Bhanupratap Singh Chouhan

Integrin evolution: from prokaryotes to the diversification within chordates
Bhanupratap Singh Chouhan was born on September 04, 1985 in Indore, M.P., India. He did his Bachelor of Science in Bioinformatics from Government Autonomous Holkar Science College, Devi Ahilya University in Indore. He then graduated from the University of Turku in 2009 with a Master of Science in Bioinformatics. This Ph.D. thesis work has taken place under the supervision of Professor Mark S. Johnson and Docent Konstantin Denissiouk at the Structural Bioinformatics Laboratory (SBL), Åbo Akademi University, Turku, Finland during 2010-2015.
Integrin evolution: From prokaryotes to the diversification within chordates

Bhanupratap Singh Chouhan

Biochemistry
Faculty of Science and Engineering
Åbo Akademi University
Åbo, Finland
2016
Supervised by:
Professor Mark S. Johnson
Docent Konstantin Denessiouk
Dr. Alexander Denesyuk
Åbo Akademi University
Åbo, Finland

Reviewed by:
Professor Donald Gullberg,
University of Bergen, Norway.

Docent Janne Ravantti,
University of Helsinki, Finland.

Opponent:
Docent Olli Pentikäinen,
University of Jyväskylä, Finland.

ISBN 978-952-12-3448-4
Painosalama Oy – Turku, Finland 2016
# Table of contents

1. Review of the literature 1
   1.1 Introduction 1
   1.2 Integrin bidirectional signaling 6
   1.3 Relationship among integrin α and β subunits 12
   1.4 N-terminal region: The β-propeller domain 15
   1.5 N-terminal region: The vWA domain 17
   1.6 Collagens and vertebrate collagen receptors 19
   1.7 Structure of collagen-binding and ICAM-binding integrins 21

2. Aims of the study 29

3. Methods 30
   3.1 Online databases 30
   3.2 Protein sequence analyses 31
   3.3 3D modelling and structural analyses 35

4. Results 37
   4.1 Conservation of the human integrin-type β-propeller domain in bacteria 37
      4.1.1 X-ray structures of αVβ3 and αIIbβ3 and unique structural characteristics 37
      4.1.2 β-turns and torsion angles 39
      4.1.3 Ca\(^2+\) binding motif (DxDxDG) 42
      4.1.4 Bacterial sequence dataset 43
   4.2 Evolutionary origin of the αC helix in integrins 45
   4.3 Early chordate origin of the human-type integrin I-domains 47
      4.3.1 Phylogenetic analyses 49
      4.3.2 Functional residues shared between human and lamprey α I-domains 51
      4.3.3 Sea lamprey α I-domains recognize different mammalian collagen types 55

5. Discussion 57
   5.1 Publication I 57
      5.1.1 Sequence alignment and phylogenetic tree 57
      5.1.2 3D comparative modelling 58
      5.1.3 Secondary structure prediction 58
      5.1.4 Possible functions 59
   5.2 Publication II 59
      5.2.1 Genome searches 60
      5.2.2 Sequence alignment and secondary structure prediction 60
      5.2.3 Importance of the αC helix 61
   5.3 Publication III 61
      5.3.1 Genome searches 61
      5.3.2 Phylogenetic analyses and multivariate plot 62
      5.3.3 Structural studies 63
      5.3.4 3D comparative modelling 64

6. Conclusions and future directions 65

7. References 67

8. Original publications 85
**Original publications**


Contribution of the author

The author has performed multiple sequence alignments, phylogenetic analyses, secondary structure predictions, 3D comparative modelling and structural studies, local genome searches as well as various online database searches during the course of studies that constitute this thesis work. The author also participated in the design, implementation, execution and writing of the publications listed as I-III (original publications) and IV-V (Additional publications).

Additional publications


Acknowledgements

This thesis work was carried out at the Structural Bioinformatics Laboratory (SBL), Faculty of Science and Engineering, Åbo Akademi University, Finland during 2010-2015. I would like to acknowledge and express my gratitude to all the people who contributed towards the completion of this thesis.

First of all I would like to extend my deepest gratitude to my supervisors, Professor Mark S. Johnson and Docent Konstantin Denessiouk for granting me the opportunity to conduct my doctoral studies at SBL. Thank you Mark and Konstantin for your guidance, patience, support and constant encouragement which enabled me to complete this doctoral thesis. I would also like to thank Dr. Alexander Denesyuk for his valuable insight, keen supervision and positive feedback. I am very thankful to Professor Donald Gullberg and Docent Janne Ravantti for taking time out of their busy schedule to read my thesis and providing constructive criticism in order to improve the quality of the text. I am also very thankful to Professor Jyrki Heino, Dr. Jarmo Käpylä and Dr. Lari Lehtiö for agreeing to be members of my thesis committee and providing valuable ideas regarding this thesis work. I would also like to acknowledge and thank my collaborators/co-authors from Professor Jyrki Heino’s laboratory at the University of Turku. I would also like to extend my gratitude to Docent Olli Pentikäinen for agreeing to be my opponent.

I would also like to thank my seniors and colleagues from SBL: Lenita Viitanen, Mikko Huhtala, Mikko Vainio, Santeri Puranen, Tomi Airenne, Outi Salo-Ahen (Thank you so much Outi for the abstract translations), Anni Kauko, Eva Bligt-Lindén, Käthe Dahlström, Jukka Lehtonen and Professor Tiina Salminen. Thank you all for the wonderful discussions and coffee room memories. A very special thanks to Frederik Karlsson for being very supportive and helpful with all the practical matters throughout my doctoral studies at SBL. I am also very grateful to the secretarial and technical staff at Department of Biochemistry for all their help. I am very thankful to the National Doctoral Programme in Informational and Structural Biology (ISB) for accepting me as a member of this amazing graduate school. Thanks also go out to my colleagues and other members of ISB who made the annual meetings fun and inspirational.
Apart from my wonderful colleagues, I would also like to thank my close friends and partners in crime at SBL (and outside SBL as well): Nitin ‘Marwaadi’ Agrawal, Vipin Ranga-’Dada’, Avlokita ‘Tikka’ Tiwari, Rakesh ‘Raakshas’ Purushottama, Athresh ‘Ornob’ Shigaval and Parthiban ‘Saaree’ Marimuthu. I will always cherish our wonderful time together especially the long weekend lunch sessions and the blunt/crude humour. This Ph.D. journey would not have been as fun without you guys. Thank you also to my band member Mohit ‘MC Elite’ Narwal for the wonderful friendship and entertainment over the years, I will always remember our long ‘rap sessions’ during the weekend which provided me with a break from the lab work (“Lyrical Prophets”). I owe my warmest gratitude to my dear friends outside my workplace for all the wonderful memories: Hasan, Aman, Raghu, Yashu, Vinay, Neeraj ji, Tamogyna, Debanga, Rishabh, Ayush, Abhishek, Roshan, Sachin and the rest of my south-asian friends (Sorry if I missed someone, but you know who you are).

I would like to thank my family for their endless support and constant encouragement which has been a source of great strength. I would especially like to thank my parents for believing in me and letting me chase after my dream. It’s been a long journey which could not have been possible without your support. Thank you Mummy and Papa, I sincerely appreciate all the sacrifices you have made, it means a lot to me. I would also like to thank my loving wife Neha, your constant support over the past several months has kept me very focused and inspired me to work harder. A special thanks to all my wonderful cousins and family members especially grandpa (Nano-sa) for always being concerned with my well-being and happiness.

I would like to dedicate this thesis to my loved ones who are no longer with me but I always remember them very fondly: my grandma (nani-sa), my grandparents (dado-sa and dadi-sa) and my best friend Prabhakar Sharma (Prabhu) who wanted to pursue his own Ph.D. but his untimely demise is something I find difficult to come to terms with, even today so I dedicate this thesis especially to all of you – “you may be gone but you are never over”.

The financial support for this thesis work has been absolutely indispensable. Therefore, I wish to acknowledge and thank Åbo Akademi University, ISB, Sigrid Juselius Foundation, Tor, Joe and Pentti Borg Foundation, Magnus Ehrnrooth Foundation and Stiftelsen för Åbo Akademi for the vital financial support.

- Bhanupratap Singh Chouhan,
Åbo, September 2016
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>β-propeller</td>
<td>Beta Propeller Domain</td>
</tr>
<tr>
<td>ADMIDAS</td>
<td>Adjacent to MIDAS site</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CATH</td>
<td>Class Architecture Topology Homologous superfamily database</td>
</tr>
<tr>
<td>CDD</td>
<td>Conserved Domain Database</td>
</tr>
<tr>
<td>COGs</td>
<td>Clusters of Orthologous Groups</td>
</tr>
<tr>
<td>DDR</td>
<td>Discoidin Domain Receptor</td>
</tr>
<tr>
<td>ECM</td>
<td>Extra Cellular Matrix</td>
</tr>
<tr>
<td>EDTA</td>
<td>EthyleneDiamineTetraacetic Acid</td>
</tr>
<tr>
<td>‘FG-GAP’</td>
<td>Phe-Ala-Gly-Ala-Pro</td>
</tr>
<tr>
<td>‘GFOGER’</td>
<td>Gly-Phe-Hyp-Gly-Glu-Arg (Hyp = Hydroxyproline)</td>
</tr>
<tr>
<td>‘GLOGEN’</td>
<td>Gly-Leu-Hyp-Gly-Glu-Asn</td>
</tr>
<tr>
<td>GOR</td>
<td>Garnier-Osguthorpe-Robson method</td>
</tr>
<tr>
<td>GPVI</td>
<td>Glycoprotein VI</td>
</tr>
<tr>
<td>I-domain</td>
<td>Inserted domain</td>
</tr>
<tr>
<td>I-EGF</td>
<td>Integrin Epidermal growth factor-like domain</td>
</tr>
<tr>
<td>JTT</td>
<td>Jones Taylor Thornton method</td>
</tr>
<tr>
<td>LAIR</td>
<td>Leukocyte-Associated IG-like Receptor-1</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MIDAS</td>
<td>Metal Ion Dependent Adhesion Site</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Components Analysis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PFAM</td>
<td>Protein Families Database</td>
</tr>
<tr>
<td>PIR</td>
<td>Protein Information Resources</td>
</tr>
<tr>
<td>PRF</td>
<td>Protein Research Foundation</td>
</tr>
<tr>
<td>PSI domain</td>
<td>Plexin-Semaphorin-Integrin domain</td>
</tr>
<tr>
<td>PSSMs</td>
<td>Position Specific Scoring Matrices</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Deviation</td>
</tr>
<tr>
<td>RPS-BLAST</td>
<td>Reversed Position Specific BLAST</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification Of Protein database</td>
</tr>
<tr>
<td>SMART</td>
<td>Simple Modular Architecture Research Tool</td>
</tr>
<tr>
<td>SyMBS</td>
<td>Synergistic Metal Ion Binding Site</td>
</tr>
<tr>
<td>ICAM</td>
<td>Inter Cellular Adhesion Molecule</td>
</tr>
<tr>
<td>TM</td>
<td>TransMembrane</td>
</tr>
<tr>
<td>UniprotKB</td>
<td>Uniprot KnowledgeBase</td>
</tr>
<tr>
<td>VCAM</td>
<td>Vascular Cell Adhesion Molecule</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular Endothelial Growth Factor</td>
</tr>
<tr>
<td>vWA</td>
<td>von Willebrand Factor A domain</td>
</tr>
<tr>
<td>WAG</td>
<td>Whelan And Goldman method</td>
</tr>
</tbody>
</table>
Abstract

Integrins are a family of large multi-domain cell surface receptors responsible for bidirectional signaling in response to cell-cell, cell-extracellular matrix interactions as well as intracellular interactions. Intrgrins are implicated in a wide variety of functions such as inflammatory responses, adoptive antigen-specific immunity, tissue remodelling and cell adhesion, proliferation and differentiation. Furthermore, integrins are also known to be associated with a wide variety of diseases and health issues, such as tumour metastasis, immune dysfunction, inflammation, viral infections and osteoporosis – to name a few – making them one of the most complex cell adhesion molecules. Integrin are heterodimers which are composed of an α and a β subunit and humans are known to express 18 α subunits and 8 β subunits which associate non-covalently to form 24 αβ heterodimers out of 144 possible combinations. Orthologues of mammalian integrins are observed throughout vertebrates including the bony fish (osteichthyes), however integrins extracted from early chordates like the tunicates Ciona intestinalis or Halocynthia roretzi are not direct mammalian orthologues. Even though integrins are observed throughout metazoans, studies have reported that integrins and their signaling machinery are located in unicellular eukaryotes. In addition, the 3D folds of the constituent domains from the integrin α and β subunits have also been detected in bacteria.

The major aims of the research work described in this thesis were to answer the following questions: i) When did the constituent domains from the integrin α and β subunits originate? ii) When did the α II-domains get integrated into the integrin heterodimer and when did the collagen-binding integrin α II-domains originate in the vertebrates? iii) When did the mammalian-type integrin orthologues originate in vertebrates?

In order to address these questions, we analysed the available sequences, genomic data as well as structural data, all of which are discussed in detail in the three studies comprising this thesis. During the course of this thesis: i) We have addressed the origin of pivotal integrin constituent domain like the N-terminal 7-bladed β-propeller domain from the α-subunit by investigating the extent of similarities between the sequences and structures of different integrin domains and similar gene products and protein sequences from bacteria. ii) We have identified characteristic structural features or motifs like the αC helix in order to understand the evolutionary process of collagen-binding integrin α II-domain in vertebrates. iii) Recent advancements in the genome assembly process of organisms like the sea lamprey (agnathostome) and the elephant shark (chondrichthyes) has helped us in understanding the origin and evolution of mammalian-type integrin orthologues. In conclusion, the studies presented in this thesis present novel insights into the evolutionary patterns of the integrins

Key Words: Integrin; molecular evolution; β-propeller; α II-domain; collagen-receptor; αC helix;
**Sammanfattning**


De viktigaste målsättningarna för forskningsarbetet presenterat i denna avhandling var att besvara följande frågor: i) När uppkom huvuddomänerna i integrinernas α- och β-subenheter? ii) När integrerades α I-domänerna i integrinheterodimeren och när uppkom de kollagenbindande integrin α I-domänerna i ryggradsdjur? iii) När uppkom integrinortologerna av däggdjurstyp hos ryggradsdjur?

För att svara på dessa frågor analyserade vi de tillgängliga sekversona, genetisk data samt strukturell data, vilket allt diskuteras i detalj i de tre studier som denna avhandling omfattar. Inom ramen för denna avhandling: i) Har vi studerat ursprunget av en central huvuddomän av integriner, såsom den N-terminala 7-bladiga β-propellerdomänen från α-subenheten, genom att undersöka omfattningen av likheterna mellan sekversona och strukturerna av olika integrindomän och liknande genprodukter och proteinsekvenser hos bakterier. ii) Har vi identifierat de karakteristiska strukturella dragen eller motiven, såsom αC-helix, för att förstå den evolutionära processen av kollagenbindande integrin α I-domänen hos ryggradsdjur. iii) Har nya framsteg i monteringsprocessen av genomet hos organismer som havsnejonöga (Agnathostome) och australisk plognos (Chondrichthyes) hjälpt oss att förstå ursprunget och evolutionen av däggdjurtyp integrinortologer. Sammanfattningsvis kan konstateras att de studier som presenteras i denna avhandling presenterar nya insikter i de evolutionära mönstren av integriner.

**Nyckelord:** integrin; molekylär evolution; β-propeller; α I-domän; kollagenreceptor; αC-helix
1. Review of the literature

1.1 Introduction

Integrins are a large family of bidirectionally signaling, heterodimeric, trans-membrane cell surface receptors that are involved in cell-cell, cell-ECM and even cell-pathogen interactions (Eble and Kühn 1997; Hynes 2002; Legate et al., 2009; Takada et al., 2007). Extracellular integrin domains interact with a wide variety of ligands like collagens, fibronectin and laminins and receptor immunoglobulin fold domains of intracellular adhesion molecules (ICAMs) (Table 1 – Johnson et al., 2009); while the intracellular cytoplasmic domains communicate with the signaling molecules located inside the cell (Luo et al., 2007) and play a pivotal role in the formation of focal adhesions (Legate et al., 2009). These interactions are central to the regulation of cell migration, phagocytosis, cell growth and development (Arnaout et al., 2007). In addition, integrins are also implicated in diseases and health issues like tumor progression (Shin et al., 2012), recognition of pathogens (Ulanova et al., 2009), immune dysfunction (Kishimoto et al., 1987), inflammation (Gahmberg et al., 1998) and osteoporosis (Teitelbaum, 2005).

Integrins are composed of two subunits, α and β subunits (Figure 1). Mammalians express at least 18 different integrin α subunits and 8 different β subunits, which are known to form 24 α/β heterodimeric combinations in humans out of 144 possible combinations (Figure 2) (Hynes, 2002; Shimaoka et al., 2003). The reason behind this observed pairing pattern is still not known but it is quite clear from Figure 2 that certain integrin subunits are more promiscuous than others. For example, the subunits β1, β2 and αV form more heterodimeric associations than rest of the subunits. Interestingly, integrin β1 subunit is the most promiscuous subunit as it can form heterodimeric association with 12 integrin α subunits. Table 1 (adapted from the review of Johnson et al., 2009) provides an insight into these receptors and their contributions in the development and sustenance of mammals.

