Decoding the structure and function of WWP2 in the TGFβ signalling pathway

Lloyd Christian Wahl

A thesis submitted for the degree of Doctor of Philosophy at the School of Biological Science, University of East Anglia, March 2016

Abstract

The secret to the specificity of the ubiquitin-proteasome system lies in the protein-protein interaction domains of the diverse group of E3 enzymes. WWP2 is one such E3 enzyme, and the relevant protein-protein interaction domain is the WW domain. WWP2 has four WW domains which are used to interact with proline-rich motifs found in the sequences of the Smad signalling proteins that propagate or inhibit the TGFβ pathway, and in so doing, allows WWP2 to regulate its Smad targets. WWP2 has three isoforms that are known to participate in regulation of TGFβ signalling, but, even amongst isoforms of the same E3, they exhibit different specificities for components of the pathway. The reason for this is unknown, but it is likely to be due to the different domain composition of WWP2, since each of the isoforms has a different combination of WW domains.

The aim of this thesis is to investigate the structure of the domains of WWP2, and to explore how this relates to the selectivity of different isoforms in the TGFβ pathway. Overexpression of recombinant WWP2 domains in a bacterial host, and affinity and size-exclusion chromatography have been used to produce pure, high concentration protein samples. Both NMR spectroscopy and crystallography have been used in an attempt to elucidate the structure of WWP2 domains. NMR spectroscopy, the more successful of the two approaches, has allowed the elucidation of the structure of the fourth WW domain of WWP2. By observing ligand interaction using NMR, the binding site of WW4 is revealed and the substrate preference of WW4 and WW3 domains is observed on a molecular level. Evidence of phospho-regulation of substrate selectivity is presented, and a structural basis for this selectivity is proposed. In addition, a further layer of complexity is added to the WWP2 isoform-mediated regulation of the TGFβ signalling pathway, as a new isoform is discovered.

