

CHAPTER 1

Introduction

Proteins are crucial components of life and are involved in virtually all biochemical processes. Their importance has been appreciated since the early 19th century and is reflected by their name which is derived from the Greek word for ‘of first rank’. Proteins have scaffolding functions (fibroin, keratin, lamin, and collagen) or are enzymes, catalyzing most of the vital reactions and thereby ‘encoding’ lipid and carbohydrate structures. In addition to the catalytic activities and static assemblies provided by proteins, transient, non-enzymatic interactions are also crucial for the cell. Adaptation to the cellular environment and the need to respond to extracellular signals require the cell to sense these stimuli and rapidly react in the appropriate way. The signal transduction processes depend on protein complexes, which assemble via temporary interactions². The dynamic macromolecular complexes serve as ‘computing modules’ that process and transmit input stimuli from receptors or activated enzymes to effectors ultimately generating the adequate output response such as induction of gene expression^{3,4}.

Homology searches combined with the increasing number of available three-dimensional protein structures indicated that recurring shapes are present in different proteins^{5,6,A}. These protein domains are central to the modular organization principle of many proteins. Domains are autonomous folding units of approximately 35–200 amino acids and frequently function independently of the protein context^{1,7,8}. A subclass that posses binding properties but lack enzymatic activities is defined as adaptor domains. The term adaptor domain is coined for intracellular domains and has not yet been extended to extracellular domains with similar functions. In analogy to Lego blocks⁹, adaptor domains are thought to be nature’s building units for new pathway connections in the complex signaling network of the cell^{4,5,8}. Analysis of the properties of individual (adaptor) domains will help to understand the function of the whole protein, just like translating individual words of a sentence is crucial to the understanding of its meaning. In this work, the binding specificities of GYF adaptor domains have been analyzed to ascribe potential biological functions to the GYF domain containing proteins.

^A www.mshri.on.ca/pawson/domains.html

1.1 Adaptor Domains

The first adaptor domain identified was the Src homology 2 (SH2) domain¹⁰, which binds to phosphotyrosines¹¹⁻¹⁷. Since then, the number of identified adaptor domains and their respective interaction partners has expanded considerably^{18-20,A} (Table 1.1 and Table 1.2). The importance of adaptor domains for higher metazoa is reflected by the correlation between complexity of the organism and their frequencies of occurrence (Table 1.1).

1.1.1 Binding Properties of Adaptor Domains

Adaptor domains are defined as non-enzymatic modular recognition domains which usually bind their target molecules with moderate to low affinity^{21-25,A}. The biological relevance of such weak interactions within the cell is increasingly acknowledged²⁶. Binding partners comprise the major classes of bioactive molecules such as nucleic acids, lipids, and proteins¹⁷. For protein-binding

Table 1.1: Abundance of selected adaptor domains in different species

Abundance of selected adaptor domains in yeast, different metazoan, and one plant species as identified by BLAST²⁷ searches or given in different databases and publications. The presence in bacteria or archaea is indicated by an x. Abbreviations for the species are: *At*: *Arabidopsis thaliana*, *Ce*: *Caenorhabditis elegans*, *Dm*: *Drosophila melanogaster*, *Hs*: *Homo sapiens*, *Mm*: *Mus musculus*, *Sc*: *Saccharomyces cerevisiae*. P4HBD is the substrate-binding domain of prolyl 4-hydroxylase. Genome sizes (given in mega base pairs [Mbp]) and frequencies of individual adaptor domains were obtained from the following resources: ^a Ensemble (v.33 – Sep 2005)^{28,B}, ^b Liu et al.²⁹, ^c Ball et al.¹⁷, ^d TBLASTN searches^{27,C}, ^e SMART database^{30,D}, ^f Human Protein Reference Database^{31,E}, and ^g TAIR (The Arabidopsis Information Resource)^{32,F}.

Organism		Number of Genes ^a	Genome size [Mbp]	PDZ	SH2 ^b	SH3	WW	EVH1 ^c	GYF ^d	Profilin	P4HBD ^d
Fungi	<i>Sc</i> ^e	6700	12	2	1	28	9	1	3	1	-
Metazoa	<i>Ce</i> ^a	20600	100	73	59	76	29	4	4	3	2
	<i>Dm</i> ^a	14400	133	93	35	83	37	5	3	1	1
	<i>Mm</i> ^a	27000	2268	296	120	259	77	8	3	4	2
	<i>Hs</i> ^f	24200	3272	235	120	260	83	10	3	5	2
Plantae	<i>At</i> ^g	~ 40000	125	22	3	5	22	-	14	7	-
Bacteria		-	-	x	-	x	-	-	-	x	-
Archaea		-	-	x	-	-	-	-	-	x	-

^B www.ensembl.org

^C www.ncbi.nlm.nih.gov/BLAST

^D smart.embl.de

^E www.hprd.org

^F www.arabidopsis.org

^G www.arabidopsis.org

adaptor domains, the interaction site can be provided by structural epitopes as, for example, ubiquitin. Many adaptor domains, however, recognize linear peptide sequences, lacking tertiary structure, which frequently comprise proline residues (Table 1.2), post-translationally modified amino acids (e.g. phosphotyrosines, -serines, -threonines, hydroxyprolines, acylated lysines and arginines) or the C-terminal carboxy-group^{6,17,33,A}. Recent estimates predict that there are up to 400 novel linear peptide binding motifs in the human proteome²⁰. *In vitro* binding studies confirmed that many adaptor domains interact with a multitude of peptides contained within natural proteins. Since the linear recognition signatures are usually short and comprise only few key residues, they provide an intrinsically promiscuous binding platform of moderate to low affinity. *In vivo*, additional protein-ligand contacts^{19,34-37}, co-compartmentalization, and the cooperative assembly into multiprotein complexes^{3,38,39} can reduce the ligand spectrum and strengthen the binding. For example, SH3 domains, which recognize peptide ligands with affinities ranging from 1 to 100 μM⁴⁰, bind larger fragments of ligand proteins with nanomolar affinities³⁷. Further increase in specificity can arise from negative selection of ligands against binding to competing adaptor domains from the same organism⁴¹. According to this model, evolution tailored certain ligands to exploit niches in sequence space only recognized by the

Table 1.2: Binding specificities of different proline-rich sequence recognition domains

Binding motifs of proline-rich sequence recognition domains (PRD). ^a Additional recognition motifs for SH3 domains are reviewed elsewhere^{18,39}. ^b The GYF domain of CD2BP2 was shown to interact with a sequence comprising the motif PPPGHR^{23,42,43}. However, key residues in this motif and binding sites for other GYF domains were elusive. The one-letter code for amino acids is used. Ω represents aromatic, Ψ aliphatic, Φ hydrophobic, and + positively charged residues. Phosphorylated amino acids are indicated by the prefix po. Small letters stand for residues which are less conserved in the recognition signatures and x for positions with no preference for a specific amino acid.

Domain	Recognition Motif(s)	References
SH3 ^a	+xΦPxxP	44-46
	PxΦPx+	45,46
	PxΩxxPxxP	45
	PxxDY	47
WW	PPx(Y/poY)	25,48,49
	(p/Φ)P(p,g)PPpR	25,50
	(p/Φ)PPRgpPp	25,50
	PPLPp	25,51
	(p/Ψ)PPPPP	25
	(poS/poT)P	25,52,53
EVH1	FPxΦP	22,54
	PPxxFr	55
	LPPPEP	56
Profilin	PPPPP	21,57,58
P4HBD	(PPG) ₁₀	59
	PPPP	
UEV	PTAP	60
GYF ^b	PPPGHR	23,42,43

relevant adaptor domain(s). Cross-reactivity with physiologically competitive domains is minimized, despite overlapping recognition profiles. In yeast, only the SH3 domain of the osmosensor protein, SHO1, interacts with the proline-rich motif from the kinase PBS2. However, the PBS2 ligand motif is promiscuously recognized by several SH3 domains of metazoa⁴¹.

1.1.2 Functional and Evolutionary Consequences of the Binding Properties

The mechanisms to reduce the variety of binding partners *in vivo*, as proposed above, may ensure the formation of highly specific macromolecular complexes. However, moderate selectivity of adaptor domains could be advantageous for the cell. Recognition of several binding sites with moderate to low affinity has been suggested to result in dynamic, parallel interaction networks with superior signaling properties compared to linear signal transduction pathways of the format A→B→C⁵. Furthermore, such ‘open’ networks, operating with a combinatorial, probabilistic set of interactions have been assumed to facilitate evolution: acquisition of an adaptor domain or a recognition motif during evolution renders a protein a new player in the corresponding interaction network and will subtly shift its overall equilibrium. In a stepwise manner, new signaling routes could be implemented into the network, whereas a linear signaling pathway would require coevolution of a unique binding domain and a unique binding site in the respective proteins³⁶. The modular structure of proteins and the intensive reuse of adaptor domains are in line with the first scenario. Exon shuffling is a possible mechanism for domain propagation into different proteins⁶¹. Following amplification of the genetic material⁶², intronic recombination may introduce the encoded domain(s) into a novel protein context. Archetypical domains for exon shuffling comprise EGF, kringle, and Sushi domains, predominantly found in extracellular metazoan proteins involved in processes such as blood coagulation and formation of the extracellular matrix^{63,64}. Similar evolutionary processes are likely to embody the amplification mechanisms for intracellular adaptor domains as well. The frequent lack of phase-symmetry in the encoding exons, characteristic for exon shuffling⁶³⁻⁶⁶, has been proposed to be the result of intron gain⁶⁷, loss⁶⁸, and/or slipping^{62,69}.

1.1.3 Examples for Adaptor Domain Networks

An increasing number of examples for protein-protein interaction networks, established by adaptor domains, is emerging^{2,6,39}. Assembly and regulation of the NADPH oxidase complex depends on SH3 domain interactions³⁹. PDZ domains of the *Drosophila melanogaster* protein InaD organize the association of signaling complexes around photoreceptors². In yeast, the potential network built by SH3 domains revealed Las17 to be an interaction platform for at least nine proteins^{40,46}. Similarly, in T cells, a network built by proline-rich sequence recognition domains (PRD) enlaces the cytoplasmic tail of CD2. This adhesion molecule contains five cytoplasmic proline-rich sequences (PRS)⁷⁰ which are recognized by a multitude of proteins. CD2BP2 binds to CD2 via its GYF domain^{42,43}, while CD2AP/CMS^{71,72}, CD2BP1/PSTPIP1⁷³, CD2BP3/CIN85^{72,74}, and Fyn⁷⁵ employ SH3 domains. In the latter case, recruitment is accompanied by induction of kinase activity. Binding of the SH3 domain to the PRS in CD2 alleviates an intramolecular inhibitory interaction and highlights the fact that compartmentalized low affinity interactions mediated by SH3 domains are able to trigger enzymatic activity^{2,75-78}. The aforementioned proteins, on their part, contain additional adaptor domains and recognition motifs that serve as docking sites for another layer of binding partners, further ramifying the grid of interactions⁷⁹⁻⁸¹.

1.2 Special Features of Proline

The unique features that distinguish proline from the other 19 naturally occurring amino acids render PRS preferred binding sites within the proteome. Proline is the only natural imino acid (also referred to as N-substituted natural amino acid in the text), its side-chain forms a pyrrolidine ring and restricts the dihedral angles of the backbone, giving rise to the polyproline type II (PPII) helix as the favored secondary structure of PRS⁸². The left-handed PPII helix adopts backbone angles of $\Phi = -78^\circ$ and $\Psi = +146^\circ$ ⁸³, resulting in a perfect three fold rotational symmetry and a periodicity of three (Fig. 1.1a). In addition, PPII helices are pseudo-symmetric and have a triangular, prism-like shape, when viewed along the helical axis¹⁹ (Fig. 1.1b). Solvent exposure of electron-rich backbone carbonyl oxygens makes PPII helices good hydrogen bond (H-bond) acceptors whereas the accessible side-chains of prolines form a continuous, highly distinguishable hydrophobic surface stretch. The rigid conformation of PPII helices reduces the entropic cost upon binding⁸⁴.

