of a is shown. As apparent, one 'stable' cluster herein, a*$_1$*. While one entirely fuzzy cluster among the 5 clusters is expected given the increasing deformability. a*$_4$ *shows the characteristics of the fuzzy cluster. Transformations among PB c, f, d are considered geometrical transitions based on their geometric resemblance.*

In our previous work, geometrical compatibilities among PBs were assessed by considering the second best PB for every local conformation [58]. According to this concept of geometrical resemblance, similar PBs can interchange more often. Clusters a$_3$ and a$_5$ followed such geometrical transitions, since they have respectively high frequencies of PB *c* (65%) and PB *d* (76%). However, no clusters with strong evolution to PB *f* can be seen.

2.3.5.2 *Clusters of PB b*

Figure 2.9 show the details about clusters of PB *b*. The cluster b$_1$ (>97% of PB *b*) represents 79% of the original PB *b* with lowest rSA and lowest nRMSf but the second lowest nBfac (0.00 vs. -0.06 for cluster b$_3$). Interestingly none of the three following clusters have used the expected major geometrical transitions (PBs *d*, *c* and *f*). However, geometric transition changes are observed for PB *l* for cluster b_2 (22%), PB *k* for cluster b_3 (64%) and PB *a* for cluster b_4 (65%). Only cluster b_3 can be considered as comparable with cluster b_1 in terms of nRMSf and nBfac and closest rSA. Cluster b_5 showed the maximum fuzziness with transitions among ~10 different PBs.



	frq (%)	PB fr(%)	nBfac	nRMSf	rSA	*Neq*	cl. *Neq*
b_1	79.06	97.07	0.06	0.00	38.34	1.13	1.21
b_2	10.57	59.36	0.67	0.57	54.58	2.24	3.96
b_3	1.64	29.63	-0.00	0.21	42.38	2.17	2.60
b_4	2.44	29.55	0.49	0.57	56.61	2.08	2.45
b_5	6.29	20.13	0.62	1.16	53.37	2.61	8.62

Figure 2.9 ***Evolution of clusters of PB b.*** *The figure comprises of a table that provides details about the individual frequency of the cluster as well as static and dynamic structural properties. Values for both Neq are given; the average Neq (Neq) and Neq of the cluster (cl. Neq). The spread of each cluster of b is shown. As apparent, one 'stable' cluster herein, b_1. While one entirely fuzzy cluster among the 5 clusters is expected given the increasing deformability. b_5 shows the characteristics of the fuzzy cluster. Transformations among PB c and d are considered geometrical transitions based on their geometric resemblance. However, cluster b_3 shows a non-geometric transition from b to k but strikingly it has nBfac of 0 and very low nRMSf.*

2.3.5.3 *Clusters of PB f*

Figure 2.10 show the details about clusters of PB *f*. As often seen the majority cluster f_1 (>98% of PB *f*) represents 83% of original PB *b* positions had lowest nBfac, nRMSf and rSA. The expected geometrical transitions (PBs *b* and *k*) have not been used during dynamics to substitute

PB *f*. They have been replaced by PB *e* for cluster f3 (65%) and PB *d* for cluster f5 (66%). Interestingly, the transition of PB *f* to PB *d* is associated with high nBfac, high nRMSf and high rSA that is quite uncommon for PB *d*. Cluster f2 is less stable than cluster b1 with a lower PB content of PB *f* (69%).

	frq (%)	PB fr(%)	nBfac	nRMSf	rSA	*Neq*	cl. *Neq*
f_1	83.13	98.69	-0.08	-0.13	22.66	1.06	1.10
f_2	8.65	68.52	0.13	0.24	28.96	2.00	3.48
f_3	1.16	25.95	0.19	0.09	23.89	2.06	2.75
f_4	5.29	21.59	0.27	0.63	32.62	2.43	8.78
f_5	1.77	22.08	0.35	0.94	35.44	2.19	2.87

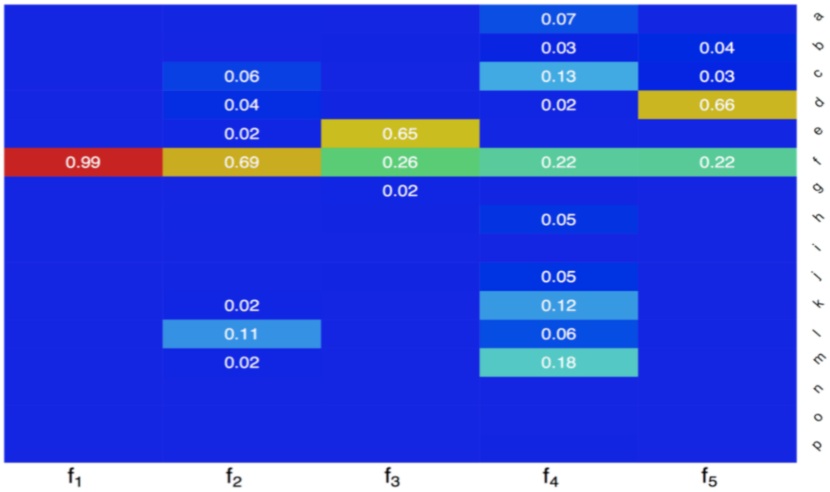


Figure 2.10 ***Evolution of clusters of PB f.*** *The figure comprises of a table that provides details about the individual frequency of the cluster as well as static and dynamic structural properties. Values for both Neq are given; the average Neq (Neq) and Neq of the cluster (cl. Neq). The spread of each cluster of f is shown. As apparent, one 'stable' cluster herein, f_1. While one entirely fuzzy cluster among the 5 clusters is expected given the increasing deformability. f_4 shows the characteristics of the fuzzy cluster. Transformations among PB f, g, h are considered geometrical transitions based on their geometric resemblance. However, cluster f_3 and f_5 shows a non-geometric transition from f to e and d.*

2.3.5.3 *Clusters of PB g*

Finally Figure 2.11 shows the details about clusters for PB *g*. Cluster g_1 (>89% of PB *g*) represents 55% of the original PB *g* positions and have the lowest rSA and lowest nRMSf but the second lowest nBfac (0.04 vs. 0.03 for cluster g_2). As seen in the previous sections, PB *g* does not stay as PB *g* as often than other PBs. The following clusters are composed of 33%, 24%, 20% and 16% of PB *g*. The first surprise cluster is cluster g_2 directed by PB *e* (57%) that is quite comparable to cluster g_1 in terms of protein flexibility characteristics (similar nBfac and nRMSf). Cluster g_3 was more or less expected as PB *c* is an expected geometrical transition; it represents 64% of the cluster. Surprisingly, the cluster g_4 is controlled by PB *p* (52%) that is not a major geometrical transition.

	frq (%)	PB fr(%)	nBfac	nRMSf	rSA	*Neq*	cl. *Neq*
g_1	55.56	89.23	0.04	-0.00	19.13	1.44	1.77
g_2	7.03	33.14	-0.03	0.08	29.38	2.41	3.00
g_3	8.58	23.64	0.30	0.46	22.99	2.32	3.00
g_4	13.78	20.24	0.21	0.41	40.04	3.17	4.19
g_5	15.05	16.21	0.31	0.67	36.18	2.87	11.69

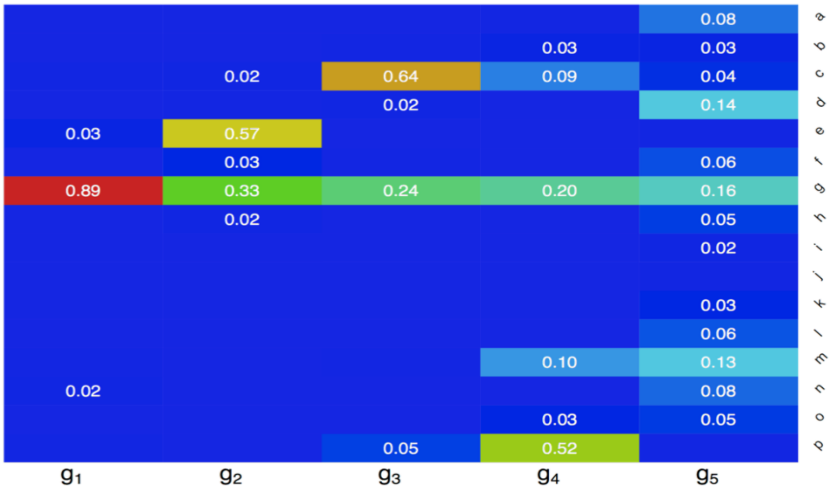


Figure 2.11 ***Evolution of clusters of PB g.*** *The figure comprises of a table that provides details about the individual frequency of the cluster as well as static and dynamic structural properties. Values for both Neq are given; the average Neq (Neq) and Neq of the cluster (cl. Neq). The spread of each cluster of g is shown. As apparent, one 'stable' cluster was expected but a slight deformable g_1 is observed. While one entirely fuzzy cluster among the 5 clusters is expected given the increasing deformability. g_5 shows the characteristics of the fuzzy cluster but g4 also appears to be slightly fuzzy.*

2.3.6 *Summary of the clustering*

For each PB, one cluster represents the initial PB with high frequency with more than 95% of initial PB, except for cluster *g* which represented PB *g* only 89.2% of times. This 'stable' cluster is not always associated to the lowest normalized B-factors and lowest mean relative solvent accessibility. One cluster among the five cluster is a fuzzy cluster with the highest average *Neq* and especially highest cluster *Neq*. The three remaining clusters are divided into:

(i) a cluster that is a degenerated version of the 'stable' cluster, often having more than mere 60% of initial PB and a mix of others eg, clusters a_2, b_2, c_2, d_2, f_2, h_2, i_2, j_2, k_2, n_2, o_2, and p_2.

(ii) Clusters that are directed by unexpected PBs, i.e. not from the major geometrical transitions.e.g. clusters b_2, b_3, b_4, c_3, e_2, f_3, f_5, g_2, etc.

(iii) The remaining clusters follows a major geometrical transitions among PBs

Comparison of the obtained clusters showed that most of the fuzzy clusters are highly similar and that most of the other clusters with unexpected PB does not cluster with their associated PB clusters. Figure 2.12 highlights that the initial local conformations can go to very different conformational behaviors. Interestingly, near no cluster from a given PB is associated to one of its related generated cluster. Such that, cluster f_1 is closest to cluster e_3, cluster f_2 is closest to d_4, cluster f_3 seems closest to e_1, cluster f_4 is a fuzzy cluster, and cluster f_5 is closest to cluster a_5. The dynamics, therefore have a strong local protein conformation impact that have been clustered and described.

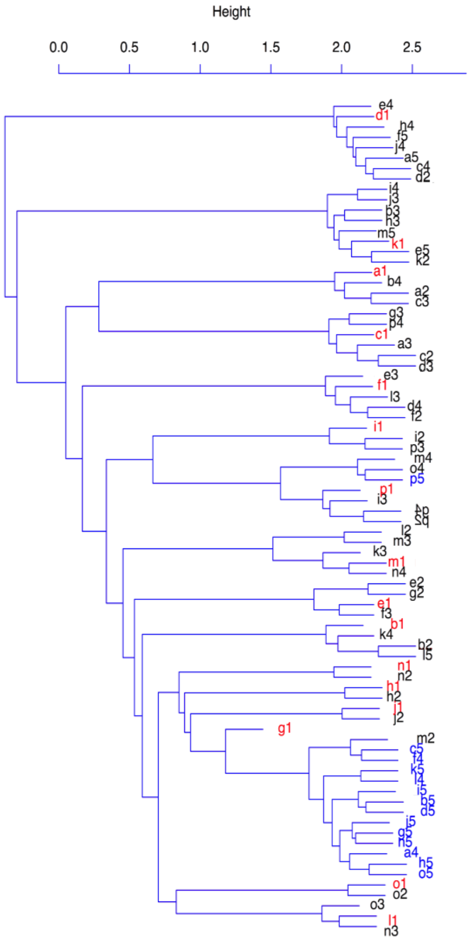


Figure 2.12 ***Hierarchical clustering of the 5-clusters of each PBs.*** *In blue are shown the fuzziest clusters and in red the first cluster associated to each PB.*

2.4 Conclusions and perspectives

In the current analysis, MD simulations were performed on a large set of 169 representative protein domains. In chapter 1, it was shown that only 76.4% of the residues associated to α-helices retain their conformation, while this tendency drops to 40.5% for 3_{10}-helices and is never seen for π-helices. Taking the logical step further the current study extends the analysis to non-helical conformations and PBs. The resulting analysis confirms the rigidity of sheets, but also underline its capacity to transform into turn conformations. While the dynamics between turns (with hydrogen bond) and bends (without hydrogen bond) have some strong similarities, the two conformations behave distinctively. The turns can transform to helical structures while bends prefer to go to extended structures.

An entire analysis of a large set of protein dynamics simulations using a structural alphabet is performed. It is done on two levels: (i) a global view in terms of PBs, (ii) performing clustering for each types of PBs. Systematic analysis of PBs provide surprising results with multiple informations. As expected a large part of the buried positions remain highly stable, but it is not an observed (fixed) rule. In fact, for at least half of the PBs, the fact to be buried or exposed does not change its dynamics, at all. The majority of PBs tend to remain as their original PB, or at least with a high frequency. Some PBs have a higher tendency to be not as rigid as others and it is particularly true for PB *g* and PB *i*. The intriguing fact is that the change from a PB to another one is not an obvious geometrical change. It is more frequent to go to an unexpected PB than an expected one (due to its geometrical compatibility).

The use of two types of classification shows the difficulty to cluster properly these dynamical properties but it indeed improves (i) our knowledge of protein dynamics and (ii) the relationship between sequence – structure and dynamics.

ACKNOWLEDGEMENTS

The current analysis and the one carried out in chapter 1 would have not been completed without the immense support from M. Nicolas Shinada and Dr. Alexandre de Brevern. A special thanks to Dr. de Brevern is required for his sincere help in shaping and compiling the chapter.

Chapter 3: Understanding local protein flexibility in light of physiological structural events: *Case studies*

Proteins are the functional currency of the biological systems. All the molecular events from DNA replication, transcription, translation, to sorting, transport, expression, to signalling involve proteins in important roles. Such diversity of functions often involves the same set of proteins but with some molecular variations. For instance, phosphorylation, glycosylation, SUMOylation, acetylation, methylation, etc, at certain sites in a structure induces conformational changes. Thus contributing to its functional versatility. Therefore, it is indeed important to understand the impact of such structural changes on local structure dynamics.

3a: Impact of post-translational modifications on protein backbone conformation

3a.1 Introduction

3a.1.1 *Physiological role of Post-Translational Modifications*

After its synthesis, a protein can undergo reversible or irreversible covalent modifications, namely Post-Translational Modifications (PTMs). The modifications alter the physicochemical properties of the proteins and thereby regulate enzymatic activity, cellular localization and intermolecular interactions [260–262]. Additionally, a protein could be modified in many ways and at different residue positions over time. The same position may also undergo changes of different kinds. However, changes may be specific to certain amino acids. For example, N-glycosylation found on asparagine in the specific consensus sequence Asn-X-Ser/Thr; where ‘X’ can be any amino acid residue but proline [263]. Also, PTMs are extremely diverse, ranging from the addition of a small group of atoms, such as phosphorylation [264], to the attachment of bulkier oligosaccharide by glycosylation [265]. PTMs are essential to regulate biological functions, such as DNA transcription by histone methylation and demethylation, acetylation or phosphorylation [266,267], nuclear-cytosolic or extra-cytosolic transport by SUMOylation [268,269] or glycosylation [270,271], tagging proteins for degradation by ubiquitination [272], and regulation of kinase activity with phosphorylation [273]. Due to implications in all major physiological functions of the cell, PTMs are often associated with major human diseases such as cancer, diabetes, cardiovascular disorders and Alzheimer's disease [274–276].

3a.1.2 *The PTM code*

In context of protein function, such diverse roles may lead to cooperative mechanisms of PTMs such as competition for serine and threonine residues between phosphorylation and O-glycosylation [277]; ubiquitination favored over phosphorylation leading to protein degradation [278], or the interactions between PTMs regulating the activity of the p53 protein and Histones [279,280]. These observations suggest towards the existence of a PTM-code [281–283], which is based on the presence and association of several PTMs leading to the realization of particular functions (Fig 3a.1). Recently, the increasing number of annotations on PTMs have assisted in understanding the cross talk or direct / indirect influences among different types of PTMs [284–286] their competition for the same residue [287], or the co-evolution of different PTMs sites within the same protein [288,289].

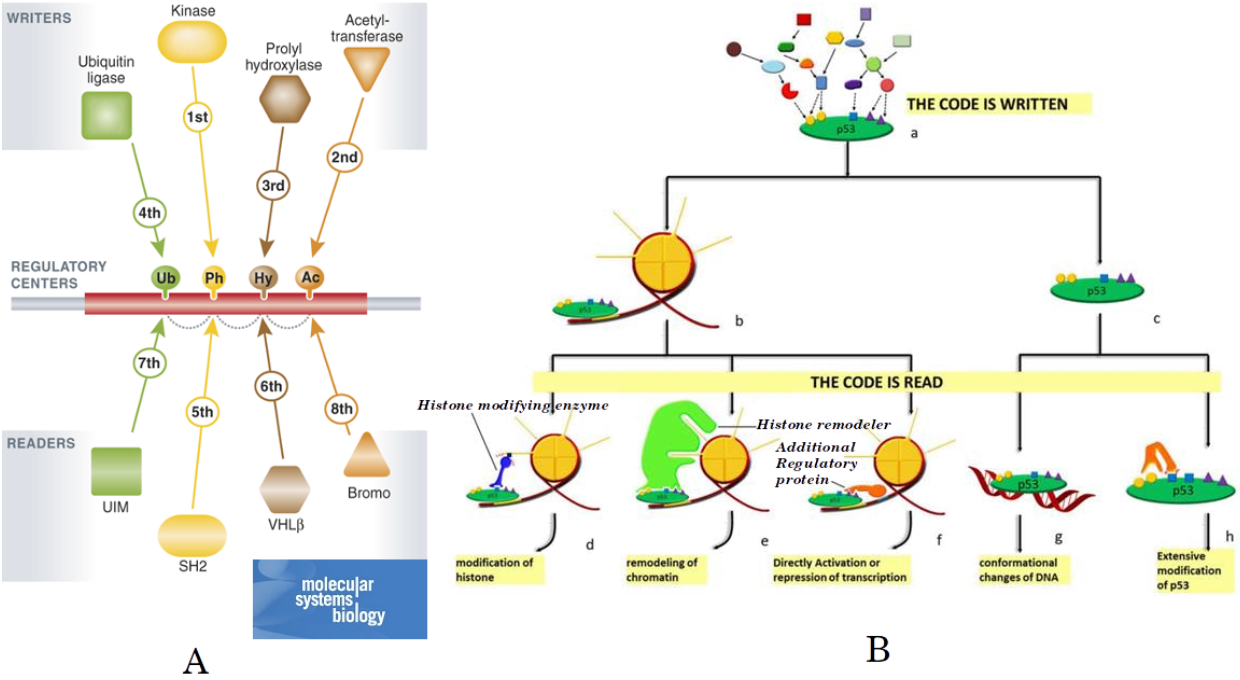


Figure 3a.1 ***PTM cross-talks and information sharing is indicative of a PTM code.*** *A) A schematic representation of cross talks between the PTM inducing proteins, like Kinases, known as writers, the PTMs like phosphorylation on regulatory centers, and the cellular function inducing readers. The dashed lines show the cross-talk between different PTMs. The sequential numbers*

depict the sequence of events. These indicate that a protein can reach multiple functional states using PTM-driven logic gates, thus indicating a PTM code. B) shows an example of such PTM code to exist using PTM network of p53 regulation. Different PTM inducing proteins modify p53 genes with their respective PTMs. The fate of p53 regulation can follow different pathways depending upon the proteins that read the PTM induced by Writters.

+*Images taken from: A) [289] and B [280]*

3a.1.3 *Effect of PTMs on protein structure*

The proteins functions and their 3D structures are intrinsically related. Hence, it is expected that PTMs, which regulate function, impact the structure of proteins as well. Several previous studies have investigated the effects that PTMs could have on the protein structure and dynamics, using X-ray data [290], and NMR data [291]. Xin and Radivojac [290] computed local and global RMSDs between modified (with at least one PTM), and unmodified PDB chains of the same protein. They concluded from the statistical analysis of their RMSDs that N-glycosylation and phosphorylation induce conformational changes, with a limited impact, at both local and at global levels, with a larger influence for phosphorylation. On their side, Gao and Xu [291] suggest that disorder-to-order transition could be induced by the modifications of phospho-serine/-threonine, various types of methyllysines, sulfotyrosine, 4-carboxyglutamate, and potentially 4-hydroxyproline.

Also intrinsically disordered protein regions have been associated with numerous PTMs, as hydroxylation, methylation, and notably phosphorylation [291–294] which was recently proposed to function as protein interaction switches in more ordered regions [295].

3a.1.4 *Computational analysis of PTMs*

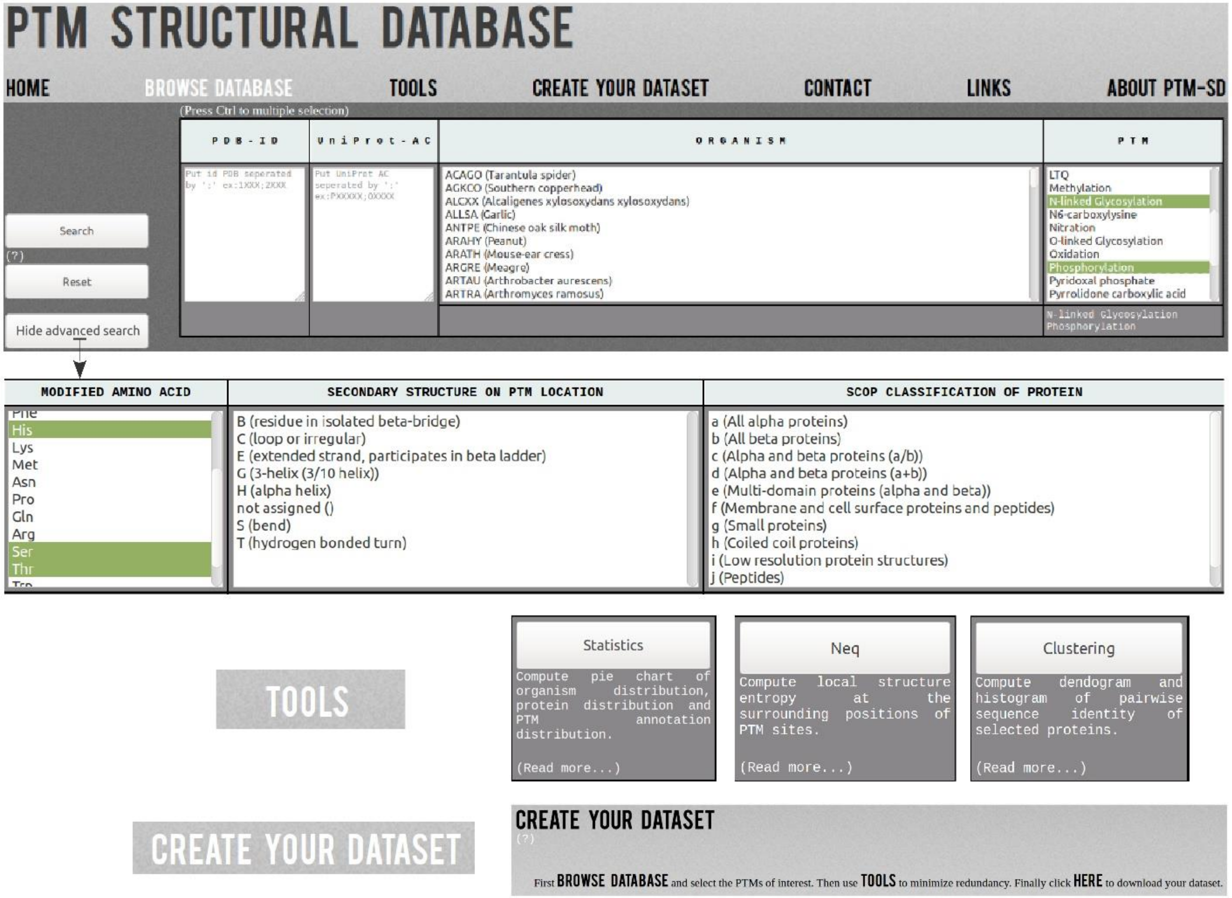
The available data on PTMs have increased drastically in the recent years due to improvements of mass spectrometry-based detection methods [296]. To acknowledge such expansion of data, many databases and prediction tools have been developed. They have enhanced the understanding of various PTMs in different organisms and simplified the analysis of complex PTM data [297]. These PTM databases contain crucial sequence annotations, specific to some PTM types and/or organisms [298], and provide related structural data thus mapping the PTM sites on corresponding structures in Protein Data Bank (PDB) [299].

Numerous machine learning methods consisting of predicting PTM sites were published recently. They mainly focus on certain types of PTM and/or organisms, and differ in their learning protocols (support vector machine, random forest, neuronal network, etc.), and in the set of descriptors extracted from the mining of the experimental data [297,300]. Few of them, used descriptors derived from structural data, such as prediction of secondary structures, disorder and accessible surface area [301,302], or from structural properties extracted from PDB [303,304].

3a.1.5 *PTM-SD*

Post Translational Modification Structural Database, abbreviated as PTM-SD (http://www.dsimb.inserm.fr/dsimb_tools/PTM-SD/) [305] was designed by Craveur et al. from our lab in 2014 [136]. PTM-SD (Fig 3a.2) is designed to give users a curated access to the proteins for which one or more Post Translational Modification(s) is (are) structurally resolved in the Protein Data Bank (PDB) and also experimentally annotated in dbPTM [299] and PTMCuration [306]. PTM-SD uses diverse set of rules to underline the discrepancies between annotation in the structure and the sequences owing to different sources. Also, PTM-SD allows the user to create customized PTM queries and perform different analyses on the returned hit. For example, computing distribution of organisms, proteins, PDB codes/chains, and PTM types, assigning PBs, computing *Neq* (section 1.6.3), highlighting discrepancies between PDB sequence and UniProt sequence, clustering for generation of non-redundant dataset, etc.

Besides a global view on PTMs, the database also provides details for each PTM and further connects to different PTM information and annotations found in other databases. Such data are very informative for studying relationship between PTMs and protein structures, for designing comparative modeling protocol, and for prediction protocol based on different approaches, for example, on secondary structure descriptors.



(Press Ctrl to multiple selection)

Put id PDB seperated by ';' ex:1XXX;2XXX

Put UniProt AC seperated by ';' ex:PXXXXX;OXXXX

ACAGO (Tarantula spider)
AGKCO (Southern copperhead)
ALCXX (Alcaligenes xylosoxydans xylosoxydans)
ALLSA (Garlic)
ANTPE (Chinese oak silk moth)
ARAHY (Peanut)
ARATH (Mouse-ear cress)
ARGRE (Meagre)
ARTAU (Arthrobacter aurescens)
ARTRA (Arthromyces ramosus)

LTQ
Methylation
N-linked Glycosylation
N6-carboxylysine
Nitration
O-linked Glycosylation
Oxidation
Phosphorylation
Pyridoxal phosphate
Pyrrolidone carboxylic acid

N-linked Glycosylation
Phosphorylation

Phe
His
Lys
Met
Asn
Pro
Gln
Arg
Ser
Thr
Trp

B (residue in isolated beta-bridge)
C (loop or irregular)
E (extended strand, participates in beta ladder)
G (3-helix (3/10 helix))
H (alpha helix)
not assigned ()
S (bend)
T (hydrogen bonded turn)

a (All alpha proteins)
b (All beta proteins)
c (Alpha and beta proteins (a/b))
d (Alpha and beta proteins (a+b))
e (Multi-domain proteins (alpha and beta))
f (Membrane and cell surface proteins and peptides)
g (Small proteins)
h (Coiled coil proteins)
i (Low resolution protein structures)
j (Peptides)

Compute pie chart of organism distribution, protein distribution and PTM annotation distribution.

Compute local structure entropy at the surrounding positions of PTM sites.

Compute dendogram and histogram of pairwise sequence identity of selected proteins.

First BROWSE DATABASE and select the PTMs of interest. Then use TOOLS to minimize redundancy. Finally click HERE to download your dataset.

Figure3a.2 ***PTM-SD, a database of structurally resolved and annotated PTM in proteins.*** *A summary of the PTM-SD database Query and Search page. Also depicting the tools that can be implemented to reduce redundancy (clustering), compute statistics and Neq (local entropy) values. Using these tools, a customized dataset can be created from the required query with PDB and/or UniProt ids, selective type of PTMs based on specific modified residues in specific organisms. Similar queries were used to generate the dataset used for studying effect of PTMs on the protein backbone conformations (selections highlighted in green).*

Since PTM-SD gives access to X-ray structures of modified residues in proteins that specifically correspond to all PTM annotations along with their statistical characterization like *Neq* [136]. It was used to investigate the impact of PTMs on the protein backbone conformations observed in crystallographic data. The currrent structural analysis is focused on understanding the following:

I. The diversity of the backbone conformations of N-glycosylated and phosphorylated regions.

II. Local and global effects on the backbones were compared between 4 specific examples of PTMs associated to a high number of experimental data.

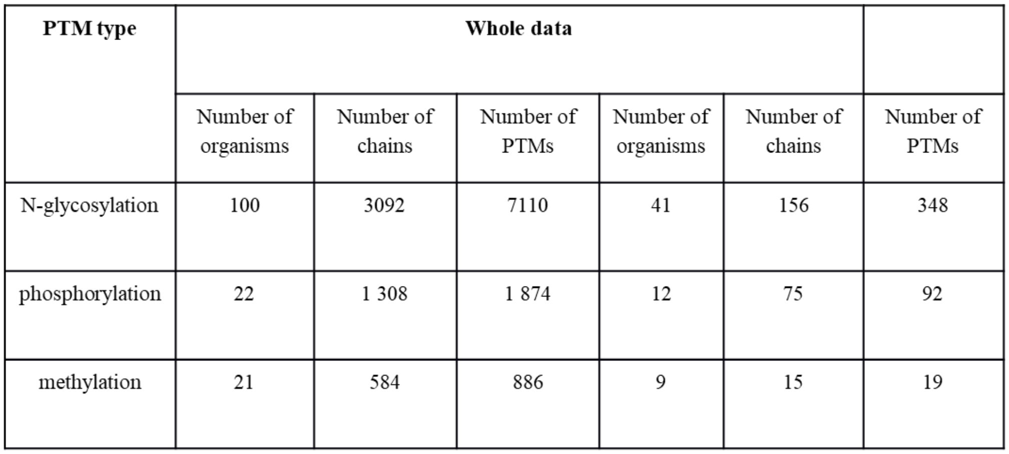
III. The conformational changes of the presence and absence of PTMs on the protein were also compared, in regards to the backbone flexibility.

3a.2. Methods

3a.2.1 <u>*Dataset preparation*</u>

The dataset was generated using PTM-SD. It comprises of structures pertaining to phosphorylation, N-glycosylation and methylation while also contains corresponding structures without a modification. Table 3a.1 summarizes the dataset composition. The comprehensive dataset included a total of 9,870 PTMs that are present on 5,948 structures. From these PTMs, 7,110 modifications are N-glycosylation while 1,874 are phosphorylation and 886 are methylations. The dataset was further refined to remove redundancy (>25% identity) using PTM-SD clustering toolkit. The percentage identity signifies that the sequences in each cluster have greater than 75% identity and the intercluster sequences will have more than 25% difference in their sequence identity. In summary, it removes the same type of PTM at the same position if the sequences are 75% identical.

Table 3a.1 ***The Dataset for PTM analysis:*** *Using PTM-SD, a comprehensive structural dataset is prepared with PTMs, N-glycosylation, phosphorylation and methylation. The table indicates the details of the dataset with diversity indicated as number of different source organisms, size depicted by the no. of chains and quality of data is indicated by the number of PTM. Similar statistic is also given for the derived non-redundant dataset (in columns 5 to 7).*

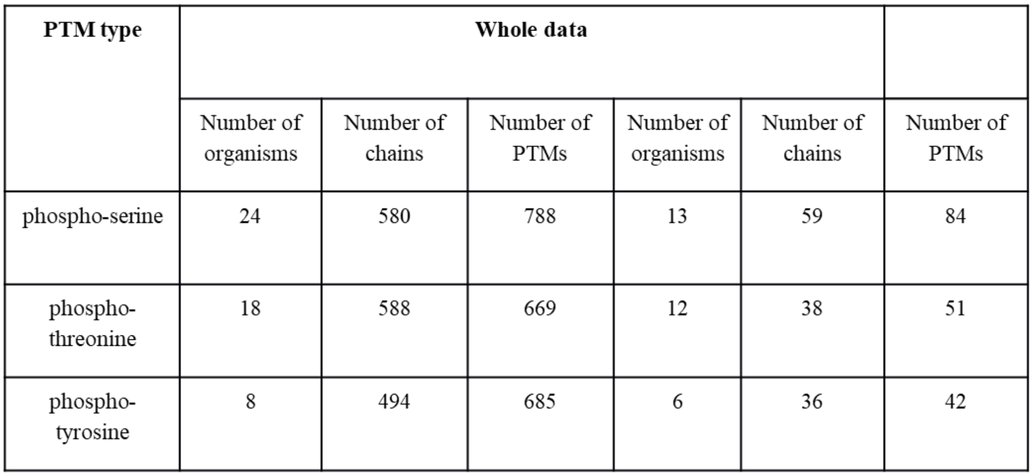


PTM type	**Whole data**					
	Number of organisms	Number of chains	Number of PTMs	Number of organisms	Number of chains	Number of PTMs
N-glycosylation	100	3092	7110	41	156	348
phosphorylation	22	1 308	1 874	12	75	92
methylation	21	584	886	9	15	19

The non-redundant dataset consisted of 348 N-glycosylation on 156 PDB chains from 41 different organisms, 92 phosphorylations on 76 structures from 12 different organisms and 19 methylations on 15 structures from 9 distinct organisms, details in Table 3a.1.

Similar datasets were also generated for the analysis of different types of phosphorylations. Dataset was selected based on the amino acid residue phosphorylated. 84 serine modifications on 59 pdb chains while 51 phosphothreonine and 42 phosphotyrosine are found on 38 and 36 unique pdb chains. Tabular details are provided in; Table 3a.2.

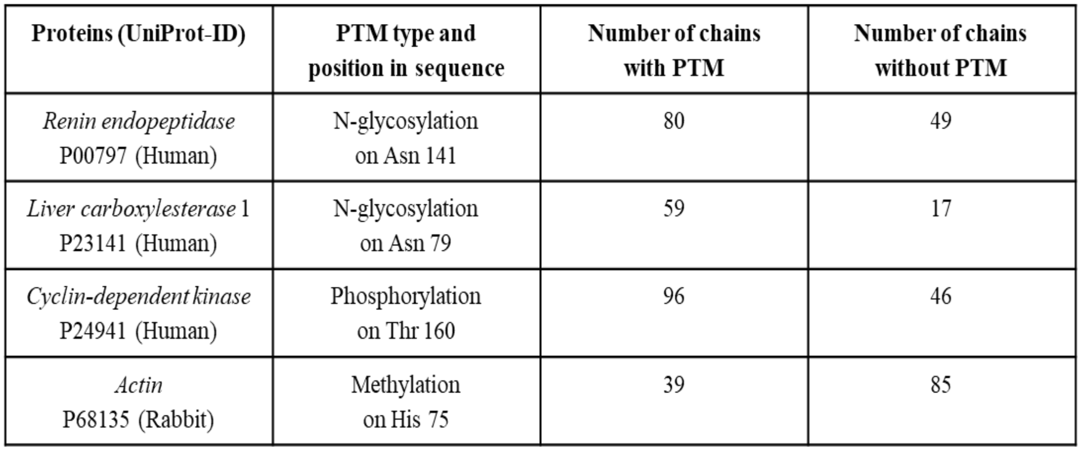
Table 3a.2 ***Dataset for phosphorylation analysis.*** *The table represents the details of the dataset comprising of different kind of phosphorylation modifications, built using PTM-SD. The diversity of the data is indicated by the number of different source organisms, size depicted by the no. of chains and quality of data is indicated by the number of PTM. Similar statistics is also given for the derived non-redundant dataset (in columns 5 to 7).*



PTM type	Whole data					
	Number of organisms	Number of chains	Number of PTMs	Number of organisms	Number of chains	Number of PTMs
phospho-serine	24	580	788	13	59	84
phospho-threonine	18	588	669	12	38	51
phospho-tyrosine	8	494	685	6	36	42

A derived dataset was also generated to assess the impact of PTM on the global structure. Therefore, a dataset comprising 4 proteins; Renin endopeptidase (N-glycosylation), Liver carboxylesterase (N-glycosylation), Cyclin dependent Kinase 2 (Phosphothreonine) and Actin (Methylation) was generated (refer to Table 3a.3).

Table 3a.3 ***Dataset to analyse local and global impacts of PTMs on 4 proteins.*** *Four proteins as listed in Column1 are selected to study the impact of PTM on the protein structure. Column 2 lists the modification taken into account while Column 3 & 4 are the no. of structures used for comparison of structural impact in presence and absence of the PTM, respectively*.



Proteins (UniProt-ID)	PTM type and position in sequence	Number of chains with PTM	Number of chains without PTM
Renin endopeptidase P00797 (Human)	N-glycosylation on Asn 141	80	49
Liver carboxylesterase 1 P23141 (Human)	N-glycosylation on Asn 79	59	17
Cyclin-dependent kinase P24941 (Human)	Phosphorylation on Thr 160	96	46
Actin P68135 (Rabbit)	Methylation on His 75	39	85

3a.2.2 *Protein Blocks (PB) assignment*

PB assignment was done using our in-house PBxplore tool [69]. The PB assignment translates a 3D structure to 1D sequence of PBs. The input is the structure coordinate file from PDB, representing an X-ray structure with or without PTM. The algorithm uses 5 residues long window for each position. For each “nth” position, 8 dihedrals ψ_{n-2}, φ_{n-1}, ψ_{n-1}, φ_{n}, ψ_{n}, φ_{n+1}, ψ_{n+1}, φ_{n+2} are compared to the reference set of 16 PBs. The comparison is performed using the RMSDA criteria (*Root Mean Square Deviation on Angular values*) [59]:

$$RMSDA\,(V_1, V_2) = \sqrt{\frac{1}{2(M-1)} \sum_{i=1}^{M-1} \quad [\psi_i(V_1) - \psi_i(V_2)]^2 + \ [\phi_{i+1}(V_1) - \phi_{i+1}(V_2)]^2}$$

where, V_1 is the 8 dihedrals vector extracted from the 5 residues long window; V_2 is the 8 dihedrals vector corresponding to the compared PBs. PB, which gets lowest RMSDA is chosen as the representing conformation observed in the window.

3a.2.3 *Local structure entropy - Neq*

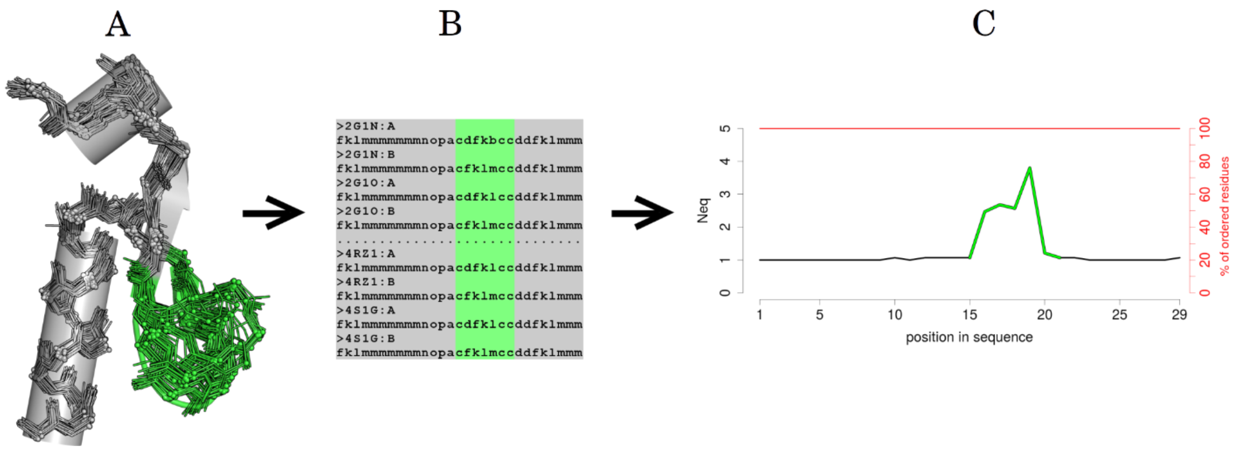
3D structures of a specific protein could be observed with different conformations in X-ray crystals, or during molecular dynamics simulations. This could be attributed to the intrinsic flexibility of the structure or the consequences of interactions with small molecules (ligand, cofactor, water molecules), or macromolecules (proteins, DNA, RNA). Under such scenarios, each of these 3D conformations would be assigned a different PB sequence (see Fig 3a.3). By analyzing the variation of PBs at each position, it is possible to investigate the local conformational changes in a protein structure.

The equivalent number of PBs (*Neq*) is a statistical measurement similar to Shannon entropy and represents the average number of PBs observed at a given position [53]. *Neq* are assigned using PTM-SD utility toolkit where *Neq* is calculated as follows:

$$Neq = exp\left(-\sum_{i=1}^{16} f_x . ln(f_x)\right)$$

where f_x is the frequency of PB *x* (*x* goes from *a* to *p*). A *Neq* value of 1 indicates that only one type of PB is observed, while a value of 16 is equivalent to a random distribution.

For example, *Neq* value around 6 would indicate that at the current position of interest, 6 different PBs are observed. An *Neq* exactly equal to 6 would mean that 6 different PBs are observed in equal proportions (1/6). By plotting the computed *Neq* value at each residue position (Fig 3a.3), it is possible to locate which protein regions have local conformational change, or in other words, which region of the structure represents backbone deformation.



```
>2G1N:A
fklmmmmmmmnopacdfkbccddfklmmm
>2G1N:B
fklmmmmmmmnopacfklmccddfklmmm
>2G1O:A
fklmmmmmmmnopacdfklccddfklmmm
>2G1O:B
fklmmmmmmmnopacfklmccddfklmmm
.............................
>4RZ1:A
fklmmmmmmmnopacdfklccddfklmmm
>4RZ1:B
fklmmmmmmmnopacfklmccddfklmmm
>4S1G:A
fklmmmmmmmnopacdfklccddfklmmm
>4S1G:B
fklmmmmmmmnopacfklmccddfklmmm
```

Figure 3a.3 ***Using Neq to understand protein backbone flexibility.*** *A) The protein backbone is assigned with PB sequence. The structures are superposed in 3D thereby yielding a superposition in 1D, as a sequence of PBs (shown in B). C) The Neq is then calculated as the equivalent number of PB at a given position. The green color in the plot maps on to the green highlighted region in the (A) and (B).*

3a.2.4 Normalization of the B-factor values

B-factor values are partly dependent on the resolution of the crystal and of the refinement process [147,307,308]. Also crystallographic contact packing and addition of stabilizing molecules can impact the B-factor values. Thus, in order to compare B-factor from several X-ray structures of the same protein, it is required to normalize the value. Raw B-factor values were normalized as recommended by Smith et al [309], starting by removing outliers values detected with the median-based approach - median absolute deviation (MAD).

The MAD was calculated from the median of the B-factor values using cran-R utility (www.cran.r-project.org/package=R.utils), *mad()* with 1.4826 as the consistency scaling constant [310,311]. The outliers were removed by defining upper and lower limit (median ± 2.5*MAD) of the B-factor data. Finally, the refined set of B-factors was normalized using:

$$nBfac \ = \frac{x-\mu}{\sigma}$$

where μ and σ are the mean and the standard deviation of the B-factor values (without outliers) respectively and x is the raw B-factor values of Cα (except outliers).

Most of the statistical analyses were done using Python programming language and R software [312].

3a.3 Results and discussions

3a.3.1 Impact of PTM on overall protein backbone conformational diversity

Using PTM-SD [136], the two most frequent PTMs were focused upon, N-glycosylation and phosphorylation. 3,092 and 1,307 chains were found containing 7,110 N-glycosylations and 1,873 phosphorylations in 100 and 22 organisms respectively. A non-redundant dataset, with less than 25% of identity between the corresponding UniProt sequences, was generated, resulting in the selection of 348 N-glycosylations (for 156 protein chains in 41 organisms) and 92 phosphorylations (for 75 protein chains in 12 organisms, see Table 3a.1).

3a.3.1.1 *Neq analysis*

Based on 16 PBs, *Neq* underlines the diversity of local conformation in a finer manner than the classical secondary structures (see Methods 3a.2.3). Figure 3a.4 shows the variations of PBs around the two PTMs ー N-glycosylation and phosphorylation. It is observed that the PTM sites do not exhibit any significant preferences for a particular local structure conformation. The *Neq* values are very high, ranging from 9.03 to 11.44 for N-glycosylation, and from 5.95 to 11.41 for phosphorylation, implying that these two modifications are observed in widely diverse structural contexts. Nonetheless, it is interesting to note that both types of PTMs have an overall different *Neq* profiles (see black curve in Figure 3a.4).

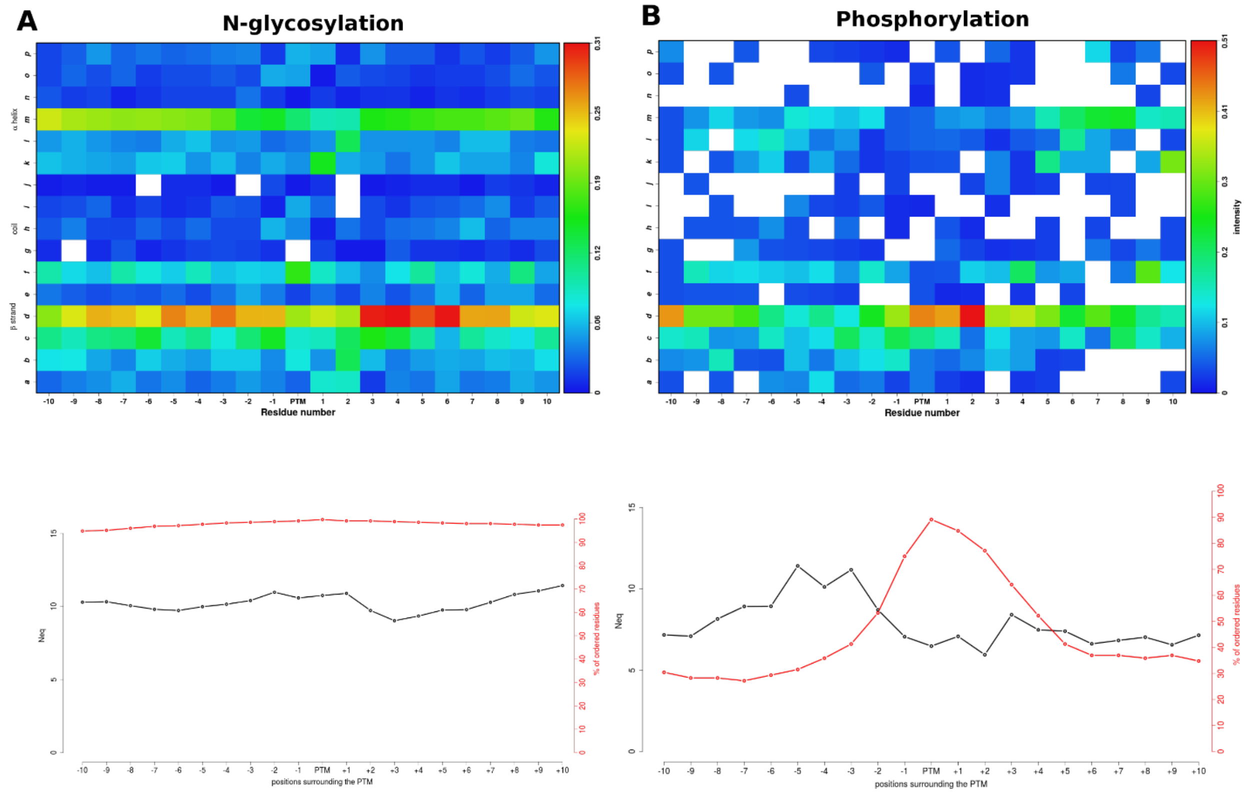


Figure 3a.4 ***Comparisons of PTM sites of N-glycosylation and Phosphorylation.*** *The top panels show the PB profiles of A) N-glycosylation and B) Phosphorylation. The PBs are plotted on the Y-axis and PTM position in the chain at X-axis. The colors are encoded according to the intensities as mentioned by the vertical bar on the left with blue depicting the least and red depicting the max. intensities. The white color or absence of a PB intensity corresponds to the missing region in the PDB file. The lower panels show the Neq analysis of A) N-glycosylation and B) phosphorylation. The Neq values are plotted on X-axis. The red curve indicate the amount of data used to compute Neq values, or in other words the percentage of ordered residues at each position in the X-ray crystal.*

For N-glycosylation, the PTM site position presents an *Neq* = 10.76 which is extremely high. This suggests that N-glycosylated residues have backbone conformation as diverse as 2/3 of the backbone conformations observed in proteins. Additionally, the surrounding positions of the PTM sites show the same level of diversity, with *Neq* values fluctuating around 10.

For phosphorylation, the *Neq* profile is quite different. First of all, as indicated by the red curve on Figure 3a.4, the surrounding positions of phosphorylation sites are mainly disordered.

The farther the positions are from the PTM sites, the higher is the disorder in the structure; suggesting that

less residues were available at these positions in the PDB chains to be used for the PBs assignments and the *Neq* computation. However, the data used is diverse enough to reach high level of *Neq* (6.48) computed at the PTM position. Preceding positions, i.e. upstream -8 to -2 show even higher diversity. It is important to confirm that the absence of data in the surrounding positions is not the consequence of phosphorylation sites located at the N- or C- terminus; indeed, only 12 of them (out of 92) are close to the protein extremities.

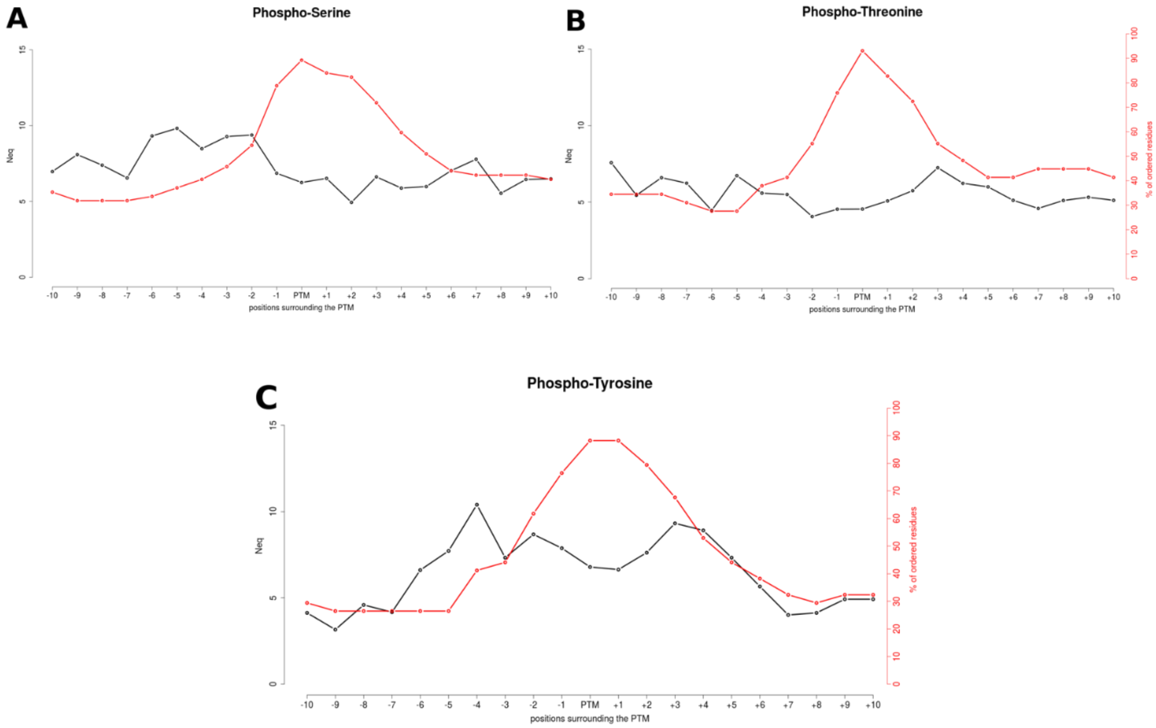
3a.3.1.2 *Analysing structural conformations using PBs*

A more precise analysis of the distribution of each type of PBs is depicted in (see Fig 3a.4 top panels). The intensity of the color at each position depicts percentage derived from the frequency of the local conformations occurring at the position. Resulting color underlines that N-glycosylation and phosphorylation sites are observed for all types of local conformations, almost any kind of PBs (except PBs *g for both,* and, *h, j,* and *p* for phosphorylation).

The conformations of the N-glycosylation sites and their surrounding residues are mainly associated with the PBs *d* and *m*. However, this proportion does not exceed 31%. It is interesting to note that the positions +3 to +6 downstream of the N-glycosylation sites are significantly observed in a PB *d* conformation. This illustrates the fact that, ~1/3 of the times N-glycosylation site precedes a β-strand conformation.

For phosphorylation, the modification sites have a preference of PB *d*, the cores of β-strands, in a little over 40% of the cases. The vicinity of the phosphorylation sites is also observed with a wide variety of conformations, however a slight preference was observed for the PBs *b*, *c*, *d*, *f*, *l* and *m*. It should be noted that more than 50% of the phosphorylation sites are separated by two residues of a PB *d*.

It is important to understand that data used here provides information on the backbone conformation of PTM sites when the modifications are present, but do not obviously reflects the backbone in the absence of modifications. Additionally, while phospho-serine and phospho-threonine share similar PB profiles, they are distinct from phospho-tyrosine (see Fig 3a.5 and Table 3a.2). The modified residues were observed in a large set of backbone conformations for all three



cases, but the preferences for the core β-strand conformation (PB *d*) is greater in the case of Ser and Thr. On the contrary, Tyr does not show any clear preference for a local structure.

Figure 3a.5 ***Flexibility profile for different types of Phosphorylation.*** *Neq distribution curves (black) gives an insight into the extent of local structural changes at the phosphorylation site and its sequential neighborhood. The red curve represents the percentage of available data for calculating Neq at a position. Higher the percentage better is the confidence. As can be seen that the Neq profiles of phospho-serine (A) and phospho-threonine (B) are similar in topology while they both differ from the Neq profile of phospho-tyrosine (C).*

3a.3.2 <u>Local backbone diversity compared to global backbone diversity in modified structures.</u>

In order to compare the flexibility of the PTM region with the rest of the protein, we selected a large number of 3D chains corresponding to the same protein. Each chain was solved with a single PTM at identical sequence positions. 4 different proteins were studied, covering 3 types of PTMs: N-glycosylation in renin endopeptidase and liver carboxylesterase, phosphorylation in cyclin-dependent kinase 2 (CDK2), and methylation in actin. A total of 471 PDB chains were used for the analysis.

3a.3.2.1 *Neq analysis*

The *Neq* profiles of modified sites and surrounding positions were compared with those of all other positions in the proteins. Figure 3a.6 shows the example of one N-glycosylated residue, at position 141, in renin endopeptidase. Figure 3a.6A is a zoom around the PTM site, while Figure 3a.6B shows the *Neq* all along the protein. In this example, the maximum entropy is found at position 234, with a *Neq* value of 7.13. This position and its surroundings are associated with the maximum number of missing residues (red curves in Figure 3a.6B), suggesting a highly flexible region. It corresponds to what Zhang *et al.* 2007 defined as a Dual Personality Fragments (DPF): a protein region, which can appear as either ordered or disordered in crystal structures. It is suggested that DPFs are potential targets of regulation by allostery or PTMs [313]. Herein, the flexible region from position 230-238 (see red fragment in Fig 3a.6C) is not annotated as PTM site and also does not interact with ligands in the structures; but interestingly the positions 230 to 234 are known to be missing in a second isoform of this protein.

In comparison, the backbone of the modified residue is always ordered and presents slight deformations with *Neq* of 1.94. Its immediate neighbor positions are in the same range, with slightly higher values in positions -6, -1, and +1 (*Neq* values 2.58, 2.10, and 2.53 respectively). In Figure 3a.6A, the PTM site seems to be slightly more deformable than majority of its surrounding positions. Using a larger scale (Fig 3a.6B) this deformation does not seem to be significantly different than other deformable parts along the sequence. To quantify it precisely, statistical tests were performed for each case (see Table 3a.4 and Fig 3a.7).

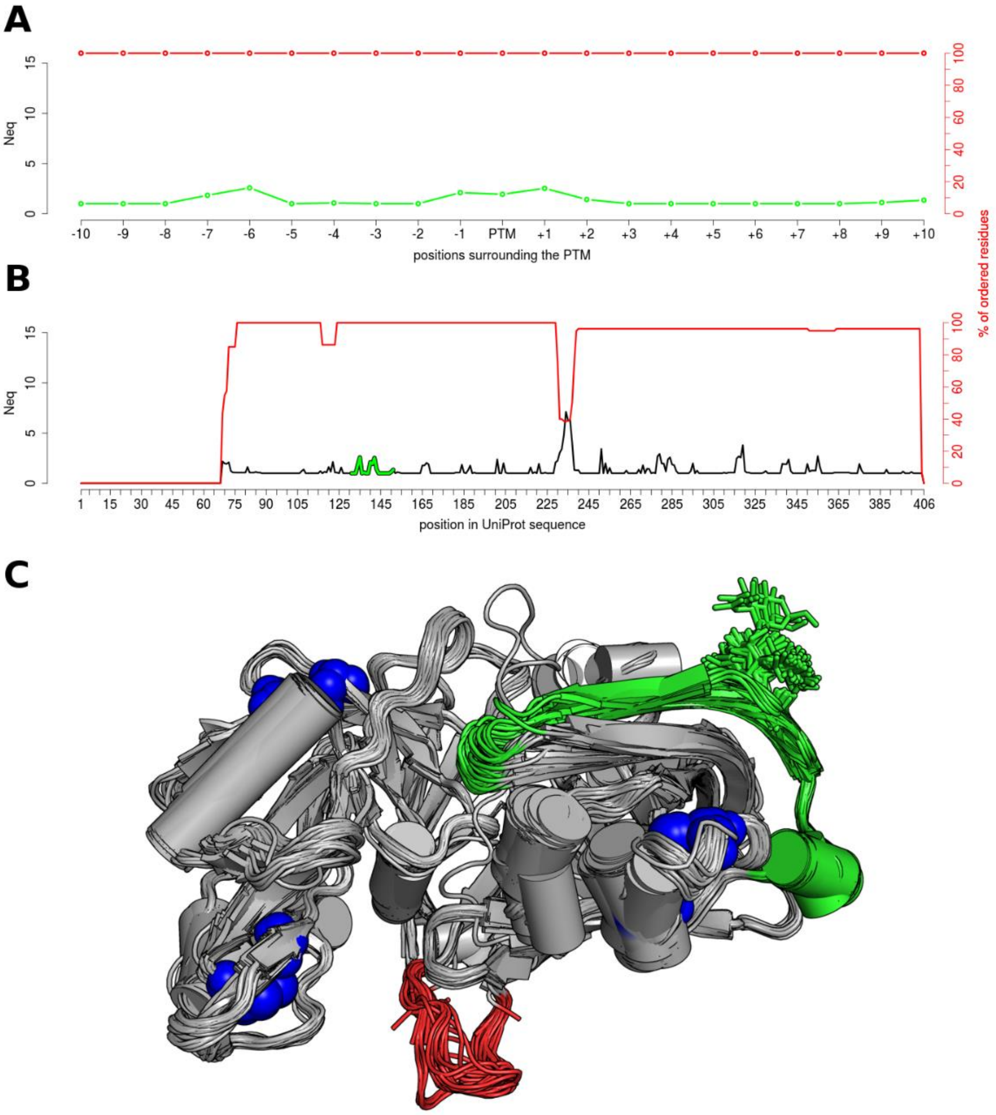


Figure 3a.6 ***N-glycosylatation on the Asn141 of the human renin endopeptidase (P00797).*** *The Neq profiles are given at a local scale (**A**), for the surrounding positions of the PTM site (colored in green), and at a global scale (**B**), for all sequence positions. (**C**) The 80 structures used for the computation were aligned on the backbone, and represented in cartoon. The glycosylated position is shown in green sticks, the disulfide-bridges in blue spheres, and the DPF loop colored in red.*

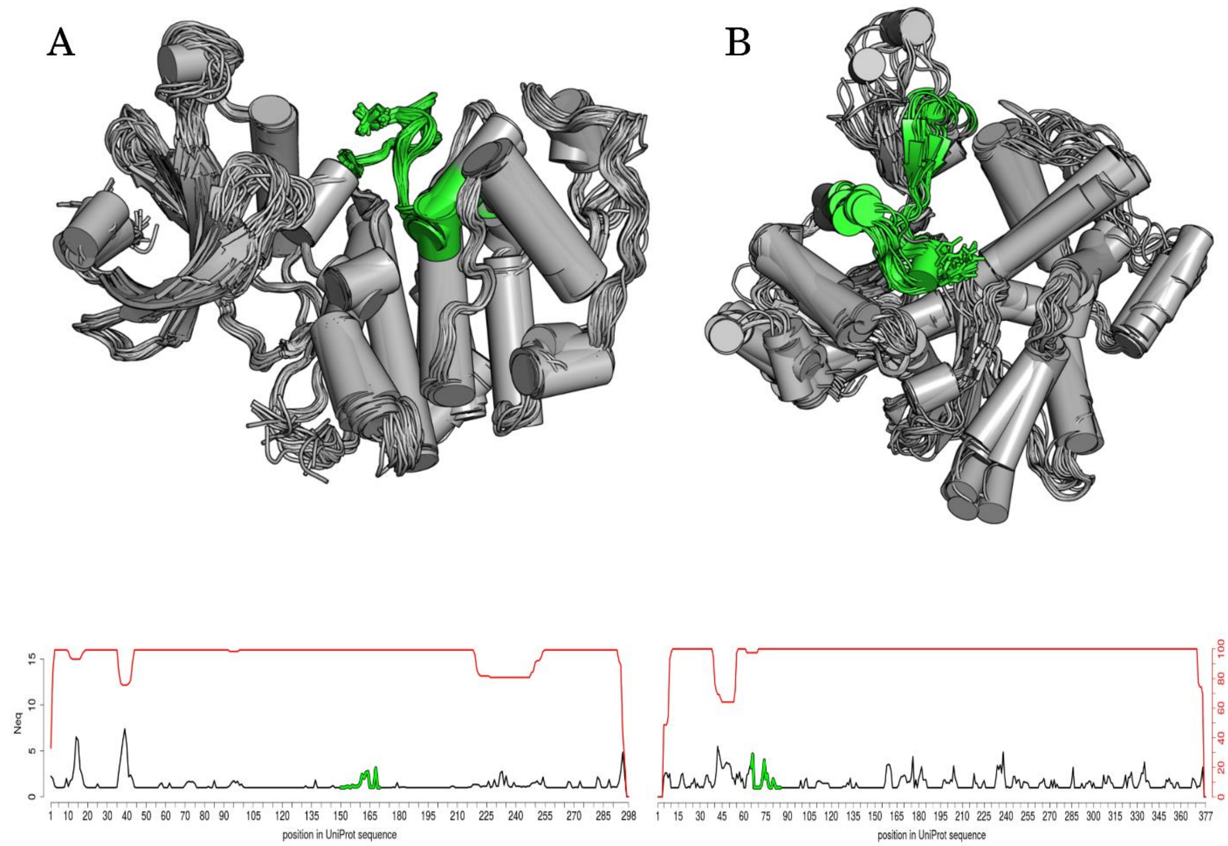


Figure 3a.7 ***Local structure comparison of phosphorylation and methylation.*** *Top panel shows the superposed chains containing phosphorylation (left) and methylation (right). The green color in the cartoon represented structure depicts the PTM. Lower panel, shows the Neq analysis of all chains containing phosphorylation and methylation. On x-axis is the Neq and residue positions (UniProt) is plotted on the y-axis. The red curve shows the %age of structural information available (ordered residues). The green trace marks the PTM and its neighbors. A) Threonine 160 phosphorylation in CDK2. B) Histidine 75 methylation in Actin.*

3a.3.2.2 *Statistical significance of differences in local and global flexibility*

Firstly, the Shapiro-Wilk (SK) test provides extremely low *p*-value, in all the cases, forcing the rejection of the null hypothesis (see columns 3 and 4 of Table 3a.4). This underlines that *Neq* values for PTM-region and the rest of the protein does not follow a normal distribution. Therefore, the nonparametric Mann-Whitney-Wilcoxon (MWW) test was used to see if *Neq* profiles observed in the PTM-region are significantly different from those observed in the rest of the protein. With a significance level, risk α = 5%, only the phosphorylated Thr-160 in the Cyclin-dependent kinase 2 protein and its neighboring positions have a significantly different *Neq* profile than the rest of the protein; the *p*-value being equal to 0.0194.

It should be noted that in both cases of N-glycosylation, no significant differences were observed between the *Neq* profile of the PTM-region and the *Neq* profile of the rest of the protein.

Table 3a.4 ***Statistical tests for the 4 proteins.*** *Shapiro-Wilk (SW) test checks if data follows Normal law distribution, while Mann-Whitney-Wilcoxon (MWW) is a nonparametric test that compared mean values. Are indicated the size of samples (n), the calculated statistics (stats), and the p-values. The chosen risk α is equal to 5%, the significant p-values allow dismissing the hypothesis H_0 and are colored in red.*

		SW *Neq* PTM region vs Normal law	SW *Neq* rest of the protein vs Normal law	MWW *Neq* PTM region vs *Neq* rest of the protein	SW *Neq* all protein with 0 PTM vs Normal law	SW *Neq* all protein with 1 PTM vs Normal law	MWW *Neq* 0 PTM vs *Neq* 1 PTM
Renin endopeptidase P00797 (Human) N-glycosylation	n	21	316	21/316	364	337	364/337
	statistic	0.6790	0.4312	3571	0.5276	0.4429	59812.5
	p-value	**1.52E-05**	**2.16E-30**	**5.08E-01**	**4.52E-30**	**5.34E-31**	**5.13E-01**
Liver carboxylesterase 1 P23141 (Human) N-glycosylation	n	21	507	21/507	529	528	529/528
	statistic	0.5522	0.4383	5582	0.4762	0.4436	138025.5
	p-value	**6.40E-07**	**6.52E-37**	**6.19E-01**	**1.26E-36**	**2.12E-37**	**6.60E-01**
Cyclin-dependent kinase P24941 (Human) phosphorylation	n	21	275	21/275	296	296	296/296
	statistic	0.6792	0.4102	3687	0.4927	0.4299	39350
	p-value	**1.53E-05**	**5.10E-29**	**1.94E-02**	**3.06E-28**	**1.44E-29**	**1.32E-02**
Actin P68135 (Rabbit) methylation	n	21	350	21/350	371	371	371/371
	statistic	0.7082	0.7201	3758	0.6745	0.7174	78367.5
	p-value	**3.49E-05**	**8.32E-24**	**8.53E-01**	**4.34E-26**	**1.42E-24**	**7.17E-04**

3a.3.3 <u>B-factors and comparison of backbone mobility with and without modifications</u>

In comparison to *Neq*, the B-factor does not give a measure of the deformation of the backbone, but could be used to represent its mobility in the crystal context. For each of the 4 proteins of interest, the B-factors of the Cα were extracted from every PDB chain. After normalization (see section 3a.2.4), the B-factors were averaged for each, structurally available, position along the sequence. The same statistical analyses, as applied to *Neq*, were performed with the B-factors in order to compare the backbone mobility in the PTM areas, and in the rest of the protein (Fig **3a.8**). It is observed, using B-factors that N-glycosylations do not have a significant impact on the protein structures backbone. While phosphorylation of Thr-160 residue in CDK2 have a smaller B-factor compared to non-phosphorylated Thr-160. These trends are in accordance with the trends observed with *Neq* distribution.

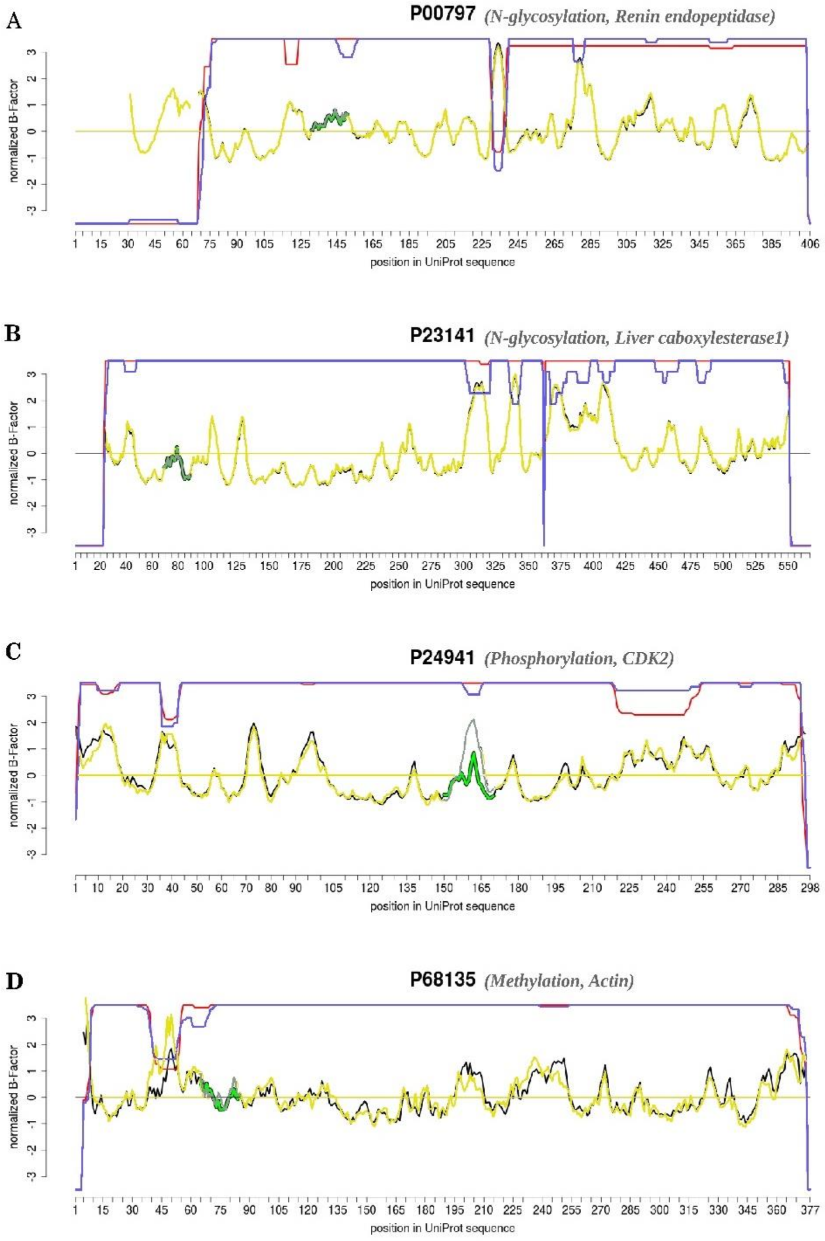


Figure 3a.8 ***Normalized B-Factor distribution for different proteins with and without PTMs***. *Each plot depicts normalized b-factor distribution of the protein with PTM (black line) and without PTM (yellow trace). The green colors highlights the PTM site and its neighborhood in structures with PTM (light green) and without PTM (pale green trace). The blue and red lines shows the position-wise data availability in PDB. The b-factor trends match with the trends observed in Neq analysis. Major differences are seen in B-factor values during phosphorylation while least difference is observed in N-glycosylations.*

3a.4 Conclusions and perspectives

PTM-SD and its user-friendly query search and utility toolkit helped in datasets extraction for different PTMs and compare them with the PTM- ‘null’ X-ray structures. Thus the effects of specific PTMs on protein backbone conformation was analysed. The conformational backbone diversity of modified residues and its close neighborhood positions is compared for the two most frequent PTMs; N-glycosylation and phosphorylation. Secondly, special case studies are performed on 4 different proteins that individually undergo an N-linked glycosylation, phosphorylation and methylation. For these, the local and global backbone diversity observed in X-ray data when a single PTM is present is compared. Finally, the backbone diversity with and without the PTM was compared for all four of the case studies, Table 3a.2.