Contents

List of Tables

List of Figures

List of Abbreviations

4HD - Four helix bundle domain

ALS - Amyotrophic lateral sclerosis

AMP - Adenosine monophosphate

APBS - Adaptive Poisson-Boltzmann Solver

APC/C - Anaphase-promoting complex/cyclosome

APS - Ammonium persulphate

ATP - Adenosine triphosphate

BARD1 - BRCA1 associated RING domain 1

bHLH - Basic-helix-loop-helix

BRCA1 - Breast cancer susceptibility gene 1

C/EPBβ - CCAAT/enhancer-binding protein-β

CDK - Cyclin-dependent kinase

CFTR - Cystic fibrosis transmembrane conductance regulator

CHIP - C-terminus of Hsc70 interacting protein

CLASP2 - Cytoplasmic linker-associated protein 2

CTGF - Connective-tissue growth factor

Dsh - Dishevelled

E6AP - E6-associated protein

ECM - Extracellular matrix

EDTA - Ethylenediaminetetraacetic acid

EGFR - Epidermal growth factor receptor

EMT - Epithelial to mesenchymal transition

ENaC - Epithelial sodium channel

ESRP - Epithelial splicing regulatory protein

EST - Expressed sequence tag

FID - Free induction decay

FITC - Fluorescein isothiocyanate

GSK - Glycogen synthase kinase

HECT - Homologous to the E6-associated protein Carboxyl Terminus

HECW - HECT C2 and WW domain-containing protein

HIF-1 - Hypoxia-inducible factor 1

HPV - Human papillomavirus

HRV - Human Rhinovirus

HSQC - Heteronuclear single quantum coherence

IMAC - Immobilised metal ion affinity

IPTG - Isopropyl β-D-1-thiogalactopyranoside

i-Smad - Inhibitory Smad

ITC - Isothermal titration calorimetry

LAP - Latency-associated peptide

LB - Lysogeny broth

LIP - Liver-enriched inhibitory protein

LLC - Large latent complex

LTBP - TGFβ-binding proteins

MALDI-TOF - Matrix-assisted desorption/ionization time-of-flight

MDM2 - Mouse double minute 2

MH - Mad-homology

MHC - Major histocompatibility complex

MMP - Matrix metalloproteinase

MWCO - Molecular weight cut-off

NEDD4 - Neural precursor cell expressed developmentally downregulated protein 4

NMR - Nuclear magnetic resonance

NOE - Nuclear Overhauser Effect

NOESY - Nuclear Overhauser Effect Spectroscopy

NTA - Nitrilotriacetic acid

O.D. - Optical density

OCT4 - Octamer-binding transcription factor 4

PAGE - Polyacrylamide gel electrophoresis

PBS - Phosphate buffered saline

PDGFRα - Platelet-derived growth factor receptor-α

PI3K - Phosphatidylinositol-3-kinase

Pin1 - Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1

PPII - Polyproline II

ppm - Parts per million

PSVS - Protein structure validation suite

PTEN - Phosphatase and tensin homolog

RF - Radio frequency

RING - Really Interesting New Gene

RMSD - Root mean square deviation

ROS - Reactive oxygen species

r-Smad - Receptor Smad

SARA - Smad anchor for receptor activation

SBE - Smad binding elements

SDS - Sodium dodecyl sulphate

SDS-PAGE - Sodium dodecyl sulphate-polyacrylamide gel electrophoresis

SH2 - Src homology 2

SMURF - Smad Specific E3 Ubiquitin Protein Ligase

TAE - Tris-Acetate-EDTA

TEMED - Tetramethylethylenediamine

TFA - Trifluoracetic acid

TGFβ - Transforming growth factor β

TβR-I - TGFβ receptor-I

TβR-II - TGFβ receptor type II

Ub - Ubiquitin

UBA1 - Ubiquitin-activating enzyme E1

UBC - Ubiquitin-conjugating domain

UEV - Ubiquitin E2 variant

ULP-1 - Ubiquitin-like protease-1

UPS - Ubiquitination-proteasome system

VCP - Valosin-containing protein

VEGF - Vascular endothelial growth factor

WWP2 - WW domain containing protein 2

YAP - Yes kinase-associated protein

Acknowledgements

I'd like to thank my supervisor, Dr Andrew Chantry, for his guidance and support, his craft beer (coming to a pub near you) and all the mackerel (so many mackerel). Most of all I would like to thank him for his patience during the periods in which I have been writing. I'm sure I've caused a great deal of anxiety for Andrew, but he has allowed me to follow my process and hopefully this completed thesis will be a repayment for the trust he has shown.

Acknowledgements must go to Dr Tharin Blumenschein for using her expertise to guide me through my project. Her ability to convey the complex world of NMR at a level a biologist can understand has been invaluable, and is greatly appreciated. I'd also like to acknowledge my secondary supervisor, Dr Andrew Hemmings. His wealth of knowledge and experience helped guide the endeavours in crystallography of this project, and has helped me in my approaches towards protein purification.

I'd also like to thank Tiffany Yim for driving me mad, and to Danielle De Bourcier and Jess Watt for having a good go at it as well. All joking aside, I should say something nice.

Special thanks must go to Dr James Tolchard, who, over the past couple of months, has assumed the role of life coach. James graduated from his doctorate in 2014 and, although there has been no obligation for him to do so, has continued to support me both as a friend and in my academic research as an expert in NMR. My suspicions are that he wanted to make it into my Thesis acknowledgements section, so I guess congratulations are also in order, it worked.

Finally, I'd like to thank my family and remaining friend. It is common for my dad to be much more worried about the challenges in my life than I am, and this thesis, and PhD in general, is no exception. I don't show it, but I appreciate your unwavering support, and now it's in writing. I should probably also thank caffeine, I feel like caffeine had a greater hand in this thesis than I did, so thank you.

1. Introduction

The driving force behind the life of cellular and multicellular organisms is proteins, produced from genes that are encoded in the DNA template. DNA is the mother code, responsible for accurate and extended reproduction of the protein workers, and for propagating the code to descendant cells. Information is stored in DNA by the specific sequential patterns of four different nucleotides. The explicit hydrogen bonding pattern of each nucleotide ensures complementarity with its specific partner, and allows transfer of genetic information to messenger RNA and then to a set of transfer RNA decoders that allow proteins to be assembled from the code. The proteins that are produced are responsible for virtually every activity in the cell. The environment created by the abundance and activity of different proteins dictates cell fate, and is dependent on the external environment. In order to sense the external environment, cells have a multitude of different receptors with complex and interconnected cascades of signal propagator proteins. This network of proteins determines which proteins are active and which are inactive, and feeds back to DNA to alter expression levels to determine which proteins should be present in abundance, and which should be removed. Hidden in the complex network of proteins is a code (analogous to the complementarity of nucleotide base pairs), that dictates which proteins interact with which and, therefore, how they function. This code has several layers of complexity, written by the primary amino acid sequence, and convoluted by secondary and tertiary folding which creates an interface at the protein surface that selects binding partners based on electrostatic and hydrophobic complementarity. This interface is regulated further by post-translational modification, which itself is dependent on protein activity. Post-translation modification has the ability to alter the properties of the binding site, either by directly changing the electrostatic profile or indirectly by altering the structural conformation by allosteric effects. Post-translational modification is also harnessed to destroy proteins, and thus, alter the protein environment of the cell. Protein destruction and protein expression are two sides of the same coin, and the balance between both of these ultimately decides cell fate.