The extracellular region consists of ligand binding N-terminal domains followed by the leg or stalk region domains, transmembrane domains and the cytoplasmic domains for both the subunits respectively (Nermut et al., 1988). The first X-ray crystal structure highlighting the ectodomain region of the integrin heterodimer was solved in 2001 by Xiong et al and it clearly showcased the multi-domain structure of the α and β subunit.
<table>
<thead>
<tr>
<th>Integrin</th>
<th>Ligand</th>
<th>Location of defects in knock-outs and/or main expression sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>α7β1</td>
<td>Laminins</td>
<td>Muscle (Mayer et al., 1997)</td>
</tr>
<tr>
<td>α6β4</td>
<td>Laminins</td>
<td>Skin (hemidesmosomes) (Stepp et al., 1990)</td>
</tr>
<tr>
<td>α6β1</td>
<td>Laminins, ADAMs</td>
<td>Gametes, macrophages, platelets (Georges-Labouesse et al., 1996)</td>
</tr>
<tr>
<td>α3β1</td>
<td>Laminins, (Collagen)</td>
<td>Skin, kidney, lung, cortex (Kreidberg et al., 1996; DiPersio et al., 1997)</td>
</tr>
<tr>
<td>αIIβ3</td>
<td>Fibrinogen (‘RGD’, ‘GAKQAGDV’), Fibronectin, vitronectin (‘RGD’)</td>
<td>Platelets (Pytela et al., 1986)</td>
</tr>
<tr>
<td>αVβ8</td>
<td>Vitronectin (‘RGD’)</td>
<td>Vascular development (Müller et al., 1997; Littlewood and Müller, 2000)</td>
</tr>
<tr>
<td>αVβ6</td>
<td>Fibronectin, TGF-β-LAP (‘RGD’)</td>
<td>Skin, lung (collagen accumulation) (Munger et al., 1999)</td>
</tr>
<tr>
<td>αVβ5</td>
<td>Vitronectin (‘RGD’)</td>
<td>Eye (retinal phagocytosis), bone (osteoclastogenesis) (Nandrot et al., 2004; Lane et al., 2005)</td>
</tr>
<tr>
<td>αVβ3</td>
<td>Fibrinogen, fibronectin, vitronectin, tenasin C, osteopontin, bone sialoprotein (‘RGD’), MMP-2</td>
<td>Bone (osteoclasts) (Hodivala-Dilke et al., 1999; McHugh et al., 2000)</td>
</tr>
<tr>
<td>αVβ1</td>
<td>Fibronectin, vitronectin (‘RGD’)</td>
<td>In vivo role in tissue fibrosis (Bodary et al., 1990; Vogel et al., 1990; Reed et al., 2015)</td>
</tr>
<tr>
<td>α8β1</td>
<td>Fibronectin, vitronectin, tenasin C, osteopontin, nefronectin (‘RGD’)</td>
<td>Kidney, inner ear (Zhu et al., 2001)</td>
</tr>
<tr>
<td>α5β1</td>
<td>Fibronectin (‘RGD’)</td>
<td>Embryonic development (blood vessels) (Pytela et al., 1985; Yang et al., 1993)</td>
</tr>
<tr>
<td>α9β1</td>
<td>Tenasin C, osteopontin, ADAMs, factor XIII, VCAM, VEGF-C, VEGF-D</td>
<td>Lymphangiogenesis (Huang et al., 2000; Vlahakis et al., 2005)</td>
</tr>
<tr>
<td>α4β7</td>
<td>Fibronectin, VCAM, MadCaM</td>
<td>Peyer's patch (Yang et al., 1995)</td>
</tr>
<tr>
<td>α4β1</td>
<td>Fibronectin, VCAM</td>
<td>Embryonic development (Yang et al., 1995)</td>
</tr>
<tr>
<td>α11β1</td>
<td>Collagens</td>
<td>Periodontal ligament (Popova et al., 2007)</td>
</tr>
<tr>
<td>α10β1</td>
<td>Collagens</td>
<td>Cartilage (Bengtsson et al., 2005)</td>
</tr>
<tr>
<td>α2β1</td>
<td>Collagens, tenasin C, (laminins)</td>
<td>Platelets, epithelium, mast-cells (mesenchymal tissues) (Zutter et al., 1990; Chen et al., 2002; Holtkötter et al., 2002; Senger et al., 2002; Edelson et al., 2004; Auger et al., 2005; Zhang et al., 2008)</td>
</tr>
<tr>
<td>α1β1</td>
<td>Collagens, semaphorin 7A, (laminins)</td>
<td>Mesenchymal tissues (Pozzi et al., 1998; Gardner et al., 1999; Ekholm et al., 2002; Conrad et al., 2007)</td>
</tr>
<tr>
<td>αDβ2</td>
<td>ICAM, VCAM</td>
<td>Eosinophils (Grayson et al., 1999; Vieren et al., 1999)</td>
</tr>
<tr>
<td>αMβ2</td>
<td>ICAM, VCAM, iC3b, factor X, fibrinogen</td>
<td>Leukocytes (phagocytosis) (Altieri et al., 1990; Diamond et al., 1990; Elemer and Edgington 1994; Barthel et al., 2006; )</td>
</tr>
<tr>
<td>αLβ2</td>
<td>ICAM</td>
<td>Leukocytes (recruitment) (Kolanus et al., 1996)</td>
</tr>
<tr>
<td>αXβ2</td>
<td>Fibrinogen, plasminogen, heparin, iC3b</td>
<td>Leukocytes (Micklem and Sim, 1985; Loike et al., 1991; Davis, 1992; Diamond et al., 1993)</td>
</tr>
<tr>
<td>αEβ7</td>
<td>E-cadherin</td>
<td>Skin, Gut (immune system) (Higgins et al., 1998)</td>
</tr>
</tbody>
</table>
Figure 1 illustrates the schematic architecture of the non-covalently associated integrin heterodimer. Integrin α subunits (1000-1190 residues) are generally larger than the β subunits (770-800 residues) but there is an exception i.e. integrin β4, as it extends to nearly 1800 residues with the additional ~1000 residues located within the intracellular C-terminal region. The integrin β subunit ectodomain region consists of eight domains, which includes the βI-like domain, hybrid domain, PSI domain (Plexin, Semaphorin, Integrin), I-EGF modules 1-4 and the β-tail domain. These domains are followed by the transmembrane domain (TM) and the cytoplasmic tail region. The βI-like domain (together with the β-propeller of the α subunit) binds extracellular integrin ligands in the absence of the α I-domain) adopts a Rossmann fold and it is inserted in the hybrid domain. The PSI domain is split into two regions (Xiao et al., 2004; Xiong et al., 2004) that are known to be connected by a disulphide bond (Cys13 to Cys435 in integrin β3 subunit). The integrin-type epidermal growth factor modules i.e. I-EGF 1-4 domains are cysteine rich regions and each I-EGF domain contains eight cysteines that are bonded together in C1-C5, C2-C4, C3-C6, C7-C8 pattern with the exception of I-EGF1 that lacks the C2-C4 disulphide bond (Beglova et al., 2002; Takagi et al., 2002; Zhu et al., 2008). The β-tail domain consists of an α+β fold and it is proposed to play an important role in integrin activation (Arnaout et al., 2005). The interaction between the integrin α subunit and β subunit TM helices results in the resting state of the heterodimer (Adair and Yeager, 2002; Luo et al., 2004; Partridge et al., 2005; Wegener and Campbell, 2008). The cytoplasmic domains (of both α and β subunits) can form interactions with multiple proteins and these domains play a significant role in keeping the dimer in a resting state as well as inside-out activation of integrins (Wegener and Campbell, 2008; Legate et al., 2009).

**Figure 1.** Schematic architecture of a typical integrin α/β heterodimer. In this figure ‘TM’ represents the transmembrane domain, ‘PSI’ represents plexin-semaphorin-integrin domain, ‘HD’ represents the hybrid domain and I-EGF represents Immunoglobulin like folds.
The integrin α subunit consists of four or five extracellular domains and these include the seven bladed β-propeller domain (discussed in detail later in this thesis), which may host an inserted I-domain between its second and third repeat, followed by the thigh, calf-1 and calf-2 domains (that constitute the stalk region), a transmembrane domain and a cytoplasmic tail region. The thigh and calf domains adopt a similar structure as they both share a β-sandwich fold (Xiong et al., 2001).

The human integrin α subunits can be subdivided into two major groups: those that contain an additional inserted α I-domain (α1, α2, α10, α11, αD, αX, αL, αM and αE) and those without α I-domain, Figure 2 (α4, α9, α6, α7, α3, αV, α5, α8 and αIIb) (Larson et al., 1989). The α I-domain is about 200 residues long and it is a homolog of the β I-like domain present in all β subunits that adopts a Rossmann fold, which basically consists of six parallel β-strands surrounded by two pairs of α-helices and the α I-domain buds out from a loop located between the second and third repeat of the β-propeller domain, which in turn is located in the head region of the integrin α subunit. I-domains have a highly-exposed binding site and can thus recognize bulkier integrin ligands like collagens and ICAMs (Intercellular Adhesion Molecule) through binding of the acidic residue glutamate to its divalent cation (Mg^{2+}) at the binding site called ‘MIDAS’ (Metal Ion Dependent Adhesion Site) (Lee et al., 1995). MIDAS is a characteristic of all the integrin α I-domains and it (MIDAS) is observed all the way back even in the urochordates (Miyazawa et al., 2001; Ewan et al., 2005; Huhtala et al., 2005). MIDAS is also present in all βI-like domains and the site for binding an aspartate of an exposed loop from the protein ligand.

Functionally speaking, integrins with the α I-domains segregate integrins two classes on the basis of their ligand specificity; in the case of the human integrins α1β1, α2β1, α10β1 and α11β1, they are known to bind collagen and contribute to the structural integrity of cells and tissues (Knight et al., 1998, 2000; Zhang et al., 2003). The human integrins αXβ2, αDβ2, αMβ2, αLβ2 and αEβ7 are known to play a key role in the interaction of leukocytes with endothelial cells and other matrix structures (Van der Vieren et al., 1996, 1999; Grayson et al., 1999; Garnotel et al., 2000; Noti et al., 2000; Hynes, 2002; Solovjov et al., 2005; Takada et al., 2007). Considerable insight into the structural basis for the function of individual integrin α I-domains has been provided in the variety of structural studies from the past 20 years (Integrin alpha 1 PDB IDs: 1QC5: Rich et al., 1999; 1QCY: Kankare et al., 2003 (Deposition author), Nymalm et al., 2003; 1PT6: Nymalm et al., 2004, Integrin alpha 2 PDB IDs: 1AOX: Emsley et al., 1997, 1DZI: Emsley et al., 2000, 1V7P: Horii et al., 2004; Integrin alpha L PDB IDs: 1LFA: Qu and Leahy, 1996, 1DGQ: Legge et al., 2000, 1CQP:
Kallen et al., 2000, 1MQ8: Shimaoka et al., 2003, 1T0P: Song et al., 2005 3BN3: Zhang et al., 2008; Integrin alpha X PDB ID: Vorup-Jensen et al., 2003; Integrin alpha M PDB IDs: Lee et al., 1995; Xiong et al., 2002 McCleverty and Liddington, 2003). In comparison to integrins without α I-domains where the ligand must have an exposed flexible loop, the insertion of the α I-domain in some integrin α subunits has led to a dramatic shift in the ligand recognition site thereby providing unprecedented access to heavier ligands like collagen fibres bundled into large macroscopic structures and the immunoglobulin-fold ICAM domains.

Figure 2. Heterodimeric association pattern of integrin α and β subunits; α subunits marked with an ‘*’ contain the additional I-domain.

All of the earliest diverging integrins do not contain the α I-domain. Consequently, integrins without the α I-domain recognize ligands like laminin and fibronectin in a different way, at a narrow cleft at the junction of β-propeller domain (of the integrin α subunit) and the β I-like domain (of the integrin β subunit) (Xiong et al., 2004). The β I-like domain (also known as the β A-domain) and its MIDAS is located in the head portion of the integrin β subunit and, in the absence of the α I-domain, it plays a central role in ligand recognition, e.g. of the arginine-glycine-aspartate sequence on the loops of ligands. Extracellular ligand recognition via both of these mechanisms, with or without the α I-domain, results in signal transduction, but integrins are known to be activated via internal cytoplasmic interactions as well.
1.2 Integrin bidirectional signaling

The early structural view of integrins was based primarily on electron microscopy data (Nermut et al., 1988) as well as the analysis of integrin sequences (Nermut et al., 1988; Arnaout, 1990) but implementation of techniques like X-ray crystallography and NMR spectroscopy has led to the solution of multiple three-dimensional integrin structures ranging from the first structures of \( \alpha I \)-domains, and later headpieces and ectodomains, the transmembrane (TM) helices and cytoplasmic tail regions, thereby providing a more comprehensive understanding of the structure and function of the integrin heterodimer.

At present, the most widely accepted integrin activation mechanism consists of three distinct confirmations as shown in Figure 3 (Arnaout et al., 2005 and Luo et al., 2007). A resting state (or bent state) where the integrin heterodimer exists in an inverted bent V-shape with respect to the location of the membrane, a low-affinity intermediate state (or extended closed state) and an activation state (or extended open state) which is associated with an extended conformation wherein the hybrid domain has swung outward in contrast to the initial resting position and the cytoplasmic tail regions from the \( \alpha \) and \( \beta \) subunits do not interact anymore. Several studies have already focused on integrin structural characterisations and activation mechanisms (Vinogradova et al., 2000, 2002, 2004; Li et al., 2001, 2005; Lu et al., 2001; Ulmer et al., 2001; Xiong et al., 2001, 2002, 2004; Beglova et al., 2002; Takagi et al., 2002, 2003; Weljie et al., 2002; Kim et al., 2003; Luo et al., 2003, 2004; Litvinov et al., 2004; Tng et al., 2004; Xiao et al., 2004; Iwasaki et al., 2005; Mould et al., 2005; Shi et al., 2005, 2007; Nishida et al., 2006) but the bent structure is generally seen in structures of the ectodomains. The current integrin activation model clearly suggests that conversion from a low affinity resting state to a high affinity extended state is pivotal for ligand binding and signalling, certain studies have indicated that integrins can bind ligands in a bent state as well (Calzada et al., 2002; Xiong et al., 2002; Chigaev et al., 2003, 2007; Adair et al., 2005; Askari et al., 2010).

Integrins can be activated from the outside as well as from the inside of the cell, which results in a large and reversible conformational change in the heterodimer. When the activation takes place due to interaction between the integrin ectodomain and extracellular ligands (like collagen, ICAM, fibronectin etc.) it is known as ‘outside-in’ signaling, whereas integrin activation taking place due to signals from within the cell are initiated by interactions between cytosolic proteins like talin and kindlin and the integrin \( \beta \) subunit.
cytoplasmic ‘tail’. This is known as ‘inside-out’ signaling. Herein is presented a very brief look at integrin bidirectional signaling.

a) Inside-out signaling: Talin and kindlins bind to the cytoplasmic tail of the integrin β-subunit, which in turn contributes to integrin activation (Tadokoro et al., 2003; Legate and Fässler, 2009; Moser et al., 2009). The tail region of the integrin β-subunit consists of two ‘NPxY’ motifs (where N is aspartagine; P is proline; x is any residue and Y is tyrosine) and it has been shown that talin interacts with the membrane proximal ‘NPxY’ motif, while kindlins are known to interact with the membrane distal ‘NPxY’ motif (Calderwood et al., 2002; Calderwood 2004; Banno and Ginsberg 2008; Moser et al., 2009; Calderwood et al., 2013). Talin participates in this interaction through its F3 PTB domain (phospho-tyrosine binding) and this interaction facilitates competitive binding between a lysine from Talin and an arginine from the integrin α-subunit, resulting in the disassociation of a salt bridge linking the α and β subunits, which activates the integrin (Anthis et al., 2009; Lau et al., 2009; Zhu et al., 2009). Integrin activation results from the separation of the cytoplasmic domains that leads to the extension of the extracellular domains causing a switchblade-like extension at the ‘genu’ domain region (Beglova et al., 2002; Takagi et al., 2002)

Figure 3. The integrin heterodimer can exist in three conformational states i.e. A) bent conformation where the two subunits are inactive and they are held together by a tight salt bridge; B) extended intermediate closed conformation and C) extended open conformation facilitating bidirectional signaling. The cover illustration for this thesis work includes the pivotal α I-domain.
Kindlins interact with the membrane distal ‘NPxY’ motif of integrin β-tail region through the FERM (Four-point-one, Ezrin, Radixin, Moesin) domain. Even though kindlins are not directly responsible for integrin activation they do play a pivotal role as co-activators (in partnership with talin). It has been suggested that inhibiting the binding of kindlin and integrins can be disruptive towards the talin-mediated activation of integrins, thereby exhibiting the importance of integrin-kindlin association for proper integrin activation (Montanez et al., 2008; Harburger and Calderwood, 2009; Federico et al., 2013). Furthermore, it has been shown that excessive expression of kindlins (in cell culture systems) does not lead to integrin activation but when kindlins are co-expressed with the talin head region then kindlins can help trigger the activation of integrin αIIbβ3 (Montanez et al., 2008; Moser et al., 2008). Structural data from an NMR study has further highlighted the fact that kindlins function as integrin co-activators (Bledzka et al., 2012). When it comes to inside-out signaling then the talin mediated integrin activation mechanism has been well studied at the molecular level but the details of the co-operation mechanism between kindlin and talin during the integrin activation process still remains to be elucidated. In conclusion here, it can be said that signals received by other receptors result in the binding of the talin and kindlin to the cytoplasmic tail region of integrins (Watanabe et al., 2008).

b) Outside-in signaling: occurs when integrins bind extracellular ligands and the resulting signal transduction takes place across the bilayer membrane towards the interior of the cell leading to the gene expression, cytoskeletal organization and modulation of the cell cycle (Yamada and Geiger, 1997). This occurs through several signaling pathways associated with outside-in integrin signaling (Ridley et al., 2003; Grashoff et al., 2004; Guo and Giancotti, 2004; Shattil and Newman, 2004). The α I-domain (when present), the β I-like domain and the β-propeller domain play a pivotal role in the recognition and binding of extracellular ligands. The α I-domain and the β I-domain share a few common features, they are both structurally homologous as they both adopt a Rossmann fold (a common feature of all vWA domains), which consists of central β-sheets surrounded by α-helices. Both of these domains bind ligands via direct interaction with a divalent metal ion bound at MIDAS; the β I-like domain also contains two additional binding sites known as SyMBS (Synergistic metal ion-binding site) and ADMIDAS (Adjacent to metal ion-dependent adhesion site), both of which are known to bind a Ca$^{2+}$ cation. In addition, the β-propeller domain plays a pivotal role in integrin function as it participates in ligand recognition either directly (in association with the β I-like domain) or indirectly (through the α I-domain). As the name suggests, the 7-bladed integrin β-propeller domain consist of seven repeats of ~60 residues that are arranged
toroidally or radially around a central pore to form seven ‘blades’ resembling the shape of a propeller wherein each blade is comprised of a four-stand antiparallel β-sheet.