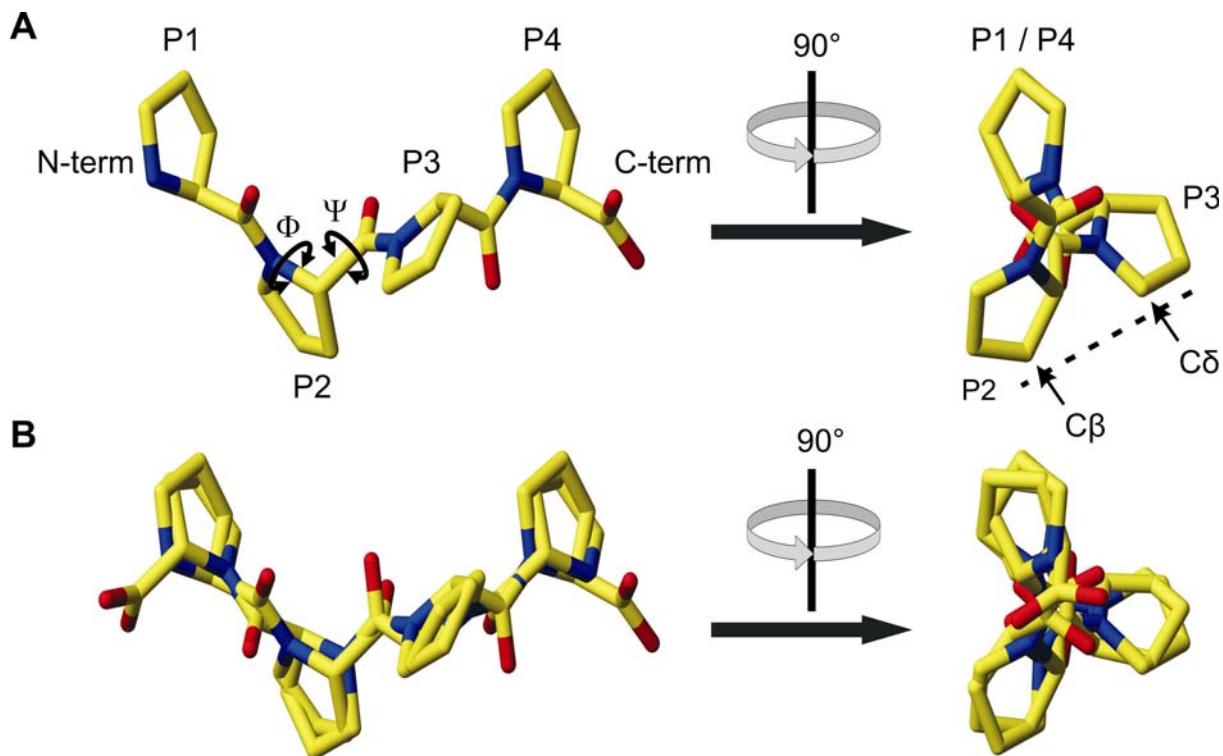


Fig. 1.1: Structure of the polyproline type II helix

(A) Polyproline type II helix conformation of a tetraproline peptide. The peptide is shown from the side in an N→C orientation (left) and from the front (right). Due to the three fold rotational symmetry, every first and fourth proline residue (P1 and P4) have identical orientations and superimpose exactly. The dotted line indicates the site of the xP dipeptide (here PP) in proline-rich ligands which interacts with the binding pocket of SH3, WW, and GYF domains (see Chapter 1.3.2). Formation of the pyrrolidine ring by covalent bonding of the delta carbon and the backbone nitrogen atom results in the substituted amide group of prolines and places the delta methylene group ($C\delta$) into a unique orientation. The arrangement of a beta methylene group ($C\beta$) and $C\delta$ of two consecutive residues x and proline, respectively, makes the site highly distinctive albeit tolerant to amino acid exchanges which preserve a $C\beta$ in the first position. (B) Overlay of two tetraproline peptides, the first one with identical orientation as shown in (A) and the second one rotated by 180° around the y-axis. The high degree to which the N→C and C→N oriented peptides overlap highlights the pseudo-symmetry of the helix. Carbon atoms are colored yellow, oxygen atoms red, and nitrogen atoms blue. Hydrogen atoms are omitted.

1.3 Proline-Rich Sequence Recognition Domains

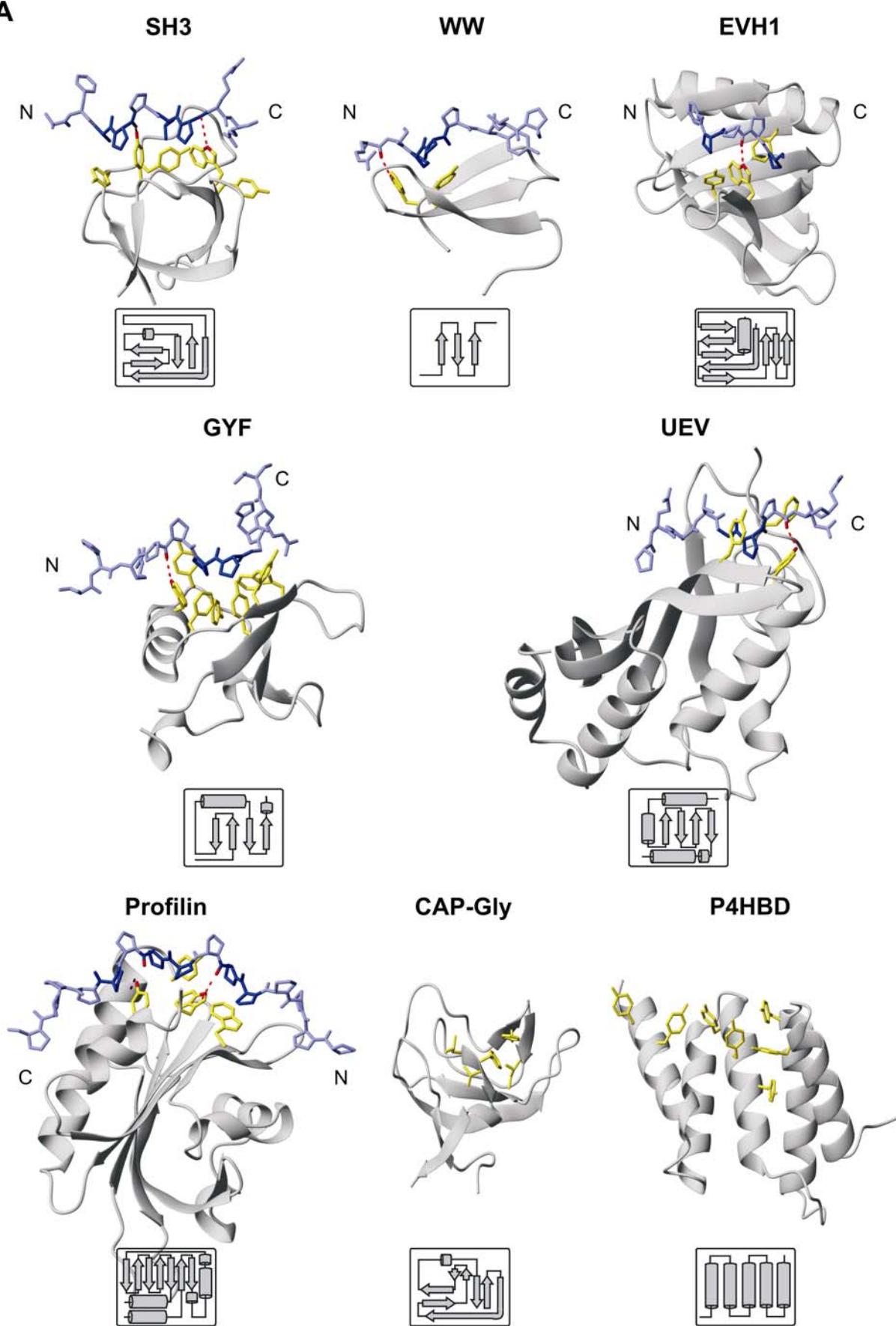
1.3.1 Association of Proline-Rich Sequence Recognition Domains with Human Diseases

PRS are among the most common peptide sequence motifs, as revealed by the analysis of different eukaryotic genomes^{85,G}. Their favorable binding properties are exploited by a limited number of PRD. Today, the superfamily of PRD comprises SH3^{86,87}, WW⁸⁸, Ena/VASP homology 1 (EVH1)⁵⁴, GYF^{42,43}, and ubiquitin E2 variant (UEV) domains^{60,89} as well as profilin⁵⁷. Further candidates are the substrate binding domain of prolyl 4-hydroxylase (P4HBD)^{59,90} and the CAP-Gly (cytoskeleton-associated protein, Gly-rich) domain, that displays an SH3-like fold⁹¹. The involvement of PRD in fundamental developmental and regulatory processes is evidenced by their implication in human diseases. WW domain-mediated interactions have been associated with cancer^{92,93} and disorders, such as Liddle's⁹⁴ and Rett's syndrome⁹⁵, Duchenne or Becker muscular dystrophy⁹⁶ as well as Alzheimer's^{93,97,98} and Huntington's diseases^{99,100}. Mutations in the SH3 domain of human nephrocystin causes juvenile nephronophthisis, an autosomal recessive, inherited kidney disease¹⁰¹⁻¹⁰³ while defects in the EVH1 domain of the WASP protein give rise to the Wiskott-Aldrich syndrome^{56,104}.

1.3.2 Binding Modes of Proline-Rich Sequence Recognition Domains

Convergent evolution has shaped the binding site of PRD similarly for optimal recognition of central prolines within the ligand (Table 1.2). Unrelated in their folds, common recognition features could be observed by structural comparison (Fig. 1.2a). All PRD identified so far share the use of stacked aromatic amino acids, frequently referred to as aromatic cradle, to exploit the characteristic features of prolines within the PPII helical conformation for binding^{19,105}. The exposed aromatic residues account for the hydrophobic nature of the binding pocket(s) and allow coplanar stacking with the proline side-chains. A conserved tryptophan (tyrosine in the case of the UEV domain) also forms an H-bond to a carbonyl oxygen of the ligand backbone^{23,58,60,106-111} (Fig. 1.2a). Differences in the shapes of the binding epitopes and the arrangement of the proline-rich ligands in the complexes led to the definition of two recognition principles and three binding models¹⁹ (Fig. 1.2b, c, and d).

^G www.ebi.ac.uk/interpro

A

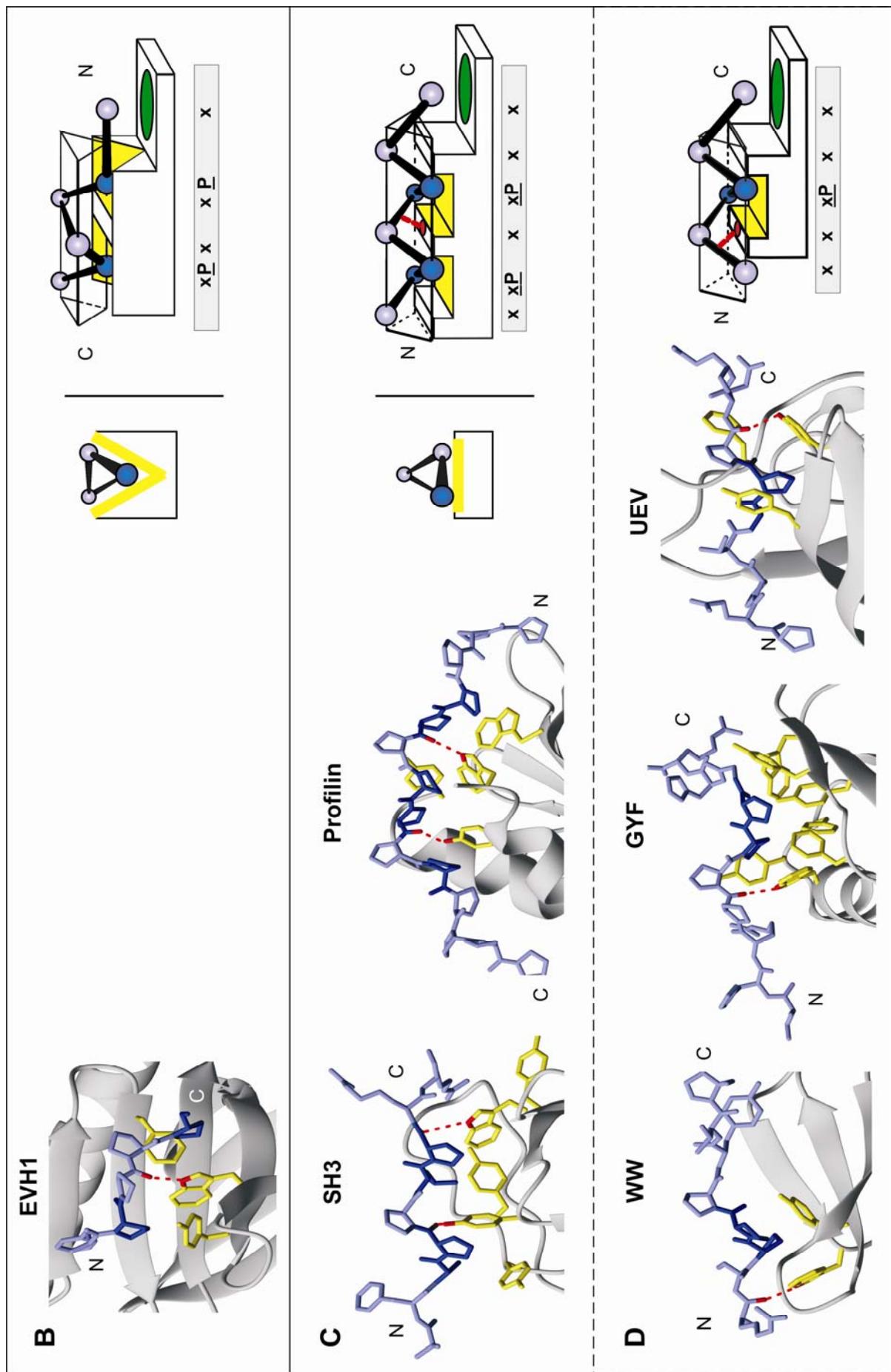


Fig. 1.2: Structures and binding mechanisms of proline-rich sequence recognition domains