One such post-translational modifier is WWP2 (WW domain containing protein 2), and its method of modification is by ligation of ubiquitin monomers, to form polymeric chains that causes its substrate to be destroyed. This thesis aims to explore how this protein functions, by decoding the secondary and tertiary folds formed by the primary amino acid sequence. The aim of this is to start to define how WWP2 fits in to the network of signalling cascades, and to therefore gain further insight in to how it might influence cell fate. The relevant environment-sensing signalling cascade here is transforming

growth factor-β (TGFβ), as WWP2 belongs to a family of proteins that are intimately involved in regulating TGFβ. These proteins interface with their binding partners using simple protein-protein interaction modules called WW domains, and there will be a strong focus on WWP2 WW domains.

This introduction will begin by outlining the different types of protein degradation present in the cell. Further attention will then be given to how the ubiquitin system functions, examples of misregulation, and then special consideration will be given to the three classifications of enzymes that constitute the ubiqiutin cascade. Starting at the top of the cascade, the E1 and E2 enzymes will be described, and then the two main classes of E3 enzymes; the RING family E3 ligases will be discussed and then the focus will move to HECT domain E3 ligases. Then the NEDD4 (Neural precursor cell expressed developmentally downregulated protein 4) E3 ligases, the WW domain-containing E3 family to which WWP2 belongs, will be briefly introduced. To put the function of these E3 ligases into context, canonical TGFβ signalling and its components will be described, including TGFβ in cancer. The domain composition of NEDD4 E3 ligase family members will be described in more detail and in particular, the structural functional relationship of the HECT domain and the WW domains. What is known about the function of WWP2 will be described and the aims of the experimental section of this thesis will be outlined.

1.1 Protein Degradation

Control over protein levels is a critical asset in virtually all cellular processes, and while a great deal of research has focused around regulation of protein synthesis, less attention has been given to the other half of the story - protein degradation. It is logical to consider over-activity at a gene promoter to hold a similar significance as an inability to degrade the gene product and, as is the way with most biological systems, protein synthesis and protein degradation are intimately linked, with each affecting the other. Unsurprisingly, proteolysis has been found to be an essential component in a diverse range of applications, from cell-cycle control to cell-repair, signalling cascades to memory formation. The core of protein turnover capabilities in the cell is composed of two systems: autophagy and the ubiquitin system.

Autophagy is the process by which cytoplasmic material is delivered to the lysosome where acid hydrolases decompose the materials in to their constituent parts which are then recycled (Mizushima & Komatsu 2011; Glick et al. 2010). There are three types of autophagy, macroautophagy, microautophagy and chaperone-mediated autophagy (Mizushima & Komatsu 2011). In macroautophagy a lipid membrane called an autophagosome forms around cytosolic materials which are then transported to the lysosome, the membrane then fuses with the lysosome and deposits its cargo (Mizushima & Komatsu 2011). This is a bulk process and cargo is not limited to proteins but also organelles and pathogens (Dodson et al. 2013; Glick et al. 2010). In microautophagy, materials encountered at the lysosome lipid membrane are consumed and degraded independent of an autophagosome. In chaperone-mediated autophagy, chaperones locate substrates to a transmembrane receptor of the lysosome, the substrate is subsequently unfolded and degraded (Mizushima & Komatsu 2011).

While there is growing evidence for a certain level of selectivity in autophagy (Johansen & Lamark 2011; Green & Levine 2014), autophagy has historically been considered a largely non-selective process that functions to replenish the nutrients required by the cell to maintain normal physiology. In contrast to the autophagy-lysosome system, the ubiquitin system provides a mechanism by which proteins are degraded on a highly selective and tightly regulated basis, utilising the proteasome instead of the lysosome as the proteolytic body. The specificity with which the ubiquitin system functions presents more research challenges in order to understand its selectivity, and will be discussed here in further detail.