The backbone analysis of the two examples of N-glycosylation, showed that the addition of the glycan neither impact the local nor the global backbone conformation of the proteins. However, the methylation on actin structure induces a local increase of the backbone diversity at the PTM site region, highlighting a higher deformation of this part of the protein. However, no effect on the intrinsic mobility of this region has been observed (*same B-factor profiles with or without the PTM*). Unfortunately, the large variability of ligands found associated with actin in X-ray data used in this study, does not allow to propose an effect of the methylation at a global scale, see Figure 3a.9.

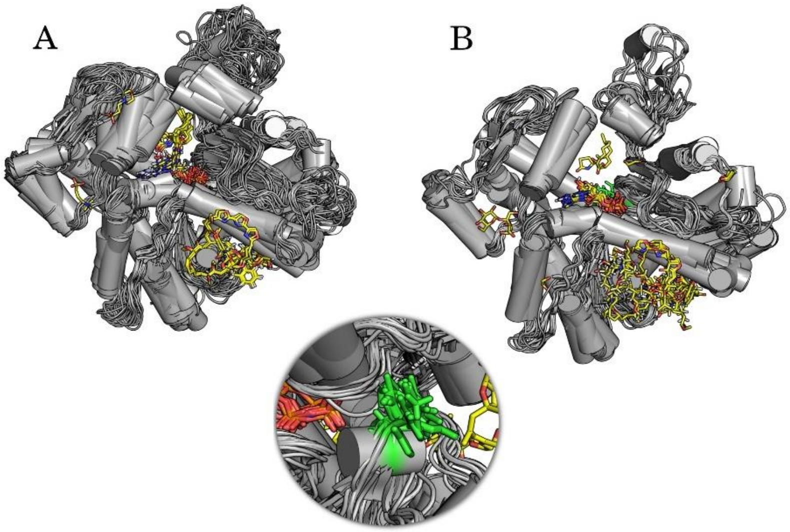


Figure 3a.9 ***Impact of methylation on Actin and its ligand binding.*** *Superimposed Actin structure along with its binding ligands. A) Actin with its ligand, without the methylation of H75. The ligands are represented as stick models. B) Actin with ligands while Histidine at 75 is methylated. Subtle*

changes can be observed in binding pattern of the ligand due to the presence of PTM. However, the data-set is too small and due to variety of ligands precise the nature of such changes cannot be commented upon.

It is clearly observed through *Neq* as well as normalized B-factor values that the phosphorylation site and its neighborhood positions display a backbone diversity that is significant. The comparison of modified structures of CDK2 with the unmodified ones reveals that the phosphorylation on the activation loop at Thr 160 have several local effects. It rigidifies the backbone (lower *Neq* and lower B-factor) locally while increasing the deformation of two other regions, near Thr14 - Tyr15, and near Thr39. These sites pertain to three other phosphorylation sites related to CDK2 activity (T14, Y15) and subcellular localization (T39). The observed rigidity in the backbone is in agreement with the proposition made by Xin and Radivojac [290]. They proposed that phosphorylation, by introducing new H-bond and salt bridges in the local neighborhood leads to a conformational shift to the lowest valley in the energy landscape of the protein. This decrease of energy was also observed by Groban and coworkers in their attempt to computationally predict the conformational changes of the CDK2 activation loop induced by the phosphorylation [314]. Finally Gao and Xu [291] suggest, by analyzing NMR structures, that disorder-to-order transition might be introduced by Threonine phosphorylation.

Despite the intrinsic link between PTM and protein function, the molecular effects of the modifications on the protein structures and dynamics remains poorly understood. Our study, like previous systematic studies of structural data of modified and unmodified protein [290,291], shows that these effects could be of multiple types (stabilization and destabilization), at different scales (at the local PTM region, in other part of the protein as allosteric effect, or at a global level), and depend of the PTM types. However, in order to propose general rules for the molecular impact of each type of PTMs, additional structural data related to the large amount of PTM annotations already available is needed. In the scope of a systematic study, these data have to be used carefully. Indeed, many factors, independent of the presence of PTMs, could have affected the structure of the proteins, such as the crystallographic packing, the presence of engineer mutations or cross-links to help crystallization process, the presence of ions, ligands and protein partners in contact with the protein structure of interest.

Molecular modeling of PTMs combines with molecular dynamic simulation is an interesting alternative. Some recent computational studies have investigated the effect of PTMs [300] on the stability of specific proteins. However, the success of such simulations also rely upon the growing number of experimental data, for the development of accurate PTM force field parameters. Once standardized, such molecular dynamics protocols can be of great use to understand impact of multiple PTMs on the structure of a protein.

A critical caution for any systematic, PDB based structural analysis is the uncertain nature of missing regions. As expected, in our datasets numerous PDB structures lack coordinates for some regions, which is depicted by a dip of the red curves in *Neq* plots. These particular regions mainly correspond to disorder regions in protein, which diversify the functional spectrum of proteins [315] by expanding their protein protein interactome. The selectivity of interacting partners and order-disorder transition of the protein structures is regulated by PTMs, and most of the times by phosphorylations. [316]. During the analysis of CDK2, identical structures with missing coordinates in the catalytic loop (functional domain) are found to be flagged as a Dual Personality Fragment [313,316]. However, this region gets ordered based on the phosphorylation of Thr 160. It may be suggested that this selectivity of interacting partners for CDK2 may well also be impacted by the number of phosphorylations in and around the catalytic domain.

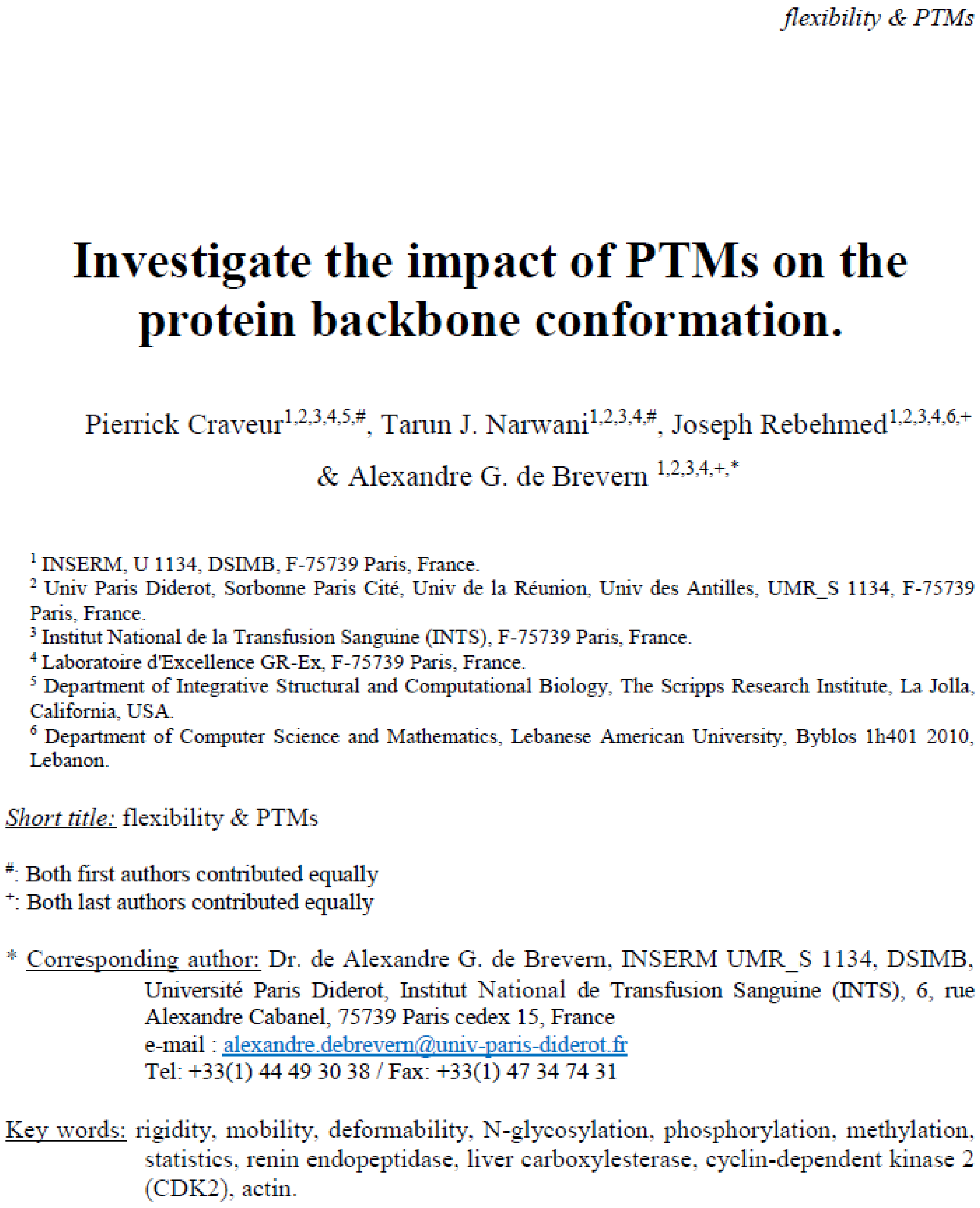
ACKNOWLEDGEMENTS

I would like to thank Dr. Pierrick Craveur for developing PTM-SD that is quintessential to this study. I also thank him and Dr. Joseph Rebehmed for extracting the data and their critical advices and suggestions during the analysis.

Dissemination of results

The results of the chapter 3b have been published as a scientific poster at ADELIH conference, PTM- from bench to bed side, held in Paris in Oct, 2016. It got the best poster award.

The latest development is that the manuscript for a research article consisting of results from Chapter 3a have been written and revised. We are waiting for one of the authors final comments before sending it to journals. The work will be published as:



flexibility & PTMs

Investigate the impact of PTMs on the protein backbone conformation.

Pierrick Craveur[1,2,3,4,5,#], Tarun J. Narwani[1,2,3,4,#], Joseph Rebehmed[1,2,3,4,6,+] & Alexandre G. de Brevern [1,2,3,4,+,*]

[1] INSERM, U 1134, DSIMB, F-75739 Paris, France.
[2] Univ Paris Diderot, Sorbonne Paris Cité, Univ de la Réunion, Univ des Antilles, UMR_S 1134, F-75739 Paris, France.
[3] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.
[4] Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.
[5] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA.
[6] Department of Computer Science and Mathematics, Lebanese American University, Byblos 1h401 2010, Lebanon.

Short title: flexibility & PTMs

[#]: Both first authors contributed equally
[+]: Both last authors contributed equally

* Corresponding author: Dr. de Alexandre G. de Brevern, INSERM UMR_S 1134, DSIMB, Université Paris Diderot, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
e-mail : alexandre.debrevern@univ-paris-diderot.fr
Tel: +33(1) 44 49 30 38 / Fax: +33(1) 47 34 74 31

Key words: rigidity, mobility, deformability, N-glycosylation, phosphorylation, methylation, statistics, renin endopeptidase, liver carboxylesterase, cyclin-dependent kinase 2 (CDK2), actin.

3b: Characterization of Dual Personality Fragments - DPF

3b.1 Introduction

A major problem when using X-ray crystal data from Protein Data Bank (PDB) is related to missing regions or regions with no coordinates. The length of such regions range from a single amino acid to 20 to 25 residues and sometimes larger. Missing regions in a crystal structure arise when the X-ray diffraction pattern from the crystal is too ambiguous for a crystallographer to resolve to certain atom or molecule. Thus signifying regions in the protein structure that are highly mobile or deformed that they cannot be snapped by X-ray or CryoEM. These highly mobile, deformed regions existing as an interconverting ensemble of structures in a protein and are functionally attributed as natively unfolded or intrinsically unstructured or intrinsically disordered regions (IDRs) [317,318]. The IDRs have been shown to be of crucial importance in protein-protein interactions since they can interact with more than one partners, given their structural malleability [319,320]. Some prominent examples of proteins that are natively unfolded are: Tau protein [321], MAP2 [322], α-synuclein [317], and Myelin Basic Protein [323].

However, there lies a subtler side of disordered proteins that pertains to regions that are either ordered or disordered based on their environmental context or interacting partners [315,317]. Such regions in proteins are called conditionally disordered regions. Many enzymes and viral peptides behave in a similar manner where the structural orderliness changes with respect to their binding partner. Between these structurally disordered proteins and those with well defined three dimensional structure lies a conceptual boundary thus defining the structure–disorder continuum [315].

3b.1.1 *Dual Personality Fragments*

Assumed to be lying at the boundary of structure-disorder continuum, these protein fragments can transit from order to disorder and therefore exhibit properties of both the states. Thus many of such protein regions are visible in crystal structures, for eg. the catalytic loop of an enzyme. These fragments have been described by various names. They are called “dual personality fragments” by Zhang *et al*., 2007 [313], “Ambiguous regions” by Le Gall *et al*., 2007 [324], and “twilight zone” by Szilágyi *et al*., 2008 [315,325]. Due to the presence of such structures, the earlier version of order and disorder as binary states is being challenged. Recent research defines the structural space as a continuum with structure and disorder being the two extremes [315]. The DPF lie at the center of this structure continuum. Figure 3b.1 shows an example that undergoes order - disorder transition [313].

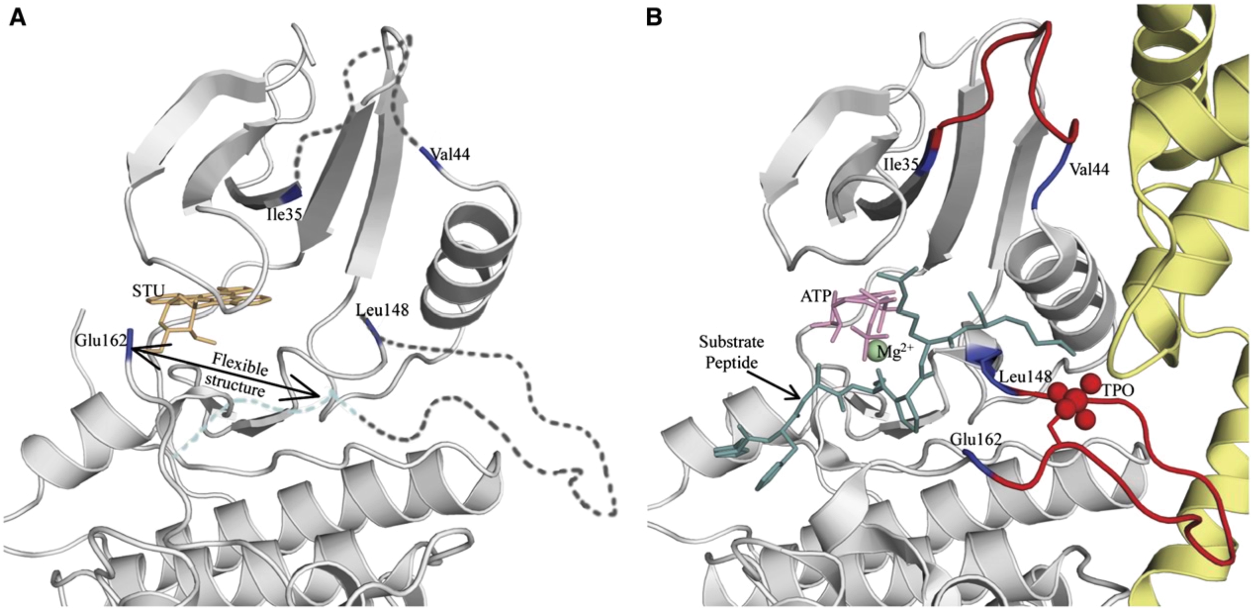


Figure 3b.1 ***An example of disorder to order transition.*** *Shown here are the two crystal structure of same protein, Cyclin dependent Kinase 2 (CDK2). A) shows the CDK2 bound to inhibitor staurosporine (STU) thus deactivating the kinase. Thus PBDid 1AQ1:A is CDK2 in inactive state. Two regions are seen missing (represented by dashed lines). B) An active state structure of CDK2, identifiable by presence of its substrate peptide and ATP. The regions missing in A) can be seen in the active state (shown in red).* $^{+}$*taken from Zhang et al, 2007, Structure [313]*

3b.1.2 *Identification of DPF*

Disorder refers to a highly flexible ensemble of structures co-existing such that a definitive structure cannot be identified. Therefore, these are highly difficult to detect. NMR can show the ensembles with large deviation in certain regions but the technique is limited by the small size of proteins. Therefore, there is not enough data generated for a systematic analysis. On the other hand, X-ray crystallography can indirectly indicate disorder. Due to their highly flexible state, the disordered region would be shown as a distortion or noise on the diffraction pattern. Therefore, such region would be missing in the X-ray coordinate files. Since, DPF are structures that can transit between order and disorder, therefore some of the ordered state of a DPF would be crystallized [313].

The rationale behind using the X-ray data is founded on the generation of crystals for crystallography. It is common to observe multiple crystal structures of the same protein. Often, with motive of improving the resolution, crystallographer alters the crystallization conditions, or introduce mutations, or add co-factors to the crystal solution, etc. From a crystallographer's perspective, these will yield different crystals with better resolution than the other. However, these conditions also become ideal to identify DPFs. With changes in crystallization conditions, the protein behavior will also change and many a times, this can be a trigger for order to disorder or vice versa transformation of a DPF [313].

Therefore, DPFs can be identified by using different crystal structures of the same protein and then comparing them.

3b.1.3 *Characterization of DPF*

Due to lack of a definitive structure, the disorder regions (IDRs or IDPs) are characterized based on the sequence related properties. However, DPF can also express as structured regions and therefore can be characterized based on sequence as well as structural features. So far, only Zhang *et al.* have systematically characterized DPF in 2007 that is more than a decade ago. Following is a summary of their results and observations:

A) 92.3% of DPF are less than 10 amino acid residues in length.

B) No specificity in neighborhood of a DPF was observed as 50% of the dataset had a DPF in vicinity of a disordered region while rest 50% had an ordered region next to DPF.

C) Structural analysis was performed just with DSSP assignments. It was observed that most of the DPF have been assigned as 'C' (coil) by DSSP. However, a striking 27% was assigned in regular secondary structures with 20% assigned as helices (collectively H, G, I) and 7% as sheets.

D) Amino acid propensities were identified for ordered, disordered and DP fragments. Smaller and hydrophilic amino acids like, Ser, Ala, Lys, Glu, and Gly were found abundant in disordered regions and deficient in ordered regions. Many polar and charged residues like, Asp, Thr, Gln, Pro, and Arg are found to have similar preferences for ordered and disordered regions. But these amino acids have higher propensities to be found in DPF regions as well. Ordered regions also some often occuring residues associated to them like, Iso, Phe, Typ, Tyr, His, Met, and Cys. Of these, Thr, Arg, Gly, Asn, Pro, and Asp were found exclusively in DPF.

E) From a similar analysis of clustered amino acids based on their physicochemical properties, it was determined that disordered regions have affinity towards polar residues while DPF have affinity towards hydrophobic and charged residues.

3b.1.4 *Functional importance of DPF*

After characterising DPF based on certain sequence and structure features, Zhang *et al.*, 2007, also characterized functional importance of DPF. They used ScanProSite to predict functionally important sites in their dataset. DPF were shown to be strongly associated with post-translational modifications. About 70% of DPF were tagged with predicted PTM sites by ScanProSite. Moreover, it was found to be more likely that in 20% of cases a PTM site is to be found within 5 residues of DPF.

3b.1.5 *Dynamics in DPF*

So far, to the best of our knowledge, there have been no attempt to understand the role of flexibility in DPF or structural space lying at the center of the structural continuum. In the current context of our objectives, it becomes fitting to study the inherent role of flexibility in highly dynamic structures such as DPF. The analysis of local structure behavior in the DPF transitions can be very useful and pivotal in our understanding of structural flexibility. Therefore, it was decided to redo the structural and sequence characterization of DPF. Thereafter model the missing regions in disordered DPF to perform short molecular dynamics on both states of DPF to catalogue the local structure behaviors.

Moreover, as can be seen in preceding sections that the study from 2007 does not detail in to the structural aspect of DPF and characterized mostly based on sequence features.

3b.2 Methods

3b.2.1 *Dataset preparation*

A redundant dataset of crystal structures were extracted from PDB (2016) hosted at www.rcsb.org [134]. Each chain in a pdb entry was treated separately as a single structure. Short chains having 8 residues or less were removed. Only crystal structures having resolution better than 2.5 Å and R-factor higher than 0.25 were selected. Thus, a total of 192163 structures (individual chains) were used for all analysis.

3b.2.2 *Feature assignment*

All the structures were assigned with different structural and sequence features in order to systematically analyse the DPF, order and disorder regions. Besides amino acid sequence, various features assigned for each structure includes:

a) *DSSP*- All the 8 states of DSSP [144] was assigned for each residue. For methodological details, please refer to section I.6.2.1.
b) *Segno*- Segno assignments do differ a lot from DSSP as both have an agreement of only ~82% [182]. Segno was used specifically for the assignment of Polyproline helices which is one of its nine state assignment. For methodological details, please refer to section S1.3.
c) *ProMotif*- A 1995 program [186] is based on DSSP-like approach and efficiently assigns many structural motifs like, β and ɣ turns, β-bulges, β-hairpins, Ψ loops, β-α-β units and disulfide bridges. It uses various geometrical properties based on hydrogen bonding pattern and distances between consecutive C_α atoms to identify these structural motifs. ProMotif v2.0 was used to specifically include β-turns information in DPF regions.
d) *Protein blocks*- Besides, secondary structures PBs can provide precise approximation of the local structure which can be highly useful in characterizing DPF. Therefore, all the structures were also assigned with 16 PBs. For methodological details, please refer to section I.6.3 or 3a.2.2.
e) *B-factors*- Crystallographic temperature factors were extracted from pdb files for all the structured regions.

f) *Relative solvent accessibility*- The solvent accessibility of each structured residue was calculated using Naccess v2.1.1 (www.bioinf.manchester.ac.uk/naccess/) which provides both absolute as well as relative solvent accessibility for each residue.

g) *Disorder*- As it is the unstructured region, the amino acid sequence was extracted by comparing the SEQRES and ATOM records of the pdb coordinate file.

3b.2.3 *Pairwise sequence alignment*

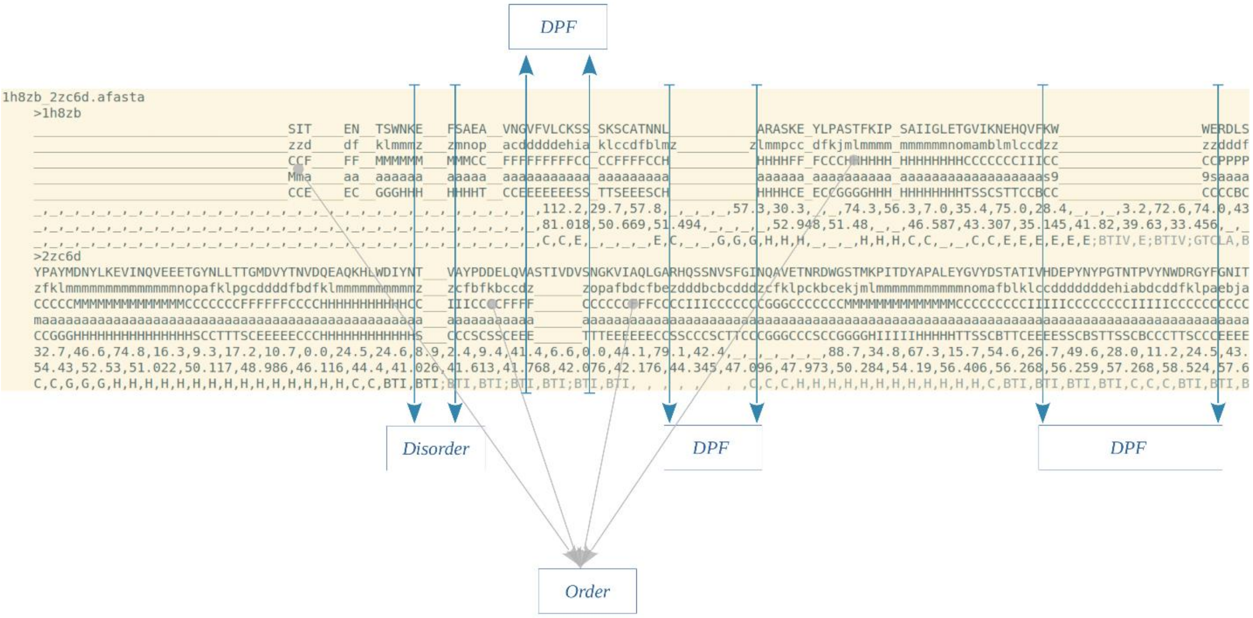
Pali v3.0 structural alignment database (Phylogeny and Alignment of homologous protein structures) [326] contains the structure based, domain level pairwise alignment of all the homologous proteins. The proteins are decomposed into domains according to SCOPe v2.04. Therefore, alignments for 192163 structures were extracted from Pali. A total of 2168 pairwise alignments were obtained from Pali.

3b.2.4 *Identification of DPF, Order and Disorder regions*

The different regions of interest namely, DPF, Order, Disorder were extracted from the Pali pairwise alignments. A group of Python methods were written to extract these regions by simply defining the following criterion:

If a given alignment has PB ‘z’ represented more than twice in between an alignment with gaps in one structure (zz-----z) but a defined region in the counter structure in the alignment (bcddddef). It was marked as a DPF. If the counter structure also lacks a defined region, then if will be classified as Disorder. The rest of the alignment where both the regions consists of well formed PBs and no ‘z’ is present, they were marked as Order. Figure 3b.2 represents a schematic example of the logic.

All the analyses were performed using python, R and bash scripts. Images were generated using R.



```
1h8zb_2zc6d.afasta
    >1h8zb
    ________________________________SIT____EN__TSWNKE___FSAEA__VNGVFVLCKSS_SKSCATNNL___________ARASKE_YLPASTFKIP_SAIIGLETGVIKNEHQVFKW__________________WERDLS
    ________________________________zzd____df__klmmmz___zmnop__acddddehia_klccdfblmz__________zlmmpcc_dfkjmlmmmm_mmmmmmnomamblmlccdzz__________________zzdddf
    ________________________________CCF____FF__MMMMMM___MMMCC__FFFFFFFFFCC_CCFFFFCCH___________HHHHFF_FCCCHHHHHH_HHHHHHHHCCCCCCCIIICC__________________CCPPPP
    ________________________________Mma____aa__aaaaaa___aaaaa__aaaaaaaaaaa_aaaaaaaaa___________aaaaaa_aaaaaaaaaa_aaaaaaaaaaaaaaaaaas9__________________9saaaa
    ________________________________CCE____EC__GGGHHH___HHHHT__CCEEEEEEESS_TTSEEESCH___________HHHHCE_ECCGGGGHHH_HHHHHHHHTSSCSTTCCBCC__________________CCCCBC
    _,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,112.2,29.7,57.8,_,_,_,_,57.3,30.3,_,_,_,74.3,56.3,7.0,35.4,75.0,28.4,_,_,_,3.2,72.6,74.0,43
    _,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,81.018,50.669,51.494,_,_,_,_,52.948,51.48,_,_,46.587,43.307,35.145,41.82,39.63,33.456,_,_
    _,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,C,C,E,_,_,_,_,E,C,_,_,G,G,G,H,H,H,_,_,_,H,H,H,C,C,_,_,C,C,E,E,E,E,E,E;BTIV,E;BTIV;GTCLA,B
    >2zc6d
    YPAYMDNYLKEVINQVEEETGYNLLTTGMDVYTNVDQEAQKHLWDIYNT___VAYPDDELQVASTIVDVSNGKVIAQLGARHQSSNVSFGINQAVETNRDWGSTMKPITDYAPALEYGVYDSTATIVHDEPYNYPGTNTPVYNWDRGYFGNIT
    zfklmmmmmmmmmmmmmmmnopafklpgcddddfbdfklmmmmmmmmmz___zcfbfkbccdz______zopafbdcfbezdddbcbcdddzcfklpckbcekjmlmmmmmmmmmmmnomafblklccdddddddehiabdcddfklpaebja
    CCCCCMMMMMMMMMMMMMMMCCCCCCCFFFFFFCCCCHHHHHHHHHHCC___IIICCCCFFFF______CCCCCCFFFCCCCIIICCCCCCCGGGCCCCCCCMMMMMMMMMMMMMCCCCCCCCCIIIIICCCCCCCCCIIIIICCCCCCCCCCC
    maaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa___aaaaaaaaaaa______aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
    CCGGGHHHHHHHHHHHHHHHSCCTTTSCEEEEECCCHHHHHHHHHHHHS___CCCSCSSCEEE______TTTEEEEEEECCSSCCCSCTTCCCGGGCCCSCCGGGGHIIIIIHHHHHTTSSCBTTCEEEESSCBSTTSSCBCCCTTSCCCEEEE
    32.7,46.6,74.8,16.3,9.3,17.2,10.7,0.0,24.5,24.6,8.9,2.4,9.4,41.4,6.6,0.0,44.1,79.1,42.4,_,_,_,_,_,_,88.7,34.8,67.3,15.7,54.6,26.7,49.6,28.0,11.2,24.5,43.
    54.43,52.53,51.022,50.117,48.986,46.116,44.4,41.026,41.613,41.768,42.076,42.176,44.345,47.096,47.973,50.284,54.19,56.406,56.268,56.259,57.268,58.524,57.6
    C,C,G,G,G,H,H,H,H,H,H,H,H,H,H,H,H,H,H,H,C,C,BTI,BTI;BTI,BTI;BTI,BTI;BTI,BTI, , , , , , ,C,C,C,H,H,H,H,H,H,H,H,H,H,H,H,C,BTI,BTI,BTI,BTI,C,C,C,BTI,BTI,B
```

Figure 3b.2 ***Schematic flow of DPF, Order and Disorder identification from alignment.*** *The image shows a pairwise structural alignment derived from Pali between PBDIDs- 1H8Z:B and 2ZC6:D chains. The PB sequence (present as the second row in each PDB) is scanned for PB z that signifies undefined PB. If a region bound by two z is empty both in the hit and at the corresponding positions in the second structure, then it is identified as Disorder. If the 'z___z' is corresponded by a PB sequence in the counter-alignment, it is identified as a DPF. When the regions in both the pairs of alignment have a complete pb sequence, it is treated as order. Thus, in the above example, there are 1 disorder, 3 DPF (2 same strand, 1 on second pb strand), and 4 Order states. Rest of the annotations are extracted using the demarcations provided by PBs.*

3b.3 Results and discussions

3b.3.1 Data set statistics

Of the total PDB structures in the dataset (192163), 34893 consists of DPF. This shows that 18% of PDB consist of structures that have DP regions while 39% of structures contain disordered regions. The distribution of the lengths of DPF varies immensely, although most of the DPF have a length between 1 to 7 residues, see Fig 3b.3. The length distribution is highly skewed towards shorter DPFs with an isolated peak at 3 residue length. However, there are some DPF with an enormous length of 139 residues also. Although these structures also have a very long sequences yet regions of 133 to 139 residue length that can transition between order and disorder are interesting case study. Especially when, such long regions are usually observed in intrinsically disordered proteins. It is also observed that there are hundreds of structures in our redundant data-set that contain DPF of the lengths 16 to 22.

Besides, each structure did not contain a single such region. There are cases where more than one DPF are observed in a structure. Rather there are 35 such structures that contain more than 10 DPF per structure; data represented in Figure 3b.4. However, maximum number of structures contain single DPF. Occurences of more than 10 DPF in a structure can provide insights about its function. Such structures must have multiple interacting partners which puts pressure on such regions to remain disordered natively. Thus encompassing enough flexibility to interact with multiple partners. When any interacting partner is in vicinity, the DPF can become structured. Many enzymes and Molecular recognition features (MoRFs) are known to function in such a manner [327].

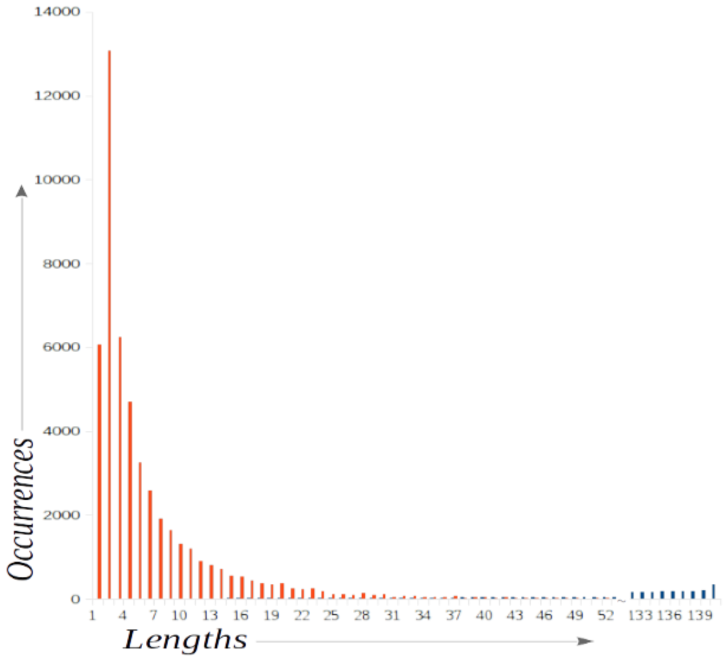


Figure 3b.3 ***Length distribution of DPF across the dataset.*** *The histogram shows the length of different DPFs identified on x-axis. The y-axis depicts the raw number of occurrences. The lengthiest DPF identified is 139 residues long while most DPFs have a length of 3 residues. The distribution is skewed towards smaller DPF lengths.*

3b.3.2 *Amino acid distributions*

Given the interdependence of sequence, structure and function, the amino acid sequence can provide indications towards a protein's structure and function. In the amino acid survey of DPF, order and disorder regions, it was observed that there is a considerable overlap among many residues which have equal or near equal propensities in either structure. However, there are certain residues which can be specifically attributed to DPF, order and disorder. Cys, Gly, Asp, and Lys shows strong propensities for DPF. According to plots in Figure 3b.5, Ala, Leu, and Gln are favourable to the ordered state of a protein. Moreover, Phe, Ser, Val, and Typ were observed to achieve high values for disorder.

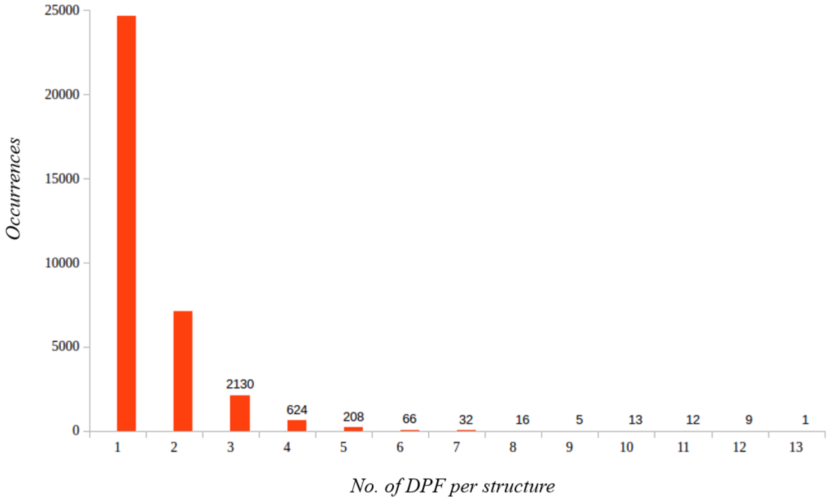


Figure 3b.4 ***Number of DPF per structure.*** *The histogram shows the number of DPFs per structure on x-axis while y-axis shows the raw occurrences. Mostly, there is a single DPF per structure but there can be 2 to 3 DPFs per structure. The most number of DPFs in a structure, in the dataset is 13. Only one structure have 13 DPFs in one pairwise alignment-* 2XVN:C *and* 4B3L:F

Figure 3b.5 shows the amino acid distributions of all the 3 states of structural continuum (order, DPF, disorder) in one plot along with earlier trends as observed by DeForte and Uversky in 2016 [315]. These trends help understand the preferred amino acid propensities of the three structural states. A region consisting of high frequency of C, G, D, and K with low R, T, V, I frequencies can be a potential DPF. The DPF favored residues are a mix of hydrophobic and hydrophilic tendencies while except for Cys, rest are flexible. These can explain the diverse nature of DPF with some regions being found in membrane proteins and others in globular proteins. More flexibility is needed for maintaining different interactions.

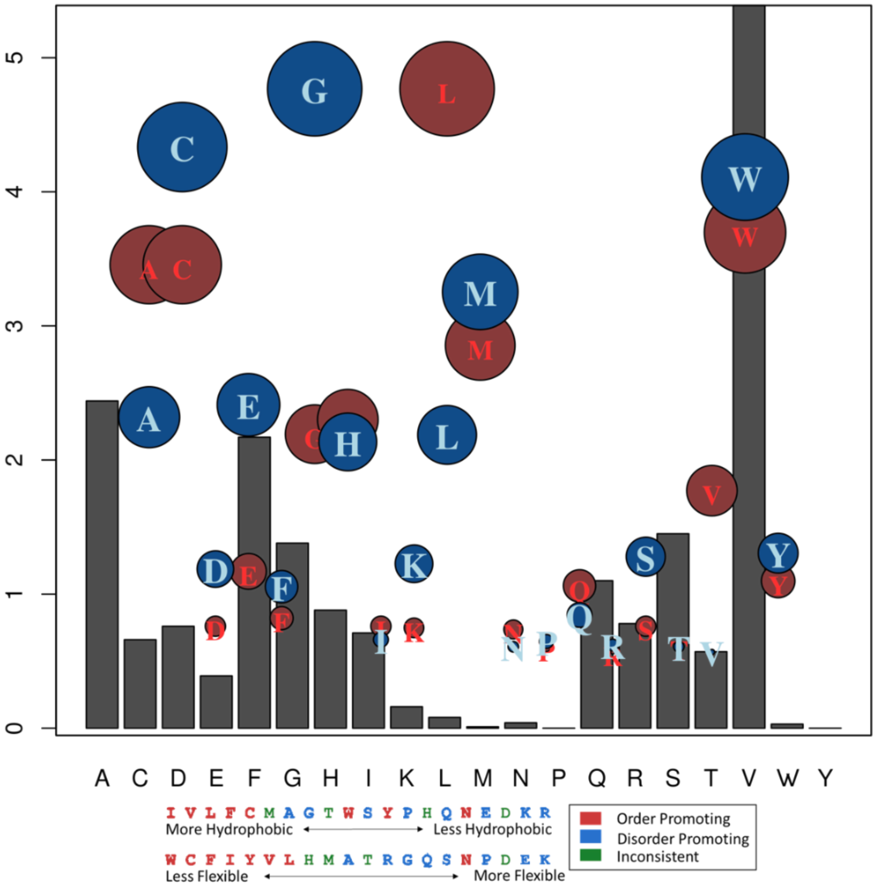


Figure 3b.5 ***Amino acid distributions in Order, DPF and Disorder.*** *The plot shows three normalized frequency values (absolute) for each amino acid on x-axis. The gray bars represent Disorder, the red circles represent Order and blue circles represent DPF. The width of the circles is proportional to their normalized frequency value. C, G, D, E, K are the most abundant amino acids observed in DPF regions. Below the plot can be seen a legend with two amino acid scales for hydrophobicity and flexibility based on Kyle and Doolittle hydrophobicity and Vihinen's flexibility scales. The color scheme in the legend depicts order or disorder promoting residues. Please note that in legend the blue color represents disorder while in plot, blue represents DPF and gray represents Disorder.*

+Image graphic in lower portion is taken from [315]

3b.3.3 Secondary structure distributions

A comparison between secondary structures present in DPF and ordered regions was carried out using DSSP, Segno and ProMotif. This provided more detailed analysis of the secondary structure space of the regions with inclusion of conformations like: β-turns, ɣ-turns, β-bulges and PPII. The results indicate specific secondary structure preference for DPF as well as ordered regions. For instance, α-helix and β-turns are observed to be highly prominent in DPF while extended β-strands are seen more often in ordered state along with π-helices and PPII helices (see Figure 3b.6). It is quite interesting to see the PPII are observed more in ordered state, albeit they are a known conformation in denatured proteins.

The propensities of amino acids (more C, G, D) and PB (in subsequent section) also helps to pinpoint the type of β-turn. According to de Brevern, 2016 [150] novel β-turn type IV_1 and type IV_3 are potentially the most occuring conformations in DPF regions (see Figure 3b.6). Although PPII are not the abundant most in DPF yet, there is a considerable occurrence of PPII helices in DPF regions. Overall, most of the secondary structures known to occur frequently in flexible regions are also observed significantly in DPF regions. This is important observation because the current structural analysis is based on the structured state of DPF and not the disordered state. Therefore, the DPF are analysed in a state of induced rigidity yet most of the secondary structures associated mostly with moibility are observed. This provides crucial insights into their structure and function relation.

3b.3.3.1 *Protein Block distributions*

Protein Blocks provide much closer approximations of protein backbone than regular secondary structures do [18]. Therefore, it is fitting to compare protein block distributions for order and DPF regions. Figure 3b.7 shows the PB distribution of DPF (blue) and order (red) residues. Both the DPF and order regions are highly populated with core of α-helical conformation as seen by abundance of *m* on both plots (Fig 3b.7). However, DPF relatively have more PBs *k* and *l* which ideally represent N-caps of an α-helix but will also qualify for β-turns since they also lie at terminus of α-helices. Both the states, DPF and order, have significant representation of β-sheets (PB *c*, and *d*) but order have more β-sheet character. A major difference is seen in the frequencies of PB *f*. DPF consists of more PB *f* conformation which vaguely resembles a β-sheet's C-cap but also resembles closely with β-bulges. Bulges are found located at the ends of antiparallel β-sheets and are important players in flexibility of the local structures. Contrary to the observations made with secondary structure distribution, β-bulges are prominently seen in DPF than in order. Thus making them a characteristic of DPF along with β-turns type IV_1 and type IV_3.

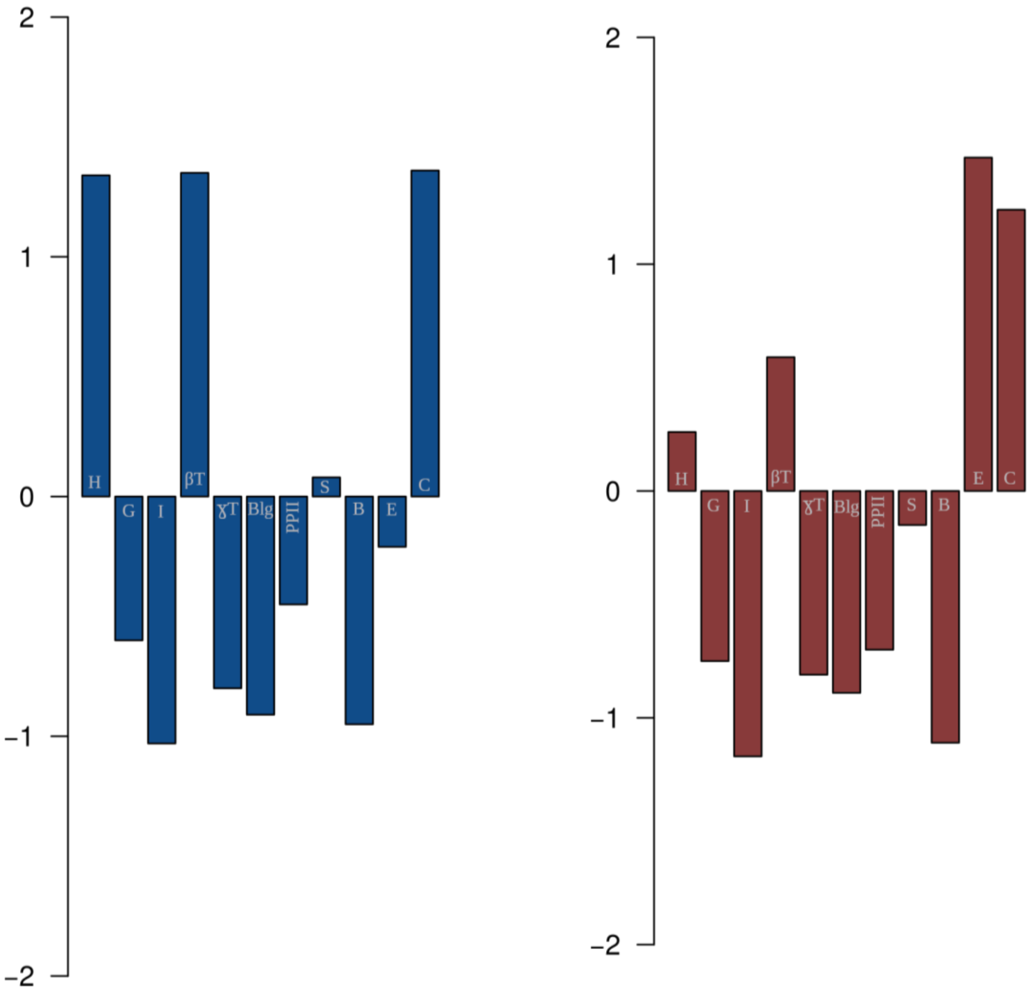


Figure 3b.6 ***Secondary structure distribution between DPF and order regions****. The x-axis represents the various secondary structure motifs like; different types of helices, β-turns, ɣ-turns, bulges, PPII, Bend, β-bridge, Extended strand and coil. These The y-axis represent the normalized frequencies of each secondary structure in DPF (blue) and Order (red). Helices and turns are collectively the most dominant secondary structures in DPF.*

Counter-intuitively, the PBs *g*, *h*, *i*, and *j (approximately loops)* are less abundant in DPF than in order regions. Perhaps, protein blocks were able to resolve turns and bulges from the assigned state of coil (C) by DSSP and others.

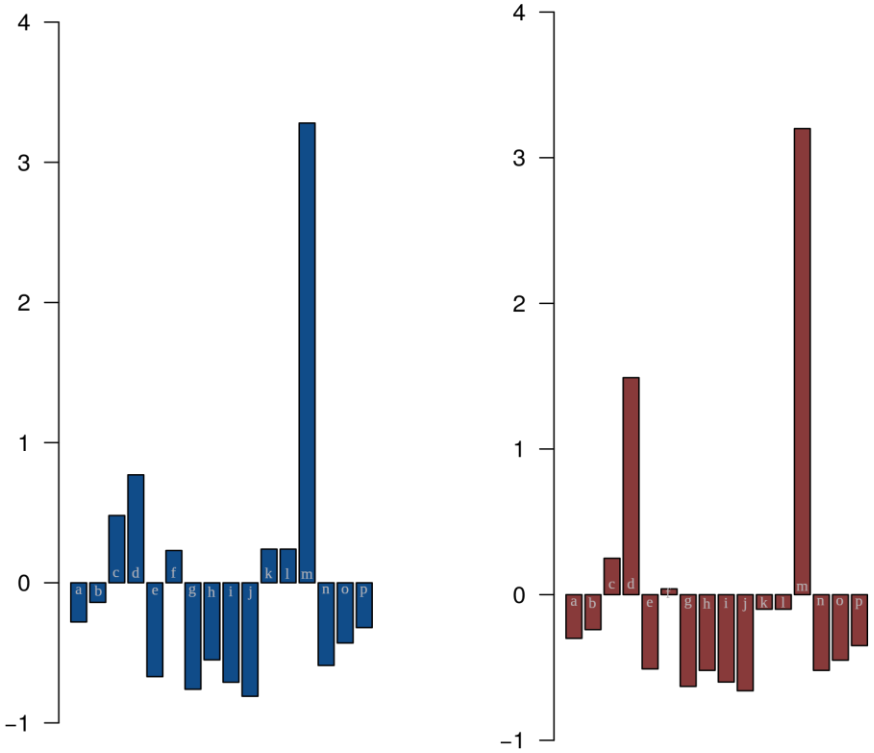


Figure 3b.7 ***Protein blocks distribution of DPF and order regions.*** *The x-axis represents the 16 PBs. The y-axis represents the normalized frequencies of each PB in DPF (blue) and Order (red). PB m is abundant in both the cases. While DPF has more representation of PB f, k, and l, Order regions have more number of sheets (PB c, d) besides having high helical content with PB m.*

3b.3.4 B-factors distributions

So far, in the preceding sections, indirect attempts have been made to understand flexibility by using secondary structures and PBs. However, a direct method for assessing flexibility is B-factors. Since DPF are captured in their structured state, B-factors are available from X-ray data. Figure 3b.8 shows a linear comparison of normalized B-factor values for both DPF and order state. Both shows a similar gumbel curve differing in their maximum and minimum values. The B-factor values for DPF suggests relative rigidity when compared with those of the ordered state. This is indeed counter-intuitive and needs more verification. However, crystal structures are known to suffer from crystal contacts that badly alters the B-factors of a protein structures. Therefore, attempts are underway to generate short molecular dynamics for random set of 200 structures from both states having equal representation. This will help in bypassing the problems, if any, due to crystal packing effects and will provide a more robust analysis on flexibility of backbone.

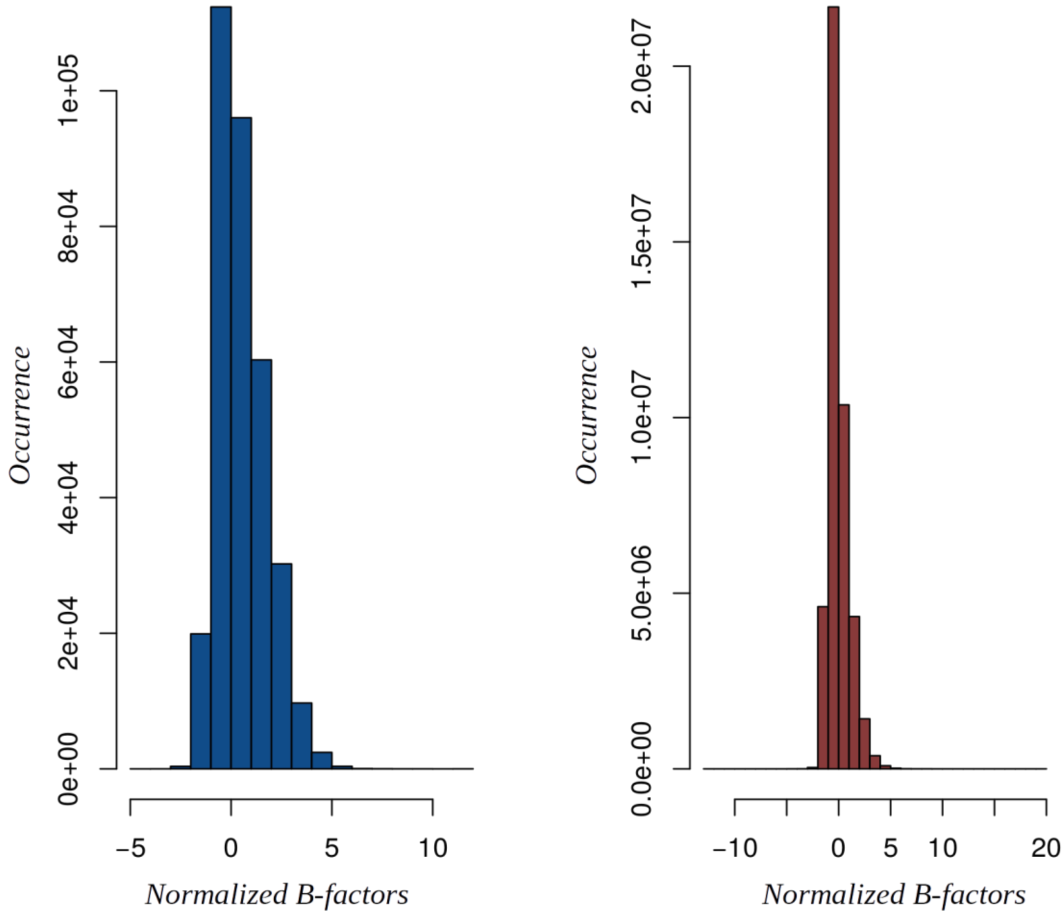


Figure 3b.8 ***Normalized B-factor distribution of DPF and Order.*** *The blue color represents the distribution of B-factors for DPF regions and red colored distribution represents B-factor values for Order. They both have a gumbel distribution with Order having slightly higher spread of data.*

3b.3.5 *Trends in Relative solvent accessibility (rSA) values*

Figure 3b.9 shows the distribution of rSA for DPF (blue) and order (red) states. The two medians from the box plots suggest that DPF have more accessibility than ordered regions. Although the median of DPF lies at ~25%, the first quartile limit (Q1) is just above 10% which is at the border of accessibility and buried areas (< 7%). The interquartile range (IQR) of both the plots are comparable with DPF having an IQR of 28 while order have an IQR of 20. The Q1 of ordered region lies at 2% suggesting some deep buried regions, Figd 3b.9. The non-outlier data in the box and whiskers of DPF (blue) have much more surface accessibility compared to the non-outlier data of order. However, the outliers to the boxplot of ordered region shows intense clustering from

~55% to 75% thus suggesting some high accessible conformations. The outliers to the DPF boxplot are lesser in number yet they have higher rSA values (>80%).

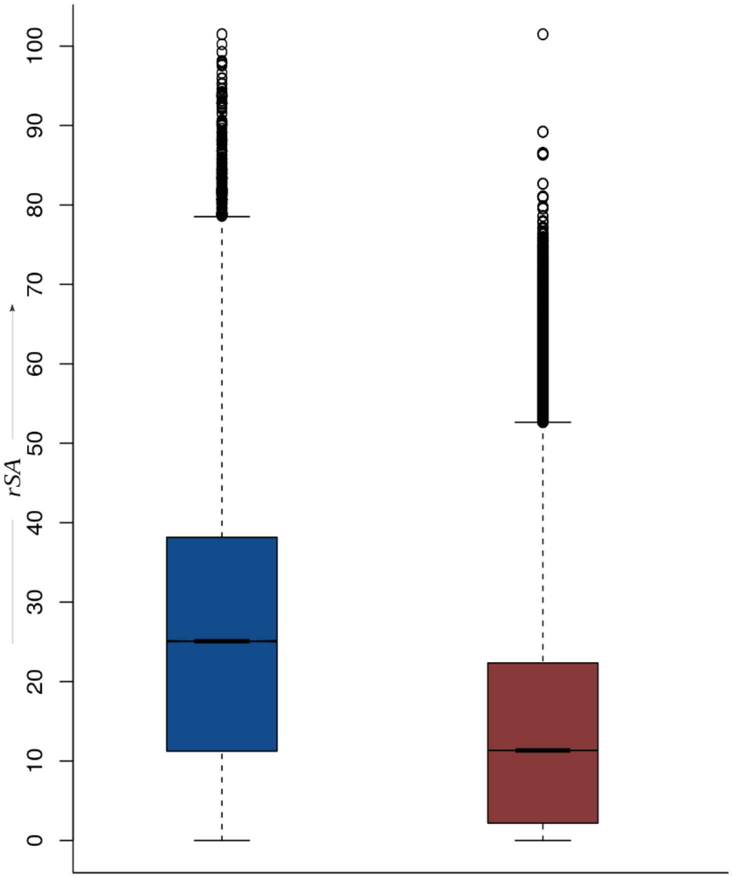


Figure 3b.9 ***Relative solvent accessibility distribution for DPF and Order.*** *The blue colored box represents the distribution of rSA for DPF regions and red colored boxplot represents rSA for Order. Overall, the range and spread of rSA for Order is smaller than DPF. It has very low Q1, around 2 that signifies deeply buried residues. While most of the distribution for DPF lies above 10% accessibility, i.e. all of the residues in DPF regions are accessible than compared to first quartile of Order having very low rSA.*

From these analyses it can be concluded that DPF have more rSA than order regions. This can be supported by the need to interact with multiple partners and solvent, since order can be triggered by environmental changes such as temperature, pH, etc. The lower Q1 and smaller range of order regions can be explained by the strong relation between rigidity and buryness of the regions in the structure.

3b.4 Conclusions and perspectives

Subsequent to the analysis of the effect of PTM on protein backbone, another unique event of protein life was analysed ─ Dual personality fragments. DPF are regions in a protein structure that can transform between disorder and order structural states. This makes them quite important as such structures support conceptualizing the structural continuum that suggests that structural states are more fluid than rigid. However, they are not very well characterized given that there has been only one systematic analysis on them, that too in 2007. The study by Zhang et al focuses on sequence characterization of DPF and that too based on identical structures alone. As DPF are essentially the disordered fragments that transit to structured state, surprisingly, their structural data remains unexploited. Therefore, it was decided to design a systematic analysis of DPF sequence and structural properties and comparison with those of order and disorder states. This could provide insights into the structure and function of DPFs and could also be suggestive of the structural properties of the otherwise denatured state.

As suggested by Zhang *et al*, that DP fragments differ from the disorder and order in their specific sequence composition. The DPF characteristic amino acid signature, as proposed by Zhang *et al.* is, 'T,R,G,N,P, and D' [313]. However, there can a caveat in the analysis as the propensities they take into consideration are solely from the corresponding data-set and not from other studies. Therefore, while analyzing amino acid distribution for DPF, order and disorder previously known propensities for order and disorder were also considered. Instead of a suggestive sequence motif, the analysis proposed characterization by frequency. Such that, if a given region has high frequency of Cys, Gly, Asp, and Lys then it can be an indicative of a DPF region. The rationale behind such an approach is two fold. A) from the analysis, Cys, Gly, Asp, Lys turn out to be specifically high for DPF regions. Also, Asp is labelled as inconsistent in being either order or disorder promoting and Gly, Asp are also part of the proposed signature by Zhang et al. B) two of the residues, Cys and Gly are rigid and moderately flexible while rest two are highly flexible. Also, Cys and Gly are hydrophobic while Asp and Lys are hydrophilic. Since, they differ in their properties, the probability of a region having high frequency of these four residues can be a reliable indicator of a DPF. Besides, the structural features of the region should also be considered. For instance, having a region with high occurrence of C, G, D, and K that has higher alpha helical and beta turn content can be a DPF. Such characterization can be used as a motivation to develop machine learning tools to predict DPF from sequences alone by using PSI-PRED or jPred.

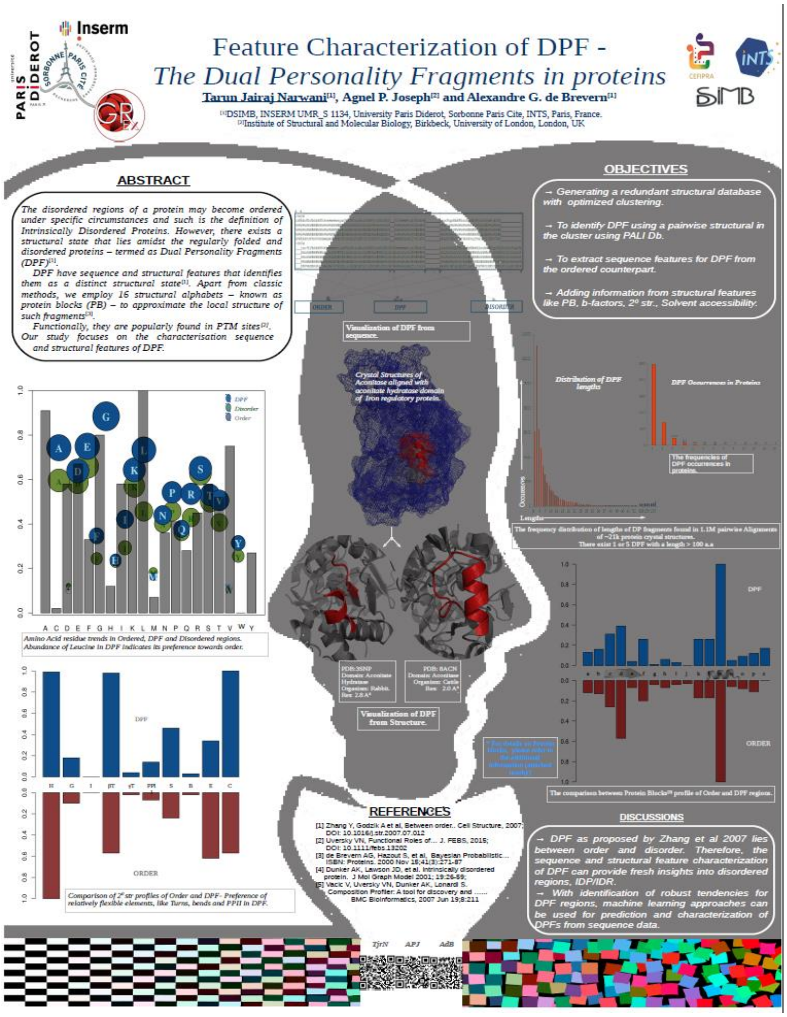
The rSA analysis shows that DPF are much more accessible than ordered regions and fittingly so. DPF have been shown to contain the site of a PTM or located near a PTM site. They have also been proposed to be active regions in Molecular recognition features (MoRFs) and enzymes [327], both of which requires interactions with multiple partners. Therefore, functionally high flexibility and more solvent accessibility is beneficial for DPF. However, there have been certain ambiguities in the B-factor analysis for flexibility. These can be attributed to the crystal contacts due to packing defects. Therefore, a logical step is to randomly select ~200 structures from the dataset and perform short MD simulations to understand the role of flexibility in DPF. Such an analysis is expected to provide much better perspective on the structural biology of DPF and may as well on the folding of protein structures.

ACKNOWLEDGEMENTS

The systematic analysis would have not been possible without the data-set prepared by Dr. Agnel Praveen Joseph and preliminary data refinement by Dr. Pierrick Craveur. The fruitful suggestions and insights into statistical methods by Dr. Alexandre de Brevern have been instrumental in the development of the study.

Dissemination of results

The resuts from chapter 3b were presented in the form of a scientific poster at IDP-2017 (Intrinsically Disordered Proteins) held at Mohali, India during December 2017. The dissemination it the form of poster helped having fruitful discussion with prominent researchers in the field like Vladmir Uversky, Peter Tompa, Rohit Pappu and Richard Kriwacki. The poster is published as: Narwani TJ, Joseph AP and de Brevern AG. Feature characterization of DPF: the dual personality fragments in proteins [version 1; not peer reviewed]. *F1000Research* 2017, **6**:2186 (poster) (doi: 10.7490/f1000research.1115178.1)



Feature Characterization of DPF - *The Dual Personality Fragments in proteins*

Tarun Jairaj Narwani[1], Agnel P. Joseph[2] and Alexandre G. de Brevern[1]

[1]DSIMB, INSERM UMR_S 1134, University Paris Diderot, Sorbonne Paris Cite, INTS, Paris, France.
[2]Institute of Structural and Molecular Biology, Birkbeck, University of London, London, UK

ABSTRACT

The disordered regions of a protein may become ordered under specific circumstances and such is the definition of Intrinsically Disordered Proteins. However, there exists a structural state that lies amidst the regularly folded and disordered proteins – termed as Dual Personality Fragments (DPF)[1].

DPF have sequence and structural features that identifies them as a distinct structural state[1]. Apart from classic methods, we employ 16 structural alphabets – known as protein blocks (PB) – to approximate the local structure of such fragments[3].

Functionally, they are popularly found in PTM sites[2]. Our study focuses on the characterisation sequence and structural features of DPF.

OBJECTIVES

→ Generating a redundant structural database with optimized clustering.

→ To identify DPF using a pairwise structural in the cluster using PALI Db.

→ To extract sequence features for DPF from the ordered counterpart.

→ Adding information from structural features like PB, b-factors, 2° str., Solvent accessibility.

Amino Acid residue trends in Ordered, DPF and Disordered regions. Abundance of Leucine in DPF indicates its preference towards order.

Comparison of 2° str profiles of Order and DPF- Preference of relatively flexible elements, like Turns, bends and PPII in DPF.

Visualization of DPF from sequence.

Crystal Structures of Aconitase aligned with aconitate hydratase domain of Iron regulatory protein.

PDB:3SNP
Domain: Aconitase Hydratase
Organism: Rabbit.
Res: 2.8 A°

PDB: 8ACN
Domain: Aconitase
Organism: Cattle
Res: 2.0 A°

Visualization of DPF from Structure.

The frequencies of DPF occurences in proteins.

The frequency distribution of lengths of DP fragments found in 1.1M pairwise Alignments of ~21k protein crystal structures.
There exist 1 or 5 DPF with a length > 100 a.a

* For details on Protein Blocks, please refer to the additional information (attached media)

The comparison between Protein Blocks[3] profile of Order and DPF regions.

REFERENCES

[1] Zhang Y, Godzik A et al, Between order.. Cell Structure, 2007; DOI: 10.1016/j.str.2007.07.012
[2] Uversky VN, Functional Roles of… J. FEBS, 2015; DOI: 10.1111/febs.13202
[3] de Brevern AG, Hazout S, et al, Bayesian Probabilistic… ISBN: Proteins. 2000 Nov 15;41(3):271-87
[4] Dunker AK, Lawson JD, et al. Intrinsically disordered protein. J Mol Graph Model 2001; 19:26-59;
[5] Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: A tool for discovery and ……. BMC Bioinformatics, 2007 Jun 19;8:211

DISCUSSIONS

→ DPF as proposed by Zhang et al 2007 lies between order and disorder. Therefore, the sequence and structural feature characterization of DPF can provide fresh insights into disordered regions, IDP/IDR.
→ With identification of robust tendencies for DPF regions, machine learning approaches can be used for prediction and characterization of DPFs from sequence data.

TjrN APJ AdB

During the writing of the thesis, the chapter have also been compiled as a manuscript, it is expected to be published by end of year 2018.

Chapter 4: Local structural dynamics in multidomain proteins- A case study of Integrin $\alpha_{IIb}\beta_3$

4.1 Introduction

While studying the impact of modifications on protein backbone, certain long range interactions were observed. Such interactions are delocalized, i.e the point of impact is structurally distant from the region of the observed effect [328]. Usually, such long range interactions occur in multidomain proteins which goes through structural transitions essential for their regulatory functions [329]. Transitioning of a structure from one structural state to another involves the changes in flexibility at a large scale. However, the driving forces during such transitions are the local structural regions. Subtle changes in these regions accumulates into large transitioning effects. Therefore, the next objective is to understand the behavior of structural flexibility in long range interactions in a multidomain protein.

One of our close collaborator Dr. Vincent Jallu from the Platelet Lab, INTS, works with Integrins $\alpha_{IIb}\beta_3$ proteins that are implicated in rare bleeding disorders like Glanzmann Thrombasthenia (GT) and Fetal Neonatal Alloimmune Thrombocytopenia (FNAIT). The Integrin protein is a multi-domain heterodimer that is expressed on the platelet cells. It undergoes structural transition from closed to open state upon activation to bind with clotting factors and aggregate. Thus it plays an important role in the clotting pathway. During, GT the defects in Integrin $\alpha_{IIb}\beta_3$ leads to the failure in transition that results in absence of clotting, thus the patient can bleed to death in event of an injury. Besides, Integrin $\alpha_{IIb}\beta_3$ are also involved in FNAIT, another thrombocytopenic defect occurring in neonatal or fetal stage. It arises due to polymorphisms of amino acids. The polymorphism can cause the expression of Human Platelet Antigen (HPA) in embryo. If the mother lacks the antigen, her placental immune system will generate antibodies against HPA that will lead to destruction of platelet cells. Due to its structural properties and pathophysiology, Integrin $\alpha_{IIb}\beta_3$ becomes the protein of interest for studying local structural dynamics in multi-domain complexes.

4.1.1 *Integrins*

Integrins are composed of a large family of heterodimeric complexes involved in cell adhesion that are expressed in different cell types. The heterodimer comprises of two large non-covalently associated, single-span type I transmembrane α and β subunits comprised of approximately 1000 and 800 residues respectively [330]. In humans, the integrins protein superfamily consists of 24 heterodimeric receptors resulting from different combinations of 18 α and 8 β subunits. Figure 4.1 depicts a schematic representation of integrins superfamily and different kinds of integrins. The extracellular domain (ectodomain) of integrins comprise of recognition sites for extracellular matrix proteins and counter receptors. The specific binding of ectodomain to such extracellular proteins and receptors lead to aggregation, cell-matrix adhesion and cell-cell adhesion, respectively [330]. While on the intracellular side, short C-terminal cytoplasmic domains link ectodomain to the cytoskeleton. Thus leading to bidirectional transmission of force through single span type I transmembrane helices. Therefore, integrins functions are crucial to embryonic development, tissue repair, host defence, homeostasis as well as haemostasis.

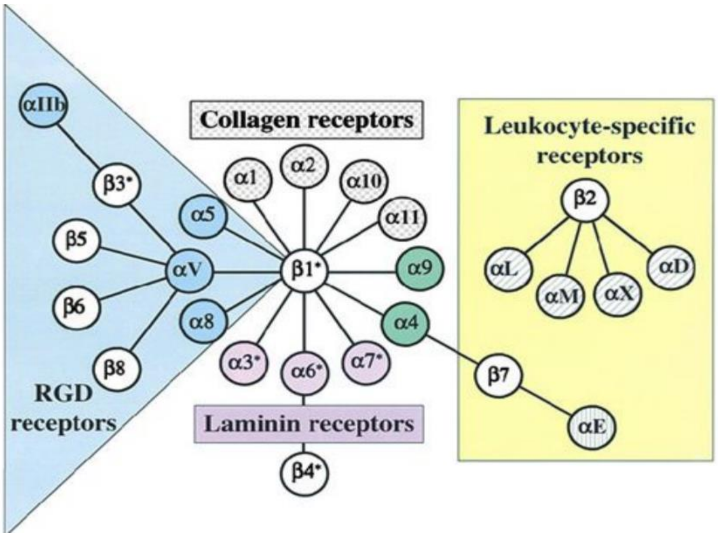


Figure 4.1. Family schematics of Integrins. *Integrins are family of proteins involved in adhesion and aggregation functions. An Integrin is identified by its heterodimeric structural assembly comprising of an alpha and a beta subunit. Overall, there are 18 α and 8 β subunits which are shown here. Based on the subunits involved, Integrins are classified into collagen receptors, Laminin receptors, Leukocyte specific receptors and RGD (Arg, Gly, Asp) receptors. Integrin $\alpha_{IIb}\beta_3$ is a member of RGD receptors. Two β subunits (β1 and β7) bind to α subunits across classes.*

$^+$Image taken from [330]

Some well-known human integrin structures are: $\alpha_x\beta_2$, $\alpha_v\beta_3$ (cell-matrix adhesion) and $\alpha_{IIb}\beta_3$ (cell-cell adhesion). α_v can associate to $\beta_{1,3,5,6,8}$ subunits while α_{IIb}, that is specific to platelets and megakaryocytes cells, only associates to β_3 [330].