(A) Structures of different PRD in complex with proline-rich ligands. For SH3, WW, EVH1, and profilin folds, only one of the two opposite ligand binding orientations is exemplified. Aromatic residues which form part of the hydrophobic binding site are depicted in yellow. Residues of the ligands are colored in shades of blue with dark blue highlighting residues that interact with the hydrophobic pocket(s). The H-bonds between backbone carbonyl oxygens of the ligands and conserved aromatic residues (tryptophans and tyrosines) of the domains are represented as red, dotted lines. The topologies of the different PRD are shown below the structures. Arrows stand for β -strands, cylinders for helices. Protein Data Base (PDB) accession codes for presented structures are: 1PRM (SH3), 1EG4 (WW), 1EVH (EVH1), 1L2Z (GYF), 1M4P (UEV), 1CJF (profilin), 1IXD (CAP-Gly), and 1TJC (P4HBD). For the latter two, no complex structures are available yet. (B, C, and D) Binding sites of domains shown in (A). The binding pockets of PRD in complex with a proline-rich ligand are enlarged and arranged according to their binding mechanisms. Cartoons on the right delineate the binding mechanism of the PRD shown on the left. The first cartoons depict the view from the specificity patch along the ligand axis (B and C), the second depict side views (B, C, and D). (B) Binding site of Mena-EVH1 in complex with the peptide from ActA. EVH1 domains accommodate the pointed site of ligands in prism-like shaped PPII helical conformation. (C) Binding sites of SH3 domains and profilins. These PRD comprise two hydrophobic pockets and accommodate ligands with the motif xPxP. P4HBD and CAP-Gly are anticipated to comprise two hydrophobic binding pockets as well (not depicted). (D) WW, GYF, and UEV domains employ a single hydrophobic pocket to recognize xP dipeptides. The hydrophobic binding grooves are formed by aromatic residues depicted in (A–D) in yellow and are colored correspondingly in the models. Residues of the schematic ligands are shown as spheres and colored as in (A). Specificity patches are presented in green, the H-bonds to conserved aromatic residues as red, dotted lines.

Recognition of the Pointed End of PPII Helices—EVH1 Domains

EVH1 domains use a concave, V-shaped binding surface to accommodate the apex of the triangular, prism-like shaped PPII helix (Fig. 1.2b). The underlying recognition principle largely focuses on the conformational properties of the ligand (PPII helix with two hydrophobic proline side-chains in defined positions, see Chapter 1.2) rather than the specific identity of residues arrayed along the ligand peptide chain¹¹⁰. Four different families of EVH1 domains are distinguished based on the protein organization, the biological functions, their structures, and their binding specificities (Table 1.2). Ena/VASP is representative of the first subfamily and recognizes the signature, FPx Φ P²², where x denotes any residue, Φ represents hydrophobic residues, and the other letters correspond to the one-letter code for amino acids. Other non-natural imino acids preserve the binding capacity of the ligand¹¹², suggesting that aliphatic N-substitution, in the context of a PPII helix, is a hallmark for binding. The binding sites of the other three EVH1 subfamilies, typified by WASP, Homer/Vesl, and Spred, lack one of the three aromatic residues (Tyr 16 in the Ena/VASP family member Mena) that are involved in ligand binding in Ena/VASP EVH1 domains¹⁷. In WASP and Homer/Vesl subfamilies, an aromatic residue at position 14 compensates for the loss of aromaticity at position 16. Similar to Ena/VASP, the EVH1 domain of WASP interacts with two residues, separated by a PPII helical turn. Ligand binding, however, occurs in the opposite orientation and with differently shaped pockets (comprising only two of the three aromatic residues). Therefore, the domain selects for

an LPxP motif within a longer ligand which wraps around the domain and contacts a second binding site⁵⁶. Homer/Vesl EVH1 domains have a single proline binding pocket and recognize the signature, PPxxFr, with the two consecutive prolines in PPII helical conformation^{113,114}, while the binding properties for the Spred subfamily remain elusive.

Recognition of xP Dipeptides—SH3, WW, GYF, and UEV Domains

The second recognition principle is based mainly on the amide N-substitution of proline residues^{19,115}. The building unit of the binding sites is the so called xP pocket^{19,115} (Fig. 1.2c, d), a shallow, hydrophobic groove partially defined by the conserved aromatic residues. The pockets simultaneously contact the side-chain C β and C δ methylene groups of two consecutive residues x and P, respectively, on one face of the PPII helix (Fig. 1.1). The N-substituted proline, preceded by a C α -substituted residue together form a highly discriminatory recognition site for xP dipeptides without conferring high binding affinity¹¹⁵. Coplanar stacking between the proline pyrrolidine ring and an aromatic residue¹⁰⁵ in the pocket may further contribute to ligand recognition.

According to the number of xP pockets in SH3, WW, GYF, and UEV domains as well as profilin, two binding models were distinguished. WW, GYF, and UEV domains comprise a single xP pocket^{19,105}. The particular shape of their xP pockets allow WW and GYF domains to preferentially recognize two consecutive prolines (Table 1.2, Fig. 1.2d, Chapters 6, 8, and 9). SH3 domains and profilins employ two xP pockets to accommodate two xP dipeptide motifs on the same face of the PPII helix, separated by a single turn (one residue between two xP motifs; Fig. 1.2c). This arrangement dictates the recognition signature xPxxP. P4HBD and the CAP-Gly domain may reveal a similar binding mode. Both interact with PRS via exposed tyrosine residues, forming shallow pockets that could possibly accommodate two xP dipeptides^{90,91,116}, but the structures of the domain–ligand complexes have not been solved yet.

The pseudo-symmetry of the PPII helix (Fig. 1.1) suggests that ligand binding in N→C and C→N orientation is possible, using the same domain interface. SH3, WW, EVH1, and profilin folds have been found to support both orientations of binding but individual domains within these families generally show distinct preferences for one of the two orientations¹⁰⁵. Domain regions that flank the aromatic cradle recognize non-proline residues, thereby introducing specificity for different PRS, restricting the orientation of binding, and defining the register for ligands comprising longer proline stretches^{19,21,105} (Fig. 1.2).

1.4 Sequence Alignment of GYF Domains

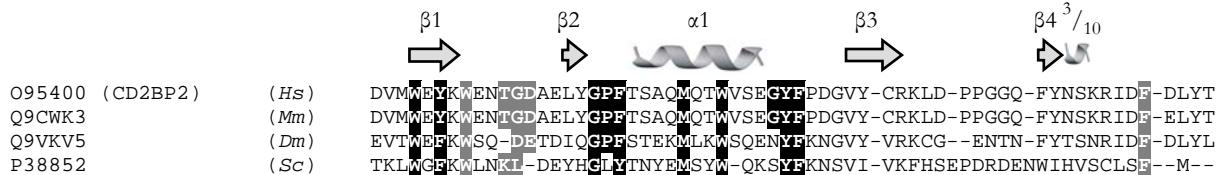
CD2BP2 was identified in a yeast two-hybrid screen as a binding partner of the T cell adhesion protein, CD2. Further analysis revealed that a C-terminal fragment of CD2BP2 and two membrane-proximal PPPPGHR motifs of the CD2 cytoplasmic tail are solely responsible for the interaction between the two proteins. NMR studies showed that the last 62 amino acids of CD2BP2 fold into a compact domain, the GYF domain^{42,43}. Database searches with the GYF domain of CD2BP2 and sequence alignments highlighted the amino acid signature W-x-Y-x₆₋₁₁-GP[F, Y]-x₄-[M, I, L, V]-x₂-W-x₃-GYF as a characteristic feature of GYF domains^{42,43} (Fig. 1.3). The functional importance of some of these residues was confirmed by an alanine-scan in conjunction with yeast two-hybrid analysis⁴².

The highly conserved residues Trp 4, Tyr 6, Gly 18, Pro 19, Phe 20, Met 25, Trp 28, Gly 32, Tyr 33, and Phe 34 (CD2BP2-GYF numbering; Fig. 1.3a), together with Trp 8, Tyr 17, and Phe 58 constitute the hydrophobic core and the ligand binding site of the domain (Fig. 1.4). Conservation of Trp 8 and Phe 58 is typical for a subgroup of GYF domains. Trp 8 is located at the beginning of an extended loop between β-strands β₁ and β₂ in these domains and thereby defines the CD2BP2 subfamily of GYF domains (Fig. 1.3a). Additional characteristics of CD2BP2-type GYF domains are their strict localization to the very C-terminus and their absence from plant proteins.

The majority of GYF domains are localized mostly in the center of the respective proteins and share a shorter loop between strands β₁ and β₂. Furthermore, they predominantly contain aspartate at position 8 instead of tryptophan and lack phenylalanine at position 58 (Fig. 1.3b).

Fig. 1.3: Sequence alignments of GYF domains and related folds (on the right)

GYF domains of the CD2BP2 subfamily (A) and the SMY2 subfamily (B), identified by BLAST searches. (C) Putative GYF domains according to SUPERFAMILY. These sequences form a hypothetical third subfamily. The postulated GYF domains are localized within diglyceride acyltransferase (DAGAT) domains. Yeast protein, Q08650, does not comprise a classified GYF domain but contains a homologous DAGAT domain. The fragment corresponding to putative GYF domains in mammalian DAGAT domains is shown. () indicates an insertion. (D) Sequences derived from the DALI database and from a DALI search using the CD2BP2-GYF domain as query structure. The alignment is structure-based. Structures of depicted sequences share a Z-score > 2 and a RMSD < 3 Å and are listed according to decreasing Z-score values. The sequences of CD2BP2-GYF and the best hit, Q9FT92, a SMY2-type GYF domain, are included. Swiss-Prot, TrEMBL, and EnsEMBL entry names and the origins of the proteins are indicated. At, Ce, Dm, Hs, Mm, Os, and Sc stand for species *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Oriza sativa*, and *Saccharomyces cerevisiae*, respectively. In (D) the PDB entry names of the structures of aligned sequences are shown in brackets. Conserved amino acids that are characteristic for GYF domains and residues with similar physico-chemical properties in these positions are depicted as white bold letters on black background. Other amino acids in these positions are represented as bold letters in (D). White bold letters on grey background represent residues or positions specific for subfamilies.