1.2 The Ubiquitin System

In the ubiquitin system, the small ubiquitin protein is covalently attached to target substrates. Ubiquitin modification is performed for a variety of purposes, including altering protein activity or to label the protein for degradation at the proteasome. The process by which proteins are ubiquitinated (outlined in Figure 1.2.1), was the subject of work that resulted in the award of the Nobel Prize in chemistry in 2004. In a series of papers between 1978 and 1983, Ciechanover, Hershko and Rose collectively worked towards defining an ATP-dependent proteolytic system that involved the previously discovered ubiquitin molecule which they originally named active principle of fraction 1

(Ciechanover et al. 1978; Hershko et al. 1979; Ciechanover et al. 1980; Hershko et al. 1980; Ciechanover et al. 1981; Hershko et al. 1981; Haas & Rose 1982; Ciechanover et al. 1982; Hershko et al. 1983).

Their first step was to identify the factor (ubiquitin) responsible for the previously observed characteristic of cellular ATP-dependent proteolysis (Ciechanover et al. 1978), then they identified a high molecular weight component believed to be the proteolytic proteasome (Hershko et al. 1979). Following this, they discovered that multiple ubiquitin units polymerised on to lysozyme and caused its degradation, and even identified a deubiquitinating action (Hershko et al. 1980). In 1981, an activation step was characterised in which an enzyme (the E1) primed ubiquitin by ATP hydrolysis, which then formed a thioester with a cysteine residue side chain (Figure 1.2.1, first step) (Ciechanover et al. 1981). Using a novel ubiquitin-Sepharose affinity column three enzymes were then identified as being essential to the ubiquitination reaction, those were the E1 ubiquitin-activating enzyme, the E2 conjugating enzyme and the E3 ligase enzyme (Hershko et al. 1983; Ciechanover et al. 1982). Each play an important role in the cascade (depicted in Figure 1.2.1), and will be discussed below.

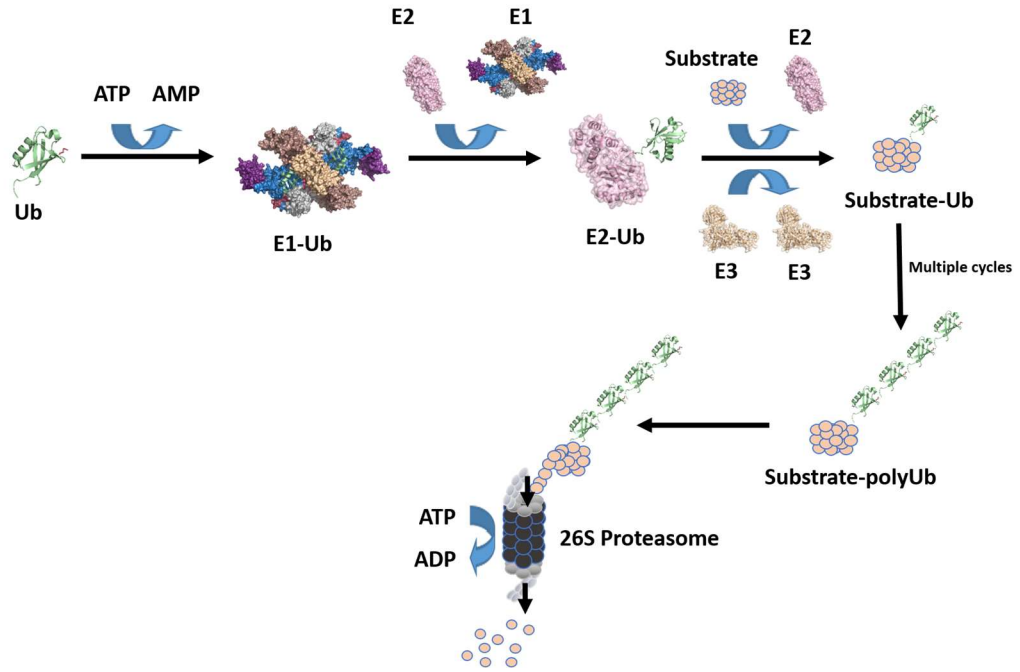


Figure 1.2.1 - The ubiquitin-proteasome system, showing the activation of ubiquitin (Ub) by ATP and the E1 enzyme, transfer of ubiquitin to the E2 conjugator, ligation of ubiquitin to a substrate by the E3 enzyme, multiple cycles resulting in polyubiquitination, followed by degradation at the proteasome. Adapted from (Corn 2007).