In the case of the non I-domain containing integrins, the β I-like domain plays a pivotal role in ligand recognition in conjugation with the β-propeller domain as seen in the X-ray crystal structures of the extracellular region of αVβ3 in complex with an ‘RGD’ ligand (PDB ID: 1L5G; Xiong et al., 2002) and αIIbβ3 headpiece bound to fibrinogen gamma peptide ‘LGGAKQRGDV’ (PDB ID 2VDO, and 2VDR; Springer et al., 2008). (Figure 5). In the case of the αVβ3 structure, one of carboxylate oxygen atoms from the aspartate of the ‘RGD’ tripeptide ligand directly binds to the metal ion, in this case Mn$^{2+}$ at MIDAS and the second carboxylate oxygen hydrogen bonds with the mainchain amide NH from Tyr122 and Asn215. While the guanidinium group of arginine forms two salt bridges, one with Asp218 and the other with Asp150 (both from the β-propeller domain), the glycine residue resides at the interface between the α/β subunits, adds conformational flexibility to the loop. This particular ligand recognition mechanism is relatively strict in comparison to the α I-domain recognition system since the recognition site is positioned at the interface of the α subunit (β-propeller) and β subunit (β I-like domain). One of the shortcomings of this recognition mechanism is the inability to recognize and accommodate bulkier ligands and the recognition sequences e.g. RGD present on ligands need to be on loop regions that can insert into the narrow cleft between the β I-like domain and the β-propeller domain (Xiong et al., 2002). This mechanism is probably further stabilized by the insertion of a positively-charged residue, an arginine, from the β I-domain into the central cavity region of the β-propeller domain. Arg261 from integrin β3 is inserted at the core of the β-propeller cavity (from αV) and is surrounded by aromatic side chains e.g. Phe21, Phe159, Tyr224, Phe278, and Tyr406 from the propeller ring residues. Additionally, residues located in the upper ring – Tyr18, Trp93, Tyr221, Tyr275, and Ser403 – interact with residues from the lower ring in order to create a hydrophobic pocket for residues adjacent to Arg261 in the $3_{10}$-helix (Xiong et al., 2001).
In the case of the I-domain containing integrin α subunits, the divalent metal ion is located at MIDAS and is coordinated by other residues of the α I-domain and by water molecules. Ligands such as collagens and ICAMs coordinate to the metal cation via a negatively charged residue, glutamate, by displacing a water molecule and the remaining coordination positions are taken up by the highly conserved MIDAS residues such as serine and threonine residues. This complex involving a negatively-charged glutamate of the ligand with the metal ion at MIDAS is observed in the experimental 3D structures of the α2 I-domain in complex with the collagen-like triple-helical ‘GFOGER’ peptide (PDB ID: 1DZI; Emsley et al., 2000), the α1 I-domain in complex with triple-helical ‘GLOGEN’ peptide (PDB ID: 2M32, Chin et al., 2013) and ICAM3 in complex with the αL I-domain (PDB ID: 1T0P; Song et al., 2005). It is proposed that on ligand binding (e.g. to collagens, ICAMs) that the I-domain undergoes a dramatic conformational shift causing the intrinsic ligand, a glutamate

\[
\begin{array}{|c|c|c|}
\hline
\text{MIDAS residues} & \text{αC Helix} & \text{Intrinsic ligand} \\
\hline
\alpha_2 & \text{DYYVDCDSSGNYP...LTNTF...VTGEH...ILPFGIAVLYSRANKLTDKLNLIKE...FSIET} & \text{} \\
\alpha_1 & \text{DYYVDCDSSGNYP...QTMTA...VTGEH...TQRFIAIQLGSYGNLSTKEFVVEE...FAEAT} & \text{} \\
\alpha_{10} & \text{DYYVDCDSSGNYP...EYKTA...VTGEH...VTRYGIAVGLYGRDQPSSLRE...FGLES} & \text{} \\
\alpha_{11} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...VTRAYAVLQYRNQGINPETFLNE...FSIET} & \text{} \\
\alpha_D & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----HAFQGTPAQRG...YAVGST} & \text{} \\
\alpha_X & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----LAFQQRNWSRK...FAEQT} & \text{} \\
\alpha_M & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----DAPFSREKSLQG...FAEQT} & \text{} \\
\alpha_L & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----KHCFQKESQET...YVEAT} & \text{} \\
\alpha_S & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----EAFEKRAARTS...ISQGST} & \text{} \\
\text{-} & \text{Hr01} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----EYDISEF...FVLSEQ} \\
\text{-} & \text{Cio1} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----KVMG...ASLSEQ} \\
\text{-} & \text{Cio2} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----EANEKEF...FVLSEQ} \\
\text{-} & \text{Cio3} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----LSDA...SSEQQ} \\
\text{-} & \text{Cio4} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNSNE...} \\
\text{-} & \text{Cio5} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----DKNJE...VKLGQ} \\
\text{-} & \text{Cio6} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----ERVOE...VKLGD} \\
\text{-} & \text{Cio7} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----ERVOE...VKLGD} \\
\text{-} & \text{Cio8} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNDNN...LTDKSN} \\
\text{-} & \text{Cio9} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNDNN...LTDKSN} \\
\text{-} & \text{Cio10} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNDNN...LTDKSN} \\
\text{-} & \text{Cio11} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNDNN...LTDKSN} \\
\text{-} & \text{Cio12} & \text{DYYVDCDSSGNYP...EYKTA...ITGDKE...IIRYAIAGV-----SNDNN...LTDKSN} \\
\hline
\end{array}
\]

Figure 4. A sequence alignment of the integrin α I-domain region highlighting the ligand-binding divalent metal site MIDAS, the signature αC helix of the collagen receptors and the proposed intrinsic ligand (‘E336’ in the α2 I-domain) present in integrins containing the α I-domain. The human sequences are α1, α2, α10, α11, αD, αX, αL, αM and αE; while the ascidian sequences are denoted as Halocynthia Hra1 and Ciona Cia1-8. (‘E336’ in the α2 I-domain) to bind to MIDAS at the β I-domain thereby resulting in outside-in signal transduction (Alonso et al., 2002; Yang et al., 2004; Jokinen et al., 2010; Xie et al., 2010). Some of the features described here are shown in Figure 4; the α I-domain based ligand recognition mechanism is discussed in detail later under ‘Structure of collagen-binding and ICAM-binding integrins’. It is also worth mentioning that there are ligands that
do not adhere to MIDAS but bind in a metal-independent manner e.g. the cholesterol lowering drug lovastatin binding to αL I-domain (Kallen et al., 1999), a snake venom metalloproteinase (Ivaska et al., 1999; Pentikäinen et al., 1999) and echovirus 1 (Bergelson et al., 1994; King et al., 1997; Xing et al., 2004; Jokinen et al., 2010) and human recombinant collagen IX to the Ciα1 I-domain of C. intestinalis (Tulla et al., 2007).

Figure 5. A) Ligand binding at the junction of β-propeller domain (of the integrin α subunit in cyan) and the β I-like domain (of the integrin β subunit in grey). The ‘RGD’ tripeptide has been highlighted in yellow (Integrin αVβ3, PDB ID: 1L5G, Xiong et al., 2002). B) Interaction between the ligand ‘RGD’ (yellow) with integrin residues (β-propeller domain residues in cyan and β I-like domain residues in grey). Here Mn^{2+} ions are highlighted as spheres coordinating residues of MIDAS, ADMIDAS and SyMBS residues.
1.3 Relationship among integrin α and β subunits

Even though several multicellular organisms like fungi and plants do not express integrins (Whittaker and Hynes, 2002; Nichols et al., 2006), they (integrins) probably played a pivotal role in the development and diversification of multicellular animals, the metazoans. The integrin heterodimer has an early origin that most probably predates the first appearance of metazoans (Sebé-Pedrós et al., 2010), being present in a single-cell eukaryote and with most of the constituent domains identifiable in bacterial sequences (Ponting et al., 1999; Johnson et al., 2009). Additionally, the integrin subunits have been detected across the invertebrates (Burke, 1999) and several earliest-diverging animals (Brover et al., 1997; Müller, 1997; Pancer et al., 1997; Reber-Muller, 2001; Nichols et al., 2006; Srivastava et al., 2008; Schierwater et al., 2009; Knack et al., 2008; Srivastava et al., 2010).

Phylogenetic relationships among different integrin subunits have been extensively studied and represented over the past two decades (DeSimone and Hynes, 1988; Hughes 1992; Fleming et al., 1993; Burke 1999; Hynes and Zhao, 2000; Hughes 2001; Johnson and Tuckwell 2003; Ewan et al. 2005; Huhtala et al. 2005; Takada et al., 2007; Johnson et al., 2009) and they are mostly in agreement with each other. Figures 6 and 7 represent a schematic representation of the phylogenetic distribution among the integrin α and β subunits respectively.

In Figure 6, The integrin α subunits are segregated into two major groups: one group contains the additional I-domain (+ I-domain) while the other group does not (- I-domain). The non I-domain containing group is further segregated based on the Drosophila melanogaster integrin α subunits (Hynes and Zhao, 2000; Hughes 2001): the 'PS1' clade (laminin receptor clade), 'PS2' clade (fibronectin receptor clade), 'PS3' clade (invertebrate specific clade) and α4/α9 clade. This classification scheme is indicative of the ‘Position Specific’ antigens from D. melanogaster that were utilized to define several of the integrin α sequence clusters (Adams et al., 2000). The PS1 clade primarily consists of laminin receptor integrins like α3, α6 and α7 from vertebrates; as well as α9 and α10 from C. Intestinalis (Ewan et al., 2005); ina-1 from C. elegans and αPS1 from D. melanogaster (this is where each 'PS' group gets its naming from). The PS2 clade includes the fibronectin receptor integrins like αIIb, αV, α5 and α8 from vertebrates and α11 from C. intestinalis, α2 from H. roretzi, pat-2 from C. elegans and αPS2 from D. melanogaster. The PS3 clade is a special group as it consists of invertebrate sequences exclusively and the integrins αPS3, αPS4 and αPS5 from D. melanogaster are known to cluster within this group. The α4/α9 clade, as the
name suggests is comprised of the α4 and α9 integrins from vertebrates. One common feature shared by the non I-domain containing integrin α-subunits has been the consistency in overall topology of the integrin heterodimer. The subsequent insertion of the α I-domain in some integrins has led to further diversification and specialization of α subunits to cope with the evolution of the chordates. Since vertebrates are characterised by the complexity of the body plan, which includes a closed pressure circulatory system, central nervous system, immune system, cartilage and skeletal system that lends support to higher organisms, the α I-domains were probably a much needed invention in order to accommodate these functions. The I-domain containing integrin α subunits segregate into three distinct clades: collagen-receptor clade, the leukocyte receptor clade and ascidian clade. The collagen receptor clade consists of four subunits: α1, α2, α10 and α11; while the leukocyte surface clade consists of five subunits: αD, αX, αL, αM and αE.

The ascidian (or urochordate) clade is a monophyletic group and it consists of integrin sequences from *C. intestinalis* (Cia1-Cia8) and *H. roretzi* (Hra1) (Sasakura et al., 2003; Huhtala et al., 2005; Ewan et al., 2005). Although, not much can be said about their functional role at this point of time there have been a few pivotal studies that have certainly attempted to provide an insight. For instance, the Hra1 sequence from *H. roretzi* is known to adopt a function that is analogous to the vertebrate complement receptor 3 (CR3, Miyazawa et al., 2001). While, Cia1 I-domain does not bind to collagens I-IV or the ‘GFOGER’ tripeptide it has been shown to bind strongly to collagen IX, but in a metal-ion/MIDAS independent manner (Tulla et al., 2007). Additionally, it is worth mentioning that several species diverging close to the urochordates are now physically extinct (Donoghue and Purnell, 2005), thereby making it extremely hard to pinpoint the true origin of mammalian integrin orthologues. This problem has been further compounded by a lack of genomic data from the extant species thereby creating a knowledge gap within the evolutionary framework of integrins. Additionally, the α I-domains have not been detected in the lancelet genome (Heino et al., 2008; Putnam et al., 2008). Since genome-wide searches have already revealed that integrins with α I-domains are not observed in the lancelet, or in earlier-diverging invertebrate like the echinoderms (Heino et al., 2008; Johnson et al., 2009), this implies that the insertion of the α I-domain took place after the divergence of deuterostomes and after lancelet.
Figure 6. A schematic representation of the phylogenetic distribution among the integrin α subunits, Figure modified from publication V: ‘Evolution of Integrin I Domains’ by Johnson and Chouhan 2014. Sequences indicated as ‘Ciα’ belong to C.intestinalis; PS3 clade consists of sequences exclusively from invertebrates and other sequences are from higher vertebrates.

Figure 7. A schematic representation of the phylogenetic distribution among the integrin β subunits, Figure modified from publication V: ‘Evolution of Integrin I Domains’ (Johnson and Chouhan 2014).
Integrin β-subunits cluster into three major clades (Figure 7): the vertebrate clade, the chordate clade and the invertebrate clade. Vertebrate integrin subunits β1, β2 and β7 constitute the vertebrate clade while the vertebrate integrins β3, β6 and β8 along with Ciβ5 from *C. intestinalis* (the only invertebrate sequence to cluster within the chordate clade) constitute the chordate clade. The invertebrate clade as the name suggests is composed mostly of integrin β subunit sequences from the invertebrates like βPS and βv from *D. Melanogaster* but also including the ascidian sequences from ciona and halocynthia. In addition, integrin β subunit sequences from the placozoan *T. adhaerens*, poriferan *A. queenslandica* and choanozoan *C. owczarzaki* form the outliers to three major clades. Here, the β subunit sequence from the unicellular eukaryote *C. owczarzaki* has been utilized to root the tree.

### 1.4 N-terminal region: The β-propeller domain

The origin of the integrin constituent domains likely predates the integrin heterodimer itself and integrin constituent domains (like the 7-bladed β-propeller domain or the β I-like domain) have been reported to be a component of different prokaryotic proteins with unknown functions. For instance, in 1999 May and Ponting first reported similarities between bacterial sequences and integrin sequences when a PSI-BLAST run highlighted similarity between the cytoplasmic region from the human integrin β4 subunit and a hypothetical protein ‘slr1403’ from the prokaryote *Synechocystis* sp. PCC680 (cyanobacteria). In addition, May and Ponting also reported that the sequences ‘slr1403’, ‘slr0408’ and ‘slr1028’ contain at least 13 of the repeats found in β-propeller repeats from integrin α subunits. Another study from 2002 highlighted sequence similarity between a clone ‘M3G149’ from *Gemmata obscuriglobus* and the Ca$^{2+}$-region from integrin αV (Jenkins et al., 2002).

This observation of similarity between bacterial sequences and human integrin sequences was further addressed by our own research group when several prokaryotic sequences were identified that aligned surprisingly well with certain regions from the integrin subunits (Johnson et al., 2009). Some of these examples are as follows: a sequence ‘YP 721619’ from the cyanobacteria *Trichodesmium erythraeum* showed considerable similarity (over ~450 residues) with N-terminal region of the integrin β-subunit including the β I-like domain, but it lacked similarity with the stalk region domains (e.g. the I-EGF domains) or the TM region domain. Similarly, certain bacterial sequences showed similarity with the N-terminal region of various integrin α-subunits, especially the repeats corresponding to the β-
propeller domain but as expected these sequences did not show any similarity to the stalk region domains either (like the Thigh, Genu, Calf-1 and Calf-2 domains) or the TM region domains. To make things even more complex studies have also reported the presence of 7 (non-integrin type), 8, 6 and 4-bladed β-propeller domains in bacteria (Adindla et al., 2007; Quistgaard et al., 2009). Additionally, individual domains such as I-EGF, Ig and Rossmann folds have been seen in many proteins and in prokaryotes, but the earliest known intact integrin subunits are those reported by Sebé-Pedrós et al., in 2010 in which they highlight the presence of integrin-like sequences in the genomes of unicellular eukaryotes like C. owczarzaki and T. trahens (Sebé-Pedrós et al., 2010). Here we briefly discuss the two N-terminal regions from the integrin α-subunit and β-subunit i.e. 7-bladed β-propeller domain and β I-like domain.

According to the SCOP (Murzin et al., 1995) and CATH (Sillitoe et al., 2015) databases, the β-propeller fold can exist as 4, 5, 6, 7 or 8-bladed antiparallel β-sheets. The basic repeating unit observed in the β-propeller is a β-meander structural motif which is a series of antiparallel β-strands linked together by hairpin loops. These anti-parallel β-strands are arranged in a toroidal or radial fashion around a central pore resembling the blades of a propeller. The β-sheets pack tightly in order to give rise to a closed structure regardless of the β-propeller superfamily and fold type (i.e. 4, 5, 6, 7 or 8-bladed). According to one study the β-propeller families display unique characteristics that differentiate them from other repetitive folds (Chaudhuri et al., 2008). These features include immense diversity in β-propeller sequences thereby indicating multiple amplifications based on a single blade at different times, display of a single symmetry i.e. based on the amplification of a single blade and conservation in sequence motifs across the blades of a single β-propeller. Additionally, in many β-propeller structures, including the integrin 7-bladed β-propeller fold, a ‘velcro’ closure arrangement is observed where the N-terminal blade replaces the C-terminal blade of the last repeat (Fülöp and Jones, 1999).

According to the SCOP database, the integrin 7-bladed β-propeller fold is one of the fourteen protein superfamilies that adopt the same fold. Some of the other notable superfamilies are: the galactose oxidase central domain, nitrous oxide reductase N-terminal domain, WD40 repeat-like domain, clathrin heavy-chain terminal domain, peptidase/esterase 'gauge' domain, tricorn protease domain 2, putative isomerase YbhE, oligoxyloglucan reducing end-specific cellobiohydrolase and nucleoporin domain. One of the unique features observed in the integrin 7-bladed β-propeller fold is the presence of a consensus structural repeat known as the ‘cage motif’ (Xiong et al., 2001) or the ‘FG-GAP
motif’ (Springer, 1997; Loftus et al., 1994) with three or four Ca$^{2+}$-ion binding sites located at blades 4-7. Figure 8 below depicts the presence of Ca$^{2+}$-ions in the loops located at the bottom phase of blades 4-7 from the crystal structures of integrins αVβ3 (PDB ID: 1JV2) and αIIbβ3 (PDB ID: 2VDR) (Xiong et al., 2002; Springer et al., 2008).

![Figure 8](image_url)

**Figure 8.** Screenshots of the N-terminal β-propeller domain region from integrins αVβ3 (left panel, PDB ID: 1JV2) and αIIbβ3 (right panel, PDB ID: 2VDR) with Ca$^{2+}$ ions highlighted as spheres.

Integrin-like sequences have also been reported in heterokonts (or stramenopiles) which are considered to be a major alternate evolutionary line of eukaryotes containing more than 100,000 known species (including unicellular diatoms). A previous study has reported the presence of integrin-like sequences in the heterokont, a filamentous brown algae, *Ectocarpus siliculosus* (Cock et al., 2010). Some of the examples from the *Ectocarpus* genome are as follows: a sequence has been reported (Genbank/EBI sequence code: CBN77719.1) that shares around 40% sequence identity with the N-terminal region of human integrin αV subunit and its C-terminal region lacks the characteristic integrin ‘KXGFFXR’ motif which is not present in the *Ectocarpus* genome; instead a glycine and alanine-rich low complexity region is observed. Another sequence (EBI sequence ID: CBJ33612.1) has been reported that shares more than 20% sequence identity with the N-terminal region of human integrin β3 subunit, including MIDAS region.

### 1.5 N-terminal region: The vWA domain

The α I-domain and β I-like domain (from α-subunit and β-subunit respectively) adopt the vWA/Rossmann fold and they match with sequences from prokaryotes as well as from *Ectocarpus* (Johnson et al., 2009; Cock et al., 2010). vWA-domain containing proteins have been observed across all domains of life and they are known to be incorporated into proteins.
that adopt a wide variety of functions: collagens, complement factors, matrillins, integrins, copines, magnesium chelatase, ion channels etc. (Ponting et al., 1999; Whittaker and Hynes, 2002; Johnson and Tuckwell, 2003). The majority of these proteins having vWA domain are extracellular but the most ancient ones are located in eukaryotes and are intracellular proteins. The vWA domains adopt a classic α/β Rossmann fold where the central parallel β-sheets are flanked by α helices (CDD database, Marchler-Bauer et al., 2015). Studies have shown that the vWA domains have been most represented in proteins that play a pivotal role in the immune system or they are associated with proteins involved in cell-cell and cell-ECM recognition (Colombatti and Bonaldo, 1991; Colombatti et al., 1993; Whittaker and Hynes, 2002). Even though a wide array of knowledge exists in reference to the vWA domains it has not been possible to pinpoint or identify the source of insertion of these domains into the integrin repertoire (Tuckwell, 1999; Johnson and Tuckwell, 2003). But it can be stated that the integrin β-subunits were predicted to contain the vWA domain (Tuckwell, 1999; Ponting et al., 2000), which was later confirmed by X-ray crystal structure study of the αvβ3 integrin heterodimer (Xiong et al., 2001).