4.1.1.1 *Inside-out signaling*

Besides their mechanical functions, integrins ectodomain can undergo conformational changes in response to the molecular or chemical signals that interact with the cytoplasmic tails. The induced conformational changes in the ectodomains lead to selective affinity for extracellular ligands [331]. For instance, the conformational changes that occur in the ectodomain of Integrin $\alpha_{IIb}\beta_3$ resulting in enhanced affinity for clotting factors like, Fibrinogen and Von Willebrand factor (vWF).

4.1.1.2 *Outside-in signaling*

Integrins also transmit chemical signals into the cell providing information about their adhesive state, vascular location, local environment, etc. These endocrine, paracrine, and autocrine signals further illicit other membrane proteins like G-proteins coupled receptors, especially chemokine receptors that can elicit numerous variable responses. Such crosstalk results in cellular migration, differentiation, survival, and motility. This type of signaling wherein chemicals released by an Integrin can cause changes in distant cells via different membrane receptors or transporters is known as outside-in signaling [332].

4.1.2 *Integrin $\alpha_{IIb}\beta_3$*

The integrin $\alpha_{IIb}\beta_3$ is a fibrinogen receptor expressed at the platelet surface. It also binds to vWF in case of severe injuries. As evident from the nomenclature, it consists of an α_{IIb} subunit non-covalently bound with β_3 subunit. It is responsible for platelets aggregation, a key process in primary haemostasis and thrombus formation [333]. Integrin $\alpha_{IIb}\beta_3$ has been shown to get activated by both outside-in and inside-out signaling [334]. Studies have shown that elevated levels of cytoplasmic Ca^{+2} leads to the binding of Talin protein on the cytoplasmic domain of β_3 subunit. This binding causes allosteric changes in the ectodomain that transitions from closed to open state conformation thus making it activated to bind with fibrinogen or vWF. Multiple integrin $\alpha_{IIb}\beta_3$

aggregates upon a fibrinogen leading to thrombus formation. Figure 4.2 shows a diagram of the inside-out signaling of integrin $\alpha_{IIb}\beta_3$ [334].

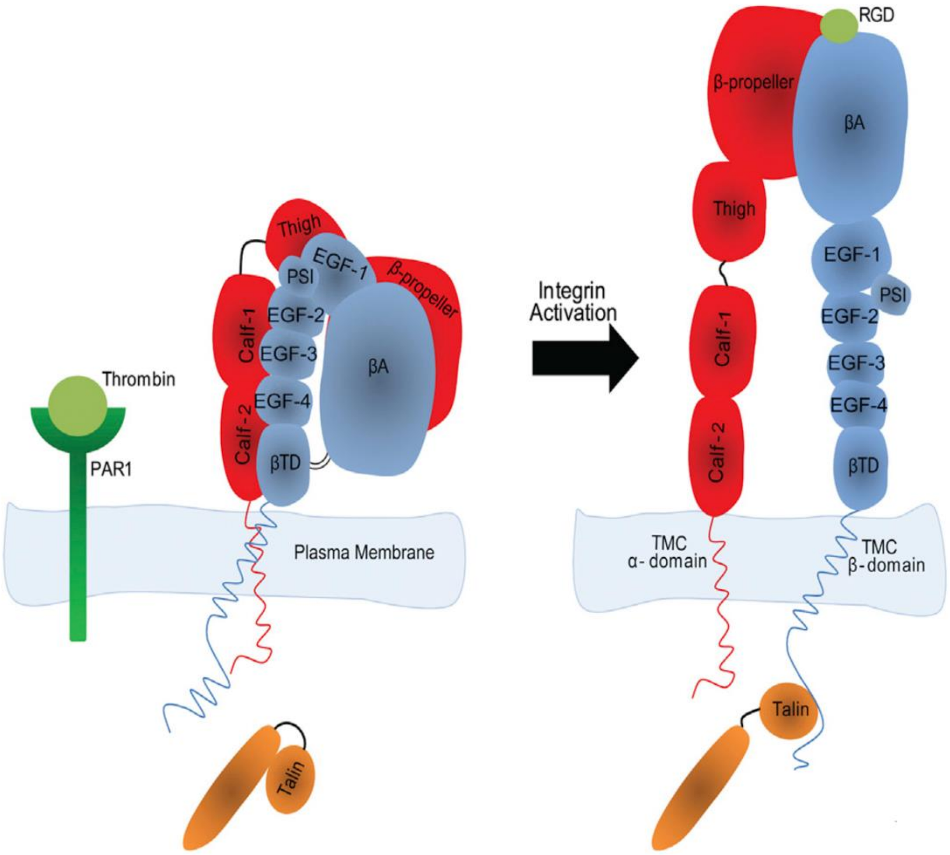


Figure 4.2. Inside-out signalling in Integrins. *The image depicts the inside out signalling in Integrin αIIbβ3. First Talin is activated due to binding of Thrombin at PAR1 receptor. Talin then binds to cytoplasmic tails of β3 subunit that causes allosteric changes in the propeller domain that contains RGD. These events lead to opening of the structure.*

$^+$Image taken from [334]

On the other hand, the concentration of divalent ions in the extracellular matrix triggers the activation of integrin $\alpha_{IIb}\beta_3$. The headpiece domains of integrin $\alpha_{IIb}\beta_3$ consists of metal ion dependent adhesion sites (MIDAS), adjacent to MIDAS (ADMIDAS) and ligand induced metal binding site (LIMBS) that are coordinated by Ca^{+2} ions. Therefore, increased concentration of Ca^{+2} ions has shown to reinforce the bent (closed) form of the ectodomain thus keeping it in the inactive state while increased concentration of Mn^{+2} favours integrin $\alpha_{IIb}\beta_3$ activation and leads to the extended (open) conformation. Thus integrin $\alpha_{IIb}\beta_3$ have been shown to be activated by both types of signalling [334].

4.1.3 *Multi-domain structure of Integrin $\alpha_{IIb}\beta_3$*

The $\alpha_{IIb}\beta_3$ structure is organized into 3 distinct regions; an N-terminus extracellular ectodomain, a single spanning transmembrane (TM) region and a C-terminus cytoplasmic region. The cytoplasmic region in α subunit is very swift (~20 residues) while in β subunit it extended up to 46 residues in length and constitutes an important node for signalling. Talin binds at the cytoplasmic domain of β subunit. Both α and β subunit TM regions are single spanning, and consist of 22 residues each. The ectodomain is relatively huge with 959 and 693 residues in α and β subunits respectively. Figure 4.3 depicts transition steps from the inactive conformations of $\alpha_{IIb}\beta_3$ (crystallized closed structure, PBDid: 3FCS) to its theoretical open liganded active form. A complete structure of the open forms of the ectodomain with or without ligand remains to be crystallized. The ectodomain is further divided into four regions: headpiece, knee, legs and tails.

Headpiece: The headpiece that carries the ligand-binding site consists of the β-propeller domain of α_{IIb} subunit and the β-I domain of β_3 subunit. The $\alpha_{IIb}\beta_3$-propeller domain consists of a 7 bladed fold with four Ca^{2+} ions coordinated with β-hairpin loops connecting the antiparallel β-strands (see Fig 4.3A). The β-I domain of β_3 mainly consists of α helices and loops with coordinated metal ions Ca^{2+} and Mg^{2+} constituting a MIDAS (Metal Ion Dependent Adhesion Sites) with an ADMIDAS (Adjacent to MIDAS) and SyMBS (Synergistic Metal Binding Site). These sites play critical role in opening the $\alpha_{IIb}\beta_3$ binding site and helps in ligand binding [335].

α_{IIb} Leg and the Knee: Downstream the β-propeller is the α_{IIb} leg, composed of the Thigh domain, the Genu (knee), the rigid Calf-1 and Calf-2 domains. The short loop of α_{IIb} Genu coordinates with a divalent calcium [336]. This metal ion might help in stabilizing Calf-1 and Thigh domain interface during the opening of the structure following the activation process (angular shift at Genu). The α_{IIb} leg is rigid and provides a framework to the entire ectodomain.

B_3 Leg and Knee: The β_3 β-I domain of the headpiece is succeeded by the Hybrid, PSI, and 4 IEGFs (Integrin Epidermal Growth Factor) domains. A short knee joins the IEGF-1 with IEGF-2 domains. The β_3 leg consists of IEGF-2 to IEGF-4 whose C-terminus ends in the ankle domain (tail). α_{IIb} and β_3 transmembrane and cytoplasmic domains are not shown.

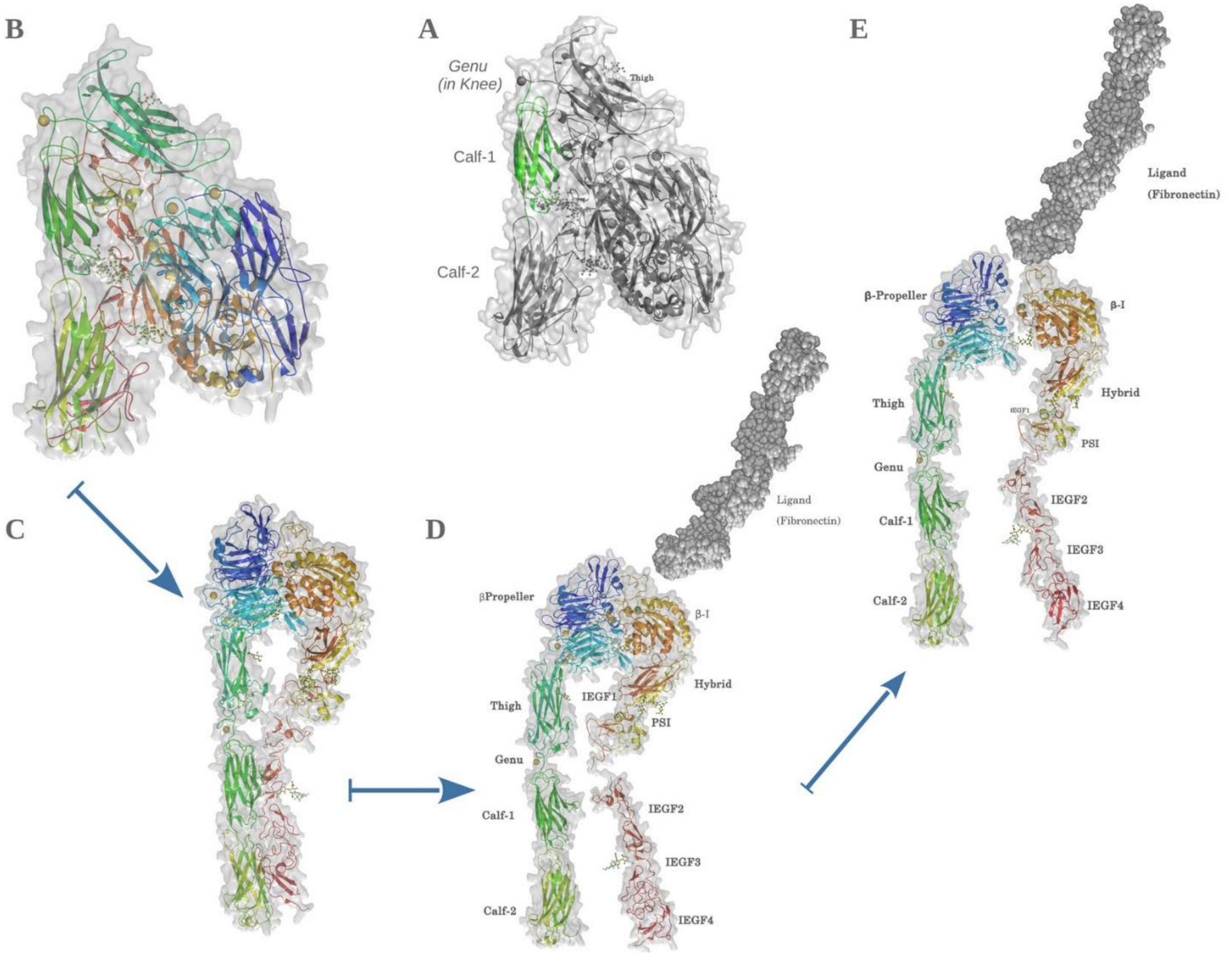


Figure 4.3 ***Closed to open transition of αIIbβ3.*** *(A) The closed form of $\alpha_{IIb}\beta_3$ ectodomain, with Calf-1 domain highlighted in green. Rest of the structure is depicted in dull grey to bring clarity. Structural organization of ectodomain is labelled. The rainbow schema of colours on secondary structures represents the $\alpha_{IIb}\beta_3$ structure. Regions in green-blue spectra mark the α_{IIb} subunit and yellow-red spectra mark the β_3 subunit. (B) Closed inactive form of $\alpha_{IIb}\beta_3$. The structure is bent along the plane of knee domains. (C) Extended α_{IIb} headpiece with β_3 leg resting alongside the α_{IIb} leg. (D) Extended β_3 conformation: The β_3 headpiece has intrinsic conformational changes at C-terminus leading to an outward pull of the β_3 leg. (E) Extended $\alpha_{IIb}\beta_3$ conformation: In the last stage, β_3 headpiece pulls out creating a ligand-binding cavity between the two headpieces. Mg^{2+} constituting MIDAS can be seen as a green sphere in the cavity, while the ligand Fibrinogen (dull*

grey dots) approaches the glycoprotein. All metal ions are shown as solid spheres with golden ones representing Ca^{2+} while green representing Mg^{2+}. Polysaccharides (N-acetyl glucosamines and Mannose) are shown in ball and stick representations. Please notice that the open forms had been modelled from the closed structure according to expected conformations using Modeller_v_9.16 [70] and images are generated by PyMol_v_1.7.0 [352].

4.1.3.1 *Structural transition from bent to extended state*

The activation state of $\alpha_{IIb}\beta_3$ controlled by inside - out signalling results from platelet activation by multiple exogenous factors (physiological plasmatic agonists, exposed sub-endothelial matrix) leading to the binding of Talin at β_3 cytoplasmic tail. The α_{IIb} headpiece opens up creating an angular shift between Thigh and Calf-1 domains (Fig 4.3C) meanwhile, the β_3 leg and tail remains parallel to the α_{IIb} leg. Thereafter, opening of the β_3 headpiece pulls the β_3 legs outward resulting into an extended open conformation (Figs 4.3D and 4.3E) that can bind plasmatic fibrinogen at the MIDAS, which is constituted by elements of both the headpieces [337]. Finally, fibrinogen binding leads to outside-in signaling and directing the platelet cells into close proximity of other platelets. Multiple platelets expressing $\alpha_{IIb}\beta_3$ binds to fibrinogen thus forming a thrombus leading to clot formation.

4.1.4 *Defective expression of integrin $\alpha_{IIb}\beta_3$*

Upon activation the integrin $\alpha_{IIb}\beta_3$ binds plasmatic fibrinogen leading to platelet aggregation. However, a defect in the expression of $\alpha_{IIb}\beta_3$ or failure to open up or specific mutations can have disruptive results. Defective platelet aggregation leads to two severe life-threatening bleeding disorders: Glanzmann thrombasthenia (GT) and Fetal / neonatal alloimmune thrombocytopenia (FNAIT). GT is a rare autosomal recessive genetic disease associated with defective expression and / or function of $\alpha_{IIb}\beta_3$ [338] while FNAIT results from fetal / neonatal platelet destruction by maternal alloantibodies in mothers lacking the fetal platelet alloantigens inherited from the father. Clinical consequences of FNAIT range from no symptoms to intracranial hemorrhages with a risk of neurological sequel and/or fetal/neonatal death [339]. Both diseases result from α_{IIb} and β_3 gene polymorphisms.

In GT, more than 300 mutations have been identified in α_{IIb} or β_3 genes. Most of them are reported in GT specific database: https://sinaicentral.mssm.edu/intranet/research/glanzmann. These mutations have distinguished effects on the $\alpha_{IIb}\beta_3$ phenotype. Many missense mutations cause defective expression of $\alpha_{IIb}\beta_3$ on the platelet cells. While certain silent mutations do not affect the phenotype instead can change the allosteric propagation of the transition sequence leading to lack of affinity for fibrinogen.

However, in FNAIT, neither the expression nor function of $\alpha_{IIb}\beta_3$ is affected but single nucelotide polymorphisms (SNP) resulting in amino acid (aa) variations lead to sequence that defines Human Platelet Antigens (HPA). The effects of these amino acid substitutions on $\alpha_{IIb}\beta_3$ structure remain largely unknown. Most of the human platelet alloantigens are described in the HPA database http://www.ebi.ac.uk/ipd/hpa.

4.1.5 *Domains of interest*

Given the enormous size of the ectodomain of integrin $\alpha_{IIb}\beta_3$ and high number of mutations for GT and FNAIT pathologies, it seems logical to study one or two domains at a time. Previously, the collaboration with Dr. Jallu on FNAIT and GT have been fruitful. Using closed state ectodomain crystal structure from PBDID: 3FCS, it was shown that the β_3 Lys253Met GT mutation impaired key ionic interactions between the α_{IIb} β-propeller and the β_3 β-I like domain [340]. Nonetheless, static models cannot depict all mutation-induced effects on a highly dynamic structure like integrins. Therefore, molecular dynamics (MD) simulations were used to study L33P substitution located in the PSI domain of β_3 subunit. The L33P substitution is responsible for the HPA-1 system, clinically the most important one in Caucasian populations [255]. The dynamics of structures of PSI domain as L33 and P33 variants was compared to find that the mutation does have an impact on the conformations [255,341]. Later, a third variant with a Valine at position 33 (L33V) was also studied along the L33P mutation [341]. Although the 3 variants mostly shared common conformations, the P33 variant showed a higher mobility and specific conformations of IEGF-1, IEGF-2, and PSI domains. As shown in Figure 4.4, the L33V substitution mainly displaced a dynamic equilibrium between common structures that could explain a variable reactivity of different anti-HPA-1a sera with the two β_3 forms [341].

As discussed in section 4.1.3 that the leg region of the α_{IIb} subunit plays anchoring role during the structural transition and is known to consist the most rigid domains of the $\alpha_{IIb}\beta_3$ structure. However, Nussinov lab have discussed multiple times that rigid domains can have underlying allostery [342,343]. Therefore, the apparent rigidity of the leg domains, Calf-1 and Calf-2 became our domains of interest.

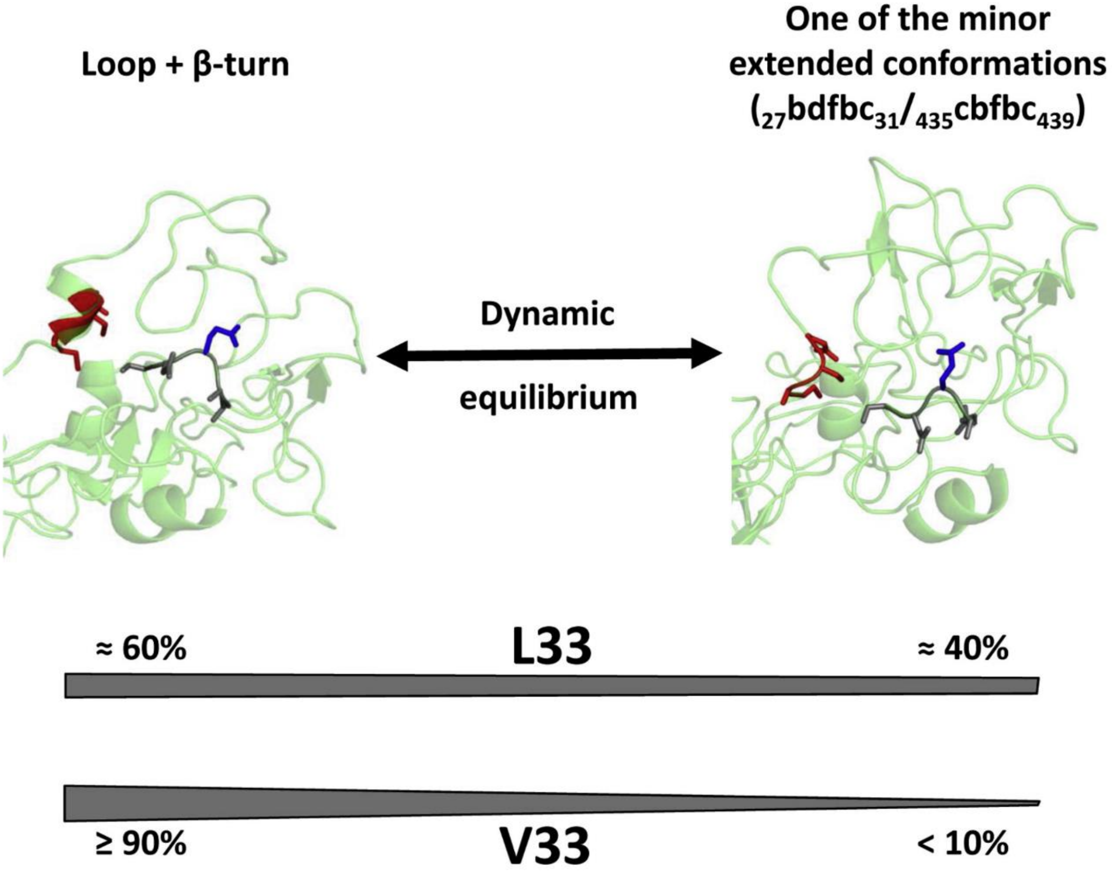

Loop + β-turn
One of the minor extended conformations
($_{27}$bdfbc$_{31}$/$_{435}$cbfbc$_{439}$)
Dynamic
equilibrium
≈ 60%
L33
≈ 40%
≥ 90%
V33
< 10%

Figure 4.4. Most frequent structures in L33-β3 and V33-β3 of residues 27 to 31 and 435 to 438. *Secondary structures (light green) of the β3 knee of L33 and V33 variants are shown. Worm lines correspond to loops or extended conformations, arrows to β sheets, and ribbons to α-helices or β-turns. Left panel, the dominant structure formed with residues 27-31 (grey) and 435-439 (red) adopting PB sequences 27bfkbc31 (loop) and 435fklmm439 (β-turn). The side chain of residue E29 is colored in blue. Right panel, one of the minor competing structures adopting extended conformations (shown here PB sequences: 27bdfbc31 and 435cbfbc439) for L33-β3 and in a lesser extent for V33-β3. The grey double arrows visualize for L33-β3 and V33-β3 the balance (and frequencies) existing between the dominant structure and all the minor structures whose only one is shown here-; (the thickness of the bar is proportional to the frequency).*

+Image taken from [341]

4.1.5.1 *Calf-1 domain*

The first domain in the leg region of α_{IIb} subunit extends from residues 603 to 743 (numbered as in PDBid 3FCS). It consists of 9 consecutive β-strands connected by 8 loops (Fig 4.5A). Loops 1 and 10, located at the N- and C-terminals of Calf-1 connects it with N-ter Thigh and C-ter Calf-2 domains, respectively.

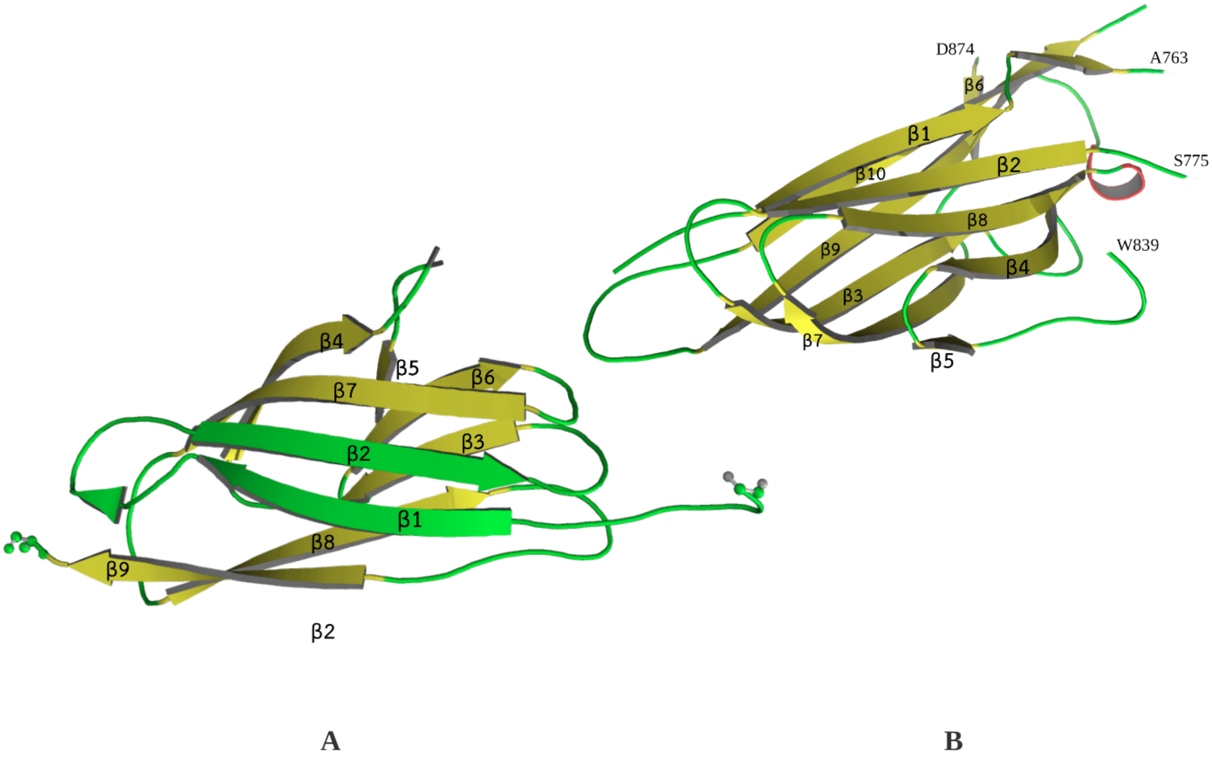


Figure 4.5 ***Calf-1 and Calf-2 domains demarcation and structural organization.*** *Shown are the isolated individual domains A) Calf-1 and B) Calf-2. Both the domains have a beta sandwich fold with anti-parallel running beta strands connected through loops. Calf-1 has a small missing region between* β4 and β5, *colored in gray. Calf-2 has two big regions of missing atomic coordinates, as marked by* β1 - β2 (11 residues) and β5 – β6 (34 residues).

4.1.5.2 *Calf-2 domain*

The last domain in the α_{IIb} ectodomain extending from residues 744 to 959 (numbered as in PDBid 3FCS). Calf-2 is made of 10 consecutive β-strand connected by 11 loops as shown in Figure 4.5B. The C-ter of Calf-2 is binded with the single spanning TM-helix of α_{IIb} subunit.

Therefore, the prime objective will be to understand the role of inherent flexibility in the domains of Calf-1 and Calf-2. Additionally, to study the structural changes induced by GT and FNAIT mutations specific to these domains.

4.2 Methods

4.2.1 *Structural data*

The α_{IIb} Calf-1 and Calf-2 domains were extracted from a 2.55 Å resolution crystal structure of the $\alpha_{IIb}\beta_3$ integrin (PBDID 3FCS) [344]. Calf-1 is a domain of 141 residues [positions 603–743] while Calf-2 spans over 216 residues from position 744 to 959. Both have a mainly beta (2) sandwich (2.60) protein with an immunoglobulin-like (2.60.40) topology as described in CATH database (no: 2.60.40.1510 and 2.60.40.1530 in http://www.cathdb.info/version/latest/domain/3fcsA03).

Some missing atoms in side chains of residues 667 and 668 of Calf-1 were completed using Modeller software v.9.14 [70]. However, Calf-2 had important missing regions which posed a challenge. Two regions of length 11 residues (position: 763-775) and 34 residues (position: 840-873) were missing and therefore were difficult to model using classical modeller protocol.

4.2.1.1 *Modeling the missing regions in Calf-2*

Homologs for $\alpha_{IIb}\beta_3$ integrin (PBDID 3FCS) were searched in PDB database using blastp v2.6.0 [401] that returned 8 structures. Sorting based on low e-value and high query coverage reveals two proteins having the highest percent identity of 38%. PBDids 4G1E and 3IJE were selected from primary results. Although both the $\alpha_V\beta_3$ integrin structures have the missing regions yet 4G1E is selected since it has a missing region of 17 residues compared to 34 residues in 3FCS and 28 residues in 3IJE. Therefore, selecting 4G1E will at least make the gap covered by half the length. Moreover, the Calf-2 domain of 4G1E is structurally closer to that of 3FCS with an RMSD of 0.72 Å and TM-score of 0.95.

Apart from sequence based homology search, structurally similar proteins were also sought after using FATCAT (Flexible structure AlignmenT by Chaining Aligned Frames Fragment Allow Twists) [345]. The top hit being a leukocyte specific receptor, $\alpha_X\beta_2$ integrin (PBDID 4NEH) having 24% identity. However, the Calf-2 domain in 4NEH is complete without any gaps, having a total length of 190 residues. Although the $\alpha_X\beta_2$ integrin is a distant homolog yet using two more related templates for modelling $\alpha_{IIb}\beta_3$ Calf-2 domain can be useful.

4.2.2 *Selected mutations and their structural variants*

Calf-1 variants- The Seven Calf-1 domain variants studied herein were involved in GT and are reported in the GT database, https://sinaicentral.mssm.edu/intranet/research/glanzmann. Variants L653R[345,346], L721R (only reported in the GT database), L721V[347], R724P[347,348], R724Q[349], and P741R[347,348] severely impaired $\alpha_{IIb}\beta_3$ expression (less than 5% expressed) while variant C674R[347,348][350] allowed a 10% residual expression (type I and II GT [351]).

Calf-2 variants- Five mutations were selected in Calf-2 domain. Two of these have implications in GT; H798P, S926L [340] while the rest three are polymorphisms that lead to HPA system implicated in FNAIT. The polymorphisms involved in FNAIT are: V837M, L841M, I843S. It is noted that four of the five mutations lie in the 34 residue long missing region in Calf-2 (840 - 873). These are suggestive of the underlying importance of the region. The selection of these variants in Calf-2 is suggested by the collaborators from platelet lab in INTS.

The seven GT aa substitutions were introduced in the structures by *in-silico* mutagenesis using PyMOL software [352] and the SCWRL method [353]. The effects of all mutations were studied exclusively.

4.2.3 *Molecular Dynamics*

MD simulations were done using GROMACS 5.1.1 software [354] with Gromos96 54a7 force-field [355]. WT and variant forms were soaked in a rhombic dodecahedral simulation box with TIP3P water molecules and neutralized with Cl- ions. The MD protocol is similar to the ones used in our previous works [255,341]. After 1 nsec of equilibration (with position restraints on the protein), each system was simulated through 10 independent dynamics for a total of 1 microsecond (10 × 100 nsec). Molecular conformations were saved every 100 psec for downstream analysis. The first 5 nsec of each MD simulation were discarded considering the noise generated by residues at the extremities.

Trajectory analyses were done with the GROMACS software, in-house Python and R scripts. Root mean square deviations (RMSD) and root mean square fluctuations (RMSF) were calculated on Cα atoms only. Residues interactions were analysed using the online tool PIC (Protein Interactions Calculator) [356].

Two important computational resources were used for running MD simulations. Our in-house super cluster, Serenity, having 48 compute nodes with 16 cores per node, thus generating a computational power of 768 cores. Also, CINES national supercomputer OCCIGEN was used under allocation no. A0010707621.

4.2.3.1 *Trajectory analysis using Protein Blocks*

Protein Blocks (PBs) are very efficient in tasks such as protein superimpositions [61]and MD analyses [241]. They are labelled from a to *p*: PBs *m* and *d* can be roughly described as prototypes for core of α-helix and central β-strand, respectively. PBs a to *c* primarily represent β-strand N-ter and PBs *e* and *f* representing β-strand C-ter; PBs a to *j* are specific to coils; PBs *k* and *l* to α-helix N-ter while PBs *n* to p to α-helix C-ter. PB assignment was carried out using PBxplore tool developed by our team and freely available at GitHub (https://github.com/pierrepo/PBxplore) [69]. PB were assigned for each residue of the domains and over every snapshot extracted from MD simulations. The equivalent number of PBs (*Neq*) is a statistical measurement similar to entropy that represents the average number of PBs for a residue at a given position. For details on *Neq*, please refer Introduction 1.6.3.

To underline the main differences between the wild-type (WT) and a variant for each position, Δ*Neq* value is computed. Δ*Neq* is the absolute difference between corresponding Neq values. However, a same Δ*Neq* value can be obtained with different types of blocks in similar proportions. Therefore, to detect a significant change in PBs profile, a ΔPB value was calculated. It corresponds to the absolute sum of the differences for each PB between the probabilities of a PB *x* to be present in the WT and the variant forms (*x* goes from PB a to PB *p*). ΔPB is calculated as follows:

$$\Delta PB \ = \sum_{x-1}^{16} \quad | f_x^{WT} - \ f_x^{var}|$$

where, f_x^{WT} and f_x^{var} are the percentages of occurrence of a PB x in respectively the WT and the variant forms of Calf-1 structures. A value of 0 indicates perfect PBs identity between WT and variant, while a score of 2 indicates a total difference.

4.3 Results and Discussions

The rigid, anchor region of the α_{IIb} subunit leg that comprises of two domains of Calf-1 and Calf-2 is under investigation for inherent flexibility. A core β sandwich fold consisting of 8 to 9 antiparallel β strands connected with loops. Usually, a β sandwich fold is found in anchoring roles in the structures, for eg, in heavy chains (V_H) of antibodies [23], Flaf protein in Archeal cell envelope [42] and therefore has more rigidity associated to it. Moreover, the inside-out as well as outside-in signaling primarily interacts with the β_3 subunit. Therefore, it will be interesting to understand the role of inherent flexibility in the dynamics of these domains. Also, the dynamics of wild type Calf domains will be compared with that of different variants (structural) implicated in GT and FNAIT.

4.3.1 *Completing the missing regions in Calf-2 domain*

With huge gaps of 11 and 34 residues, it will not be possible to understand the dynamics of Calf-2. The selected templates, 4G1E and 4NEH were used exclusively to model the missing regions. While the overall scaffold of the Calf-2 domain is provided by the $\alpha_{IIb}\beta_3$ self-template structure of 3FCS. Thus three template structures are used having PBDids; 3FCS ($\alpha_{IIb}\beta_3$), 4G1E ($\alpha_V\beta_3$), and 4NEH ($\alpha_X\beta_2$). However, the generated model with best DOPE score did not have convincing conformation for the 34 residue missing region. It modelled it as a highly disordered loop (*based on the loop conformation in all the 100 models*) which exhibited self-interactions. Given that the loop consists of FNAIT variants that will lead to expression of HPA and that the loop might interact with IEGF domains of β_3 subunit, the loop confirmation is unacceptable. Therefore, based on the shorter yet complete loop structure of the leukocyte specific integrin, $\alpha_X\beta_2$ (4NEH) and the very small structural distance between G840 and Q873, structural constraints are designed for the missing region. Each 5th residue in the loop should have a distance of 10Å while two 20Å distance restraints are put between 840th to 850th residue and 863rd to 873rd residue. The principle schematic of the restraints is shown in Figure 4.6A. The principle of such structural restraints is to avoid the self interaction of the loop and an expanded conformation given its interactions with IEGF domains.

Thus the final selected model has two loops and both exhibit small helical component at the most distal part of the loops, Fig 4.6B. Short molecular dynamics of 50 nsec confirmed the stability of the loop, although the loop 7 is highly more deformable.

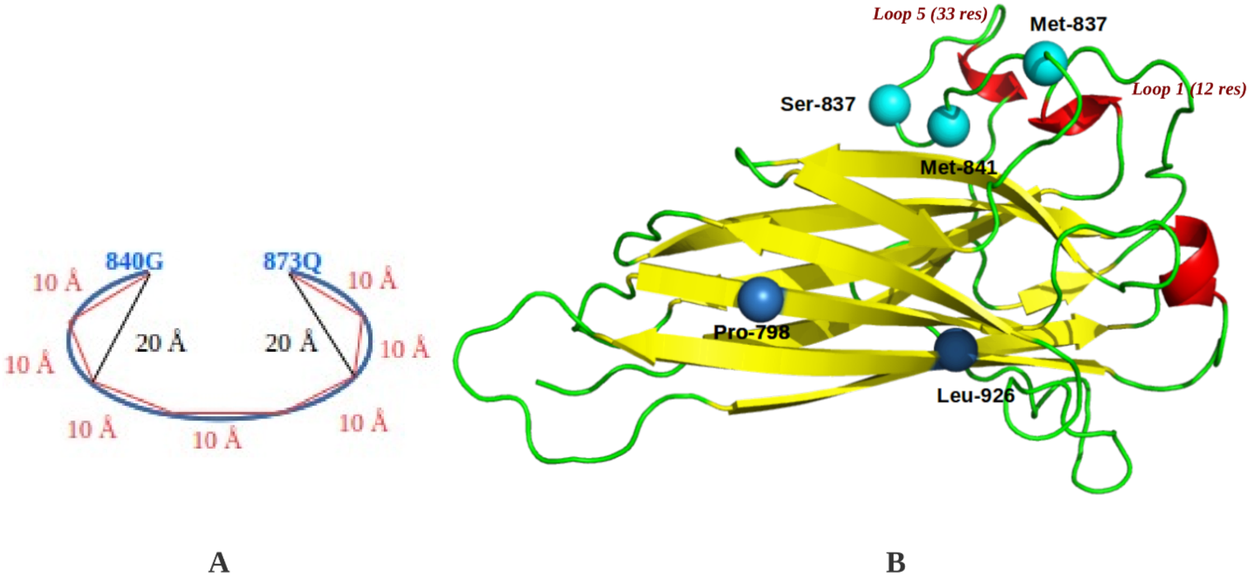


Figure 4.6 ***Completing the missing regions in Calf-2.*** *A) schematic of restraints applied. B) the modelled loops based on the structural restraints.*

4.3.2 <u>*Structural analysis of the Calf-1 domain*</u>

Calf-1 domain extends from residues 603 to 743 of the α_{IIb} integrin subunit. This domain is an all β structure adapting an Immunoglobulin-like β-sandwich fold with 9 consecutive β-strands connected by 8 loops (the loops position can be seen in Figure 4.7D) [357]. Loops 1 and 10, located at the N- and C-terminals of Calf-1 connects it with N-ter Thigh and C-ter Calf-2 domains, respectively. RMSD from all MD simulations reach a steady state at 2.5 nsec (Fig 4.8) that is maintained in longer runs of 100 nsec indicating stable and reproducible independent dynamics. According to the high B-factor values obtained from crystallographic data, loops 2, 3, 4 and 5 are the most flexible regions of Calf-1 (Fig 4.7A). Residues 622, 643, 710 and residues 667/668 (of loop 5 that contains missing atoms) presented the highest B-factor values in their respective loops. On average β-strands are more rigid than loops [34,35] although some of their residues represent relatively high B-factor values in Calf 1. As it is known that B-factors are strongly influenced by the crystal packing of the structure [40] therefore, it was checked and B-factors are confirmed to be not influenced by crystal packing contacts.

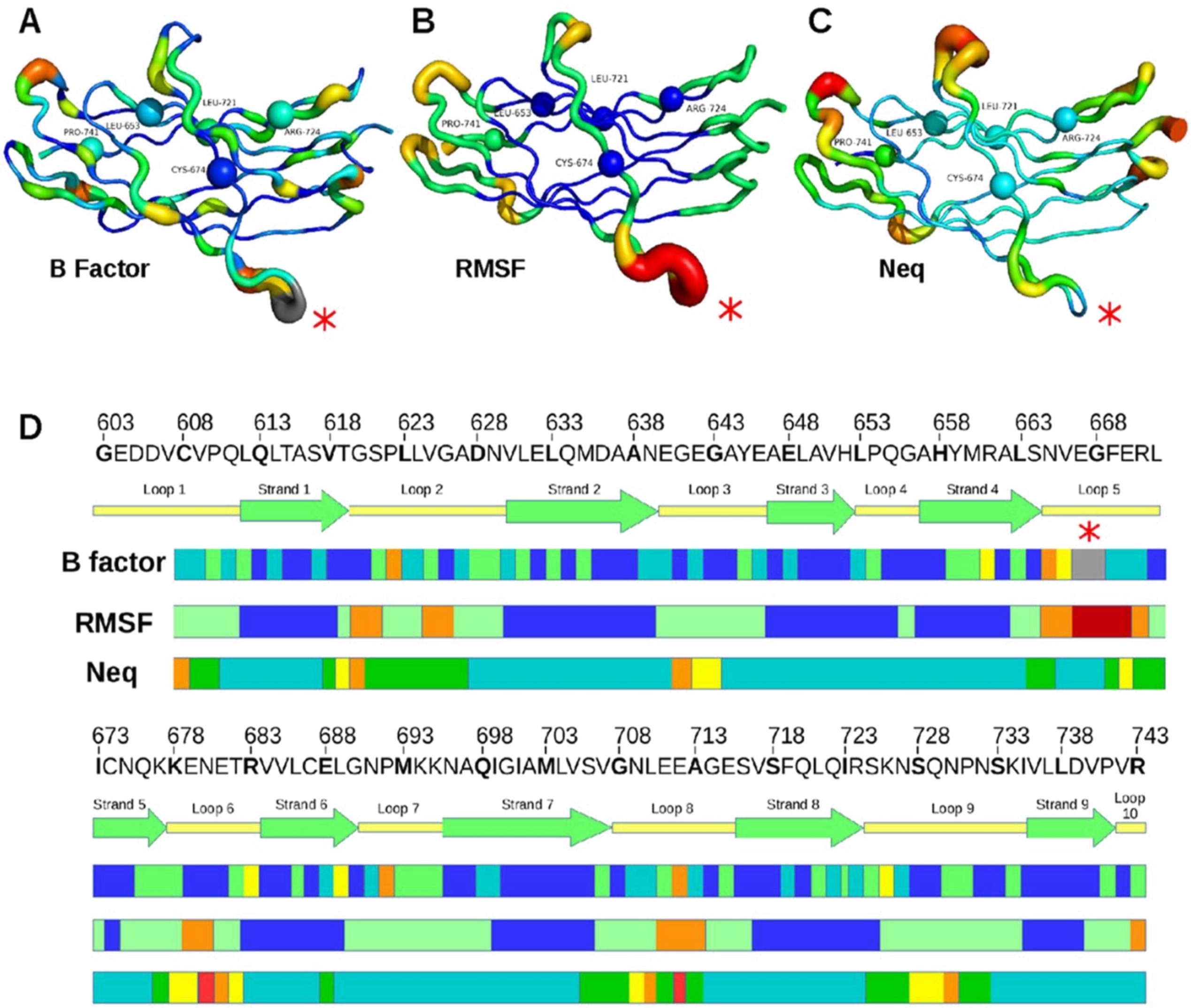


Figure 4.7 ***Comparison of the protein flexibility of Calf-1 through different metrics.*** *3D structures of Calf-1domain represented through (A) B-factor values, (B) RMSF values, and (C) N eq values. Local structure is ranked from rigid (thin blue line, a value of 0.0) to flexible (thick red line, a value of 4.0). Residues with completed missing atoms are in grey in the B-factor cartoon (A). (D) The Calf-1 amino acid sequence is placed in regards to its secondary structures assignment and to protein flexibility according to the B-factor, the RMSF or the Neq values. Blue, green, yellow, orange and red colours scale the structure from rigid to flexible. The loops are: loop1 (size: 9, positions 603–611), loop 2 (size: 10, positions 620–629), loop 3 (size: 7, positions 640–646), loop 4(size: 4, positions 653–656), loop 5 (size: 8, positions 665–672), loop 6 (size: 6, positions 678–683), loop 7 (size:6, positions 690–695), loop 8 (size: 8, positions 708–715), loop 9 (size: 11, positions 725–735), and loop 10 that begins at position 742.*

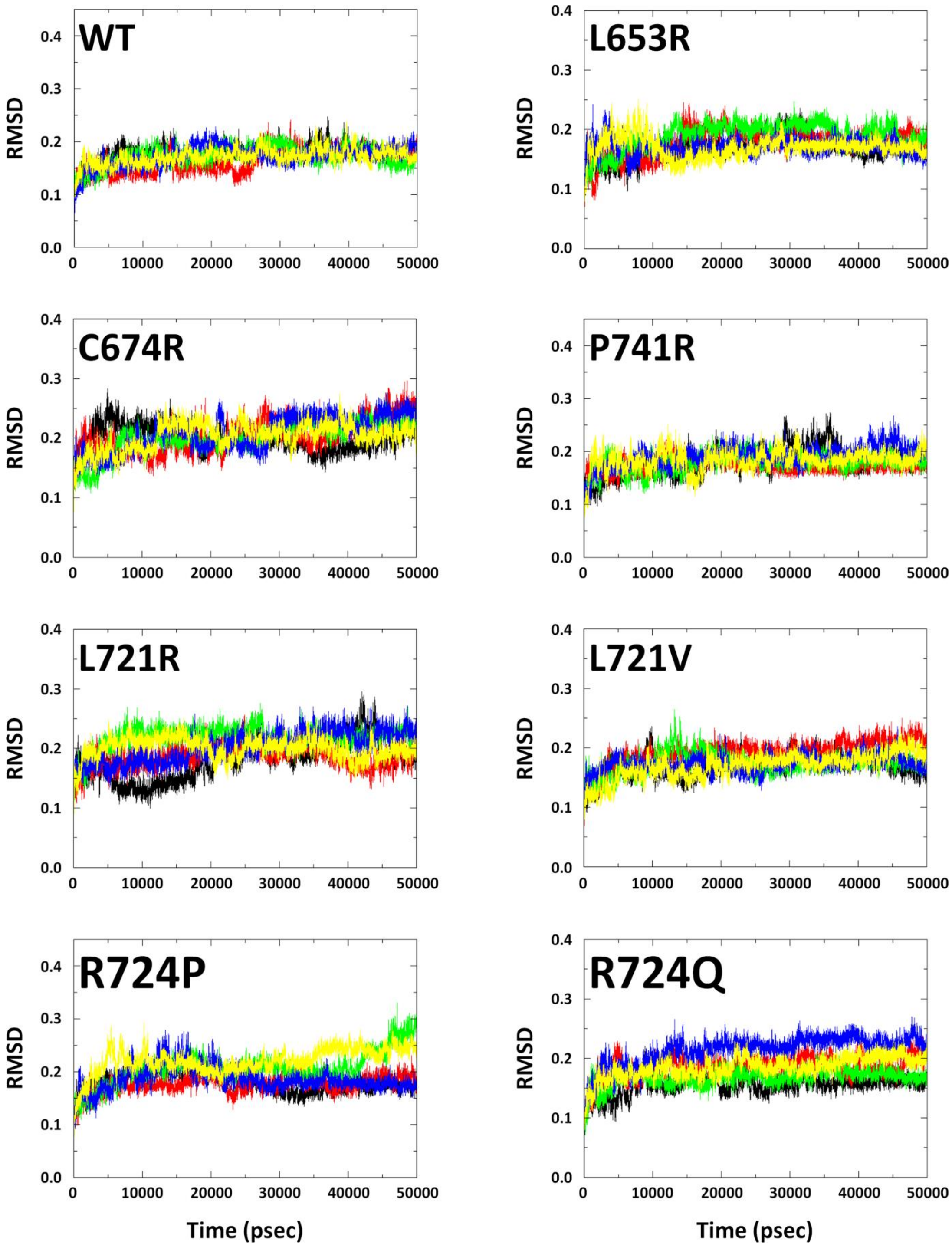


Figure 4.8 ***RMSD curves of the WT form of Calf-1 domain.*** *Shown are curves of the 5 MD simulations performed for 50 nsec. All curves converge at 25000 picoseconds to reach a steady state.*

4.3.3 *Inherent flexibility in the Calf-1 domain*

Some protein moieties that are very flexible in solution might seem to be rigid only because they are involved in the solid-state packing. RMSF values computed from MD simulations measure the mobility of each residue around its median position in the structure and allow assessing protein flexibility (Fig 4.7B). High RMSF values are often associated with loops and sometimes with C-ter of β-strands. As defined by high RMSF values, loop 2 (residues 619–620 and 625–626), loop 5 (residues 665–671) and loop 8 (residues 711–713) are flexible regions, with loop 5 being the most flexible. The rest of the structure is relatively rigid.

RMSF and B-factor values are correlated for loops 2, 5 and 8 (Fig 4.7D). Some points are noteworthy: (a) the limits of flexible positions can show some little differences between RMSF and B-factor and (b) loop 3 is associated to high B-factor but low RMSF values although it binds a Ca^{2+} (not included during simulations) in the crystal structure. Similar correlation between B-factor and RMSF values have been reported previously [238]. Figure 4.7[B to D] also indicate a good correlation between RMSF and *Neq* values. Indeed, highest *Neq* values are associated to flexible regions (as defined by B-factor and RMSF) with residues K678-T682 (loop 6) and N709-E712 (loop 8), but also with T619 (loop 2). Expectedly, some regions show higher *Neq* for some residues; G641-G643 (loop 3) and S728-N730 (loop 9). On the other hand, highly flexible region can also represent high local rigidity in terms of PBs, for instance, residues V666-F669 and E670 in loop 5 (Fig 4.7C and 4.9).

4.3.3.1 *Flexible yet rigid: Resolving ability of Neq*

Direct comparison of RMSF and *Neq* values (Fig 4.9A) clearly shows that E667 represents a high RMSF but a low *Neq*. This can be explained by its PB distribution (Fig 4.9B): E667, G668 and F669 representing the highest RMSF values (and also B-factors), mainly adopted the PB sequence "*hia*" with respective occurrences of 86.2, 82.9 and 61.6%. A series of PB "*hia*" is a classical loop conformation but this region (in blue rectangle on Fig 4.9C) maintains a single conformation and is not really flexible. This apparent discrepancy can be explained by the insertion of the rigid stretch E667-F669 in a larger flexible (or more precisely deformable) loop N665-L672. Interestingly, the results reveal that a locally rigid aa stretch (few possible conformations/low *Neq*) can be a part of a large mobile loop involved in the global structural motions of the protein (high RMSF). Overall, the results show a good correlation between experimental data (B-factor), RMSF and *Neq* obtained from MD simulations. Although some discrepancies did exist, they are explained by local structure singularities. As expected in an all-β domain, rigid β-strands are linked by flexible loops.

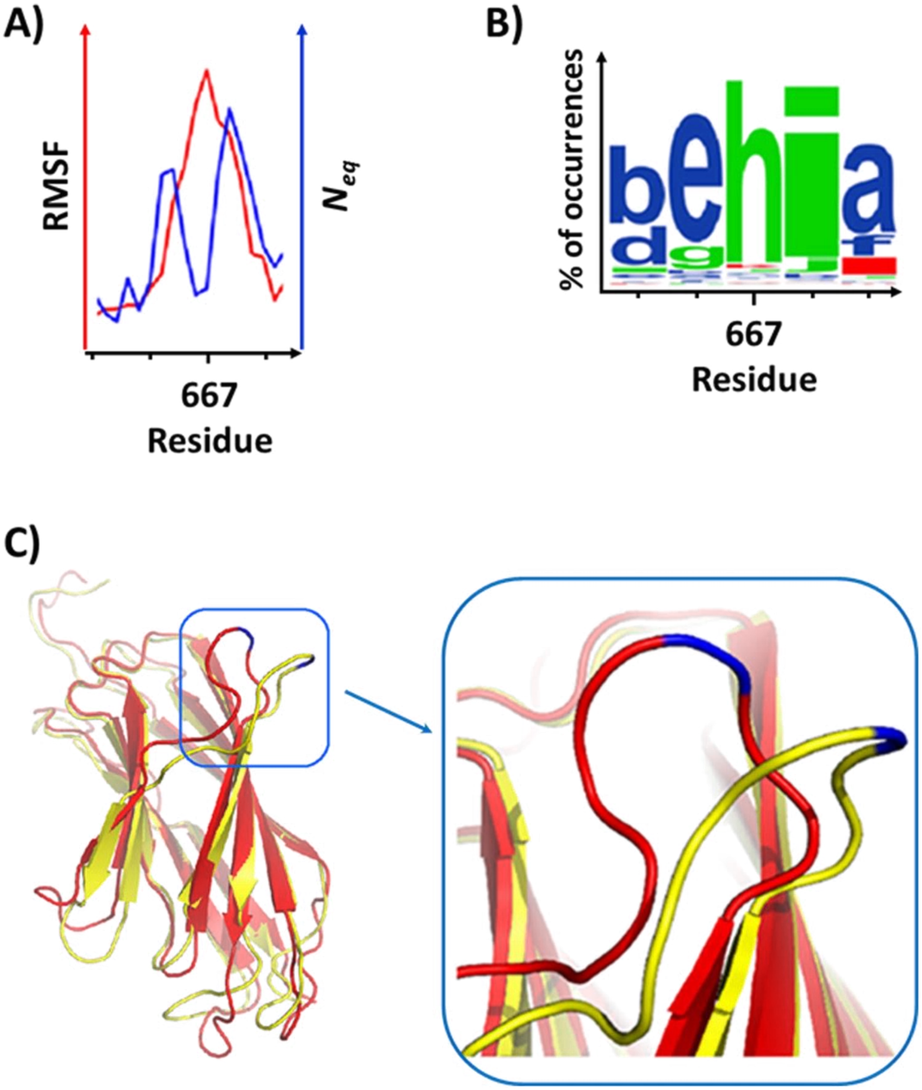


Figure 4.9 ***Local rigid conformation in a deformable loop, low Neq versus high RMSF.*** *(A) Superimposed RMSF and Neq values (red and blue curves respectively) from residues N665 to G668, (B) The WebLogo 49 indicates the frequency of occurrences with respect to the PBs adopted (size of the letter) by a residue in MD simulations. Here, residues V666 to F669 mainly adopted the PBs profile "ehia" corresponding to low Neq for them. (C) 3D model of the Calf-1 domain and the frame magnified of two adopted by the loop conformations (red and yellow worm-lines) carrying the residue E667 (in blue) that keep a rigid structure relative to the mobile loop.*

4.3.4 Comparisons of dynamics between GT variants and WT Calf-1

The $\alpha_{IIb}\beta_3$ integrin was cut into compact structural domains through Protein Peeling [358] that correlate the delineations found in literature [344]. As shown in Figure 4.10, the variant residues under investigation are mostly located at β-strands presenting low flexibility with the exception of residue 653 localized near the β-strand 3 C-ter. Similar to WT system, the 7 variants (structural mutants) were studied with 10 independent MD simulations performed to a complete timing of 1 µsec and with parameters similar to Jallu et al., 2014 [341]. Each system reached a plateau after 5 nsec with an average RMSD of 2Å (beginning of loop 1 and end of loop 10 excluded). All energetic and geometric parameters show a good evaluation for the 70 different simulations used in this study; no clashes are found. The Calf-1 domain stays consistent during the whole dynamics. Average RMSF from each variant and the WT were comparable (Fig 4.11). The most important variations observed in loop 2 (V625), loop 5 (E670), loop 8 (A713) and loop 9 (N732) did not lead to disordered patterns. Some variants showed specific higher or lower RMSF for some restricted positions like for C674R and L721V variants (Fig 4.11).

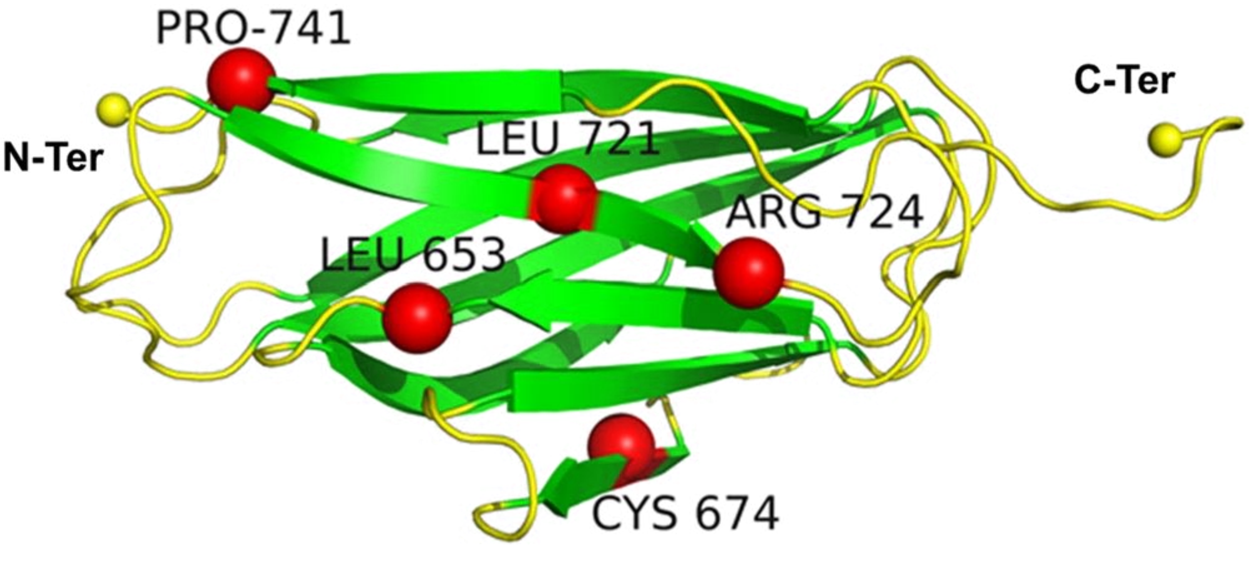


Figure 4.10 ***Ribbon model of the Calf-1 domain showing the location of the studied variant residues.*** *Strands are coloured in green and loops in yellow. Variant residues are identified as red balls. N-ter and C-ter ends are shown as yellow balls.*

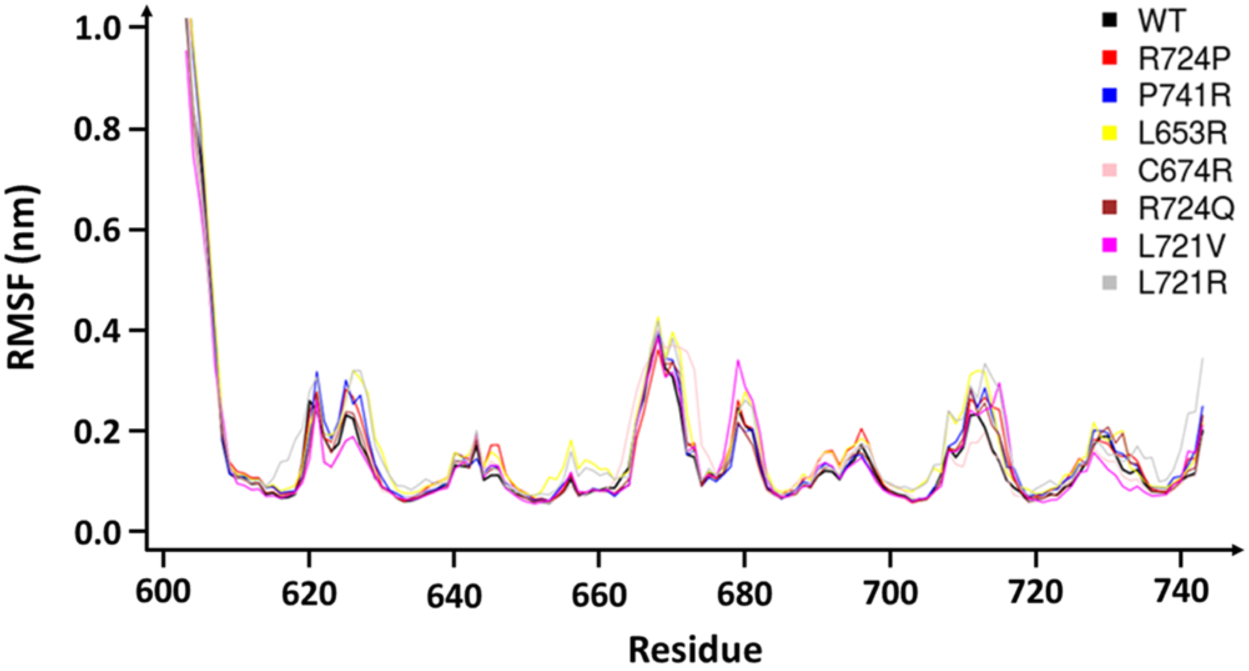


Figure 4.11 ***Calf-1 RMSF of the different systems.*** *By comparison, Calf-1 variant structures mainly behaved like the WT form (black curve). The noisy peaks for the N-ter first residues were discarded in the majority of the analyses since, in nature they lay at conjunction to the neighboring domain.*

4.3.5 *Protein block analysis of WT and variant dynamics*

To resolve deformed region from rigid and flexible regions, PBs analyses of the MD trajectories is performed. PBs analyses revealed striking local structure alterations, but distant from the variant sites. Three variants R724Q, L653R and C674R are found to be representative of all behaviors observed for the 7 variants.

4.3.5.1 *R724Q*

This aa variation is located at β-strand number 8. In regards to the WT structure (Fig 4.12), the highest *Neq* differences are at S621 (beginning of loop 2), A644 (loop 3) and L710 (loop 8). These loops that are naturally flexible are even more so in the variant. Therefore, an increase in flexibility is observer. Conversely, residues L624 to D628 have a lower *Neq* value thus indicating that loop 2 represents a dual behavior, with increased deformability at its beginning and enhanced stability in its C-ter part. Surprisingly, the mutant residue Q724 (β-strand 8) conserved the same

Neq (Fig 4.12_A) with a low ΔPB of 0.09 (Fig 4.12B) indicating that local β-strand conformation is conserved, i.e. PB *d*.

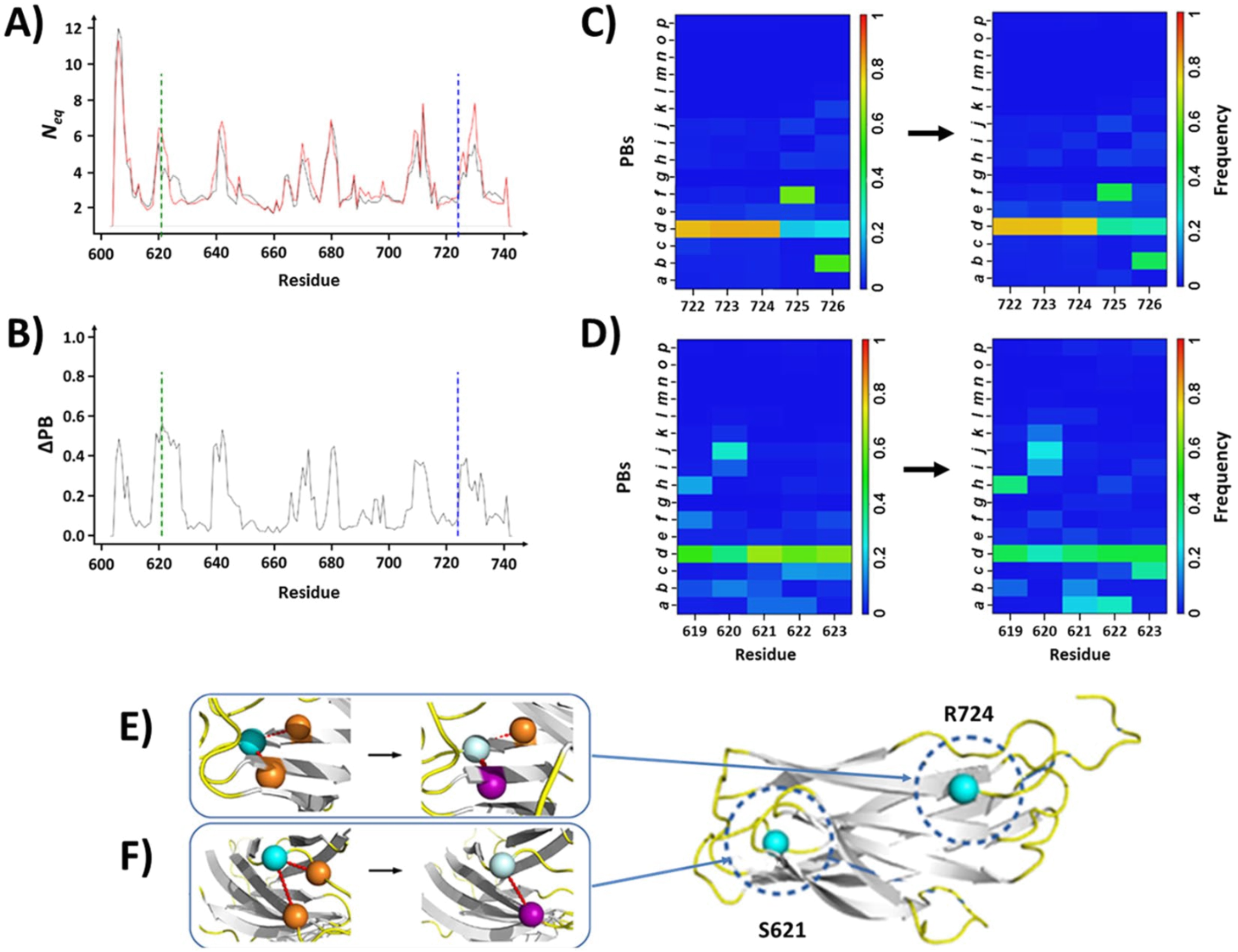


Figure 4.12 ***Variant R724Q.*** *(A) Neq values of residues from the WT system (black curve) and of the R724Q variant (red curve). Positions of the residues 724 and S621 that presented the highest RMSF alteration are respectively indicated by blue and green dots lines. (B) Curve of the ΔPB values computed from the difference between the two systems. (C) PB maps from residues 722 to 726 for the WT (left) and the variant (right)systems. (D) PB maps for residues 619 to 623. Color scales indicate the frequency of occurrences of the PBs in the map. (E) Molecular interactions made by the residues 724 in the WT and the variant systems (left and right cartoons respectively). (F) Molecular interactions made by the residues 621 in the WT and the variant systems (left and right cartoons respectively). Residues 724 and S621 are shown as cyan balls in the WT form, and as light cyan balls the variant. Orange balls indicate residues that conserved their interactions with the residues724 or S621 while magenta balls correspond to residues with modified interactions. A cartoon of the Calf-1domain shows the respective locations of the residues 724 and 621*

Regarding the structure, the polar amino acid arginine contains a longer aliphatic side-chain than glutamine, an uncharged hydrophilic polar amino acid. Q724 conserves the backbone - backbone interaction with E648 as observed with R724 (β-strand 3, see Fig 4.12E). Besides, Q724 lost the ionic bond and the side chain - side chain interactions with E648 but made new hydrogen bonds through side chains interactions with E722. This showcase a classic example of structural compensation that maintained the local conformation of the residue through different interactions. The highest Δ*Neq* (2.71) that is also associated with the highest ΔPB (0.57), is observed for S621 (Fig 4.12A, B). S621 is located at the opposite side of the domain in reference to residue 724 (Fig 4.12E). In the variant structure, S621 mostly remained in a PB *d* (i.e., β-strand) conformation with however, a decreased frequency of occurrences. Besides, downstream P622 and L623 presented some lost conformers with increased frequencies of PBs *e* and *h* respectively. Very few typical backbone - backbone interactions of S621 with L623 and backbone - side chain interactions with N629 are replaced by a single bond between side chains with N629. Adding to this high mobility, S621 did not do consistent and sustainable interactions. This behavior is amplified in the Q724 variant with the most stable residue S621 in a naturally flexible region (loop 2), became one of the most deformable positions.

4.3.5.2 *L653R*

This GT variant results from a L653R substitution in loop 4. The highest *Neq* variations (Fig 4.13A) affected residues G620 - P622 (loop 2), V630 - L631 (β-strand 2), E646 (loop 3), R671 (loop 5) and L710 (loop 8). As observed with the R724Q substitution, residues G620 - L623 gained slightly more flexibility. Conversely, residues L624 - D628 shows increased flexibility but with a limited impact (average ΔPB = 0.23) on the most frequent PBs (PB *e* for L624, *h* for V625 and *i* for G626 in Fig 4.13B). The mutated residue in position 653 (loop 4) is not subjected to any *Neq* modification. It conserves a strong local structural stability (Fig 4.13C) similar to its direct environment. The PB series at this position "*dddeh*" is even slightly more common in the variant than in the WT (64% and 59%, respectively).

In the R653 variant, the 8 hydrophobic bonds of L653 disappeared in favour of new interactions between the R653 backbone and A657 and E676 side-chains (Fig 4.13E). The backbone – backbone interaction with R683 remained conserved. The mutation zone showed no conformational change as the loss of important specific interactions were partly compensated by new ones. Of the 9 original interactions only 1 is conserved while 3 new are created.