A CD2BP2 subfamily of GYF domains**B SMY2 subfamily of GYF domains**

SDG2_1	(<i>At</i>)	LGKWFYLDYY---GTEHGPARLSDIKALMEQGILFSDHM-IKHSDDNNRWLNVNPPEAPGNLLEDIAD
SDG2_2	(<i>At</i>)	IGDWFYTDGA---GQEQQGPLSFSFQQLKLVKEKGFIKSHSS-VFRKSDKIWWVPTISITKSPETIAMLR
Q9FIH7_2	(<i>At</i>)	HACWFVLVDGE---GRNHGPHSILEFLSWQQHGYVSDAL-IRDGENKLRLPITLASLIGVWRVKCGD
048697	(<i>At</i>)	DFLFLYIDPQ---GVIQGPFIGSDIIISWFEQGFFGTDLQ-VRLANAPE---GTPFQDLGRVMSYL
Q9F7J2	(<i>At</i>)	EFLFLYIDPQ---GVIQGPFIGSDIIISWFEQGFFGTDLQ-VRLASAPE---GTPFQDLGRVMSYL
Q9FMM3 (GYN4)	(<i>At</i>)	ELSLIYYKDQ---GLIQGPFGSGSDIIIGWFEAGYFGIDL-LVRLASAPN---DSPFSLLGDVMPHL
Q02875 (SYH1)	(<i>Sc</i>)	ESQWKYIDS---GNIQGPFGTNMMSQWYQGGYFTPTLQICRLATSPPEPGVNDRFLIRLGELTTLV
P32909 (SMY2)	(<i>Sc</i>)	ESSWRYIDTQ---GQIHGPFTTQMMWSQWYIGGYFASTLQISRLGSTPETLGINDIFITLGELMTKL
075137 (PERQ2)	(<i>Hs</i>)	MQKWWYKDPQ---GEIQGPFFNNQEMAEEWFQAGYFTMSLL-VKRCACDESFPQLGDIMKMWGRVPFSP
Q8C585 (GIGYF2)	(<i>Mm</i>)	MQKWWYKDPQ---GEIQGPFFNNQEMAEEWFQAGYFTMSLL-VKRCACDESFPQLGDIMKMWGRVPFSP
075420 (PERQ1)	(<i>Hs</i>)	ARKWFYKDPQ---GEIQGPFTTQEMAEWFQAGYFTMSLL-VKRCGDEGFQPLGEVIKMWGRVPFAP
Q99MR1 (GIGYF1)	(<i>Mm</i>)	ARKWFYKDPQ---GEIQGPFTTQEMAEWFQAGYFTMSLL-VKRCGDEGFQPLGEVIKMWGRVPFAP
Q9V482	(<i>Dm</i>)	NELWFYRDPQ---ANVQGPFSAVEMTEWYRAGYENENL-F-VRRYSDNRFRLGELIKFCCHGNMFT
Q09482	(<i>Ce</i>)	PVQFYYYMDPT---ETRRGPFPKQDMNWFKAGYFTIDESLRVQRGENGEYKTIGDLKKLHGSSTPFE
Q9XVJ6	(<i>Ce</i>)	DTKWHYLPD---SEKYGPYMSKDMFLWQLAGYFNNDGLQ-LKTENEPMNYHTLGEWSQLLGTHPFSM
Q9FHL0	(<i>At</i>)	KVMWFYEYPK---GKTHGPFSLTDLKWTWSDEEYFVGVPD-FKVWKTGESAVLLTKLLSQIKT---
Q9LP87	(<i>At</i>)	QLLLFYIEPT---GRTHRPFSLTSMRKWWDQGYFDGFPN-LKFLPSLNF-----
CL026838	(<i>Os</i>)	EKWHYKDP---GNVQGPFTLVLQSKW---TSYFPRDRM-VWLTFESEERSLLLTEVLSKQPKDFG
Q9FW12	(<i>Os</i>)	EKWHYKDP---GSVQGPFTLQLSKW---AAYFPHQLV-LMMSGPLEATLVNGTIVRIVVGEHSHW
CL036702	(<i>Os</i>)	EKIWQYMDPT---GKIQGPFSIVQIRKWNNGS GYF PPNLK-IWKSTEKQDDSIILTALLGRFEKDL
Q9SIV5	(<i>At</i>)	EKIWHYKDP---GKVQGPFSMAQIRKWNNT GYF PAKLE-IWKANESPQDSVLTLDALAGLFQKQT
Q9SD34	(<i>At</i>)	SEIWHYRDP---GKTQGPFSMVQIRRWKFSGHFPPYLR-IWRAHENQDESVLTLDALAGRFDKAT
Q9SL38	(<i>At</i>)	NMVWLYGDP---GKIHGPFSLYNIRQWNSSGHEPEPEL-IWRLGEQHQHSSILLTDALNGQFHKTG
CL009775	(<i>Os</i>)	ASVWYYNDPQ---GDEQGPFLPLTIRHWSKAGYFEDFR-VWRTGQSCDSAIIKKDALLLTS---
Q9FT92	(<i>At</i>)	KLNWLYKDP---GLVQGPFSLTQKAWSDAEYFTKQFR-VWMTGESMESAIIKKDALLLTS---
Q9LF02	(<i>At</i>)	DVGWYIILGEN---QCNLQGPYTFSETCNHFRNGYLLETTL-WVADGRSEWQPLSAIPDLMRSRISGAE
Q966F5	(<i>Ce</i>)	ELEIIFYIDDE---DNVQGPYGAHKVLGWYRNGHFFHDDHQ-FKIVDCAQHGEVLTVEAYLGDILSRF
P34520	(<i>Ce</i>)	DITVFYTDDE---GTVQGPYGASTVLDWYQKGYFSDNHQ-MRFTDNGQRIGNLFTYETTLGEMKAR
Q9FIH7_1	(<i>At</i>)	ASGWMYGNQQ---GQMC GP TQQQLYDGLSTNFPLEDLLVYPIINGYTANSVPLKYFKQFPDHVAT
Q9VKV2	(<i>Dm</i>)	KQTWPNK---EDFFPYSSDSHSYW---T GYF TSRPTQKRFHRDGHNFFQTVKQLSVLANLSGT

C Putative GYF domains (third subfamily)

ENSMUSP00000033572	(<i>Mm</i>)	--VWIAYDWN---THIQDGRRSAWRNWTLW KYF QSYFP-VKILKTKDLSPESENYIMGVPHGILT
ENSP00000198801	(<i>Hs</i>)	-AAWWYLDR---KPROGGRHIQAIRWWTI KYF KMDYFP- I SLVKTAELDPSRNYIAGFPHGVLA
ENSP00000264412	(<i>HS</i>)	-LMWLYFDW---TPERGGRSSWIKWTW KYF KDYFP-IHLIKTQDLDPSHNYIFGFPHGVIMA
ENSMUSP00000064041	(<i>Mm</i>)	-ATWWYLDWD---KPRQGGRPIQFERRLAIW KYF KMDYFP-SVLVKTAELDPSRNYIAGFPHGVLA
ENSMUSP00000012331	(<i>Mm</i>)	-LWFWYWDWR---TPEQGGRRNWVQSWP VW KYFKEYFP-ICLVKTQDLDPGHNYIFGFPHGVIFV
ENSMUSP00000033001	(<i>Mm</i>)	-FTWLAFDW---TPKKGGRRSQWV RN AVWRYF RD YFP-IQLVKTHNLLTRNYIFGYPHGIMG
ENSP00000228027	(<i>HS</i>)	-FTWLFVFDW---TPKKGGRRSQWV RN AVWRYF RD YFP-IQLVKTHNLLTRNYIFGYPHGIMG
ENSMUSP00000036845	(<i>Mm</i>)	--VWIAYDWN---THIQDGRRSAWRNWTLW KYF QSYFP-VKLVKTHDLSPKHNYIILSHPHGILS
ENSP00000328036	(<i>HS</i>)	--WLYFDW---THSQGGRSSA W RVNWTLW KYF NRNFP-VKLVKTHDLSPKHNYIIANHPHGILS
ENSP00000223114	(<i>HS</i>)	-LWFWYVWDW---TPNQGGRSEW RN RRAW RN WARRHVAW KYF YAFSYFP-LRLIKTADLPDRNYV L GAPHGMIC
W01A11.2	(<i>Ce</i>)	-AVWFWYDFD---TPKKASRRWNWARHVAW KYF YAFSYFP-LRLIKTADLPDRNYV L GAPHGMFS
F59A1.10	(<i>Ce</i>)	-AWWYLYDRE---SPRRGGYRD N W R SLH K W A YFP-VKLHKTAE D LPQN Y LF G YPH H ILG
Y53G8.B.2	(<i>Ce</i>)	-ACWYFYDMD---SPRRGGYSSD W R V R N D W FA Q YFP-INLHKTAE D LPQN Y LF G YPH H ILG
Q08650	(<i>Sc</i>)	YMI Y FFFDRSPA-TGEVVNRYSLR R STPI W KW C DYFP-ISLIKTVNLKPT()YLF G YPH H GIGA

D GYF domain related folds (DALI)

CD2BP2	(1GYF)	DVMWEYKWE--NTGD--AELY- GP ---FTSA-QMOTWVSE GYF --PDG-VYCRKLDPPGGQ-FYNSKRIDE-DFDLYT
Q9FT92	(1WH2)	KLNWLYKDP---Q--GLVQ- GP ---FSLT-QKAWSDAE YF --TKQ-FRVWMT-GESMESAIIKKDALLLTS
P06786	(1BGW)	TPIIKVSIT--K---PTKNTI-- A ---FYNMPDMEK W REE- ES --HKFTWKQKYYKG
CAA46264	(1K25)	ATSYN V YAV----ISFGSKG- NG ----ITYA-NMMAIKKE E TAEV KG -IDFTTSPN
Q9UX16	(1TLJ)	SGRITIVDAEMPWDRKNSTII- F KNHLR I TEQ-D E DVL-SKN Q --VRR-LW L IV--
Q08288	(1WJV)	MVF F TCNA-----C-GESV----KK-I-Q VE KHV-SNC R --N-C-ECLSCIDCG--K-DFWGDDY--K-SHV K

Therefore, these GYF domains constitute a second subgroup termed SMY2 subfamily of GYF domains, referring to its most prominent member, suppressor of myo2-66 (SMY2; Chapters 1.5.3 and 9).

A third subfamily of GYF domains can be postulated on the basis of sequence alignments provided by the SUPERFAMILY server^{117,H} (Fig. 1.3c), using the GYF superfamily definition of the Structural Classification of Proteins (SCOP) database^{118,I}. SUPERFAMILY sequence alignments are based on a hidden Markov model¹¹⁹ to extend the list of potential members of protein domain families. The suggested third subfamily of GYF domains is highly conserved and not present in plant proteins. Similar to the SMY2 subfamily, the domains of this group contain aspartate at position 8 and have a shorter loop between strands β_1 and β_2 , but the otherwise conserved triplets, especially the first one (GPF), are largely missing. All these putative GYF domains are localized within metazoan diglyceride acyltransferase (DAGAT) domains, which have also been identified in plants, fungi, and bacteria. The DAGAT domain of the yeast protein, Q08650, aligns well with the corresponding region in metazoan DAGAT domains which form the hypothetical third subfamily of GYF domains (Fig. 1.3c). The existence of non-GYF domain containing, homologous DAGAT domains in other species stresses the hypothetical character of the postulated GYF domains. Furthermore, for this hypothetical subfamily, no structures are available to show these protein fragments to adopt the typical GYF fold.

1.5 Structure of GYF Domains

1.5.1 Structure of CD2BP2-GYF

The structure of the CD2BP2-GYF domain was solved by NMR spectroscopy⁴³ (Fig. 1.4a) and recently the crystal structure of CD2BP2-GYF in complex with the protein U5-15K has been deposited in the Protein Data Base^J (PDB; accession code: 1SYX¹²⁰; Fig. 1.4b). Both structures are highly similar and associate the conserved amino acid signature with a characteristic helix-bulge motif, whereby the two amino acid triplets GP[F, Y] and GYF flank the helix. The second triplet forms the bulge and lent the domain its name due to its particular structure, its conservation, and its implication in ligand binding.