In integrins not having an α I-domain, the β I-like domain from the β-subunit plays a pivotal role in binding ligands through its MIDAS where the metal ion at the top crevice coordinates a carboxylate from the ligand sequence e.g. RGD, a feature very commonly observed among most vWA domains. The vWA domains of the integrin β-subunits are considered to be the most ancient among the ones that are involved in cell adhesion (Whittaker and Hynes, 2002). Already from the earliest diverging integrins, their ligands were most likely bound via interactions between the β I-like domain from the β-subunit and the β-propeller domain from the α-subunit (Wimmer et al., 1999). While, the insertion of the vWA domains in the α-subunit is a chordate-specific feature (Heino et al., 2008; Johnson et al., 2009). As discussed earlier this insertion event created three specific groups of I-domain containing integrin α-subunits: urochordate integrin with α I-domains, collagen-binding α I-domains and leukocyte-specific α I-domains; the latter two groups of α I-domains bind their respective ligands in a MIDAS-dependent manner. It is noteworthy that certain vWA domains can bind collagen in a MIDAS-independent manner too (Colombatti et al., 1993; Bienkowska et al., 1997; Huizinga et al., 1997; Romijn et al., 2001; Nishida et al., 2003).
1.6 Collagens and vertebrate collagen receptors

The Extra Cellular Matrix is complex and consists of a large number of biomolecules secreted by cells and the amount and interactions among biomolecules and with cells are tightly regulated for the proper functioning and fate of the cells that the ECM surrounds (Lin and Bissell, 1993). The ECM acts as a support platform that the cells can anchor to and subsequently form specialized tissues, therefore the ECM is essential for the organization, maintenance and remodeling of the body tissues. The vertebrates are characterized by having highly specialized support tissues such as the bone and cartilage of the skeletal system. A family of ECM proteins known as collagens play a significant role in maintaining the structure of those tissues among others; collagens are also involved in functions like cell adhesion, migration, tissue remodelling, differentiation, morphogenesis and wound healing (Myllyharju and Kivirikko, 2004). Collagen molecules are composed of three polypeptide α chains that can assemble to form homotrimers (identical α chains) or heterotrimers (different α chains). The three α chains form a left-handed helix, which is twisted to form a right-handed triple helix and give rise to a rod-shaped coiled-coil structure. This is a common structural feature shared by different collagens. The triple helical sequences are composed of a very basic repetitive 'GXY' motif, where 'G' signifies glycine while 'X' and 'Y' represent any given residue. It has been observed that the residue at position 'X' is often proline, while the residue at 'Y' is hydroxyproline, leading to a frequent repeat of 'GPO' where O represents hydroxyproline (Beck and Brodsky, 1998; Myllyharju and Kivirikko, 2004; Heino et al., 2008).

Collagens in humans are numbered from I to XXVIII, which are further categorized into various subfamilies based on sequence similarities and the complexes they form (Prockop and Kivirikko 1995; Boot-Handford et al., 2003; Boot-Handford and Tuckwell 2003, Myllyharju and Kivirikko, 2004; Ricard-Blum and Ruggiero 2005). Collagen subgroups similar to the human subgroups are found already in early chordates like *C. intestinalis* (Ewan et al., 2005). Some of the different types of collagens expressed in humans are fibril-forming collagens (i.e. collagens I, II, III, V, XI, XXIV and XXVII), beaded-filament forming collagen IV, anchoring fibril collagen VII, network collagens (IV, VIII and X) and TM domain containing α subunit collagens (collagens XIII, XVII, XXII and XXV). Additionally, Fibril Associated Collagens with Interrupted Triple helices or FACITs refer to probably the largest subgroup among the collagens (collagens IX, XII, XIV, XVI, XIX, XX, XXI and XXII). Interestingly, even a primitive organism like the freshwater sponge *E.*
*milleri* has at least two distinct collagens: one of them is fibril-forming (Exposito and Garrone, 1990) while the other one is non-fibrillar (Exposito *et al.*, 1990). The genome of *M. brevicolis*, a choanoflagellate, encodes at least two genes that consist of the repetitive collagen-like sequence with 'GXY' motif. Furthermore, one laminin-like and several integrin-like genes have also been reported from this same genome (King *et al.*, 2008), suggesting that they could have played an important role in metazoan evolution given that both collagens and integrins function in cell-cell and cell-matrix interactions.

Vertebrates have developed specific mechanisms to recognize and bind different collagen types through various collagen receptors. Besides integrins some other TM collagen receptors are: Discoidin domain receptors (DDR1 and DDR2), Glycoprotein VI (GPVI), Leukocyte-associated IG-like receptor-1 (LAIR-1), Mannose receptor family (MR) and urokinase-type plasminogen activator associated protein (uPARAP) and Endo180 (Heino *et al.*, 2008). All of these receptors clearly belong to different structural groups (Leitinger and Hohenester, 2007) but, just like integrins, they are all multi-domain collagen receptors and exhibit collagen specificity. However, the Mannose receptor family including the Endo180 are endocytosis receptors and they can bind denatured collagen and are thus not classical collagen receptors. Additionally, a recent review also mentions OSCAR and GPR56 as collagen receptors (Zeltz and Gullberg, 2016).

The most abundant and widely distributed collagen-binding integrin subunits are α1β1 (localized on mesenchymal cells, endothelial cells, smooth muscle cells, neurons, fibroblasts etc.) and α2β1 (localized on a variety of epithelial cells, platelets, endothelial cells, keratinocytes, fibroblasts etc.). During embryonic development α1β1 and α2β1 have the broadest tissue expression of the collagen receptors (Barczyk *et al.*, 2010; Leitinger, 2011). Integrin α10β1 appears to be expressed in cartilage, heart, trachea, lung, aorta and spinal chord and plays a critical role in skeletal development (Camper *et al.*, 2001; Lundgren-Åkerlund and Aszödi, 2014). Knockout studies have provided insight into the function of collagen receptor integrins and these studies have reported mild phenotypes where embryonic development was not hampered. For instance, knockout studies have shown that the α1 deficient mice may develop normally (Gardner *et al.*, 1996) while α2 deficient mice display defects in mammary gland branching morphogenesis as well as adhesion of platelets to collagen (Chen *et al.*, 2002, Holtkötter *et al.*, 2002). Meanwhile, α10β1 is more localised on the chondrocytes so it is well worth noticing that α10 knockout mice showcase a defect in the growth plate (Bengtsson *et al.*, 2005) while an α10 truncation in dogs causes a
chondrodysplasia (Kyöstilä et al., 2013). The expression of integrin \(\alpha 1\beta 1\) is more prominent in the regions that are rich in interstitial collagen-networks (Tiger et al., 2001) and it is implicated in regulation of periodontal ligament function in the erupting mouse incisor (Popova et al., 2007). It has also been reported that a loss of matrix component binding integrin like \(\alpha V\), \(\alpha 3\) and \(\alpha 8\) results in more serious defects as compared to a loss of a collagen-receptor integrin (Hynes, 2002).

One of the key areas where the collagen-binding integrins have been studied quite well is collagen-specificity (Kern et al., 1993, Tuckwell, 1995, Tiger et al., 2001 and Tulla et al., 2001). Integrin \(\alpha 1\) prefers collagens IV and VI along with the fibril forming collagens, while the \(\alpha 10\) subunit shares a similar specificity as \(\alpha 1\) but it can also bind collagen II. Meanwhile integrins \(\alpha 2\) and \(\alpha 11\) preferentially bind the fibril-forming collagens. The two major driving factors behind the success of collagen-binding integrin studies are: firstly, the possibility to crystallize isolated integrin I-domains and their ability to retain specificity towards their respective ligand collagens (Kamata and Takada 1994; Tuckwell et al., 1995), secondly, the synthetic tripeptides (for instance ‘GFOGER’ and integrin \(\alpha 1\)) have been greatly instrumental in identifying integrin-specific binding sites (Knight et al., 1998, 2000). Additionally, two more integrin binding sites (‘GLOGER’ and ‘GASGER’) were reported for integrins \(\alpha 1\) and \(\alpha 2\); ‘GFOGER’ has also been reported to be a binding site for integrin \(\alpha 11\) (Zhang et al., 2003). There are two important integrin \(\alpha\) I-domain structures available from \(\alpha 2\) and \(\alpha 1\) that showcase binding to ‘GFOGER’ and ‘GLOGEN’ tri-peptides respectively (PDB ID: 1DZI, Emsley et al., 2000; PDB ID: 2M32, Chin et al., 2013).

### 1.7 Structure of collagen-binding and ICAM-binding integrins

The earliest structural revelations on the integrin heterodimer were from electron microscopy reconstructions, which showed a bent structure (Carrell et al., 1985; Nermut et al., 1988), as well as the analysis of integrin sequences (Nermut et al., 1988; Arnaout, 1990). The I-domain was already a subject of great interest early on (Larson et al., 1989; Arnaout, 1990) and some of the earliest solved integrin crystal structures are that of \(\alpha\) I-domain from the \(\alpha\)-subunit e.g. \(\alpha I\)-domain of the immune system: \(\alpha M\) (PDB ID: 1IDO and 1JLM, Lee et al., 1995a,b) and \(\alpha L\) (PDB ID:1LFA, Qu and Leahy 1995); and \(\alpha\) I-domain of the collagen-receptor integrin-type: \(\alpha 2\) without (PDB ID: 1A0X, Emsley et al., 1997) and with (PDB ID: 1DZI, Emsley et al., 2000) collagen-like triple-helical ‘GFOGER’ tri-peptide bound.
Subsequently, the first integrin headpiece structures were also solved, for instance: extracellular segment of integrin αVβ3 (PDB ID: 1JV2, Xiong et al., 2001), integrin αIIbβ3 headpiece segment bound to a fibrinogen peptide (PDB ID: 2VDL; Springer et al., 2008), α4β7 headpiece complexed with Fab ACT-1 (PDB ID: 3V4P, Yu et al., 2012), α5β1 integrin headpiece in complex with ‘RGD’ peptide (PDB ID: 3VI4, Nagae et al., 2012) recently two I-domain containing integrin ectodomain structures were solved, that of αXβ2 (resolution: 3.5 Å; PDB ID: 3K6S, Xie et al., 2010) and αLβ2 (resolution: 2.15 Å; PDB ID: 5E6S, Sen and Springer, 2016). In addition, quite a few structures deposited in the PDB repository correspond to the integrin TM and cytoplasmic region. Some of these structures are: NMR structure of αIIbβ3 cytoplasmic domain (PDB ID: 1M8O, Vinogradova et al., 2002), NMR structure of the cytoplasmic domain of integrin αIIb in DPC micelles (PDB ID: 1S4W, Vinogradova et al., 2004), platelet integrin αIIbβ3 transmembrane-cytoplasmic heterocomplex (PDB ID: 2KNC, Yang et al., 2009), structures and interaction analyses of the integrin αMβ2 cytoplasmic tails (PDB ID: 3V4P, Yu et al., 2012), integrin αIIbβ3 transmembrane complex (PDB ID: 2K9J, Lau et al., 2009) to name a few.

The aforementioned two structures (αXβ2 and αLβ2) that include leukocyte ectodomains show the relationship of this inserted domain to the β-propeller from which it buds out of the α subunit, and relationship to the β I-like domain and the β subunit. The insertion of the α I-domain has functioned to substantially enhance the ligand binding capacity of integrins through easy access to more bulky ligands, like collagen fibers and ICAM Ig-fold domains, in comparison to the flexible loops on ligands recognized by non-I domain integrins. As mentioned earlier the α I-domain and the β I-like domain are homologous as they adopt the same fold (Rossmann fold) and they have been categorized as members of the same family (vWA ECM family) but specifically they belong to the vWA ECM protein subfamily along with seven other similar domains (Conserved Domain Database: CDD, Marchler-Bauer et al., 2015). Apart from the ECM subfamily there are at least eighteen different intracellular vWA subfamilies like midasin, copine, complement factors, collagen, trypsin inhibitor and magnesium chelatase to name a few. The vWA domains are mainly associated with proteins involved in cell-cell and cell-ECM recognition as they are key constituent domains of receptor proteins as well as pivotal ECM proteins like collagen (Colombatti and Bonaldo, 1991; Colombatti et al., 1993; Whittaker and Hynes, 2002). One significant feature that is common to both the α I-domain and the β I-like domain is the presence of MIDAS where a divalent cation (like Mg$^{2+}$ or Mn$^{2+}$ but not Ca$^{2+}$ due to its large size) is localized towards the top surface of the domain. Furthermore, the β I-like domain has two additional Ca$^{2+}$-binding
sites located near MIDAS, they are ADMIDAS (Adjacent to MIDAS) and LIMBS (ligand-associated metal-binding site).

In the structure of α2 I-domain bound to ‘GFOGER’ tri-peptide (Emsley et al., 2000) the metal ion is coordinated by highly conserved residues: D151 (via a water molecule), S153 (via hydroxyl oxygen) and S155 (via hydroxyl oxygen) located on loop 1, T221 (via hydroxyl oxygen) located on loop 2 and D254 (via a water molecule) and E256 (via a water molecule) located on loop 3. The remaining coordination positions are taken up by water molecules and a negatively charged residue (like glutamate) from the ligand which displaces the water molecule in order to bind to the metal ion. Additionally, phenylalanine from the middle strand (of the ligand GFOGER tripeptide) rests atop the side chains of Q215 and N154 (from the I-domain), phenylalanine from the trailing stand is engaged in hydrophobic interactions with Y157 and L286 (from the I-domain) and arginine is located in an acidic pocket close to E256 and no salt bridge is formed. Furthermore N154 and Y157 from loop 1, H258 from loop 3 form hydrogen bonds with the collagen.

A comparison between the ‘GFOGER’ tripeptide collagen-bound integrin (PDB ID: 1DZI, Emsley et al., 2000) and unliganded integrin α2 I-domain (PDB ID: 1AOX; Emsley et al., 1997) reveals that movement is observed in the MIDAS loops and α1 helix due to the direct coordination of the metal ion, a ‘slinking’ motion is observed between the αC helix and the α6 helix resulting in the formation of the collagen binding groove coupled with a large displacement of the α7 helix relaying the signal downwards (secondary structure names for α helices and loops are derived from Emsley et al., 2000). Therefore, in the case of ligand binding at the α2 I-domain, it was observed that the metal ion is displaced about 2.6 Å closer to the T221 in order to establish a direct bond, this movement of the metal ion is closely followed by the MIDAS loop 1 from the I-domain to maintain the direct bonds via S153 and S155. The MIDAS loop 3 is also rearranged and this causes the loss of direct bond between the side chain of D254 and the metal ion and E256 forms a water molecule mediated bond with the metal ion. A shift observed in loop 1 and 3 causes a rearrangement of the αC and α7 helix, where the α7 helix experiences a downward displacement of about 10Å and this movement in turn results in breaking a salt bridge between E318 from α7 and R288 from the αC helix. This causes the αC helix to unwind and in turn enables the connecting loop at the N-terminus of helix 6 to form an additional turn. Also, R288 shifts closer to MIDAS and forms a water-mediated salt bridge with D254 (Emsley et al., 2000).
These are some of the broad changes that take place with the transition of the α2 I-domain from unliganded to liganded.

A similar movement can also be observed in the case of the αM I-domain where conformational shifts are observed with respect to the change in metal coordination. In one of the crystals (Lee et al., 1995a) of the αM I-domain the metal ion is coordinated by a glutamate from the neighboring I-domain in the crystal lattice thereby completing the coordination sphere of the metal ion and it was proposed that the glutamate behaved as a ‘ligand-mimetic’. A parallel can be drawn between the GFOGER tripeptide bound α2 I-domain and the ‘ligand-mimetic’ αM I-domain structure as MIDAS structure is quite identical and glutamate coordinates the metal ion. While, another structure of the αM I-domain obtained under different conditions which lacked the ‘ligand-mimetic’ has a similar MIDAS motif as the unliganded α2 I-domain (Emsley et al., 1997). A comparison of these two αM I-domain structures (with and without the ‘ligand-mimetic’) with the liganded and unliganded α2 I-domain structure reveals the broad similarities and the conformational changes observed in the I-domains. Both cases detail a movement in the metal coordination and resulting in a bond with a threonine and a loss of bond with an aspartic acid. A similar MIDAS mediated ligand recognition mechanism can also be observed in the recently solved 3D crystal structure of α1 I-domain bound to ‘GLOGEN’ (PDB ID: 2M32, Chin et al., 2013) where the glutamate from the collagen-like triple-helical peptides binds at a coordinating position to the divalent metal cation. Similarly, the negatively charged glutamate from ICAM1 (PDB ID: 1MQ8, Shimaoka et al., 2003; PDB ID: 3TCX, Kang et al., unpublished), ICAM3 (PDB ID: 1TOP, Song et al., 2005) and ICAM5 (PDB ID: 3BN3, Zhang et al., 2008) also bind to MIDAS of the αL I-domain.

One of the key diagnostic features of the collagen receptor integrin is the presence of an additional αC helix which is located towards the carboxy-terminus of the I-domain (‘284GYLNR288’ from PDB ID: 1A0X; Emsley et al., 1997, Figure 9). The αC helix can be only observed in the ‘closed’ (inactive) confirmation of the α I-domain as the ‘open’ (active) confirmation forces this helix to unwind and create a groove where the collagen molecule binds and coordinates the divalent metal ion at the MIDAS. Additionally, this αC helix is not present in the I-domains of the leukocyte integrins, I-domains of the tunicates and the β I-like domain from the β-subunit. Apart from the obvious functional role of the αC helix, it serves as a marker that distinguishes between the two functionally-distinct sets of I-domains found throughout the vertebrates. Deletion of the αC helix from the α2 I-domain
does not inhibit binding to collagen I but the affinity is certainly reduced (Käpylä et al., 2000). Recombinant Ciα1 from the ascidian C. intestinalis cannot bind the ‘GFOGER’ motif from collagen or the fibril-forming collagens. But, it does bind to Collagen IX in a MIDAS and metal ion independent manner (Tulla et al., 2007). These results show that the metal ion dependent and MIDAS mediated adhesion of collagen to integrins is a characteristic of the vertebrates and the evolution of this specialised collagen receptor has played a crucial role in the development of bones, cartilage and the blood vessel system.

Integrins without the α I-domain bind simple signature sequences e.g. “RGD” on solvent exposed loops of their ligands, thereby limiting the ligands that can be recognized. However, with the insertion of the highly solvent-exposed α I-domain integrins have gained access to unprecedented and bulkier ligands and this insertion event has forced the β I-like domain to take on a new responsibility as it binds a conserved ‘intrinsic ligand’ glutamate (E336 in α2I) which in turn aids in stabilizing the integrin active conformation and modulating signal transduction. Structural and experimental data support a mechanism whereby when a ligand binds to MIDAS of the α I-domain then the intrinsic ligand glutamate binds to MIDAS of the β I-like domain as part of the downward integrin signaling mechanism (Alonso et al., 2002; Yang et al., 2004; Xie et al., 2010; Jokinen et al., 2010). But unfortunately, there is no direct structural evidence available to confirm this at this point because the structure is in the inactive bent conformation. Mutational analyses have shown that the mutation of this glutamate can lead to abolishment of integrin activation (Huth et al. 2000; Alonso et al. 2002). Another supporting observation in favor of this statement is the flexibility displayed by the α I-domain, especially the α7 helix located towards the C-terminal region. The downward displacement of the α7 helix could definitely provide the necessary push for interdomain interaction between the α I-domain and the β I-like domain (Xie et al., 2009).
**Figure 9.** MIDAS residues coordinating the divalent metal ion (green spheres) in A: Integrin α2 I-domain in complex with ‘GFOGER’ tri-peptide (PDB ID: 1DZI), B: Integrin α1 I-domain in complex with ‘GLOGEN’ tri-peptide (PDB ID: 2M32) and C: Integrin αL I-domain in complex with ICAM-3 (PDB ID: 1T0P). Key water molecules are highlighted as red spheres and MIDAS residues are displayed as sticks in cyan. The I-domain binding glutamate from the three ligands (ICAM, ‘GFOGER’ or ‘GLOGEN’) is shown as Glu11, Glu212 and Glu37 in panels A, B and C respectively (page 27).