The ubiquitin-proteasome system (UPS) functions to rapidly and specifically turn over proteins. This action is used for many purposes including but not limited to: the regulation of gene transcription through the targeting of transcription factors, the removal of misfolded, mutated or damaged proteins, generating antigens for presentation by the major histocompatibility complex (MHC) class I, regulating signalling pathways and controlling progress through the cell cycle. The anaphase-promoting complex/cyclosome (APC/C) is a large 13 subunit E3 ubiquitin ligase that is activated by Cdc20 family members during M phase of the cell cycle and ubiquitinates separase, various cyclins and Cdc20, promoting exit from M phase (McLean et al. 2011). APC/C remains active during G1 phase to suppress cyclin activity and prevent mitosis. Cystic fibrosis is caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR), and can often result in protein misfolding. Misfolded CFTR is subsequently ubiquitinated and degraded at the proteasome before it reaches the cell surface, demonstrating the quality control mechanism of the UPS (Ward et al. 1995).

Gain of function or loss of function in the UPS plays an important, if complex, role in disease, with the lines often blurred between cause and effect. This is particularly true of neurodegenerative disorders such as Huntington's disease, Alzheimer's disease, Parkinson's disease and amyotrophic lateral sclerosis (ALS), in which ubiquitinated proteins are found in protein aggregates, the hallmark of these types of disease. The inability of a dysfunctional cell to clear these protein aggregates has historically been considered to be due to dysfunction of protein degradation by the ubiquitin system resulting in the accumulation of aggregation-prone proteins (Ciechanover & Brundin 2003; Waelter et al. 2001), but ubiquitin-proteasome dependent protein degradation has been seen to remain active (Dantuma & Bott 2014). Despite components of the ubiquitin system such as the E3 ligase Parkin and the proteasome shuttle factors ubiquilin and valosin-containing protein (VCP) being linked to neurodegenerative diseases, there is evidence suggesting secondary functions such as coordination with the autophagy system and degradation-independent ubiquitination are of equal significance (Dantuma & Bott 2014). It is likely that a coordinated age-related decline in both proteasomal and lysosomal degradation is responsible for progression of the neurodegenerative phenotype (Martinez-Vicente et al. 2005; Löw 2011).

A defective UPS can contribute to the development of an oncogenic phenotype by circumventing the quality control mechanisms, and allowing the accumulation of mutated and oncogenic proteins. This can cause misregulation of signalling pathways and

can prevent the effective repair of DNA. The p53 tumour suppressor is targeted for degradation at the proteasome by mouse double minute 2 (MDM2), which is overexpressed in a number of cancers (Fulda et al. 2012). Subsequently, the cell cycle checkpoint and apoptotic mechanisms in response to DNA damage are circumvented. Breast cancer susceptibility gene 1 (BRCA1) is an E3 ligase that forms a dimer with another E3 BRCA1 associated RING domain 1 (BARD1) and ubiquitinates cyclin B and Cdc25C in response to DNA damage (Shabbeer et al. 2013). Loss of function mutations in the BRCA1 gene cause a high predisposition for breast cancer development and are present in more than 50% of hereditary breast cancers (Shi & Grossman 2010).

1.2.1 The Proteasome

Substrate fate depends on the type of ubiquitination; monoubiquitination often alters the activity of the protein while polyubiquitination of a protein typically marks it for its demise at the multi-subunit 26S proteasome. The proteasome is formed by a large complex of multiple proteins, and is large enough to observe by electron microscopy (Figure 1.2.2). Polyubiquitinated proteins are recognised by a complex of the 26S proteasome called the 19S proteasome (Figure 1.2.2), which prevents non-specific degradation of untagged cellular proteins by functioning as a gate to the proteolytic 20S complex at the core (Glickman & Ciechanover 2002). When proteins encounter the 19S proteasome, the ubiquitin chain is cleaved into monomers and recycled by a combination of deubiquitinating enzymes associated to the regulatory structure (Lee et al. 2011). The substrate is unfolded and passes in to the core where a mixture of proteases with trypsin, chymotrypsin, and caspase-like activity cleave the protein into small peptides to be recycled by the cell (Glickman & Ciechanover 2002; Nussbaum et al. 1998; Adams 2003).