^H supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

^I scop.mrc-lmb.cam.ac.uk/scop/

^J www.rcsb.org/pdb

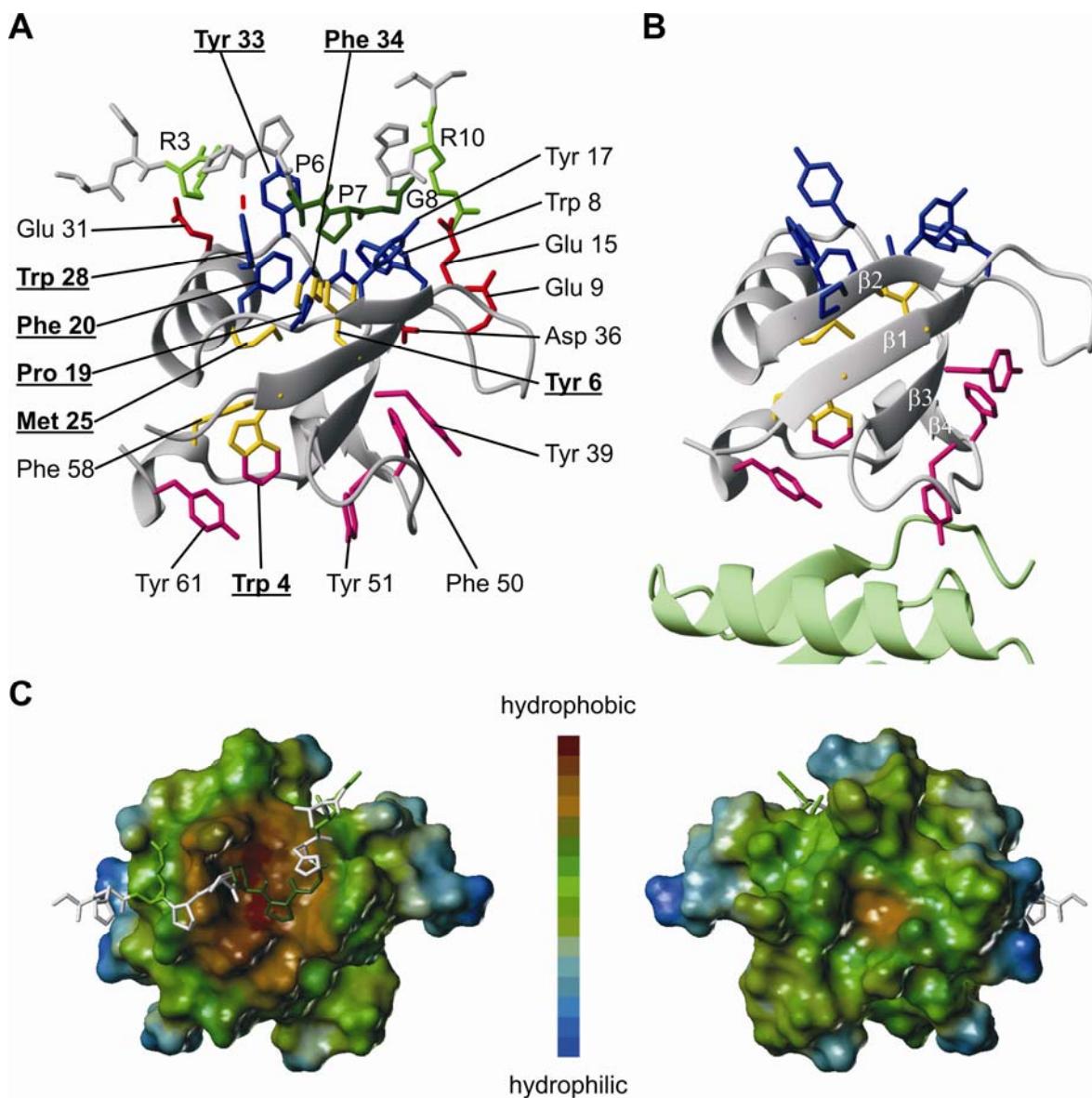


Fig. 1.4: Structure of the CD2BP2-GYF domain

(A) Structure of CD2BP2-GYF in complex with the CD2 peptide SHRPPPGHRV. Side-chains of conserved residues are labeled in bold, underlined letters. Residues are color coded according to their function. The side-chains depicted in blue form the aromatic PRS binding pocket and the CD2 peptide residues P6–G8, shown in dark green, are in closest contact. Negatively charged residues of the domain and the interacting, positively charged ligand residues are represented in red and green, respectively. Yellow side-chains take part in the hydrophobic core of the domain and aromatic residues in pink are partly involved in binding to U5-15K. Partial coloring indicates multiple functions. The H-bond between P4 carbonyl oxygen and Trp 28 side-chain amine hydrogen is shown as red, dotted line. (B) Structure of CD2BP2-GYF in complex with the protein U5-15K. Coloring of CD2BP2-GYF residues is identical to (A). β1–4 indicate the strands of the sheet which folds against the helix. A portion of U5-15K, comprising the CD2BP2 binding site is shown in green. (C) Lipophilic surface potential of CD2BP2-GYF. The binding surface for CD2 peptide (left) and U5-15K (right) are depicted, colored according to the lipophilic potential. Hydrophobicity scaling is from brown (most hydrophobic) to blue (hydrophilic). The structures are related by a 180° rotation around a vertical axis and comprise the CD2 ligand for orientation. Note that only the CD2 peptide binding site is largely hydrophobic in nature.

The GYF domain of CD2BP2 comprises 62 amino acids and has a β_1 - β_2 - α - γ - β_3 - β_4 topology (β : β -strand, α : α -helix, γ : bulge) with the four β -strands organized as a twisted antiparallel sheet (Fig. 1.4b). The α -helix is tilted away from the sheet and the side-chains of Trp 4, Tyr 6, Met 25, Phe 34, and Phe 58, as part of the hydrophobic core, are tightly packed between helix and sheet (Fig. 1.4a). Trp 8, Tyr 17, Pro 19, Phe 20, Trp 28, and Tyr 33 together with Tyr 6 and Phe 34 form a contiguous hydrophobic depression on the surface of the domain (Fig. 1.4a and c). Phe 20, Trp 28, and Tyr 33 represent three walls of this pocket; the floor is formed by Tyr 6 and Phe 34. The fourth wall is slightly more open and is composed of Trp 8, Tyr 17, and Glu 15. Similar to other PRD (Chapter 1.3.2), the hydrophobic pocket of CD2BP2-GYF confers binding to PRS^{23,121} (see below).

A third group of at least partly solvent exposed aromatic residues exists in the CD2BP2-GYF domain opposing the major hydrophobic hot spot (Fig. 1.4a). It contains Tyr 39, Phe 50, Tyr 51, and Tyr 61. The side-chain of Trp 4 is partly solvent exposed and is therefore included, too. The crystal structure of CD2BP2-GYF in complex with U5-15K defines parts of this site as additional protein-protein interface (see Chapter 11.4.2). Similar to other domains, amino- (N-) and carboxy- (C-) termini of the GYF domain are juxtaposed in space¹²², at a position in the domain, allowing for integration of GYF domains within existing proteins, without compromising either PRS or U5-15K binding in CD2BP2.

1.5.2 Structure of CD2BP2-GYF in Complex with the CD2 Peptide

The structure of the CD2BP2-GYF domain in complex with the CD2 derived peptide, SHRPPPPGHRV, revealed that the four prolines in the ligand (P4–P7) adopt a PPII helical conformation in the bound form^{23,121} (Fig. 1.4a). The hydrophobic pocket of the domain accommodates P6 and P7 with coplanar packing of the P6 pyrrolidine ring and the Trp 28 indol ring of the domain. Trp 28 further contributes to ligand binding by an H-bond between its amine group hydrogen atom and the backbone carbonyl oxygen of the ligand residue P4. The unrestrained dihedral angles of glycine allow for a sharp kink in the ligand at position 8 ($\Phi = 76^\circ$, $\Psi = 80^\circ$) which terminates the PPII helix. This conformation orients H9 towards the solvent and thereby prevents a collision with the side-chain of Trp 8. The R10 side-chain faces the domain, its aliphatic region forms hydrophobic interactions with Trp 8 and the positively charged head group interacts with the negatively charged side-chain of Glu 9 and/or 15. Further electrostatic attractions are probably operative between residues R3 and R10 of the ligand and the residues Glu 31 and Asp 36 in the domain, respectively. These residues could contribute to long range steering effects which facilitate ligand encounter. The GYF domain and other PRD share the use

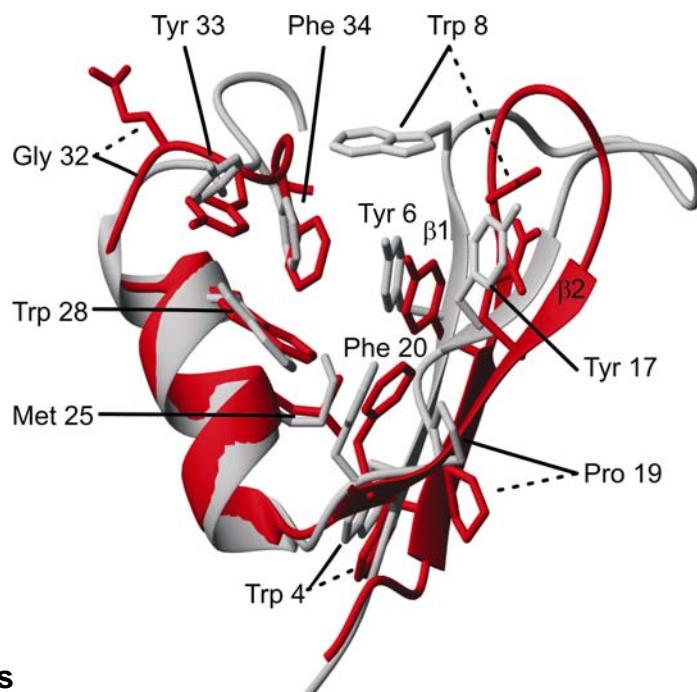
of (i) an aromatic cradle as major constituent of the binding site and (ii) a conserved tryptophan/tyrosine as an H-bond donor (Fig. 1.4a) to accomplish binding of proline-rich ligands (Chapter 1.3.2). The analogy is particularly striking for GYF and WW domains. Both contain a single xP dipeptide binding pocket that preferably accommodates two consecutive proline residues (Fig. 1.2d and Chapter 10). Association of CD2BP2-GYF with its ligand occurs without major rearrangements of the domain^{23,43,120}, similar to ligand binding observed for SH3^{107,123,124} and WW domains¹²⁵.

1.5.3 Structure of SMY2-type GYF Domains

In addition to CD2BP2-GYF, the structure of a member of the SMY2 subfamily has been determined by NMR, the GYF domain of the *Arabidopsis thaliana* protein Q9FT92 (PDB: 1WH2)¹²⁶. Both domains have a very similar overall fold and residues of the GYF domain signature superimpose well, suggesting that PRS binding properties are conserved (Fig. 1.5). There are significant differences, however, in both the sequence and the structure of the C-terminal regions of the two domains (Fig. 1.3 and Fig. 1.6).

Fig. 1.5: Comparison of CD2BP2-GYF and Q9FT92-GYF

The conserved N-terminal halves of both GYF domains (residue 1–36 in CD2BP2-GYF) are superimposed. These fragments comprise all conserved residues of GYF domains, the first two β -strands, and the helix-bulge motif. Side-chains at positions 8, 17, and of the residues which partake in the GYF signature are depicted and labeled according to residues in CD2BP2-GYF (continuous lines). Dashed lines indicate corresponding residues in Q9FT92-GYF. Hydrogen atoms are omitted for clarity and hence glycine residues cannot be seen. Gly 32 is replaced by Glu in Q9FT92-GYF. The C-terminal parts of the domains are not shown.



1.5.4 GYF Domain Related Folds

Several other protein shapes reveal similarities to GYF domains (DALI^{127,K} results; Fig. 1.6). These proteins have low sequence homology to GYF domains and lack the conserved residues (Fig. 1.3d). The characteristic bulge succeeding the helix is mostly missing, but the protein folds

^K www.ebi.ac.uk/dali/

share a similar topology with GYF domains, with a helix packing against an anti-parallel β -sheet (Fig. 1.6). Unlike most GYF domains, which probably represent autonomous folding units within the respective full-length proteins, some of the folds are an integral part of larger domains. The absence of conserved, exposed aromatic residues is likely to render the candidate domains incapable of binding to proline-rich ligands.