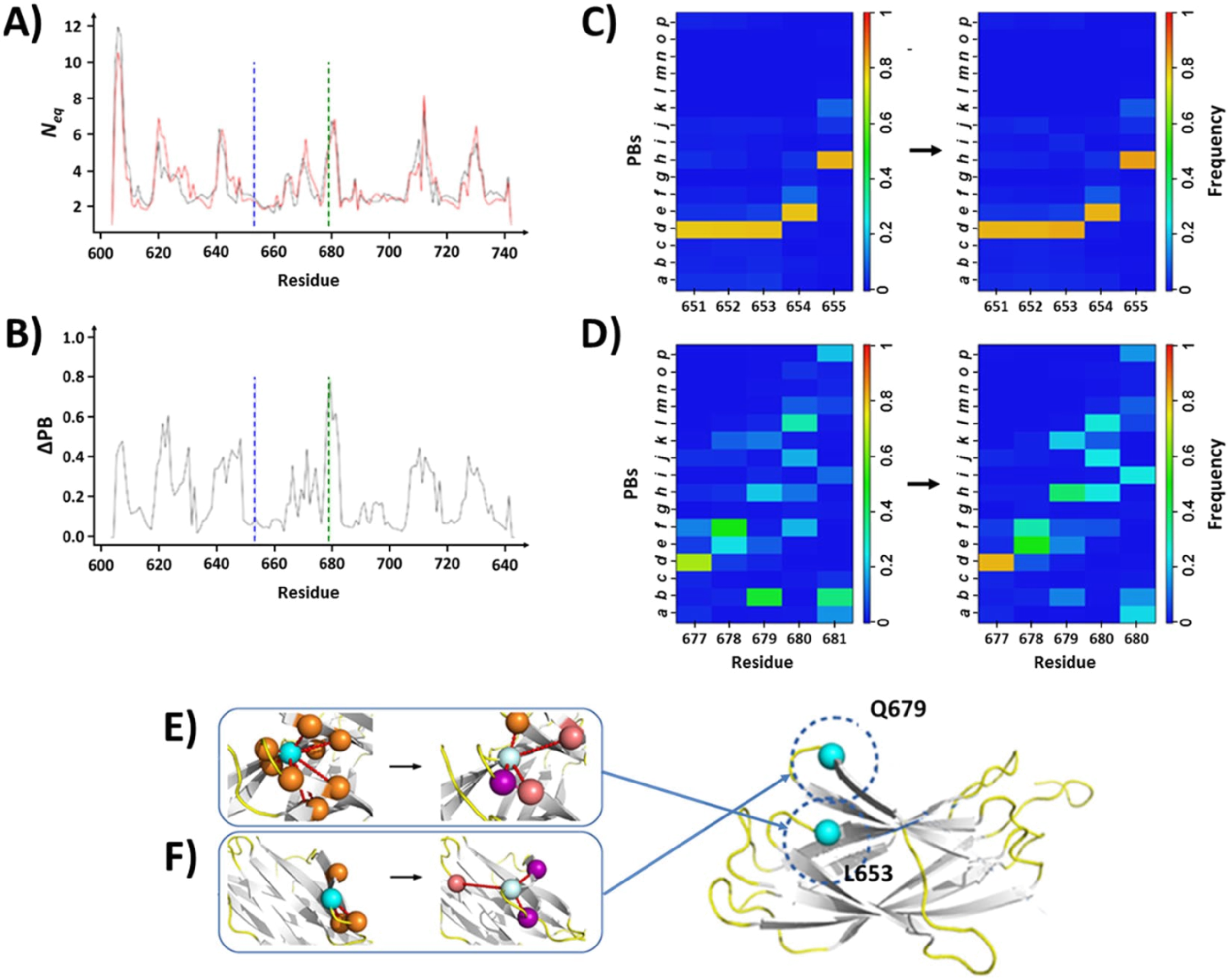


Figure 4.13 ***The variant L653R.*** *Panels (A and B) respectively show the Neq and the ΔPB curves of the L653 (WT) and R653 (variant) systems. Panels (C, D) respectively show the PB maps of residues 651 to 655 and 677 to 681 with the WT at the left and the variant form at the right. (E) Molecular interactions made by the residues L653or R653 and (F) by E679. For colour scales and residue presentation, see the legend of Figure 4.12.*

Q679 (loop 6) is a very interesting case where Δ*Neq* is negligible while the ΔPB is the highest (0.78). The most frequent PB *b* (N-ter of β-strand) is replaced by a PB *h* (loop structure) in regards to their frequency of occurrences (Fig 4.13D).

Hydrogen interactions with T682 and K677 are retained but the backbone - backbone interaction with E681 is lost and replaced by side chain and ionic side chain interactions with R724 in loop 9. In the variant structure, this region has high fluctuations in PBs, mainly associated to loops that even affected the C-ter of the β-strand 5 located the above loop 9.

4.3.5.3 *C674R*

This variant is associated with a C674R substitution in β-strand 5. An *Neq* profile (Fig 4.14A) similar to that occurring in the R724Q substitution is observed (see section 4.3.5.2). Loop 2 presented the same increased deformation at its beginning (S621), followed by a stiffening in its centre (residues L624-D628). The same PB series "*ehiac*" (L624 - D628) is found in greater proportion than in the Q724 and R674 variants, reinforcing the local stiffening of the loop in this region. The main destabilization was far upstream of residue 674 (Fig 4.14C).

Loss of a disulfide bridge: With the C674R substitution, the residue 674 not just lost its covalent disulfide bond with C687 located at the end of β-strand 6, but also its aromatic interaction with Y659 in β-strand 4 (Fig 4.14D). However, the mutated R674 made an ionic bond with E688 located at end of β-strand 8 that strengthened a backbone – backbone interaction. The 80% frequency of PB *d* (the highest) in WT decreased to 49% in the variant. Surprisingly, N675 and Q676 located downstream the substitution remained structurally stable with similar PB occurrences.

The highest *Neq* variation affected R671 as shown by the strongest Δ*Neq* (5.02) and ΔPB (0.91). The side chain of R671 is mainly exposed at the domain surface and forms a single ionic interaction with the neighboring E670, like in WT. But in the variant conformation, the R671 side-chain can occasionally turn towards loop 8 to make ionic side chain interactions with E688 (Fig 4.14E). The frequency of PB *d* (the highest) drastically decreased in the variant leading to an increased disorganization of the neighborhood.

Experimentally, the C674R mutation severely impaired the $\alpha_{IIb}\beta_3$ complex expression with only 10% of the integrin expressed at the surface of the patient's platelets and transiently transfected CHO cells. However, the C674R mutation did not impair pro-α_{IIb} synthesis but affect the stability of the complex that is not correctly matured and/or expressed at the cell membrane.

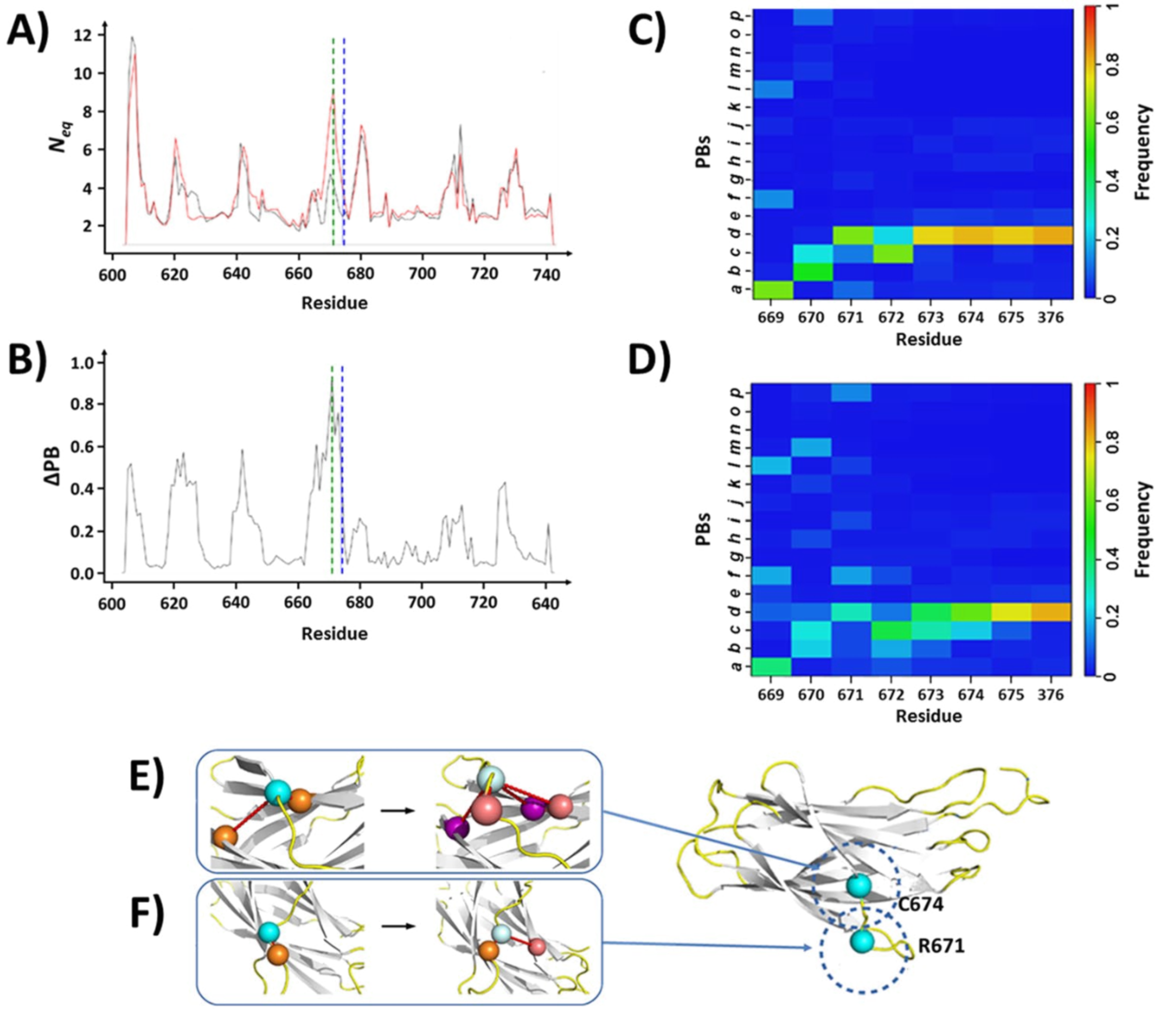


Figure 4.14 ***The variant C674R.*** *Panels (A and B) respectively show the Neq and the ΔPB curves of the C674 (WT) and R674 (variant) systems. Panel C show the PB maps of residues 669 to 676 for the WT (above) and the variant forms (below). (D) Molecular interactions made by the residues C674 or R674 and (E) R671. For colour scales and residue presentation, see the legend of Fig 4.12.*

4.3.5.3 *Other variants*

Proline is the aa known to cause the most drastic change in conformations [359]. Indeed, the P741R substitution (Fig 4.15) inverse the PB profile going from 55% of PB *d* (β-strand) and 29% of PB *f* (C-cap of β-strand) to 24% of PB *d* and 59% of PB *f*. This case was associated with a low Δ*Neq* (0.15) while the ΔPB was high (0.70). In P741R substitution two hydrophobic interactions were lost and R741 formed ionic and side chain - side chain interactions shortening the β-strand.

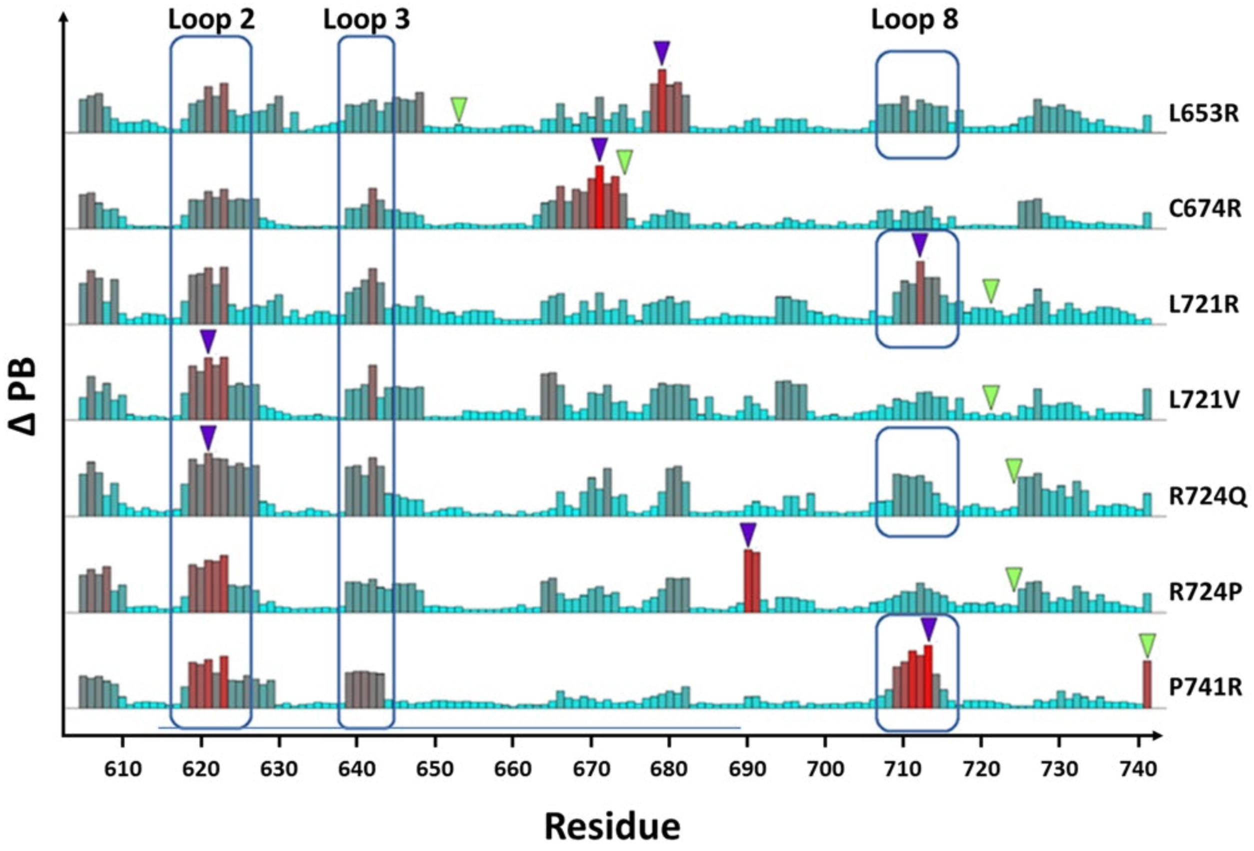


Figure 4.15 ***Comparative study of the ΔPB values for all WT-variant pair****s. Histogram schema presenting the ΔPB value computed for each residue position (abscissa) from each variant and the WT systems. Green triangle indicates the aa variation position while the purple one position shows the position of maximal ΔPB. Residues from loops 2, 3 and 8 presenting common high ΔPB for all variants are boxed.*

In the remaining 5 variants studied, compensation mechanisms were also observed. Most interactions formed by WT residues are replaced by new ones, allowing conservation of the local structure. Surprisingly, regions displaying significant changes (high ΔPB) are distant from substitution sites without any contact/interaction with the substituted aa. These regions contribute towards increasing the deformability and are usually located at interfaces adjacent to neighborhood β-propeller, Calf-2 or Thigh domains. These results depict changes resulting from substitutions in distant regions suggesting long-range mechanism to be at play.

Different variants with common mutation sites. L721R and L721V showed quite different results (Fig 4.15). Compared to L721R, the L721V substitution had very little impact on RMSF, apart for the end of loop 8, which is a highly flexible region. This is particularly true for E712 (loop 8), whose Δ*Neq* were respectively, 3.33 and 0.

4.4 Conclusions and future perspectives

The MD simulations of Calf-1 domain allowed to demonstrate more or less pronounced structural changes in the wild type structure as well as the impact of GT variants. The analysis gets huge enhancement by using protein blocks statistical measures like Δ*Neq* and ΔPB. These helped in closely evaluating the regions that comprised of a local regions of rigidity inside otherwise deformable regions, for instance as analysed in case of Glu 667, Fig 4.10. Flexibility profile of the Calf domains showed that although their anchoring role demands them to have a rigid core yet the connecting loops contribute to the structural dynamics of the core. This principle gets even more profound by studying the effect of GT variants on the Calf-1 structure. Overall in the structural mutants, the beta-strand core tends to maintain or regain rigidity which can be attributed to its structural role in the big integrin complex. However, the impact of GT variants that may disturb the core are systematically compensated by the loops. The energy gain or loss due to lost interactions in mutants is shown to be compensated by new interactions and the residual energy is apparently transferred to the loops. This causes the long range effects of the impact of mutation, as observed at residues L653, L721, and R724.

While, mutation at C674 and P741 variants displayed conformational changes at the mutated site, predominantly. In the case of the C674R substitution, the resulting loss of the disulfide linkage relaxes the structure and introduces significant structural alterations (Fig 4.14 and 4.13). Such an effect is largely suggestive that the structural-functional context of the structure influences the rigidity. Thus, inherent flexibility is important and crucial to the conservation of the core.

For P741R it should be noted that residue 741 is located at two residues upstream from the C-term of Calf-1 and is normally in contact with the Calf-2 domain. Thus, the absence of the neighboring domain in Calf-1 MD simulations can impact the observations. To resolve, MD simulations of the complete domains have been performed. However, the technical failure of our computational cluster inhibits the inclusion of results from Calf-2, Calf-1 + Calf2 + knee + Thigh domains. Nonetheless, similar observations were also made in the dynamics of these domains as well. The calcium containing domain, ‘Genu’ (knee) seems to play a key role in assisting the flexible domain to be stabilized during structural changes from leg to thigh regions.

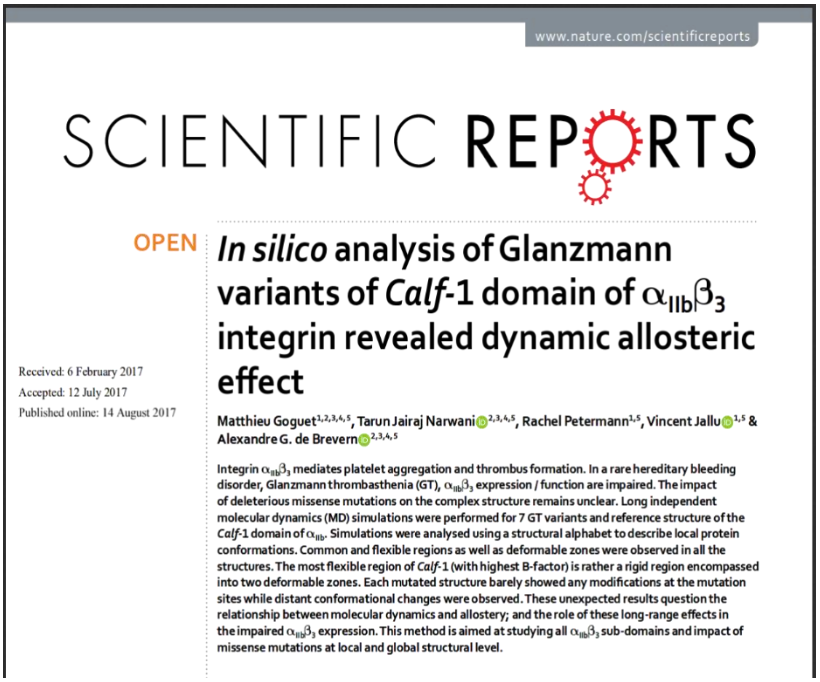
Although the primary objective was to profile the inherent flexibility in all-beta, rigid Calf domains yet the evaluation of the dynamics of GT variants enhanced our understanding of local structure dynamics. With deservingly expected developments in the project, it will be interesting to compare and test the inferences from leg region with apparently flexible domains of β_3 subunit.

ACKNOWLEDGEMENTS

All the MDs were run by four master students, Matthieu Goguet (M2), Sali Anies (M2), Sarah Kaddah (M1), and Soubika Bisoo (M2) under my supervision. Without their sincere efforts, the simulations might have not been finished during my thesis tenure. Therefore, a very warm thanks to them. Sincere hardwork done by Dr. Vincent Jallu in finding and selecting different variants is incredible and is fundamental to the project.

Dissemination of results

The results from Calf-1 showing effects of 7 GT variants on the wild type structure of integrin $\alpha_{IIb}\beta_3$ have been published as: Goguet M.*, Narwani T.J.*, Petermann R., Jallu V., de Brevern A.G. In silico analysis of Glanzmann variants of Calf-1 domain of $\alpha_{IIb}\beta_3$ integrin revealed dynamic allosteric. Sci Rep (2017) 7(1):8001



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

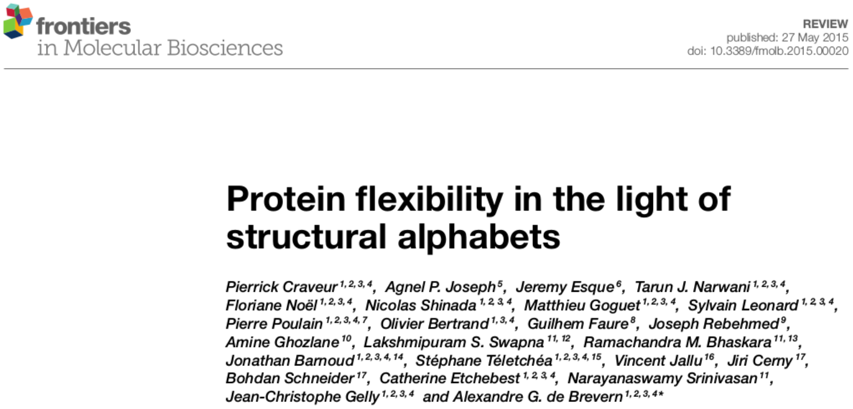
In silico analysis of Glanzmann variants of _Calf_-1 domain of $\alpha_{IIb}\beta_3$ integrin revealed dynamic allosteric effect

Received: 6 February 2017
Accepted: 12 July 2017
Published online: 14 August 2017

Matthieu Goguet[1,2,3,4,5], Tarun Jairaj Narwani[2,3,4,5], Rachel Petermann[1,5], Vincent Jallu[1,5] & Alexandre G. de Brevern[2,3,4,5]

Integrin $\alpha_{IIb}\beta_3$ mediates platelet aggregation and thrombus formation. In a rare hereditary bleeding disorder, Glanzmann thrombasthenia (GT), $\alpha_{IIb}\beta_3$ expression / function are impaired. The impact of deleterious missense mutations on the complex structure remains unclear. Long independent molecular dynamics (MD) simulations were performed for 7 GT variants and reference structure of the *Calf*-1 domain of α_{IIb}. Simulations were analysed using a structural alphabet to describe local protein conformations. Common and flexible regions as well as deformable zones were observed in all the structures. The most flexible region of *Calf*-1 (with highest B-factor) is rather a rigid region encompassed into two deformable zones. Each mutated structure barely showed any modifications at the mutation sites while distant conformational changes were observed. These unexpected results question the relationship between molecular dynamics and allostery; and the role of these long-range effects in the impaired $\alpha_{IIb}\beta_3$ expression. This method is aimed at studying all $\alpha_{IIb}\beta_3$ sub-domains and impact of missense mutations at local and global structural level.

The use of derived statistical tool from PBs, Δ*Neq and* ΔPB in assessing structural flexibility have also been summarized in the form of a review article. The article is titled as: Craveur P., Joseph A.P., Esque J., Narwani T.J., *et al*. Protein flexibility in the light of structural alphabets. Frontiers in Molecular Biosciences - Structural Biology (2015).



frontiers
in Molecular Biosciences

REVIEW
published: 27 May 2015
doi: 10.3389/fmolb.2015.00020

Protein flexibility in the light of structural alphabets

Pierrick Craveur[1,2,3,4], Agnel P. Joseph[5], Jeremy Esque[6], Tarun J. Narwani[1,2,3,4], Floriane Noël[1,2,3,4], Nicolas Shinada[1,2,3,4], Matthieu Goguet[1,2,3,4], Sylvain Leonard[1,2,3,4], Pierre Poulain[1,2,3,4,7], Olivier Bertrand[1,3,4], Guilhem Faure[8], Joseph Rebehmed[9], Amine Ghozlane[10], Lakshmipuram S. Swapna[11,12], Ramachandra M. Bhaskara[11,13], Jonathan Barnoud[1,2,3,4,14], Stéphane Téletchéa[1,2,3,4,15], Vincent Jallu[16], Jiri Cerny[17], Bohdan Schneider[17], Catherine Etchebest[1,2,3,4], Narayanaswamy Srinivasan[11], Jean-Christophe Gelly[1,2,3,4] and Alexandre G. de Brevern[1,2,3,4]*

Another manuscript emphasizing on the modelling approach for long missing regions of Calf-2 and its dynamics is under preparation.

Chapter 5: ***Protein dynamics in structural assemblies- An affair of ACKR1 and Plasmodium vivax***

5.1 Introduction

As have been seen from subsequent chapters that local protein structures are context dependent. For instance, in chapter 3, differences in dynamics of secondary structures can be observed between PTM and DPF dependent contexts. Moreover, in chapter 4 the dynamics of backbone is preserved, although key residues lying in beta-strands are mutated. Both the Calf domains in Integrin $\alpha_{IIb}\beta_3$ have a structural role of anchoring the chain to the cytoplasmic membrane. Therefore, the ambitious mutation in the core of the domain had compensatory effects to preserve its function. Thus the objective is to understand protein flexibility in a more complex structural organization that is, a multimeric assembly. Also, in the preceding chapters the domains and local structures under investigation can be encapsulated as having a globular nature. Therefore, to add contrast to the structural contexts studied so far, a transmembrane protein is selected as a case study to understand protein dynamics in structural assemblies.

The selected protein is Duffy Antigen / Chemokine Receptor (DARC). DARC has a physiologically promiscuous behavior in humans while being corruptly implicated in Malaria. DARC is a transmembrane GPCR and thus expectedly have scarce information about its structure. Therefore, the primary challenge will be to generate a robust structural model and consequently investigate the dynamics of its important structural regions. Being a GPCR implicated in a pathology like Malaria, its structural dynamics can be exploited to design effective inhibitors for Malarial transmission.

5.1.1 *Malaria*

Four Plasmodium species *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale,* and *Plasmodium malariae*, are the cause of malaria in *Homo sapiens sapiens* while a simian parasite *Plasmodium knowlesi* may also be able to infect [360]. Of these *Plasmodium falciparum* malaria mostly lead to fatalities while the rest leads to milder yet recurring and severe infections. Although, *Plasmodium vivax* infections are not as fatal as *Plasmodium falciparum,* yet it is the most widespread malaria causing species in Asia, Europe and Americas [361,362][363] as shown in the

Figure 5.1A. *Plasmodium vivax* has a high morbidity rate in the developing countries of south-east Asia, southern America and, Africa with upto 140 million cases of *Plasmodium vivax* malaria per year [361]. Of these 80% of infections are reported from Asia and south America, while only 12.4% are acquired in Africa. The populations of western sub-saharan Africa are resistant to *Plasmodium vivax* infections, Figure 5.1B [364]. This is attributed to a silencing mutation that selectively abolishes the expression of Duffy Antigen/Chemokine Receptor (DARC) on erythrocytes (Red Blood Cells).

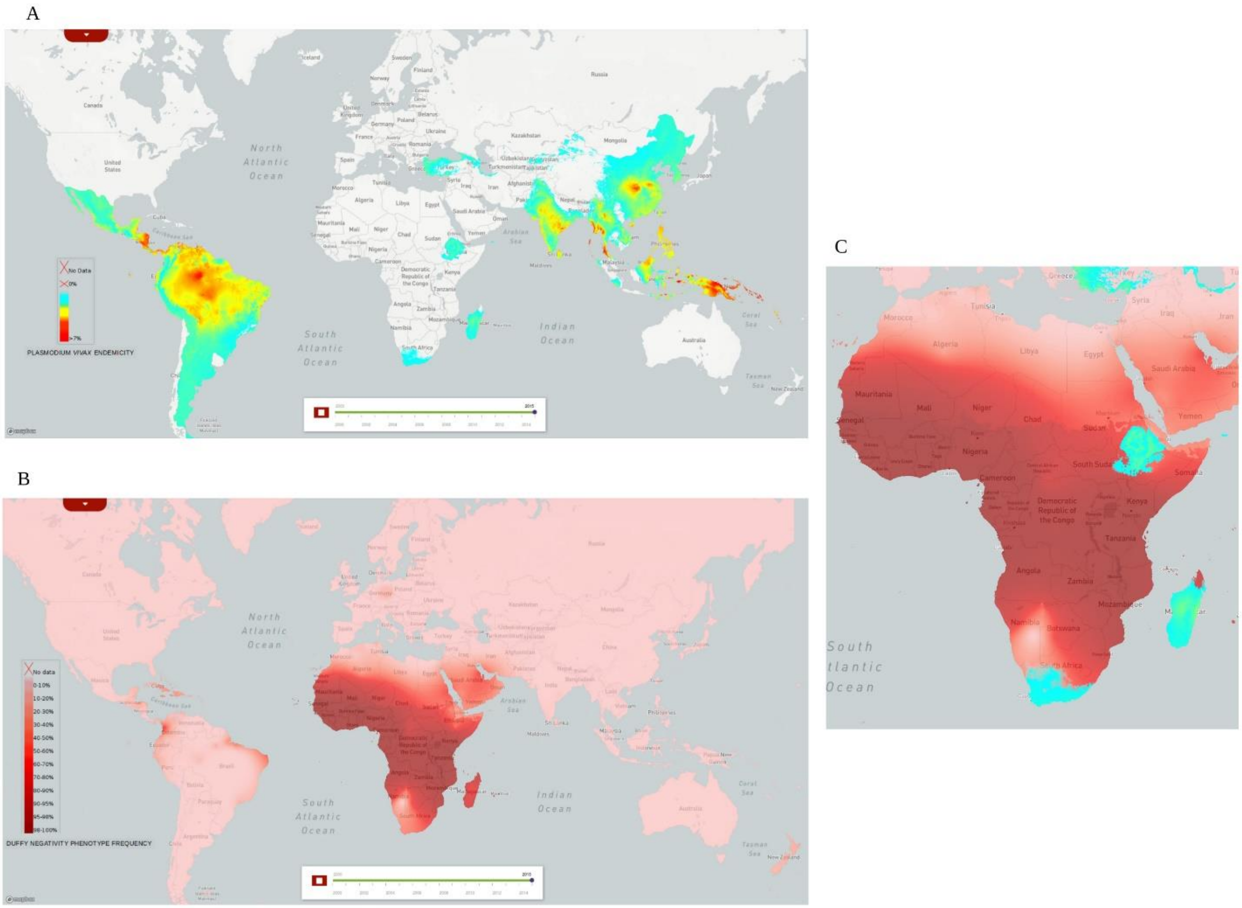


Figure 5.1 *An affair of Plasmodium vivax and ACKR1*. *A) shows the endemicity of the Plasmodium vivax infected malaria with red depicting highest and cyan depicting lowest no. of cases. Most of the incidents are reported in the tropical and sub-tropical climate zones, except most of African continent. B) depicts the spatial distribution of Duffy negative population across the world. Ranges are shown as gradient of color red with pink being negligible while more red indicated more Duffy negative population. As can be seen from C) that the Duffy negativity totally complements Plasmodium vivax malaria trends, especially in African continent.*

+The images are generated from the Malaria Atlas project (https://map.ox.ac.uk/).

5.1.2 *Plasmodium vivax*

Plasmodium vivax, like other plasmodium parasites, have a dual phased, digenetic life cycle. The sexual cycle or schizogony is carried out in the host, female *Anopheles* ending in the generation of sporozoites [365]. The blood meal of the mosquito on human commences the asexual cycle or sporogony. This culminates with the production of gametocytes that are sucked by the host mosquito during blood meal. While in the asexual phase, the *Plasmodium vivax* merozoites undergo two stages of development in liver and three stages of infection in erythrocytes called erythrocytic cycle. Malarial merozoites must invade erythrocytes to begin the infection stage and this makes the invasion a critical step in the life cycle of *Plasmodium vivax*. A bottleneck in the merozoites entry process is the interaction of their micronemes with a septa-helical transmembrane, glycoprotein on erythrocytes known as DARC. However, in liver, the merozoites can also enter a dormant stage called hypnozoites. Figure 5.2 shows a schematic representation of *Plasmodium vivax* life cycle.

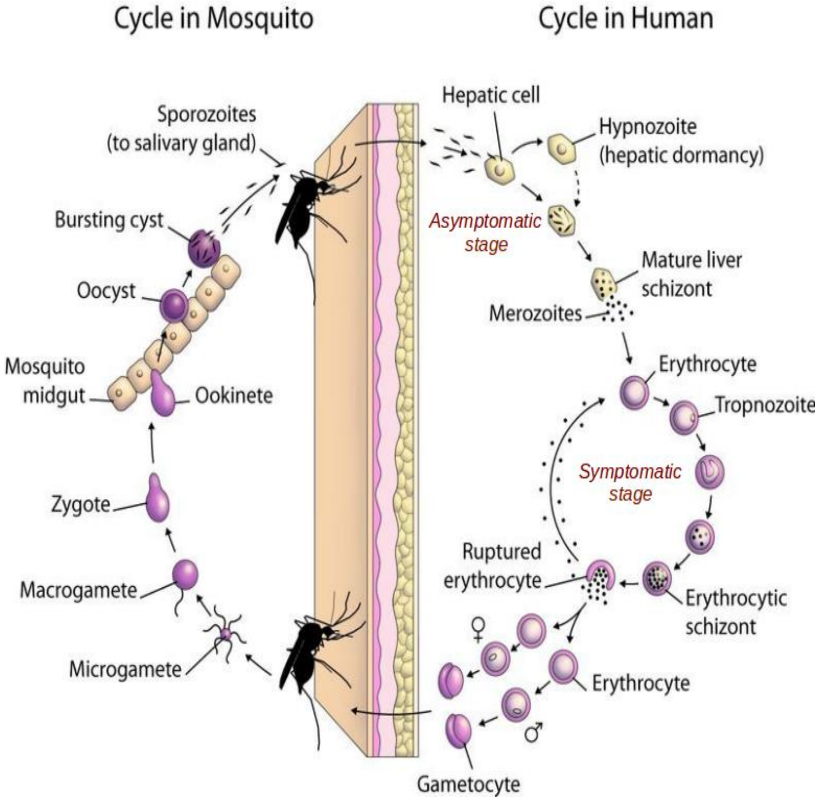


Figure 5.2 ***Life cycle of Plasmodium.*** *Parasites of Plasmodium genus have digenetic life-cycle. The sexual part of the life-cycle is carried out in vectors, mostly mosquito. The infectious stage is executed in the host cells, mostly Humans. After a blood meal the parasite is transferred to the host via salivary glands. The parasite multiplies in the hepatic cells but is asymptomatic and*

therefore difficult to diagnose malaria. In 2-3 weeks, the parasites rupture the hepatic cells to enter blood stream where they infect Erythrocytes. The erythrocyte infection leads to malarial symptoms. The parasite cells multiply in the erythrocytes and spread over and thence when the vector takes another blood meal the parasite enters the blood stream of vector to carry on the sexual cycle.

$^{+}$*Image credits to Stephan Kappe, PhD, University of Washington*

5.1.3 An introduction to DARC

Duffy Antigen / Chemokine Receptor is a minor blood group antigen that expresses Human alloantigens, Fy^{a} and Fy^{b} in its N-terminal extracellular domain [366]. It was discovered in western Africa and is allegedly named after the individual whom it was discovered in [367]. Its official denomination is Atypical Chemokine Receptor 1 (ACKR1) while alternatively it had been previously termed as Fy glycoprotein (Fy) or Cluster of Differentiation 234 (CD234). ACKR1 is encoded by a single copy gene, DARC, located on chromosome 1 [368]. DARC gene exists as two co-dominant alleles Fy^{a} and Fy^{b} arising due to a base mutation, G125A [369]. In Fy^{a}, the mutation in the 42nd codon leads to the encoding of glycine while in Fy^{b} it encodes aspartic acid, described by polymorphism- G42D [370]. These alleles are immunologically distinct and therefore would result into four Duffy blood group phenotypes: Fy^{a+b+}, Fy^{a+b-}, Fy^{a-b+} and Fy^{a-b-}. Some minor phenotypes have also been characterized namely, Fy3, Fy4, Fy5, Fy6 and Fy^{x} (weak expression of Fy^{b}) [371]. Fy^{a-b-} also called Fy-null phenotype, arises due to a polymorphism in Fy^{b} at 46th nucleotide (T46C) in the erythroid regulatory element of the DARC promoter region. The mutation leads to the disruption of the binding site for erythroid transcription factor, GATA1, in erythrocytes derived from hemopoietic lineage [372], as depicted in Fige 5.3. Therefore, the Duffy -ve or Fy-null blood group phenotype would abolish the expression of ACKR1 on erythrocytic membrane. While, the Duffy -ve individuals can have ACKR1 expressed on the endothelial lining of postcapillary venules, epithelial cells of renal collecting ducts [373] and Purkinje cells [374]. Since *Plasmodium vivax* merozoites invade reticulocytes by interacting with ACKR1 (DARC), the duffy -ve population can avoid this parasitic invasion thus making them resistant to *Plasmodium vivax* infected malaria, Fig 5.1C.

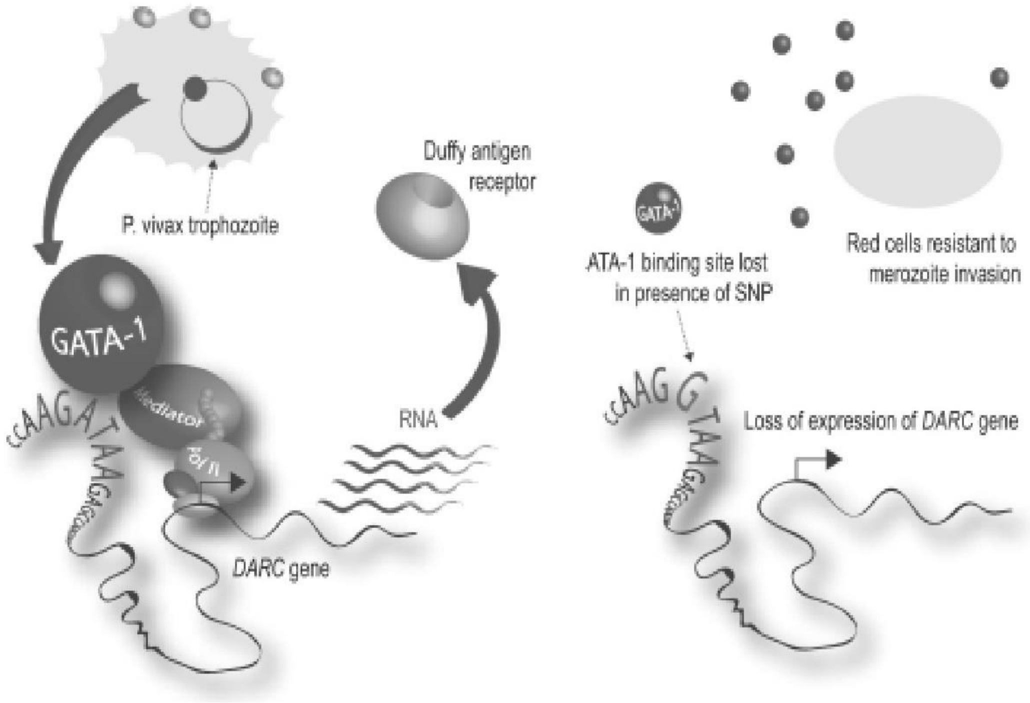


Figure 5.3 ***Disruption of GATA-1 box.*** *A schematic description of the expression of DARC gene in normal phenotypes (left) and in Duffy -ve phenotypes (right). In the promoter region of the gene, the SNP T46C alters the recognition site of GATA-1 thus leading to loss of expression of DARC gene. The SNP is predominant in the Fy^b allele lineage that expresses DARC on reticulocytes. Therefore, loss of DARC on RBC, will hinder the interaction with Plasmodium vivax DARC binding proteins.*

$^+$Image taken from [372]

5.1.3.1 *Sequence of ACKR1 (DARC)*

As it has been established that ACKR1 interaction is the primal point of access for *Plasmodium vivax* merozoites, it becomes crucial to understand the sequence-structure-function of ACRK1 and its mechanism of interaction with *Plasmodium vivax* micronemes. ACKR1 is a transmembrane protein belonging to Chemokine Receptor family which is a member of class A GPCR gene family [375]. Therefore, with a protein sequence of 336 amino acids (in Humans) ACKR1 adapts a rhodopsin-like structure identified by the classical 7-transmembrane helices linked by 6 connecting loops; 3 on the extracellular region (extracellular domains or ECD) while 3 on intracellular (intracellular domains or ICD). The extracellular region from N-terminal to the start of first helix is termed as ECD1 and the cytosolic C-terminal region is termed as ICD4. The first 60 amino acid residues of ACKR1 sequence constitutes the ECD1. The residues 60 - 316 forms the TM domain and residues 316 - 336 constitutes the ICD4 [241], see **Figure 5.4**. The members of Chemokine Receptors are structurally conserved especially in the 7-TM structure. The major variation arises in the length and amino acid composition of ECD1. It comprises of epitopes specific to the chemokine ligands and Fy blood group antigens. Besides, the ECD1 in ACKR1 also contains specific residues which are highly conserved given their roles in important interactions. A detailed description of such residues is provided in the subsequent sections in lieu of a fitter context.

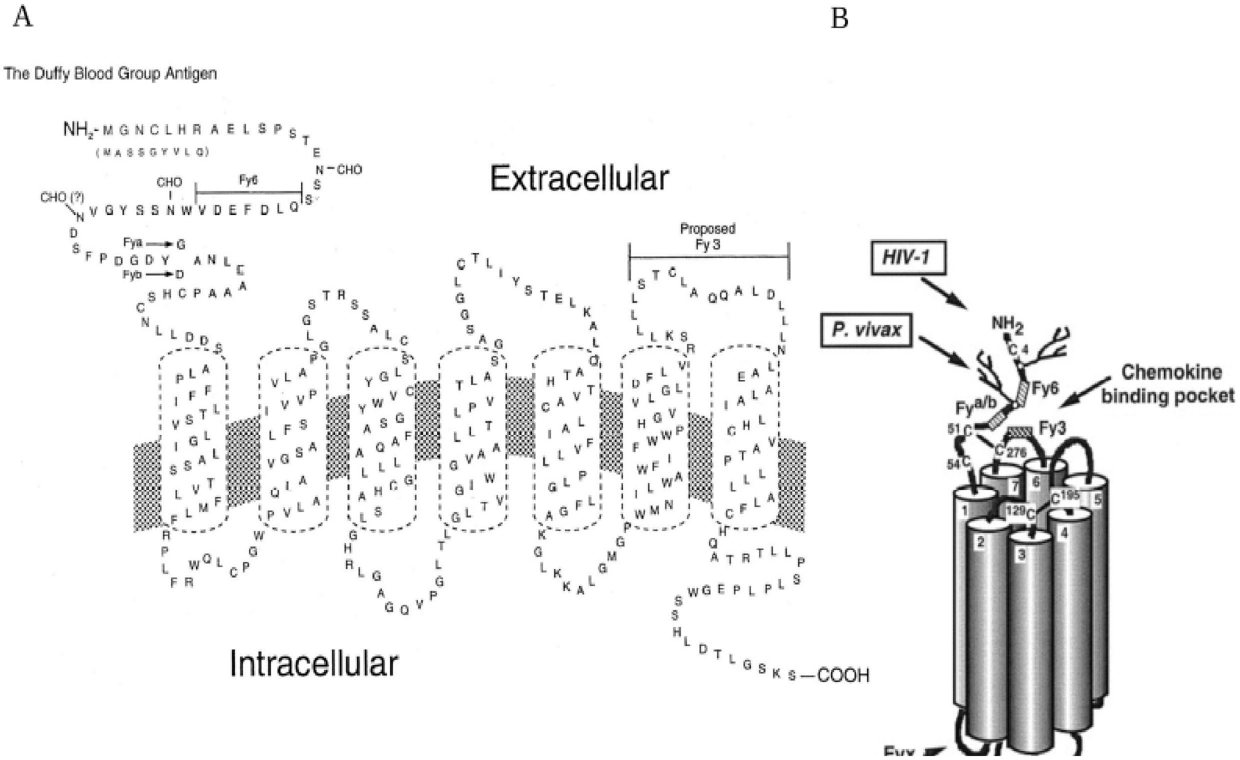


Figure 5.4. ACKR1. *A) a schematic representation of ACKR1 (uniprot ID: Q16570) highlighting the important epitopes for blood Fy antibodies. The SNP that leads to Fya and Fyb allele lineages is also shown. The conserved cysteines that forms the disulfide bridges are highlighted in green. The intracellular loop 2 is labelled as ICD2 and as seen, DRYLAIV motif is missing. B) shows a tentative model structure of ACKR1 monomer that shows the structural placement of conserved motifs and epitopes. G-proteins natural docking site is also shown but as ACKR1 lacks the DRY motif, G-proteins do not couple.*

+Image credits: A) taken from [374], B) adapted from [430]

5.1.3.2 *Physiology of ACKR1 and Chemokine Receptors*

Chemokine receptors are classified by the type of chemokine ligand it interacts with. The receptors which interact with CC class of chemokines are called CCR while those interacting with CXC chemokines are termed as CXCR. In their inactive state chemokine receptors are coupled with heterotrimeric complex of G-proteins (Guanine binding proteins). G-proteins are further bound to a guanosine diphosphate (GDP) molecule [376]. The binding of the chemokine ligand predominantly happens at the N-terminal face with extracellular regions (loops and parts of helices) forming the binding pocket. Major portion of the ECD1 is responsible for ligand recognition and specificity. Chemokine binding illicit a conformational change across extracellular to intracellular faces of the 7TM leading to the exchange of GTP for GDP molecule. This exchange dissociates the heterotrimeric complex of G-proteins into Gα and dimeric Gβɣ [376]. Activated G-proteins having a GTP molecule attached, triggers a series of regulatory pathways using secondary messengers for downstream propagation. Most used secondary messengers are phosphatidylinositol 4,5-bisphosphate (PIP_2), phosphatidylinositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). A summarising description of chemokine signaling response is depicted in Figure 5.5. The allosteric motions set into place by chemokine binding leads to outward tilting of the intercellular face of the chemokine receptor [377]. The motif that is responsible for coupling G-proteins is located in the ICD2 [378]. It is represented by a conserved amino acid sequence of aspartate, arginine, tyrosine followed by slightly less conserved sequence of leucine, alanine, isoleucine and valine- 'DRYLAIV'. The sequence DRY has been observed to be very crucial for the activation of the G-proteins, especially the Arg residue that interacts directly with the G_α protein in active state [379]. In the inactive state, Arg forms an ionic lock with the Asp [380]. Upon activation, the pKa of the arginine changes consequently leading to the disruption of the ionic lock. The loss of the ionic interaction is compensated by new interactions with the well conserved Tyr residue of the following intracellular loop- ICD3 and Gα subunit of the G-proteins [381]. Therefore, the Arg works as a molecular switch and DRY motif is crucial for G-protein mediated signaling upon activation.

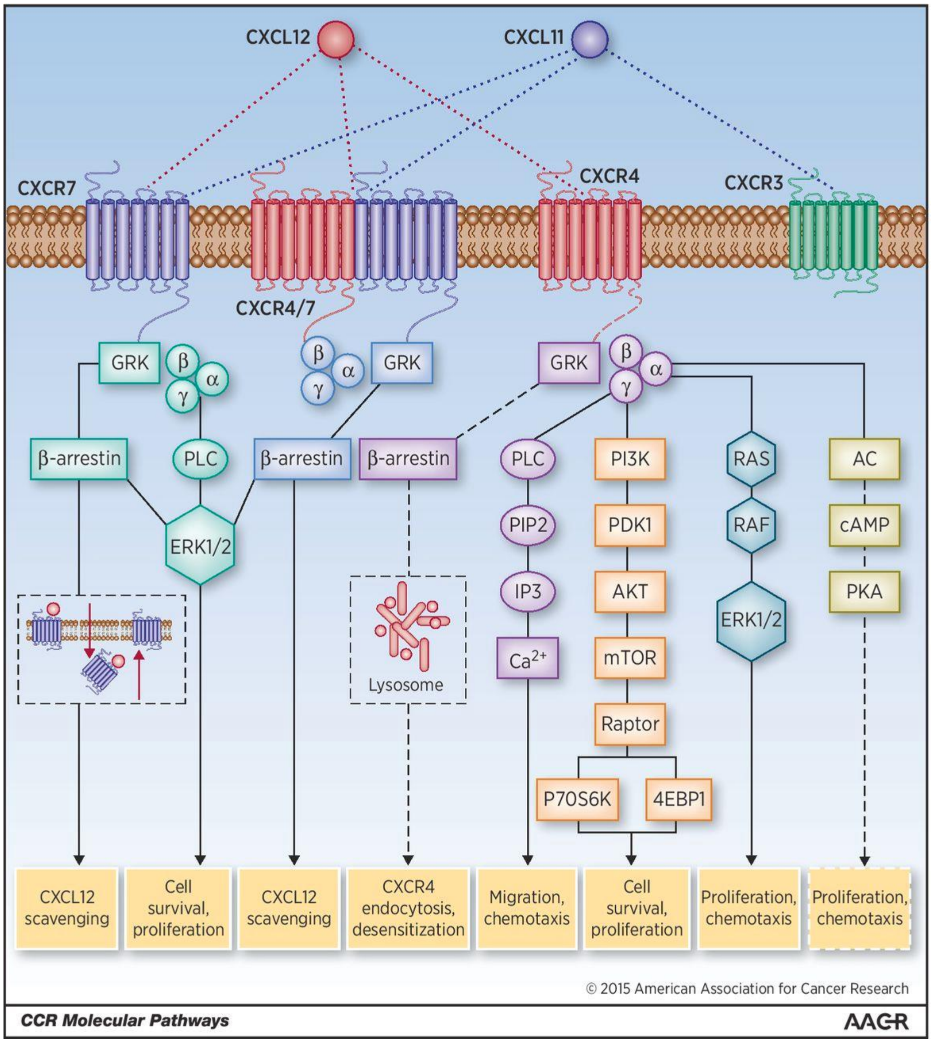


Figure 5.5 ***Chemokine system signalling.*** *The cartoon picture shows the various pathways that can be triggered by chemokine binding to a chemokine receptor. Shown are the chemokine receptors, CXCR3, CXCR4, a dimer of CXCR4 and CXCR7 (ACKR3), and CXCR7 (ACKR3). All these receptors can be activated by either CXCL12 or CXCL11. The signalling of ACKR3 shows no Gαi signalling as it has variations in the DRY motif.*

The Atypical Chemokine Receptor (ACKR) contain variations in this DRYLAIV motif and therefore cannot transduce signal after binding to a chemokine. Some ACKRs for eg, D6 (ACKR2), CXCR7 (ACKR3), (GPR35) contains the motif with some variation and are able to transduce signal by independent pathways [382,383]. The DRY motif is completely absent in ACKR1 and therefore it cannot couple with G-proteins and thus no signal transduction is observed, see Fig 5.4B. Besides, ACKR1 is also the only member in chemokine receptor family that binds non-specifically to chemokines. It binds to inflammatory chemokines of both types: CXCL as well as CCL. ACKR1 have been reported to interact with: CXCL1, CXCL8, CCL2 and CCL5 in erythrocytes. This makes ACKR1 behavior to be an 'atypical' one amongst the subfamily of atypical chemokine receptors. Among chemokine receptors, surprisingly less information is available about ACKR1 while it is the oldest known chemokine receptor [384]. The physiological function of ACKR1 in erythrocytes is yet unclear. As it binds to 20 chemokines of different types, it supposedly functions as a scavenger and regulate the inflammatory pathways [385][386]. Due to lack of an active response to chemokine binding, ACKR1 is also termed as a silent or decoy receptor. In non-erythrocytic cells like those in venular endothelium, cerebellar neurons and Purkinje fibers, ACKR1 is expressed by the allele Fy^a [373,387]. In endothelial cells, it mediates chemokine transcytosis. Wherein, ACKR1 internalizes the chemokines and migrate them from luminal to extravascular space, to induce leukocytes migration and thus regulate inflammatory response [388]. While the physiological function of ACKR1 in erythrocytes is still under consideration, it plays a critical part in entry of *Plasmodium vivax* micronemes [389]. It has been shown that DARC dimerizes during its interaction with *Plasmodium vivax* proteins. This makes ACKR1 a.k.a DARC, a protein of interest for our case study.

5.1.3.3 *Structure of ACKR1*

A decoy receptor like ACKR1 that is also the point of entry for *Plasmodium vivax* in Humans, is indeed an intriguing case of host-pathogen biology. Thus, enriching literature is available about the cell biology of ACKR1 and *Plasmodium vivax* DARC Binding Protein (*Pv*DBP). Numerous studies also identifies critical residues and motifs involved in both the proteins [390]. However, scarce information is available about the molecular structure of ACKR1 while structure of *P.knowlesi* DBP was published in 2005 by Singh *et al,* highlighting the crucial insights into DARC - DBP interactions [390]. The lack of structural information can be attributed to the challenge of crystallizing membrane bound proteins like GPCR. Currently, in PDB, 115 crystal structure of GPCR proteins are available [391], of which only 20 have a resolution lesser than 2.5 Å. This is indeed a small number for popular drug targets like GPCR. Availability of only 115 of ~800 GPCR proteins also estimates the challenges of experimental structure determination for membrane bound GPCRs. Chemokine receptors are Class A GPCR and have only 8 structures available in PDB with only 1 qualifying the resolution and r-free value thresholds.

5.1.4 *Homology modeling of DARC*

In the absence of an experimentally determined structure of DARC, it is logical to perform homology modelling to study the structural aspects of DARC. It is mentionworthy that our group have generated a structural model for DARC in 2005 and is the only one since then. It has a monomeric assembly modelled on a template from a very distant relative (sequence identity 12%) belonging to Rhodopsin family (Bovine Rhodopsin, PBDid: 1F88:A) [241]. With technological advancement in molecular biophysics, there is much more information available about chemokine receptors than in 2005 [377]. Therefore, we decided to remodel DARC using comparative modeling coupled with knowledge based restraints. In the last decade, evidence of oligomerization of chemokine receptors have gained enormous support. It has been reported that chemokine receptors often exist as homo- or heterodimers as well as oligomers with members outside chemokine receptor family [377]. Chemokine receptors have conserved 7-TM helices and a variable length N-terminal domain, ECD1, which is mostly disordered [392]. The disorder in ECD1 accounts for the functional diversity along with promiscuous binding network of chemokine ligands with chemokine receptors [377]. However, failure of drug candidates due to non-specificity towards target chemokine receptor leads to the notion of redundancy in chemokine receptors. In the last decade, these notions have been critically challenged by reports of oligomerization in chemokine receptors that diversifies the functional spectrum of the family [377]. There are profoundly three types of oligomeric structures in chemokine receptors: CC oligomers, CXC oligomers, and heteromers formed predominantly by the members of either CCRs or CXCRs and with other TM receptors [377]. However, an XCR and two CX3CR also exist. The major difference among chemokine receptors is among the sequential difference between cysteines and a tyrosine residue that can undergo sulfonation [393]. The dimerization of chemokine receptors is believed to be influenced by the class of the chemokines involved. Since, DARC binds to both classes of chemokine ligands with similar affinities, it poses a challenge to identify a correct template for comparative modelling. Consequently, three questions emerge: 1) Does DARC exist as a monomer or an oligomer? 2) If as oligomer, then is it a heteromer or homomer? 3) Which of the chemokine receptor(s) can be effectively used as a template(s) structure? Fortunately, in 2010 first structure of chemokine receptor was solved and released as a series of 4 structures with different antagonist ligands [394–397]. All the structures were crystallized as homodimers thus diminishing the alleged role of crystal packing. This was followed by a series of NMR and mutagenesis studies which assesses most CXCRs and DARC as a dimer [377]. Chakera and collegues in 2008 [398], had shown that DARC exists as a heteromer in cells proposing a CCR5/DARC complex. While these evidence were supporting DARC to be modelled as a dimer rather than the previous monomeric model. The question on homo- or heteromeric state was still unanswered because the heterodimeric studies were based on case specific analysis. DARC dimerizes with CCR5 to restrict the conformational changes in CCR5 that favors chemotaxis, thus acting as a trans-inhibitory partner [398]. DARC's physiological state on erythrocytes as well as during *Plasmodium vivax* contact was still not clear.

In 2014 Batchelor *et al*. [389], crystallized the DARC-PvDBL interaction and showed the assembly to be a heterotetrameric and heterotrimeric. In both the structures, homodimer of *Pv*DBL was shown interacting with either monomer (2:1) or dimer (2:2) of DARC based on which it is termed as trimeric or tetrameric. ITC (Isothermal Titration Calorimetry) results postulate the trimeric assembly as an intermediate structure in a stepwise binding process. Thus proposing a hypothesis that *PvDBL* homodimerizes under the effect of ACKR1 as depicted in Figure 5.6A. The structure provides crucial details about the interaction of ECD1 (DARC's N-terminus) and *Pv*DBL but do not have coordinates for the transmembrane structure or other interface and non-interface domains. The 2.6 Angstrom structure shows the dimerization of the regions of two distinct *Pv*DBL namely *Pv*DBL-RII (cysteine rich region II) and dimerization of ECD1 of DARC. Since the interaction takes place at the central region of DARC's ectodomain (ECD1) therefore, only residues 19-30 have sufficient electron density, Figure 5.6B. This also validates that the N-terminal ECD1 is indeed an IDR (Intrinsically Disordered Region) that acquires structure upon dimerization and interaction with *Plasmodium vivax*. The work of Batchelor *et al,* 2014 is highly pivotal in finding answers to the questions about homology modelling of DARC. Based on the findings above, DARC was decided to be modelled as a homodimer.

The selection of an effective template structure still remained at large. To decide whether to use a CXCR or CCR structure or both as a template, a crude phylogenetic approach was adopted. An understanding of the phylogenetic placement of DARC with other human chemokine receptors should provide sophisticated support in selection of a structural template. Therefore, a phylogenetic tree was plotted using the sequences available for 21 chemokine receptors [399,400]. Based on the tree topology and branch lengths, it is observed that DARC is highly distant from the rest of the clades while CXCR4 is suggested as a potent structural template for modelling DARC.

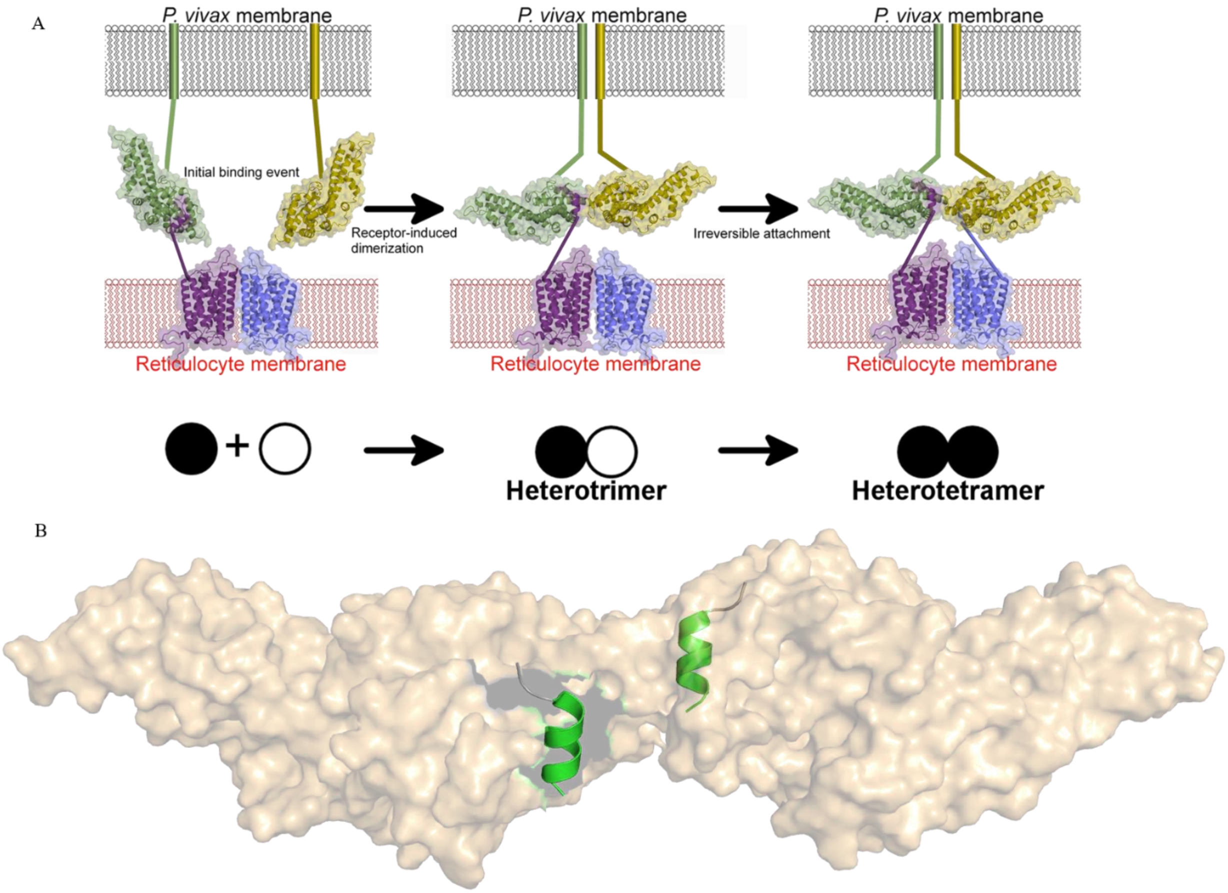


Figure 5.6 *Plasmodium vivax and ACKR1 interaction.* *A) shows the pathway followed by Plasmodium vivax DARC binding ligand (PvDBL) to bind with ACKR1 expressed on the reticulocytes of Duffy positive individuals. A monomer of PvDBL interacts with homodimer of ACKR1 that causes the dimerization of the PvDBL. Once dimerized, the N-terminal residues 19-30 of the ECD1 of ACKR1 docks irreversibly into binding pockets of PvDBL dimer. This is termed as a hetero-tetramer assembly (PBDID: 4NUV). B) shows the zoomed in version of the heterotetramer interaction. The otherwise disordered residues 19-30 of ACKR1 forms a well defined helix while rest of the ECD1 does not have definitive electron density.*

$^{+}$Image adapted from [389]

5.2. Methods

5.2.1 Structural template selection

5.2.1.1 *Template search*

The sequence of the ACKR1 or DARC was extracted from UniProtKB/Swiss-Prot database (accession number: Q16570). The sequence corresponding to the isoform 2 was selected as it has been annotated as the physiological form. The sequence so obtained was used as a query in blastp v2.6.0+ [401] as well as pHMMER v3.1b2 [402] to search against PDB database. For blast, amino acid substitution matrix BLOSUM62 [403] was used along with *word_size*: 6, *window_size*: 40. The composition based scoring adjustments (*comp_based_stats*) were used with an *e value* threshold of 1e^{-5}. pHMMER was run with default values using BLOSUM62 substitution matrix. e-value restrictions for domains, -*domE* and -*incdomE,* were used to focus the search on the 7TM domains and thus remove false positives.

The blastp search extracted only two hits, pertaining to ECD1 of ACKR1 in PDB structures: ids 4NUV and 4NUU. These PBD represent the heterotetrameric and heterotrimeric assembly of *Pv*DBL and DARC, respectively [389]. pHMMER also has these two as top hits and since -*domE* option was used; other TM proteins were also hit. However, a conclusive result was not obtained. Although, the matched subjects have >90% identity but they had critically less query coverage (< 8%). This noise in the results was created by the relatively long ECD1 of DARC.

Therefore, the query sequence was re-submitted after clipping the first 50 residues. Using the same parameters as before, with blastp and pHMMER; the query fetched significant number of hits containing chemokine receptors as well as other GPCR structures too. It is noteworthy, that only 50 residues instead of 60 residues of ECD1 were clipped in order to maintain the effect of ECD1 in the structure. The results were scanned and resubmitted to PSI-blast (after blastp) and JACKHMMER (after pHMMER). The iterations were stopped after third run (including the first run) with the final set containing 15 pdbs. An attempt to enrich the potential template dataset was made using FATCAT [345] at default configuration. The structure search method did not find any new hits that could be added to the 15 hits found by the sequence search methods.

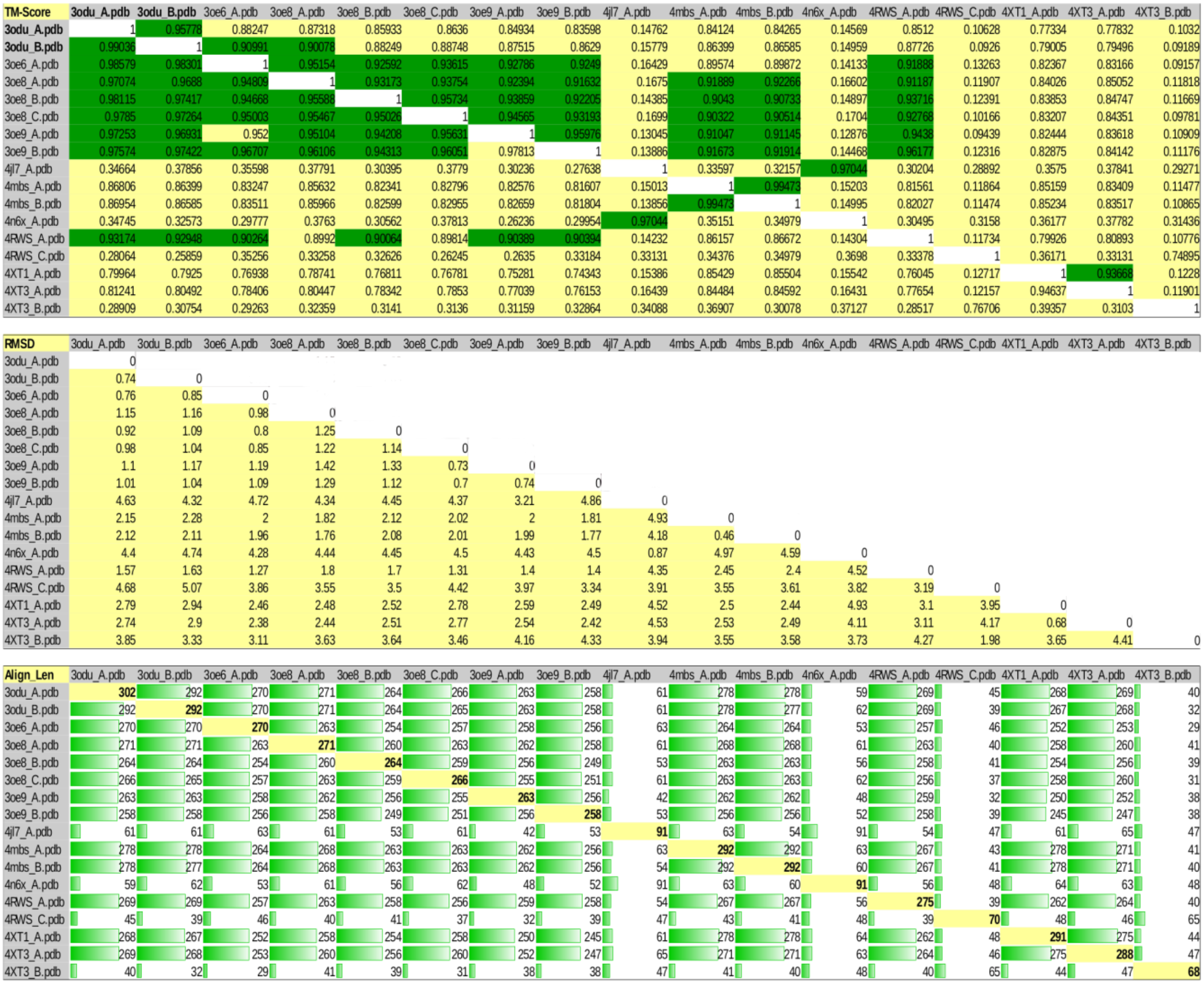
5.2.1.2 *Template selection*

Of the selected 15 PBDids, 5 were removed given the missing residues in their structures (*based on REMARK 465 of PDB file*), higher R-free value and lower resolution (> 3Å). The final set of potential templates contained 10 protein structures representing CXCR4, CCR5, CXCR2 and a viral GPCR protein. The PBDids for these structures are: CXCR4: {3ODU, 3OE6, 3OE8, 3OE9, 4RWS}, CXCR2: {4JL7, 4N6X}, CCR5 {4MBS}, vGPCR: {4XT1, 4XT3}. To verify their structural integrity, a conformational space analysis was performed using a three layered structural analysis, which was called a Three Tier Method (TTM). The TTM comprises of three simple metrics namely, TM-score [404], RMSD (root mean square deviation) and alignment coverage.

All the structures were first isolated into individual chains and heteroatoms were removed. The boundaries of TM-helices of each chain were identified and were pair-fitted using ProFit Version 3.1 [405]. This step validated the structural integrity of the GPCRs. After this preliminary analysis, TM-align [404] was used to perform all v/s all pairwise alignments of individual chains, as TM-align is length independent. The statistics extracted from the TM-align were used in the TTM. Table 5.1, shows the output of the TTM with a comparison of all vs all chains. The use of PDBs 3OE6, 3OE8, 3OE9 in the analysis acted as a positive control for the TTM. The first section analyse the TM-score, which depends on the reference structure selected during alignment. Higher the TM-score, more related are the structures. Second section establishes structural relatedness using RMSD: lower the value, closer are the structures. The second section acts as a validation for the first section. In the third section, alignment lengths or coverage was used to assess the structural comparison.

These steps along with the preliminary analysis done using ProFit proved crucial in identification and exclusion of the outliers like, 4JL7, 4N6X, and parts of 4RWS and 4XT3. The TTM helped in narrowing the templates from 10 to 2. The selected templates were 3ODU (CXCR4) and 4MBS (CCR5). 3ODU exists as a dimer in the asymmetric unit while 4MBS contains a monomer in its asymmetric unit (*based on REMARK 290 and 350 of their PDB files*).