**Figure 10.** A) Representative structures from both known conformations of the α2 I-domain are superposed i.e. the liganded conformation (with ‘GFOGER’ tripeptide in wheatish tint; PDB ID: 1DZI, Emsley et al., 2000) and the unliganded conformation (in grey and the αC helix ‘284GYLNR288’ in green; 1A0X; Emsley et al., 1997) where the αC helix has unwound (RMSD = 0.73 Å). B) A close up shot in panel B. C) Superposition of the unliganded α2 I-domain structure (in grey and the αC helix in green; PDB ID: 1A0X) with the unliganded αL I-domain structure (in lightblue tint; PDB ID: 1ZON; Qu and Leahy, 1996) where the αC helix region is clearly absent (RMSD = 1.24 Å). D) A close up shot in panel D. E) Collagen-like ‘GFOGER’ tripeptide bound to the α2 I-domain through a metal ion coordinating Glu11 (from the tripeptide) located at the top phase of the α2 I-domain and the green highlighted region represents the unwound αC helix (PDB ID: 1DZI). F) ICAM-3 bound to the αL I-domain through a metal ion coordinating Glu37 from (ICAM-3) located at the top phase of the αL I-domain and the green highlighted region represents the absence of an αC helix (PDB ID: 1DZI) (page 28).
Figure 9.
Figure 10.
2. Aims of the study

The specific aims of this thesis were:

i) To determine the possible origin of the key functional integrin domains, especially the vWA-like α I-domain and the 7-bladed β-propeller domain.

ii) To identify when during evolution that the α I-domain was incorporated into some integrin α subunits along with the possible source of that domain.

iii) To identify how distinct are the integrin α I-domains from the other vWA domains.

iv) To determine the point of origin for the human or mammalian integrin orthologues.
3. Methods

3.1 Online databases

Protein sequences and structures used in this thesis work were obtained from online databases. Protein sequences of interest were extracted from online databases like the NCBI protein database (http://www.ncbi.nlm.nih.gov/protein), Uniprot knowledgebase (http://www.uniprot.org) and the Ensembl genome project (http://www.ensembl.org). Elephant shark sequences were downloaded from the Elephant shark genome project (http://esharkgenome.imcb.a-star.edu.sg/). The NCBI protein database consists of a vast collection of protein sequences with additional information from sources like RefSeq, Swiss-Prot, GenPept, PIR, PDB, and PRF. The Ensembl project hosts a wide variety of genome databases ranging from early eukaryotic organisms to higher vertebrates and the Uniprot knowledgebase (UniprotKB) contains functional information on protein sequences. Functional information in UniprotKB is derived from a combination of manual annotation from Swiss-Prot as well as automatic annotation from TrEMBL. The emphasis behind every UniprotKB entry is to capture as much annotation information as possible like protein description, functional information, and taxonomic data along with external database references. The BLAST (Basic Local Alignment Search Tool: http://blast.ncbi.nlm.nih.gov/Blast.cgi) service at the NCBI web page was utilised to perform searches across various sequence databases in order to identify and create a dataset of homologous sequences.

Protein sequences were also referenced against databases like CDD and Pfam (Protein families database: http://pfam.xfam.org/) for validation. CDD is a large collection of well annotated multiple sequence alignment models made publically available as PSSMs (Position Specific Scoring Matrices) which helps in identifying protein domains through RPS-BLAST. CDD also includes external data from sources like NCBI curated domains which contains information derived from experimentally solved 3D structures and it helps in identifying domain boundaries. Furthermore, domain models are incorporated from a variety of different sources like SMART (Simple Modular Architecture Research Tool), PRK (PRotein K(c)lusters), COGs (Clusters of Orthologous Groups of proteins), TIGRFAM (The Institute for Genomic Research’s database of protein families) and Pfam (Protein families). The Pfam database is a collection of curated protein families and each protein family is defined by sequence alignments and profile Hidden Markov Model (HMM). The computer program HMMER is implemented to build profile HMMs and subsequently searches are
conducted against a large sequence database based on UniProtKB. It is also noteworthy that Pfam is composed of two major sections: Pfam-A which consists of manually curated entries and it covers majority of the entries present in the database while Pfam-B is composed of automatically generated entries.

All the protein structures were obtained from RCSB Protein Data Bank (Berman et al., 2000). PDB is a key repository that contains experimentally solved protein structures, nucleic acids and complexes which are determined by techniques like X-ray crystallography and Nuclear magnetic resonance. Furthermore, SCOP database (Structural Classification of Proteins: http://scop.mrc-lmb.cam.ac.uk/scop) was consulted in order to assign protein families and folds. SCOP database is manually curated in assistance with automated tools and it provides pivotal information in relation to structural and evolutionary relationships shared by proteins with a known structure. The SCOP classification of proteins covers several levels of hierarchy but the key levels are family, superfamily and fold.

3.2 Protein sequence analyses

Multiple sequence alignments were constructed in order to closely assess: the conservation level of key residues among the constituent sequences, inspect the secondary structural elements among the target sequences and provide distance values between all pairs of sequences in order to reconstruct phylogenetic trees. Computer programs like T-COFFEE (Notredame et al., 2000) and ClustalW (Larkin et al., 2007) were utilized for creating alignments which were then manually curated for obvious errors. Some of these refined alignments (unpublished) were also utilised for phylogenetic analyses. Pairwise alignments for 3D-modeling were constructed using using Malign (Johnson and Overington, 1993) from the Bodil package (Lehtonen et al., 2004). Secondary structure prediction (SSP) techniques are aimed at predicting secondary structural elements like α-helices, β-sheets and turns within proteins sequences based on the information derived from the known primary protein sequence. Computer program implemented in SSP are known to perform with a high accuracy therefore we used a combination of various SSP methods namely PHD from PredictProtein (Rost B et al., 2004), PSIPRED (McGuffin LJ et al., 2000), PROF (Ouali and King, 2000), Jpred (Cole et al., 2008), GOR (Sen et al., 2005) and Porter (Pollastri and McLysaght, 2005) in order to predict secondary structural elements for our target sequences and fragments. The Weblogo server (Crooks et al., 2004) was utilised to examine the amino acid frequency at pivotal positions in the β-propeller repeat.
**Table 2.** (For publication I) Sequences used in the alignment of the β-propeller domain with two representative sequences from human and five representative sequences from different bacterial species.

<table>
<thead>
<tr>
<th>Protein accession code</th>
<th>Organism</th>
<th>Protein name</th>
<th>PDB ID</th>
<th>UNIPROT</th>
<th>NCBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>2VDR (PDB)</td>
<td>Human</td>
<td>Integrin αIIb</td>
<td>2VDR</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1JV2 (PDB)</td>
<td>Human</td>
<td>Integrin αV</td>
<td>1JV2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A3VFV0</td>
<td>Rhodobacterales bact. HTCC2654</td>
<td>Hyp. Protein</td>
<td>-</td>
<td>A3VFV0</td>
<td>ZP_01013484</td>
</tr>
<tr>
<td>ASL751</td>
<td>Vibrionales bact. swat 3</td>
<td>Hyp. Protein</td>
<td>-</td>
<td>ASL751</td>
<td>ZP_01816147</td>
</tr>
<tr>
<td>Q3JAB2</td>
<td>N. oceani ATCC 19707</td>
<td>Integrin α-like</td>
<td>-</td>
<td>Q3JAB2</td>
<td>YP_343764</td>
</tr>
<tr>
<td>A0YNP0</td>
<td>Lyngbya sp. pcc 8106</td>
<td>Hyp. Protein</td>
<td>-</td>
<td>A0YNP0</td>
<td>ZP_01620661</td>
</tr>
<tr>
<td>Q31NK2</td>
<td>S. elongatus pcc 7942</td>
<td>Integrin α-like</td>
<td>-</td>
<td>Q31NK2</td>
<td>YP_400354</td>
</tr>
</tbody>
</table>

In publication I, two X-ray crystal structures (αIIbβ3 PDB ID: 2VDR (Springer et al., 2008) and αVβ3 PDB ID: 1JV2 (Xiong et al., 2001) were studied in order to understand the ‘FG-GAP’ (Pfam01839) or the ‘cage’ motif. Upon establishing a consensus repeat pattern within the β-propeller domain, we narrowed down our dataset from 562 bacterial sequences to nine sequences from five different bacterial species which share the seven consensus repeat pattern with the human β-propeller domain. These bacterial sequences were subjected to secondary structure prediction using three different prediction methods (PHD, PSIPRED and PROF) and were subsequently aligned manually against β-propeller domain sequences from αV and αIIb in order to highlight the conservation level of the ‘FG-GAP’/cage motif as well as the secondary structural elements.

In publication II, 16 chordate sequences were aligned across the span of integrin α I-domain region in order to highlight the area of extensive similarity as well as the presence or absence of the characteristic αC helix which is a hallmark of the collagen-binding integrins. In our dataset we included two human collagen-binding integrin α I-domain sequences with
the characteristic αC helix region; α1 and α11, two leukocyte surface integrin α I-domain sequences which lack the αC helix; αM and αD, along with all the known α I-domain sequences from *C. intestinalis*. Secondary structure prediction for the sea lamprey sequences

**Table 3.** (For publication II) Sequences used in the alignment of the I-domain region with 16 representative sequences from different chordate species.

<table>
<thead>
<tr>
<th>No.</th>
<th>Protein name</th>
<th>Organism</th>
<th>Protein coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Integrin α1</td>
<td>Human α1</td>
<td>NP_852478 (NCBI)</td>
</tr>
<tr>
<td>2.</td>
<td>Integrin α11</td>
<td>Human α11</td>
<td>NP_001004439.1 (NCBI)</td>
</tr>
<tr>
<td>3.</td>
<td>Integrin αM</td>
<td>Human αM</td>
<td>NP_000623.2 (NCBI)</td>
</tr>
<tr>
<td>4.</td>
<td>Integrin αD</td>
<td>Human αD</td>
<td>NP_005344.2 (NCBI)</td>
</tr>
<tr>
<td>5.</td>
<td>Pma_f1</td>
<td>Sea lamprey f1</td>
<td>Scaffold GL479139</td>
</tr>
<tr>
<td>6.</td>
<td>Pma_f2</td>
<td>Sea lamprey f2</td>
<td>Scaffold GL477642</td>
</tr>
<tr>
<td>7.</td>
<td>Pma_f3</td>
<td>Sea lamprey f3</td>
<td>Scaffold GL501125</td>
</tr>
<tr>
<td>8.</td>
<td>Ebu_f</td>
<td>Hagfish</td>
<td>BJ655520.1 (NCBI)</td>
</tr>
<tr>
<td>9.</td>
<td>Cin_ α1</td>
<td>Sea squirt α1</td>
<td>ci0100131118</td>
</tr>
<tr>
<td>10.</td>
<td>Cin_ α2</td>
<td>Sea squirt α2</td>
<td>ci0100149446</td>
</tr>
<tr>
<td>11.</td>
<td>Cin_ α3</td>
<td>Sea squirt α3</td>
<td>ci0100130596</td>
</tr>
<tr>
<td>12.</td>
<td>Cin_ α4</td>
<td>Sea squirt α4</td>
<td>ci0100130838</td>
</tr>
<tr>
<td>13.</td>
<td>Cin_ α5</td>
<td>Sea squirt α5</td>
<td>ci0100152002</td>
</tr>
<tr>
<td>14.</td>
<td>Cin_ α6</td>
<td>Sea squirt α6</td>
<td>ci0100131399</td>
</tr>
<tr>
<td>15.</td>
<td>Cin_ α7</td>
<td>Sea squirt α7</td>
<td>ci0100152615</td>
</tr>
<tr>
<td>16.</td>
<td>Cin_ α8</td>
<td>Sea squirt α8</td>
<td>ci0100130149</td>
</tr>
</tbody>
</table>

(Pma_f1, Pma_f2 and Pma_f3) were conducted with the help of five prediction methods Jpred, GOR, Porter, Prof_seq and PSIPRED (Jones, 1999). X-ray crystal structure of integrin α1 I-domain (PDB ID: 1PT6 (Nymalm *et al*., 2004) was used to assign secondary structural elements to the sequence alignment. This multiple sequence alignment clearly
highlights the presence of MIDAS residues as well as the αC helix in the lamprey sequences.

**Table 4.** (For publication III) Chordate genomes and EST assemblies utilized for the creating three multiple sequence alignment datasets. (Also refer supplementary table1 in publication III). Here ‘-’ indicates classification not available.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Sequence code used</th>
<th>Scientific name</th>
<th>Subphylum / Superclass / Class / Subclass / Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Hsa</td>
<td><em>Homo sapiens</em></td>
<td>Vertebrata / Tetrapoda / Mammalia / Theria / Primates</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Ptr</td>
<td><em>Pan troglodytes</em></td>
<td>Vertebrata / Tetrapoda / Mammalia / Theria / Primates</td>
</tr>
<tr>
<td>Horse</td>
<td>Eca</td>
<td><em>Equus caballus</em></td>
<td>Vertebrata / Tetrapoda / Mammalia / Theria / Perissodactyla</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mmu</td>
<td><em>Mus musculus</em></td>
<td>Vertebrata / Tetrapoda / Mammalia / Theria / Rodentia</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gga</td>
<td><em>Gallus gallus</em></td>
<td>Vertebrata / Tetrapoda / <em>Aves</em> / - / Galliformes</td>
</tr>
<tr>
<td>African clawed frog</td>
<td>Xtr</td>
<td><em>Xenopus laevis</em></td>
<td>Vertebrata / Tetrapoda / Amphibia / - / Anura</td>
</tr>
<tr>
<td>Green spotted pufferfish</td>
<td>Tni</td>
<td><em>Tetraodon nigroviridis</em></td>
<td>Vertebrata / Osteichthyes / Actinopterygii / Neopterygii / Tetraodontiformes</td>
</tr>
<tr>
<td>Nile tilapia</td>
<td>Oni</td>
<td><em>Oreochromis niloticus</em></td>
<td>Vertebrata / Osteichthyes / Actinopterygii / Neopterygii / Perciformes</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Dre</td>
<td><em>Danio rerio</em></td>
<td>Vertebrata / Osteichthyes / Actinopterygii / Neopterygii / Cypriniformes</td>
</tr>
<tr>
<td>Common carp</td>
<td>Cca</td>
<td><em>Cyprinus carpio</em></td>
<td>Vertebrata / Osteichthyes / Actinopterygii / Neopterygii / Cypriniformes</td>
</tr>
<tr>
<td>Elephant shark</td>
<td>Cmi</td>
<td><em>Callorhinchus milii</em></td>
<td>Vertebrata / Chondrichthyes / Chondrichthyes / Holocephali / Chimaeriformes</td>
</tr>
<tr>
<td>Inshore hagfish</td>
<td>Ebu</td>
<td><em>Eptatretus burgeri</em></td>
<td>Vertebrata / - / Myxini / - / Myxiniformes</td>
</tr>
<tr>
<td>Sea lamprey</td>
<td>Pma</td>
<td><em>Petromyzon marinus</em></td>
<td>Vertebrata / - / Cephalaspidomorphi / - / Petromyzontiformes</td>
</tr>
<tr>
<td>Vase tunicate</td>
<td>Ci</td>
<td><em>Ciona intestinalis</em></td>
<td>Tunicata / - / Asciadiacea / - / Enterogona</td>
</tr>
<tr>
<td>Sea pineapple</td>
<td>Hro</td>
<td><em>Halocynthia roretzi</em></td>
<td>Tunicata / - / Asciadiacea / - / Pleurogona</td>
</tr>
</tbody>
</table>
In publication III, three different multiple sequence alignments datasets (unpublished) were created based on the length of sequence alignment and subjected to phylogenetic analyses (also refer the next section):

a) 69 sequences with the nearly full length (Pma_f3) sea lamprey integrin α-sequence,
b) 72 sequences with a coverage of the common region (406-409 residues) including the three lamprey sequences Pma_f1-3 and
c) 73 sequences with a coverage of the α I-domain region (~200 residues) including the three lamprey sequences Pma_f1-3 and the hagfish fragment (Ebu_f).

Phylogenetic studies: Phylogenetic studies deal with evolutionary relationship among species based on molecular sequence data (like DNA or protein) or morphological data. In the case of molecular sequence data, a well-refined multiple sequence alignment is supplied to a phylogenetic program in order to establish evolutionary relationship among the candidate species form the sequence alignment. This evolutionary relationship can be assessed through various phylogenetic approaches like Neighbour Joining (NJ), Maximum Likelihood (ML), Maximum Parsimony (MP) and Bayesian Method (BM).

Phylogenetic analyses were performed using MEGA (Tamura et al., 2011), MrBayes (Huelsenbeck et al., 2001) and Phylip (Felsenstein, 1989). The NJ phylogenetic test was based on the Jones-Taylor-Thornton (JTT) distance matrix, which was made using MEGA and phylip. The ML phylogenetic test was based on the Whelan and Goldman (WAG, Whelan et al., 2001) matrix, which was selected by ProtTest (Darriba et al., 2011) and MEGA to be the best-fit model based on the Bayesian Information Criterion (BIC) which is implemented for selecting the best statistical model (among a set of finitie statistical models) for any given data. Felsenstein’s bootstrap method (Felsenstein, 1985) was implemented to validate the tree topology. BM phylogenetic analyses were performed based on the WAG matrix with MCMC analysis. The Bayesian posterior probability was used to assess the confidence level of the tree nodes.

In publication III, three datasets were created (mentioned above) and each dataset was subjected to NJ, ML and BM phylogenetic analyses.

3.3 3D modelling and structural analyses

3D modelling or comparative modelling is a very useful technique for studying proteins that lack an experimentally solved 3D structure. As the name suggests the target protein sequence is modelled to an atomic resolution based on a template structure (which is
actually an experimentally solved 3D structure) belonging to a homologous protein family. Since related proteins share a higher sequence identity along with similar 3D folds, this makes it possible to generate a quality working model. However, the lower the sequence identity between the template and the target sequence, the lower the quality of the model. Where the sequence alignment is wrong, the resulting model will be wrong too, so much effort is made to evaluate the alignment and estimate levels of reliability.