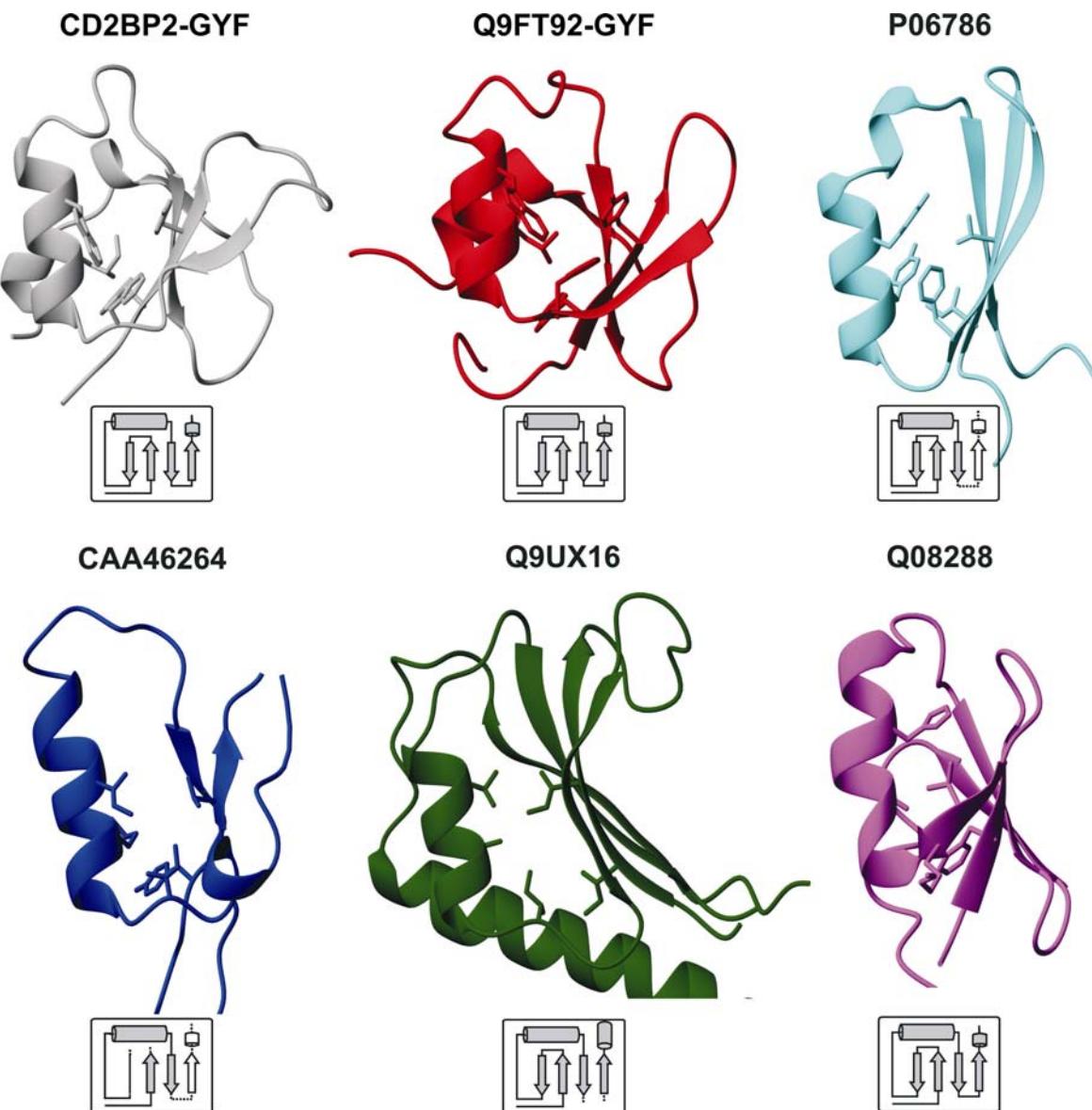


Fig. 1.6: Comparison of GYF domains and related folds

The GYF domains of CD2BP2 and Q9FT92 in comparison to structures, identified by DALI as GYF-like folds (sequences and PDB entry names are given in Fig. 1.3d). Indicated folds share a Z-score > 2 and a RMSD < 3 Å and were either determined by NMR (CD2BP2, Q9FT92, and Q08288) or by X ray diffraction (P06786, CAA46264, and Q9UX16). The latter two folds are part of larger structural organizations within the respective proteins. The topology of each structure is shown below. Arrows stand for β -strands, cylinders for helices. Filled and empty symbols indicate the presence and absence of the corresponding secondary structure elements, respectively. Insertions, not essential for the GYF fold, are represented by two dotted lines. The GYF domain characteristic residues at position 4, 6, 20, 25, and 28 (CD2BP2-GYF numbering) and the corresponding side-chains in the other folds (see Fig. 1.3d) share similar physico-chemical properties and are therefore depicted.

1.6 Functional Context of GYF Domains

At the time when work for this thesis commenced, information about the biological role of proteins containing a GYF domain was very limited. A potential role in T cell signaling originated from the pioneering work on CD2BP2 and its GYF domain^{42,43}. Screening for genetic suppressors of a *Saccharomyces cerevisiae* myosin mutant and a *Schizosaccharomyces pombe* kinase mutant identified the GYF domain containing proteins SMY2 (suppressor of myo2-66)¹²⁸ and MPD2 (multicopy suppressor of pld1 2)¹²⁹, respectively. The functional connection to splicing or splicing-associated processes arose from results of this work and from other groups during the last 4 years. An introduction to this topic is included here as well.

1.6.1 Involvement of CD2BP2 in T cell Signaling

CD2 is an adhesion molecule on T lymphocytes, thymocytes, and natural killer cells¹³⁰⁻¹³⁴ while the GYF domain containing interaction partner, CD2BP2, is expressed in different tissues^{135,L}. The extracellular N-terminus of CD2 comprises two immunoglobulin-like domains, the first of them interacting with the human counter receptor CD58 (CD48 in mice) on antigen presenting cells (APC)¹³⁶. The CD2 and CD58 extracellular domains are juxtaposed and span the ~ 15 nm distance between the T cell and the APC membrane as it is defined by the central interactions of T cell receptors with peptide-loaded major histocompatibility complex (MHC) molecules¹³⁷. Multiple CD2–counter receptor interactions stabilize the T cell–APC contact, in concert with other adhesion molecules, such as LFA-1 and ICAM-3 on T cells and ICAM-1, -2, and -3 on the APC¹³⁸. Thereby, CD2 engagement reduces the activation threshold of T cells¹³⁹, enhances interleukin 12 (IL-12) responsiveness of activated T cells¹⁴⁰⁻¹⁴², and induces T cell polarization¹⁴³. In addition to the scaffolding function, CD2 also augments TCR signaling via its cytoplasmic tail¹⁴⁴. The signal transduction capacity of CD2 has been shown by cross linking experiments using the T11₂ / T11₃ pair of antibodies, which can induce IL-2 production and T cell proliferation, without additional TCR stimulation¹⁴⁵. The cytoplasmic tail of CD2 recruits essential signaling molecules via five conserved PRS¹³⁴ (see Chapter 1.1.3), amongst them, the Src kinase Fyn^{75,146}, CD2BP1⁷³, CD2AP^{71,72}, and CD2BP2⁴². Fyn is important for T cell activation, since it phosphorylates a number of relevant signaling molecules after its activation^{75,147}. CD2BP1 recruits PTP-PEST to CD2, thereby enhancing the motility of cells⁷³. CD2AP is involved in T cell polarization and cytoskeletal rearrangements^{71,72}. The two membrane proximal PRS of signature PPPGHR have been shown to be crucial for IL-2 signaling^{148,149} and, intriguingly, these

^L Unigene at www.ncbi.nlm.nih.gov

are the CD2BP2 interaction sites. Overexpression of CD2BP2-GYF enhances CD2-triggered IL-2 production in Jurkat cells whereas transcription of an antisens construct has the opposite effect⁴². Fyn-SH3 and CD2BP2-GYF compete for binding to CD2 *in vitro*²³, but only Fyn is stably associated with the cytoplasmic membrane *in vivo*^{42,150,151}. Upon stimulation, CD2 partially translocates to the detergent-insoluble fraction of T cell lysates¹⁵². It has been suggested that the replacement of the GYF domain-mediated CD2BP2–CD2 complex by an SH3 domain driven Fyn–CD2 interaction may act as a potential trigger for downstream signaling²³. In contrast to Fyn and its SH3 domain, the implications of the CD2BP2–CD2 interaction for downstream signaling in T cells are elusive.

1.6.2 Spliceosomal Functions of GYF Domains

The coding regions of most genes in higher eukaryotes are interrupted by introns^{153,153}. Their removal from primary transcripts is an elaborate process known as splicing¹⁵⁴ and is a prerequisite for correct translation. The origin of introns is still a matter of intense debate. The introns-early theory regards the non-coding regions as ancestral elements of genome architecture, originally separating exons encoding short amino acid modules. Genome streamlining during evolution eliminated introns in non-eukaryotic organisms^{155,156}. According to the introns-late theory, introns were absent in ancestral organisms and arose late in evolution^{62,63,157}. The importance of introns for protein evolution, however, is undoubtedly. Introns are believed to accelerate protein evolution by facilitating the recombination of exons⁶³ (see Chapter 1.1.2). Introns themselves can also evolve new functions. For example, they can code for small nucleolar RNAs (snoRNAs)¹⁵⁸ and micro RNAs (miRNAs)¹⁵⁹ or comprise entire genes, encoding maturases or transposases¹⁶⁰. The alternative use of intron-excision sites gives rise to several protein isoforms of distinct functionality from a single gene transcript^{161,162}. In fact, about 50 % of the human genes are spliced to different isoforms^{163,164}. Alternative splicing is therefore a rich source of protein diversity in vertebrates^{165,166} and potentiates the complexity of the proteome.

Mechanism of Splicing

Removal of introns occurs in two consecutive transesterification reactions (Fig. 1.7a). Precise excision requires specific sequence signatures to mark the correct cleavage sites. The 5' and 3' splice site motifs flank the intron while the branch site is an internal conserved region, followed by a polypyrimidine tract in higher eukaryotes^{167,168}. In the first splicing reaction, the 5' splice site is cleaved, generating a free 5' exon. Since the hydroxyl group initiating the reaction is provided by an internal nucleotide from the branch site (predominantly an adenosine), the intron is

converted into a lariat intermediate with the branch-point nucleotide connected to its 5' end. In the second step, the free 3' hydroxyl group of the 5' exon attacks the 3' splice site resulting in the fusion of the two exons and the release of the intron as a lariat structure¹⁶⁹⁻¹⁷².

The splicing reaction is catalyzed by the spliceosome, a large protein–RNA complex (50–60 S) of dynamic composition. It comprises all together five uridine-rich small nuclear RNAs (U snRNAs; U1, U2, U4, U5, and U6; Fig. 1.7b) and about 300 proteins¹⁷³⁻¹⁷⁵. The spliceosome is likely to be a ribozyme¹⁷⁶, with the catalytic RNA moieties supported by the protein framework.

Assembly and Composition of the Spliceosome

The U snRNAs are present in form of protein–RNA complexes, small nuclear ribonucleoprotein particles (snRNPs), which are the central building blocks of the spliceosome. Maturation of the U1, U2, U4, and U5 snRNP involves nuclear-cytoplasmic shuttling and modification of the snRNAs including their 5' cap structure in higher eukaryotes¹⁷⁷. In the cytoplasm, seven Sm core proteins form a ring structure around the conserved uridine-rich Sm site in the snRNAs¹⁷⁸⁻¹⁸⁰. Back in the nucleus, association of snRNP-specific proteins completes snRNP assembly. The U6 snRNA is exceptional in having a different cap structure, lacking the Sm site and associating with Lsm (Sm-like) instead of Sm proteins. Moreover, the U6 snRNP maturation process is devoid of a cytoplasmic phase¹⁷⁷.