Table 5.1 ***Three Tier Method for filtering template structures.*** *All vs all structural comparison of tentative template structures is done using TM-align and ProFit. Tier 1: indicates the TM-scores of pairwise structural alignments. A TM-score above 0.8 indicates same family and a TM-score > 0.9 indicates highly similar structures (highlighted in green). Tier 2: indicates the RMSD in (Å). Half matrix is shown as the RMSD does not change with change of reference structure. The low RMSD values of the range 0.65 – 4.97 are shown. 3ODU and 4RWS are potential templates given their overall low RMSD values. Tier 3: is the alignment coverage during the pairwise structural alignment. It helps in justifying the high structural deviation values in Tier 1 and 2.*



TM-Score	**3odu_A.pdb**	**3odu_B.pdb**	3oe6_A.pdb	3oe8_A.pdb	3oe8_B.pdb	3oe8_C.pdb	3oe9_A.pdb	3oe9_B.pdb	4jl7_A.pdb	4mbs_A.pdb	4mbs_B.pdb	4n6x_A.pdb	4RWS_A.pdb	4RWS_C.pdb	4XT1_A.pdb	4XT3_A.pdb	4XT3_B.pdb
3odu_A.pdb	1	0.95778	0.88247	0.87318	0.85933	0.8636	0.84934	0.83598	0.14762	0.84124	0.84265	0.14569	0.8512	0.10628	0.77334	0.77832	0.1032
3odu_B.pdb	0.99036	1	0.90991	0.90078	0.88249	0.88748	0.87515	0.8629	0.15779	0.86399	0.86585	0.14959	0.87726	0.0926	0.79005	0.79496	0.09189
3oe6_A.pdb	0.98579	0.98301	1	0.95154	0.92592	0.93615	0.92786	0.9249	0.16429	0.89574	0.89872	0.14133	0.91888	0.13263	0.82367	0.83166	0.09157
3oe8_A.pdb	0.97074	0.9688	0.94809	1	0.93173	0.93754	0.92394	0.91632	0.1675	0.91889	0.92266	0.16602	0.91187	0.11907	0.84026	0.85052	0.11818
3oe8_B.pdb	0.98115	0.97417	0.94668	0.95588	1	0.95734	0.93859	0.92205	0.14385	0.9043	0.90733	0.14897	0.93716	0.12391	0.83853	0.84747	0.11669
3oe8_C.pdb	0.9785	0.97264	0.95003	0.95467	0.95026	1	0.94565	0.93193	0.1699	0.90322	0.90514	0.1704	0.92768	0.10166	0.83207	0.84351	0.09781
3oe9_A.pdb	0.97253	0.96931	0.952	0.95104	0.94208	0.95631	1	0.95976	0.13045	0.91047	0.91145	0.12876	0.9438	0.09439	0.82444	0.83618	0.10909
3oe9_B.pdb	0.97574	0.97422	0.96707	0.96106	0.94313	0.96051	0.97813	1	0.13886	0.91673	0.91914	0.14468	0.96177	0.12316	0.82875	0.84142	0.11176
4jl7_A.pdb	0.34664	0.37856	0.35598	0.37791	0.30395	0.3779	0.30236	0.27638	1	0.33597	0.32157	0.97044	0.30204	0.28892	0.3575	0.37841	0.29271
4mbs_A.pdb	0.86806	0.86399	0.83247	0.85632	0.82341	0.82796	0.82576	0.81607	0.15013	1	0.99473	0.15203	0.81561	0.11864	0.85159	0.83409	0.11477
4mbs_B.pdb	0.86954	0.86585	0.83511	0.85966	0.82599	0.82955	0.82659	0.81804	0.13856	0.99473	1	0.14995	0.82027	0.11474	0.85234	0.83517	0.10865
4n6x_A.pdb	0.34745	0.32573	0.29777	0.3763	0.30562	0.37813	0.26236	0.29954	0.97044	0.35151	0.34979	1	0.30495	0.3158	0.36177	0.37782	0.31436
4RWS_A.pdb	0.93174	0.92948	0.90264	0.8992	0.90064	0.89814	0.90389	0.90394	0.14232	0.86157	0.86672	0.14304	1	0.11734	0.79926	0.80893	0.10776
4RWS_C.pdb	0.28064	0.25859	0.35256	0.33258	0.32626	0.26245	0.2635	0.33184	0.33131	0.34376	0.34979	0.3698	0.33378	1	0.36171	0.33131	0.74895
4XT1_A.pdb	0.79964	0.7925	0.76938	0.78741	0.76811	0.76781	0.75281	0.74343	0.15386	0.85429	0.85504	0.15542	0.76045	0.12717	1	0.93668	0.1228
4XT3_A.pdb	0.81241	0.80492	0.78406	0.80447	0.78342	0.7853	0.77039	0.76153	0.16439	0.84484	0.84592	0.16431	0.77654	0.12157	0.94637	1	0.11901
4XT3_B.pdb	0.28909	0.30754	0.29263	0.32359	0.3141	0.3136	0.31159	0.32864	0.34088	0.36907	0.30078	0.37127	0.28517	0.76706	0.39357	0.3103	1

RMSD	3odu_A.pdb	3odu_B.pdb	3oe6_A.pdb	3oe8_A.pdb	3oe8_B.pdb	3oe8_C.pdb	3oe9_A.pdb	3oe9_B.pdb	4jl7_A.pdb	4mbs_A.pdb	4mbs_B.pdb	4n6x_A.pdb	4RWS_A.pdb	4RWS_C.pdb	4XT1_A.pdb	4XT3_A.pdb	4XT3_B.pdb
3odu_A.pdb	0																
3odu_B.pdb	0.74	0															
3oe6_A.pdb	0.76	0.85	0														
3oe8_A.pdb	1.15	1.16	0.98	0													
3oe8_B.pdb	0.92	1.09	0.8	1.25	0												
3oe8_C.pdb	0.98	1.04	0.85	1.22	1.14	0											
3oe9_A.pdb	1.1	1.17	1.19	1.42	1.33	0.73	0										
3oe9_B.pdb	1.01	1.04	1.09	1.29	1.12	0.7	0.74	0									
4jl7_A.pdb	4.63	4.32	4.72	4.34	4.45	4.37	3.21	4.86	0								
4mbs_A.pdb	2.15	2.28	2	1.82	2.12	2.02	2	1.81	4.93	0							
4mbs_B.pdb	2.12	2.11	1.96	1.76	2.08	2.01	1.99	1.77	4.18	0.46	0						
4n6x_A.pdb	4.4	4.74	4.28	4.44	4.45	4.5	4.43	4.5	0.87	4.97	4.59	0					
4RWS_A.pdb	1.57	1.63	1.27	1.8	1.7	1.31	1.4	1.4	4.35	2.45	2.4	4.52	0				
4RWS_C.pdb	4.68	5.07	3.86	3.55	3.5	4.42	3.97	3.34	3.91	3.55	3.61	3.82	3.19	0			
4XT1_A.pdb	2.79	2.94	2.46	2.48	2.52	2.78	2.59	2.49	4.52	2.5	2.44	4.93	3.1	3.95	0		
4XT3_A.pdb	2.74	2.9	2.38	2.44	2.51	2.77	2.54	2.42	4.53	2.53	2.49	4.11	3.11	4.17	0.68	0	
4XT3_B.pdb	3.85	3.33	3.11	3.63	3.64	3.46	4.16	4.33	3.94	3.55	3.58	3.73	4.27	1.98	3.65	4.41	0

Align_Len	3odu_A.pdb	3odu_B.pdb	3oe6_A.pdb	3oe8_A.pdb	3oe8_B.pdb	3oe8_C.pdb	3oe9_A.pdb	3oe9_B.pdb	4jl7_A.pdb	4mbs_A.pdb	4mbs_B.pdb	4n6x_A.pdb	4RWS_A.pdb	4RWS_C.pdb	4XT1_A.pdb	4XT3_A.pdb	4XT3_B.pdb
3odu_A.pdb	**302**	292	270	271	264	266	263	258	61	278	278	59	269	45	268	269	40
3odu_B.pdb	292	**292**	270	271	264	265	263	258	61	278	277	62	269	39	267	268	32
3oe6_A.pdb	270	270	**270**	263	254	257	258	256	63	264	264	53	257	46	252	253	29
3oe8_A.pdb	271	271	263	**271**	260	263	262	258	61	268	268	61	263	40	258	260	41
3oe8_B.pdb	264	264	254	260	**264**	259	256	249	53	263	263	56	258	41	254	256	39
3oe8_C.pdb	266	265	257	263	259	**266**	255	251	61	263	263	62	256	37	258	260	31
3oe9_A.pdb	263	263	258	262	256	255	**263**	256	42	262	262	48	259	32	250	252	38
3oe9_B.pdb	258	258	256	258	249	251	256	**258**	53	256	256	52	258	39	245	247	38
4jl7_A.pdb	61	61	63	61	53	61	42	53	**91**	63	54	91	54	47	61	65	47
4mbs_A.pdb	278	278	264	268	263	263	262	256	63	**292**	292	63	267	43	278	271	41
4mbs_B.pdb	278	277	264	268	263	263	262	256	54	292	**292**	60	267	41	278	271	40
4n6x_A.pdb	59	62	53	61	56	62	48	52	91	63	60	**91**	56	48	64	63	48
4RWS_A.pdb	269	269	257	263	258	256	259	258	54	267	267	56	**275**	39	262	264	40
4RWS_C.pdb	45	39	46	40	41	37	32	39	47	43	41	48	39	**70**	48	46	65
4XT1_A.pdb	268	267	252	258	254	258	250	245	61	278	278	64	262	48	**291**	275	44
4XT3_A.pdb	269	268	253	260	256	260	252	247	65	271	271	63	264	46	275	**288**	47
4XT3_B.pdb	40	32	29	41	39	31	38	38	47	41	40	48	40	65	44	47	**68**

5.2.1.3 *Phylogenetic Analysis for template selection*

To select between a CXCR and a CCR template to model DARC was a challenge, given that DARC binds to both classes of chemokine ligands. Therefore, phylogenetic information of chemokine receptors was used in the selection process. Sequences of 21 chemokine receptors out of 24 known Human chemokine receptors were selected. 3 receptors, namely CX3CR1, XCR1, and ACKR6 were excluded as they do not bind to the canonical chemokine ligands [399,400]. Multiple alignment of these 21 sequences was performed using MAFFT v7.2 [406], with default parameters. The multiple sequence alignment (MSA), in Stockholm format, was annotated with the helical boundary information of 7TM helices. Sequence editor, JalView v16 [407] was used to refine the MSA based on the conservation of cysteine residues and DRY motif.

5.2.1.4 *Tree generation and visualization*

The resulting MSA was submitted to IQ-Tree v1.4.2 [408]- an efficient tree reconstruction algorithm based on maximum likelihood. A phylogram was generated using JTT exchange rate matrix having free rate heterogeneity and empirical frequencies from the data: (JTT+F+R10). The tree was bootstrapped at 1000 exchanges. The resulting phylogram was visualized using iterative Tree of Life (iTOL) [409]. For details on the tree building parameters, see section 6.2.4.

5.2.2 *Structural Modelling*

5.2.2.1 *Sequence Alignment*

The sequence of the template structure and ACKR1 (without the ECD1) was aligned using Promals3D [410]. In addition to the global alignment of sequences, Promals3D also takes into account the secondary structure elements (SSE) of template and the predicted sse using PSIpred [49]. Therefore, it helped in easy visualization of the alignment in the TM regions. However, the global pairwise alignment was also annotated with the assigned SSE of template using DSSP. It was made sure that there were no critical in/dels in the helical regions. The final alignment was saved in PIR format in the alignment.ali file. The alignment had the complete sequence of DARC minus the 49 residues at N-terminal, namely ECD1.

5.2.2.2 *Structural Modelling and assessment*

Modeller v9.16 [75] was used to generate structural models using the template and the alignment file. Modeller script, model-default.py was customized by adding snippets to define spatial restraints and different model assessment scores. Two constraints were added; the transmembrane boundaries and two disulfide bonds. Three assessment scores were given to assess the best model among the 100 models generated. Apart from DOPE/molpdf [75,411], and GA341 (model reliability), normalized DOPE score (nDOPE) was also calculated in order to compare models from different templates. For instance, to compare model quality generated from a monomeric template and a dimeric template because DOPE works only with single chains.

Since, the query is a TM protein and assessment scores like DOPE/molpdf, GA341 are optimized for globular proteins, it was required to assess the models differently. Therefore, MAIDEN was used [412]. Model quality Assessment for Intramembrane Domains using an ENergy criterion- MAIDEN is a statistical potential optimized using structural information of membrane proteins from PDBTM (xml file) [413]. The energy potential of TM proteins is supplied as an 'intp' file with option '-*e*'. These potentials are also used to calculate an approximation of the free energy which is given as the raw potential in the output. MAIDEN uses sequence based decoys to calculate Z-scores, similar in principle to e-value calculations. An important feature of MAIDEN is that it uses globular potentials for calculations of the extra-membrane portions. Further, a conformational space for the models were sampled using TM-score of 101*101 pairwise alignment of all models against all models (100) and the template (1). This step helped in understanding the expanse of models from the template as well as among themselves.

A top twenty approach or "T20 test" was devised to select the best model. Top twenty models were selected from each of the scoring functions. An intersection of the 4 sets, each containing 20 models was calculated. The set can be represented as (MAIDEN ∩ (GA341 ∩ (nDOPE ∩ DOPE))), where the set names represent the 20 best models according to the scoring function. Since, only one of the scores is optimized for TM proteins, MAIDEN set was given preference over nDOPE, DOPE and GA341 sets. For instance, if a model appears lower in T20 of nDOPE and DOPE/molpdf but ranks as the best model according to MAIDEN score, that model will be selected. Thus a MAIDEN score is weighted more than the rest of the assessment scores. This information is then clubbed with the TM-score analysis and a best model is selected. In summary, the best model should be ranked amongst the top 20 models in each scoring scheme and should be a representative of the collective structural space of template and models.

5.2.3 Membrane Building

The structural model should be embedded in a lipid bilayer. This step is important for the analysis of the pathophysiological role of DARC. Therefore, the first step was to resolve the protonation states of asp, lys, glu and his residues at the physiological pH 7.2, which was done using proPka [414] and PDB2PQR server [415]. The dimeric model was then submitted to PPM server [414,416] to estimate the extent of the lipid bilayer. PPM server was used after specifying that the N-term of the model lies on the extracellular side. PPM server hosted online by OPM database that hosts the membrane orientations of 3426 TM proteins from PDB. The structural model was then submitted to the Membrane Builder of CHARMM-GUI [417]. The objective was to embed the structural model of DARC in a mimic reticulocyte membrane. The composition of the RBC membrane during reticulocyte stage as well as mature stage, was estimated by doing extensive literature survey. Following information was extracted about the RBC membrane:

a) *The physiological pH of the membrane ranges from 6.3 to 7.9 [418].*
b) *Cholesterol (CHL) seems equally distributed between the inner and outer halves [419].*
c) *Phospholipids are asymmetrically distributed between the leaves and is crucial for red cell physiology [420]. Changes in membrane lipids can affect the RBC shape by perturbing the balance in inner and outer leaflet lipids [421,422].*
d) *Fluidity of the bilayer is determined by molar ratio of cholesterol to phospholipids, degree of unsaturation of phospholipid acyl chains, and phosphatidylcholine to sphingomyelin ratio [423]. In reticulocyte membrane, since the phospholipid and cholesterol is synthesized from glycerol and acetate, their concentration are expected to be in nearly equal proportions [424].*
e) *While phosphatidylcholine (PC) forms highly fluid lipid regions, sphingomyelin (SM) induces rigidity [423].*

f) *PC and SM are located in the outer leaflet while; phosphatidylinositols (PI), phosphatidylethanolamine (PE), and phosphatidylserine (PS) occur mostly in the inner leaflet* [425].
g) *Most of all PC lipids are in the outer leaflet while most of all PS lipids are in the inner leaflet [422]*,[418].
h) *Redistribution of membrane phospholipids may trigger clotting cascade [426]*.

The concentration, type of lipids, and distribution ratio was calculated strictly to satisfy these findings from literature. Therefore, the selection of heterogeneous lipids in step 02 of Membrane builder was the bottleneck of the protocol. Out of 182 different lipids, POPC, POPE, POPS, POPI and PSM were chosen. Since all of these are derivatives of the same acyl chain, Palmitoyl Oleoyl- and therefore offer consistency in acyl unsaturation and chain lengths. Therefore, the effect of acyl chain unsaturation on membrane fluidity can be effectively normalized. Further, these lipids are added in definitive ratios. Cholesterol (CHL1) is added in equal amount at 25:25 (outer: inner) satisfying the ratio 1:1.

5.3. Results and Discussion

Our interest in studying the structure of ACKR1 (previously named DARC) is two fold. Primarily we wanted to understand the inherent dynamics of a protein structure in a multimeric assembly; a homodimer in this case. Secondly, DARC is rather an interesting protein that does not transduces signal, unlike other chemokine receptors and plays a crucial role in pathophysiology of *P.vivax* malaria. Absence of a robust structure for DARC makes it challenging to understand its structural biology as well as its interaction with the *Pv*DBP. Therefore, it becomes elementary to model the structure of DARC so as to achieve our objectives.

5.3.1 *Phylogenetics based selection of the template*

As explained in the methods section of this chapter, to enhance the template selection procedure, phylogenetic information is used. A phylogram for 21 Human chemokine receptors was generated as shown in Figure 5.7. DARC (mentioned as ACKR1 in the tree), has the longest distance from the root at 5.97 branch length units. Only one receptor comes closer to DARC with 5.20 units distance from the root, GPR35 or potentially CXCR8 [427]. However, GPR35 does not belong to the same clade as DARC. The distance of DARC from its nearest clade is 4.92 units. According to the tree topology, CCR10 seems closest to DARC, in the clade. Interestingly, the distance of CCR10 from DARC is 3.10 units which quantifies the observation that there is huge evolutionary gap between DARC and its closest neighbor. This observation also intensifies the curiosity in the evolution of DARC which is handled later in details (see chapter 6). Based on the branch lengths and tree topology, CXCR3 (3.97 units away from DARC), CXCR5 (3.70 units away) and CXCR4 (4.16 units away) are the next closest neighbor of DARC, in that particular order. Of these, only CXCR4 has a crystal structure available in PDB. Since, CCR10 is the closest neighbor of DARC and the crystal structure of CCR5 is known, the proximity between CCR10 and CCR5 was also tested. Unfortunately, CCR5 is 1.42 units and seven clades away from CCR10 while it has a distance of 5.52 units from DARC, see Figure 5.7. Therefore, CXCR4 clearly qualifies as the selected template for modelling a structure for DARC.

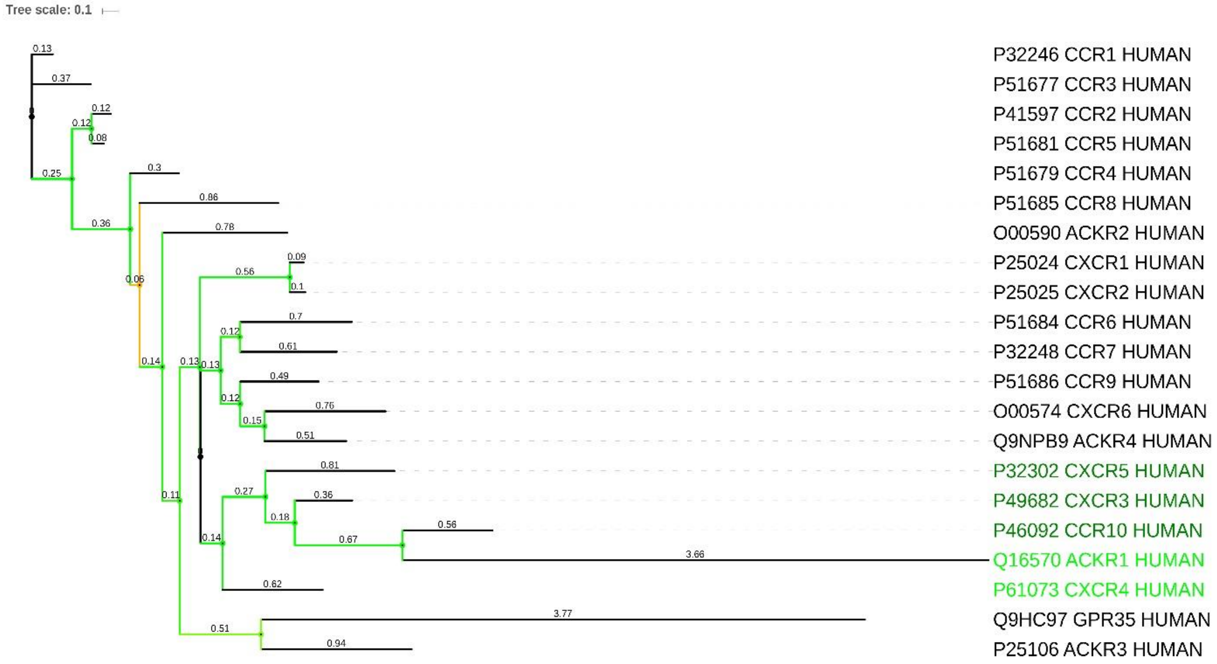


Figure 5.7 ***Phylogenetic placement of ACKR1.*** *The tree topology as generated by IQtree. The green colour of branches indicates high bootstrap values (of 1000). The branch lengths are indicated on the branches approximated to two decimal places. The green coloured labels indicate members belonging to the clade of ACKR1. The green asterisks mark the availability of structure in PDB. As observed, ACKR1 is highly distant from rest of the clade at a branch length of 3.66*

units. It's closest neighbor in the clade is CCR10 however, the structure for CCR10 does not exist. The neighbor also having a crystal structure is CXCR4 at a branch length of 4.16 units from ACKR1.

5.3.2 <u>*Structural model for Duffy Antigen / Chemokine Receptor*</u>

5.3.2.1 *T20 test for the selection of the best model*

Structural template search accompanied by a robust template selection protocol along with phylogenetic placement of DARC in Human chemokine receptors and using the gathered molecular biology information as anchors, the homodimeric structure is modelled as depicted in Figure 5.8. Of the 100 models generated, the model no. 85 was selected based on the intersection set of the 'T20 test' and proximity analysis with the template structure using RMSD and TM-score. The intersection set contained models numbered: m33, m59, and m85. m85 has the least energy in MAIDEN calculations and ranked 16th in the nDOPE set. However, it appears last in the GA341 set. m33 tops the nDOPE, DOPE and GA341 sets but ranks 9th in MAIDEN scores. m59 like m85 ranks 2nd in MAIDEN set but appears in the last quarter of nDOPE, DOPE and GA341 sets. Since, both m59 and m85 are ranked higher in the MAIDEN set they were weighted higher. Of these two, GA341 score becomes a 'decider', with m85 having a score of 0.15 in contrast to higher score of 0.21 for m59. GA341 is a score to estimate the accuracy of the model using percent sequence identity between individual chains of model and template. An ideal model would be scored 1.00 thus having 100% sequence identity with all the template chains. However, our approach is to find a divergent structure from the CXCR4 template and thus m85 is selected on this argument. The RMSD between m85 and m59 is 0.56Å while structural deviation from template is 2.14Å and 2.17Å for m85 and m59 respectively. The deviation from template structure is also quantified by average TM-score (mean of chain-wise scores) as 0.812 and 0.815 for m85 and m59 respectively.

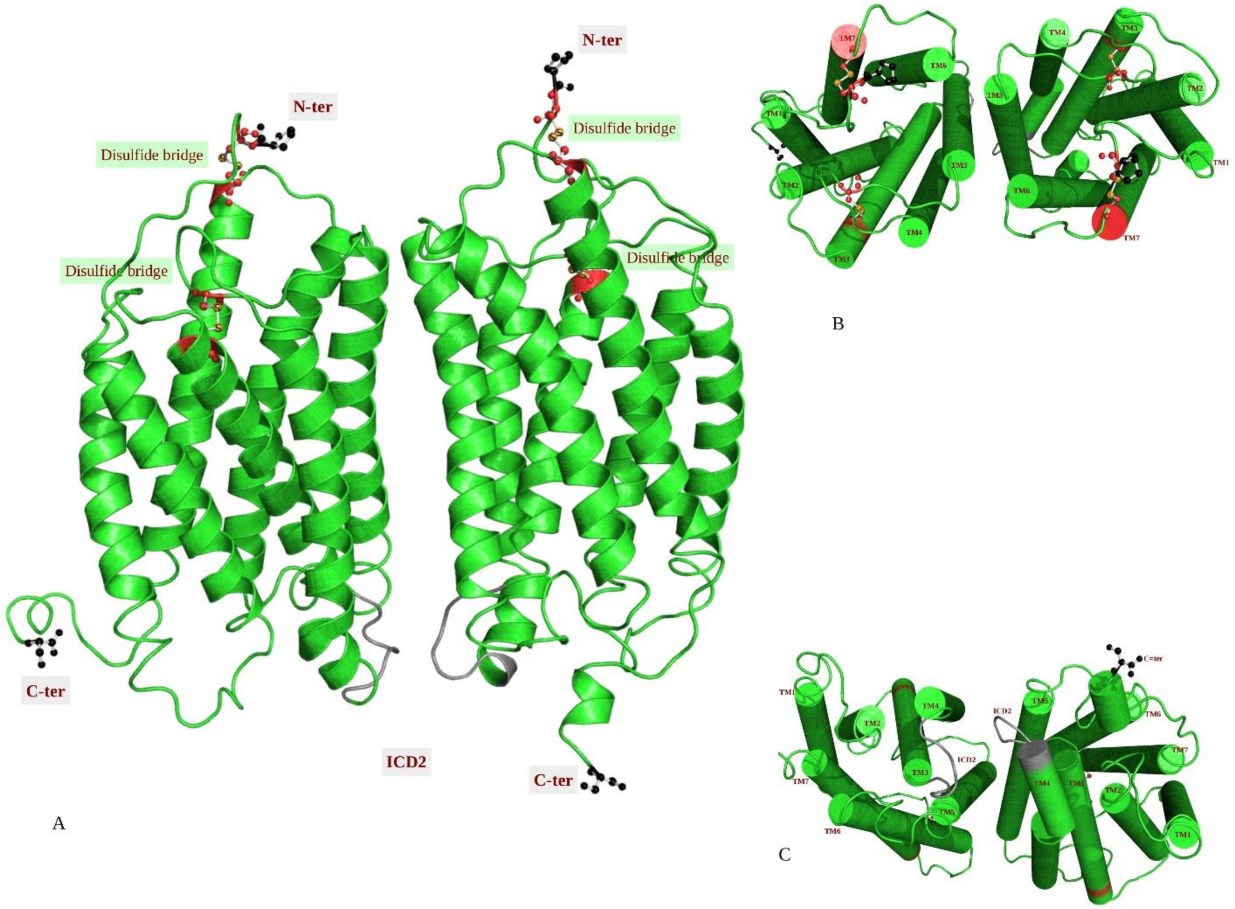


Figure 5.8 ***Structural model for ACKR1.*** *With the help of T20 approach for model selection, best model was selected and validated. A) The important cysteines are found to be conserved and forms the disulfide bridges (Cys in red and S-S bond is shown in orange). The terminals are shown in black, ball and stick models. ICD2 (loop that contains DARC) is shown in grey colour. B) shows the top view of the dimer with respective helices marked. The orientation symmetry of the helices is mirrored and the ECD1 seems to move towards each other, as expected. The top view also shows the open interface and binding pocket for the ligand. C) Bottom view of the dimer shows the inwards tilting and gives a better perspective on the orientation of the ICD2 (grey).*

5.3.2.2 *Structural orientation in the bilayer*

Modeller protocol does not take into account the effect of solvent on the structural model. The conformations may change with solvent properties, especially in the case of TM protein structure; wherein some region interacts in a hydrophobic environment while the rest resides in a hydrophilic environment. DARC is a TM protein that is expressed on the RBC membranes and therefore it is required to embed the obtained structural model in a lipid bilayer. However, before building the membrane, it is essential to mark the boundaries of the TM regions and estimate the orientation of the TM helices. Therefore, PPM web server was used [416]. The PPM protocol that uses the alignment of protein structure's z-axis with the normal to the bilayer and minimization of the transfer free energies of the embedded amino acid residues, provided a membrane orientation for the homodimeric model of DARC. The assembly have a ΔG for transfer energies as 107.9 kcal/mol. The thickness of the hydrophobic bilayer is estimated as 30.6 ± 0.8Å which matches with the average bilayer depth of TM proteins. Moreover, PPM server also reports a tilt in the TM helices at 4. ± 2°. The TM boundaries are also provided for the seven helices. These results provide some elementary understanding of the structure of the DARC homodimer.

5.3.3 Comparison with the old computational model of DARC

As it is mentioned in section 5.1.4 that our lab has generated a homology model of DARC in 2005 therefore it would be logical to compare the old model and the new. Although, there are robust reasons to believe that the two structural models are not fittingly comparable. In 2005, the protein data bank had only a very few crystal structure of 7TM fold and even rare would have been to find a crystal structure for Class A GPCR. This can be put into context by acknowledging that the first crystal structure of chemokine receptor was published in 2010. Therefore, the comparison of the two structural models is merely a conventional exercise.

The structural model from 2005 (for reference will be called, DARC$_{old}$ hereafter) is modelled based on a bovine Rhodopsin GPCR template. The bovine rhodopsin (PBDid 1F88:A) had a sequence identity of less than 20% in the TM regions. According to de Brevern et al, 2005, an ensemble of structural models were generated but only two were selected as they followed the spatial restraints [241]. The major difference is that both the models are monomeric. And as during the current study, a lot of time was invested in finding our the correct oligomeric state of DARC in erythrocytes. An explanation for the monomeric model could be that the Rhodopsin is a non-erythroid protein and therefore DARC$_{old}$ might represent the structural model of DARC from vascular lineage (Fy^{a}). DARC is known to express as monomer in epithelial cells. Therefore, in order to compare the current model with DARCold, one subunit of the dimer have to used. The RMSD between two subunits of DARC dimer is 0.12Å. Figure 5.9 shows the structural superposition between the old models and single subunit of the current dimeric model. Since, the DARCold was modelled as a whole structure therefore, both the models consist of ECD1 which causes a lot of noise in comparing the TM-scores and RMSD. Therefore, alignment was done using current model as the reference and ECD1 from $DARC_{old}$ was ignored. The RMSDs between the $DARC_{old}$ and current monomer is 6.25Å and 5.84Å with the ECD1. The TM-scores among the two $DARC_{old}$ models and current model are 0.73 and 0.76 depicting that the TM domains belong to the family. However, upon a quick observation of Figure 5.9, it can be seen that the TM helices of $DARC_{old}$ does not show a well defined outward tilting at their N-terminal face. Therefore, the major deviation in TM-region among models can be seen at the N-terminal face of TM-helices. Another striking difference among models is that the old models does not have the disulfide bridges among conserved cysteines. This can also explain the less defined to absent outward tilting of $DARC_{old}$. However, the TM-score among the old models is 0.82 and they have an RMSD of 2.63Å.

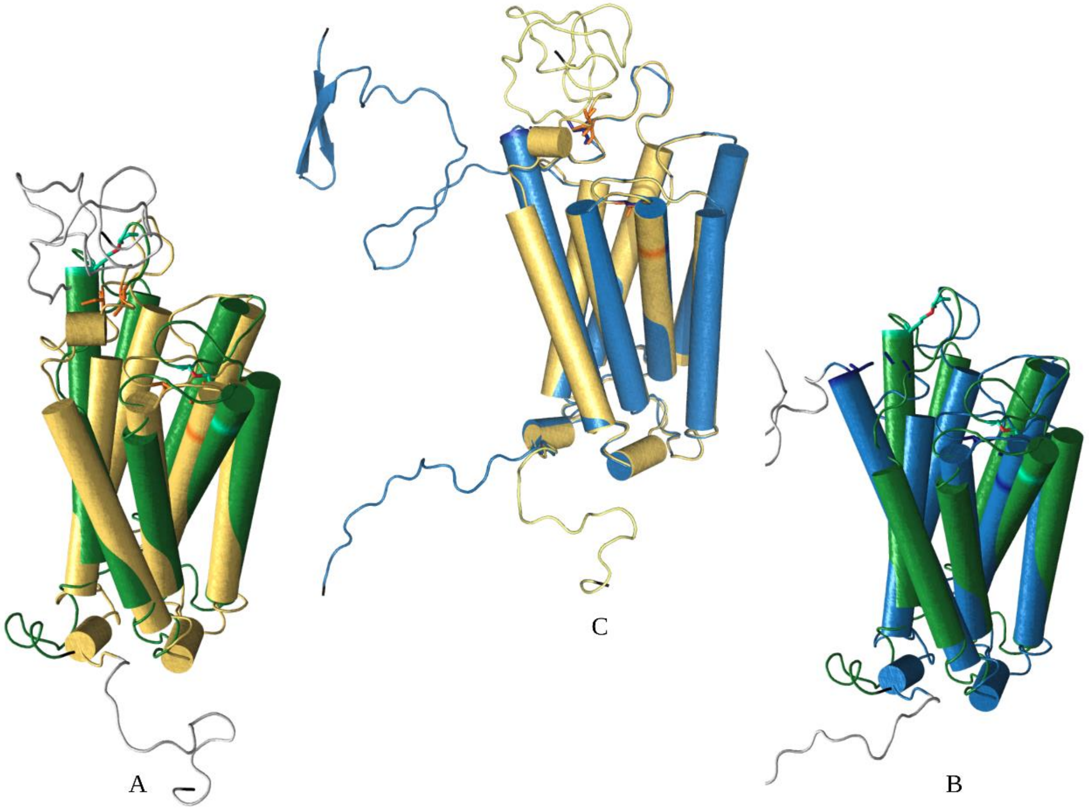


Figure 5.9 ***Structural comparison between old and new models of DARC***. *A) Superposed structural model of DARCold (yellow) and ACKR1monomeric model (green). The gray color of the ECD1 and ICD4 signifies that these were not included in the alignment using TM-align. B) Superposed structural models of ACKR1 monomer (green) and more extended of the DARCold models. The loops of ECD1 are not not shown here since they were not used in the alignment. These loops can be seen in C) forming an anti-parallel beta sheet (in blue). In C) the helices orientation can be appreciated. In A and B, the deviation of yellow and blue model from green*

model is mostly inward and therefore, they do not have a well defined outward tilting. The cysteine residues are shown in Red.

5.3.3 <u>*Generation of a mimic RBC bilayer*</u>

While the effect of a lipid membrane on the structure can be estimated using a lipid membrane embedded structure of DARC using CHARMM-GUI (see methods), the bottleneck is the choice of lipids in the membrane. Since, the second objective of the project is to apprehend the DARC - *Pv*DBL interactions, the type of membrane is an important factor. The membrane builder from CHARMM-GUI is a sophisticated software but like every other computational tool, the biological significance of its result depends on the given inputs. Therefore, it was important to generate an *in-silico* mimic of a physiological RBC membrane. However, it becomes really complicated because the reticulocytes (young RBC) are slightly different from erythrocytes (mature RBC) [424]. Unfortunately, there are no standardized concentration of lipids available for reticulocytes but some RBC centered research articles mention lipid distribution among cytosolic and extracellular halves of the lipid membrane [424,426,428]. Such articles were mined (details in Methods) and an estimate of the lipid concentrations and distributions amongst the two leaflets of reticulocyte membrane was proposed.

After estimating the orientation of the structural model and estimating the reticulocyte membrane composition, it was subjected to the membrane building protocol at CHARMM-GUI web-server. The structure was inspected manually for the TM boundaries, helix orientations and disulfide bridges, at each step of the six-step long protocol. After the first step, the oriented structure showed a beta strand in ECD3 between TM helix 4 and 5 while the input structure has a β-hairpin loop. It was found that the appearance of β-strand is due to the difference in the visualization platforms as the coordinates of both the files were same. In step 2, after verifying the cross-sectional areas, heterogeneous lipids were added with the system having a water cross-section of 10 Å (height) on both sides of the bilayer. The system size was calculated based on the number of lipids added to the system. A total of 121 lipids were added to the upper leaflet while 110 lipids were added to the lower leaflet. This is to maintain the asymmetry of the reticulocyte membrane. The final structural model was assembled with the generated membrane using replacement method (Fig 5.10). The resulting dimeric structure is minimized for removal of bad contacts, especially between cholesterol ring and the lipid tails, using PME and SHAKE algorithm in Gromacs [429].

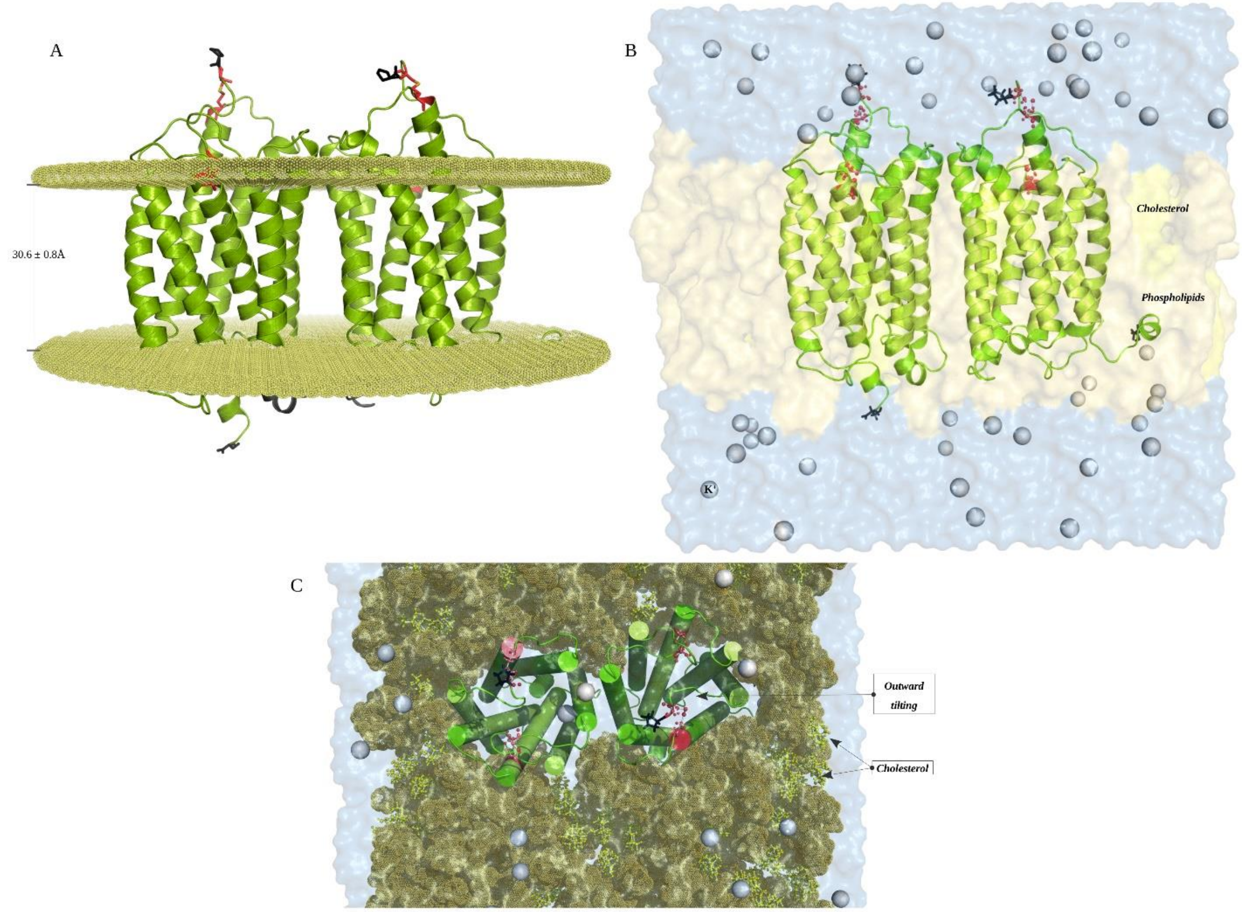


Figure 5.10 ***Membrane embedded ACKR1 dimer.*** *A) shows the orientation of the helices in a dummy membrane boundary (shown as cluster of dots). The width of the TM region is also shown. Red colour highlights the disulfide bridges and grey colour on the intracellular face depicts ICD2. B) The complete model system with dimer embedded in an RBC mimic membrane (shades of yellow), sandwiched by water layers of thickness 10Å containing 22 neutralizing K+ ions. Cholesterol is coloured separately from the phospholipids to depict its equal distribution across leaves. Terminals are coloured in black. C) A top view of the membrane system with SS bridges highlighted in red colour. The top view gives a better perspective of outward tilting and binding pocket. Cholesterol is shown in yellow as ball and stick model.*

5.3.4 *Validation of the structural model in bilayer*

Disulfide bridges in the extracellular domains is a signature feature of chemokine receptors. These SS bonds brings the ECD1 (N-ter to 50 residues) closer to the extracellular face of the receptor. Since ECD1 contain motifs for ligand identification and specificity, the SS-bond assist in formation of the binding pocket. The other disulfide bond associates ECD2 with TM4 thus providing anchorage to the changed conformation [430]. Due to these disulfide bridges, the inward tilting of the helices is observed. The inward tilting causes an outward tilting of the helices supported by the microswitch containing motifs, CWxp, DRY, and NPxxY. All chemokine receptors have conserved disulfide bridge forming cysteines, with an exception of CXCR6.

The disulfide bridges are found conserved in the structural model. The orientation of helices was modified after the addition of lipids. Chain A helices has an additional tilt of 2° and the helices of the chain B have a tilt of 13°. However, the overall tilt of the dimer matches the initial tilt estimated by PPM server, i.e. 4. ± 2°. The transmembrane regions are also in validation to the PPM server estimations.

5.3.5 *Structural interface*

The interface of the homodimer membrane system is calculated using PIC web server [356] and ViP approach. ViP is developed by our collegue, Dr. Jeremy Esque based on his work on VLDP to describe structures using Voronoi and Laguerre tessalations. ViP has been standardized for membrane bound interfaces and therefore provides confidence to the calculations from PIC. The interface interactions are observed to be predominantly hydrophobic in nature with few important polar interactions. Residues from TM5 and TM 6 of both the monomeric units forms majority of the interface. However, a smaller interface lying closer to the cytosolic leaflet of the bilayer is also observed. Fig 5.11 shows the structural interface of the ACKR1 homodimer.

The larger interface lies towards the extracellular face of ACKR1 comprising mostly of hydrophobic residues like, Ala, Leu, Val, Phe, and Ile. A total of 23 residues, majorly from TM5, forms the interface. The electrostatic interactions (in red, Fig 5.11) are made by Glu present at the N-terminal region of TM5, Arg and Lys belonging to TM6, and Arg residing in ICD2. The arginine residues from ICD2 makes the smaller interface at the intracellular side. The interface in the ICD2 is most surprising because ICD2 is the location of Arg microswitch of DRY motif (in gray color, Fig 5.11). The Arg in DRY either forms an ionic bond with Asp of DRY in inactive state or interacts with G-proteins during activation. However, DRY is completely absent in ACKR1 which makes this observation interesting.

Apart from these two polar interactions are also observed involving Thr from TM5 and Ser from TM6 (in blue, Fig 5.11). Therefore, a total of 23 residues from TM5, TM6, and ICD2 makes the structural interface between DARC dimer.

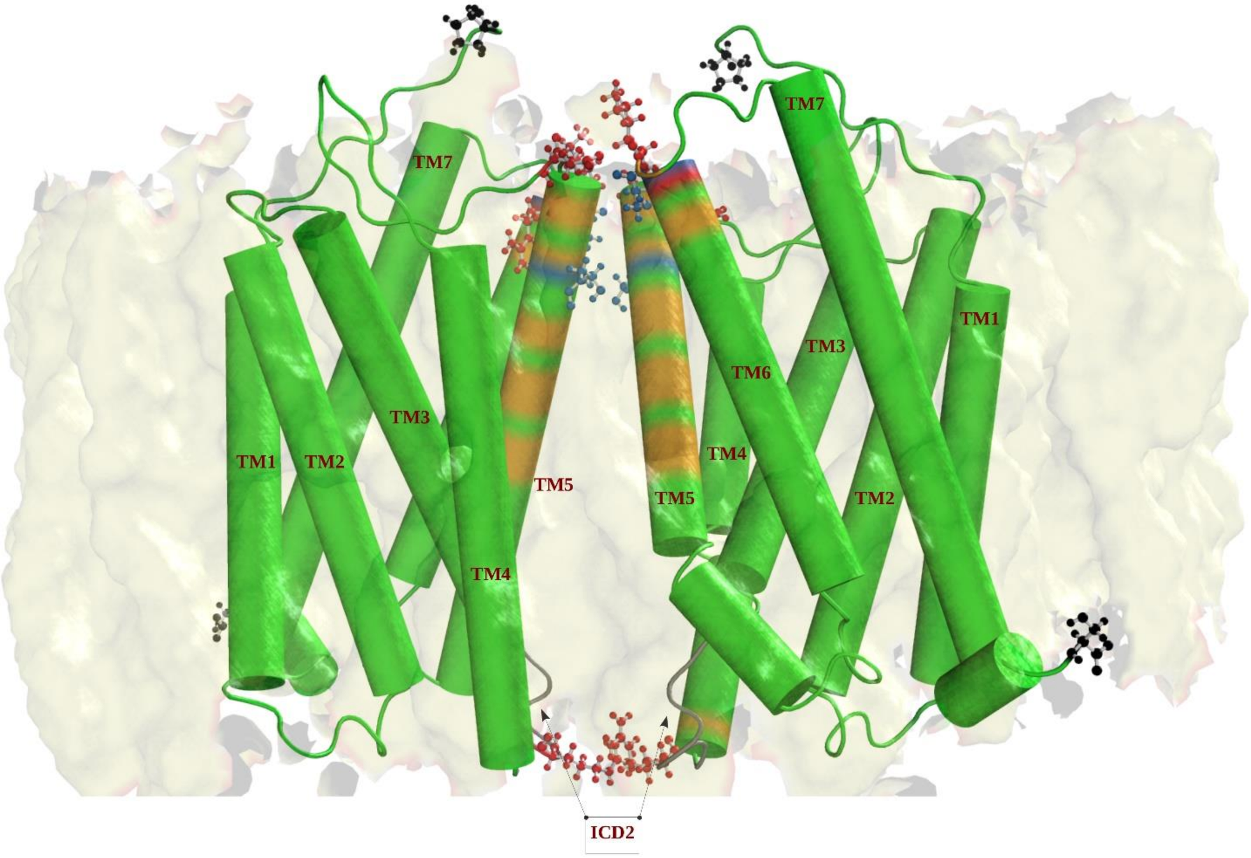


Figure 5.11 ***Dimer interface.*** *The figure shows 7 + 7 TM-helices as cylinders embedded in phospholipid membrane (shown as transparent surface). The helices are marked with their numbers. The interacting residues are coloured in yellow (hydrophobic), blue (polar), and red (charged interactions). TM5 and TM6 covers the major interface area on extracellular side. However, ICD2 also have charged interaction (Arg) forming a smaller interface at intracellular face.*

5.4. Conclusions and future perspectives

ACRK1 is the oldest known among chemokine receptors. However, it was identified as a blood group antigen system in 1950 and called 'Duffy' in reference to the patient in whose sera, the antigen was found. Therefore it was called Duffy Antigen [431]. It would be decades later in 1993 that Duffy antigen would be found to have structural properties of a chemokine receptor and acts as a receptor for malarial parasite, *P.vivax* [432]. Since 1993, there have been numerous studies on Duffy antigen's role as a *P.vivax* receptor [433–435] but a few discusses about its role as a chemokine receptor [366,430,436]. Later in 1998, the contribution of Duffy antigen towards chemokine system would be identified as a receptor of chemokine ligands belonging to different classes [437]. Thus Duffy antigen was renamed as Duffy Antigen Chemokine Receptor (DARC).

DARC is identified as a mammalian chemokine receptor that can bind to inflammatory chemokines across classes. Besides able to bind effectively to different chemokines, it does not transduce the signal as it lacks the motifs that couple with G-proteins. Therefore, in 2014 International Union for Pharmacology (IUPHAR), updated the nomenclature and replaced DARC with Atypical Chemokine Receptor 1 (ACKR1) [384]. Among the atypical chemokine receptors, ACKR1 is the only one that exhibit promiscuous binding with chemokines and lacks the DRYLAIV motif completely. Also, ACKR1 serves as a receptor for *Plasmodium vivax* merozoites leading to the symptomatic infectious stage of malaria.

These makes ACKR1 an important protein from physiological, pathological and evolutionary perspectives. Still, there is none to scarce information about the structure of ACRK1 except for the residues 19-30 in its N-terminal extracellular domain. Therefore, we decided to build a structural model for ACKR1 integrating the physiological, pathological and evolutionary information available about ACKR1. The physiological and pathological properties of chemokine receptors assisted in identifying the key residues. Structural information from other chemokine receptors would provide the basic scaffold for modeling ACKR1. However, chemokine receptors exist in various oligomeric states and therefore it was challenging to decide the oligomeric state as well as homo- or hetero- composition of ACKR1. These questions were addressed with the information gathered on oligomerization of chemokine receptors and phylogenetic analysis of 21 human chemokine receptors. Thus, ACKR1 was decided to be modelled as homodimer based on its closest homolog (with available structure), CXCR4. The modelling procedure was provided with the knowledge about the transmembrane boundaries and disulfide bridges. Since, the structure under consideration is a membrane protein caution was taken in selecting the best structural model. One such approach is to use T20 Test, where sets containing 20 top ranked models from different assessment scores are intersected to obtain best model(s). The generated model is validated for conservation of important residues and structural features. Moreover, comparison with 12 years old model show important difference underlying the needed to propose novel one. As the structure under study is a membrane protein, the structural model is embedded in an *in-silico* membrane system. Given that ACKR1 expressed on reticulocytes acts as the receptor for *P.vivax*, the embedded membrane system mimics the real RBC membrane composition. The interfacial residues are identified from the dimer and they are in accordance to the physiological data available for chemokine receptors. This enhances confidence in the structural model of ACKR1.

In terms of understanding the behavior of local structural flexibility, we notched up to a more complex structural organization with a dimer formation in a phospholipid membrane system. The primary objective is to understand the dynamics of local secondary structures and protein blocks at the interface region as well as at the sites of conserved motifs, like ICD2. Thereafter, a perturbation response study of key residues in the dynamic local structures can help us understand the role of allostery in the 7TM structure of ACKR1. Therefore, conclusive remarks on the dynamics of the local structures in the homodimeric, membrane embedded, assembly of ACKR1 will require all atom molecular dynamics. However, given the enormous size of the system the computational cost is expensive. Therefore, while the 1 microsecond range simulations are running on the cluster, a primary study of the motions using ANM based normal mode analysis (NMA) is designed. The NMA of the ACKR1 dimeric model will also be used for perturbation studies.

The impact of this study will be two-fold. Besides understanding the role of structural flexibility in a membrane protein assembly can reveal insights about the behavioral changes in local structures depending upon the context. While, a molecular model of *P.vivax* DARC Binding Ligand region II (*Pv*DBL-RII) is under process, the molecular modelling and dynamics protocol designed for ACKR1 can be directly applied to *Pv*DBL-RII. The docking pose of ACKR1 dimer and PvDBL-RII is already known from the PBDID 4NUV. Therefore, molecular dynamics of the complex with an RBC mimic membrane is expected to be impactful towards identifying key residues in the complex and targeting them to inhibit *P.vivax* binding to DARC in reticulocytes.

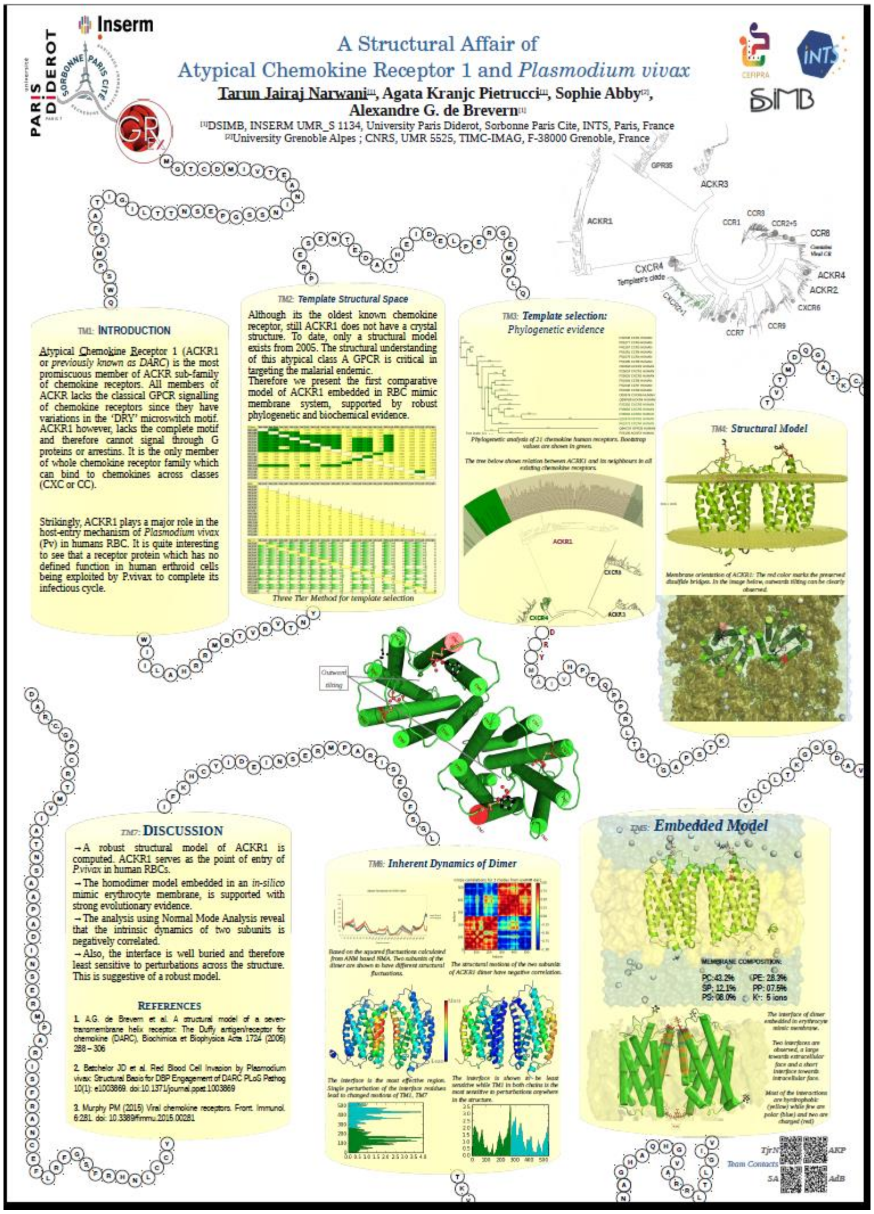
Our new colleague Dr. Agata Kranjc Pietrucci has already started working towards modelling the *Pv*DBL and ECD1 of ACRK1.

ACKNOWLEDGEMENTS

Given that almost little to nothing is known about the structure of ACKR1, I would really like to acknowledge the patience and support given by Dr. Sophie Abby, Dr. Alexandre de Brevern, Dr. Agata Kranjc Pietrucci. A special mention is needed for Dr. Olivier Bertrand for sharing his immense experience with DARC through insightful discussions.

Dissemination of results

The results from the structural study of ACKR1 have been published as a scientific poster at ISMB-3dSIG conference held at Prague in July 2017. A more recent and updated poster is to be presented at EMBL BioMalPar XIV conference to be held at Heidelberg in May 2018. The conference hosts a section dedicated to computational approcahes and therefore can nourish intense discussion on our results. The response to the poster at ISMB yielded fruitful discussions with researchers dedicated to GPCR biology, like Dr. Ravinder Abrol of California State University. Following is the updated poster: Narwani TJ, Pietrucci AK, Abby S and de Brevern AG. DARC shade of chemokine receptors [version 1; not peer reviewed]. *F1000Research* 2017, **6**(ISCB Comm J):1269 (poster) (doi: 10.7490/f1000research.1114529.1)



A Structural Affair of Atypical Chemokine Receptor 1 and *Plasmodium vivax*

Tarun Jairaj Narwani[1], Agata Kranjc Pietrucci[1], Sophie Abby[2], Alexandre G. de Brevern[1]

[1]DSIMB, INSERM UMR_S 1134, University Paris Diderot, Sorbonne Paris Cite, INTS, Paris, France
[2]University Grenoble Alpes ; CNRS, UMR 5525, TIMC-IMAG, F-38000 Grenoble, France

TM1: INTRODUCTION

Atypical Chemokine Receptor 1 (ACKR1 or *previously known as DARC*) is the most promiscuous member of ACKR sub-family of chemokine receptors. All members of ACKR lacks the classical GPCR signalling of chemokine receptors since they have variations in the 'DRY' microswitch motif. ACKR1 however, lacks the complete motif and therefore cannot signal through G proteins or arrestins. It is the only member of whole chemokine receptor family which can bind to chemokines across classes (CXC or CC).

Strikingly, ACKR1 plays a major role in the host-entry mechanism of *Plasmodium vivax* (Pv) in humans RBC. It is quite interesting to see that a receptor protein which has no defined function in human erthroid cells being exploited by P.vivax to complete its infectious cycle.

TM2: Template Structural Space

Although its the oldest known chemokine receptor, still ACKR1 does not have a crystal structure. To date, only a structural model exists from 2005. The structural understanding of this atypical class A GPCR is critical in targeting the malarial endemic.
Therefore we present the first comparative model of ACKR1 embedded in RBC mimic membrane system, supported by robust phylogenetic and biochemical evidence.

Three Tier Method for template selection

TM3: Template selection: Phylogenetic evidence

Phylogenetic analysis of 21 chemokine human receptors. Bootstrap values are shown in green.

The tree below shows relation between ACKR1 and its neighbours in all existing chemokine receptors.

TM4: Structural Model

Membrane orientation of ACKR1: The red color marks the preserved disulfide bridges. In the image below, outwards tilting can be clearly observed.

TM7: DISCUSSION

→A robust structural model of ACKR1 is computed. ACKR1 serves as the point of entry of P.vivax in human RBCs.

→The homodimer model embedded in an *in-silico* mimic erythrocyte membrane, is supported with strong evolutionary evidence.

→The analysis using Normal Mode Analysis reveal that the intrinsic dynamics of two subunits is negatively correlated.

→Also, the interface is well buried and therefore least sensitive to perturbations across the structure. This is suggestive of a robust model.

REFERENCES

1. A.G. de Brevern et al. A structural model of a seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC), Biochimica et Biophysica Acta 1724 (2005) 288 – 306

2. Batchelor JD et al. Red Blood Cell Invasion by Plasmodium vivax: Structural Basis for DBP Engagement of DARC PLoS Pathog 10(1): e1003869. doi:10.1371/journal.ppat.1003869

3. Murphy PM (2015) Viral chemokine receptors. Front. Immunol. 6:281. doi: 10.3389/fimmu.2015.00281

TM6: Inherent Dynamics of Dimer

Based on the squared fluctuations calculated from ANM based NMA. Two subunits of the dimer are shown to have different structural fluctuations.

The structural motions of the two subunits of ACKR1 dimer have negative correlation.

The interface is the most effective region. Single perturbation of the interface residues lead to changed motions of TM1, TM7

The interface is shown to be least sensitive while TM1 in both chains is the most sensitive to perturbations anywhere in the structure.

TM5: Embedded Model

MEMBRANE COMPOSITION:

PC: 43.2% PE: 28.3%
SP: 12.1% PP: 07.5%
PS: 08.0% K+: 5 ions

The interface of dimer embedded in erythrocyte mimic membrane.

Two interfaces are observed, a large towards extracellular face and a short interface towards intracellular face.

Most of the interactions are hydrophobic (yellow) while few are polar (blue) and two are charged (red)

Team Contacts

Chapter 6: An evolutionary perspective on Chemokine Receptors.

6.1 Introduction:

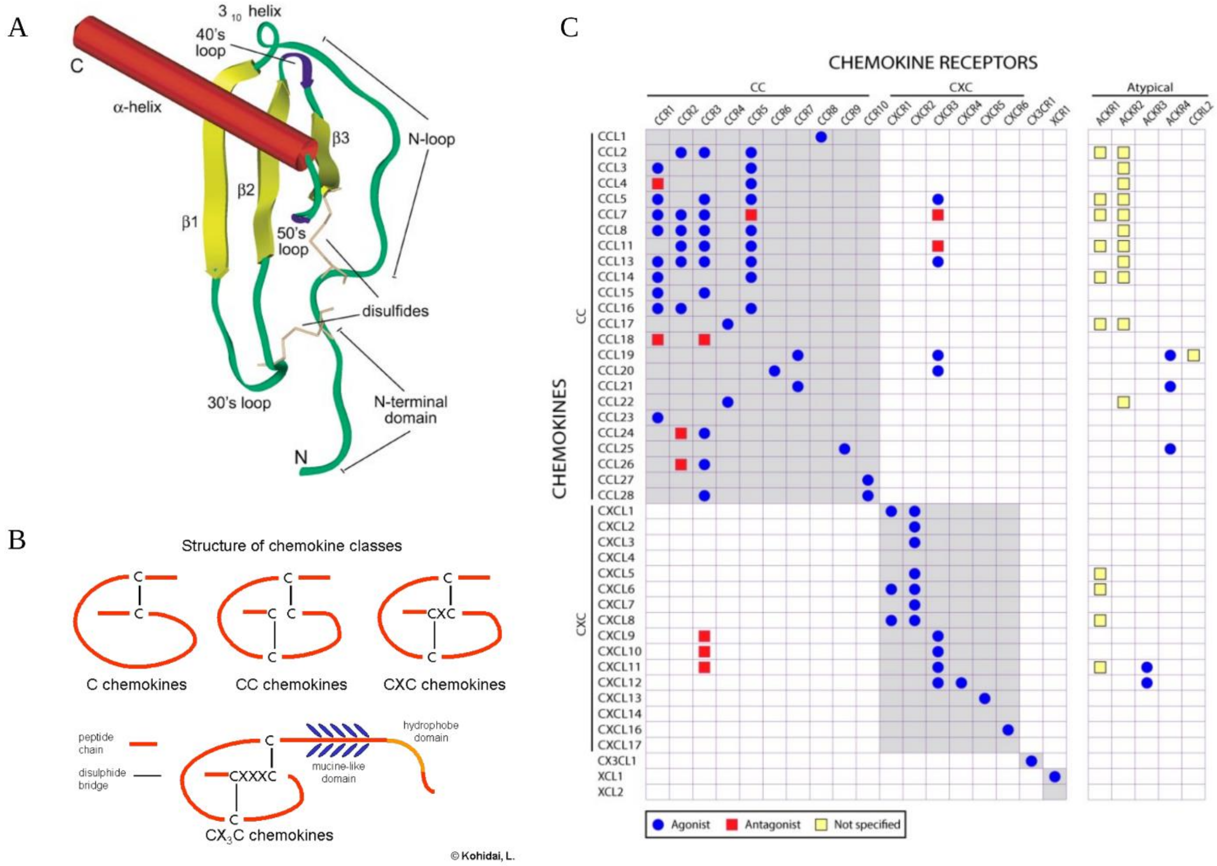
As discussed in chapter 5, the template selection protocol for modelling ACKR1, was supported by the phylogenetic placement of ACKR1 among human chemokine receptors. The resulting tree topology depicted ACKR1 to be highly distant from rest of the clade. Thus leaving a big phylogenetic gap between ACKR1 and its closest neighbor. Also, recently it has been established that the silent mutation in Duffy negative of Western African population have reached fixation levels [438]. This mutation grants them natural immunity against *Plasmodium vivax* Malaria and thus is seen as a striking example of natural selection of genetic traits.

The function of ACKR1 on erythroid cells as well as the underlying mechanisms of its promiscuous behavior towards chemokines and *P.vivax* DBL is not clearly established. Moreover, the chemokine receptor family consist of many anomalies. Besides, ACKR1 interaction with *Plasmodium vivax*; CCR5, CCR3 and CXCR4 also plays an important role in the entry of the virus during HIV-1 infection [439,440]. Interestingly, there are reported incidents of gene piracy in the chemokine receptors. The large DNA viruses copy the encoding regions of host chemokine receptors and use them against the host machinery to either bypass the immune response or cellular reprogramming or cell entry [441]. Several viral homologues of chemokine receptors have been identified in Humans, mostly encoded by the members of *Poxviridae* and *Herpesviridae* families [442,443].

Therefore, a comprehensive phylogenetic analysis of chemokine receptors is designed to understand some, if not all, peculiarities of chemokine receptors.

6.1.1 Chemokines:

Chemokines are an abbreviated form of chemotactic cytokines and are defined as the cytokines that induce chemotaxis by binding to GPCRs. They are involved in major immunological and homeostatic pathways and thus form the largest family of cytokines [444][445]. The chemokine concentration acts as a chemoattractant to guide the migration of cells, mostly leukocytes (also called homing). Chemokines are small proteins weighing 8 to 10 kilodaltons characterized mainly by 4 invariant cysteine residues [445]. Functionally, chemokines can be characterized as inflammatory and homeostatic. The homeostatic chemokines are constitutively secreted and perform functions like leukocyte trafficking out of bone marrow, and across blood and lymphatic vessels. Therefore, the homeostatic chemokines are important for immunosurveillance and immune tolerance of the organism [446] [447]. Moreover, the memory of an immune response is also dependent on the leukocyte localisation during an immune response [448]. The migration of leukocyte also happens during diseases that deregulate the immune system. For example, inflammation during atherosclerosis, chronic allergies in autoimmune diseases, multiple sclerosis and many others. The chemokines that traffic such leukocytes are called, inflammatory leukocytes. A complete classification of chemokines can be seen in Figure 6.1C



CHEMOKINE RECEPTORS

		CC										CXC								Atypical				
		CCR1	CCR2	CCR3	CCR4	CCR5	CCR6	CCR7	CCR8	CCR9	CCR10	CXCR1	CXCR2	CXCR3	CXCR4	CXCR5	CXCR6	CX3CR1	XCR1	ACKR1	ACKR2	ACKR3	ACKR4	CCRL2
CC	CCL1								●															
	CCL2		●	●		●														□	□			
	CCL3	●				●															□			
	CCL4	■				●															□			
	CCL5	●		●		●								●						□	□			
	CCL7	●	●	●		■								■						□	□			
	CCL8	●	●	●		●															□			
	CCL11		●	●		●								■						□	□			
	CCL13	●	●	●		●								●							□			
	CCL14	●				●														□	□			
	CCL15	●		●																				
	CCL16	●	●			●																		
	CCL17				●															□	□			
	CCL18	■		■																				
	CCL19							●						●									●	□
	CCL20						●							●										
	CCL21							●															●	
	CCL22				●																□			
	CCL23	●																						
	CCL24		■	●																				
	CCL25									●													●	
	CCL26		■	●																				
	CCL27										●													
	CCL28			●							●													
CXC	CXCL1											●	●											
	CXCL2												●											
	CXCL3												●											
	CXCL4																							
	CXCL5												●							□				
	CXCL6											●	●							□				
	CXCL7												●											
	CXCL8											●	●							□				
	CXCL9			■										●										
	CXCL10			■										●										
	CXCL11			■										●						□		●		
	CXCL12													●	●							●		
	CXCL13															●								
	CXCL14																							
	CXCL16																●							
	CXCL17																							
	CX3CL1																	●						
	XCL1																		●					
	XCL2																							

● Agonist ■ Antagonist □ Not specified

Figure 6.1 ***Chemokines structure and classification.*** *A) shows the structural scaffold of chemokines. The sites of two disulfide bridges is also shown. N-terminal domain ends before the first Cys residue. N-loop plays important role in binding to the receptor. B) schematic diagrams of chemokines according to their classes. The number of residues between the two Cys between N-termimal domain and N-loop forms the basis of classification of chemokines. C) The table shows*

all the known chemokines and their receptors. The agonist and antaonist activities are in reference to the type of response illicit by the receptor after binding to chemokine. ACKR1 is the only receptor that binds to chemokines from both the classes. Also ACKR1 binds only to inflammatory chemokines.

$^{+}$Image A taken from [444], $^{+}$Image C taken from [378]

All chemokines adopt a similar structure with 3 antiparallel beta-strands and an alpha helix as depicted in Figure 6.1A. The N-terminus is a long deformed region that contains two of the four cysteine residues. The position of the two cysteines in the N-terminal loop are important in the nomenclature of the chemokines. The loop region from second cysteine residue uptil a short, single turn 3_{10}-helix is called as an N-loop, followed by three antiparallel beta strands (β1, β2, β3) succeeded by a C-terminal α-helix. The turn connecting β1 to β2 is termed as 30's loop, while the turn between β2 and β3 is called 40's loop. The 50's loop joins the β3 strand to the α-helix. The 30's turn contains the third cysteine and 50's loop contains the fourth cysteine. These cysteines form two disulfide bridges with each of N-terminal cysteine forming a S-S bond with 30's loop cysteine and 50's loop cysteine, respectively [444].

The chemokines are classified by the local sequence of the N-terminal cysteines. Presence of an amino acid between the two cysteines represented as a local motif -CxC- is classified as *α*-chemokine or CXC chemokine Figure 6.1B. When the two cysteines are adjacent, the chemokine is classified as CC- or *β*-chemokine. These two classes represent majority of the chemokines with 27 *a*- and 17 *β*-chemokines known. However, some smaller groups like CX3C chemokines and C chemokines exist. Of these, CX3C or *ɣ*-chemokines are represented by the presence of three amino acids between the N-terminal cysteines. CX3C chemokine is unusual as it is a part of a surface receptor and contains only one member CX3CL1 or fractalkine (Fig 6.1B). The C-type chemokines are also unusual because it contains only two cysteines, one at N-terminal and another one downstream. Therefore, C-chemokines contain only one disulfide bridge (Fig 6.1B). To date, only two C-chemokines known as XCL1 and XCL2 exist. With exception of two C-chemokines, the rest three chemokine families have four cysteines albeit they are not positionally conserved. Therefore, the identity between chemokine sequences can vary from 20% to 90% [444].

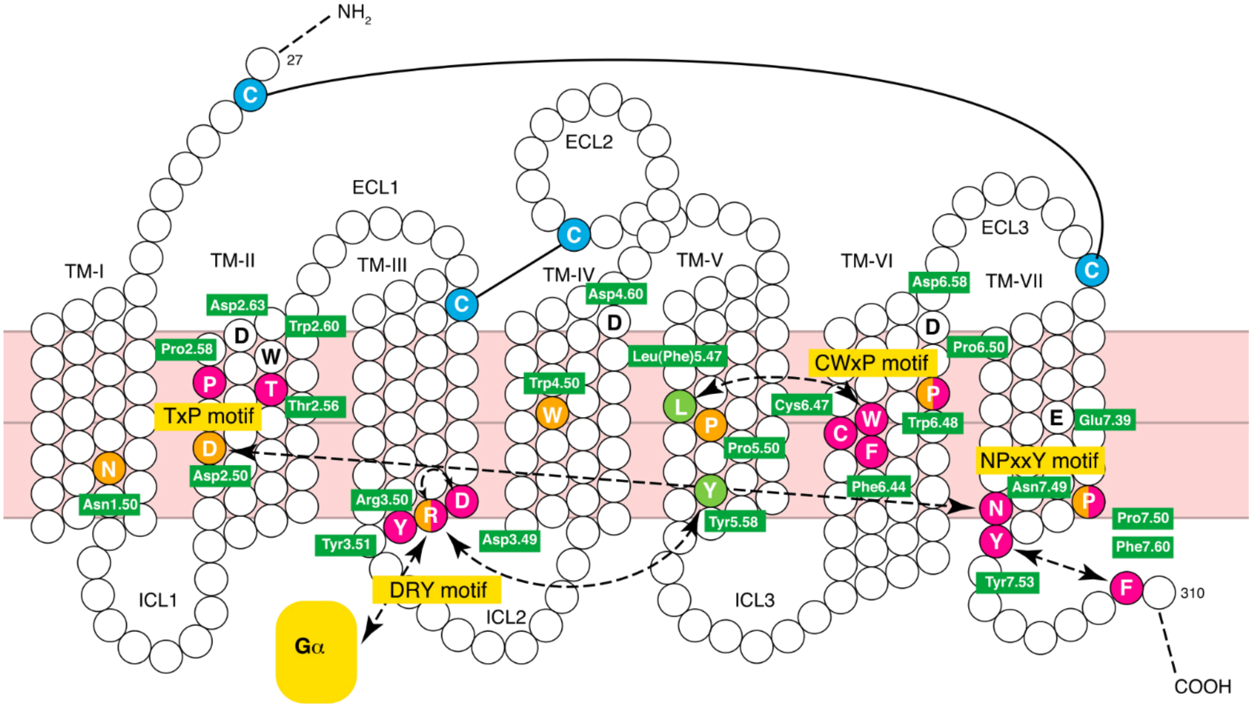


Figure 6.2 ***Chemokines receptors.*** *The schematic representation of basic chemokine receptor scaffold. Important residues are marked with green, orange and pink color. Blue denotes the residues involved in disulfide bridges. Pink represents the most conserved residue in a TM helix while orange indicates the microswitch residues. Yellow tabs show the important motifs involved during the activation of the chemokine receptor. Arrows indicate the interactions, (solid- permanent, dotted- transient).*

$^+$Image taken from [381]

6.1.2 Chemokines and their receptors:

NMR studies have shown that the formation of the disulfide bridges lead to decrease in deformability of the N-terminal loop but the N-loop stays flexible [449]. The N-terminal loop of *a*-chemokines (CXC) is responsible for the activation of leukocytes. The motif composed of residues glutamate, leucine, and arginine (ELR) present in the N-terminal loop (preceding the first cysteine) is the key player in activation of neutrophils and sometimes eosinophils [450]. It has been established that the *a*-chemokines consisting an ELR motif (called ELR$^+$ chemokine) induces leukocyte migration and are therefore involved in homeostasis [451]. However, the ELR$^-$ *a*-chemokines (eg. CXCL8, previously, interleukin-8) are involved in migration of lymphocytes that leads to inflammation. The ELR and other residues in the N-terminal loop like lysine, serine, and methionine are crucial for the activation of the receptor [452].