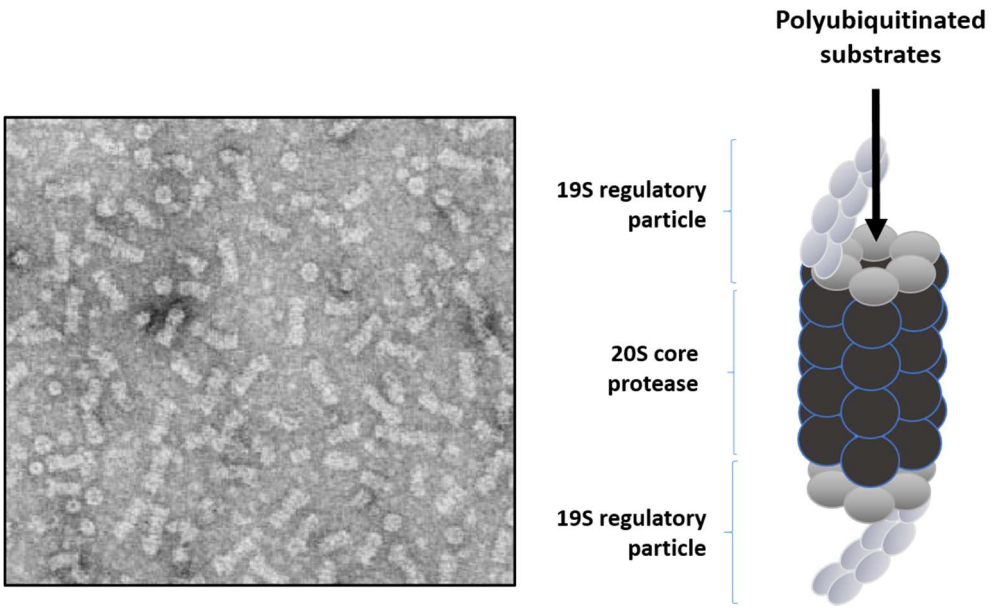


Figure 1.2.2 - Electron micrograph of the 26S proteasome of Drosophila melanogaster, the circular objects are the proteasome viewed head on, and the oblong objects are the proteasome viewed from the side (Walz et al. 1998), and a schematic of the 26S proteasome adapted from (Sullivan et al. 2003).

The proteasome has already proven to be a legitimate therapeutic target with bortezomib, a first generation reversible proteasome inhibitor that binds two of the 20S proteasome subunits and inhibits chymotrypsin and caspase-like activity (Dou & Zonder 2014). This proteasome inhibitor is approved for use as a treatment against multiple myeloma and mantle cell lymphoma, and causes growth suppression and apoptosis through altered regulation of NFκB, Bcl-2 proteins and p53 (Dou & Zonder 2014). There are limitations to bortezomib treatment including drug resistance, poor activity against solid-tumours and off-target cytotoxicity; subsequently, a range of second and third generation proteasome inhibitors are at various stages of development to overcome these problems (Dou & Zonder 2014). One of the complications with this sort of treatment is the broad range of proteins affected by proteasome inhibition which could have adverse and unpredictable effects.

1.2.2 Ubiquitin

Ubiquitin is a 76 amino acid protein that has an evolutionarily conserved tight fold (Figure 1.2.3), with a high percentage of hydrogen bonding which contributes to its

resistance against digestion, and its stability over broad pH and temperature ranges (Vijay-kumar et al. 1987). The C-terminal glycine of ubiquitin is covalently attached to the lysine amide group of the target substrate forming an isopeptide bond (Hershko et al. 1981). Polymerisation of ubiquitin occurs at one of seven lysine residues across the surface of the protein. Chain formation at different lysine residues can result in open or closed conformations that regulate ubiquitin polymer recognition by ubiquitin-interacting proteins (Ye et al. 2012). Subsequently, polymerisation at different lysines is associated with different substrate fates including DNA repair, trafficking, kinase modification, lysosomal and proteasomal degradation (Komander 2009).