The five snRNPs are arranged in at least six distinct spliceosomal complexes: E, A, B, BΔU1, B*, and C in higher eukaryotes (seven in *Saccharomyces cerevisiae*; Fig. 1.7c)¹⁸¹⁻¹⁸⁵. Assembly begins with the recognition of the 5' splice site, presumably concurrent with transcription. The U1 snRNA base pairs with the 5' splice site (commitment complex 1 (CC1) in yeast)^{181,186-188} and contacts to the branch site lead to the E complex (commitment complex 2 (CC2) in yeast)¹⁸¹⁻¹⁸³. ATP-dependent formation of the pre-spliceosome (complex A) involves U2 snRNA base pairing with the branch site. Thereby the branch nucleotide is bulged out from the branch site–U2 snRNA duplex helix^{189,190}. Association of the preformed U4/U6•U5 tri-snRNP results in the spliceosome (complex B)¹⁹¹⁻¹⁹³, which further matures, releasing the U1 snRNP, to form complex BΔU1¹⁸⁵. The dissociation of U1 snRNP allows base pairing of the U6 snRNA with the 5' splice site¹⁹⁴⁻¹⁹⁷ upon disruption of U4–U6 snRNA interactions^{198,199}. Release of U4 snRNP defines the complex B*, also known as activated spliceosome^{184,185}, and coincides with U2–U6 snRNA interactions^{198,199}. In complex B*, the 5' splice site and the branch site are juxtaposed to initiate the splicing reaction. The first transesterification converts the activated spliceosome into complex C, which then accomplishes the second reaction step. The U5 snRNA has been proposed to bring together the two exons for the second reaction to proceed²⁰⁰⁻²⁰³. After completion of the splicing reaction, the ribonucleoprotein complex disassembles, the spliced mRNA is exported

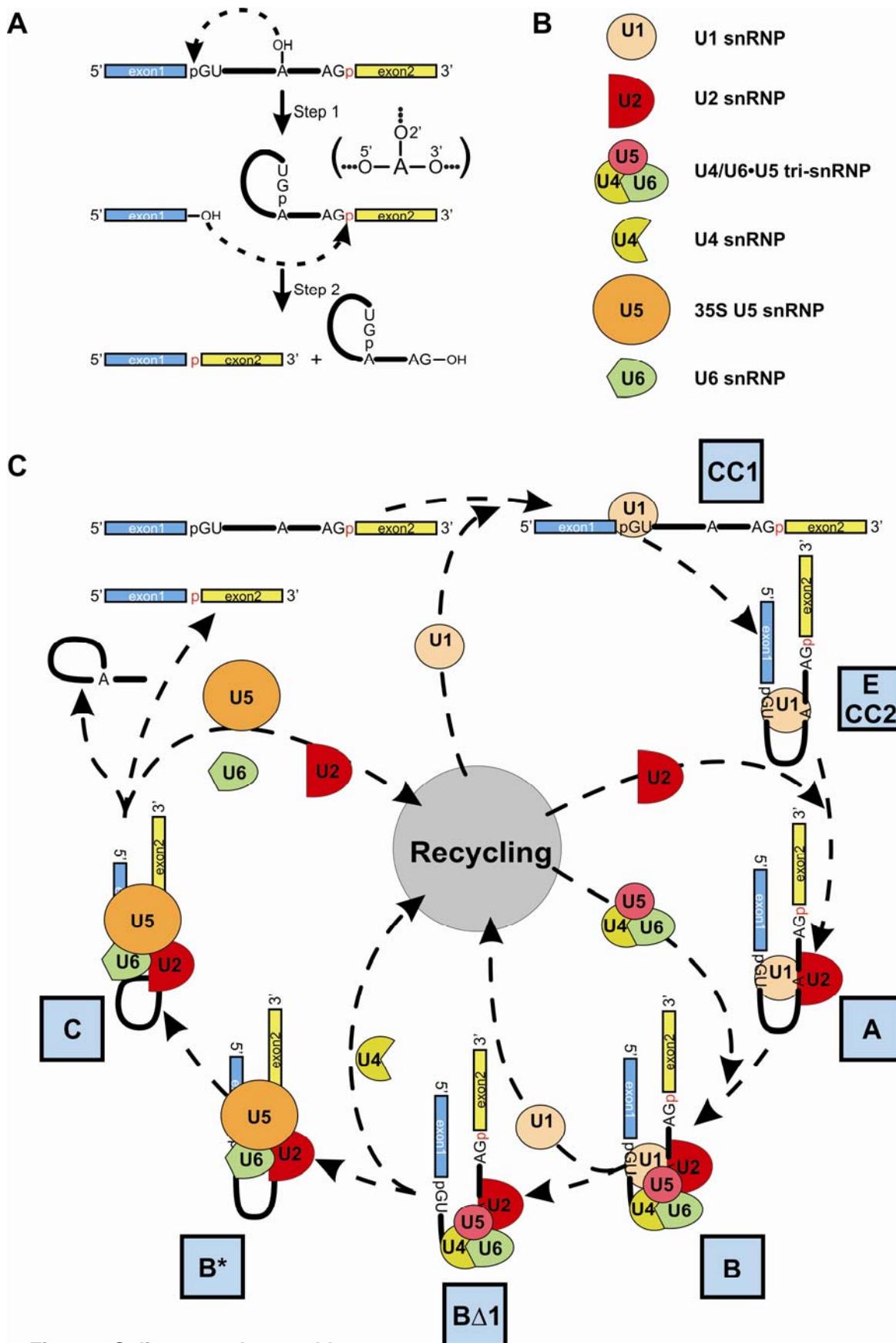


Fig. 1.7: Spliceosomal assembly

(A) Mechanism of nuclear pre-mRNA splicing. (B) snRNPs involved in splicing. (C) Schematic diagram of spliceosomal remodeling events. The model depicts the defined spliceosomal complexes (labeled with letters in blue boxes) and the association or release of U snRNPs. Other factors are omitted.

into the cytoplasm for translation and the utilized snRNP complexes are recycled¹⁷⁵. The wealth of proteins, implicated in splicing, support the RNA–RNA interaction network and dynamics^{176,204}. The underlying protein–protein contacts are frequently mediated by arginine-serine domains, RNA recognition motifs^{205,206}, and WW domains²⁰⁷. The numerous occurrence of PRS in spliceosomal proteins particularly underscores the importance of WW domains and possibly other PRD for spliceosomal assembly. The transient nature of interactions provided by these adaptor domains is well suited to support the dynamic processes during the splicing reaction.

Recently, the model of step-wise assembly of the spliceosome has been challenged by the observation of a penta-snRNP (U1•U2•U4/U6•U5) complex, purified from *Saccharomyces cerevisiae* lysates²⁰⁸. Proponents of the model of a preformed spliceosome interpret the observed, individual spliceosomal complexes as stable cores of the complete spliceosome at different stages.

The more loosely attached components at these spliceosomal stages are lost during the stringent scheme of spliceosome purification. This model might be of relevance for the detection of protein-protein interactions, mediated by PRD since they are characterized by high off rates and moderate affinities.

Involvement of GYF Domain Containing Proteins in Splicing

The yeast homolog of CD2BP2, LIN1, has been implicated in splicing, based on its identified interaction with PRP8²⁰⁹. Its functional role in splicing is especially attributed to the GYF domain since PRP8 has N-terminal PRS, reminiscent of those in CD2 that interact with CD2BP2-GYF. Further evidence for the involvement of GYF domains in spliceosomal processes stems from the finding that the representative of the second subfamily of GYF domains, SMY2, and its paralog SYH1 (SMY2 homolog 1; Swiss-Prot entry name Q02875) both bind to MSL5 (MUD synthetic-lethal 5) and MUD2 (mutant synthetic-lethal with U1 snRNA 2) in yeast two-hybrid experiments²¹⁰. MSL5, also known as yeast branch-point binding protein (BBP/ScSF1) and MUD2, as well as their human counterparts, BBP/hSF1 and U2AF⁶⁵, are components of the commitment complex 2/E complex²¹¹⁻²¹⁴. The proteins interact with each other, thereby bridging the branch-point region and U1 snRNP at the 5' splice site^{210,214,215}.

Finally, the identification of CD2BP2 in the human pre-spliceosome²¹⁶ further links GYF domains to splicing. A role for CD2BP2 in the assembly of the U4/U6•U5 tri-snRNP has been suggested due to its presence in the U5 snRNP prior to the formation of the tri-snRNP²¹⁷. In line with this assumption is the finding that PRP8, the LIN1 binding partner, is required for association of the U5 snRNP with the U4/U6 snRNP^{203,218,219}.

1.6.3 Miscellaneous Functional Contexts

Other data suggest diverse functional implications of GYF domain containing proteins. LIN1 interaction partners link the protein to chromosome cohesion or condensation and DNA repair, in addition to splicing²⁰⁹ (see above). MPD2, the SMY2 homolog in *Schizosaccharomyces pombe*, has been identified as a multicopy suppressor of the cdc7-D1¹²⁹ mutation. This finding functionally relates MPD2 to CDC7, a protein kinase that regulates replication initiation and heterochromatin-mediated cohesion²²⁰.

The mouse Grb10 protein interacts with the insulin-like growth factor receptor (IGFR) and has been shown to bind the Grb10 interacting GYF proteins (GIGYF) 1 and 2, homologs of the human PERQ (P, E, R, and Q amino acid rich with GYF domain protein) proteins 1 and 2. A proline-rich region in mouse Grb10 is thought to be responsible for these interactions²²¹.

Several lines of evidence support a functional role of proteins, comprising a GYF domain, in transport processes. The yeast protein SMY2 was originally cloned as multicopy suppressor of myo2-66, a temperature sensitive mutation within the motor protein MYO2¹²⁸, and its paralog SYH1 was found to be synthetically lethal with ric1²²². The encoded protein RIC1 has a function in vesicular transport^{223,224}. Furthermore, MPD2 suppresses the mRNA export defect of ptr1-1, a mutation within the putative HECT-type ubiquitin ligase PTR1²²⁵. Interestingly, splicing and transport processes might be linked, as it is indicated by the functional interaction of MUD2 with SUB2²²⁶, which is involved in mRNA export²²⁷.

At the time when work for this thesis commenced, the lack of information about the biological activities, additional domains or folded regions comprised in CD2BP2 and other GYF domain containing proteins¹³⁵ prevented the elucidation of their precise biological roles. For SH3, WW, and PDZ domains, detailed analysis of the binding properties allowed researchers to decipher their recognition codes (Table 1.2). In the era of proteomics, with growing numbers of completely sequenced genomes and improved protein prediction algorithms, these recognition codes have proven useful for the identification of novel interaction partners and the functional annotation of the corresponding proteins^{25,40,228-231}. The limited information about the GYF domain binding properties and about the entire proteins therefore called for a similarly systematic analysis.

1.7 Aim of the Work

The main intention of this work was to determine the recognition code of different GYF domains. We aimed to use phage display to identify the highest affinity binders, while more detailed information about the contribution of individual amino acids in peptide ligands to binding should be obtained from SPOT experiments. A further aim of the project was to describe binding by quantitative measures and to put the observed binding specificities in the context of the atomic structure of the GYF domain. Finally, yeast two-hybrid, pulldown, and cellular localization experiments were anticipated to reveal putative biological significance of the respective GYF domain-mediated interaction.

1.7.1 GYF Domains Selected for Analysis

For a comprehensive study of the binding properties of GYF domains, members of the first and second subfamily were chosen for analysis (Fig. 1.3a and b). Domains belonging to the hypothetical third subfamily were suspected to be elements of larger three-dimensional arrangements (DAGAT domains), rather than being autonomous folding units. They lack most of the exposed, conserved aromatic residues - a hallmark for PRD (Fig. 1.3c). Correspondingly, these domains were not anticipated to exhibit proline binding properties. For the same reasons GYF related folds (Fig. 1.3d and Fig. 1.6) were excluded from the study. GYF domains of the model organisms *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Drosophila melanogaster* in addition to human domains were selected because of the wealth of information about these organisms, including completely sequenced genomes and extensive proteome analysis. This choice also allowed the characterization of domains from the three major kingdoms of eukaryotic life: plantae, fungi, and animalia. The GYF domains under study were derived from the human proteins CD2BP2 and PERQ2 (Swiss-Prot entry name O75137), from the *Arabidopsis thaliana* protein GYN4 (GYF domain-containing protein binding to Not4, Swiss-Prot entry name Q9FMM3), from the *Drosophila melanogaster* protein Q9VKV5, and from all GYF domain containing proteins in *Saccharomyces cerevisiae*, namely LIN1, SMY2, and SYH1. CD2BP2-, Q9VKV5-, and LIN1-GYF are representatives of the CD2BP2-GYF subfamily, the other four domains belong to the SMY2 subfamily. Amongst them, the GYF domains of GYN4 and SYH1 were of particular interest because of a potential regulatory mechanism of their binding competence by intramolecular interactions, a well known mechanism of SH3 domains to regulate their binding to other proteins^{74,80}.

1.7.2 Approach

From the plethora of screening methods (see Chapter 2), a combination of phage display²³² and SPOT peptide arrays²³³ was employed to identify the binding properties of the selected GYF domains.

Phage display allows the screening of large peptide or protein libraries, displayed on the surface of filamentous phage^{232,234} and has been successfully applied to delineate the recognition characteristics of different adaptor domains (see Chapter 2). Based on the low affinities often observed for binding of PRD to PRS^{21–25,A}, a phagemid system²³⁵ was chosen, displaying peptides fused to the major capsid protein g8p (gene 8 protein). Multiple copies of the peptide on the surface of the phage enhance avidity effects and were expected to allow the selection of GYF domain binding partners.