The flexibility of the N-loop is determined to play a role in recognition and binding of the chemokine to its receptor known as a chemokine receptor. The residues like Tyr, Ser, Lys, Pro, Phe, Gln, Ile, Ser, Arg, Glu, and His are crucial in specificity of the chemokine ligand towards the chemokine receptor [453–456]. For instance, presence of sequences like 'YSKPF', 'LQGI', 'RFFESH' in the N-loop of CXCL8, CXCL1, CXCL12, CCL7, CCL2, CCL5 confers specificity to their respective receptors. Also, mutational studies of the β-chemokine N-loop residues, established four core residues important for receptor specificity and binding [457].

The chemokine receptors are one of the largest sub-family among class A GPCRs adopting a septa-helical transmembrane (7TM) structure. Different chemokine receptors have variation in the lengths of their N-terminal extracellular domain (ECD1) and C-terminal intracellular domain (ICD4). The ECD1 is composed of mostly acidic amino acids with a Tyr residue that can be sulfonated and at least one Cys residues [458]. The cysteine forms a disulfide bridge with ECD3 (see Figure 6.2). The S-S bridge is crucial to the binding of the chemokine ligand [430]. The N-terminal domain also consists of a site for N-acetyl glycosylation. As was observed in chapter 3, attachment of a glycan is a surface event and may not interact with the core of the structure. However, the N-glycosylation at the terminals behave differently [459]. Therefore, the glycosylation in the N-terminal loop of chemokine receptors might assist in stability of the otherwise mobile ECD1 thus forming a pocket for binding the chemokine ligand [459]. Although the Cys, Arg and Tyr residues provide identity to the ECD1 of chemokine receptors yet there is variability in the length and composition of ECD1 among various chemokine receptors. Such a variability is important for ligand specificity [460]. Unlike N-terminal domain, the C-terminal domain does not vary much in size across the chemokine receptor family [461]. It is identified by the presence of serine / threonine residues which are sites for phosphorylation by a GPCR Kinase (GRK) [462]. The phosphorylation in ICD4 is crucial to desensitize the activated chemokine receptor by preventing the recoupling of the G-proteins with the 7TM helices. The desentization is important, in order to prevent repeated stimulation of the receptor and to mark it for internalization [463,464]. Such signalling mechanism is an alternative to classical Gαi mediated signaling in GPCR.

Upon chemokine receptor activation, the S-S bridges between ECD1 and ECD4 / TM7, and TM3 and ECD3, brings the helices closer forming a pocket at the extracellular face (see Figure 6.3) for the binding of chemokine ligand. At the intracellular face, the ICD3 and ICD4 orients in close proximity to the heterotrimeric G-proteins (known as Gα, Gβ, and Gγ) with Gα binded to a molecule of GDP (guanosine diphosphate). Apart from ICD2 and ICD3, the intracellular regions of TM helices (3,5,6,7) also interact with the Gα subunit of the heterotrimeric G-proteins. The extracellular loops (ECD2, ECD3, ECD4) as well as N-terminal regions of each TM helix interacts with the binding chemokine [465]. Figures 6.2 and 6.3 highlights the interacting residues and 3-D structural orientations during chemokine ligand receptor interaction.

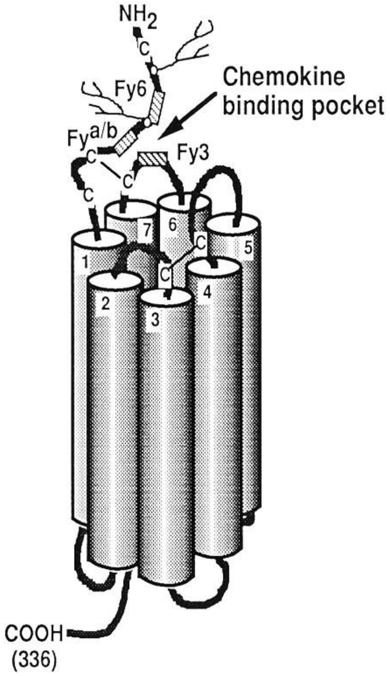


Figure 6.3 ***Structural organization of a chemokine receptor.*** *The 7TM helices forms a classical helical bundle with three interconnecting loops, facing each extracellular and intracellular sides. Helices are marked as numbers. Chemokine binding pocket is shown along with the specific epitopes present in ECD1 for ligand identification. In the current image, given the epitopes, the chemokine receptor is ACKR1. The crucial disulfide bridges is also shown. The disulfide bridge between N-terminal domain and last extracellular loop is crucial for ligand binding pocket formation.*

+Image taken from [372]

6.1.3 Chemokine signalling in brief:

The chemokine binding to the receptor illicit a conformational change across extracellular to intracellular faces of the 7TM leading to the exchange of GTP for GDP molecule. This dissociates the heterotrimeric complex of G-proteins [376]. Gβγ subunits are able to activate Phospholipase Cβ2 (PLC). In the cell, PLC cleaves phosphatidylinositol 4,5-bisphosphate (PIP_2) into phosphatidylinositol 1,4,5-triphosphate (IP_3) and diacyl-glycerol (DAG). Of these secondary messengers, IP_3 mobilizes the cytosolic free calcium while DAG employs Ca^{2+} to activate various Protein Kinase C (PKC). The activated PKC and free calcium ions forms the basis of various subsequent pathways that would lead to transendothelial migration as well as paracellular and transcellular migration of leukocytes [466]. The selection of the subsequent pathway will be dependent on the kind of chemokine ligand attached to the chemokine receptor. A homeostatic chemokine of CXC type, like CXCL1, CXCL8, CXCL6 could follow either of Ras-Rho, MAPK or RTK pathways. Binding of an inflammatory chemokine like CCL11, CCL7, CXCL9, CXCL10 will employ different pathways for inflammatory responses. Although competitive binding to chemokine receptors across chemokine family is rare yet there are promiscuous receptors that bind to both types of chemokines [467]. Figure 6.4 shows different pathways by which chemokine receptor signal.

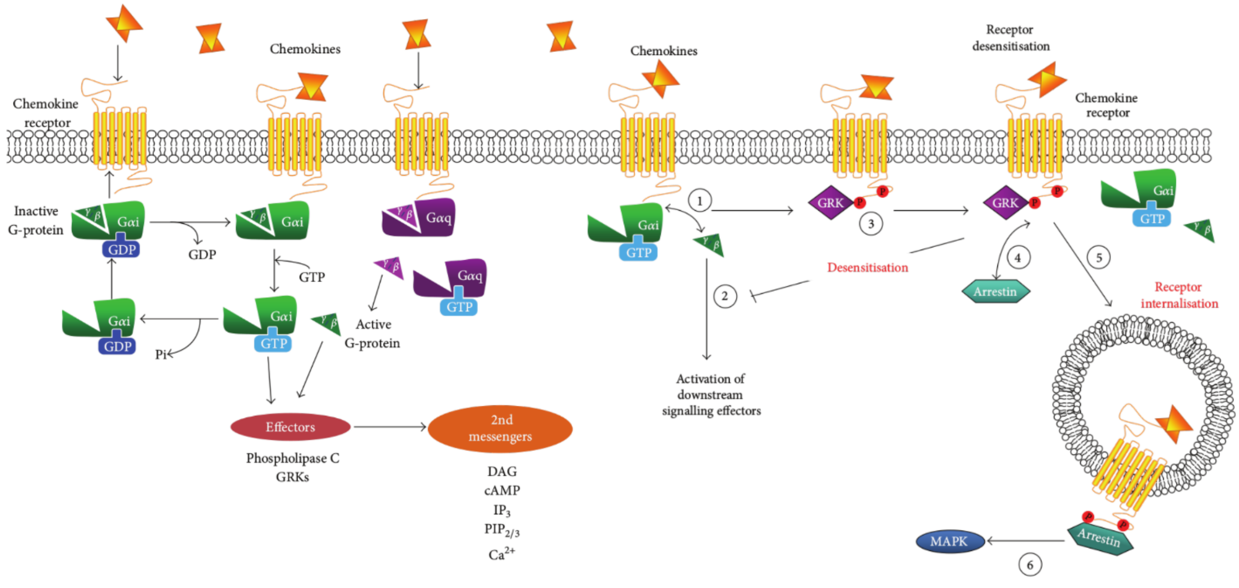


Figure 6.4 ***Chemokine system signalling pathway.*** *The complete signalling scheme for chemokine system. The left side shows normal G-protein signalling via Gαí, Gβɣ, and Gαq pathways. Upon activation, the GDP is exchanged for GTP thus dissociating the heterotrimeric G-proteins. The*

right side receptors show atypical or G-protein independent signaling. The steps 1,2,4 depicts the biased signalling usually leading to receptor internalization.

+Both Image are taken from [376]

6.1.4 Conservation of microswitches in chemokine receptors:

The availability of the crystal structures of active state class A GPCRs (like Bovine Rhodopsin) by the advent of new millenia, encouraged numerous biomolecular and biophysical studies. These lead to a practical understanding of the activation mechanism of GPCRs. Later, in 2010 the first crystal structure of a chemokine receptor was solved [395]. Numerous studies supported that the activation of the chemokine receptor by the chemokine binding is accompanied by a see-saw tilting of the 7TM helices [468,469]. The extracellular portions of the helices supported by the disulfide bridges lead to an inward tilting of the helices. Thereby, causing an outward tilting of the intracellular regions of the TM helices. This leads to the interaction of the ICDs with G-proteins. The key residues involved in such a switching of the structure, post activation are called molecular microswitches. Following are the well known microswitches in chemokine receptors.

(i) *DRY motif*- The conserved three residue motif composed of Asp, Arg and Tyr is located at the junction of TM3 and ICD2. The motif is positionally conserved throughout the chemokine receptor family with one exception. The arginine residue in the motif functions as a microswitch. In the inactive state, it forms an ionic lock with the Asp [380]**.** Post ligand binding, the pKa of the arginine changes consequently leading to the disruption of the ionic lock. The loss of interaction is compensated by new interactions with the well conserved Tyr residue of the following intracellular loop- ICD3 and Gα subunit of the G-proteins [381]. This leads to the outward tilt of the TM3 helix towards the heterotrimeric G-proteins. The Arg residue is highly conserved throughout the chemokine receptors as well as the class A GPCRs. However, in some class A GPCRs like Rhodopsin, the 'D' of the motif is replaced by Glu (E) residue. Therefore, the motif is also known as E/DRY motif [381].

(ii) *CWxP motif:* Located in the TM6 helix the short motif is composed of residues Cys, Trp succeeded by any residue and a Pro. As mentioned in chapter 1.2 that proline residues are helix breaker due to their cyclic backbone. Therefore Pro in case of CWxP motif causes a kink in the helix due to its singular presence and hydrophobic environment [469]. This kink is crucial to the helix movement during activation. Upon activation, the Trp residue interacts with the Phe of the neighboring TM5, as marked by arrows in Figure **6.2**. This interaction accompanied by the disulfide bridge between TM3 and ECD3 leads to the tilt in the TM5 helix. Thus Trp functions as a microswitch in CWxP motif. However, the Cys residue of the motif is also claimed to work as a microswitch [470]. The Cys residue interacts with the Asn of TM7. This interaction keeps the helix in position. Upon activation the interaction breaks leading to the outwards movement of intracellular region of TM6. This makes CWxP motif a crucial microswitch with Pro residue inducing kink in the TM6. However, the motif is not found highly conserved [381]. There are two variations in the form of xWxP and xQxP where in the Cys is usually substituted with Thr, Phe, Ser, or Leu. The expected reason for such substitution is the non-conservative nature of the Asn residue in TM7 with which Cys of CWxP interacts [381].

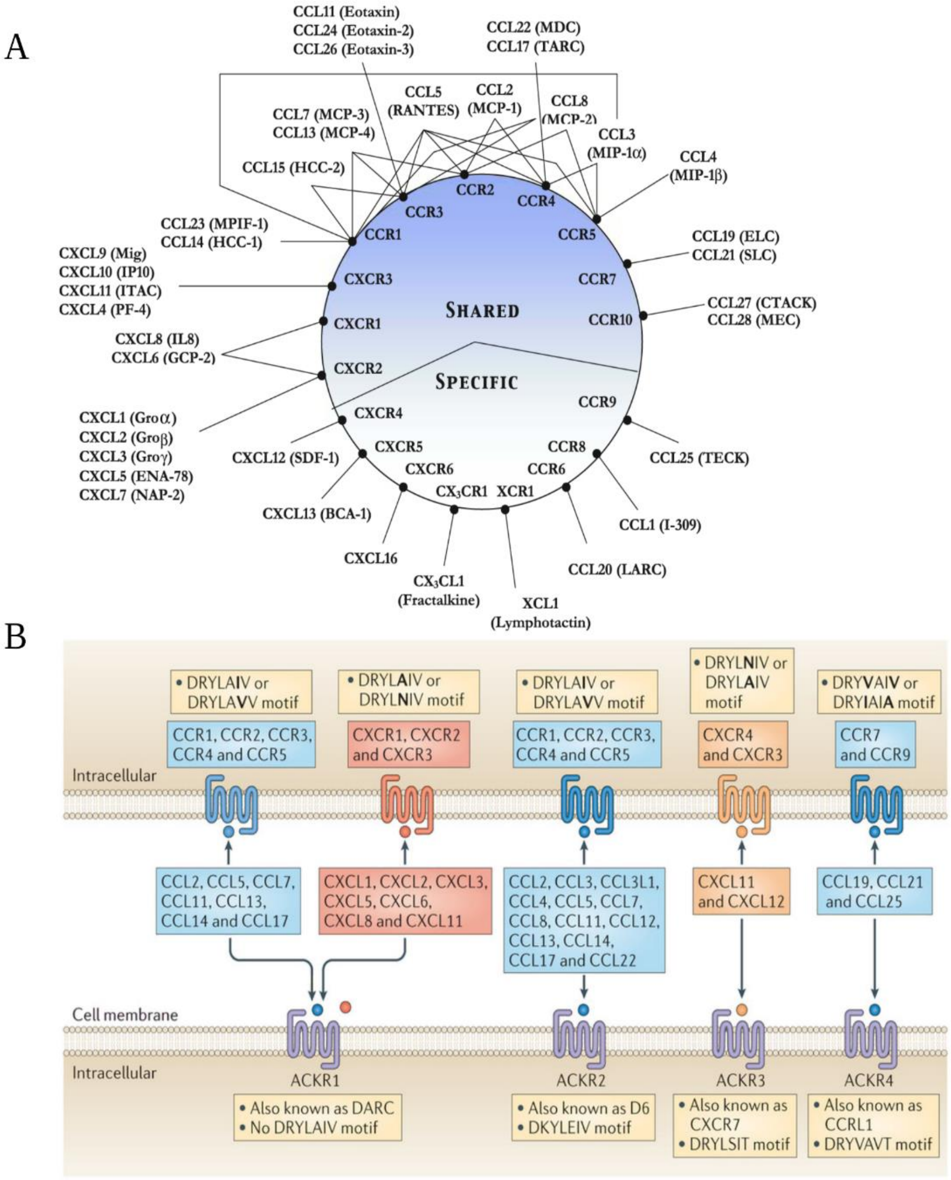
(iii) *NPxxY motif:* The motif comprised of residues Asn, Pro succeeded by two amino acids and a Tyr. The Tyr of the motif interacts with a conserved Phe residue 5 to 6 residues downstream of Tyr [381]. Therefore, it can also be represented as NPxxY(x)$_{5,6}$F motif. Upon ligand binding, the Tyr-Phe interaction breaks and Tyr interacts with a hydrophobic cluster between TM6 and TM7 thus stabilising the tilt of TM6, as marked in Figure **6.2**. The loss of interaction between Cys of CWxP and Asn (TM7) results in the outward tilt with Tyr of NPxxY motif balancing the movement. This may suggest that microswitches might be working in concert.

Apart from these three prominent motifs, there are certain residues that also function as molecular microswitches in the chemokine receptor activation process [381]. However, these are not as conserved as the DRY, CWxP and NPxxY motifs. Such important residues are Phe in TM6, Glu in TM7, Trp in TM2, and a TxP (Thr, any res, Pro) motif in TM2 and interact with one of the DRY, CWxP or NpxxY to induce their effects [381].

6.1.5 *Nomenclature of chemokine receptor*

For receptor activation, the chemokine ligand has to bind with its specific chemokine receptor. Similar to the N-terminal loop of chemokines, the N-terminal region or ECD1 of chemokine receptor is majorly responsible for recognition and binding of the chemokine. As discussed earlier, the ECD1 of chemokine receptors have an acidic composition besides having a sulfonated tyrosine and N-glycosylation. The ECD1 of different chemokine receptors contain different epitopes. These epitopes are essential for the recognition and binding specificity of the chemokine ligand and are thus unique [471]. The epitopes therefore enhance the variability of the ECD1 as well as the specificity of the binding chemokine(s). Thus, forming the basis of nomenclature of chemokine receptors [472].

There are 24 known chemokine receptors identified by the class and type of chemokine ligand that binds to it. The chemokine receptors that interact with a-chemokines are called receptors, while those receiving β-chemokines as their ligands are named, CC-receptors. Thus following the trend, there are primarily four categories of chemokine receptors; CXC-receptor (CXCR), CC-receptor (CXCR), CX3C-receptor or fractalkine receptor (CX3CR), and XC-receptor (XCR). Of these four, the CXCR and CCR have the most number of members with CCR having 11 and CXCR having 7 receptors [472]. They are named as CCR1 through 11 and CXCR1 through 7. Since the γ- and δ-chemokines have only one and two members, therefore, CX3CR have a single receptor (CX3CR1) while XCR have two receptors that are isoforms. Since the number of CXC- as well as CC- types chemokines is way bigger than the number of CXCRs and CCRs respectively. Therefore, the same receptor is used by more than one chemokine ligand using slight variations in the N-loop. For instance, CXCL8 binds to both CXCR1 and CXCR2. The amino acid sequence ‘YSKPF’ in the N-loop of CXCL8 makes it more specific towards CXCR1 over CXCR2 [444]. Nevertheless, the class of the chemokine ligand defines the receptors. Such that the CXC- type receptors will bind to CXC- chemokines and CCRs will bind to only CC-chemokines, as shown in Figure 6.5A. However, certain chemokine receptors like, CXCR7 and CCR11 does not illicit signal via G-proteins upon binding of the chemokine ligand.



Nature Reviews | Immunology

Figure 6.5 ***Chemokines and their receptors.*** *A) shows the share space of the chemokine and their receptors. Some of the receptors are shared among many chemokines for eg. CCL5 binds with 5 receptors. While lower sector of (A) shows the receptors specific to single chemokine. B) shows details about the ACKRs. ACKR1 is the only motif that completely lacks the DRY motif and also binds to two classes of chemokines.*

+Image A is taken from [467]

6.1.5.1 *Atypical chemokine receptors*

The chemokine receptor binds to a chemokine ligand and illicit cellular pathways using G-proteins that finally lead to leukocyte migration for homeostasis or inflammation. A chemokine receptor that bind to chemokines but does not signal is termed as an atypical chemokine receptor (ACKR) [384]. There can be more than one reasons for the failed signal transduction for example, lack of microswitches, or mutations in such motifs, or alternate signaling using β-arrestins. Among the known chemokine receptors, it was found that CXCR7 and CCR11 cannot signal through G-proteins and therefore fail to induce leukocyte migration [473]. Apart from these, Duffy antigen discovered in 1950 would be later classified as an ACKR along with a β-chemokine receptor D6 [386,474]. These comprises a sub-family of four 7TM receptors under canonical chemokine receptors. The four ACKR are named as ACKR1 (previously Duffy Antigen for Chemokine Receptors), ACKR2 (previously D6 or CCBP2), ACKR3 (previously CXCR7), and ACKR4 (previously CCR11).

All ACKR are expressed on non-leukocyte cells. While ACKR1 is also expressed on vascular endothelium and erythrocytes, the rest of ACKR members are expressed on lymphatic endothelial cells [473]. All ACKRs except ACKR1, are known to signal via G-proteins independent pathways like biased signaling using β-arrestins [475]. Therefore, ACKR1 cannot signal either through canonical or alternate signaling mechanism. The reason for such behavior could be attributed to ACKR1 complete lack of the DRY motif, especially the Arg microswitch [381]. Therefore, it cannot couple to G-proteins or β-arrestins. In other ACKRs the Arg microswitch is conserved although the overall DRY-LAIV motif contain variations. Another starking difference between ACKR1 and rest of the ACKRs is that ACKR1 is the only member that binds to both α- and β-chemokines. Figure 6.5B provides details about the four different ACKRs and their ligands.

The ACKRs expressed on endothelial cells are known to perform ligand scavenging functions, wherein the bound chemokine ligand is internalized and subjected to lysosome activity [476]. In vascular endothelium, ACRK1 is involved in ligand presentation. After internalizing of the chemokine by the vascular ACKR1, the ligand is transferred to the laminar side and presented for the leukocyte migration [388]. Such scavenging activity by ACKRs is of importance in inflammatory responses whereby the excessive chemokines can be internalized and hyper-inflammation can be prevented. Therefore, the chemokines binding to ACKRs are mostly inflammatory in nature rather than homeostatic. Fig 6.1C and Fig 6.5A provides an overall summary of different chemokine receptors and nature of their chemokine ligands and their interactions.

6.1.5.2 *Virus encoded chemokine receptors (vChemR)*

Three members of the chemokine receptor family are known to be involved in pathogenesis. ACKR1, besides being an atypical chemokine receptor is also the point of entry for *Plasmodium vivax* into human red blood cells [432,477]. CCR5 and reportedly CXCR4 interacts as a co-receptor with glycoprotein CD4 for HIV-1, during the virus's entry into T-lymphocytes [478]. Such microbial exploitation of the chemokine system can be attributed to its fundamental role in the cell-mediated immunity. Thus, some mammalian DNA viruses have evolved strategies to evade the system. The mammalian dsDNA (double stranded) viruses like poxviruses and herpesviruses employ gene piracy during their host infection and encode the chemokine receptor as well as chemokines and cytokines [479]. The viral cytokines and viral chemokines are then used to compete with host chemokines to desensitize their receptors and thus bypass the immune detection. A vChemR however, attracts the host chemokines that would otherwise will lead to an immune response. To date, 10 virally encoded chemokines and chemokine receptors have been identified. Some of such viruses are; Human Cytomegalovirus (HCMV, a β-herpesvirus), Tanapox virus (a poxvirus), Yaba-like disease virus (YLDV, a poxvirus), Human Herpes Virus (U12 and U51 families of HHV), etc.

The vChemR are structurally similar to the chemokine receptors given their 7TM structure but they vary significantly in their sequence. Most of the vChemR have different conservation profile for the Arg (DRY), Tyr (NPxxY), and Asn (in TM7). Moreover, they have different selective pressures on the DRY-LAIV motif [461]. Such design ables them to employ different pathways to signal, scavenge, internalize or act as a sink for chemokines. Also, most of the vChemR are known to have constitutive activity i.e they can transduce signal through different pathways without a ligand thus remaining in a constant active state [480,481]. However, some receptors like, ORF74 (an HHV8) exhibit high specificity for the ligands of CXCR1 and CXCR2. This have been proposed to be a receptor specific exploitation to gain control over the host's regulatory mechanisms [479]. The ORF74 is peculiar as it has DTW motif instead of DRY motif but can have normal G-proteins signalling. It has been shown that the chemokine binding and signaling is directed by the C-terminal helix of ORF74. Therefore, the evolutionary pressures on the C-terminal of ORF74 will be totally different from canonical and atypical chemokine receptors [482,483]. A mutational study reports that, the C-terminal region of the vChemR are usually shorter and this helps them evade the normal internalization procedure of a chemokine receptor. All of the vChemR except YLDV and ORF74, behaves in a promiscuous manner and can bind to multiple chemokines depending on their host's immune response. Therefore, it becomes compelling to understand such biological hacking of genes that exploits the immune mechanisms, otherwise designed to identify and kill such viruses.

Our prime interest in understanding chemokine receptors was due to ACKR1's (DARC) promiscuous behavior and its relation to *Plasmodium vivax*. Since ACKR1 branched out as an outlier amongst Human chemokine receptors phylogenetic tree (see section 5.3.1). The study was initially directed to find evolution of ACKR1. However, given the complex yet interesting relationships of the chemokine system a number of attempts have been made to understand the phylogeny of the chemokine system, including the vChemR [381,461,484]. All these studies have a caveat that they are centered around the mammalian phylogeny since virally encoded chemokine receptors effect the mammalian hosts. Therefore, in order to answer our questions about evolution of ACKR1 and understanding the phylogenetic perspective of the chemokine receptors, a comprehensive protocol was designed. The objective of this protocol would be to understand the exhaustive phylogenetic relationships among the chemokine receptors.

6.2 **Methods:**

6.2.1 *Extraction of homologs and building the dataset:*

The 21 sequences of chemokine receptors (as obtained in chapter 5) were subjected to pHMMER v3.1b2 [402] with SWISS-PROT database as the target. Output was controlled using e-value (-E) and domain wise e-value (--domE) cutoffs at 10e^{-5}. After filtering the hits by high e-value, short sequences and the domains with low coverage values were also removed. Hits which had significantly large bias were also removed. As a result, a refined set of 118 homologs were obtained.

A multiple sequence alignment of the 118 sequences was generated using alignment program MAFFT v7.27 (Multiple Alignment using Fast Fourier Transform) [406] with its iterative alignment method G-INS-i at 1000 iterations. The method is recommended for global alignment of sequences of similar lengths and works best on a set of <= 200 sequences. The generated MSA was checked for unnecessary gaps using a python script provided by our collaborator Dr. Sophie Abby. The refined alignment was used to build an HMM profile that would be used to search the nr database (March 2016 release). Using 118 sequences instead of the 21 sequences assure that the HMM profile have enough diversity to match distant homologs. Figure 6.6 shows the conservation profile of functionally important sites in 118 chemokine receptors. All the default settings were used for hmmbuild v3.1b2 with --amino tag. Using the HMM profile generated from a carefully curated MSA enhances the confidence in the quality of the hits obtained by hmmsearch (v3.1b). A total of 10332 hits were obtained. The output was filtered based on high domain wise e-values and high bias to score ratios. After removing 3936 sequences, a data-set of 6404 sequences was obtained.

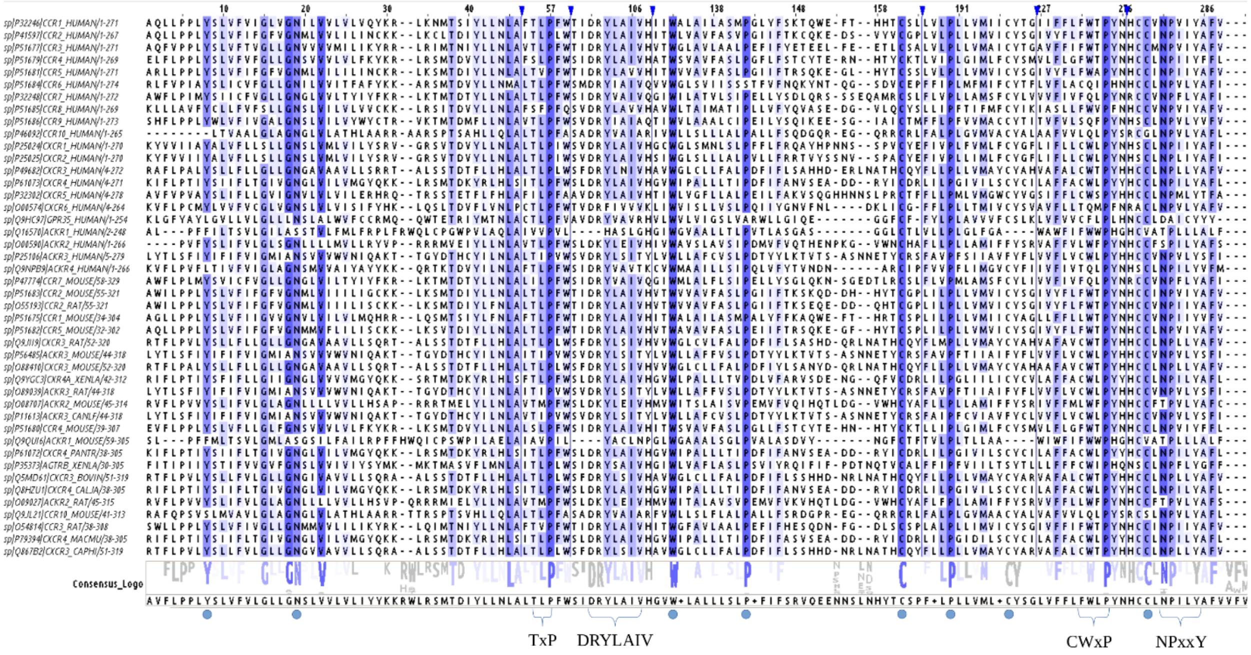

10
20
30
40
57
106
138
148
158
191
227
276
286
Consensus_Logo
AVFLPPLYSLVFVLGLLGNSLVVLVLIYYKKRWLRSMTDIYLLNLALTLPFWSIDRYLAIVHGVW+LALLLSLP+FIFSRVQEENNSLNHYTCSPF+LPLLVML+CYSGLVFFLFWLPYNHCCLNPILYAFVVFV
TxP
DRYLAIV
CWxP
NPxxY

Figure 6.6 ***MSA used for HMM building.*** *The multiple sequence alignment of 118 chemokine receptor homologs. The alignment is colored according to residue identity. The intensity of blue color signifies the extent of conservation of a single amino acid. Conservation of important microswitch motifs is also shown. The blue dots signify high conservation of important residues for chemokine receptors, e.g. Proline.*

6.2.2 *Data-set optimization*:

Visualizing 6404 sequences as a single phylogram would have been difficult and noisy. Therefore, the sequences were clustered at 65% sequence identity using SiLiX clustering program [485], resulting into a non-redundant dataset. The threshold for clustering was chosen based on the consideration that a non-redundant database (NCBI_nr) clustered at 100% identity was used. Also, the average percent identity between class A GPCRs is 26% [397,398]. The clusters were selected based on SiLiX Family_networks_builder (silix-fnet) utility. It gives weighted edges that describes the network between predefined families. In this case, the set of pre-defined families (FILE.FAMS file) include the 21 Human chemokine receptors and the utility was used without -strict option. The clusters were therefore selected based on their size (minimum 100 sequences) and weights of the edges. Thus by clustering, 3277 sequences were obtained. Of these, 148 were singletons, i.e a cluster that have only one sequence. Therefore, a final sequence set comprising of distant homologs of chemokine receptors was obtained, containing 3129 sequences.

6.2.3 *Multiple Sequence Alignment (MSA)*:

MAFFT v7.27 was used to align 3129 homologs of chemokine receptors. Since all the sequences obtained are expected to belong to class A GPCR family, the 7 transmembrane helices provide a strong control. To exploit this feature, an MSA of 21 human chemokine receptors, including ACKR was generated. The alignment was manually edited to conserve the 7TM boundaries as extracted from CXCR4 crystal structure [394] and JPred secondary structure predictions [486]. This alignment was used as a seed for MSA of 3129 sequences. A seed alignment given to MAFFT is expected to anchor the alignment for 7TM conservation. However, it should be noted that it does not conflict with the substitution rates of each position, as described by the substitution matrix. MAFFT was used with a progressive refinement method, FFT-NS-i that uses a rough guide tree. It should be noticed that use of a progressive alignment method gives huge boost to the speed while accuracy can be affected. However, use of the more accurate iterative methods with ~ 3000 sequences would have been computationally expensive. Therefore, the iterative alignment method (G-INS-i) used during MSA for HMM building and the presence of the seed sequences will provide close approximations for the guide tree. The default BLOSUM62 substitution matrix was used because of the SiLiX clustering performed at 65% during data-set optimization stage. The resulting MSA was edited using python scripts to remove empty gaps and noisy alignment positions. The final MSA was visualized using JalView application [407]

6.2.4 *Generating the phylogenetic tree:*

A robust maximum likelihood (ML) tree was generated from the MSA of 3129 sequences, using IQ-Tree v1.4.2 [487]. Since, the data-set comprises of distant homologs with no *a priori* information regarding their evolutionary rates, using an ML based method is fitting. IQ-Tree provides many advanced options to optimize the tree and therefore many options were supplied to the IQ-Tree command. Following options were provided with the command:

Amino acid substitution matrix	`-m GPCRtm+F+R10`
Additional protein structure matrices	`-madd EHO,EX2,EX3,UL2,UL3,EX_EHO,LG4X`
Tree refinement options	`-wbtl, -bb 1000, -abayes, -con`

All the default substitution matrices available in IQ-Tree like Dayhoff, JTT, DCMut, Poisson, WAG, etc, are optimized on globular proteins. Shortly after the first JTT matrix in 1992, Jones et al published another matrix in 1994 that was based on transmembrane proteins and showed that the substitution rates are different than the classical Dayhoff matrices [488]. Therefore, a new substitution matrix based on class A GPCR specific substitution rates, GPCRtm, was used. The GPCRtm had been shown to outperform JTT-tm matrix [489].

The frequency rate change of amino acids for TM proteins can vary significantly, given the low sequence identity among GPCRs. Thus, an empirical calculation of frequencies from the data (F) was selected. The model for the rate of heterogeneity was selected by using IQ-Tree's ModelFinder utility with (-m MFP) option. Therefore, before selecting the heterogeneity rate, a test run was performed on the data and a free rate of heterogeneity across sites (R) was suggested. R_{10} signifies a free rate model with 10 sites (at given time) being allowed to evolve at different rates. The selection of the rate category is a computationally expensive step for IQ-Tree and 10 is the maximum value available for large data-sets.

Additional important matrices that are derived from the protein structures were also given with option -madd. These matrices are based on structural properties like, extended, helix, other sites (EHO), 2 state and 3 state models for solvent accessibility: exposed and buried sites (EX2), and exposed, intermediate and buried sites (EX3). Ex_EHO combines the EHO and EX2 models while UL are the unsupervised trained models of EX2 and EX3. All the matrices might not be used during tree assembly, however in case of TM proteins these matrices can be of assistance.

Finally, the ML tree was bootstrapped with 1000 steps. Ultrafast option (-bb) was selected given the computationally exhaustive process otherwise. IQ-Tree was also commanded to write the bootstrapped tree along with individual branch lengths. This tree will be used for final visualization. Another statistical measure, Bayes probability test (-abayes) was also used to test individual branches along with bootstraps. These provide confidence to the tree topology.

6.2.5 *Tree visualization*:

The bootstrapped unrooted tree was visualized using interactive Tree of Life (iTOL) [409]. The topology was changed to circular and scaled according to the branch lengths for better visualization. The tree topology was analyzed for the clade organizations. In-built iTOL utilities were exploited to isolate different taxa based on branch lengths while color-marking CCRs, CXCRs, ACKRs and others. All the nodes having branch lengths < 0.05 were collapsed for improved visualization. The collapsed nodes are shown as circles proportional to the size of each node. Labels are shown at the tree circumference.

6.3 **Results and Discussions**

6.3.1 *Composition of the dataset*:

The raw dataset obtained from HMMer searches contained 3810 sequences that were annotated as a predicted chemokine receptor or GPCR, or uncharacterized, or hypothetical proteins, or synthetic proteins. Contemporary annotations are mostly automated and thus can be potentially misleading [490,491]. Therefore, sequences were not removed from the dataset based on their annotations.

Initially there were 3810 sequences that were annotated as a predicted chemokine receptor. 862 sequences were annotated as hypothetical proteins while 33 sequences had titles with “synthetic” keywords in them. After refinement of the first stage, based on Hmmer domain wise e-values and bias to score ratio, 892 sequences that annotated as ‘predicted’, ‘unnamed’, and ‘partial’ were removed. 552 of 862 ‘hypothetical’ proteins were removed while only 4 synthetic proteins were left. When the rejected proteins were analysed, their average length was 134 residues which directly translates to the reason for their high domain e-values and high bias. Post clustering, the number of such sequences was reduced to 2029; of which 19 are annotated as unnamed while 57 were labelled as ‘hypothetical’. Therefore, 1953 proteins with annotations containing ‘predicted’ proteins were selected in a dataset of 3129 sequences. This amounts to ~62% of the dataset, thus giving away the state of the contemporary automated annotation methods.

6.3.1.1 *Viral Chemokine Receptors*:

The initial dataset contained 1074 viral chemokine receptor hits. However, 1051 were rejected based on short lengths and could not pass the first stage of filtering. The rejected receptors mostly belonged to *Equine herpesvirus* (EPV), *Sheep poxvirus* (CPV), and *Fowl poxvirus* (APV) while most of the Human pox and herpes virus (like HCMV, HHPV) were among the 23 receptors selected. Post clustering, these were reduced to 7 sequences as 9 were rejected as singletons. The singletons also belonged to *Human Epstein Barr virus* (BILF-EBV), *Equine* EPV, *Capri* CPV and *Aves* APV classes. The BILF receptor from EBV was investigated for its rejection and it was found out that it is a non-chemokine receptor encoded in the virus genome. Therefore, it’s rejection as a singleton can be justified.

The final 7 viral chemokine receptors belonged to *Yaba-like disease virus*, *Yaba monkey tumor virus*, *Tanapox virus*, *Human cytomegalovirus* (HCMV), and *Swine poxvirus*. Except HCMV, rest all belong to poxviridae family. Two receptors from HCMV (beta herpesvirus), unique short 28 (US28) and unique long 33 (UL33) were present in the final dataset.

6.3.2 *Multiple Sequence Alignment*

Before proceeding with the generation of the phylogenetic tree, it is logically fitting to validate the MSA on which the tree will be built. The alignment has huge gaps especially at the terminal regions. Such gaps are well expected given the known diversity of N- and C-terminal regions among class A GPCRs, decoy and viral chemokine receptors. The terminals do not have a previously known conserved position except for the C-terminal microswitch containing motif, NPxxY. However, no conserved positions are observed downstream NPxxY and therefore these regions are removed in consideration of the loss of information to the noise they generated. Similarly, the N-terminal positions are highly variable and no highly conserved column is observed before the first cysteine residue in the sequence of the seed chemokine receptors. This corresponds to the 45th to 55th residues in the human chemokine receptor, given the variable length of their N-terminal deformed region. The high conservation of the first cysteine is supported by its role in the important disulfide bridge formation with TM7. The SS bridge may reduce the flexibility of the N-terminal loop and thus forming a binding pocket for the chemokine ligand.

The rest of the MSA also have some enormous gaps and the TM boundaries are also not completely conserved. This is also expected given the average identity of class A GPCR family is ~26%. Yet the important residues such as most of the cysteine, proline, tyrosine, tryptophan threonine, aspartate, and asparagine residues are found to be highly conserved if not completely conserved, as shown in Figure 6.7. This finding validated the MSA as all these residues will be under selective evolutionary pressure given their important role in the function. Apart from these, the motifs containing microswitches, including TxP, are also well conserved. However, these motifs have a single or double insertions in them. As this is further investigated, it is found that these insertions are contributed by a single sequence in each case as can be seen in Figure 6.7 consensus row. For instance, in motif NPxxY, there are two gaps between the 'x' and 'x' making the motif as NPx--xY. The insertion includes a glycine (G) and tyrosine (Y) residue from a predicted GCPR35 receptor in nine banded Armadillo (NCBI Acc: XP_012378710.1). It is also noticed that the motif is not at all conserved in the GCPR35 sequence, where it reads: DAxGYxY instead of NPxxY. Therefore, it can be safely considered as an artifact especially when removing such sequences reveal absolutely conserved motifs like NPxxY, CWxP, TxP. The most validating observation about the MSA is very high or near complete conservation of DRY motif with D, R, and Y occuring more than 90% of times at the aligned position. Moreover, the DRY motif is completely absent in all the ACKR1 sequences and have substitutions like E, K, and F (respectively) in some viral and decoy receptors. The conservation profiles of these important regions is shown in logo in consensus row and quality row in Figure 6.7.

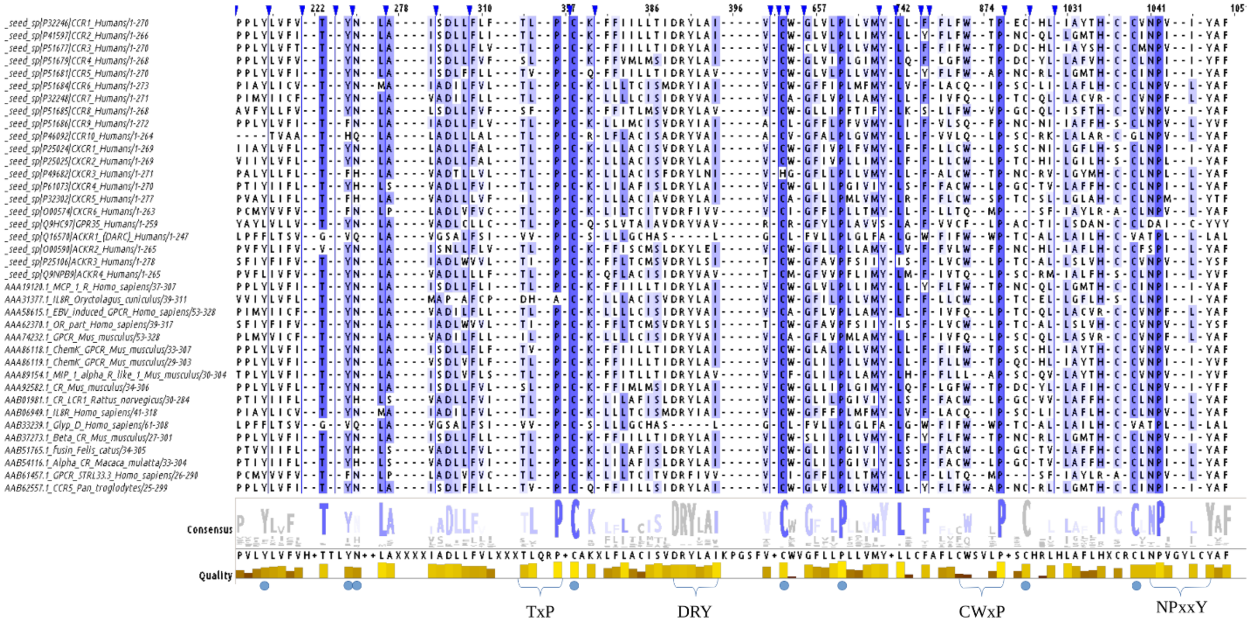


Figure 6.7 ***MSA used for tree generation.*** *The multiple sequence alignment of 3129 homologs of human chemokine receptors. The intensity of blue color signifies the extent of conservation of a single amino acid. The blue ticks on the top denotes the collapsed alignment between two points. The overall conservation is depicted as consensus logo along with the quality index of alignment. Conservation of important microswitch motifs is shown. The blue dots signify high conservation of important residues for chemokine receptors, e.g. Proline.*

6.3.3 <u>Tree topology</u>

The circular tree is referentially rooted at CCR1 as shown in the Figure 6.8. Overall the tree topology is in accordance with the classical nomenclature scheme of chemokine receptors. Most of the CCR receptors makes a single super-clade while most of CXCRs also lie in clades adjacent to one another. The ACKRs, however, do not form isolated clades, rather are found to be placed according to their old names. For instance, ACKR2 is taxonomically related to CCR super-clade containing CCR6, CCR7, CCR9, and CCR10 while ACKR4 is placed next to CXCR6. ACKR3 is phylogenetically most related to GPR35 which have been characterized as CXCR8 and interestingly the old name of ACKR3 is CXCR7. These observations indicate the selective evolutionary pressures on the chemokine receptors based on their functions, i.e, binding to chemokines. The classical nomenclature of chemokine receptors is based on the type of chemokines that bind to a receptor. The phylogenetic placement of these receptors seems to follow the trend. For instance, the clade of CXCR2 has high overlaps with the adjacent clade of CXCR1 so much so that it is difficult to consider CXCR2 as a separate clade, see Fig 6.7. It is also known that CXCL1 (IL-8) can bind to both CXCR1 as well as CXCR2 with different affinities [444]. Moreover, the ligand pool of CXCR1 and CXCR2 consists of CXCL2, CXCL1 and CXCL8; the genes of these chemokines lie on Human chromosome 4. Similar overlaps are also observed in the clades of CCR2 and CCR5. While both CCR2 and CCR5 have exclusive set of ligands but their encoding genes are located on Human chr 17. Besides, all the ligands in ligand pool of CCR2 and CCR5 have macrophage regulatory function. Therefore, further validations on their chemokine signaling can propose a merging of the two chemokine receptors. A caveat of proposing it from the current study is that the chemokine signaling of all the chemokine receptors have to be analysed in context of their individual species.

6.3.3.1 *Evolutionary placement of ACKR1*

ACKR1 forms a distant isolated clade placed at extremely large branch length of more than 7 units from the root. The closest clade to ACKR1 is GPR35 (or CXCR8) at 5.4 branch units from the root (as seen in Fig 6.8). Therefore, ACKR1 is highly distant from rest of the tree with a distance of 4.7 units from its branching node. The super-clade (sharing the same node) of ACKR1 consists of three clades ACKR1, GPR35, and ACKR3 and can be represented as (ACKR1(GPR35, ACKR3)). The clade of CXCR4 lies adjacent to this super-clade sharing the parent node with it. It can be represented as ((ACKR1(GPR35, ACKR3)) CXCR4). Since, no crystal structure is available for either GPR35 or ACKR3, CXCR4 becomes the closest clade that have its molecular structure determined, experimentally. This observation supports the choice of CXCR4 as a potential structural template for modelling ACKR1 structure. The same observation was also made from the phylogenetic tree generated with 21 seed sequences.

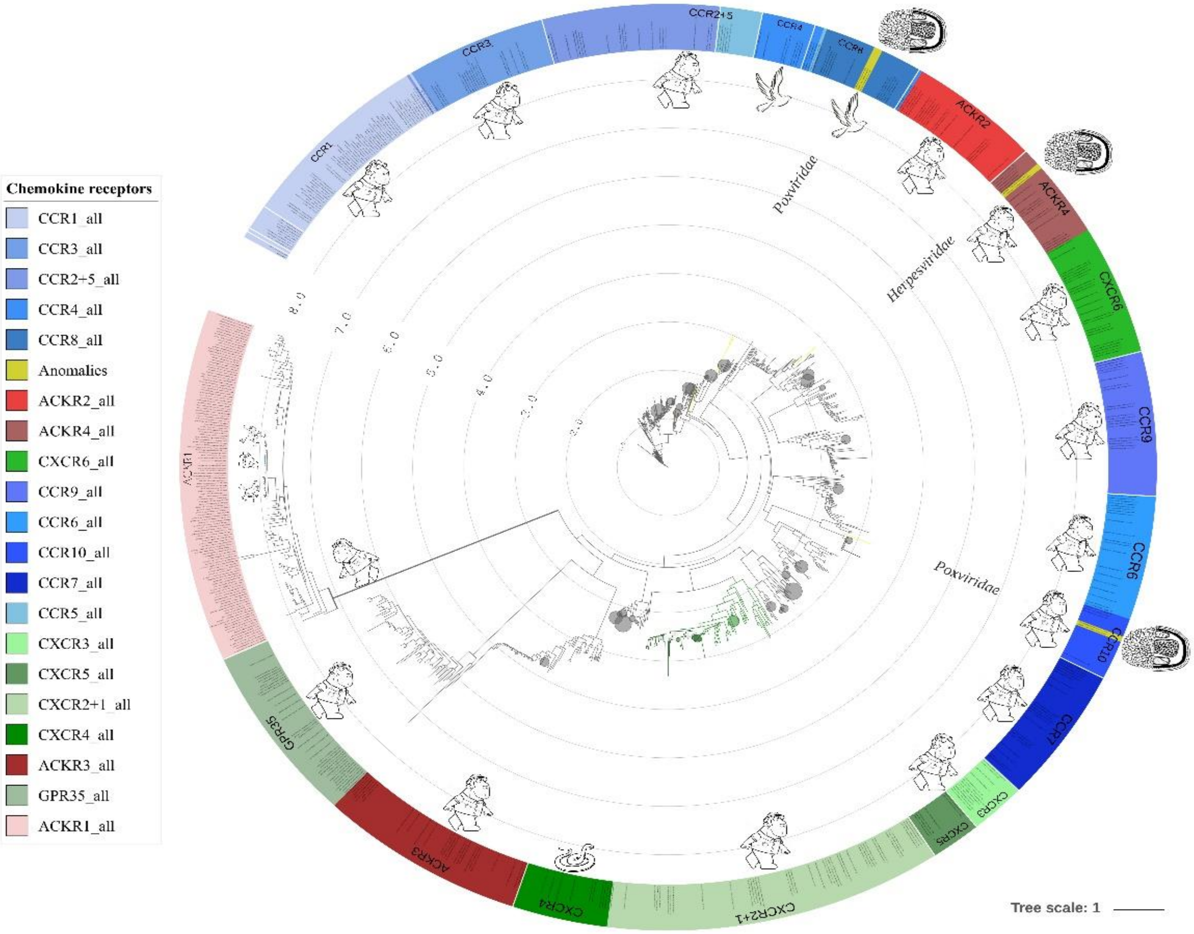


Figure 6.8 ***Evolutionary perspective on chemokine receptors.*** *The tree topology generated from IQtree for 3129 chemokine receptor sequences. The different clades are colored under the shades of same color. For instance, all CCR have shades of blue while CXCR and ACKR have shades of green and red, respectively. The clade distances are based on the bootstrapped branch lengths. Most populous species from each clade is shown as a cartoon. The walking hippo denote Mammals. The ACKR1 clade is expanded to fullest to show the location of primates in the tree. Yellow coloured branches signify the presence of viral chemokine receptors. They form outlier groups in each of the CCR8, ACKR4, and CCR10 clades.*

6.3.3.2 *Occurence of Viral encoded chemokine receptors*

As discussed before, in the section 6.3.1.1 that there are 7 viral chemokine receptors that are included for the MSA. vChemR are guided by evolutionary pressures different from those on chemokine receptors and therefore, it is interesting to see their placement in the tree. Rather than forming a separate clade, the vChemR are found in the clades of chemokine receptors like CCR8, CCR10 and ACKR4 placed as outliers in their respective clades. Yatapox viruses like Tanapox virus (TPV), Yaba-like disease virus (YDV), and Yaba monkey tumour virus (YTC) are clustered in the same clade located in the CCR8 clade. TPV and YDV are immediate neighbors sharing the same branch as shown in Figure 6.8. While, YTV branches off just before the branching node of TPV, YDV forming a representation as (YTV (TPV, YDV)). The clade is placed at a distance of 1.8 units from the average distance of CCR8 members from the root thus making it an outlier to the group. vChemR belonging to Human Cytomegalovirus (HCMV) US28 and UL33 are located within the clade of ACKR4 at a distance of 0.8 away from the average distance of ACKR4 from the root. The phylogenetic distance between US28 and UL33 is 0.24 units while the distance to their closest ACKR4 neighbor is 0.25. The other pox viruses, i.e, the swinepox virus are found in the CCR10 clade occupying a single isolated branch. The branch is 0.5 units away from the average distance of CCR10 members from the root and 0.05 units away from closest CCR10 neighbors. The two swinepox viruses are present very close to each other at a distance of 0.005 units.

The viruses in the CCR8 clade are quite distant from rest of the clade while the HCMV and SPV in ACKR4 and CCR10 resp, are located at a minimal distance from their respective clades. Given that CCR10 previously included ACKR2 and the SPV forms a distinct clade in CCR10 could be an indication that vChemR are closer to decoy receptors than canonical receptors. This is also supported by the fact that most of the vChemR are known to function as scavengers, a property ACKR shares as well. More detailed analysis of ACKR and vChemR will be required to conclude such a hypothesis, especially when the studies on human and mouse chemokine receptors have shown otherwise [381,461,484].

6.3.4 *Taxonomic distribution of the clades*

The taxonomic contribution of the sequences in a phylogenetically scaled tree can reveal the evolution of the query protein/gene. Therefore, an analysis is performed by taxonomically identifying each sequence in the tree from its source organism and thereafter observing the stages in their evolution. Overall, the tree is populated by the taxonomic class Mammalia as majority while there are substantial number of chemokine receptors from classes Aves, Reptilia, and Fishes. Moreover, Mammalia is not the majority population in all the clades. Chemokine receptor sequences of CCR4 and CCR8 have predominantly Avian contribution and CXCR4 have majority of sequences originating from class Reptilia. Figure 6.9 provides information about the sequence contributions by different taxonomic classes to each clade. Also, Figure 6.8 depicts the taxonomic distribution by denoting a representative cartoon belonging to the most dominant taxa in each clade.

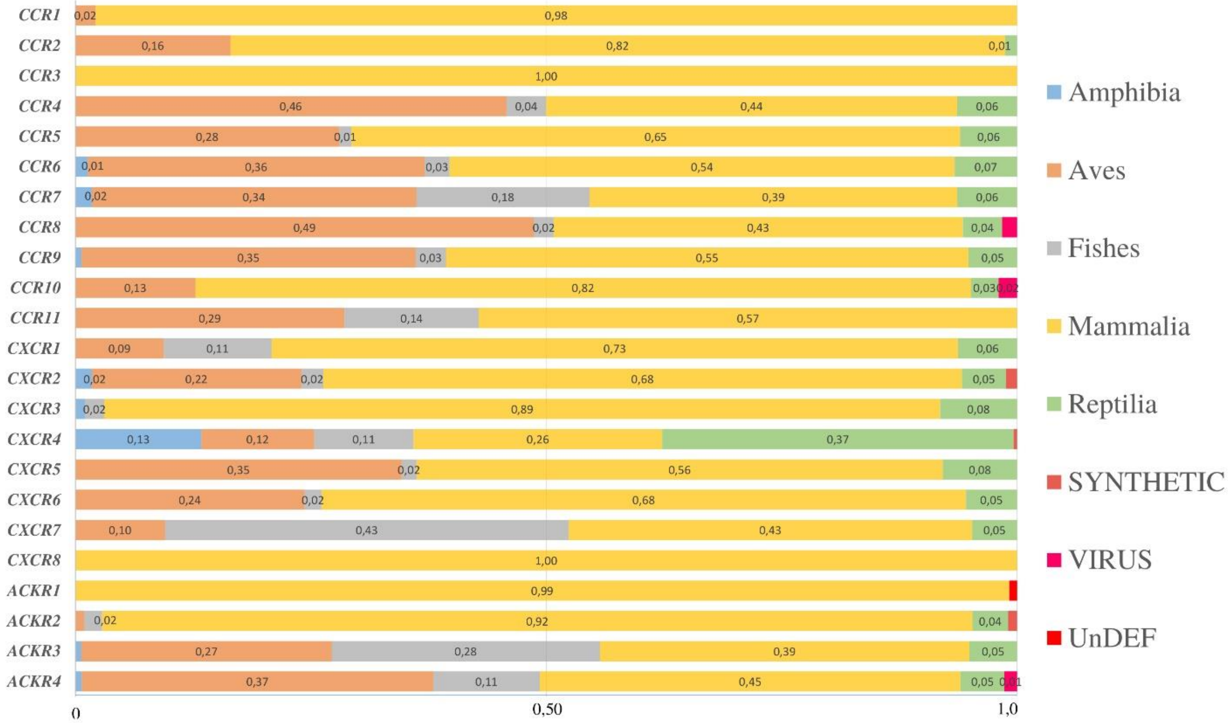


Figure 6.9 ***Taxonomic distribution of the different chemokine receptor clades.*** *Relative frequencies of different taxas is plotted on each clade. Mammalia is shown to be the most populist group. However, in CXCR7, CXCR4, CCR8, CCR4, dominant groups are Fishes, Reptilia, and birds respectively. ACKR1 has occurrence of a sequence that is annotated as "UnDefined protein". Few synthetic constructs are also observed in CXCR4 and CXCR2 clades. The values for CCR11 should be clubbed with those of ACKR2.*

In the clade CCR4, the population of Aves chemokine receptor sequences (at 46%) is nearly succeeded by the mammalian chemokine receptors at 44%. While the clade has small number of sequences from class Fishes and Reptilia contributing to 4% and 6% to the CCR4 sequence pool. Similarly, in clades of CCR8, 49% of receptors are avian while 43% are from Mammalia and 8% from Amphibia, Fishes, and Reptilia combined. Sequence pool of CXCR4 has significant contributions from all the major taxonomic classes. Reptilia have dominating contribution to CXCR4 sequences with 37% succeeded by 26% from Mammalia. CXCR4 also have 13% of sequences belonging to Amphibia which is also the most dominant contribution from class Amphibia to a chemokine receptor clade. In other clades, Amphibia contributes only 1% to 2% to the sequences. Clade CXCR7 have an equal population of Fishes and Mammalia with 43%

of sequences belonging to each class. The class Mammalia completely dominates the sequence pool of CCR1, CCR2, CCR3, CCR10, CXCR3, CXCR8, ACKR1, and ACKR2 with more than 80% contribution to their sequences.

Mammalian species distribution in clades: Mammalia is either the major contributing or second major contributing class towards chemokine sequences in all the clades. It indicates that major evolution of chemokine receptors is confined to mammals and therefore it is required to explore it at the species level to understand the evolutionary trends. The sequences of mammalian chemokine receptor in the tree belong to a diverse set of mammalian species; from Bears (*Ailuropoda* and *Ursus*), to hedgehogs (*Erinaceidae*), to the exotic species of Pika (*Ochotona*), to Marsupials like Tasmanian devil (*Dasyuromorphia*), to egg laying mammal like Platypus (*Ornithorhynchus*), to Sea cow (*Sirenia*), to Whales, to Primates, to Apes and to Humans. Therefore, the mammalian chemokine receptors are diversified in almost all the genera of class Mammalia. Figure 6.10 shows the absolute contribution of different genera of class Mammalia to chemokine sequences of each clade.

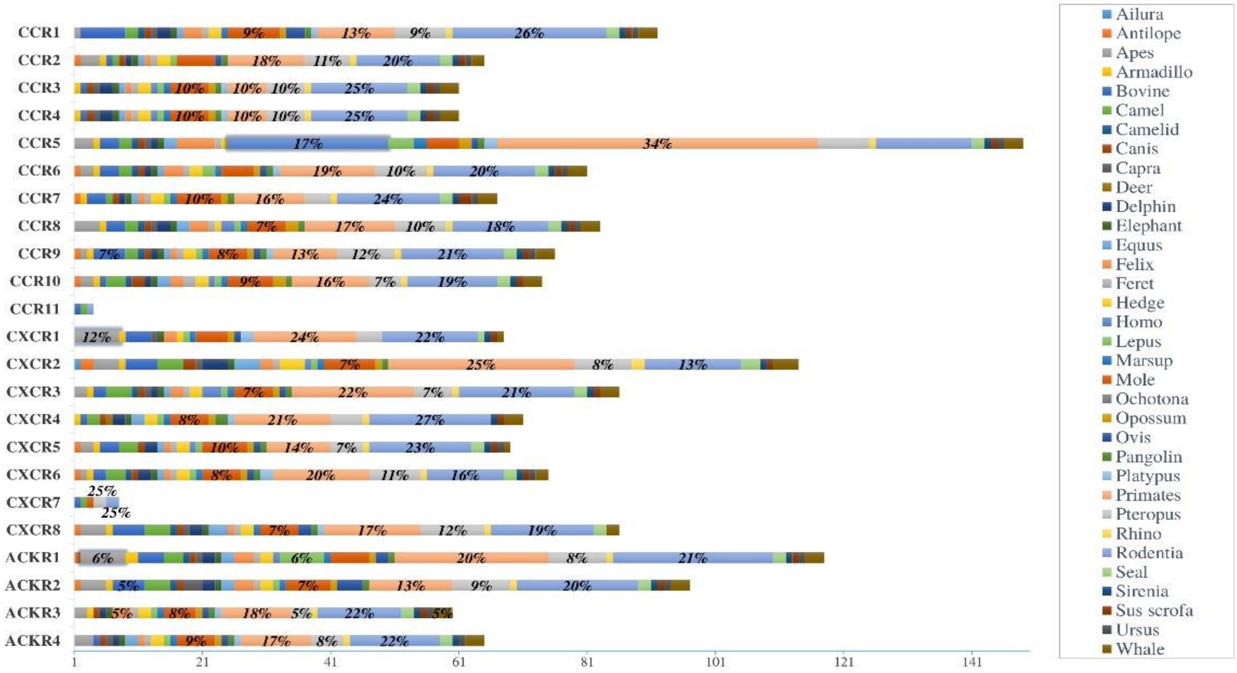


Figure 6.10 ***Mammalian contribution to different chemokine receptor sequences***. *Absolute distribution of mammalian species in different clades. Mostly, few genera dominate all clades. Such genera are those of: Primates, rodents, chiroptera (bats), and Moles. The occurrence of another dominant genus in a clade is signified by a highlighted border, for instance Humans have*

17% or second largest contribution to the CCR5 sequences. The values for CCR11 should be clubbed with those of ACKR2.

Although the diversity in mammalian chemokine receptors is considerable yet there are some species that dominate others by their contribution to the sequence pool. Therefore, in almost all the clades, four major species are observed to have majority contribution to the chemokine receptor pool. Most of the sequences in each clade belong to Primates, Rodentia (*Muridae* family), Moles (*Talpidae* family), and Bats (order *Chiroptera*). These four species contribute towards 50% of the sequence pool in each clade. Also, overall the most contributing genera are: Rodentia with ~20% of sequences belonging to it and ~19% of receptor sequences originating from Primates. With 3.12% and 2.08% of sequences belonging to Apes and Humans, the *hominidae* contribution to the chemokine receptor sequences amounts to 5.2%.

In CCR1, 26% of sequences are contributed by Rodentia while 13% belong to Primates and 9% from Bats (*Chiroptera*) and Moles each, thus comprising more than 50% of the clade population. Similar trend is observed in other clades with slight exceptions in CCR5, CXCR1, CXCR2 and ACKR1 where some other species contributes significantly, if not equally to the clade's receptor pool. For instance, in CCR5 17% of sequences belong to *homo sapiens*, preceded by 34% of sequences originating from Primates. Also, it is the highest contribution of *homo sapiens* to any clade in the tree. These statistics from CCR5 are indicative of the number of studies performed on CCR5 possibly due its involvement in HIV-1 infection. On the contrary, there only two sequences belonging to Humans in ACRK1, one of which happens to be the query seed sequence. Similarly, a single human chemokine receptor is observed in CCR1, CCR2, CCR6, CCR7, CCR9, CXCR1, CXCR4, CXCR6 through CXCR8, and all atypical receptors, except ACKR1. These statistics are indicative of the conservation in the human chemokine receptors and different evolutionary pressures on Human CCR5 and ACKR1.

6.3.5 *Tracing evolution of ACKR1*

In clade of ACKR1, the phylogenetic clade is mapped with the information of the source species to understand the evolution of ACKR1. Figure 6.11 shows the mapped species, as cartoons, on the clade of ACKR1 along with their taxonomic grouping (numbered circles).

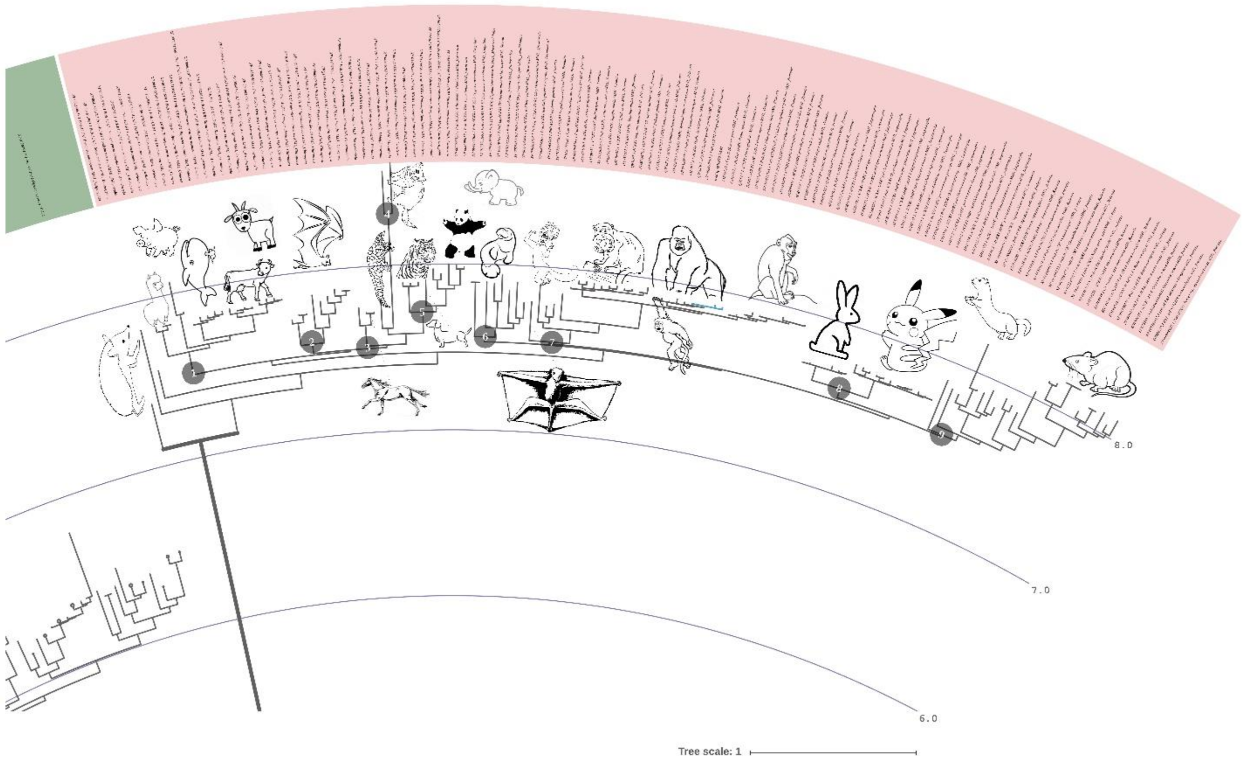


Figure 6.11 ***Topology of the ACKR1 clade.*** *Figure shows the phylogenetic relationships among class Mammalia for ACKR1. Taxonomic orders are depicted as cartoon of its most known species. The branch distance from the root can be estimated using the radiating scale values given on the right side. The numbers shown on the nodes signifies that the nodes after that can be collapsed under same order. For eg. [1] corresponds to superorder Cetartiodactyla, [2] will collapse all bats species under chiroptera. Although the ACKR1 clade has a distance of more than 7.0 from the root, yet the whole clade, except marsupials, have evolved within 1.0 units.*

Although the clade is phylogenetically remote to rest of the chemokine receptors the branches within the clade are very closely related to each other. If the clade is collapsed at 0.05 branch lengths, more than half of the clade is clustered. Moreover, as discussed in the sections above that Mammalia is the dominant population and contains enormous diversity within it. Thus, it was difficult to understand the evolutionary history of ACKR1 just by mapping the taxonomic information. Therefore, the clades were collapsed under their taxonomic orders to simplify the visualization of, otherwise very closely related sub-clades, Figure 6.12A. One taxonomic order is expanded each time to analyse the evolutionary progress, Fig 6.12[B to D]. The branch lengths are converted to evolutionary ages by subtracting individual branch lengths from the root distance. This will reveal the youngest leaf with an age of 0 and closer the leaf gets to the root, the age value increases. Figure 6.12A depicts the phylogram focused on the ACKR1 clade extracted from the whole tree. All leaves are shown collapsed at their respective taxonomic orders, except *hominidae*. For instance, all the chemokine receptor sequences belonging to camels, cattle, whales, horses, sheep, pig, and other even-toed ungulates are collapsed together under Cetartiodactyla taxonomic order. Some of the leaves cannot be clustered given their taxonomic placement in the clade. For instance, *Dermoptera* also known as the flying lemurs lies as a single node linking Primates to order Afrotheria. The sub-clade of *hominidae* is not collapsed because it contains the seed ACKR1 sequence and can serve as a reference point for tracing ACKR1's evolution.