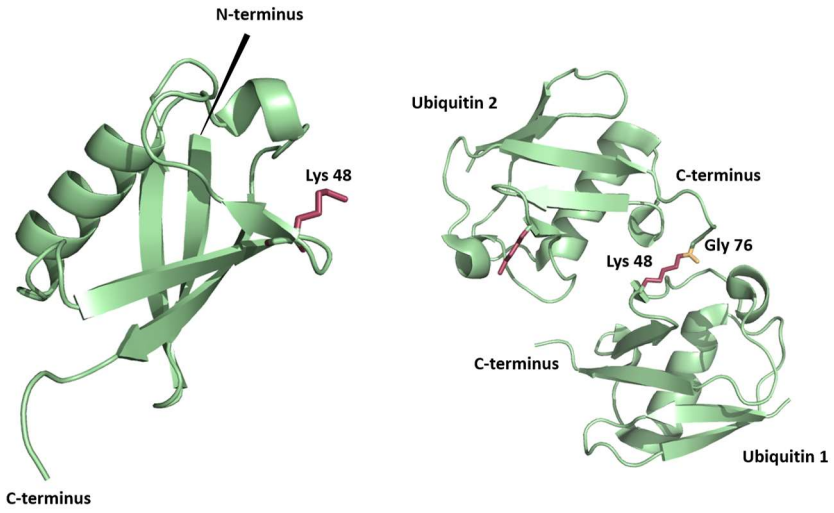


Figure 1.2.3 - The structures of monomeric ubiquitin (PDB: 1UBQ) and lysine 48-linked diubiquitin (PDB: 3AUL), with lysine 48 (shown in red) covalently bonded to the C-terminal glycine (shown in orange) of a second ubiquitin molecule (Vijay-kumar et al. 1987; Hirano et al. 2011). Image generated using PyMol molecular graphics software (Delano 2002).

Degradation at the proteasome is associated with polymerisation at lysine 48 (Figure 1.2.3). At least 4 ubiquitin units are required for proteasomal recognition, this increases proteasomal affinity 100-fold over ubiquitination of only 2 units, while increasing the number of ubiquitin units to 8 only increases the affinity a further 6.6-fold (Thrower et al. 2000). This non-linear pattern suggests the increase in affinity is not the result of an increase in ubiquitin concentration but is the result of a conformation that only occurs when at least 4 units are present, likely to involve the position of ubiquitin surface hydrophobic residues important in proteasome binding (Thrower et al. 2000; Pickart 2000; Beal et al. 1996).

1.3 E1 Ubiquitin Activator

The first step in the conjugation of ubiquitin to target substrates is the activation of ubiquitin by the E1 ubiquitin activator (UBA1). In contrast to the enzymes in the rest of the ubiquitin conjugation cascade, there is only one gene in the human genome that codes for the ubiquitin activator. The *uba1* gene codes for two isoforms of UBA1 called UBA1a and UBA1b which are 1058 and 1018 amino acids long, respectively. UBA1b is missing 40 amino acids at the N-terminal due to translation at an alternative initiation codon, and as a result UBA1b localises predominantly in the cytosol while UBA1a localises to the nucleus (Shang et al. 2001). This is probably a significant feature during cell-cycle progression where localisation changes of UBA1 are observed depending on which phase the cell is in (Grenfell et al. 1994).

Missense and synonymous mutations in the E1 gene are associated with X-linked infantile spinal muscular atrophy, with the synonymous mutation causing a reduction in expression through an altered DNA methylation pattern (Ramser et al. 2008). Loss of function mutations of the E1 in *Drosophila* result in motor impairment and reduced lifespan, and outline the significance of impairment of the UPS in disease (Liu & Pfleger 2013).

UBA1 contains one active, and one inactive adenylation domain (AAD and IAD), a first and second catalytic-cysteine half-domain (FCCH and SCCH), a four helix bundle domain (4HD) and a C-terminal ubiquitin fold domain (UFD). The structure, bound to ubiquitin, is shown in Figure 1.3.1, solved by X-ray diffraction (Lee & Schindelin 2008). The FCCH and the four helix bundle are inserted in to the middle of the IAD sequence at the N-terminal, the AAD follows with the SCCH insert and the UFD is at the C-terminal (Lee & Schindelin 2008). The AAD, FCCH and SCCH form a large canyon which recruits ubiquitin, burying 33% of its surface area by hydrophobic interactions and hydrogen bonds at the AAD and FCCH (Figure 1.3.1) (Lee & Schindelin 2008).