3D modelling was performed using Modeller (Sali and Blundell, 1993) and Homodge from Bodil (Lehtonen et al., 2004). Modeller generates 3D models by optimally satisfying spatial restraints derived from the alignment, structures related to the target, protein structures in general and expressed as probability density functions (pdfs) for the features restrained. The program also incorporates limited functionality for ab initio structure prediction of loop regions of proteins, which are often highly variable even among homologous proteins and therefore difficult to predict by homology modelling. The Homodge program is located in the Bodil package and it helps in generating a 3D model by relying on information from the template structure with minimum alterations. This method is useful especially in case of proteins with high sequence similarity. The generated model does not undergo minimization which makes the process quite fast and quick to assess the quality of the model and prompt any further action like refinement of the model or the sequence alignment. The models constructed with homodge were subjected to energy minimization with the Charmm forcefield (Brooks et al., 1983) from the Discovery Studio package (http://accelrys.com/). All of the models were visualized and analysed with Bodil and the side-chain confirmations were investigated using the rotamer option. VERTAA from the Bodil package was used to superimpose 3D models on to their respective template structure. In publication III, 3D models for the protein sequences Pma_f1, Pma_f2 and Pma_f3 (from sea lamprey) were created in the ligated and closed conformation respectively. Models for Pma_f1, Pma_f2 and Pma_f3 were created based on the crystal structure of the integrin α2 I-domain (open and ligated form in complex with the collagen-like ‘GFOGER’ tripeptide, PDB ID: 1DZI (Emsley et al., 2000); Closed form, PDB ID: 1AOX (Emsley et al., 1997). Surf2 program (Prof. Mark S. Johnson, unpublished) was implemented to study the interactions between α2 I-domain and ‘GFOGER’ tripeptide. In publication I, crystal structures of the extracellular region from integrin αIIbβ3 (PDB ID: 2VDR) and αVβ3 (PDB ID: 1JV2) were studied using Sybyl (Tripos Associates, Inc., St. Louis, MO) and Bodil. The torsion angles of amino acids (psi and phi) from the β-propeller (for αIIb and αV) domain were calculated using Sybyl.
4. Results

4.1 Conservation of the human integrin-type β-propeller domain in bacteria

Integrins have been extensively studied over the past 30 years and several crystal structures have been solved and deposited in the PDB repository. However, there was a limited amount of information on the prokaryotic origin of individual domains that constitute the integrin heterodimer. The aim of our first study (publication I) was to study the origin of constituent domains from integrin α and β subunits prior to the divergence of multicellular organisms. Certain integrin-type constituent domains can be detected in bacteria and the 7-bladed β-propeller domain (located in the α-subunit) is one such example. The available X-ray structures of integrin extracellular region were studied in order to identify characteristic structural motifs and map them on to a selected set of bacterial sequences to identify the origin of the metazoan-type 7-bladed β-propeller domains in prokaryotes.

4.1.1 X-ray structures of αVβ3 and αIIbβ3 and unique structural characteristics

During the course of this project work, ectodomain regions from the available crystal structures i.e. integrins αVβ3 (PDB ID: 1JV2, 3.10 Å resolution) and αIIbβ3 (PDB ID: 2VDL, 2.40 Å resolution) were studied. Additionally, an in-depth analysis of the remaining 16 integrin α subunit sequences was done which was helped in identifying structural characteristics that distinguish human integrin-type 7-bladed β-propeller super family from the remaining 13 super families that adopt the 7-bladed β-propeller fold. Furthermore, 32 representative structures from 13 superfamilies were also examined that lead to the identification of four structural characteristics that are uniquely specific to the human integrin-type 7-bladed β-propeller super family and these structural characteristics can be utilized in order to identify sequences that may adopt the same fold. These structural characteristics have been summarized in Figure 11 and described as follows:

i). Presence of the ‘FG-GAP’ motif or the ‘cage’ motif: The ‘FG-GAP’ motif was identified based on a combination of sequence similarities and a structural model of the integrin 7-bladed β-propeller domain. According to this model each propeller domain basically consists of seven repeating blades and each blade consists of four antiparallel β-strands wherein the ‘FG’ (Phe-Gly) pair is located on the first β-strand while the ‘GAP’ (Gly-Ala-Pro) tripeptide is located on the second β-strand (Springer 1997). The cage motif was identified as ϕϕGϕX_{13–20} PX_{2–15} GX_{5–8} (ϕ - aromatic residue; G - glycine; X - any residue; P -
proline) consensus sequence in the ectodomain X-ray crystal structure of integrin αVβ3 (Xiong et al., 2001).

ii). Type I and type II β-turns: are observed in each blade/repeat of the integrin 7-bladed β-propeller domain and they are located on the adjacent loop regions i.e. segment A (β-turn type II) and segment B (β-turn type I) in Figure 11. In depth analysis of 32 representative structures from the remaining 13 superfamilies clearly shows that at least 12 structures do not contain even a single pair of adjacent β-turns in their 7-bladed repeats while the remaining 20 structures contain at least a pair of β-turns in one of the seven blades. Therefore, apart from the integrin 7-bladed β-propeller domain there are no other known representative structures from the 13 superfamilies that contain neighbouring β-turns on adjacent loops in all the repeats of the 7-bladed β-propeller domain.

iii). An intricate H-bonding network: Another integrin specific characteristic is the presence of an intricate H-bonding (hydrogen bonding) network which occurs due to interaction among the two adjacent β-turns (type I and type II). A closer look reveals that five residues from segment A, five residues from segment B and two residues from segment C are involved in creating this elaborate H-bonding network. Furthermore, this characteristic feature is not observed in any of the other representative structures because the β-turns which are pivotal for creating this interaction network are easily located beyond the H-bonding distance.

iv). Presence of a Ca$^{2+}$ binding motif: A Ca$^{2+}$ binding motif ‘DxDxDG’ (D - aspartate; x - any residue; G - glycine) is located on the opposite end of the β-turns and ‘FG-GAP’ motif and it grants stability to the β-blade/repeat. This motif spans over two loops (loops two and four) that come together to co-ordinate the divalent cation (Ca$^{2+}$), although this motif is observed only in four or five blades out of the seven it is a distinguishing characteristic of the integrin-type 7-bladed β-propeller domain. These structural characteristics are discussed in detail below.
Figure 11. Schematic architecture of a typical β-propeller repeat (left panel), which in turn is comprised of four anti-parallel β-strands (β-strands 1-4) and four connecting loops (Loop1-4). The pivotal A, B and C segments are located on loops 1 and 3 respectively while the Ca$^{2+}$-binding ‘DxDxDG’ motif is located on loops 2 and 4. The ‘cage’ motif and the ‘FG-GAP’ motif describe the same structural motif; the key residues constituting the cage motif are localised in the segments A, B and C while the key residues that comprise the ‘FG-GAP’ motif are localised in the segments A and B. An extensive hydrogen bond network (right panel) exists between the segments A, B and C which links five residues from segment A (A0-A4), five residues from segment B (B0-B4) and two residues from segment C (C1,C2). Highly conserved glycine and proline residues at positions A3 and B2 respectively are also clearly highlighted. Figures in the left and right panel are from publication I: Chouhan et al., 2011 reprinted with permission.

4.1.2 β-turns and torsion angles

The β-turns are commonly occurring secondary structural elements that are observed in protein structures and they are classified as coils since they are non-repetitive structures. Longer repetitive structures like the α-helices and β-strands are characterised by the presence of successive residues that have similar torsion angles (Φ and Ψ angles), while non-repetitive structures like β-turn are characterised by different torsion values for each residue. Based on a theoretical conformational analysis β-turns were first described in 1968 where the available conformational freedom for a four residue system was studied which could be stabilised by a hydrogen bond between the CO group of the $n$ residue with the NH group of the $n+3$ residue (Venkatachalam, 1968). It is noteworthy that a β-turn is essentially composed of four residues and based on the torsion angles there are different types of β-
turns: type I, type II, type I', type II', type VIa1, type VIa2, type VIb, type VIII and type IV (Richardson, 1981; Hutchinson and Thornton, 1994).

Among the β-turns, type I and type II are the most commonly occurring and results from publication I have clearly indicated that segment A located on loop 1 consists of a type II β-turn while the segment B on loop 3 consists of type I β-turn. As mentioned earlier the β-turn is essentially composed of four residues and they have been annotated in publication I as A1-A4 for segment A (type II β-turn) and B1-B4 for segment B (type I β-turn). Some distinguishing characteristics of a type II β-turn are: i). torsion angles values ΦA2 = -60°, ψA2= +120°, ΦA3= +90°, ψA3=0° (A2 and A3 are residue positions in segment A); ii). Glycine occupies the position at 3 (here annotated as A3) and this position is very well conserved; iii). A hydrogen bond exists between the main chain oxygen atom (from the carbonyl group) of the residue at position A1 and the main chain nitrogen atom (from the amino group) at the position A4. Ramachandran plots were prepared (for residue positions A2 and A3 of the β-propeller blades from both the X-ray structures of αIIb and αV) to further substantiate the observations that almost all the residues at position A2 and A3 in the beta propeller blades have torsion angles that correspond to the values of a typical type II β-turn. Furthermore, conservation level of glycine at position A3 (supplementary information provided along with publication I) and the presence of a stabilizing H-bond clearly indicates the presence of a type II β-turn.

As seen in Figure 12, the segment B corresponds to a type I β-turn and the torsion angles are ΦB2 = -60°, ψB2= -30°, ΦB3= -90°, ψB3=0° (B2 and B3 are residue positions in segment B). A conserved proline residue is located at the position B2 in most of the repeats of the β-propeller domain (supplementary table provided along with publication I). Due to structural and functional constraints imposed on certain positions (like A3 and B2 in type I and type II β-turns) they are taken up by certain specific residues which also results in high level of conservation like glycine and proline (at A3 and B2 respectively). It is noteworthy that these two residues are common to both, the cage motif as well as the ‘FG-GAP’ motif. However, certain residues do not strictly adhere to the torsion angle values and display deviations, like in the case of type II β-turn residues A2 and A3 from blade 6 of integrin αV subunit are known to deviate from the standard torsion angle values. While blade I from integrin αV subunit has a non-standard conformation with respect to a type I β-turn; ΦB2 P41 = -57°, ΨB2 P41 = 163°, ΦB3 K42 = 64°, ΨB3 K42 = 32° (Figure 12).
In addition to a high level of sequence conservation across the blades of the two β-propeller domains (from human integrin αV and αIIb subunits) the number, geometry and orientation of the hydrogen bonds that join segments A, B and C, are identical in each of the seven blades from the two structures. The secondary structure elements of segments A, B and C are also identical in all seven blades of the β-propeller domains of integrins αV and αIIb.

**Figure 12.** Ramachandran plots A through D highlight the fourteen pairs of torsion angles for residues A2 and A3 from segment A; and residues B2 and B3 from segment B derived from the two integrin structures αV and αIIb. The amino acid composition along with their respective torsion angle values (Ψ and Φ) for second and third amino acid position within the segments A and B correspond to β turn types II (A2 and A3) and I (B2 and B3) respectively. Plots were also prepared for two aforementioned exceptions i.e. blade 6 residues (A2 and A3) from segment A from integrin αV subunit (C) and blade 1 from segment B from integrin αV subunit. Figure from publication I: Chouhan et al., 2011 reprinted with permission.
4.1.3 Ca$^{2+}$ binding motif (DxDDDG)

The divalent Ca$^{2+}$ binding ‘DxDDDG’ motif is located on loops 2 and 4 (‘DxDDD’ on loop 2 and ‘G’ on loop 4) but on the opposite end of ‘FG-GAP’ repeat/Cage motif and the β-turns. This Ca$^{2+}$ binding motif is not observed in all the blades of the β-propeller domain but it is present in certain repeats (3-4 repeats depending on the integrin α subunit). Interestingly, glycine from the ‘DxDDDG’ motif which is located on loop 4 coordinates the divalent Ca$^{2+}$ cation through a water molecule (which is bound to the Ca$^{2+}$ cation). It is also worth mentioning here that this ‘DxDDDG’ motif is also present in quite a few calcium binding proteins that are completely unrelated, for e.g. anthrax protective antigen and human thrombospondin (Rigden and Galperin, 2004). Furthermore, the Ca$^{2+}$ cation is tightly coordinated by residues from the calcium binding loops 2 and 4 (Figure 13). The two motifs: i.e. calcium binding motif (top phase) and the ‘FG-GAP’ motifs (bottom phase) are responsible for granting stability to the β-propeller domain as they are two key anchor points on the opposite ends of the β-blade.

![Figure 13. The calcium (Ca$^{2+}$) binding motif ‘DxDDDG’ present in certain repeats of the human integrin β-propeller domain. A strong network of ionic interactions exists between the divalent cation and the side chains of the conserved residues located between the β-strands 1 and 2; while main chain atoms from the residue located between the β-strands 3 and 4 interact with Ca$^{2+}$ through a conserved water molecule and side chain of an aspartate residue interacts directly. Figure from publication I: Chouhan et al., 2011 reprinted with permission.](image-url)
4.1.4 Bacterial sequence dataset

In order to identify bacterial sequences that could be most similar to the human β-propeller domain, the Pfam sequence motif with accession code ‘pfam01839’ was utilized and all the bacterial sequences that showed similarity to at least one copy of the structural motif (which is a characteristic of the human-type 7-bladed β-propeller domain) were extracted. The sequence motif ‘pfam01839’ in the Pfam database incorporates information on the secondary structural motif ‘FG-GAP’ as well as the Ca$^{2+}$ binding motif which led to an identification of 1093 sequences (473 eukaryotic sequences and 620 bacterial sequences). These protein sequences were then manually curated to remove any outdated or obsolete sequence records resulting in 464 eukaryotic sequences and 562 bacterial sequences. This dataset was investigated for the presence of at least seven consecutive ‘FG-GAP’/cage motif signatures using the ‘sequence search’ option which resulted in the identification of 229 sequences. Interestingly, some of the sequences displayed a presence of total 14 such signatures which indicates the presence of at least two tandem copies. Subsequently, the sequences that were shorter than the required length to incorporate all the structural elements of the motifs were removed from the dataset. This further reduced our bacterial sequence dataset down to 35 sequences from 21 different bacterial species which consist of seven full-length segments and each of the seven full-length segments carries the Pfam-defined ‘FG-GAP’/Cage consensus signature.

These 35 sequences were further examined to identify the presence of ‘FG-GAP’/Cage motif in each of the seven repeats which further reduced the dataset down to nine sequences from five different bacterial species. As seen in Figure 14, five out of nine representative bacterial sequences have been aligned with the structural alignment of human integrin αV and αIIb subunits. Furthermore, these five bacterial sequences have been listed in the materials and methods section and they were also subjected to secondary structure prediction with the help of PHD, PSIPRED and PROF prediction methods. Our results agree well with the topology and distribution of the β-strands in the available human β-propeller domain structures.
Figure 14. Sequence alignment between the β-propeller domain repeats of the human integrin αIIb and αV subunits with seven bacterial sequences obtained from five different bacterial species. Each bacterial sequence contains seven ‘FG-GAP’ repeats; secondary structural elements for αIIb and αV are highlighted in black while the same are highlighted in grey for the bacterial sequences (based on secondary structure prediction methods).

The three glycine residues as well as the proline residue (from the ‘FG-GAP’/cage motif) are very well conserved in the bacterial sequences but surprisingly the Ca$^{2+}$ binding motif is also highly conserved in each and every repeat of the bacterial sequences. These results indicate that the bacterial sequences reported in our dataset are similar to the human-type 7-bladed β-propeller domain and they do not share any similarities or features with the remaining 13 families of β-propeller fold proteins. Furthermore, our results also point towards the origin of an ancestral fold that was much more conserved but with evolution this β-propeller fold was adopted for its functions in integrins with the loss of few of its Ca$^{2+}$ binding sites.
4.2 Evolutionary origin of the αC helix in integrins

The nine human integrin α I-domain sequences were utilized to perform comprehensive searches across all the available genomes and EST libraries from organisms that appeared between the divergence of tunicates and the appearance of osteichthyes (Figure 15). Our searches covered the genomes of organisms like *Petromyzon marinus* (sea lamprey), *Callorhinchus milii* (elephant shark), *Leucoraja erinacea* (little skate), *Squalus acanthias* (dogfish shark) and *Eptatretus burgeri* (inshore hagfish).

During these searches three full length integrin α I-domain sequences from the sea lamprey genome (Pma_f1-f3) were identified, the shark/skate/ray genomic data did not yield any integrin α I-domain sequence and one short EST fragment from the hagfish genome (Ebu_f) was identified. These sequences were aligned with human integrin α1 and α11 (two out of four αC helix containing, collagen-binding human α I-domain sequences) and human integrin αM and αD (two out of five αC helix lacking human leukocyte specific α I-domain sequences) along with all the known α I-domain sequences from the tunicate *Ciona intestinalis*. The resulting dataset consisted of a total of 16 sequences (listed in materials and methods section) and the multiple sequence alignment clearly highlights regions of extensive sequence similarity spanning across the entire length of the integrin α I-domain (Figure 17). The secondary structure elements (β strands A-E, α helices C and 1-7) corresponding to the human integrin α1 I-domain (PDB ID: 1PT6) are indicated at the top of the sequence alignment. Additionally, the alignment highlights the conserved MIDAS
residues (which are highlighted as black columns) and regions of high sequence similarity (which are highlighted as grey shade boxes).

It is quite clear from the alignment that apart from the collagen-binding integrin α I-domains (α1 and α11) only the sea lamprey sequences (Pma_f1-3) contain the αC helix region which is a characteristic signature of the collagen-binding integrins while the α I-domain sequences from the leukocyte specific integrins and the tunicates lack this αC helix region. The presence of αC helix region in the lamprey sequences (Pma_f1-f3) was further substantiated by results obtained from the secondary structure prediction computer programs like Jpred (Cole et al., 2008), Gor (Sen et al., 2005), Porter (Pollastri and McLysaght, 2005), PROF (Rost and Sander, 1994) and Psipred (Mcguffin et al., 2000) which agree with the formation of an α helix corresponding to the region where the αC helix is located in the human α I-domains (Figures 16 and 17). The EST fragment (Ebu_f) from inshore hagfish is also shown which terminates just prior to the αC helix region but it does share certain features with the other α I-domains like the presence of MIDAS residues and a strong sequence identity.

<table>
<thead>
<tr>
<th>Has_al</th>
<th>QDCEDENIQRFSIAGLGSYNQGNLST EKF VEEEIKSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pma_f3</td>
<td>KDCNDGGITRFGAVLYYISSLNVKEKLQAEIRSI</td>
</tr>
<tr>
<td>Jpred</td>
<td>---------------EEEEHHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>Gor</td>
<td>cccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Porter</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
<tr>
<td>Prof_sec</td>
<td>---------------EEEEHHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>PSIPRED</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
<tr>
<td>Pma_f2</td>
<td>AAcERDNITRYAIAVLGYKYKNIIPSNFISELKAIC</td>
</tr>
<tr>
<td>Jpred</td>
<td>---------------EEEEHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>Gor</td>
<td>cccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Porter</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
<tr>
<td>Prof_sec</td>
<td>---------------HHHHHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>PSIPRED</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
<tr>
<td>Pma_f1</td>
<td>DQCEERGITRYAIAVLRSYSSNADDVARLINEVRSI</td>
</tr>
<tr>
<td>Jpred</td>
<td>---------------EEEEHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>Gor</td>
<td>ccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Porter</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
<tr>
<td>Prof_sec</td>
<td>---------------HHHHHHHHHH--------HHHHHHHHH--------</td>
</tr>
<tr>
<td>PSIPRED</td>
<td>CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
</tr>
</tbody>
</table>

Figure 16. Secondary structure predictions performed for the three lamprey sequences which are based on the five prediction programs (Jpred, GOR, Porter, Prof and PSIPRED) and these predictions are compared against human integrin α1 I-domain sequence. Figure from publication II: Chouhan et al., 2012 reprinted with permission. The αC helix region (corresponding to the human integrin α1 I-domain) is highlighted in grey.
Therefore, the αC helix definitely serves as a solid marker that can help in distinguishing collagen-binding integrin α I-domains from the remaining two clades (of I-domain) present in the chordates. At the time of this study the amount of available genomic data was quite limited in terms of quality and quantity. The next step was to investigate the lamprey sequences in more detail by conducting binding studies wherein these three sequences are expressed as recombinant fusion proteins and their binding affinities are tested against different types of collagens. Additionally, we were also interested in uncovering the phylogenetic relationship these lamprey sequences share with integrin sequences from the rest of the chordates.