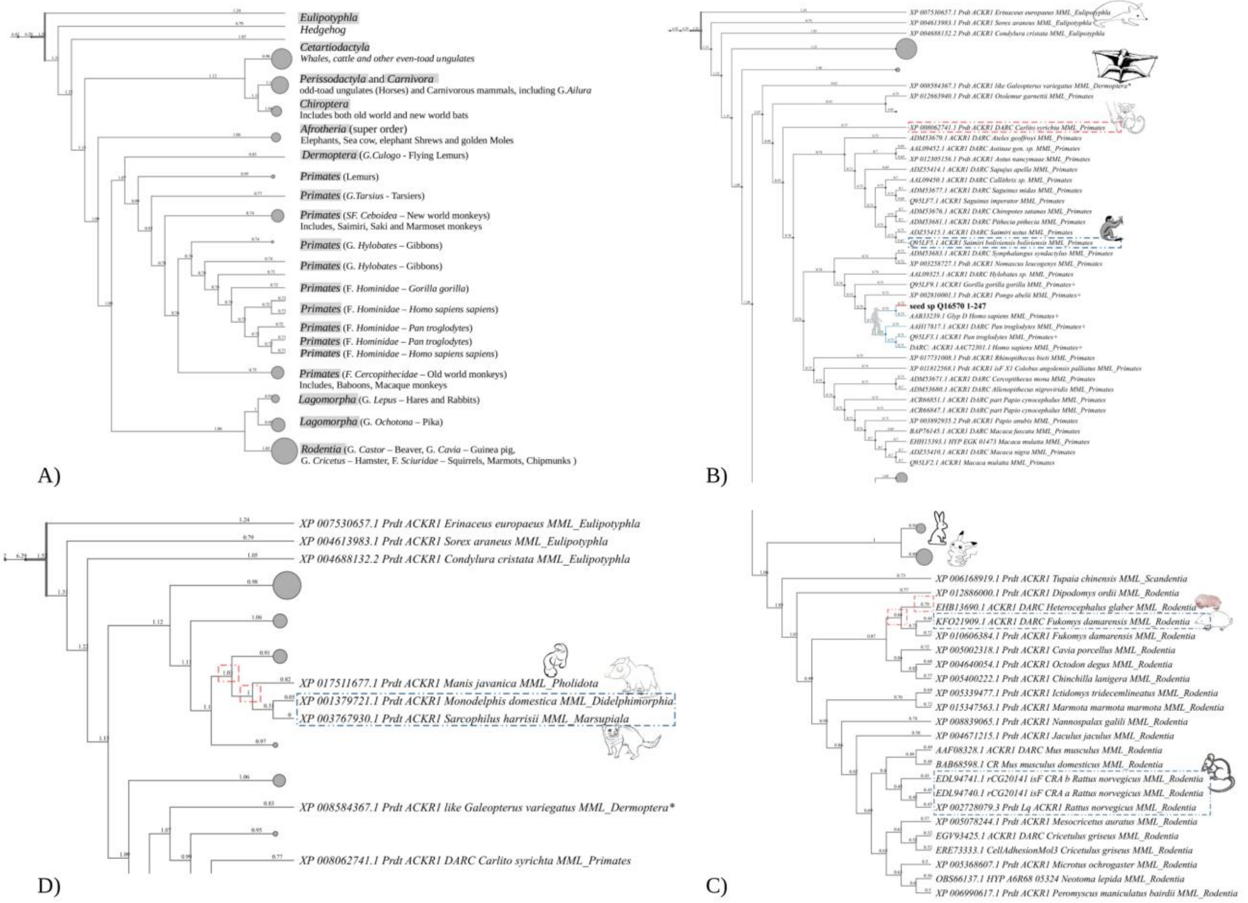

Eulipotyphla
Hedgehog
Cetartiodactyla
Whales, cattle and other even-toad ungulates
Perissodactyla and Carnivora
odd-toad ungulates (Horses) and Carnivorous mammals, including G.Ailura
Chiroptera
Includes both old world and new world bats
Afrotheria (super order)
Elephants, Sea cow, elephant Shrews and golden Moles
Dermoptera (G.Culogo - Flying Lemurs)
Primates (Lemurs)
Primates (G.Tarsius - Tarsiers)
Primates (SF. Ceboidea – New world monkeys)
Includes, Saimiri, Saki and Marmoset monkeys
Primates (G. Hylobates – Gibbons)
Primates (G. Hylobates – Gibbons)
Primates (F. Hominidae – Gorilla gorilla)
Primates (F. Hominidae – Homo sapiens sapiens)
Primates (F. Hominidae – Pan troglodytes)
Primates (F. Hominidae – Pan troglodytes)
Primates (F. Hominidae – Homo sapiens sapiens)
Primates (F. Cercopithecidae – Old world monkeys)
Includes, Baboons, Macaque monkeys
Lagomorpha (G. Lepus – Hares and Rabbits)
Lagomorpha (G. Ochotona – Pika)
Rodentia (G. Castor – Beaver, G. Cavia – Guinea pig,
G. Cricetus – Hamster, F. Sciuridae – Squirrels, Marmots, Chipmunks)
A)
XP 007530657.1 Prdt ACKR1 Erinaceus europaeus MML_Eulipotyphla
XP 004613983.1 Prdt ACKR1 Sorex araneus MML_Eulipotyphla
XP 004688132.2 Prdt ACKR1 Condylura cristata MML_Eulipotyphla
XP 008584367.1 Prdt ACKR1 like Galeopterus variegatus MML_Dermoptera*
XP 012663940.1 Prdt ACKR1 Otolemur garnettii MML_Primates
XP 008062741.1 Prdt ACKR1 DARC Carlito syrichta MML_Primates
ADM53679.1 ACKR1 DARC Ateles geoffroyi MML_Primates
AAL09452.1 ACKR1 DARC Aotinae gen. sp. MML_Primates
XP 012305156.1 Prdt ACKR1 Aotus nancymaae MML_Primates
ADZ55414.1 ACKR1 DARC Sapajus apella MML_Primates
AAL09450.1 ACKR1 DARC Callithrix sp. MML_Primates
ADM53677.1 ACKR1 DARC Saguinus midas MML_Primates
Q95LF7.1 ACKR1 Saguinus imperator MML_Primates
ADM53676.1 ACKR1 DARC Chiropotes satanas MML_Primates
ADM53681.1 ACKR1 DARC Pithecia pithecia MML_Primates
ADZ55415.1 ACKR1 DARC Saimiri ustus MML_Primates
Q95LF5.1 ACKR1 Saimiri boliviensis boliviensis MML_Primates
ADM53683.1 ACKR1 DARC Symphalangus syndactylus MML_Primates
XP 003258727.1 Prdt ACKR1 Nomascus leucogenys MML_Primates
AAL09325.1 ACKR1 DARC Hylobates sp. MML_Primates
Q95LF9.1 ACKR1 Gorilla gorilla gorilla MML_Primates+
XP 002810001.1 Prdt ACKR1 Pongo abelii MML_Primates+
seed sp Q16570 1-247
AAB33239.1 Glyp D Homo sapiens MML_Primates+
AAH17817.1 ACKR1 DARC Pan troglodytes MML_Primates+
Q95LF3.1 ACKR1 Pan troglodytes MML_Primates+
DARC: ACKR1 AAC72301.1 Homo sapiens MML_Primates+
XP 017731008.1 Prdt ACKR1 Rhinopithecus bieti MML_Primates
XP 011812568.1 Prdt ACKR1 isF X1 Colobus angolensis palliatus MML_Primates
ADM53671.1 ACKR1 DARC Cercopithecus mona MML_Primates
ADM53680.1 ACKR1 DARC Allenopithecus nigroviridis MML_Primates
ACR66851.1 ACKR1 DARC part Papio cynocephalus MML_Primates
ACR66847.1 ACKR1 DARC part Papio cynocephalus MML_Primates
XP 003892935.2 Prdt ACKR1 Papio anubis MML_Primates
BAP76145.1 ACKR1 DARC Macaca fuscata MML_Primates
EHH15393.1 HYP EGK 01473 Macaca mulatta MML_Primates
ADZ55410.1 ACKR1 DARC Macaca nigra MML_Primates
Q95LF2.1 ACKR1 Macaca mulatta MML_Primates
B)
XP 007530657.1 Prdt ACKR1 Erinaceus europaeus MML_Eulipotyphla
XP 004613983.1 Prdt ACKR1 Sorex araneus MML_Eulipotyphla
XP 004688132.2 Prdt ACKR1 Condylura cristata MML_Eulipotyphla
XP 017511677.1 Prdt ACKR1 Manis javanica MML_Pholidota
XP 001379721.1 Prdt ACKR1 Monodelphis domestica MML_Didelphimorphia
XP 003767930.1 Prdt ACKR1 Sarcophilus harrisii MML_Marsupiala
XP 008584367.1 Prdt ACKR1 like Galeopterus variegatus MML_Dermoptera*
XP 008062741.1 Prdt ACKR1 DARC Carlito syrichta MML_Primates
D)
XP 006168919.1 Prdt ACKR1 Tupaia chinensis MML_Scandentia
XP 012886000.1 Prdt ACKR1 Dipodomys ordii MML_Rodentia
EHB13690.1 ACKR1 DARC Heterocephalus glaber MML_Rodentia
KFO21909.1 ACKR1 DARC Fukomys damarensis MML_Rodentia
XP 010606384.1 Prdt ACKR1 Fukomys damarensis MML_Rodentia
XP 005002318.1 Prdt ACKR1 Cavia porcellus MML_Rodentia
XP 004640054.1 Prdt ACKR1 Octodon degus MML_Rodentia
XP 005400222.1 Prdt ACKR1 Chinchilla lanigera MML_Rodentia
XP 005339477.1 Prdt ACKR1 Ictidomys tridecemlineatus MML_Rodentia
XP 015347563.1 Prdt ACKR1 Marmota marmota marmota MML_Rodentia
XP 008839065.1 Prdt ACKR1 Nannospalax galili MML_Rodentia
XP 004671215.1 Prdt ACKR1 Jaculus jaculus MML_Rodentia
AAF08328.1 ACKR1 DARC Mus musculus MML_Rodentia
BAB68598.1 CR Mus musculus domesticus MML_Rodentia
EDL94741.1 rCG20141 isF CRA b Rattus norvegicus MML_Rodentia
EDL94740.1 rCG20141 isF CRA a Rattus norvegicus MML_Rodentia
XP 002728079.3 Prdt Lq ACKR1 Rattus norvegicus MML_Rodentia
XP 005078244.1 Prdt ACKR1 Mesocricetus auratus MML_Rodentia
EGV93425.1 ACKR1 DARC Cricetulus griseus MML_Rodentia
ERE73333.1 CellAdhesionMol3 Cricetulus griseus MML_Rodentia
XP 005368607.1 Prdt ACKR1 Microtus ochrogaster MML_Rodentia
OBS66137.1 HYP A6R68 05324 Neotoma lepida MML_Rodentia
XP 006990617.1 Prdt ACKR1 Peromyscus maniculatus bairdii MML_Rodentia
C)

Figure 6.12 ***Tracing ACKR1 evolutionary developments.*** *The branch lengths of the ACKR1 clade is converted to age values (from the root). The larger the value, more ancient is the branch leaf. A) shows the evolution of ACKR1 in order primates with respect to other taxonomic families. B) shows the detailed evolution of ACKR1 protein in primates. The youngest species that acquired ACKR1 is highlighted in blue dotted lines and its representative is shown as cartoon. The most ancient species is marked by red dotted lines. Squirrel monkeys (new world monkeys) are the youngest while Tamarins are the oldest in primate sub-clade. C) Similar analysis for Rodentia.*

Rats and Damaraland mole rat have acquired ACKR1 recently. Interestingly, the mole-rat branch is accompanied by the eldest species in the clade, naked mole rat. D) Age wise analysis of the youngest species in the ACKR1 clade.

It is observed from the branch ages that the ACKR1 sequence originating from the Marsupial species of Tasmanian devil (*sarcophilus harrisii*) and Opossum (*monodelphis domestica*) are the youngest with an age value of 0 and 0.5, respectively. Please refer to Figure 6.12D. ACKR1 sequence from the *Eulipotyphla* species (Hedgehogs) is the oldest and given the clade topology the common ancestor of *Eulipotyphla* and rest of the mammalia is the oldest link traceable to ACKR1 evolution (age 1.5). ACKR1 seems to have first evolved from Hedgehogs to non-primate mammals like species from orders Carnivora and Rodentia. Based on the branch ages, the delayed transfer of ACKR1 from Eulipotyphla to other clades (nodal age difference = 0.37) is accompanied by a rapid speciation events. ACKR1 sequences from genera *Equus* (Horse), *Lepus* (Rabbit) and Ochotona (Pika) have similar ages of the order of 1.0 to 0.9 thus indicating a parallel evolution, see Figure 6.12. ACKR1 is later found in genera *Camelus* (Camel), *Pteropus* (old world bats) and marks its entry in primates with first occurrences in *Dermaptera* species (flying lemurs) and then in Lemurs from genus *Propithecus* and *Microcebus, see* Fig 6.12B. The reference age of these species lies between 0.90 to 0.80. Once, ACKR1 sequence have proliferated the primates, it soon enters the *Hylobates* species and the *Platyrrhini* (new world monkeys) while evolving parallelly from *Lagomorpha* to *Heterocephalus* (Mole-rats), Figure 6.12[B,C]. The node ages of Artiodactyla also shows tha ACKR1 sequence may have been evolving parallely in the sub-clade with bovine ACKR1 sequences having an age of 0.79. Evolution of ACKR1 is seen simultaneously in many order like Rodentia, Artiodactyla and Chiroptera (bats) having node ages of 0.75.

Thereafter, the ACKR1 homologs in Hylobates may have speciated to old world monkeys like *Cercopithecus* and *Rhinopithecus,* as shown in Figure 6.12B. Within an age difference of 0.03 ACKR1 homologs are observed in all old world monkeys and all the great apes of *hominidae*. This might be an example of parapatric or sympatric speciation of ACKR1 gene leading to its evolution in four different species of hominids in a very short age difference of 0.01 (0.72 to 0.71). The ACKR1 homologs from genera *Canis* (Dogs) and *Felix* (cats) also have the same age as 0.71.

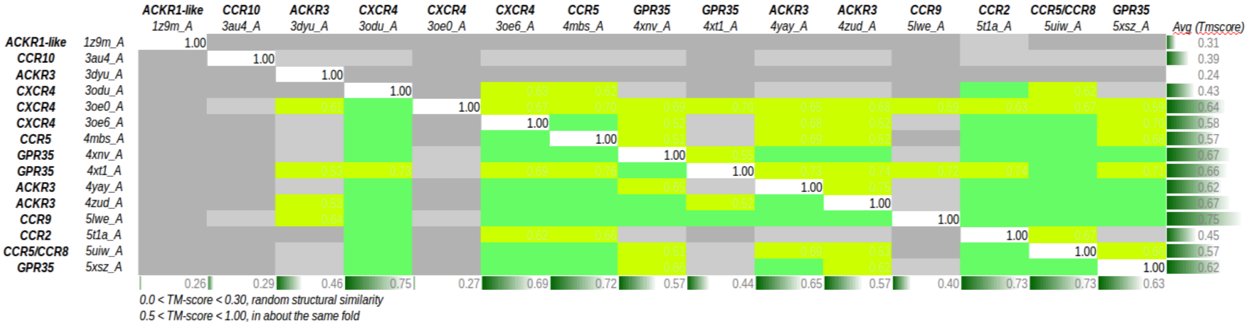
Meanwhile, the homologs of ACKR1 in *odobenidae* (walrus), *carnivora* (Weasels, Bears, and Pandas), *Loxodonta* (Elephant) and new world monkeys keep evolving further. The most recent advances in the speciation of ACKR1 has been shown in genera *Sus* (Pig), and many species of Rodentia. Amongst Rodentia (Fig **6.12C**), ACKR1 belonging to genera *Mus* and genera *Rattus* are shown to be the youngest with branch age of 0.49 and 0.45, respectively. Surprisingly, Marsupials (Tasmanian Devil and Opossum) seems to have acquired ACKR1 gene most recently with a branching age of 0.31, Figure 6.12D.

Although, the reference time of occurence of ACRK1 in Primates is comparatively later from the origin of ACKR1 in *mammalia* yet the speciation in primates is very rapid. Most of the topology of ACKR1 clade resembles that of the mammalian tree of life but with some critical exceptions, like placement of Eulipotyphla. If validated, such exceptions can give insights into the evolution of ACKR1 among Mammalia.

6.3.6 *Structural relatedness between clades*

The sequences of chemokine receptors have high amount of diversity with exception of a few functionally conserved regions. However, from a structural perspective the receptors share a common 7TM GPCR fold. Therefore, a quick protocol is designed to extract structural information from the tree. All the sequences in each clade was used as a query against the PDB database using Blastp. The resulting hits were sorted based on high query coverage and low e-value and the top hit was selected. It was performed in order to get one structural representative from each clade.

However, it is observed that only 12 unique PDB ids are represented as hits for the 3129 sequences. These PDB ids are: 1Z9M, 3AU4, 3DYU, 3ODU, 3OE0, 3OE6, 4MBS, 4NUV, 4XNV, 4XT1, 4YAY ,4ZUD, 5LWE, 5T1A, 5UIW, and 5XSZ. Of these 3OE6, 3ODU, and 3OE0 are crystal structures of CXCR4 and 4NUV is the structure of residues 19-30 of ACKR1. Therefore, 4NUV was rejected from the list. Since all the blast hits matched chain A of the PDB proteins, chain A of these 11 structures was extracted. From the blast results the PBDids were reverse mapped on to their query sequence and thence to their respective clades in the tree. At last, the clades that have a structural representative are: CCR2 (5T1A), CCR5 (4MBS), CCR8 (5UIW, *shared with CCR5*), CCR9 (5LWE), CCR10 (3AU4), CXCR4 (3ODU, 3OE6, 3OE0), GPR35 or CXCR7 (4XT1, 4XNV, 5XSZ), ACKR1 (1Z9M), and ACKR3 (3DYU, 4YAY, 4ZUD). The chain of the structures is aligned using TM-Align to assess their structural relatedness. Table 6.1: summarizes the closeness profile of these structures.



		ACKR1-like	CCR10	ACKR3	CXCR4	CXCR4	CXCR4	CCR5	GPR35	GPR35	ACKR3	ACKR3	CCR9	CCR2	CCR5/CCR8	GPR35	
		1z9m_A	3au4_A	3dyu_A	3odu_A	3oe0_A	3oe6_A	4mbs_A	4xnv_A	4xt1_A	4yay_A	4zud_A	5lwe_A	5t1a_A	5uiw_A	5xsz_A	Avg (Tmscore)
ACKR1-like	1z9m_A	1.00															0.31
CCR10	3au4_A		1.00														0.39
ACKR3	3dyu_A			1.00													0.24
CXCR4	3odu_A				1.00		0.63	0.62							0.62		0.43
CXCR4	3oe0_A			0.61		1.00	0.67	0.70	0.69	0.70	0.65	0.68	0.59	0.63	0.67	0.59	0.64
CXCR4	3oe6_A						1.00		0.52		0.68	0.52				0.70	0.58
CCR5	4mbs_A							1.00	0.51		0.69	0.52				0.68	0.57
GPR35	4xnv_A								1.00	0.55							0.67
GPR35	4xt1_A			0.53	0.73		0.69	0.76		1.00	0.73	0.74	0.72	0.74		0.71	0.66
ACKR3	4yay_A								0.65		1.00	0.75					0.62
ACKR3	4zud_A			0.53						0.52		1.00					0.67
CCR9	5lwe_A			0.64									1.00				0.75
CCR2	5t1a_A						0.62	0.66						1.00	0.67		0.45
CCR5/CCR8	5uiw_A								0.51		0.69	0.53			1.00	0.66	0.57
GPR35	5xsz_A								0.66			0.62				1.00	0.62
		0.26	0.29	0.46	0.75	0.27	0.69	0.72	0.57	0.44	0.65	0.57	0.40	0.73	0.73	0.63	

0.0 < TM-score < 0.30, random structural similarity
0.5 < TM-score < 1.00, in about the same fold

Table 6.1. Structural relatedness among different clades. *The TM-score profile of 18 PDB structures representing major clades of the tree. For many clades, no PDB structure was found. The PDB id and the chain id used is represented on both axes under the name of the clade it represents. These represent 9 clades from the tree. The gradient green, from light to dark indicates higher structural similarity or better TM-score. The intensity of grey shows least TM-score and no relation. Both row-wise and column-wise averages are calculated and represented as a dark green bars*.

PBDID: 1Z9M is the structure of a Nectin like cell adhesion molecule which does not have a 7TM structure. Therefore, 1Z9M cannot be treated as an ACKR1 representative. Similarly, 3AU4 also does not have a 7TM structure being a Netrin receptor involved in apoptosis. Netrin receptor has a large number of helices interspersed by two beta sheets but does not belong to 7TM GPCR family and therefore, it cannot be a representative of CCR10. However, according to blast results, these PDB ids are related to the ACKR1 and CCR10 clade and therefore kept in the analysis. These might be useful as a positive and negative control for structural alignment of 7TM using TM-Align.

As revealed by the green color in the Table 6.1 that all but ACKR1 and CCR10 are structurally related. The relatedness of the structures is given by the TM-Score; a value of 0.5 and above implies that the structures are related while a value below 0.3 reflects random structures. ACKR3 have very good structural similarity with CCR5, CXCR4, GPR35, CCR8 (*shared with CCR5*), and CCR2 with an average TM-Score of 0.72. CXCR4 have structure relatedness with CCR2, CCR8, and CCR5 with an average TM-Score of 0.63. GPR35 have an average score of 0.67 while CCR5 have an average score of 0.57. However, CCR2 have an average score of 0.49

when CCR2 is used as the reference but when rest of the structural representatives are aligned with CCR2 the average TM-Score is 0.73. Therefore, CCR2 is structurally related to only CXCR4 (score 0.86) but not with rest of the structures. However, they show high local similarity with the CCR2 structure when they are aligned to CCR2. Such an analysis is also important to assess the structural diversity of chemokine receptors.

6.4 **Conclusions and Perspectives**

Chemokine receptor family is comprised of a diverse set of sequences classified into various sub-groups like CCR-, CXC-, CX3C, or XC- Receptors. The current classification is based on the class of chemokine ligand binding to the receptor, therefore an α-chemokine binding receptor is named CXCR while a β-chemokine binding receptor is termed CCR. Such a premise, however may change with the enhancing knowledge about the hetero-oligomerization in chemokine receptors and discovery of more virally encoded chemokine receptors. Therefore, it may become necessary to understand the evolutionary perspective of chemokine receptors to assess their phylogenetic relations along side their functional relationships. The presented phylogenetic study is based on the most comprehensive data (till date) on chemokine receptors. The data-set contains 3129 sequences of chemokine receptors and few other class A GPCRs. Strict controls and filters have prevented the contamination of the dataset by unrelated or highly redundant sequences. The tree is also supported by a robust multiple sequence alignment founded on a seed alignment of 118 known chemokine receptors. Moreover, the MSA is also validated by the conservation of all the important functional and structural sites. These initial checks and validations accompanied by the use of GPCR specific substitution model for tree building enhances the confidence in the phylogram.

Strong evidence provided by the overlaps in the clades of CCR2 and CCR5 as well as CXCR1 and CXCR2 is suggestive of similar evolutionary pressures among the two pair of receptors. The similarity between the gene locations of their respective ligands as well their functional similarity further supports the merging of the two clades. However, detailed analysis of their ligand's genetic and biochemical profiles as well as receptor cross-talks during hetero-oligomerization have to be performed for confirming the hypothesis. Series of works by Zlotnik, Nomiyama, Yoshie, et al explores the genomic organization and evolution of chemokines going back to agnathan fishes [492–495]. The future intend of our project is to utilize such information to deduce parallels between evolution of chemokines and that of chemokine receptors. Such a study can inform us about the evolutionary pressures of chemokines on chemokine receptors and vice versa.

The current tree is rooted on CCR1 which is not an outlier to the phylogram and therefore may arise doubts. However, the rooting of the tree in this particular study does not matter because the principle question to address is the phylogenetic relationships among chemokine receptor and evolution of individual clades. Therefore, an unrooted tree can also be used instead of a rooted one. Given that ACKR1 is the most distant clade and thus can be treated as an outlier, based on the distance and not on function, a tree rooted on ACKR1 was also generated. The tree topology remained the same with few node rotations. Thus the current tree topology having each branch tested with bootstrap and bayesian probability has a high confidence value.

The great diversity among the sequences of chemokine receptors is complemented by the huge diversity of the species in the tree. The species information, in reference to the tree of life, is used to understand the evolution of chemokine receptors. One such deduction is carried out in the clade of ACKR1. The evolution of ACKR1 is traced by the virtue of their branch lengths converted to ages. The age of each node helps in identifying the first and recent occurence of ACKR1 in species and also in tracing the gaps in between. However, unfortunately there are no ACKR1 sequences identified outside mammalia yet and thus limits the access of our analysis. The absence of non-mammalian ACKR1 gene might be the reason for the huge distance of ACKR1 clade from rest of the clades that have sequences from Amphibia, Fishes as well as Reptilia.

However, there can be another explanation to the absence of ACKR1 in non-mammalian classes. A strong hypothesis might be that ACKR1 is not a chemokine receptor, based on its atypical behavior and non-specific binding to different chemokine classes. Such behaviors are noticed in some viral chemokine receptors too but ACKR1 differs from viral chemokine receptors also. The virally encoded receptors are in a constitutively active state while ACKR1, especially in reticulocytes does not signal or scavenge at all. Therefore, it might have been a viral chemokine receptor that was genetically pirated from the host during infection. During a subsequent infection by the virus (possibly a ssDNA or retrovirus) the altered receptor gene was back-pirated into the host genome. However, this is a mere speculation and would require rigorous analysis to test it; for which the current evolutionary tree will be highly useful.

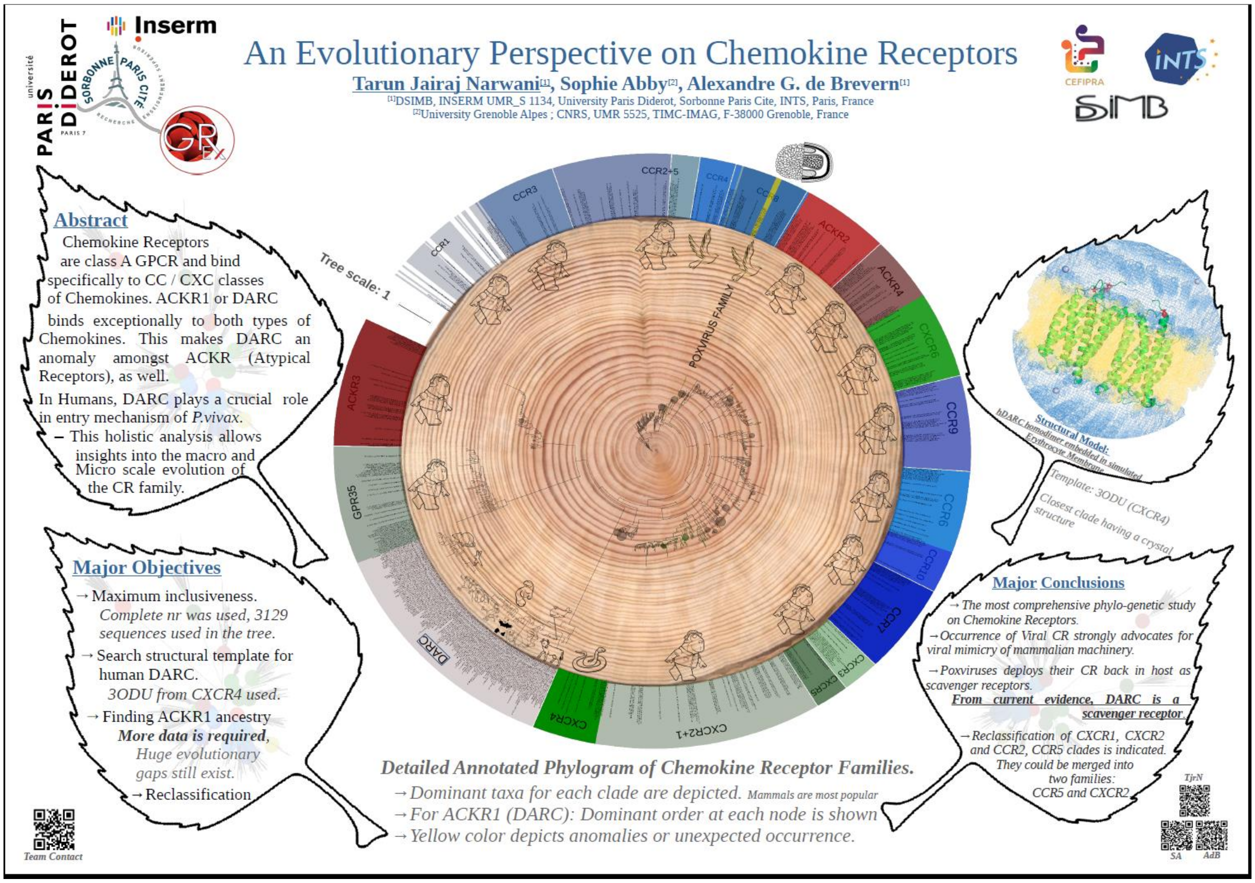
A caveat of studying evolution of viral chemokine receptors along side canonical and decoy chemokine receptors is the choice of the substitution matrix. Although the vChemR are genetically pirated from the vertebrate hosts yet they are expressed as a viral protein and therefore, virus specific substitution models should be preferred. Comparing tree topologies of ACKR and viral chemokine receptors generated using a virus specific substitution matrix to the current tree topology can be insightful.

ACKNOWLEDGEMENTS

All the support and guidance required for the phylogenetic analysis was provided by our collabolator and friend from Austria, Dr. Sophie Abby. Her inputs have been pivotal in designing and shaping the work. It is her expertise in molecular phylogenetics that suggested using MSA of 118 chemokine receptors as seed for the MSA of 3129 sequences. It is my pleasure to have found a mentor like Sophie.

Dissemination of the results

The results from chapter 6 have been published in the form of a scientific poster at ISMB/ECCB conference held at Prague in July 2017. The poster garnered overwhelming response with people asking interesting questions about the viral chemokine receptors. The poster was awarded as best poster at the conference. Narwani TJ, Abby S and de Brevern AG. An evolutionary perspective on chemokine receptor family [version 1; not peer reviewed]. *F1000Research* 2017, **6**(ISCB Comm J):1271 (poster) (doi: 10.7490/f1000research.1114530.1)



An Evolutionary Perspective on Chemokine Receptors

Tarun Jairaj Narwani[1], Sophie Abby[2], Alexandre G. de Brevern[1]

[1]DSIMB, INSERM UMR_S 1134, University Paris Diderot, Sorbonne Paris Cite, INTS, Paris, France
[2]University Grenoble Alpes ; CNRS, UMR 5525, TIMC-IMAG, F-38000 Grenoble, France

Abstract

Chemokine Receptors are class A GPCR and bind specifically to CC / CXC classes of Chemokines. ACKR1 or DARC binds exceptionally to both types of Chemokines. This makes DARC an anomaly amongst ACKR (Atypical Receptors), as well.

In Humans, DARC plays a crucial role in entry mechanism of *P.vivax*.

– This holistic analysis allows insights into the macro and Micro scale evolution of the CR family.

Major Objectives

→ Maximum inclusiveness.
Complete nr was used, 3129 sequences used in the tree.

→ Search structural template for human DARC.
3ODU from CXCR4 used.

→ Finding ACKR1 ancestry
More data is required,
Huge evolutionary gaps still exist.

→ Reclassification

Team Contact

Detailed Annotated Phylogram of Chemokine Receptor Families.

→ Dominant taxa for each clade are depicted. Mammals are most popular
→ For ACKR1 (DARC): Dominant order at each node is shown
→ Yellow color depicts anomalies or unexpected occurrence.

Structural Model:
hDARC homodimer embedded in simulated Erythrocyte Membrane

Template: 3ODU (CXCR4)
Closest clade having a crystal structure

Major Conclusions

→ The most comprehensive phylo-genetic study on Chemokine Receptors.

→ Occurrence of Viral CR strongly advocates for viral mimicry of mammalian machinery.

→ Poxviruses deploys their CR back in host as scavenger receptors.
From current evidence, DARC is a scavenger receptor.

→ Reclassification of CXCR1, CXCR2 and CCR2, CCR5 clades is indicated. They could be merged into two families: CCR5 and CXCR2.

TjrN
SA AdB

During the compilation of the chapter, a manuscript has been written and we expect to get it published by end of the year 2018.

Conclusive Outline:

Before starting my PhD with Alex, I was gifted a book from my previous scientific advisor, Dr. Srikrishna Subramanian. The book was a series of inspiring lectures from famous physicist Dr. Richard Feynman titled ‘The pleasure of finding things out’. The book gives insights into Dr. Feynman’s approach towards science that can be summarized as simply as curiosity.

During the course of these three years of research work under PhD tenure, I got to learn more than just the meaning of pleasure of finding things out. Perhaps, Alex designed the flow of my PhD, titled Dynamics of protein structures and its impact on local structural behaviors, in this specific manner to help me learn and grow. The first chapter focuses on a portion of structural space, Helices and asks very simple question on how do they behave in dynamics. While chapter 5 that forms the penultimate chapter for my thesis deals with a complex structure assembly that too in a membrane environment. In hindsight, I can see the underlying plan of my thesis as I realize that each chapter exploits the information gained from its preceding chapters. The inclusion of chapter 6 was unavoidable because of its close association with chapter 5 of the thesis but it also helped me employ different *in silico* techniques to answer pertinent questions.

Chapter 1 is based on the objective to understand local structural behaviors in helices. Therefore, the first task was to check the persistence of helices in their original or starting conformation (during MD simulations). It was observed that more than 3/4th of α-helices persist thus indicating the order in their structures. However, 3_{10}-helices changed much frequently with more than 40.5% of time the residues assigned as 3_{10}-helices changed to either a helical or non-helical conformation. The π-helices were observed to be the most deformed as very few π-helices persisted as π-helices during the collective simulation time of 150ns. The α-helix showed good correlation among their stability and flexibility in terms of B-factors, RMSf and surface accessible area. The unsupervised clustering of different helical conformations and use of PB and related statistics, *Neq* showed that the α-helix also have a higher tendency to assume β-turn conformations than either of the two other helical forms. The individual clusters of 3_{10}- and π-helix revealed their tendency to transit to α-helix. However, the 3_{10}-helix that transformed to α-helix showed different characteristics. It depicted higher B-factor and RMSf values than the average values in its cluster, thus revealing that

3_{10}-helix are dynamic than α-helices. The residues associated with π-helices were found to be also closely associated with β-turn and bend rather than other helices. A counterintuitive finding was that π-helices that showed low B-factors but high accessibility. Thus defined them as very flexible/deformable also supported by their very high RMSF and *Neq* values. Such dynamic behavior of π-helices may be characteristics of post nucleation, cooperative protein folding effect of protein folding.

Since π-helices, due to their high deformability were being assessed for their involvement in disorder and folding process, we found out the that Polyproline II helical conformations dominates other helical forms in less structured space. Moreover, based on its geometry, sequence and structure it should be a part of regular secondary structure elements. PPII has a left-handed geometry unlike the right-handedness of popular protein helices. Therefore, in an attempt to understand PPII conformation, we reviewed the recent advances made in PPII. We found out that there has been a sudden surge in publications related to PPII. An interesting personal learning was that it is not necessary for PPII helical conformation to be comprised of proline residues at all. Rather, the amino acid composition of PPII can change depending on its context and so does its length. This provided interesting insights into inherent flexibility PPII helices contain.

After the analysis of dynamic behavior of helices, the study was extended to non-helical conformations as well. The resulting analysis confirmed the rigidity of sheets, but also underline their capacity to transform into turn conformations. While the dynamics between turns (with hydrogen bond) and bends (without hydrogen bond) showed some strong similarities, the two conformations behave quite distinctively. These revealing results about the dynamics of DSSP secondary structure states motivated us to analyze the structures using a structural alphabet – The Protein Blocks (PB). Systematic analysis of PBs provided surprising results with multiple information. An important one was in regards to the relationship between solvent accessibility, stability and dynamics. While a large part of buried residues remained stable, important discrepancies were observed. For at least half of the PBs (16 in number), the fact to be buried or exposed did not affect their dynamics, at all. Majority of PBs persisted as their original PB. Some PBs showed higher tendency to be not as rigid as others, particularly PB *g* and PB *i*. The changes amongst PBs in their clusters were assessed based on their geometrical compatibility. More frequently they tend to exchange with an unexpected PB than an expected one. Thus depicting the inherent flexibility in protein backbone using simple molecular dynamics. These results with PBs and DSSP states provided a very basic understanding that protein structures are much more dynamic than we usually assume due to our exposure to static X-ray crystal data. The importance of having such inherent flexibility can be attributed their involvement in regulation of cellular function.

Such regulatory processes require protein structures to modulate their behavior in different contexts and at molecular level, it is achieved by post translational modifications (PTM). Therefore, we analyzed PDB data extracted from PTM-SD database in order to find the impact of PTMs on protein backbone. N-glycosylation, phosphorylation and methylation were selected as PTMs of interest based on the sufficient data that exist in PDB. Besides their global analyses, specific example proteins were chosen for the three PTMs – N-glycosylation in Liver carboxylesterase 1 and Renin endopeptidase, threonine phosphorylation in Cyclin dependent kinase 2 and histidine methylation in Actin. The backbone analysis using PB derived entropy function (*Neq)* of N-glycosylation showed that the addition of the glycan neither impact the local nor the global backbone conformation of the proteins. However, the methylation on actin structure induced a local increase of the backbone diversity at the PTM site region, thus highlighting a higher deformation of this part of the protein. However, no effect on the intrinsic mobility of this region was observed as the structure with and without PTM had same B-factor profiles. *Neq* as well as normalized B-factor values revealed that the phosphorylation site and its neighborhood positions display a significant backbone diversity. The comparison among modified and unmodified structures of CDK2 revealed that the phosphorylation on the activation loop at Thr 160 have several local effects. It rigidifies the backbone locally while increasing the deformation at two distant regions both of which are also important sites for PTM.

Despite the intrinsic link between PTM and protein function, the molecular effects of the modifications on the protein structures and dynamics remain poorly understood. Therefore, molecular modeling of PTMs combined with molecular dynamic simulation is an interesting alternative. It is mention-worthy here that I also work on understanding dynamics of active to inactive transformation in protein kinases in collaboration with Prof. N. Srinivasan of IISc Bangalore. I have already completed the structures of kinases with and without PTM, using molecular modelling. Moreover, we also recently submitted a molecular dynamics analysis of active and inactive protein kinase A. Therefore, these data can be used for a more global understanding of the impact of PTM on protein backbone.

A major concern while dealing with PTM structures are the missing regions in PDB crystals that mostly pertains to disorder in protein. The disorder helps protein structures to expand their protein-protein interactome. The selectivity of interacting partners and order-disorder transition of the protein structures is also regulated by PTMs, and most of the times by phosphorylation thus explaining the missing regions in our analyses. While managing these missing regions in phosphorylation data, we stumbled upon a rather interesting structural event.

A unique structural event of protein life known as Dual personality fragments (DPF) was identified and analyzed subsequent to the analysis of the effect of PTM on protein backbone. DPF are regions in a protein structure that can transform between disorder and order structural states and is expected to be lying at the core of structural continuum. Almost scarce information is available for DPF as only a single research article by Adam Godzik's lab in 2007. They tried to identify and characterize DPF and proposed that DPF differs from the disorder and order in their specific sequence composition. The DPF characteristic amino acid signature, as proposed by Zhang *et al.* 2007 is, Thr, Arg, Gly, Asn, Pro, and Asp. Though a one of its kind and a benchmark study, they focus mostly on sequence based characterization of DPF like most disorder related articles. DPF transit from disorder to order and thus will have structural information available. We decided to exploit this structural information to characterize DPF along with sequence features. Based on our analyses with PB, secondary structures, B-factors, and solvent accessibility we characterize DPF.

High frequency of Cys, Gly, Asp, and Lys in a region can be an indicative of a DPF. Two of these residues, Cys and Gly are rigid and moderately flexible while rest two are highly flexible. Also, Cys and Gly are hydrophobic while Asp and Lys are hydrophilic. This information coupled with involvement of DPF in multi-partner interactions and MoRFs (Molecular recognition features), the presence of high propensity of these residues makes sense. Additionally, having a region with high occurrence of C, G, D, and K that also have a higher alpha helical and beta turn content is also an indicative of a DPF. Although the definite characterization of DPF and using that information to predict DPF from sequence seems distant for now yet the information acquired about the protein local structures is enriching.

The finding of a region in CDK2 that showed distant effect upon phosphorylation indicated allostery or at least long range interactions in protein structures. This motivated us to move further ahead from analysis of secondary structures to more complex multi-domain proteins like integrin $\alpha_{IIb}\beta_3$. The integrin structure undergoes structural transition from a bend to open conformation and allostery have been shown to play a major role. We decided to study two rigid domains in the structure which functions are the anchor during the structural transformation. Therefore, the intrinsic dynamics of these rigid beta sandwich domains, Calf-1 and Calf-2, would be quite exciting to study. Especially since the $\alpha_{IIb}\beta_3$ is implicated in a rare bleeding disorder like Glanzmann Thrombasthenia (GT) and pregnancy related disorder Fetal / Neonatal Alloimmune Thrombocytopenia (FNAIT). Protein blocks statistical measures like ΔNeq and ΔPB were used to analyze the impact of GT mutations on the Calf-1 domain. The significance of using these measures is their ability to resolve local rigid regions encompassed inside an otherwise deformable region.

It was observed that the impact of GT variants that may disturb the core β-strands are systematically compensated by the loops. The energy gained or lost due to loss of interactions in mutants was shown to be compensated by new interactions with the residual energy being transferred to the loops. Interestingly, of the seven GT variants studied only two, C674 and P741, displayed conformational changes at the mutated site. The case of the C674R substitution was particularly enriching for me as I remembered the famous Anfinsen's experiment. The resultant loss of the disulfide bridge relaxes the structure and introduces significant structural alterations but the β sandwich architecture persist. Such an effect suggested that the structural-functional context influences the rigidity. Thus, inherent flexibility is important and crucial to the conservation of the core.

In terms of understanding the behavior of local structural flexibility, we further notched up to a more complex structural organization with a dimer formation in a phospholipid membrane system. The protein of interest was Duffy Antigen Chemokine Receptor, DARC. DARC is identified as a mammalian chemokine receptor that can bind to inflammatory chemokines across classes. Besides able to bind effectively to different chemokines, it does not transduce the signal since it lacks the motifs that couple with G-proteins during GPCR signaling. Therefore, International Union for Pharmacology (IUPHAR) updated the nomenclature and replaced DARC with Atypical

Chemokine Receptor 1 (ACKR1). Among the atypical chemokine receptors, ACKR1 is the only one that exhibit promiscuous binding with chemokines and lacks the DRYLAIV motif completely. Most characteristically, ACKR1 serves as a receptor for *Plasmodium vivax* merozoites on human RBC that leads to the symptomatic infectious stage of malaria. Although DARC was the first one to be identified among chemokine receptors yet to date no experimentally determined structure exists. Only a structural model generated using homology modelling exists that was done by Alex in 2005. Therefore, we decided to build a structural model for ACKR1 integrating the latest physiological, pathological and evolutionary information available. The physio-pathological properties of chemokine receptors assisted in identifying the key residues. Structural information from other chemokine receptors was also instrumental in establishing the basic scaffold for modelling ACKR1. Using these along with phylogenetic information of human chemokine recepors as well as structural information acquired in preceding chapters, ACKR1 was modelled as a homodimer based on the crystal structure of active CXCR4 (PDBid 3ODU). The dimeric interface was determined and key residues were identified. Most importantly, the dimer model is embedded in an erythrocyte membrane mimic system. Special caution was carried in building the membrane and was perhaps the most challenging task of all due to the specific cell structure of RBCs.

The primary objective is to understand the dynamics of local secondary structures and protein blocks at the interface region as well as at the sites of conserved micro-switch motifs. Therefore, while the 1 microsecond range simulations are ongoing on the cluster, a primary study of the motions using ANM based normal mode analysis (NMA) is designed. The NMA results have been just weeks before completion of this thesis and therefore not included in the main chapter. The preliminary NMA results indicate that the two subunits of the dimer have different structural fluctuations rather they are shown to be negatively correlated. Exploration of these results can provide insights into the individual and concerted dynamics of the dimer. Further, a perturbation response study of key residues has also been carried out. The preliminary results show that the interface residues are the most effected one and the TM1 and TM7 are the most exposed and sensitive regions. Any perturbation in the interface residue leads to increased fluctuation in the overall structure, especially in TM1 and TM7. The further analysis of these results can help us understand the role of allostery in the 7TM structure of ACKR1. Of course, the conclusive remarks on the dynamics of the local structures in the homodimeric, membrane embedded, assembly of

ACKR1 requires all atom molecular dynamics. However, given the enormous size of the system the computational cost is expensive and I hope I will have time to analyze these results in subsequent months of finishing this document.

Given the endemic that is malaria and especially the global widespread presence of *P.vivax* infected malaria, the prime objective has been extended to study the interactions of ECD1 with *P.vivax* DARC Binding Ligand.

During the template selection for modelling ACKR1, information of phylogenetic relationships among 21 human chemokine receptors was used. The resulting tree showed ACKR1 to be highly distant from rest of the family. This kindled the curiosity regarding the evolution of ACKR1. Therefore, we decided to collaborate with Dr. Sophie Abby from Austria to study the molecular phylogenetics of the chemokine receptor family. Foremost, this required a basic understanding of chemokines and their receptors, their sequence, structure and function characteristics. The current classification of chemokine receptors is based on the class of chemokine ligand binding to the receptor. An α-chemokine (CXC type) binding receptor is named CXCR while a β-chemokine (CC type) binding receptor is termed as CCR. Such a premise, however may change with the enhancing knowledge about the hetero-oligomerization of chemokine receptors and discovery of more virally encoded chemokine receptors. The information gathering process during this project was fascinating for instance, I learned for the very first time about the existence of virally encoded chemokine receptors. The concept of gene piracy and how smartly does viruses use chemokine receptor mimics of their host to escape the immune response.

The design of protocol for generating the tree was suggested by Sophie. Thus we used state of the art tools like SiLiX for clustering the sequences and IQtree for generating the maximum likelihood tree. Also, the sequence alignment was checked at each stage for conservation profiles of important functional residues. For the final multiple sequence alignment (MSA) of 3129 sequences, the manually verified MSA of 118 sequences (used for HMM profile building) was used as a seed. The most fascinating part was the analysis and deducing functional information from the tree. Strong evidence provided by the overlaps in the clades of CCR2 and CCR5 as well as CXCR1 and CXCR2 suggested similar evolutionary pressures among the two pair of receptors. The similarities between the gene locations of their respective ligands as well their functional similarity further support the merging of the two clades. The evolution of ACKR1 was traced by converting their branch lengths to ages. The age helped in identifying the first (primitive) and the most recent occurrence of ACKR1 among species of tree of life and also in tracing the gaps in between. Strikingly, there are no ACKR1 sequences identified outside class Mammalia and thus limits the access of our analysis. The absence of non-mammalian ACKR1 gene might be the reason for the huge distance of ACKR1 clade from rest of the tree that have sequences from Amphibia, Fishes as well as Reptilia. A strong hypothesis might be that ACKR1 is not a chemokine receptor and have been a viral chemokine receptor that was genetically pirated from the host during infection. However, during a subsequent infection by the virus (possibly a ssDNA or retrovirus) the altered receptor gene was back-pirated into the host genome.

Sadly, I do not have sufficient time and information at present to test this hypothesis but this question remains open. Hopefully, I may return to address this question during my research career. I do acknowledge that the test may be negative for such a strong hypothesis but such is the pleasure of finding things out.

REFERENCES

1. Nishizuka Y. The role of protein kinase C in cell surface signal transduction and tumour promotion. Nature. 1984;308: 693–698. doi:10.1038/308693a0

2. Agarwal PK. Role of Protein Dynamics in Reaction Rate Enhancement by Enzymes. J Am Chem Soc. 2005;127: 15248–15256. doi:10.1021/ja055251s

3. Gilman AG. G Proteins: Transducers of Receptor-Generated Signals. Annu Rev Biochem. 1987;56: 615–649. doi:10.1146/annurev.bi.56.070187.003151

4. Pollard TD, Weihing RR, Adelman MR. Actin And Myosin And Cell Movemen. CRC Crit Rev Biochem. 1974;2: 1–65. doi:10.3109/10409237409105443

5. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227: 561–563. doi:10.1038/227561a0

6. Dobson CM. Protein folding and misfolding. Nature. 2003;426: 884–890. doi:10.1038/nature02261

7. Jalkanen KJ, Elstner M, Suhai S. Amino Acids and Small Peptides as Building Blocks for Proteins: Comparative Theoretical and Spectroscopic Studies. ChemInform. 2005;36. doi:10.1002/chin.200505280

8. Dill KA. Dominant forces in protein folding. Biochemistry. 1990;29: 7133–7155. doi:10.1021/bi00483a001

9. Anfinsen CB. Principles that Govern the Folding of Protein Chains. Science. 1973;181: 223–230. doi:10.1126/science.181.4096.223

10. Koonin EV, Tatusov RL, Galperin MY. Beyond complete genomes: from sequence to structure and function. Curr Opin Struct Biol. 1998;8: 355–363. doi:10.1016/s0959-440x(98)80070-5

11. Sadowski MI, Jones DT. The sequence–structure relationship and protein function prediction. Curr Opin Struct Biol. 2009;19: 357–362. doi:10.1016/j.sbi.2009.03.008

12. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to Glutaredoxins/Thioredoxins and T 1 Ribonucleases 1 1Edited by F. Cohen. J Mol Biol. 1998;281: 949–968. doi:10.1006/jmbi.1998.1993

13. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007;8: 995–1005. doi:10.1038/nrm2281

14. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7: 95–99. doi:10.1016/s0022-2836(63)80023-6

15. Bolin KA, Millhauser GL. α and 310: The Split Personality of Polypeptide Helices. Acc Chem Res. 1999;32: 1027–1033. doi:10.1021/ar980065v

16. Low BW, Baybutt RB. THE π HELIX—A HYDROGEN BONDED CONFIGURATION OF THE POLYPEPTIDE CHAIN. J Am Chem Soc. 1952;74: 5806–5807. doi:10.1021/ja01142a539

17. von Heijne G. Proline kinks in transmembrane α-helices. J Mol Biol. 1991;218: 499–503. doi:10.1016/0022-2836(91)90695-3

18. Craveur P, Joseph AP, Esque J, Narwani TJ, Noël F, Shinada N, et al. Protein flexibility in the light of structural alphabets. Front Mol Biosci. 2015;2: 20. doi:10.3389/fmolb.2015.00020

19. Ramachandran GX, Chandrasekharan R. Interchain hydrogen bonds via bound water molecules in the collagen triple helix. Biopolymers. 1968;6: 1649–1658. doi:10.1002/bip.1968.360061109

20. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol. 1984;44: 97–179. doi:10.1016/0079-6107(84)90007-5

21. Richardson JS, Getzoff ED, Richardson DC. The beta bulge: a common small unit of nonrepetitive protein structure. Proceedings of the National Academy of Sciences. 1978;75: 2574–2578. doi:10.1073/pnas.75.6.2574

22. Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers. 1968;6: 1425–1436. doi:10.1002/bip.1968.360061006

23. Harrison S. DNA Recognition By Proteins With The Helix-Turn-Helix Motif. Annu Rev Biochem. 1990;59: 933–969. doi:10.1146/annurev.biochem.59.1.933

24. Lewit-Bentley A, Réty S. EF-hand calcium-binding proteins. Curr Opin Struct Biol. 2000;10: 637–643. Available: https://www.ncbi.nlm.nih.gov/pubmed/11114499

25. Sibanda BL, Thornton JM. β-Hairpin families in globular proteins. Nature. 1985;316: 170–174. doi:10.1038/316170a0

26. Teichmann SA, Levy ED, Marsh JA, De S. Evolution and dynamics of protein complexes. Acta Crystallogr A. 2011;67: C62–C62. doi:10.1107/s0108767311098527

27. Fanelli AR, Antonini E, Caputo A. Hemoglobin and Myoglobin. Advances in Protein Chemistry. 1964. pp. 73–222. doi:10.1016/s0065-3233(08)60189-8

28. Dayhoff MO. The origin and evolution of protein superfamilies. Fed Proc. 1976;35: 2132–2138. Available: https://www.ncbi.nlm.nih.gov/pubmed/181273

29. Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res. 1997;25: 236–239. doi:10.1093/nar/25.1.236

30. Johnson JE, Cornell RB. Amphitropic proteins: regulation by reversible membrane interactions (Review). Mol Membr Biol. 1999;16: 217–235. doi:10.1080/096876899294544

31. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature. 1958;181: 662–666. Available: https://www.ncbi.nlm.nih.gov/pubmed/13517261

32. Matthews BW, Remington SJ. The Three Dimensional Structure of the Lysozyme from Bacteriophage T4. Proceedings of the National Academy of Sciences. 1974;71: 4178–4182. doi:10.1073/pnas.71.10.4178

33. Baase WA, Liu L, Tronrud DE, Matthews BW. Lessons from the lysozyme of phage T4. Protein Sci. 2010;19: 631–641. doi:10.1002/pro.344

34. Cowtan K. Phase Problem in X-ray Crystallography, and Its Solution. Encyclopedia of Life Sciences. 2003. doi:10.1038/npg.els.0002722

35. Hauptman HA. The Phase Problem of X-ray Crystallography: Overview. Electron Crystallography. 1997. pp. 131–138. doi:10.1007/978-94-015-8971-0_9

36. Kelly A, Groves GW, Kidd P. Crystallography and Crystal Defects [Internet]. John Wiley & Sons; 2000. Available: https://books.google.com/books/about/Crystallography_and_Crystal_Defects.html?hl=&id=7SRoezfTO6QC

37. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics. 2008;24: 652–658. doi:10.1093/bioinformatics/btn022

38. Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. J Mol Biol. 2007;372: 774–797. doi:10.1016/j.jmb.2007.05.022

39. Janin J, Rodier F. Protein-protein interaction at crystal contacts. Proteins. 1995;23: 580–587. doi:10.1002/prot.340230413

40. Carugo O, Argos P. Protein-protein crystal-packing contacts. Protein Sci. 1997;6: 2261–2263. doi:10.1002/pro.5560061021

41. Luo J, Liu Z, Guo Y, Li M. A structural dissection of large protein-protein crystal packing contacts. Sci Rep. 2015;5. doi:10.1038/srep14214

42. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem. 2015;84: 551–575. doi:10.1146/annurev-biochem-060614-034142

43. Bai X-C, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015;40: 49–57. doi:10.1016/j.tibs.2014.10.005

44. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne J, et al. 2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor [Internet]. 2015. doi:10.2210/pdb5a1a/pdb

45. Nogales E. The development of cryo-EM into a mainstream structural biology technique. Nat Methods. 2016;13: 24–27. doi:10.1038/nmeth.3694

46. Orlova EV, Saibil HR. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. Chem Rev. 2011;111: 7710–7748. doi:10.1021/cr100353t

47. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974;13: 222–245. doi:10.1021/bi00699a002

48. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43: W389–W394. doi:10.1093/nar/gkv332

49. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000;16: 404–405. doi:10.1093/bioinformatics/16.4.404

50. Heffernan R, Dehzangi A, Lyons J, Paliwal K, Sharma A, Wang J, et al. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics. 2016;32: 843–849. doi:10.1093/bioinformatics/btv665

51. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22: 2577–2637. doi:10.1002/bip.360221211

52. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic

Acids Res. 2004;32: W500–2. doi:10.1093/nar/gkh429

53. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins. 2000;41: 271–287. Available: https://www.ncbi.nlm.nih.gov/pubmed/11025540

54. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π-helix. Protein Eng Des Sel. 2002;15: 353–358. doi:10.1093/protein/15.5.353

55. Kohonen T. Self-Organizing Maps [Internet]. 2001. doi:10.1007/978-3-642-56927-2

56. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43: 59–69. doi:10.1007/bf00337288

57. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77: 257–286. doi:10.1109/5.18626

58. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: improved prediction methods. Proteins. 2005;59: 810–827. doi:10.1002/prot.20458

59. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng. 1996;9: 833–842. Available: https://www.ncbi.nlm.nih.gov/pubmed/8931122

60. Gelly J-C, Joseph AP, Srinivasan N, de Brevern AG. iPBA: a tool for protein structure comparison using sequence alignment strategies. Nucleic Acids Res. 2011;39: W18–23. doi:10.1093/nar/gkr333

61. Joseph AP, Srinivasan N, de Brevern AG. Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies. Biochimie. 2012;94: 2025–2034. doi:10.1016/j.biochi.2012.05.028

62. Dudev M, Lim C. Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics. 2007;8: 106. doi:10.1186/1471-2105-8-106

63. Wu CY, Chen YC, Lim C. A structural-alphabet-based strategy for finding structural motifs across protein families. Nucleic Acids Res. 2010;38: e150. doi:10.1093/nar/gkq478

64. Zimmermann O, Hansmann UHE. LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. J Chem Inf Model. 2008;48: 1903–1908. doi:10.1021/ci800178a

65. Rangwala H, Kauffman C, Karypis G. svmPRAT: SVM-based protein residue annotation toolkit. BMC Bioinformatics. 2009;10: 439. doi:10.1186/1471-2105-10-439

66. Suresh V, Ganesan K, Parthasarathy S. A protein block based fold recognition method for the annotation of twilight zone sequences. Protein Pept Lett. 2013;20: 249–254. Available: https://www.ncbi.nlm.nih.gov/pubmed/22591480

67. Joseph AP, de Brevern AG. From local structure to a global framework: recognition of protein folds. J R Soc Interface. 2014;11: 20131147. doi:10.1098/rsif.2013.1147

68. Craveur P, Joseph AP, Rebehmed J, de Brevern AG. β-bulges: Extensive structural analyses of β-sheets irregularities. Protein Sci. 2013; doi:10.1002/pro.2324

69. Barnoud J, Santuz H, Craveur P, Joseph AP, Jallu V, de Brevern AG, et al. PBxplore: a tool to analyze local protein structure and

deformability with Protein Blocks. PeerJ. 2017;5: e4013. doi:10.7717/peerj.4013

70. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234: 779–815. doi:10.1006/jmbi.1993.1626

71. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. Available: https://www.ncbi.nlm.nih.gov/pubmed/9254694

72. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10: 845–858. doi:10.1038/nprot.2015.053

73. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009;77 Suppl 9: 128–132. doi:10.1002/prot.22499

74. Sali A, Overington JP. Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci. 1994;3: 1582–1596. doi:10.1002/pro.5560030923

75. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Current Protocols in Protein Science. 2016. pp. 2.9.1–2.9.37. doi:10.1002/cpps.20

76. Swapna LS, Mahajan S, de Brevern AG, Srinivasan N. Comparison of tertiary structures of proteins in protein-protein complexes with unbound forms suggests prevalence of allostery in signalling proteins. BMC Struct Biol. 2012;12: 6. doi:10.1186/1472-6807-12-6

77. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26: 1668–1688. doi:10.1002/jcc.20290

78. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem. 1983;4: 187–217. doi:10.1002/jcc.540040211

79. Christen M, Hünenberger PH, Bakowies D, Baron R, Bürgi R, Geerke DP, et al. The GROMOS software for biomolecular simulation: GROMOS05. J Comput Chem. 2005;26: 1719–1751. doi:10.1002/jcc.20303

80. Cristianini N. Gradient Descent (Steepest Descent Method). Dictionary of Bioinformatics and Computational Biology. 2004. doi:10.1002/9780471650126.dob0303.pub2

81. Steihaug T, Hestenes M. Conjugate Direction Methods in Optimization. Math Comput. 1982;38: 332. doi:10.2307/2007488

82. Cramer CJ, Truhlar DG. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. Chem Rev. 1999;99: 2161–2200. doi:10.1021/cr960149m

83. Levy RM, Gallicchio E. COMPUTER SIMULATIONS WITH EXPLICIT SOLVENT: Recent Progress in the Thermodynamic Decomposition of Free Energies and in Modeling Electrostatic Effects. Annu Rev Phys Chem. 1998;49: 531–567. doi:10.1146/annurev.physchem.49.1.531

84. Makov G, Payne MC. Periodic boundary conditions inab initiocalculations. Phys Rev B: Condens Matter Mater Phys. 1995;51: 4014–4022. doi:10.1103/physrevb.51.4014

85. Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. Chem Rev. 2010;110: 1463–1497. doi:10.1021/cr900095e

86. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins: Structure, Function, and Genetics. 1998;33: 417–429. doi:3.0.co;2-8">10.1002/(sici)1097-0134(19981115)33:3<417::aid-prot10>3.0.co;2-8

87. de Vries SJ, van Dijk M, Bonvin AMJJ. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc. 2010;5: 883–897. doi:10.1038/nprot.2010.32

88. Dominguez C, Boelens R, Alexandre M J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc. 2003;125: 1731–1737. doi:10.1021/ja026939x

89. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30: 2785–2791. doi:10.1002/jcc.21256

90. Minke WE, Diller DJ, Hol WG, Verlinde CL. The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test case. J Med Chem. 1999;42: 1778–1788. doi:10.1021/jm980472c

91. Sotriffer CA, Flader W, Winger RH, Rode BM, Liedl KR, Varga JM. Automated docking of ligands to antibodies: methods and applications. Methods. 2000;20: 280–291. doi:10.1006/meth.1999.0922

92. Rimoin DL, Pyeritz RE, Korf B. Emery and Rimoin's Principles and Practice of Medical Genetics [Internet]. Academic Press; 2013. Available: https://books.google.com/books/about/Emery_and_Rimoin_s_Principles_and_Practi.html?hl=&id=GHlTYt11KL8C

93. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066. Available: https://www.ncbi.nlm.nih.gov/pubmed/12136088

94. Tabassum Khan N, Khan NT. MEGA - Core of Phylogenetic Analysis in Molecular Evolutionary Genetics. Journal of Phylogenetics & Evolutionary Biology. 2017;05. doi:10.4172/2329-9002.1000183

95. Felsenstein J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. Evolution. 1985;39: 783–791. doi:10.1111/j.1558-5646.1985.tb00420.x

96. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012;13: 303–314. doi:10.1038/nrg3186

97. Dinh V, Darling AE, Matsen FA Iv. Online Bayesian Phylogenetic Inference: Theoretical Foundations via Sequential Monte Carlo. Syst Biol. 2018;67: 503–517. doi:10.1093/sysbio/syx087

98. Eisenberg D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A. 2003;100: 11207–11210. doi:10.1073/pnas.2034522100

99. Pauling L, Corey RB, Branson HR. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences. 1951;37: 205–211. doi:10.1073/pnas.37.4.205

100. Pauling L, Corey RB. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. Proceedings of the National Academy of Sciences. 1951;37: 251–256. doi:10.1073/pnas.37.5.251

101. Millhauser GL. Views of helical peptides: a proposal for the position of 3(10)-helix along the thermodynamic folding pathway. Biochemistry. 1995;34: 3873–3877. Available: https://www.ncbi.nlm.nih.gov/pubmed/7696249

102. Millhauser GL, Stenland CJ, Bolin KA, van de Ven FJM. Local helix content in an alanine-rich peptide as determined by the complete set of 3JHNα coupling constants. J Biomol NMR. 1996;7: 331–334. doi:10.1007/bf00200434

103. Armen R, Alonso DOV, Daggett V. The role of alpha-, 3(10)-, and pi-helix in helix-->coil transitions. Protein Sci. 2003;12: 1145–1157. doi:10.1110/ps.0240103

104. Ghozlane A, Joseph AP, Bornot A, de Brevern AG. Analysis of protein chameleon sequence characteristics. Bioinformation. 2009;3: 367–369. Available: https://www.ncbi.nlm.nih.gov/pubmed/19759809

105. Miller SE, Watkins AM, Kallenbach NR, Arora PS. Effects of side chains in helix nucleation differ from helix propagation. Proc Natl Acad Sci U S A. 2014;111: 6636–6641. doi:10.1073/pnas.1322833111

106. Chakrabartty A, Kortemme T, Baldwin RL. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Sci. 2008;3: 843–852. doi:10.1002/pro.5560030514

107. Kumar S, Bansal M. Geometrical and Sequence Characteristics of α-helices in Globular Proteins. Biophys J. 1998;75: 1935–1944. doi:10.1016/s0006-3495(98)77634-9

108. Malkov SN, Živković MV, Beljanski MV, Stojanović SĐ, Zarić SD. A Reexamination of Correlations of Amino Acids with Particular Secondary Structures. Protein J. 2009;28: 74–86. doi:10.1007/s10930-009-9166-3

109. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science. 1988;240: 1648–1652. Available: https://www.ncbi.nlm.nih.gov/pubmed/3381086

110. Pal L, Chakrabarti P, Basu G. Sequence and Structure Patterns in Proteins from an Analysis of the Shortest Helices: Implications for Helix Nucleation. J Mol Biol. 2003;326: 273–291. doi:10.1016/s0022-2836(02)01338-4

111. Presta LG, Rose GD. Helix signals in proteins. Science. 1988;240: 1632–1641. Available: https://www.ncbi.nlm.nih.gov/pubmed/2837824

112. Aurora R, Srinivasan R, Rose G. Rules for alpha-helix termination by glycine. Science. 1994;264: 1126–1130. doi:10.1126/science.8178170

113. Aurora R, Rose GD. Helix capping. Protein Sci. 1998;7: 21–38. doi:10.1002/pro.5560070103

114. Ho BK, Thomas A, Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci. 2003;12: 2508–2522. doi:10.1110/ps.03235203

115. Ermolenko DN, Thomas ST, Aurora R, Gronenborn AM, Makhatadze GI. Hydrophobic Interactions at the Ccap Position of the C-capping Motif of α-Helices. J Mol Biol. 2002;322: 123–135. doi:10.1016/s0022-2836(02)00734-9

116. Prieto J, Serrano L. C-capping and helix stability: the Pro C-capping motif. J Mol Biol. 1997;274: 276–288. doi:10.1006/jmbi.1997.1322

117. Dirr HW, Little T, Kuhnert DC, Sayed Y. A Conserved N-capping Motif Contributes Significantly to the Stabilization and Dynamics of the C-terminal Region of Class Alpha GlutathioneS-Transferases. J Biol Chem. 2005;280: 19480–19487. doi:10.1074/jbc.m413608200

118. Kuhnert DC, Sayed Y, Mosebi S, Sayed M, Sewell T, Dirr HW. Tertiary interactions stabilise the C-terminal region of human glutathione transferase A1-1: a crystallographic and calorimetric study. J Mol Biol. 2005;349: 825–838. doi:10.1016/j.jmb.2005.04.025

119. Donohue J. Hydrogen Bonded Helical Configurations of the Polypeptide Chain. Proc Natl Acad Sci U S A. 1953;39: 470–478. Available: https://www.ncbi.nlm.nih.gov/pubmed/16589292

120. Pal L, Basu G. Novel protein structural motifs containing two-turn and longer 3(10)-helices. Protein Eng. 1999;12: 811–814. Available: https://www.ncbi.nlm.nih.gov/pubmed/10556239

121. Pal L, Basu G, Chakrabarti P. Variants of 310-helices in proteins. Proteins: Structure, Function, and Genetics. 2002;48: 571–579. doi:10.1002/prot.10184

122. Barlow DJ, Thornton JM. Helix geometry in proteins. J Mol Biol. 1988;201: 601–619. Available: https://www.ncbi.nlm.nih.gov/pubmed/3418712

123. Pal L, Dasgupta B, Chakrabarti P. 3(10)-Helix adjoining alpha-helix and beta-strand: sequence and structural features and their conservation. Biopolymers. 2005;78: 147–162. doi:10.1002/bip.20266

124. Karpen ME, de Haseth PL, Neet KE. Differences in the amino acid distributions of 3(10)-helices and alpha-helices. Protein Sci. 1992;1: 1333–1342. doi:10.1002/pro.5560011013

125. Khrustalev VV, Barkovsky EV, Khrustaleva TA. The Influence of Flanking Secondary Structures on Amino Acid Content and Typical Lengths of 3/10 Helices. Int J Proteomics. 2014;2014: 1–13. doi:10.1155/2014/360230

126. Weaver TM. The pi-helix translates structure into function. Protein Sci. 2000;9: 201–206. doi:10.1110/ps.9.1.201

127. Kumar P, Bansal M. Dissecting π-helices: sequence, structure and function. FEBS J. 2015;282: 4415–4432. doi:10.1111/febs.13507

128. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins. 2002;47: 228–235. Available: https://www.ncbi.nlm.nih.gov/pubmed/11933069

129. Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Res. 2016;44: W430–W435. doi:10.1093/nar/gkw306

130. Pancsa R, Raimondi D, Cilia E, Vranken WF. Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity. Biophys J. 2016;110: 572–583. doi:10.1016/j.bpj.2015.12.028

131. Lee KH, Benson DR, Kuczera K. Transitions from alpha to pi helix observed in molecular dynamics simulations of synthetic peptides. Biochemistry. 2000;39: 13737–13747. Available: https://www.ncbi.nlm.nih.gov/pubmed/11076513

132. Goyal B, Kumar A, Srivastava KR, Durani S. Scrutiny of chain-length and N-terminal effects in α-helix folding: a molecular dynamics study on polyalanine peptides. J Biomol Struct Dyn. 2017;35: 1923–1935. doi:10.1080/07391102.2016.1199972

133. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2002;58: 899–907. doi:10.1107/s0907444902003451

134. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. Available: https://www.ncbi.nlm.nih.gov/pubmed/10592235

135. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42: D304–9. doi:10.1093/nar/gkt1240

136. Craveur P, Rebehmed J, de Brevern AG. PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. Database . 2014;2014. doi:10.1093/database/bau041

137. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013;29: 845–854. doi:10.1093/bioinformatics/btt055

138. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Struct Funct Bioinf. 2010; NA–NA. doi:10.1002/prot.22711

139. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984;81: 3684–3690. doi:10.1063/1.448118

140. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys. 1981;52: 7182–7190. doi:10.1063/1.328693

141. Hess B, Bekker H, Berendsen HJC, Johannes G E. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997;18: 1463–1472. doi:3.3.co;2-l">10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.3.co;2-l

142. Darden T, Perera L, Li L, Pedersen L. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure. 1999;7: R55–60. Available: https://www.ncbi.nlm.nih.gov/pubmed/10368306

143. Bornot A, Etchebest C, de Brevern AG. Predicting protein flexibility through the prediction of local structures. Proteins. 2011;79: 839–852. doi:10.1002/prot.22922

144. van der Kant R, Vriend G. Alpha-Bulges in G Protein-Coupled Receptors. Int J Mol Sci. 2014;15: 7841–7864. doi:10.3390/ijms15057841

145. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat. 1979;28: 100. doi:10.2307/2346830

146. Tyagi M, Bornot A, Offmann B, de Brevern AG. Analysis of loop boundaries using different local structure assignment methods. Protein Sci. 2009;18: 1869–1881. doi:10.1002/pro.198

147. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins: Struct Funct Bioinf. 2005;61: 115–126. doi:10.1002/prot.20587

148. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics. 2006;22: 891–893. doi:10.1093/bioinformatics/btl032

149. de Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly J-C. PredyFlexy: flexibility and local structure prediction from sequence. Nucleic Acids Res. 2012;40: W317–22. doi:10.1093/nar/gks482

150. de Brevern AG. Extension of the classical classification of β-turns. Sci Rep. 2016;6: 33191. doi:10.1038/srep33191

151. Richardson JS. The Anatomy and Taxonomy of Protein Structure. Advances in Protein Chemistry. 1981. pp. 167–339. doi:10.1016/s0065-3233(08)60520-3

152. Rohl CA, Doig AJ. Models for the 3(10)-helix/coil, pi-helix/coil, and alpha-helix/3(10)-helix/coil transitions in isolated peptides. Protein Sci. 1996;5: 1687–1696. doi:10.1002/pro.5560050822

153. Pauling L, Corey RB. TWO HYDROGEN-BONDED SPIRAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN. J Am Chem Soc. 1950;72: 5349–5349. doi:10.1021/ja01167a545

154. Rose GD. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature. 1978;272: 586–590. Available:

https://www.ncbi.nlm.nih.gov/pubmed/643051

155. Bornot A, de Brevern AG. Protein beta-turn assignments. Bioinformation. 2006;1: 153–155. Available: https://www.ncbi.nlm.nih.gov/pubmed/17597878

156. Cowan PM, McGAVIN S, North ACT. The Polypeptide Chain Configuration of Collagen. Nature. 1955;176: 1062–1064. doi:10.1038/1761062a0

157. Pauling L, Corey RB. The Structure of Fibrous Proteins of the Collagen-Gelatin Group. Proceedings of the National Academy of Sciences. 1951;37: 272–281. doi:10.1073/pnas.37.5.272

158. Arnott S, Dover SD. The structure of poly-L-proline II. Acta Crystallogr B. 1968;24: 599–601. Available: https://www.ncbi.nlm.nih.gov/pubmed/5756983

159. Sasisekharan V. Structure of poly-L-proline. II. Acta Crystallogr. 1959;12: 897–903. doi:10.1107/s0365110x59002535

160. Ramachandran GN, Kartha G. Structure of collagen. Nature. 1955;176: 593–595. Available: https://www.ncbi.nlm.nih.gov/pubmed/13265783

161. Rich A, Crick FHC. The Structure of Collagen. Nature. 1955;176: 915–916. doi:10.1038/176915a0

162. Subramanian E. Nat Struct Biol. 2001;8: 489–491. doi:10.1038/88544

163. Bochicchio B, Tamburro AM. Polyproline II structure in proteins: Identification by chiroptical spectroscopies, stability, and functions. Chirality. 2002;14: 782–792. doi:10.1002/chir.10153

164. Soman KV, Ramakrishnan C. Occurrence of a single helix of the collagen type in globular proteins. J Mol Biol. 1983;170: 1045–1048. Available: https://www.ncbi.nlm.nih.gov/pubmed/6644813

165. Sreerama N, Woody RW. Poly(Pro)II Helixes in Globular Proteins: Identification and Circular Dichroic Analysis. Biochemistry. 1994;33: 10022–10025. doi:10.1021/bi00199a028

166. Whittington SJ, Chellgren BW, Hermann VM, Creamer TP. Urea promotes polyproline II helix formation: implications for protein denatured states. Biochemistry. 2005;44: 6269–6275. doi:10.1021/bi050124u

167. Toal S, Schweitzer-Stenner R. Local order in the unfolded state: conformational biases and nearest neighbor interactions. Biomolecules. 2014;4: 725–773. doi:10.3390/biom4030725

168. Adzhubei AA, Sternberg MJE. Left-handed Polyproline II Helices Commonly Occur in Globular Proteins. J Mol Biol. 1993;229: 472–493. doi:10.1006/jmbi.1993.1047

169. Mansiaux Y, Joseph AP, Gelly J-C, de Brevern AG. Assignment of PolyProline II conformation and analysis of sequence--structure relationship. PLoS One. 2011;6: e18401. doi:10.1371/journal.pone.0018401

170. Adzhubei AA, Sternberg MJE, Makarov AA. Polyproline-II helix in proteins: structure and function. J Mol Biol. 2013;425: 2100–2132. doi:10.1016/j.jmb.2013.03.018

171. Creamer TP. Left-handed polyproline II helix formation is (very) locally driven. Proteins. 1998;33: 218–226. Available: https://www.ncbi.nlm.nih.gov/pubmed/9779789

172. Ferreon JC, Hilser VJ. The effect of the polyproline II (PPII) conformation on the denatured state entropy. Protein Sci. 2003;12: 447–457. doi:10.1110/ps.0237803

173. Stapley BJ, Creamer TP. A survey of left-handed polyproline II helices. Protein Sci. 1999;8: 587–595. doi:10.1110/ps.8.3.587

174. Jha AK, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF. Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library†. Biochemistry. 2005;44: 9691–9702. doi:10.1021/bi0474822

175. Cubellis MV, Caillez F, Blundell TL, Lovell SC. Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. Proteins: Struct Funct Bioinf. 2005;58: 880–892. doi:10.1002/prot.20327

176. Kumar P, Bansal M. Structural and Functional Analyses of PolyProline-II helices in Globular Proteins [Internet]. 2016. doi:10.1101/068098

177. Carugo O, Djinovic-Carugo K. Half a century of Ramachandran plots. Acta Crystallogr D Biol Crystallogr. 2013;69: 1333–1341. doi:10.1107/S090744491301158X

178. Bella J, Eaton M, Brodsky B, Berman HM. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. Science. 1994;266: 75–81. Available: https://www.ncbi.nlm.nih.gov/pubmed/7695699

179. Aksianov E, Alexeevski A. SheeP: A tool for description of β-sheets in protein 3D structures. J Bioinform Comput Biol. 2012;10: 1241003. doi:10.1142/S021972001241003X

180. Cao C, Wang G, Liu A, Xu S, Wang L, Zou S. A New Secondary Structure Assignment Algorithm Using Cα Backbone Fragments. Int J Mol Sci. 2016;17: 333. doi:10.3390/ijms17030333

181. Carter P. DSSPcont: continuous secondary structure assignments for proteins. Nucleic Acids Res. 2003;31: 3293–3295. doi:10.1093/nar/gkg626

182. Cubellis M, Cailliez F, Lovell SC. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. BMC Bioinformatics. 2005;6: S8. doi:10.1186/1471-2105-6-s4-s8

183. Dupuis F, Sadoc J-F, Mornon J-P. Protein secondary structure assignment through Voronoï tessellation. Proteins. 2004;55: 519–528. doi:10.1002/prot.10566

184. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Genetics. 1995;23: 566–579. doi:10.1002/prot.340230412

185. Hosseini S-R, Sadeghi M, Pezeshk H, Eslahchi C, Habibi M. PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive Cα atoms. Comput Biol Chem. 2008;32: 406–411. doi:10.1016/j.compbiolchem.2008.07.027

186. Hutchinson EG, Thornton JM. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci. 1996;5: 212–220. doi:10.1002/pro.5560050204

187. King SM, Johnson WC. Assigning secondary structure from protein coordinate data. Proteins. 1999;35: 313–320. Available: https://www.ncbi.nlm.nih.gov/pubmed/10328266

188. Kneller GR, Hinsen K. Protein secondary-structure description with a coarse-grained model. Acta Crystallogr D Biol Crystallogr.