Following phage display, experiments with peptide arrays, synthesized on cellulose membranes (SPOT analysis, see Chapter 2.3), were conducted to refine obtained recognition motifs and to identify potential interaction partners. Single-substitution SPOT analyses of binding peptides allowed the determination of key amino acids within recognition motifs responsible for binding, as it has proven useful for WW^{25,49,236,237}, PDZ^{230,231}, and EVH1²² domains.

The recognition signatures of different GYF domains set the basis for database searches in the proteomes of the respective organisms to identify potential interaction sites in proteins. Human, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana* proteome databases were screened for the obtained GYF domain recognition signatures and binding to the identified sites in proteins was studied by SPOT analysis, an approach recently described for SH3 domains⁴⁰.

Yeast two-hybrid screens are complementary to phage display screens⁴⁶ and were therefore incorporated into the strategy. Furthermore, pulldown, yeast two-hybrid, and colocalization experiments were conducted to verify putative GYF domain–protein interactions. Finally, NMR experiments and fluorescence titrations provided structural insight into the recognition mechanisms and allowed the determination of binding affinities of different GYF domains for various ligands (see Chapter 3).

CHAPTER 2

Selection Methods to Identify Protein–Protein Interactions

Many different methods exist to study protein-protein interactions^{17,235}. Three general types of screening methods can be distinguished according to the polypeptide libraries employed.

2.1 Endogenous Proteins from Cell Extracts

Immunoprecipitation and tag-based coprecipitation experiments represent powerful methods to isolate proteins of interest and their associated interaction partners from cell extracts^{53,238-240}. Copurified proteins are subsequently identified by immunodetection or mass spectrometry. The recent advances in mass spectrometry have allowed the characterization of large protein complexes^{216,241} and even the analysis of protein interaction networks on a proteomic scale^{242,243}.

2.2 Expression of Proteins from DNA Libraries

The common theme in this group of methods is the expression of polypeptides from encoding DNA libraries. Depending on the strategy, interactions between library members and the target protein, often referred to as prey and bait, respectively, take place *in vivo* or *in vitro*.

2.2.1 *In Vivo* Selection

In the yeast two-hybrid system, bait and prey are expressed as fusion proteins and the interaction between both restores the function of a transcriptional activator for reporter gene expression. The reporter protein in turn supports cell growth on selective media or confers a detectable enzymatic activity²⁴⁴. The method has proven very useful²⁴⁵⁻²⁴⁷ and is suitable for the analysis of whole proteomes²⁴⁸⁻²⁵¹. Numerous modifications and comparable systems have been developed^{235,247}.

2.2.2 *In Vitro* Selection

In screens of phage expression libraries²⁵², interaction between prey and bait is detected in the format of a far Western blot²³⁵. Arrays of purified proteins employ the same detection principle, but allow faster identification of preys^{228,229,253}. Display methods are based on macromolecular complexes (including intact cells) which physically link the prey protein and its encoding DNA. The selection process occurs in repetitive cycles of affinity selection and subsequent amplification of the selected sublibrary, as opposed to the other methods described above comprising a single selection round. The prototype of display methods is phage display. Polypeptides, encoded by recombinant viral DNA, are presented on the surface of a bacteriophage and determine the binding properties of viral particles to immobilized bait protein^{232,234}. The bacteriophage M13, or the closely related f1 or fd phages are often employed for presentation. They are filamentous phages, with a circular single stranded genome, that infect *E. coli*. Proteins or peptides of interest are mostly fused either to the major capsid protein g8p or the minor capsid proteins g3p. g8p-fusion allows for a multivalent display, whereas g3p-fusion limits the number of copies per virion to 3–5. The method has been used successfully to delineate the binding profiles of several adaptor domain families such as SH3⁴⁴, WW^{254–257}, SH2^{258–260}, and PDZ domains²⁶¹. The numerous applications and modifications of phage display are reviewed elsewhere^{235,262–267}. In contrast to phage display, ribosome display²⁶⁸ and mRNA display^{269,270} are solely performed *in vitro* and do not require a transformation step which can limit the diversity of the library. All steps, transcription of the DNA library, translation of mRNA, and selection, take place outside the cell. In ribosome display, the polypeptide and its encoding mRNA are linked via the ribosome while, in the case of mRNA display, a chemical linker tethers the mRNA and the polypeptide covalently²⁷¹. The inherent monovalent display precludes selection of low affinity interaction partners.

2.3 Synthetic Peptides

Methods which are based on screening of synthetic peptide libraries have been utilized extensively to identify binding sites or refine recognition codes²⁷². Two major routes exist, employing either soluble^{50,273-275} or immobilized peptide libraries. Peptide synthesis²⁷⁶ circumvents transformation or transfection steps possibly limiting the diversity of the library. Selected peptides from soluble peptide libraries are usually identified by peptide sequencing using Edman chemistry²⁷⁷. Besides the substantial amounts of peptides required for sequencing (~ 10 pmoles²³⁵), this approach solely allows the determination of a binding motif, averaged over the mixture of selected peptides and hence prevents the detailed analysis of positional interdependence within a ligand. Beads, each coupled with one type of peptide (known as a ‘one beat one compound’ library)²⁷⁸ facilitate analysis of binders. They provide enough material of each interacting peptide for separate sequencing^{107,279-281} but depend on a manual separation step. Peptide arrays²⁸² offer an elegant alternative for rapid identification of binding peptides, where sequence information is imprinted in the position of the peptide on the array, similar to protein arrays. SPOT synthesis, the highly parallel synthesis of peptides on cellulose membranes by position-specific application of defined building blocks in each synthesis cycle^{233,283}, has become a widely used tool to study molecular recognition. Although the density of peptide spots is low, compared to arrays that are based on photolithographic synthesis of peptides²⁸⁴, up to 2000 spots on a 8 x 12 cm membrane (microtiter plate size) can be synthesized²⁷². Binding peptides are identified by detecting the positions where the target protein is specifically retained on the membrane, either by radioactive or fluorescent labeling or by antibody-based recognition of the target protein. SPOT synthesis has been utilized successfully for epitope mapping (also known as peptide walking)^{54,285,286}, alanine scanning²⁸⁷, substitution analysis^{22,25,49,54,230,236,288}, screening of potential peptide ligands derived from genomic sequences^{40,54,230,231} or mutational analysis of the binding domain²³⁷. Extension or extrapolation of the sequence space to be screened is achieved by a position-wise analysis of binding peptides, neglecting their respective sequence background^{230,289}. However, similar to methods based on soluble peptides, this strategy precludes potential interdependence of neighboring positions and identifies an average recognition code.

CHAPTER 3

NMR Experiments

Following identification of interaction partners, the affinity and the molecular mechanisms of binding come to the fore. The first issue can be addressed by titration experiments, where the change of physical properties upon addition of increasing amounts of ligand allows the determination of a dissociation constant. Examples are fluorescence titration, isothermal titration calorimetry, and binding tests based on surface plasmon resonance. Information about the interaction mechanism of two molecules can be obtained from solution nuclear magnetic resonance (NMR) spectroscopy. In the case that the interaction is characterized by high off rates (k_{off}) when compared to the time scale of the NMR experiment ($k_{\text{off}} \sim 1 \text{ ms}$) – which is usually the case for interactions between PRD and PRS – both the binding affinity and the recognition epitope on the protein surface can be analyzed simultaneously. NMR spectroscopy is based on the existence of a quantum mechanical property of nuclei with an odd mass or charge number called spin. For a nuclear spin with spin quantum number I , $2I+1$ different states exist. In an external magnetic field, the energy levels of the states split (Zeemann interaction) and give rise to distinct magnetic momentums which precess around the vector of the external magnetic field with Larmor frequency ω_0 . The spins in a sample populate different energy states according to the Boltzmann distribution. Electromagnetic radiation orthogonal to the static magnetic field perturbs the distribution of this spin population and induces an observable magnetization of the sample perpendicularly to the external magnetic field. The evolution of this magnetization, called free induction decay, is measured in NMR experiments and reflects the Larmor frequencies of the excited nuclei. However, NMR spectroscopy is an insensitive method because the observable macroscopic magnetization is directly proportional to the population difference between the spin states. Typically, this is in the order of only 10 parts per million (ppm) for ^1H nuclei at 25 °C in a magnetic field of 14.1 Tesla, and for other nuclei, the difference is even smaller. The identity of the frequency of an absorbed energy quantum from the electromagnetic radiation and the Larmor frequency allows, within limits, a classical description of NMR spectroscopy. A complete description, however, requires quantum mechanical analysis. Details about the phenomenological approximation by the classical Bloch model and quantum mechanical description can be found in current text books²⁹⁰.

3.1 Chemical Shift

Nuclei ^1H , ^{13}C , and ^{15}N are central to NMR spectroscopy of organic compounds and macromolecules. Since these nuclei have spin $I=1/2$, they comprise only two energy levels and their magnetization state has a longer lifetime than that of nuclei with larger spin numbers. The Larmor frequency depends on the applied external magnetic field. The ratio of Larmor frequency and external magnetic field, called gyromagnetic ratio constant γ , is characteristic for a given nucleus. The electronic structure in the local environment modulates the effective magnetic field and correspondingly the Larmor frequency of a particular nucleus in a macromolecule. Usually, the chemical shift rather than Larmor frequency is given for a nucleus. The chemical shift represents the normalized difference of the Larmor frequency to a reference signal in ppm. Chemical shift values are dimensionless and independent from the applied magnetic field. In small to medium-sized, folded proteins (≤ 35 kDa)²⁹¹, the chemical shifts of the atomic nuclei can be discriminated in multi-dimensional spectra and allow their individual assignment.

3.2 Epitope Mapping and Determination of the Dissociation Constant

The chemical environment of nuclei in the binding site of proteins is altered upon ligand encounter and hence the chemical shifts of these nuclei change. Once the backbone resonance assignments have been obtained (Chapter 3.3), mapping of the changes onto the structure of the protein is an elegant method to determine the protein-ligand interface. Conformational changes which eventually orchestrate ligand binding, will also induce chemical shift changes and blur the interaction site footprint. Chemical shift perturbations can be observed for example using two-dimensional ^1H - ^{15}N -heteronuclear single-quantum coherence (HSQC) spectra. In this type of experiment, the chemical shifts of covalently linked ^1H and ^{15}N nuclei are determined at the same time, giving rise to resonance peaks in a two-dimensional spectrum.

The interaction kinetics affect the NMR spectra upon ligand titration. Low on and off rates of the ligand are characterized by the gradual disappearance of peaks corresponding to the free form of the domain, while those corresponding to the bound form appear during titration. Intermediate exchange rates results in substantial line broadening and prevent the detection of peaks. In a fast exchange regime, however, peaks gradually move from resonances of the free to resonances of the bound state. The relative peak position, measured by the chemical shift

changes of covalently linked ^1H and ^{15}N nuclei, reflects the average ratio of the free and bound population. Different regions of the binding epitope can display different binding kinetics.

Frequently, both chemical shift changes ($\Delta^1\text{H}$ and $\Delta^{15}\text{N}$) of each amide group are combined as:

$$[(10^*\Delta^1\text{H})^2 + (2^*\Delta^{15}\text{N})^2]^{1/2}$$

to facilitate the detection of chemical shift perturbations and the calculation of binding affinities in the case of a fast exchange regime. The weighting factors account for the different scale of chemical shift dispersion for ^1H and ^{15}N nuclei. For the GYF domain of CD2BP2, the interaction with the CD2 ligand was shown to be in the fast exchange regime⁴³. The complex structure (Fig. 1.4) confirmed the observed chemical shift changes to originate mostly from direct ligand binding rather than conformational changes^{23,121}. Therefore, NMR titration experiments with GYF domains were expected to allow the determination of both binding affinities and binding epitopes in a single experiment.

3.3 Backbone Assignment

A prerequisite for atomic resolution of the binding epitope is the assignment of individual NH group resonances according to the protein sequences. Strategies to accomplish the backbone assignment are based on the correlation of resonances of backbone nuclei via chemical bonds (J -coupling). Triple resonance experiments are used which correlated ^1H , ^{13}C , and ^{15}N resonances and thereby reduce spectral overlap. A detailed description of the different NMR spectra and the assignment strategies can be found elsewhere²⁹⁰.