4.3 Early chordate origin of the human-type integrin I-domains

As mentioned earlier searches were conducted across genomic assemblies and EST libraries from organisms that diverged between the appearance of the urochordates and osteichthytes. Apart from the three sea lamprey sequences and the hagfish fragment three very short fragments from the elephant shark survey genome (which was published and available until the end of 2013) were also obtained. Two out of three fragments shared close sequence similarity with human integrin I-domains from α1 (AAVX01128089.1; 55 residues; 76% identical) and α2 subunits (AAVX01352230.1; 55 residues; 71% identical). The third fragment showed similarity to the fifth repeat of the β-propeller domain from the human integrin α2 subunit (AAVX01625876.1; 52 residues; 63% identical).

At the beginning of year 2014 a more detailed version of the elephant shark genome was released and in-depth genomic searches were performed which yielded at least four full-length sequences (orthologous to their respective mammalian integrin counterparts): collagen-binding integrin α1, α2, α10 and leukocyte-specific integrin αE. This development provided the opportunity to create more a diverse dataset in order to perform phylogenetic analyses. Three different sets of phylogenetic trees were constructed based on the length of the sequences alignments as well as the different phylogenetic methods implemented: NJ, ML and Bayesian. Here only one set of phylogenetic trees are discussed i.e. ML trees while rest of the topologies obtained from the NJ and Bayesian methods can be observed in the supplementary information provided for publication III.
Figure 17. Sequence alignment of the integrin α I-domains and the secondary structure elements are derived from α1 I-domain; MIDAS residues are highlighted in black while the grey regions indicate similarity. Figure from publication II: Chouhan et al., 2012 reprinted with permission.
4.3.1 Phylogenetic analyses

The three sets of phylogenetic datasets were created based on the following sequence alignments (Figure 18):

a). A set of 69 chordate sequences that includes a nearly full length sea lamprey sequence (Pma_f3) and four full length elephant shark sequences (integrins α1, α2, α11 and αE).

b). A set of 72 sequences that have been trimmed down to cover the maximum common area between the three lamprey sequences (Pma_f1-f3).

c). A set of 73 sequences that span across the integrin α I-domain (~200 residues) including the sea lamprey sequences, elephant shark sequences as well as the incomplete domain fragment from hagfish (Ebu_f).

Phylogenetic trees were inferred based on: NJ method where the pairwise distances between the sequences were calculated based on the JTT matrix, ML method and Bayesian method where the WAG matrix was implemented in order to resolve the phylogenetic relationship among sequences. In addition, 3D PCA multivariate plots were prepared based on the JTT matrix in order to substantiate the observed tree topologies. Most of the clustering patterns are in agreement with previously published trees as the major observed clades are: Tunicate (or Ascidian) clade, Leukocyte specific clade and Collagen-binding clade (DeSimone and Hynes, 1988; Hughes, 1992; Fleming et al., 1993; Burke, 1999; Hughes, 2001; Hynes and Zhao, 2000; Miyazawa et al., 2001; Johnson and Tuckwell, 2003; Ewan et al., 2005; Huhtala et al., 2005; Takada et al., 2007; Johnson et al., 2009). The tunicate clade is an outlier monophyletic group that consists of integrin sequences from the vase tunicate (C. intestinalis) and sea pineapple (H. roretzi) while the other two remaining clades consist of vertebrate integrin sequences. Some important observations to arise from the phylogenetic trees are mentioned as follows:

Phylogenetic analyses based on full-length integrin sequence alignment: All three methods implicated in deciphering the phylogenetic relationships among the full length integrin sequences essentially agree with each other except in regard to the bootstrap support value between α1/α2 subunit clustering in the NJ tree which is 52% while in case of ML it is 100% and posterior probability support for Bayesian is also 100%. The nearly full-length sequence Pma_f3 from the sea lamprey genome clearly diverges prior to the integrin α10/11 cluster and in addition, the four sequences extracted from the elephant shark genome (α1, α2, α11 and αE) cluster as near outliers to their respective groups for instance it is quite
clear that they diverge prior to the osteichthyes thereby indicating the presence of vertebrate orthologues in chondrichthyes. It is also noteworthy that the bony fish sequences display the presence of isoforms (like in case of zebrafish - Dre α11A, Dre α11B; carp - Cca αL1, Cca αL2 and tilapia – Oni αM-A, Oni αM-B) but one discrepancy that still remains to be addressed is that some bony fish sequences that branch out after αE and αL clusters appear to have diverged prior to the specialization and diversification of αD, αM and αX subunits in the mammals.

Phylogenetic analyses based on aligned common sequence region: In case of largest common sequence region shared among the sea lamprey sequences all three methods essentially are in agreement with each other as the three lamprey sequences diverge prior to the α10/11 cluster in all the trees. Interestingly the ML and NJ methods produce topologies that are supported by near 100% bootstrap support. Additionally, the posterior probabilities observed at the branches in the Bayesian tree are also nearly 100% especially at the nodes located prior to the divergence of the three lamprey sequences.

Phylogenetic analyses based on alignment of the integrin α I-domain sequences: In this particular case we observed that although the quality of the sequence alignment across the ~200 residue long α I-domain region is quite good, it becomes rather difficult for the phylogenetic programs to distinguish and discriminate among the constituent sequences due to a lack of sufficient similarity differences as compared to a longer sequence alignment. Although, the representative trees here mirror the basic topology from the other trees (derived from longer sequence alignments) but the level of noise is higher and as a result more discrepancies are observed in case of the α I-domain based phylogenetic trees. But the three lamprey sequences do cluster in the collagen-binding clade albeit with certain variations along with poor bootstrap support values and this is also reflected in the 3D multivariate plots.

It is worth noticing that the short EST fragment extracted from the hagfish genome (which terminates prior to the αC helix region) branches out and clusters near the αL I-domains and this pattern is also observed in the multivariate plots. Furthermore, simple reverse BLAST searches have also revealed the αL I-domains to be the closest match for the hagfish EST fragment thereby indicating that an early homologue of the leukocyte specific integrin could exist in the agnathostomes but since the EST fragment is devoid of any additional distinguishing features.
Figure 18. Maximum likelihood phylogenetic analysis of integrin sequences based on: A) full length sequence alignment, B) largest common region sequence alignment between the three lamprey sequences and C) sequence alignment of the α I-domain region. Figure from publication III: Chouhan et al., 2014 reprinted with permission.

4.3.2 Functional residues shared between human and lamprey α I-domains

The available integrin crystal structures which are in complex with their respective ligands were studied in order to identify the functional residues pivotal for the identification of ‘GFOGER’ and ‘GLOGEN’ tri-peptides that mimic the collagen tri-peptide. This was
accomplished using the Surf2 program (Prof. Mark S. Johnson, unpublished) which was utilised to study the similarities and differences among the residues that occupy an equivalent position in other human and sea lamprey α I-domain sequences. As discussed earlier in this thesis, the α I-domain clearly provides a large exposed region for the ligands to bind which is mediated through a divalent cation like Mg\(^{2+}\) or Co\(^{2+}\) and in the case of collagen-binding integrins; the α2 and α1 I-domains recognise and bind the ‘GFOGER’ and ‘GLOGEN’ tri-peptides respectively through a glutamate (like ‘E11’ from the ‘GFOGER’ tri-peptide). A similar mechanism can also be observed in case of the αL I-domain when it is in complex with ICAM-1 D1 (PDB ID: 3TCX, Kang et al., unpublished), ICAM-3 (PDB ID:1T0P, Song et al., 2004) and ICAM-5 (PDB ID: 3BN3, Zhang et al., 2008) as the recognition and binding process takes place through a glutamate (through E34, E37 and E37 respectively).

Integrin α2 I-domain structure in complex with the ‘GFOGER’ tri-peptide (PDB ID: 1DZI) was studied with the help of SURF2 program and subsequently a table was prepared wherein all the residues from the I-domain that are located within a vicinity of 4.2 Å of the tri-peptide are displayed and comparisons were made against the other human and agnathostome α I-domain sequences (Table 5). In addition, similar tables were also prepared for α1 I-domain in complex with ‘GLOGEN’ tri-peptide and αL I-domain in complex with ICAM3 tri-peptide (not shown here, refer article III). All three tables clearly indicate that residues from the sea lamprey α I-domain sequences are more like the collagen receptor α I-domains as compared to the leukocyte specific α I-domain.

3D models were constructed for the lamprey sequences based on the α2 I-domain complex structure (PDB ID: 1DZI) to closely examine the interaction of α I-domains with the ‘GFOGER’ tri-peptide. The 3D model created for the Pma_f3 α I-domain sequence shows similarity with the α2 I-domain crystal structure as the two sequences (Pma_f3 I-domain and α2 I-domain sequences) share 44% sequence identity and there are only two deletions in the sequences alignment towards the C-terminal region. There are 16 residues from the α2 I-domain which are in the vicinity (4.2 Å) of the ‘GFOGER’ tri-peptide and two additional residues (total 18) which are a part of MIDAS constitute some of the functionally important interactions. Out of these 18 residues, 12 are identical between the Pma_f3 and the α2 I-domain and 14 residues are identical between the Pma_f3 and the α1 I-domain. One such difference can be observed in the replacement of serine with histadine, while the rest of the 2 residues (out of 3) are conserved that are pivotal for binding R12B of the ‘GFOGER’ tri-peptide to α2 I-domain.
Figure 19. The panels A and C on the left depict the α2 I-domain crystal structure which is in complex with the ‘GFOGER’ tri-peptide (PDB ID: 1DZI) while the panels B and D on the right depict the lamprey Pma_f3 sequence in complex with same tri-peptide. The α2 I-domain crystal structure and the Pma_f3 3D model were superimposed in order to highlight the relevant residue positions in the Pma_f3 model. Residue side chains from the tri-peptides are shown as CPK model while residue side chains from model as well as the crystal structure are shown as ball and stick model. Figure from publication III: Chouhan et al., 2014 reprinted with permission.

Another example is replacement for the residue equivalent to D219 in the α2 I-domain and in case of Pma_f3 it is a lysine (K219) which can potentially form electrostatic interactions with E11D (Figure 19). This interaction is particularly important as it helps in determining the collagen subtype preference and in case of human α1 I-domain and α10 I-domain an arginine is present instead of an aspartate. Pma_f1 and Pma_f2 share 9 out of 16 residues in common with human integrin α2 I-domain. But we did observed one discrepancy which remains to be tested and that is the observation that in α2 I-domain there is a threonine (T221) which is required to chelate the divalent cation at MIDAS, however in Pma_f1 there
is no conserved threonine present (even in the close vicinity) and the equivalent residue is uncertain at this point of time.

**Table 5.** Residues from the $\alpha_2$ I-domain structure (PDB ID: 1DZI) located within 4.2 Å of the bound ‘GFOGER’ tripeptide along with equivalent residues from other human, lamprey $\alpha$ I-domains and the hagfish fragment. Numbered residue indicate sequence numbering within a solved 3D protein structure; PDB ID and resolution of the structure are also mentioned. MIDAS residues (S153, S155 and T221) are indicated in italics and D151 and D254 are not listed here. ‘*’ indicates no equivalent or aligned residue; ‘?’ indicates that residue is not present in the fragment; ‘†’ indicates that alignment is uncertain at that position and ‘-’ indicates that threonine is not present in the sequence nearby. Table from publication III: Chouhan et al., 2014 reprinted with permission.

<table>
<thead>
<tr>
<th>$\alpha_1$</th>
<th>1DZI</th>
<th>S153</th>
<th>N154</th>
<th>S155</th>
<th>Y157</th>
<th>N189</th>
<th>Q215</th>
<th>G216</th>
<th>G217</th>
<th>R218</th>
<th>Q219</th>
<th>T220</th>
<th>E255</th>
<th>S256</th>
<th>H258</th>
<th>Y285</th>
<th>L286</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>1PT6</td>
<td>S152</td>
<td>N153</td>
<td>S154</td>
<td>Y156</td>
<td>E188</td>
<td>Q214</td>
<td>G216</td>
<td>G217</td>
<td>R218</td>
<td>Q219</td>
<td>T220</td>
<td>E255</td>
<td>S256</td>
<td>H258</td>
<td>Y285</td>
<td>L285</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1F1</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>Y</td>
<td>E</td>
<td>R</td>
<td>E</td>
<td>G</td>
<td>R</td>
<td>E</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>H</td>
<td>H</td>
<td>Y</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1F2</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>Y</td>
<td>E</td>
<td>Q</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>E</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>H</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1F3</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>Y</td>
<td>A</td>
<td>R</td>
<td>W</td>
<td>G</td>
<td>M</td>
<td>E</td>
<td>R</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>H</td>
<td>S</td>
</tr>
<tr>
<td>Pma_f1</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>Y</td>
<td>F</td>
<td>S</td>
<td>P</td>
<td>F</td>
<td>V</td>
<td>R</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>H</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Pma_f2</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>W</td>
<td>E</td>
<td>Q</td>
<td>G</td>
<td>G</td>
<td>K</td>
<td>V</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>S</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Ebu_af</td>
<td>S</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>S</td>
<td>Q</td>
<td>K</td>
<td>A</td>
<td>*</td>
<td>G</td>
<td>T</td>
<td>E</td>
<td>S</td>
<td>D</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>3F74</td>
<td>S139</td>
<td>M</td>
<td>140</td>
<td>S</td>
<td>141</td>
<td>Q</td>
<td>143</td>
<td>T</td>
<td>175</td>
<td>H</td>
<td>201</td>
<td>L</td>
<td>203</td>
<td>L</td>
<td>204</td>
<td>*</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1DO</td>
<td>S</td>
<td>141</td>
<td>G</td>
<td>142</td>
<td>S</td>
<td>143</td>
<td>I</td>
<td>146</td>
<td>E</td>
<td>178</td>
<td>Q</td>
<td>204</td>
<td>L</td>
<td>206</td>
<td>G</td>
<td>207</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1NY</td>
<td>S</td>
<td>140</td>
<td>G</td>
<td>141</td>
<td>S</td>
<td>142</td>
<td>S</td>
<td>144</td>
<td>N</td>
<td>176</td>
<td>Q</td>
<td>202</td>
<td>Q</td>
<td>204</td>
<td>G</td>
<td>205</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1N2</td>
<td>S</td>
<td>G</td>
<td>S</td>
<td>D</td>
<td>N</td>
<td>Q</td>
<td>K</td>
<td>G</td>
<td>*</td>
<td>L</td>
<td>T</td>
<td>Q</td>
<td>K</td>
<td>Y</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1G1</td>
<td>S</td>
<td>G</td>
<td>S</td>
<td>D</td>
<td>G</td>
<td>Q</td>
<td>G</td>
<td>S</td>
<td>*</td>
<td>V</td>
<td>T</td>
<td>G</td>
<td>I</td>
<td>F</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
4.3.3 Sea lamprey α I-domains recognize different mammalian collagen types

The sea lamprey sequences were synthesized and cloned (by our collaborators at Professor Jyrki Heino’s lab) into expression vectors pGEX-2T as recombinant GST fusion proteins in the BL21 tuner strain of *E. coli* bacteria. Despite a minor amount of detected GST the expressed proteins were pure enough to carry out experiments. A solid-phase assay was implemented in order to test the recognition and binding capacity of the sea lamprey α I-domain sequences against different types of collagens. The binding studies were conducted at a fixed concentration of the Pma α I-domains (400 nM) and the results clearly indicate that the Pma α I-domains recognize a wide variety of collagens like: rat collagen I and bovine collagen II (fibrillar collagens), mouse collagen IV (network-forming collagen), and recombinant human collagen IX (FACIT). In general, the Pma α I-domains show highest binding with the rat collagen I while Pma_f3 α I-domain exhibits the best binding with all the tested ligands (Figure 20). The lamprey α I-domains mediate their respective ligands through a metal ion but when all the Pma α I-domains were incubated with EDTA in order to test the binding levels against rat collagen I and they were clearly lower (in comparison to non EDTA collagen I binding levels) indicating the importance of MIDAS and the metal ion as well. Furthermore, the binding affinity of the three lamprey sequences was also tested against the ‘GFOGER’ tri-peptide and all three Pma α I-domains showed good binding results with Pma_f1 and Pma_f3 α I-domains exhibiting the highest binding.

In addition, comparative binding studies were also conducted against rat collagen I where the binding affinity of Pma_f3 α I-domain in comparison to human collagen receptor α2 I-domain (wild type human α2I wt and open conformation mutant human α2I E318W) was tested. Our results clearly display the lower binding levels for Pma_f3 α I-domain (in comparison to the human orthologous sequences) possibility indicating the presence of higher binding sites on rat collagen I for human α2I wt or human α2I E318W (Figure 21).

In conclusion, phylogenetic analyses, 3D comparative modelling as well as experimental testing was implemented in order to highlight the: i) first appearance of features (in the sea lamprey genome) that are a hallmark of the collagen receptor integrins, ii) the three sea lamprey sequences (Pma_f1-3) bind different types of collagens (including the mammalian rat collagen I) in a metal ion dependent manner, iii) vertebrate integrin orthologues exist even in the cartilaginous fish and iv) specialization and diversification of leukocyte integrins (αE, αL, αM, αD and αX) in the bony fish needs to be studied and addressed in more detail at this point of time.
Figure 20. Panel-A: three sea lamprey sequences Pma_f1-f3 recognize different collagen types and exhibit different levels of binding; GST as control for the lamprey sequences. Panel-B: ‘GFOGER’ tri-peptide binding to the lamprey sequences with BSA as the control. Panel-C: comparative binding of Pma_f3, wild type human α2 I-domain and human E318W mutant to rat collagen I and GFOGER tripeptide. Figure from publication III: Chouhan et al., 2014 reprinted with permission.

Figure 21. Binding affinities of the three Pma α I-domains against the rat collagen I as a function of the concentration of Pma_f al. BSA serves as the control in place of collagen I. Figure from publication III: Chouhan et al., 2014 reprinted with permission.
5. Discussion

5.1 Publication I

As discussed earlier, sequence searches performed by our own group led to the identification of different bacterial sequences that aligned surprisingly well with N-terminal domains from the integrin α and β subunits respectively (Johnson et al., 2009). This observation prompted us to study the bacterial sequences further and investigate in more detail whether these sequence were truly of the integrin-type i.e. a member of the thirteen 7-bladed β-propeller superfamilies (SCOP, Murzin et al., 1995) or a novel superfamily adopting the same fold. We performed extensive structural and sequence studies and highlighted different bacterial sequences that share similar structural and sequence characteristics with the integrins.