2015;71: 1411–1422. doi:10.1107/S1399004715007191

189. Labesse G, Colloc'h N, Pothier J, Mornon J-P. P-SEA: a new efficient assignment of secondary structure from Cα trace of proteins. Bioinformatics. 1997;13: 291–295. doi:10.1093/bioinformatics/13.3.291

190. Law SM, Frank AT, Brooks CL 3rd. PCASSO: a fast and efficient Cα-based method for accurately assigning protein secondary structure elements. J Comput Chem. 2014;35: 1757–1761. doi:10.1002/jcc.23683

191. Majumdar I, Krishna SS, Grishin NV. PALSSE: a program to delineate linear secondary structural elements from protein structures. BMC Bioinformatics. 2005;6: 202. doi:10.1186/1471-2105-6-202

192. Salawu EO. RaFoSA: Random forests secondary structure assignment for coarse-grained and all-atom protein systems. Cogent Biology. 2016;2. doi:10.1080/23312025.2016.1214061

193. Parisien M, Major F. A new catalog of protein beta-sheets. Proteins. 2005;61: 545–558. doi:10.1002/prot.20677

194. Park SY, Yoo M-J, Shin J, Cho K-H. SABA (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. BMB Rep. 2011;44: 118–122. doi:10.5483/BMBRep.2011.44.2.118

195. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins. 1988;3: 71–84. doi:10.1002/prot.340030202

196. Sklenar H, Etchebest C, Lavery R. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins: Structure, Function, and Genetics. 1989;6: 46–60. doi:10.1002/prot.340060105

197. Zacharias J, Knapp E-W. Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC. J Chem Inf Model. 2014;54: 2166–2179. doi:10.1021/ci5000856

198. Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol. 2005;5: 17. doi:10.1186/1472-6807-5-17

199. Offmann B, Tyagi M, de Brevern A. Local Protein Structures. Curr Bioinform. 2007;2: 165–202. doi:10.2174/157489307781662105

200. Srinivasan R, Rose GD. A physical basis for protein secondary structure. Proceedings of the National Academy of Sciences. 1999;96: 14258–14263. doi:10.1073/pnas.96.25.14258

201. Sugeta H, Miyazawa T. General method for calculating helical parameters of polymer chains from bond lengths, bond angles, and internal-rotation angles. Biopolymers. 1967;5: 673–679. doi:10.1002/bip.1967.360050708

202. Shakarji CM. Least-Squares Fitting Algorithms of the NIST Algorithm Testing System. J Res Natl Inst Stand Technol. 1998;103: 633–641. doi:10.6028/jres.103.043

203. Agrawal V, Kishan KVR. Promiscuous binding nature of SH3 domains to their target proteins. Protein Pept Lett. 2002;9: 185–193. Available: https://www.ncbi.nlm.nih.gov/pubmed/12144515

204. Hicks JM, Hsu VL. The extended left-handed helix: a simple nucleic acid-binding motif. Proteins. 2004;55: 330–338. doi:10.1002/prot.10630

205. Williamson MP. The structure and function of proline-rich regions in proteins. Biochem J. 1994;297 (Pt 2): 249–260. Available: https://www.ncbi.nlm.nih.gov/pubmed/8297327

206. Booker GW, Breeze AL, Downing AK, Panayotou G, Gout I, Waterfield MD, et al. STRUCTURE OF AN SH2 DOMAIN OF THE P85 ALPHA SUBUNIT OF PHOSPHATIDYLINOSITOL-3-OH KINASE [Internet]. 1994. doi:10.2210/pdb2pna/pdb

207. Koleske AJ, Buratowski S, Nonet M, Young RA. A novel transcription factor reveals a functional link between the RNA polymerase II CTD and TFIID. Cell. 1992;69: 883–894. doi:10.1016/0092-8674(92)90298-q

208. Suzuki M. SPXX, a frequent sequence motif in gene regulatory proteins. J Mol Biol. 1989;207: 61–84. Available: https://www.ncbi.nlm.nih.gov/pubmed/2500531

209. Suzuki M, Sohma H, Yazawa M, Yagi K, Ebashi S. Histone H1 kinase specific to the SPKK motif. J Biochem. 1990;108: 356–364. Available: https://www.ncbi.nlm.nih.gov/pubmed/2177468

210. Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, et al. Sequence-Specific RNA Binding by a Nova KH Domain. Cell. 2000;100: 323–332. doi:10.1016/s0092-8674(00)80668-6

211. Chevrier L, de Brevern A, Hernandez E, Leprince J, Vaudry H, Guedj AM, et al. PRR repeats in the intracellular domain of KISS1R are important for its export to cell membrane. Mol Endocrinol. 2013;27: 1004–1014. doi:10.1210/me.2012-1386

212. Blanch EW, Morozova-Roche LA, Cochran DAE, Doig AJ, Hecht L, Barron LD. Is polyproline II helix the killer conformation? a raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme 1 1Edited by A. R. Fersht. J Mol Biol. 2000;301: 553–563. doi:10.1006/jmbi.2000.3981

213. Syme CD, Blanch EW, Holt C, Jakes R, Goedert M, Hecht L, et al. A Raman optical activity study of rheomorphism in caseins, synucleins and tau. Eur J Biochem. 2002;269: 148–156. doi:10.1046/j.0014-2956.2001.02633.x

214. Adzhubei AA, Anashkina AA, Makarov AA. Left-handed polyproline-II helix revisited: proteins causing proteopathies. J Biomol Struct Dyn. 2017;35: 2701–2713. doi:10.1080/07391102.2016.1229220

215. Eiríksdóttir E, Konate K, Langel U, Divita G, Deshayes S. Secondary structure of cell-penetrating peptides controls membrane interaction and insertion. Biochim Biophys Acta. 2010;1798: 1119–1128. doi:10.1016/j.bbamem.2010.03.005

216. Franz J, Lelle M, Peneva K, Bonn M, Weidner T. SAP(E) – A cell-penetrating polyproline helix at lipid interfaces. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2016;1858: 2028–2034. doi:10.1016/j.bbamem.2016.05.021

217. Geisler I, Chmielewski J. Cationic amphiphilic polyproline helices: side-chain variations and cell-specific internalization. Chem Biol Drug Des. 2009;73: 39–45. doi:10.1111/j.1747-0285.2008.00759.x

218. Ruzza P, Calderan A, Guiotto A, Osler A, Borin G. Tat cell-penetrating peptide has the characteristics of a poly(proline) II helix in aqueous solution and in SDS micelles. J Pept Sci. 2004;10: 423–426. doi:10.1002/psc.558

219. Yamashita H, Kato T, Oba M, Misawa T, Hattori T, Ohoka N, et al. Development of a Cell-penetrating Peptide that Exhibits Responsive Changes in its Secondary Structure in the Cellular Environment. Sci Rep. 2016;6: 33003. doi:10.1038/srep33003

220. Fillon YA, Anderson JP, Chmielewski J. Cell penetrating agents based on a polyproline helix scaffold. J Am Chem Soc. 2005;127: 11798–11803. doi:10.1021/ja052377g

221. Li L, Geisler I, Chmielewski J, Cheng J-X. Cationic amphiphilic polyproline helix P11LRR targets intracellular mitochondria. J Control Release. 2010;142: 259–266. doi:10.1016/j.jconrel.2009.10.012

222. Foged C, Nielsen HM. Cell-penetrating peptides for drug delivery across membrane barriers. Expert Opin Drug Deliv. 2007;5: 105–117. doi:10.1517/17425247.5.1.105

223. Chebrek R, Leonard S, de Brevern AG, Gelly J-C. PolyprOnline: polyproline helix II and secondary structure assignment database. Database . 2014;2014. doi:10.1093/database/bau102

224. Narwani TJ, Santuz H, Shinada N, Melarkode Vattekatte A, Ghouzam Y, Srinivasan N, et al. Recent advances on polyproline II. Amino Acids. 2017;49: 705–713. doi:10.1007/s00726-017-2385-6

225. Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJE, Oliva B. Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. Proteins. 2005;60: 746–757. doi:10.1002/prot.20516

226. Hermoso A, Espadaler J, Enrique Querol E, Aviles FX, Sternberg MJE, Oliva B, et al. Including Functional Annotations and Extending the Collection of Structural Classifications of Protein Loops (ArchDB). Bioinform Biol Insights. 2009;1: 77–90. Available: https://www.ncbi.nlm.nih.gov/pubmed/20066127

227. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng. 1993;6: 377–382. Available: https://www.ncbi.nlm.nih.gov/pubmed/8332595

228. Tyagi M, Bornot A, Offmann B, de Brevern AG. Protein short loop prediction in terms of a structural alphabet. Comput Biol Chem. 2009;33: 329–333. doi:10.1016/j.compbiolchem.2009.06.002

229. Fourrier L, Benros C, de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics. 2004;5: 58. doi:10.1186/1471-2105-5-58

230. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins. 2003;51: 504–514. doi:10.1002/prot.10369

231. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins. 1989;5: 355–373. doi:10.1002/prot.340050410

232. Rooman MJ, Rodriguez J, Wodak SJ. Relations between protein sequence and structure and their significance. J Mol Biol. 1990;213: 337–350. doi:10.1016/S0022-2836(05)80195-0

233. Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, et al. A short survey on protein blocks. Biophys Rev. 2010;2: 137–145. doi:10.1007/s12551-010-0036-1

234. de Brevern AG, Valadié H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci. 2002;11: 2871–2886. doi:10.1110/ps.0220502

235. De Brevern AG, Etchebest C, Benros C, Hazout S. 'Pinning strategy': a novel approach for predicting the backbone structure in terms of protein blocks from sequence. J Biosci. 2007;32: 51–70. Available: https://www.ncbi.nlm.nih.gov/pubmed/17426380

236. de Brevern AG, Hazout S. 'Hybrid Protein Model' for optimally defining 3D protein structure fragments. Bioinformatics. 2003;19: 345–

353. doi:10.1093/bioinformatics/btf859

237. De Brevern AG, Hazout SA. Hybrid Protein Model (HPM): a method to compact protein 3D-structure information and physicochemical properties. Proceedings Seventh International Symposium on String Processing and Information Retrieval SPIRE 2000. doi:10.1109/spire.2000.878179

238. Bornot A, Etchebest C, de Brevern AG. A new prediction strategy for long local protein structures using an original description. Proteins. 2009;76: 570–587. doi:10.1002/prot.22370

239. Li Q, Zhou C, Liu H. Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. Proteins. 2009;74: 820–836. doi:10.1002/prot.22191

240. Faure G, Bornot A, de Brevern AG. Protein contacts, inter-residue interactions and side-chain modelling. Biochimie. 2008;90: 626–639. doi:10.1016/j.biochi.2007.11.007

241. de Brevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C, Etchebest C. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). Biochim Biophys Acta. 2005;1724: 288–306. doi:10.1016/j.bbagen.2005.05.016

242. de Brevern AG, Autin L, Colin Y, Bertrand O, Etchebest C. In silico studies on DARC. Infect Disord Drug Targets. 2009;9: 289–303. Available: https://www.ncbi.nlm.nih.gov/pubmed/19519483

243. Etchebest C, Benros C, Bornot A, Camproux A-C, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J. 2007;36: 1059–1069. doi:10.1007/s00249-007-0188-5

244. Zuo Y-C, Li Q-Z. Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids. 2010;38: 859–867. doi:10.1007/s00726-009-0292-1

245. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. Proteins. 2008;71: 920–937. doi:10.1002/prot.21776

246. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins. 2006;65: 32–39. doi:10.1002/prot.21087

247. Léonard S, Joseph AP, Srinivasan N, Gelly J-C, de Brevern AG. mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet. J Biomol Struct Dyn. 2014;32: 661–668. doi:10.1080/07391102.2013.787026

248. Dong Q-W, Wang X-L, Lin L. Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med. 2007;37: 1610–1616. doi:10.1016/j.compbiomed.2007.03.002

249. Thomas A, Deshayes S, Decaffmeyer M, Van Eyck MH, Charloteaux B, Brasseur R. Prediction of peptide structure: How far are we? Proteins: Struct Funct Bioinf. 2006;65: 889–897. doi:10.1002/prot.21151

250. Vetrivel I, Mahajan S, Tyagi M, Hoffmann L, Sanejouand Y-H, Srinivasan N, et al. Knowledge-based prediction of protein backbone conformation using a structural alphabet. PLoS One. 2017;12: e0186215. doi:10.1371/journal.pone.0186215

251. Mahajan S, de Brevern AG, Sanejouand Y-H, Srinivasan N, Offmann B. Use of a structural alphabet to find compatible folds for amino acid sequences. Protein Sci. 2015;24: 145–153. doi:10.1002/pro.2581

252. Ghouzam Y, Postic G, Guerin P-E, de Brevern AG, Gelly J-C. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. Sci Rep. 2016;6: 28268. doi:10.1038/srep28268

253. Ghouzam Y, Postic G, de Brevern AG, Gelly J-C. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. Bioinformatics. 2015;31: 3782–3789. doi:10.1093/bioinformatics/btv462

254. Goguet M, Narwani TJ, Petermann R, Jallu V, de Brevern AG. In silico analysis of Glanzmann variants of Calf-1 domain of αβ integrin revealed dynamic allosteric effect. Sci Rep. 2017;7: 8001. doi:10.1038/s41598-017-08408-w

255. Jallu V, Poulain P, Fuchs PFJ, Kaplan C, de Brevern AG. Modeling and Molecular Dynamics of HPA-1a and -1b Polymorphisms: Effects on the Structure of the β3 Subunit of the αIIbβ3 Integrin. PLoS One. 2012;7: e47304. doi:10.1371/journal.pone.0047304

256. Ladislav M, Cerny J, Krusek J, Horak M, Balik A, Vyklicky L. The LILI Motif of M3-S2 Linkers Is a Component of the NMDA Receptor Channel Gate. Front Mol Neurosci. 2018;11: 113. doi:10.3389/fnmol.2018.00113

257. Jonsson AL, Scott KA, Daggett V. Dynameomics: a consensus view of the protein unfolding/folding transition state ensemble across a diverse set of protein folds. Biophys J. 2009;97: 2958–2966. doi:10.1016/j.bpj.2009.09.012

258. van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, et al. Dynameomics: a comprehensive database of protein dynamics. Structure. 2010;18: 423–435. doi:10.1016/j.str.2010.01.012

259. Hensen U, Meyer T, Haas J, Rex R, Vriend G, Grubmüller H. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. PLoS One. 2012;7: e33931. doi:10.1371/journal.pone.0033931

260. Deribe YL, Pawson T, Dikic I. Post-translational modifications in signal integration. Nat Struct Mol Biol. 2010;17: 666–672. doi:10.1038/nsmb.1842

261. Zhao S, Xu W, Jiang W, Yu W, Lin Y, Zhang T, et al. Regulation of cellular metabolism by protein lysine acetylation. Science. 2010;327: 1000–1004. doi:10.1126/science.1179689

262. Duan G, Walther D. The roles of post-translational modifications in the context of protein interaction networks. PLoS Comput Biol. 2015;11: e1004049. doi:10.1371/journal.pcbi.1004049

263. Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol. 2012;13: 448–462. doi:10.1038/nrm3383

264. Humphrey SJ, James DE, Mann M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. Trends Endocrinol Metab. 2015;26: 676–687. doi:10.1016/j.tem.2015.09.013

265. Imberty A. Oligosaccharide structures: theory versus experiment. Curr Opin Struct Biol. 1997;7: 617–623. Available: https://www.ncbi.nlm.nih.gov/pubmed/9345618

266. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21: 381–395. doi:10.1038/cr.2011.22

267. Mijakovic I, Grangeasse C, Turgay K. Exploring the diversity of protein modifications: special bacterial phosphorylation systems. FEMS Microbiol Rev. 2016;40: 398–417. doi:10.1093/femsre/fuw003

268. McIntyre JC, Joiner AM, Zhang L, Iñiguez-Lluhí J, Martens JR. SUMOylation regulates ciliary localization of olfactory signaling proteins. J Cell Sci. 2015;128: 1934–1945. doi:10.1242/jcs.164673

269. Hendriks IA, Vertegaal ACO. A comprehensive compilation of SUMO proteomics. Nat Rev Mol Cell Biol. 2016;17: 581–595. doi:10.1038/nrm.2016.81

270. Imberty A, Pérez S. Stereochemistry of the N-glycosylation sites in glycoproteins. Protein Eng. 1995;8: 699–709. Available: https://www.ncbi.nlm.nih.gov/pubmed/8577698

271. Dewald JH, Colomb F, Bobowski-Gerard M, Groux-Degroote S, Delannoy P. Role of Cytokine-Induced Glycosylation Changes in Regulating Cell Interactions and Cell Signaling in Inflammatory Diseases and Cancer. Cells. 2016;5. doi:10.3390/cells5040043

272. Zhou B, Zeng L. Conventional and unconventional ubiquitination in plant immunity. Mol Plant Pathol. 2017;18: 1313–1330. doi:10.1111/mpp.12521

273. Krupa A, Preethi G, Srinivasan N. Structural modes of stabilization of permissive phosphorylation sites in protein kinases: distinct strategies in Ser/Thr and Tyr kinases. J Mol Biol. 2004;339: 1025–1039. doi:10.1016/j.jmb.2004.04.043

274. Li S, Iakoucheva LM, Mooney SD, Radivojac P. Loss of post-translational modification sites in disease. Pac Symp Biocomput. 2010; 337–347. Available: https://www.ncbi.nlm.nih.gov/pubmed/19908386

275. Martin L, Latypova X, Terro F. Post-translational modifications of tau protein: implications for Alzheimer's disease. Neurochem Int. 2011;58: 458–471. doi:10.1016/j.neuint.2010.12.023

276. Gong C-X, Liu F, Grundke-Iqbal I, Iqbal K. Post-translational modifications of tau protein in Alzheimer's disease. J Neural Transm. 2004;112: 813–838. doi:10.1007/s00702-004-0221-0

277. Zeidan Q, Hart GW. The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. J Cell Sci. 2010;123: 13–22. doi:10.1242/jcs.053678

278. Vodermaier HC. APC/C and SCF: controlling each other and the cell cycle. Curr Biol. 2004;14: R787–96. doi:10.1016/j.cub.2004.09.020

279. Latham JA, Dent SYR. Cross-regulation of histone modifications. Nat Struct Mol Biol. 2007;14: 1017–1024. doi:10.1038/nsmb1307

280. Darmawan D. Biological Communication Behavior through Information Technology Implementation in Learning Accelerated. International Journal of Communications, Network and System Sciences. 2012;05: 454–462. doi:10.4236/ijcns.2012.58056

281. Creixell P, Linding R. Cells, shared memory and breaking the PTM code. Mol Syst Biol. 2012;8. doi:10.1038/msb.2012.33

282. Minguez P, Bork P. Bioinformatics Analysis of Functional Associations of PTMs. Methods Mol Biol. 2017;1558: 303–320. doi:10.1007/978-1-4939-6783-4_14

283. Nussinov R, Tsai C-J, Xin F, Radivojac P. Allosteric post-translational modification codes. Trends Biochem Sci. 2012;37: 447–455. doi:10.1016/j.tibs.2012.07.001

284. Lu Z, Cheng Z, Zhao Y, Volchenboum SL. Bioinformatic analysis and post-translational modification crosstalk prediction of lysine acetylation. PLoS One. 2011;6: e28228. doi:10.1371/journal.pone.0028228

285. Tokmakov AA, Kurotani A, Takagi T, Toyama M, Shirouzu M, Fukami Y, et al. Multiple post-translational modifications affect heterologous protein synthesis. J Biol Chem. 2012;287: 27106–27116. doi:10.1074/jbc.M112.366351

286. van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, et al. Cross-talk between phosphorylation and lysine acetylation

in a genome-reduced bacterium. Mol Syst Biol. 2012;8. doi:10.1038/msb.2012.4

287. Danielsen JMR, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, et al. Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. Mol Cell Proteomics. 2011;10: M110.003590. doi:10.1074/mcp.M110.003590

288. Minguez P, Letunic I, Parca L, Bork P. PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. Nucleic Acids Res. 2013;41: D306–11. doi:10.1093/nar/gks1230

289. Gu B, Zhu W-G. Surf the post-translational modification network of p53 regulation. Int J Biol Sci. 2012;8: 672–684. doi:10.7150/ijbs.4283

290. Xin F, Radivojac P. Post-translational modifications induce significant yet not extreme changes to protein structure. Bioinformatics. 2012;28: 2905–2913. doi:10.1093/bioinformatics/bts541

291. Gao J, Xu D. Correlation between posttranslational modification and intrinsic disorder in protein. Pac Symp Biocomput. 2012; 94–103. Available: https://www.ncbi.nlm.nih.gov/pubmed/22174266

292. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res. 2007;6: 1917–1932. doi:10.1021/pr060394e

293. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res. 2007;6: 1882–1898. doi:10.1021/pr060392u

294. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J Proteome Res. 2007;6: 1899–1916. doi:10.1021/pr060393m

295. Betts MJ, Wichmann O, Utz M, Andre T, Petsalaki E, Minguez P, et al. Systematic identification of phosphorylation-mediated protein interaction switches. PLoS Comput Biol. 2017;13: e1005462. doi:10.1371/journal.pcbi.1005462

296. Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11: 427–439. doi:10.1038/nrm2900

297. Gianazza E, Parravicini C, Primi R, Miller I, Eberini I. In silico prediction and characterization of protein post-translational modifications. J Proteomics. 2016;134: 65–75. doi:10.1016/j.jprot.2015.09.026

298. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43: D512–20. doi:10.1093/nar/gku1267

299. Huang K-Y, Su M-G, Kao H-J, Hsieh Y-C, Jhong J-H, Cheng K-H, et al. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. Nucleic Acids Res. 2016;44: D435–46. doi:10.1093/nar/gkv1240

300. Audagnotto M, Dal Peraro M. Protein post-translational modifications: prediction tools and molecular modeling. Comput Struct Biotechnol J. 2017;15: 307–319. doi:10.1016/j.csbj.2017.03.004

301. López Y, Dehzangi A, Lal SP, Taherzadeh G, Michaelson J, Sattar A, et al. SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. Anal Biochem. 2017;527: 24–32. doi:10.1016/j.ab.2017.03.021

302. Lorenzo JR, Alonso LG, Sánchez IE. Prediction of Spontaneous Protein Deamidation from Sequence-Derived Secondary Structure and Intrinsic Disorder. PLoS One. 2015;10: e0145186. doi:10.1371/journal.pone.0145186

303. Wuyun Q, Zheng W, Zhang Y, Ruan J, Hu G. Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set. PLoS One. 2016;11: e0155370. doi:10.1371/journal.pone.0155370

304. Torres MP, Dewhurst H, Sundararaman N. Proteome-wide Structural Analysis of PTM Hotspots Reveals Regulatory Elements Predicted to Impact Biological Function and Disease. Mol Cell Proteomics. 2016;15: 3513–3528. doi:10.1074/mcp.M116.062331

305. PTM Structural Database [Internet]. [cited 6 May 2018]. Available: http://www.dsimb.inserm.fr/dsimb_tools/PTM-SD/

306. Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep. 2011;1. doi:10.1038/srep00090

307. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure. 2003;11: 1453–1459. Available: https://www.ncbi.nlm.nih.gov/pubmed/14604535

308. Hinsen K. Structural flexibility in proteins: impact of the crystal environment. Bioinformatics. 2008;24: 521–528. doi:10.1093/bioinformatics/btm625

309. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci. 2003;12: 1060–1072. doi:10.1110/ps.0236203

310. Huber PJ. Robust Statistical Procedures: Second Edition [Internet]. SIAM; 1996. Available: https://market.android.com/details?id=book-9xyux0TiJ60C

311. Jurečková J, Picek J. Robust Statistical Methods with R [Internet]. CRC Press; 2005. Available: https://market.android.com/details?id=book-dhIEwbmm1c4C

312. Hothorn T, Everitt BS. A Handbook of Statistical Analyses using R, Third Edition [Internet]. CRC Press; 2014. Available: https://books.google.com/books/about/A_Handbook_of_Statistical_Analyses_using.html?hl=&id=cuTMAwAAQBAJ

313. Zhang Y, Stec B, Godzik A. Between order and disorder in protein structures: analysis of 'dual personality' fragments in proteins. Structure. 2007;15: 1141–1147. doi:10.1016/j.str.2007.07.012

314. Groban ES, Narayanan A, Jacobson MP. Conformational changes in protein loops and helices induced by post-translational phosphorylation. PLoS Comput Biol. 2006;2: e32. doi:10.1371/journal.pcbi.0020032

315. DeForte S, Uversky VN. Order, Disorder, and Everything in Between. Molecules. 2016;21. doi:10.3390/molecules21081090

316. Hsu W-L, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, et al. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci. 2013;22: 258–273. doi:10.1002/pro.2207

317. Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT Jr. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry. 1996;35: 13709–13715. doi:10.1021/bi961799n

318. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293: 321–331. doi:10.1006/jmbi.1999.3110

319. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 2005;579: 3346–3354. doi:10.1016/j.febslet.2005.03.072

320. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem Sci. 2008;33: 2–8. doi:10.1016/j.tibs.2007.10.003

321. Schweers O, Schönbrunn-Hanebeck E, Marx A, Mandelkow E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. J Biol Chem. 1994;269: 24290–24297. Available: https://www.ncbi.nlm.nih.gov/pubmed/7929085

322. Dunker AK, Oldfield CJ. Back to the Future: Nuclear Magnetic Resonance and Bioinformatics Studies on Intrinsically Disordered Proteins. Adv Exp Med Biol. 2015;870: 1–34. doi:10.1007/978-3-319-20164-1_1

323. Thomas WH, Weser U, Hempel K. Conformational changes induced by ionic strength and pH in two bovine myelin basic proteins. Hoppe Seylers Z Physiol Chem. 1977;358: 1345–1352. Available: https://www.ncbi.nlm.nih.gov/pubmed/21842

324. Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK. Intrinsic disorder in the Protein Data Bank. J Biomol Struct Dyn. 2007;24: 325–342. doi:10.1080/07391102.2007.10507123

325. Szilágyi A, Györffy D, Závodszky P. The twilight zone between protein order and disorder. Biophys J. 2008;95: 1612–1626. doi:10.1529/biophysj.108.131151

326. Balaji S. PALI--a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res. 2001;29: 61–65. doi:10.1093/nar/29.1.61

327. Dunker AK, Keith Dunker A. Another Window into Disordered Protein Function. Structure. 2007;15: 1026–1028. doi:10.1016/j.str.2007.08.001

328. Marcos E, Crehuet R, Bahar I. Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members. PLoS Comput Biol. 2011;7: e1002201. doi:10.1371/journal.pcbi.1002201

329. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014;508: 331–339. doi:10.1038/nature13001

330. Hynes RO. Integrins: bidirectional, allosteric signaling machines. Cell. 2002;110: 673–687. Available: https://www.ncbi.nlm.nih.gov/pubmed/12297042

331. Calderwood DA. Integrin activation. J Cell Sci. 2004;117: 657–666. doi:10.1242/jcs.01014

332. Miranti CK, Brugge JS. Sensing the environment: a historical perspective on integrin signal transduction. Nat Cell Biol. 2002;4: E83–90. doi:10.1038/ncb0402-e83

333. Coller BS, Shattil SJ. The GPIIb/IIIa (integrin alphaIIbbeta3) odyssey: a technology-driven saga of a receptor with twists, turns, and even a bend. Blood. 2008;112: 3011–3025. doi:10.1182/blood-2008-06-077891

334. Mehrbod M, Trisno S, Mofrad MRK. On the activation of integrin αIIbβ3: outside-in and inside-out pathways. Biophys J. 2013;105: 1304–1315. doi:10.1016/j.bpj.2013.07.055

335. Zhu J, Zhu J, Springer TA. Complete integrin headpiece opening in eight steps. J Cell Biol. 2013;201: 1053–1068. doi:10.1083/jcb.201212037

336. Zhang K, Chen J. The regulation of integrin function by divalent cations. Cell Adh Migr. 2012;6: 20–29. doi:10.4161/cam.18702

337. Xiao T, Takagi J, Coller BS, Wang J-H, Springer TA. Structural basis for allostery in integrins and binding of ligand-mimetic therapeutics to the platelet receptor for fibrinogen [Internet]. 2004. doi:10.2210/pdb1ty7/pdb

338. Nurden AT, Pillois X, Wilcox DA. Glanzmann thrombasthenia: state of the art and future directions. Semin Thromb Hemost. 2013;39: 642–655. doi:10.1055/s-0033-1353393

339. Curtis BR. Recent progress in understanding the pathogenesis of fetal and neonatal alloimmune thrombocytopenia. Br J Haematol. 2015;171: 671–682. doi:10.1111/bjh.13639

340. Jallu V, Dusseaux M, Panzer S, Torchet M-F, Hezard N, Goudemand J, et al. AlphaIIbbeta3 integrin: new allelic variants in Glanzmann thrombasthenia, effects on ITGA2B and ITGB3 mRNA splicing, expression, and structure-function. Hum Mutat. 2010;31: 237–246. doi:10.1002/humu.21179

341. Jallu V, Poulain P, Fuchs P-F-J, Kaplan C, De Brevern A-G. Modélisation et simulation de dynamique moléculaire du variant V33 de la sous-unité β3 des intégrines : comparaison structurale avec les allèles 1a (L33) et 1b (P33) du système alloantigénique plaquettaire HPA-1. Transfus Clin Biol. 2013;20: 278. doi:10.1016/j.tracli.2013.04.087

342. Tsai C-J, del Sol A, Nussinov R. Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. J Mol Biol. 2008;378: 1–11. doi:10.1016/j.jmb.2008.02.034

343. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins. 2004;57: 433–443. doi:10.1002/prot.20232

344. Zhu J, Luo B-H, Xiao T, Zhang C, Nishida N, Springer TA. Structure of a Complete Integrin Ectodomain in a Physiologic Resting State and Activation and Deactivation by Applied Forces. Mol Cell. 2008;32: 849–861. doi:10.1016/j.molcel.2008.11.018

345. Veeramalai M, Ye Y, Godzik A. TOPS FATCAT: Fast flexible structural alignment using constraints derived from TOPS Strings Model. BMC Bioinformatics. 2008;9: 358. doi:10.1186/1471-2105-9-358

346. Pillitteri D, Pilgrimm A-K, Kirchmaier CM. Novel Mutations in the GPIIb and GPIIIa Genes in Glanzmann Thrombasthenia. Transfus Med Hemother. 2010;37: 268–277. doi:10.1159/000320258

347. D'Andrea G, Colaizzo D, Vecchione G, Grandone E, Di Minno G, Margaglione M, et al. Glanzmann's Thrombasthenia: Identification of 19 New Mutations in 30 Patients. Thromb Haemost. 2002;87: 1034–1042. doi:10.1055/s-0037-1613129

348. Vijapurkar M, Ghosh K, Shetty S. Novel mutations in GP IIb gene in Glanzmann's thrombasthenia from India. Platelets. 2009;20: 35–40. doi:10.1080/09537100802434861

349. Nurden AT, Pillois X, Fiore M, Alessi M-C, Bonduel M, Dreyfus M, et al. Expanding the Mutation Spectrum Affecting αIIbβ3 Integrin in Glanzmann Thrombasthenia: Screening of theITGA2BandITGB3Genes in a Large International Cohort. Hum Mutat. 2015;36: 548–561. doi:10.1002/humu.22776

350. Franchini M, Favaloro EJ, Lippi G. Glanzmann thrombasthenia: an update. Clin Chim Acta. 2010;411: 1–6. doi:10.1016/j.cca.2009.10.016

351. George JN, Caen JP, Nurden AT. Glanzmann's thrombasthenia: the spectrum of clinical disease. Blood. 1990;75: 1383–1395. Available:

https://www.ncbi.nlm.nih.gov/pubmed/2180491

352. Lill MA, Danielson ML. Computer-aided drug design platform using PyMOL. J Comput Aided Mol Des. 2010;25: 13–19. doi:10.1007/s10822-010-9395-8

353. Dunbrack R. SCWRL. Dictionary of Bioinformatics and Computational Biology. 2004. doi:10.1002/9780471650126.dob0654.pub2

354. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. J Comput Chem. 2005;26: 1701–1718. doi:10.1002/jcc.20291

355. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, et al. The GROMOS Biomolecular Simulation Program Package. J Phys Chem A. 1999;103: 3596–3607. doi:10.1021/jp984217f

356. Tina KG, Bhadra R, Srinivasan N. PIC: Protein Interactions Calculator. Nucleic Acids Res. 2007;35: W473–W476. doi:10.1093/nar/gkm423

357. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015;43: D376–81. doi:10.1093/nar/gku947

358. Gelly J-C, de Brevern AG. Protein Peeling 3D: new tools for analyzing protein structures. Bioinformatics. 2011;27: 132–133. doi:10.1093/bioinformatics/btq610

359. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol. 1991;218: 397–412. doi:10.1016/0022-2836(91)90721-h

360. Warrell DA, Hemingway J, Marsh K, Sinden RE, Butcher GA, Snow RW. Malaria. Oxford Textbook of Medicine. 2010. pp. 1046–1088. doi:10.1093/med/9780199204854.003.070802

361. Mendis K, Marchesini P, Carter R, Sina BJ. The neglected burden of Plasmodium vivax malaria. Am J Trop Med Hyg. 2001;64: 97–106. doi:10.4269/ajtmh.2001.64.97

362. Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, Guerra CA, et al. A long neglected world malaria map: Plasmodium vivax endemicity in 2010. PLoS Negl Trop Dis. 2012;6: e1814. doi:10.1371/journal.pntd.0001814

363. Carter R, Mendis KN. Evolutionary and Historical Aspects of the Burden of Malaria. Clin Microbiol Rev. 2002;15: 564–594. doi:10.1128/CMR.15.4.564-594.2002

364. Welch SG, McGregor IA, Williams K. The Duffy blood group and malaria prevalence in Gambian West Africans. Trans R Soc Trop Med Hyg. 1977;71: 295–296. Available: https://www.ncbi.nlm.nih.gov/pubmed/339418

365. Hulden L, Hulden L. Activation of the hypnozoite: a part of Plasmodium vivax life cycle and survival. Malar J. 2011;10: 90. doi:10.1186/1475-2875-10-90

366. Tournamille C, Le Van Kim C, Gane P, Cartron JP, Colin Y. Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism. Hum Genet. 1995;95: 407–410. Available: https://www.ncbi.nlm.nih.gov/pubmed/7705836

367. Cutbush M, Mollison PL. The Duffy blood group system. Heredity . 1950;4: 383–389. Available: https://www.ncbi.nlm.nih.gov/pubmed/14802995

368. Dracopoli NC, O'Connell P, Elsner TI, Lalouel J-M, White RL, Buetow KH, et al. The CEPH consortium linkage map of human chromosome 1. Genomics. 1991;9: 686–700. doi:10.1016/0888-7543(91)90362-i

369. Chaudhuri A, Polyakova J, Zbrzezna V, Williams K, Gulati S, Pogo AO. Cloning of glycoprotein D cDNA, which encodes the major subunit of the Duffy blood group system and the receptor for the Plasmodium vivax malaria parasite. Proceedings of the National Academy of Sciences. 1993;90: 10793–10797. doi:10.1073/pnas.90.22.10793

370. Mallinson G, Soo KS, Schall TJ, Pisacka M, Anstee DJ. Mutations in the erythrocyte chemokine receptor (Duffy) gene: the molecular basis of the Fya/Fyb antigens and identification of a deletion in the Duffy gene of an apparently healthy individual with the Fy(a-b-) phenotype. Br J Haematol. 1995;90: 823–829. Available: https://www.ncbi.nlm.nih.gov/pubmed/7669660

371. Olsson ML, Smythe JS, Hansson C, Poole J, Mallinson G, Jones J, et al. The Fyx phenotype is associated with a missense mutation in the Fyb allele predicting Arg89Cys in the Duffy glycoprotein. Br J Haematol. 1998;103: 1184–1191. doi:10.1046/j.1365-2141.1998.01083.x

372. Tournamille C, Colin Y, Cartron JP, Van Kim CL. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. Nat Genet. 1995;10: 224–228. doi:10.1038/ng0695-224

373. Peiper SC, Wang ZX, Neote K, Martin AW, Showell HJ, Conklyn MJ, et al. The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. J Exp Med. 1995;181: 1311–1317. Available: https://www.ncbi.nlm.nih.gov/pubmed/7699323

374. Horuk R, Martin A, Hesselgesser J, Hadley T, Lu ZH, Wang ZX, et al. The Duffy antigen receptor for chemokines: structural analysis and expression in the brain. J Leukoc Biol. 1996;59: 29–38. Available: https://www.ncbi.nlm.nih.gov/pubmed/8558064

375. Mohinta S, Watabe K. Duffy Antigen Receptor for Chemokines. Encyclopedia of Cancer. Springer, Berlin, Heidelberg; 2008. pp. 918–919. doi:10.1007/978-3-540-47648-1_1750

376. Patel J, Channon KM, McNeill E. The downstream regulation of chemokine receptor signalling: implications for atherosclerosis. Mediators Inflamm. 2013;2013: 459520. doi:10.1155/2013/459520

377. Stephens B, Handel TM. Chemokine receptor oligomerization and allostery. Prog Mol Biol Transl Sci. 2013;115: 375–420. doi:10.1016/B978-0-12-394587-7.00009-9

378. Stone MJ, Hayward JA, Huang C, E Huma Z, Sanchez J. Mechanisms of Regulation of the Chemokine-Receptor Network. Int J Mol Sci. 2017;18. doi:10.3390/ijms18020342

379. Scheerer P, Park JH, Hildebrand PW, Kim YJ, Krauß N, Choe H-W, et al. Crystal structure of opsin in its G-protein-interacting conformation. Nature. 2008;455: 497–502. doi:10.1038/nature07330

380. Rasmussen SG, Jensen AD, Liapakis G, Ghanouni P, Javitch JA, Gether U. Mutation of a highly conserved aspartic acid in the beta2 adrenergic receptor: constitutive activation, structural instability, and conformational rearrangement of transmembrane segment 6. Mol Pharmacol. 1999;56: 175–184. Available: https://www.ncbi.nlm.nih.gov/pubmed/10385699

381. Nomiyama H, Yoshie O. Functional roles of evolutionary conserved motifs and residues in vertebrate chemokine receptors. J Leukoc Biol. 2015;97: 39–47. doi:10.1189/jlb.2RU0614-290R

382. Heuss C, Gerber U. G-protein-independent signaling by G-protein-coupled receptors. Trends Neurosci. 2000;23: 469–475. doi:10.1016/s0166-2236(00)01643-x

383. Borroni EM, Cancellieri C, Vacchini A, Benureau Y, Lagane B, Bachelerie F, et al. β-arrestin-dependent activation of the cofilin pathway is required for the scavenging activity of the atypical chemokine receptor D6. Sci Signal. 2013;6: ra30.1–11, S1–3. doi:10.1126/scisignal.2003627

384. Bachelerie F, Ben-Baruch A, Burkhardt AM, Combadiere C, Farber JM, Graham GJ, et al. International Union of Pharmacology. LXXXIX. Update on the Extended Family of Chemokine Receptors and Introducing a New Nomenclature for Atypical Chemokine Receptors. Pharmacol Rev. American Society for Pharmacology and Experimental Therapeutics; 2014;66: 1. doi:10.1124/pr.113.007724

385. Novitzky-Basso I, Rot A. Duffy antigen receptor for chemokines and its involvement in patterning and control of inflammatory chemokines. Front Immunol. 2012;3: 266. doi:10.3389/fimmu.2012.00266

386. Rot A. Contribution of Duffy antigen to chemokine function. Cytokine Growth Factor Rev. 2005;16: 687–694. doi:10.1016/j.cytogfr.2005.05.011

387. Dzik S, Reid ME, Freedman JJ. The coding sequence of duffy blood group gene in humans and simians: Restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissue in duffy-negative individuals. Transfus Med Rev. 1996;10: 152. doi:10.1016/s0887-7963(96)80091-5

388. Pruenster M, Mudde L, Bombosi P, Dimitrova S, Zsak M, Middleton J, et al. The Duffy antigen receptor for chemokines transports chemokines and supports their promigratory activity. Nat Immunol. 2009;10: 101–108. doi:10.1038/ni.1675

389. Batchelor JD, Malpede BM, Omattage NS, DeKoster GT, Henzler-Wildman KA, Tolia NH. Red blood cell invasion by Plasmodium vivax: structural basis for DBP engagement of DARC. PLoS Pathog. 2014;10: e1003869. doi:10.1371/journal.ppat.1003869

390. Singh SK, Hora R, Belrhali H, Chitnis CE, Sharma A. Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. Nature. 2006;439: 741–744. doi:10.1038/nature04443

391. GPCR-RD: GPCR experimental restaint database [Internet]. [cited 29 Mar 2018]. Available: https://zhanglab.ccmb.med.umich.edu/GPCR-RD/

392. Raucci R, Costantini S, Castello G, Colonna G. An overview of the sequence features of N- and C-terminal segments of the human chemokine receptors. Cytokine. 2014;70: 141–150. doi:10.1016/j.cyto.2014.07.257

393. Ludeman JP, Stone MJ. The structural role of receptor tyrosine sulfation in chemokine recognition. Br J Pharmacol. 2014;171: 1167–1179. doi:10.1111/bph.12455

394. Wu B, Mol CD, Han GW, Katritch V, Chien EYT, Liu W, et al. Crystal structure of the chemokine CXCR4 receptor in complex with a small molecule antagonist IT1t in P1 spacegroup [Internet]. 2010. doi:10.2210/pdb3oe9/pdb

395. Wu B, Mol CD, Han GW, Katritch V, Chien EYT, Liu W, et al. The 2.5 A structure of the CXCR4 chemokine receptor in complex with small molecule antagonist IT1t [Internet]. 2010. doi:10.2210/pdb3odu/pdb

396. Wu B, Mol CD, Han GW, Katritch V, Chien EYT, Liu W, et al. Crystal structure of the CXCR4 chemokine receptor in complex with a cyclic peptide antagonist CVX15 [Internet]. 2010. doi:10.2210/pdb3oe0/pdb

397. Wu B, Mol CD, Han GW, Katritch V, Chien EYT, Liu W, et al. Crystal structure of the CXCR4 chemokine receptor in complex with a small molecule antagonist IT1t in P1 spacegroup [Internet]. 2010. doi:10.2210/pdb3oe8/pdb

398. Chakera A, Seeber RM, John AE, Eidne KA, Greaves DR. The duffy antigen/receptor for chemokines exists in an oligomeric form in living cells and functionally antagonizes CCR5 signaling through hetero-oligomerization. Mol Pharmacol. 2008;73: 1362–1370. doi:10.1124/mol.107.040915

399. Murphy PM. Chemokine Receptors. xPharm: The Comprehensive Pharmacology Reference. 2007. pp. 1–5. doi:10.1016/b978-008055232-3.60183-7

400. Bendall L. Chemokines and their receptors in disease. Histol Histopathol. 2005;20: 907–926. doi:10.14670/HH-20.907

401. Altschul S. Basic Local Alignment Search Tool. J Mol Biol. 1990;215: 403–410. doi:10.1006/jmbi.1990.9999

402. HMMER [Internet]. [cited 3 Apr 2018]. Available: www.hmmer.org

403. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. Nat Biotechnol. 2008;26: 274–275. doi:10.1038/nbt0308-274

404. Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309. doi:10.1093/nar/gki524

405. www.bioinf.org.uk : Dr. Andrew C.R. Martin's Group at UCL [Internet]. [cited 3 Apr 2018]. Available: http://www.bioinf.org.uk/software/profit/index.html

406. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

407. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004;20: 426–427. doi:10.1093/bioinformatics/btg430

408. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44: W232–5. doi:10.1093/nar/gkw256

409. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242–5. doi:10.1093/nar/gkw290

410. Pei J, Tang M, Grishin NV. PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36: W30–4. doi:10.1093/nar/gkn322

411. Eramian D, Shen M-Y, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci. 2006;15: 1653–1666. doi:10.1110/ps.062095806

412. Postic G, Ghouzam Y, Gelly J-C. An empirical energy function for structural assessment of protein transmembrane domains. Biochimie. 2015;115: 155–161. doi:10.1016/j.biochi.2015.05.018

413. .: PDBTM: Protein Data Bank of Transmembrane Proteins : [Internet]. [cited 3 Apr 2018]. Available: http://pdbtm.enzim.hu/

414. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput. 2011;7: 525–537. doi:10.1021/ct100578z

415. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, et al. PDB2PQR: expanding and upgrading automated preparation of

biomolecular structures for molecular simulations. Nucleic Acids Res. 2007;35: W522–W525. doi:10.1093/nar/gkm276

416. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012;40: D370–6. doi:10.1093/nar/gkr703

417. Jo S, Lim JB, Klauda JB, Im W. CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. Biophys J. 2009;97: 50–58. doi:10.1016/j.bpj.2009.04.013

418. Gedde MM, Yang E, Huestis WH. Shape response of human erythrocytes to altered cell pH. Blood. 1995;86: 1595–1599. Available: https://www.ncbi.nlm.nih.gov/pubmed/7632969

419. Bretscher MS. Asymmetrical lipid bilayer structure for biological membranes. Nat New Biol. 1972;236: 11–12. Available: https://www.ncbi.nlm.nih.gov/pubmed/4502419

420. The asymmetric arrangement of phospholipids in the human erythrocyte membrane. Biochem Biophys Res Commun. Academic Press; 1973;50: 1027–1031. doi:10.1016/0006-291X(73)91509-X

421. Steck TL. Red Cell Shape. Cell Shape. 1989. pp. 205–246. doi:10.1016/b978-0-12-664655-9.50011-7

422. Daleke DL. Erythrocyte morphology reflects the transbilayer distribution of incorporated phospholipids. J Cell Biol. 1989;108: 1375–1385. doi:10.1083/jcb.108.4.1375

423. Borochov H, Zahler P, Wilbrandt W, Shinitzky M. The effect of phosphatidylcholine to sphingomyelin mole ratio on the dynamic properties of sheep erythrocyte membrane. Biochim Biophys Acta. 1977;470: 382–388. Available: https://www.ncbi.nlm.nih.gov/pubmed/410447

424. Ballas SK, Burka ER. Pathways of de novo phospholipid synthesis in reticulocytes. Biochim Biophys Acta. 1974;337: 239–247. Available: https://www.ncbi.nlm.nih.gov/pubmed/4433549

425. Hanahan DJ. Chemical Composition of Membranes. Concepts and Models. 1978. pp. 205–237. doi:10.1007/978-3-642-46370-9_6

426. Smith JE. Erythrocyte membrane: structure, function, and pathophysiology. Vet Pathol. 1987;24: 471–476. doi:10.1177/030098588702400601

427. Maravillas-Montero JL, Burkhardt AM, Hevezi PA, Carnevale CD, Smit MJ, Zlotnik A. Cutting edge: GPR35/CXCR8 is the receptor of the mucosal chemokine CXCL17. J Immunol. 2015;194: 29–33. doi:10.4049/jimmunol.1401704

428. Tanner MJA. Erythrocyte Membrane Structure and Function. Novartis Foundation Symposia. 2008. pp. 3–23. doi:10.1002/9780470715444.ch2

429. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys. 1977;23: 327–341. doi:10.1016/0021-9991(77)90098-5

430. Tournamille C, Van Kim CL, Gane P, Blanchard D, Proudfoot AE, Cartron JP, et al. Close Association of the First and Fourth Extracellular Domains of the Duffy Antigen/Receptor for Chemokines by a Disulfide Bond Is Required for Ligand Binding. J Biol Chem. 1997;272: 16274–16280. doi:10.1074/jbc.272.26.16274

431. Cutbush M, Mollison PL, Parkin DM. A New Human Blood Group. Nature. 1950;165: 188–189. doi:10.1038/165188b0

432. Horuk R, Chitnis C, Darbonne W, Colby T, Rybicki A, Hadley T, et al. A receptor for the malarial parasite Plasmodium vivax: the erythrocyte chemokine receptor. Science. 1993;261: 1182–1184. doi:10.1126/science.7689250

433. Chitnis CE. The domain on the Duffy blood group antigen for binding Plasmodium vivax and P. knowlesi malarial parasites to erythrocytes. J Exp Med. 1996;184: 1531–1536. doi:10.1084/jem.184.4.1531

434. Hans D, Pattnaik P, Bhattacharyya A, Shakri AR, Yazdani SS, Sharma M, et al. Mapping binding residues in the Plasmodium vivax domain that binds Duffy antigen during red cell invasion. Mol Microbiol. 2005;55: 1423–1434. doi:10.1111/j.1365-2958.2005.04484.x

435. Bhardwaj R, Shakri AR, Hans D, Gupta P, Fernandez-Becerra C, Del Portillo HA, et al. Production of recombinant PvDBPII, receptor binding domain of Plasmodium vivax Duffy binding protein, and evaluation of immunogenicity to identify an adjuvant formulation for vaccine development. Protein Expr Purif. 2017;136: 52–57. doi:10.1016/j.pep.2015.06.011

436. Tournamille C, Blancher A, Le Van Kim C, Gane P, Apoil PA, Nakamoto W, et al. Sequence, evolution and ligand binding properties of mammalian Duffy antigen/receptor for chemokines. Immunogenetics. 2004;55: 682–694. doi:10.1007/s00251-003-0633-2

437. Hub E, Rot A. Binding of RANTES, MCP-1, MCP-3, and MIP-1alpha to cells in human skin. Am J Pathol. 1998;152: 749–757. Available: https://www.ncbi.nlm.nih.gov/pubmed/9502417

438. Ratner AJ. S. aureus Toxins Join the DARC Side. Cell Host Microbe. 2015;18: 272–274. doi:10.1016/j.chom.2015.08.010

439. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science. 1996;272: 872–877. Available: https://www.ncbi.nlm.nih.gov/pubmed/8629022

440. Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper SC, et al. A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. Cell. 1996;85: 1149–1158. Available: https://www.ncbi.nlm.nih.gov/pubmed/8674120

441. Rosenkilde MM, Smit MJ, Waldhoer M. Structure, function and physiological consequences of virally encoded chemokine seven transmembrane receptors. Br J Pharmacol. 2008;153 Suppl 1: S154–66. doi:10.1038/sj.bjp.0707660

442. Nakano K, Tadagaki K, Isegawa Y, Aye MM, Zou P, Yamanishi K. Human Herpesvirus 7 Open Reading Frame U12 Encodes a Functional -Chemokine Receptor. J Virol. 2003;77: 8108–8115. doi:10.1128/jvi.77.14.8108-8115.2003

443. Ahuja SK, Gao J-L, Murphy PM. Chemokine receptors and molecular mimicry. Immunol Today. 1994;15: 281–287. doi:10.1016/0167-5699(94)90008-6

444. Fernandez EJ, Lolis E. STRUCTURE, FUNCTION,ANDINHIBITION OFCHEMOKINES. Annu Rev Pharmacol Toxicol. 2002;42: 469–499. doi:10.1146/annurev.pharmtox.42.091901.115838

445. Zlotnik A, Yoshie O. The chemokine superfamily revisited. Immunity. 2012;36: 705–716. doi:10.1016/j.immuni.2012.05.008

446. Le Y, Zhou Y, Iribarren P, Wang J. Chemokines and chemokine receptors: their manifold roles in homeostasis and disease. Cell Mol Immunol. 2004;1: 95–104. Available: https://www.ncbi.nlm.nih.gov/pubmed/16212895

447. Zlotnik A, Yoshie O. Chemokines: a new classification system and their role in immunity. Immunity. 2000;12: 121–127. Available: https://www.ncbi.nlm.nih.gov/pubmed/10714678

448. Gortz A, Nibbs RJB, McLean P, Jarmin D, Lambie W, Baird JW, et al. The chemokine ESkine/CCL27 displays novel modes of intracrine

and paracrine function. J Immunol. 2002;169: 1387–1394. Available: https://www.ncbi.nlm.nih.gov/pubmed/12133963

449. Ye J, Kohli LL, Stone MJ. Characterization of binding between the chemokine eotaxin and peptides derived from the chemokine receptor CCR3. J Biol Chem. 2000;275: 27250–27257. doi:10.1074/jbc.M003925200

450. Baggiolini M, Dewald B, Moser B. Human chemokines: an update. Annu Rev Immunol. 1997;15: 675–705. doi:10.1146/annurev.immunol.15.1.675

451. Weber M. Deletion of the NH2-terminal residue converts monocyte chemotactic protein 1 from an activator of basophil mediator release to an eosinophil chemoattractant. J Exp Med. 1996;183: 681–685. doi:10.1084/jem.183.2.681

452. Simmons G, Clapham PR, Picard L, Offord RE, Rosenkilde MM, Schwartz TW, et al. Potent inhibition of HIV-1 infectivity in macrophages and lymphocytes by a novel CCR5 antagonist. Science. 1997;276: 276–279. Available: https://www.ncbi.nlm.nih.gov/pubmed/9092481

453. Butcher EC. Leukocyte-endothelial cell recognition: three (or more) steps to specificity and diversity. Cell. 1991;67: 1033–1036. Available: https://www.ncbi.nlm.nih.gov/pubmed/1760836

454. Springer TA. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. Cell. 1994;76: 301–314. Available: https://www.ncbi.nlm.nih.gov/pubmed/7507411

455. Bleul CC, Schultze JL, Springer TA. B Lymphocyte Chemotaxis Regulated in Association with Microanatomic Localization, Differentiation State, and B Cell Receptor Engagement. J Exp Med. 1998;187: 753–762. doi:10.1084/jem.187.5.753

456. Dieu M-C, Vanbervliet B, Vicari A, Bridon J-M, Oldham E, Aït-Yahia S, et al. Selective Recruitment of Immature and Mature Dendritic Cells by Distinct Chemokines Expressed in Different Anatomic Sites. J Exp Med. 1998;188: 373–386. doi:10.1084/jem.188.2.373

457. Mayer MR, Stone MJ. Identification of Receptor Binding and Activation Determinants in the N-terminal and N-loop Regions of the CC Chemokine Eotaxin. J Biol Chem. 2001;276: 13911–13916. doi:10.1074/jbc.m011202200

458. Farzan M, Mirzabekov T, Kolchinsky P, Wyatt R, Cayabyab M, Gerard NP, et al. Tyrosine Sulfation of the Amino Terminus of CCR5 Facilitates HIV-1 Entry. Cell. 1999;96: 667–676. doi:10.1016/s0092-8674(00)80577-2

459. Murphy P, Tiffany H. Cloning of complementary DNA encoding a functional human interleukin-8 receptor. Science. 1991;253: 1280–1283. doi:10.1126/science.1891716

460. Kufareva I. Chemokines and their receptors: insights from molecular modeling and crystallography. Curr Opin Pharmacol. 2016;30: 27–37. doi:10.1016/j.coph.2016.07.006

461. Jensen A-SM, Sparre-Ulrich AH, Davis-Poynter N, Rosenkilde MM. Structural Diversity in Conserved Regions Like the DRY-Motif among Viral 7TM Receptors-A Consequence of Evolutionary Pressure? Adv Virol. 2012;2012: 231813. doi:10.1155/2012/231813

462. Butcher AJ, Kong KC, Prihandoko R, Tobin AB. Physiological Role of G-Protein Coupled Receptor Phosphorylation. Handbook of Experimental Pharmacology. 2011. pp. 79–94. doi:10.1007/978-3-642-23274-9_5

463. Shenoy SK, Lefkowitz RJ. Multifaceted roles of β-arrestins in the regulation of seven-membrane-spanning receptor trafficking and signalling. Biochem J. 2003;375: 503–515. doi:10.1042/bj20031076

464. Penela P, Ribas C, Mayor F. Mechanisms of regulation of the expression and function of G protein-coupled receptor kinases. Cell Signal.

2003;15: 973–981. doi:10.1016/s0898-6568(03)00099-8

465. Kleist AB, Getschman AE, Ziarek JJ, Nevins AM, Gauthier P-A, Chevigné A, et al. New paradigms in chemokine receptor signal transduction: Moving beyond the two-site model. Biochem Pharmacol. 2016;114: 53–68. doi:10.1016/j.bcp.2016.04.007

466. Muller WA. Mechanisms of leukocyte transendothelial migration. Annu Rev Pathol. 2011;6: 323–344. doi:10.1146/annurev-pathol-011110-130224

467. Rajagopalan L, Rajarathnam K. Structural Basis of Chemokine Receptor Function—A Model for Binding Affinity and Ligand Selectivity. Biosci Rep. 2006;26: 325–339. doi:10.1007/s10540-006-9025-9

468. Elling CE, Frimurer TM, Gerlach L-O, Jorgensen R, Holst B, Schwartz TW. Metal Ion Site Engineering Indicates a Global Toggle Switch Model for Seven-transmembrane Receptor Activation. J Biol Chem. 2006;281: 17337–17346. doi:10.1074/jbc.m512510200

469. Schwartz TW, Frimurer TM, Holst B, Rosenkilde MM, Elling CE. MOLECULAR MECHANISM OF 7TM RECEPTOR ACTIVATION—A GLOBAL TOGGLE SWITCH MODEL. Annu Rev Pharmacol Toxicol. 2006;46: 481–519. doi:10.1146/annurev.pharmtox.46.120604.141218

470. Olivella M, Caltabiano G, Cordomí A. The role of Cysteine 6.47 in class A GPCRs. BMC Struct Biol. 2013;13: 3. doi:10.1186/1472-6807-13-3

471. Williams G, Borkakoti N, Bottomley GA, Cowan I, Fallowfield AG, Jones PS, et al. Mutagenesis Studies of Interleukin-8. J Biol Chem. 1996;271: 9579–9586. doi:10.1074/jbc.271.16.9579

472. Murphy PM, Baggiolini M, Charo IF, Hébert CA, Horuk R, Matsushima K, et al. International union of pharmacology. XXII. Nomenclature for chemokine receptors. Pharmacol Rev. 2000;52: 145–176. Available: https://www.ncbi.nlm.nih.gov/pubmed/10699158

473. Bonecchi R, Graham GJ. Atypical Chemokine Receptors and Their Roles in the Resolution of the Inflammatory Response. Front Immunol. 2016;7: 224. doi:10.3389/fimmu.2016.00224

474. Fra AM, Locati M, Otero K, Sironi M, Signorelli P, Massardi ML, et al. Cutting edge: scavenging of inflammatory CC chemokines by the promiscuous putatively silent chemokine receptor D6. J Immunol. 2003;170: 2279–2282. Available: https://www.ncbi.nlm.nih.gov/pubmed/12594248

475. Rajagopal S, Kim J, Ahn S, Craig S, Lam CM, Gerard NP, et al. -arrestin- but not G protein-mediated signaling by the ‘decoy’ receptor CXCR7. Proceedings of the National Academy of Sciences. 2009;107: 628–632. doi:10.1073/pnas.0912852107

476. Galliera E, Jala VR, Trent JO, Bonecchi R, Signorelli P, Lefkowitz RJ, et al. β-Arrestin-dependent Constitutive Internalization of the Human Chemokine Decoy Receptor D6. J Biol Chem. 2004;279: 25590–25597. doi:10.1074/jbc.m400363200

477. Miller L, Mason S, Dvorak J, McGinniss M, Rothman I. Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. Science. 1975;189: 561–563. doi:10.1126/science.1145213

478. Berger EA, Murphy PM, Farber JM. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol. 1999;17: 657–700. doi:10.1146/annurev.immunol.17.1.657

479. Murphy PM. Viral exploitation and subversion of the immune system through chemokine mimicry. Nat Immunol. 2001;2: 116–122. doi:10.1038/84214

480. Rosenkilde MM, Schwartz TW. The chemokine system - a major regulator of angiogenesis in health and disease. APMIS. 2004;112: 481–495. doi:10.1111/j.1600-0463.2004.apm11207-0808.x

481. Rosenkilde MM, Kledal TN, Bräuner-Osborne H, Schwartz TW. Agonists and inverse agonists for the herpesvirus 8-encoded constitutively active seven-transmembrane oncogene product, ORF-74. J Biol Chem. 1999;274: 956–961. Available: https://www.ncbi.nlm.nih.gov/pubmed/9873037

482. Kledal TN, Rosenkilde MM, Schwartz TW. Selective recognition of the membrane-bound CX3C chemokine, fractalkine, by the human cytomegalovirus-encoded broad-spectrum receptor US28. FEBS Lett. 1998;441: 209–214. doi:10.1016/s0014-5793(98)01551-8

483. Rosenkilde MM. High constitutive activity of a virus-encoded 7TM receptor in the absenceof the conserved DRY-motif (Asp-Arg-Tyr) in transmembrane helix 3. Mol Pharmacol. 2005; doi:10.1124/mol.105.011239

484. Daiyasu H, Nemoto W, Toh H. Evolutionary Analysis of Functional Divergence among Chemokine Receptors, Decoy Receptors, and Viral Receptors. Front Microbiol. 2012;3: 264. doi:10.3389/fmicb.2012.00264

485. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12: 116. doi:10.1186/1471-2105-12-116

486. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 2008;36: W197–201. doi:10.1093/nar/gkn238

487. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274. doi:10.1093/molbev/msu300

488. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Lett. 1994;339: 269–275. Available: https://www.ncbi.nlm.nih.gov/pubmed/8112466

489. Rios S, Fernandez MF, Caltabiano G, Campillo M, Pardo L, Gonzalez A. GPCRtm: An amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors. BMC Bioinformatics. 2015;16: 206. doi:10.1186/s12859-015-0639-4

490. Ning K, Chua HN. Automated Identification of Protein Classification and Detection of Annotation Errors in Protein Databases Using Statistical Approaches. Lecture Notes in Computer Science. 2006. pp. 123–138. doi:10.1007/11683568_11

491. Kahanda I, Funk CS, Ullah F, Verspoor KM, Ben-Hur A. A close look at protein function prediction evaluation protocols. Gigascience. 2015;4: 41. doi:10.1186/s13742-015-0082-5

492. Nomiyama H, Osada N, Yoshie O. The evolution of mammalian chemokine genes. Cytokine Growth Factor Rev. 2010;21: 253–262. doi:10.1016/j.cytogfr.2010.03.004

493. Nomiyama H, Hieshima K, Osada N, Kato-Unoki Y, Otsuka-Ono K, Takegawa S, et al. Extensive expansion and diversification of the chemokine gene family in zebrafish: Identification of a novel chemokine subfamily CX. BMC Genomics. 2008;9: 222. doi:10.1186/1471-2164-9-222

494. Nomiyama H, Mera A, Ohneda O, Miura R, Suda T, Yoshie O. Organization of the chemokine genes in the human and mouse major clusters of CC and CXC chemokines: diversification between the two species. Genes Immun. 2001;2: 110–113. doi:10.1038/sj.gene.6363742

495. Zlotnik A, Yoshie O, Nomiyama H. The chemokine and chemokine receptor superfamilies and their molecular evolution. Genome Biol. 2006;7: 243. doi:10.1186/gb-2006-7-12-243

Abstract in English:

Dynamics of protein structures and its impact on local structural behaviors

Protein structures are highly dynamic in nature contrary to their depiction in crystal structures. A major component of structural dynamics is the inherent protein flexibility. The prime objective of this thesis is to understand the role of the inherent dynamics in protein structures and its propagation. Protein flexibility is analyzed at various levels of structural complexity, from primary to quaternary levels of organization. Each of the first five chapters' deal with a different level of local structural organization with first chapter dealing with classical secondary structures while the second one analysis the same using a structural alphabet - Protein Blocks. The third chapter focuses on the impact of special physiological events like post-translational modifications and disorder to order transitions on protein flexibility. These three chapters indicate towards a context dependent implementation of structural flexibility in their local environment. In subsequent chapters, more complex structures are taken under investigation. Chapter 4 deals with integrin αIIbβ3 that is involved in rare genetic disorders. Impact of the pathological mutations on the local flexibility is studied in two rigid domains of integrin αIIbβ3 ectodomain. Inherent flexibility in these domains is shown to modulate the impact of mutations towards the loops. Chapter 5 deals with the structural modelling and dynamics of a more complex protein structure of Duffy Antigen Chemokine Receptor embedded in an erythrocyte mimic membrane system. The model is supported by the most comprehensive phylogenetic analysis on chemokine receptors till date as explained in the last chapter of the thesis.

Keywords:

Protein structural flexibility, Allostery, Protein Blocks, Dual personality fragments, Post-translational modifications, Integrin $\alpha_{IIb}\beta_3$, Glanzmann Thrombasthenia, Fetal/neonatal Alloimmune Thrombocytopenia, *Plasmodium vivax* malaria, Duffy Antigen Chemokine Receptors, Molecular phylogenetics

Résumé en français :

Dynamique des structures protéiques et son impact sur les comportements structuraux locaux

Les structures protéiques sont de nature hautement dynamique contrairement à leur représentation dans les structures cristallines. Une composante majeure de la dynamique structurelle est la flexibilité des protéines inhérentes. L'objectif principal de cette thèse est de comprendre le rôle de la dynamique inhérente dans les structures protéiques et leur propagation. La flexibilité des protéines est analysée à différents niveaux de complexité structurelle, du niveau d'organisation primaire au niveau quaternaire. Chacun des cinq premiers chapitres traite un niveau différent d'organisation structurelle locale avec le premier chapitre traitant des structures secondaires classiques tandis que le second analyse la même chose en utilisant un alphabet structurel - les blocs protéiques. Le troisième chapitre se concentre sur l'impact d'événements physiologiques spéciaux comme les modifications post-traductionnelles et le désordre sur les transitions d'ordre sur la flexibilité des protéines. Ces trois chapitres indiquent une mise en œuvre dépendante du contexte de la flexibilité structurelle dans leur environnement local. Dans les chapitres suivants, des structures plus complexes sont prises en compte. Le chapitre 4 traite de l'intégrine $\alpha_{IIb}\beta_3$ impliquée dans des troubles génétiques rares. L'impact des mutations pathologiques sur la flexibilité locale est étudié dans deux domaines rigides de l'intégrine $\alpha_{IIb}\beta_3$ ectodomaine. La flexibilité inhérente dans ces domaines est montrée pour moduler l'impact des mutations vers les boucles. Le chapitre 5 traite de la modélisation structurelle et de la dynamique d'une structure protéique plus complexe du récepteur des chimiokines des antigènes du groupe Duffy incorporé dans un système de membrane mimétique érythrocytaire. Le modèle est soutenu par l'analyse phylogénétique la plus complète sur les récepteurs de chimiokines jusqu'à ce jour, comme expliqué dans le dernier chapitre de la thèse.

Mots clés :

Flexibilité de la structure des protéines, allostérie, Blocs Protéiques, fragments de double personnalité, modification post-translationnelle, Intégrine $\alpha_{IIb}\beta_3$, Thrombasthénie de Glanzmann, thrombocytopénie allo-immune fœtale / néonatale, paludisme à *Plasmodium vivax*, récepteurs des chimiokines des antigènes du groupe Duffy, phylogénie moléculaire.