5.1.1 Sequence alignment and phylogenetic tree

Five sequences from five different bacterial species (of which two sequences contain an additional tandem repeat) were manually aligned against the structural alignment of the β-propeller region from the crystal structures of αVβ3 (PDB ID: 1JV2) and αIIβ3 (PDB ID: 2VDR) (Figure 14 and Supplementary Figure S1 from the publication I). This alignment clearly shows the presence of integrin specific characteristics like the conservation of key residues i.e. the three glycines and proline from the FG-GAP/cage motif and high level conservation of Ca\(^{2+}\)-binding repeat in each blade clearly suggesting that the bacterial sequences are similar to the integrin β-propeller superfamily in comparison to the other thirteen 7-bladed β-propeller superfamilies. Additionally, the conservation level of the Ca\(^{2+}\)-binding repeat in each blade suggests that the bacterial sequences may be quite similar to an ancestral fold of β-propeller domain that evolved later to serve a specific function in the integrin heterodimer thereby making the case of lateral gene transfer less likely.

In order to better understand the evolution of the 7-bladed β-propeller fold containing families, we wanted to construct phylogenetic trees wherein we would include representative sequences from species ranging from prokaryotes all the way through to humans. But, unfortunately since the families are too diverse and the representative sequences from different families lack substantial similar features it was not possible to construct a proper sequence alignment to build these phylogenetic trees.
5.1.2 3D comparative modelling

3D comparative models for the bacterial sequences of interest can be constructed by utilizing the human integrin 7-bladed $\beta$-propeller structure as the template but the models would have to be constructed very carefully owing to the key differences observed in the topology of the human $\beta$-propeller sequences in comparison to the bacterial sequences. The bacterial sequences contain more conserved features within each repeat and display shorter and more consistent loop regions. Whereas, human sequences have 3 or 4 Ca$^{2+}$ binding sites, the seven bacterial repeats each have a calcium binding site. One possible way to construct a 3D comparative model for the bacterial sequence would be to approach it in a stepwise manner i.e. model one $\beta$-repeat at a time and upon completion superpose all the constituent seven $\beta$-repeats and model the loop regions to connect them as a single 7-bladed $\beta$-propeller domain model furthermore this sort of 3D comparative model can also be utilized to study various interactions and perform some molecular dynamics studies as well. This 3D modelling study has not yet been performed but it is definitely a prospective project work at our lab.

5.1.3 Secondary structure prediction

Secondary structure predictions for the bacterial sequences of interest were carried out in a manner wherein each constituent repeat from the bacterial sequences (seven sets of seven repeats) was submitted to secondary structure prediction programs i.e. PHD from PredictProtein, PSIPRED and PROF. These methods were used in conjugation with each other to improve the accuracy of secondary structure predictions. As it can be seen in the alignment (Figure 14) the predictions align well with the structure of the $\beta$-propeller domain from integrins $\alpha$V and $\alpha$IIb. The integrin $\beta$-propeller fold is known to adopt a ‘Velcro’ fold where the last blade of the last $\beta$-repeat i.e. fourth blade of the seventh $\beta$-repeat is composed of residues located just adjacent to the first $\beta$-repeat of the first blade, thereby locking the structure together. This observation also holds true for all of the bacterial sequences, where a short $\beta$-strand was predicted just prior to the first $\beta$-repeat of the first strand. Therefore, in the absence of a 3D comparative model, the secondary structure predictions were very insightful as they helped us in studying the bacterial sequences and establishing these sequences to be of the integrin-type $\beta$-propeller domain.
5.1.4 Possible functions

At this point of time it is quite difficult to comment on the functions adopted by these β-propeller domains from the bacterial sequences. In the case of the human integrin sequences, the β-propeller domain plays a significant role in recognition and binding of ligands either directly (in conjugation with the β I-like domain) or indirectly (through the α I-domain); also, it is located towards the N-terminal region of the ectodomain of the integrin heterodimer. This is not possible in case of the bacterial sequences we retrieved from the database, since it is very likely that they are not membrane bound since they do not contain a stretch of hydrophobic residues that could indicate the presence of a transmembrane helix region and so they may adopt a functional role in the cytoplasmic region.

Another unique feature of the bacterial sequences is the presence of a well conserved Ca$^{2+}$-binding motif on all the constituent β-repeats indicating a possible role in calcium signaling (Tisa et al., 1993; Norris et al., 1996; Herbaud et al., 1998; Dominguez, 2004; Zhao et al., 2005). Through the course of evolution some of these Ca$^{2+}$-binding repeats have been lost and their number has been restricted to three or four depending upon the integrin sequence. Although integrins are not involved in calcium signaling, the three-four Ca$^{2+}$-binding motifs located on the loop regions between β-strands 1-2 and 3-4 are most probably involved in granting stability to the β-propeller structure. While the loss of Ca$^{2+}$-binding motifs on the remaining β-repeats may be attributed to acquiring flexibility in order to bind ligands or to host the α I-domain or may be for signal transduction (unlikely). But clearly, this probably could be answered by studying the bacterial and human sequences in more detail either experimentally or through molecular dynamics.

5.2 Publication II

Since the mammalian orthologues have been reported in the sequences of bony fishes and the urochordate integrins (α I-domains) do not contain the characteristic αC helix, it has been difficult to speculate on the origin of the αC containing collagen-binding α I-domains within the evolutionary framework of integrins due to physical extinctions within the early chordates coupled with absence of genomic data from the extant species resulting in a knowledge gap (Donoghue and Purnell, 2005; Huhtala et al., 2005; Jonson et al., 2009). Here we performed extensive genome searches to highlight sequences and fragments from organisms that diverged between urochordates and osteichthyes. Additionally, these sequences were subjected to secondary structure predictions in order to provide more insight into the evolution of collagen-binding integrins.
5.2.1 Genome searches

At the time of this study the available genome assemblies from two chordates *P. marinus* (sea lamprey, 5.9x coverage) and *C. milli* (elephant shark, 1.4x coverage) were downloaded and extensive local tBLASTn searches were performed using the I-domain regions from integrin α subunits. Furthermore, tBLASTn searches were also performed against incomplete genomes and ESTs from organisms like *E. burgeri* (inshore hagfish), *R. erinacea* (little skate), and *S. acanthias* (dogfish shark). The sea lamprey genome searches yielded three full length α I-domains (Pma_f1, Pma_f2 and Pma_f3) and four short fragments, one short fragment from the hagfish genome (Ebu_f) and the *C. milli* genome searches did not yield any unambiguously identifiable integrin α I-domain fragments.

5.2.2 Sequence alignment and secondary structure prediction

The sea lamprey sequences and the hagfish fragment were aligned with representative sequences from α subunits of human collagen-binding integrins (α1 and α11), leukocyte specific integrins (αM and αD) and *C. intestinalis* α I-domains (Cinα1-α8) using T-COFFEE (Notredame *et al.*, 2000). The region corresponding to the αC helix in the three lamprey sequences were subjected to secondary structure prediction using PHD from PredictProtein, PSIPRED, PROF, Jpred, GOR and Porter and these methods are in consensus with the formation of an αC helix in the lamprey sequences (Figure 16). The sequence alignment was carried out with the help of the TCOFFEE alignment program and the secondary structure elements from human integrin α1 I-domain were introduced on top of the alignment to guide and improve the quality of the alignment (Figure 17). The sequence identity levels of sea lamprey sequences with human α I-domains are shown in Table 6.
Table 6. Sequence identity of the sea lamprey sequences against the human integrin α I-domain sequences

<table>
<thead>
<tr>
<th></th>
<th>Hsa α1</th>
<th>Hsa α11</th>
<th>Hsa αM</th>
<th>Hsa αD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pma_f1</td>
<td>53%</td>
<td>51%</td>
<td>33%</td>
<td>30%</td>
</tr>
<tr>
<td>Pma_f2</td>
<td>45%</td>
<td>49%</td>
<td>25%</td>
<td>26%</td>
</tr>
<tr>
<td>Pma_f3</td>
<td>44%</td>
<td>53%</td>
<td>26%</td>
<td>28%</td>
</tr>
</tbody>
</table>

5.2.3 Importance of the αC helix

Although the αC helix serves as a distinguishing marker for the collagen-binding integrins, its presence is not absolutely necessary for integrins to bind collagens (Kamata et al., 1999; Tulla et al., 2007). This suggests that integrins could probably bind collagens much prior to their specialization as a collagen receptor group and the αC helix evolved to serve as a marker or it is possible that the evolutionary advantage of possessing an αC helix could have something to do with the conformational change taking place in the activation of α I-domain, where unwinding of the αC helix puts additional pressure on the downward shift of the α7 helix resulting in the downward transfer of signal to the β I-like domain.

Even though the sequence or the genomic data was limited it provided us with very useful insights and fortunately things began to change at the end of the year 2013 and beginning of the year 2014 as the genome assembly process for these two key genomes (elephant shark and sea lamprey genomes) were gaining momentum. The elephant shark genome was published at the beginning of 2014 (Venkatesh et al., 2014) and the sea lamprey genome (Smith et al., 2013) was also being assembled progressively thereby providing us with a very unique window of opportunity to study integrin sequences from these two genomes in greater detail.

5.3 Publication III

5.3.1 Genome searches

Two of the fragments we recovered from the C. milli genome displayed similarity to human integrin α I-domain sequence region from α1 subunit: fragment AAVX01128089.1; 55 residues; 76% identical and α2: fragment AAVX01352230.1; 55 residues; 71% identical, respectively. Another fragment showed similarity to the fifth repeat of the β-propeller domain of human integrin α2 subunit - fragment AAVX01625876.1; 52 residues; 63%
identical. Upon publication of the elephant shark genome we performed extensive searches 
and recovered at least four full-length integrin α sequences that could be potential human 
orthologues i.e., α1, α2, α11 (collagen-binding clade) and αE (leukocyte clade). 
Additionally, our searches also led to the identification of updated version of the sea 
lamprey fragments: the Pma_f1 was published with two splice variants - 
ENSPMAP00000003339, 617 amino acids; ENSPMAP0000003342, 582 amino acids, 
Pma_f2 - ENSPMAP0000008300, 478 amino acids and Pma_f3 - 
ENSPMAP0000003839, 1099 amino acids. The Pma_f3 fragment is nearly full-length and 
lacks around 120 residues towards the N-terminus region of the β-propeller domain. 
Unfortunately, we were not able to find any updates for the hagfish fragment (Ebu_f) and it 
terminates just prior to the αC helix. Also, the little skate genome was slated to get 
underway at some point of time but we could not locate any integrins sequences or ESTs 
from the skate genome. Currently, the skatebase website (http://skatebase.org) hosts data 
downloads for little skate mitochondrion and a little skate genomic contig build, which may 
be updated in the future.

5.3.2 Phylogenetic analyses and multivariate plot

As discussed earlier in the results section we prepared a total of nine phylogenetic trees; 
three sets of phylogenetic trees (ML, NJ and Bayesian based) for three sequence datasets 
based on the length of the alignment and these trees were inferred based on pairwise 
distances based on either JTT matrix or WAG matrix. In addition, we also prepared 3D 
multivariate plots (Principle Components Analysis) in order to complement the observed 
tree topologies. Some key issues observed in the clustering pattern are discussed here, for 
example placement of the hagfish fragment (Ebu_f) close to the αL cluster in the leukocyte 
integrin clade. Reverse BLAST searches using the hagfish fragment returns several αL 
integrin sequences as top hits and this observation is supported by the multivariate plots as 
they suggest placement of the fragment in vicinity to the αL cluster.

Another issue we observed during our phylogenetic analyses is the clustering pattern for the 
leukocyte fish integrins, as it seems like the fish cluster that branches out after αE and αL 
must have diverged prior to the specialization of the human αM, αD and αX cluster. This 
could imply that the fish leukocytes annotated as αM-like, αD-like and αX-like in the 
sequence database are precursors to the tetrapod leukocyte integrins. Although it clearly 
needs to be studied in greater detail, it is worth mentioning here that the genomes of the
lobe-finned fish like the coelacanth and the lungfish will be pivotal in order to gain insight into the diversification of vertebrate leukocyte integrins.

Additionally, phylogenetic trees based on an α I-domain sequence alignment have relatively lower bootstrap values compared to the full-length sequence alignment based phylogenetic trees or the common region sequence alignment based phylogenetic trees. This is due to low similarity differences across the short length of the entire domain (in contrast to the longer sequence alignments). Despite the good quality of sequence alignment behind the α I-domain region tree, there is clearly a certain level of noise (in the tree) but the topology depicts the clustering of the sea lamprey sequences close to the collagen-binding clade despite the low bootstrap scores.

### 5.3.3 Structural studies

Surf2 computer program (Prof. MS Johnson, unpublished) was used with several X-ray crystal structure complexes in order to identify, tabulate and study the key interaction residues (from α I-domains) involved in recognition and binding of ligands (within 4.2Å distance of the ligand). Furthermore, these key interaction residues were compared with equivalent residues from remaining human integrin α I-domain sequences and sea lamprey sequences (Pma_f1-f3) to highlight the similarities and differences (Refer publication III: Table 2 for α2 I-domain-GFOGER tripeptide complex, Table S2 for α1 I-domain-GLOGEN tripeptide complex and Table S3 for αL I-domain-ICAM3 complex). An interesting observation is the Pma_f1 sequence, where the conserved MIDAS threonine (T221 in case of α2 I-domain) is replaced by an arginine (from sequence pattern ‘MER’). It could be said that the glutamate located adjacent to the arginine could possibly take up the function of the threonine but it remains to be experimentally tested. Similar observations were made for the sea lamprey sequences when comparisons were done against the α1 I-domain-GLOGEN interactions (Table S2 in supplementary materials) and αL I-domain-ICAM3 interactions (Table S3 in supplementary materials). These results clearly indicated that the residues from the sea lamprey sequences are more similar to the residues from the collagen-binding integrin α I-domains as compared to the leukocyte integrin α I-domains and the lamprey sequences could potentially bind different collagen-types. In order to test this we performed 3D comparative modelling as well as experimental testing.
5.3.4 3D comparative modelling

The overall quality of a 3D model depends directly on the sequence identity shared between the target sequence with the sequence of a template structure and a good quality sequence alignment. In order to create good quality models for the sea lamprey sequences we considered the α2 I-domain X-ray crystal structure in complex with the collagen-like GFOGER tripeptide. It has a resolution of 2.1 Å and the lamprey sequences share a good level of sequence identity with the template sequence (α2 I-domain sequence). The Pma_f3 lamprey sequence shares 44% sequence similarity with the α2 I-domain sequence and deletions are found at only two position towards the C-terminal region of the domain, while Pma_f2 and Pma_f1 share 42% and 46% sequence identity with α2 I-domain sequence, respectively.

Additionally, the majority of the key functional residues about 10 out of 16 picked by Surf2 and shared between the Pma_f3 sequence and the integrin α2 I-domain sequence are identical which is higher in comparison to the other two models i.e. Pma_f2 and Pma_f1 as they share only 9 and 8 identical residues out of 16 ligand interacting residues. The absence of a threonine residue equivalent to MIDAS coordinating T221 (in α2 I-domain sequence) in Pma_f1 is an issue we were faced with again during the course of 3D modelling but it is not possible to substitute a threonine with an arginine and expect it to occupy the same 3D space as well as function. As discussed earlier the adjacent glutamate may step in to take over the function of the threonine but it remains to be tested.

In any case, the 3D models do provide us with insight into the potential interactions taking place behind the coordination between the sea lamprey sequences and the ligand GFOGER tripeptide. Additionally, the solid phase assay experimental work conducted by our collaborators confirmed the ability of the lamprey sequences (Pma α I-domain sequences) to recognize and bind different collagen-types including collagen I and collagen II (fibrillar collagens), collagen IV (network-forming collagen), collagen IX (FACIT-collagen) and GFOGER tripeptide (for technique see Tulla et al., 2008).
6. Conclusions and future directions

In publication I: we have presented evidence to highlight the origin of the integrin-type 7-bladed β-propeller domain prior to the divergence of multicellular organisms and it appears to be more regular in bacteria than the human integrin domains.

In publication II: we analysed the available genomic data (before the year 2014) from species that arose between the divergence of the urochordates and the osteichthyes in order to identify the presence of Integrin α I-domains. Here, we put forward evidence for the presence of the hallmarks of a vertebrate collagen receptor including the MIDAS motif and the αC helix in the sea lamprey Petromyzon Marinus.

In publication III: we have presented evidence to highlight the presence of orthologues of vertebrate integrin α I-domains in the agnathostomes and later diverging species. Advances in the genome assembly process for the sea lamprey genome and the elephant shark genome (at the beginning of the year 2014) helped clarify the evolutionary picture of I-domain containing integrins. Orthologues of the mammalian collagen-binding integrins extend from cartilaginous fish while the leukocyte receptor integrins need to be studied in greater detail. The sea lamprey fragments (Pma_f1, Pma_f2 and Pma_f3) that we identified share similarity with collagen-binding integrins. Additionally, experimental work from our collaborators confirmed that lamprey sequences recognize and bind mammalian collagens at MIDAS in a metal ion dependent manner. Therefore, Integrin α I-domains with vertebrate specific functions arose between the divergence of urochordates and appearance of jawless vertebrates.

Some of the possible future directions are:

3D modelling and structural studies of the β-propeller domain: the β-propeller domain from the integrin α subunit can be studied in greater detail at a structural level. The objective here would be to prepare 3D models for the bacterial 7-bladed β-propeller domain sequences and perform interaction as well as molecular dynamics studies in combination with experimental techniques to extend our understanding of the bacterial sequences. Additionally, phylogenetic studies can also be performed in greater detail to shed more light on the evolutionary pattern of the 7-bladed β-propeller domain from prokaryotes to higher vertebrates.
Evolution of leukocyte-specific integrins in vertebrates: the evolution of leukocyte integrins (αE, αL, αM, αD and αX) in vertebrates. The fish leukocyte clades are not direct mammalian orthologues and the objective here would be to perform a comprehensive phylogenetic analyses which can shed more light on the evolution and specialization of leukocyte integrins in bony fish as well as higher vertebrates.

Insight into the evolution of integrin heterodimerization pattern: in humans there are 24 integrin αβ heterodimers that are formed from the 18 α subunits and 8 β subunits, although 144 pairs are theoretically possible. The patterning of pair is nonetheless quite complex. The objective here would be to combine details from X-ray structures, observed contacts between domains, sequence variation for individual subunits over species, ligand binding and biological function which would help define the rules for the heterodimer formation with general implications for other cell-surface receptors as well.
7. References


Johnson MS, Overington JP. 1993. A structural basis for the comparison of sequences: An


Kolanus W1, Nagel W, Schiller B, Zeitlmann L, Godar S, Stockinger H, Seed B. 1996. αLβ2
integrin/LFA-1 binding to ICAM-1 induced by cytohesin-1, a cytoplasmic regulatory molecule. Cell. 86:233-242


Li Z, Zhang G, Feil R, Han J, Du X. 2005. Sequential activation of p38 and ERK pathways by cGMP-dependent protein kinase leading to


Senger DR, Perruzzi CA, Streit M, Koteliansky VE, de Fougerolles AR, Detmar M. 2002. The α1β1 and α2β1 integrins provide critical support for vascular endothelial growth factor signaling, endothelial cell migration, and tumor angiogenesis. Am J Pathol. 160:195-204.


Yang JT, Rayburn H, Hynes RO. 1995. Cell adhesion events mediated by α4 integrins are
essential in placental and cardiac development. Development. 121: 549-560.


Bhanupratap Singh Chouhan

Integrin evolution: from prokaryotes to the diversification within chordates