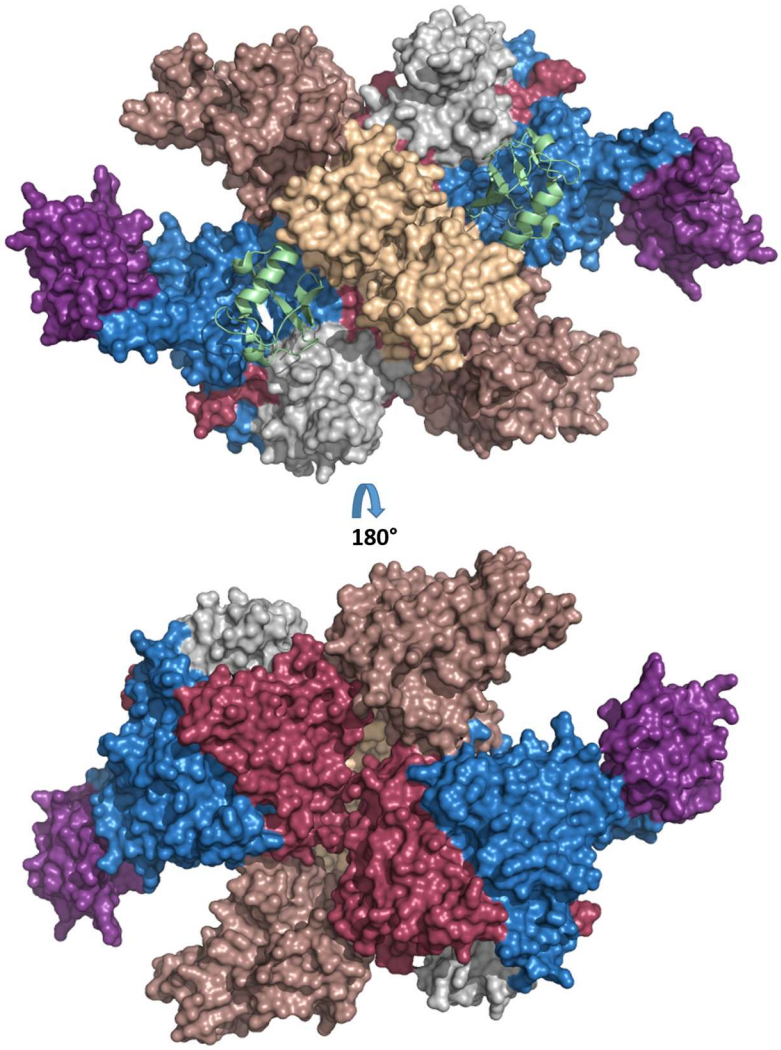


Figure 1.3.1 - Structure of the UBA1 dimer bound to ubiquitin (from Saccharomyces cerevisae) (PDB: 3CMM) (Lee & Schindelin 2008). The UFD domain is shown in purple, AAD shown in blue, IAD shown in red, FCCH domain shown in beige, SCCH domain shown in brown, 4HD domain shown in grey and ubiquitin shown in green using the ribbon graphic. Image generated using PyMol.

Upon association of the E1 activator with ubiquitin, the active adenylation domain attaches AMP to the carboxyl group of the ubiquitin C-terminal glycine by ATP hydrolysis, releasing pyrophosphate (Haas & Rose 1982). Significant conformational changes then occur in which the SCCH domain is rotated 130 degrees and the active site is remodelled, bringing the catalytic cysteine adjacent to the active site which now contains side chains required for thioester bond formation (Olsen et al. 2010). The cysteine SH group at the SCCH attacks the adenylated ubiquitin carboxyl terminus which forms a thioester bond and releases AMP (Haas & Rose 1982; Lee & Schindelin 2008). After a second ubiquitin

molecule is adenylated by the E1, the E2 conjugator is recruited to the E1 enzyme by electrostatic interactions with the ubiquitin fold domain, shown in Figure 1.3.2 (Haas & Rose 1982; Lee & Schindelin 2008; Olsen & Lima 2013).

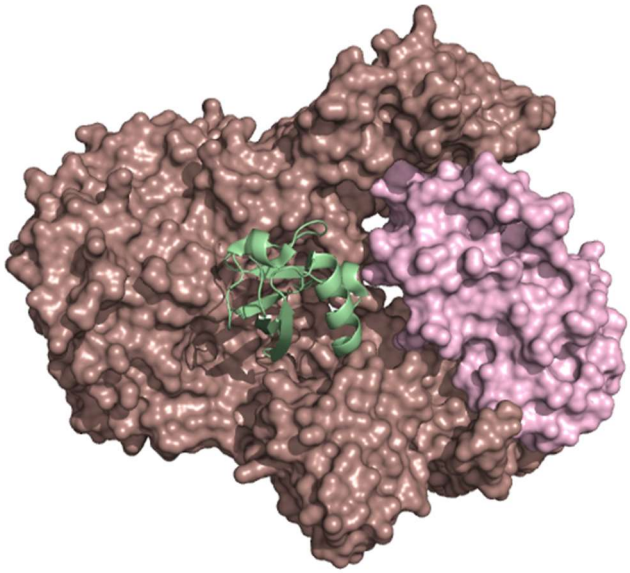


Figure 1.3.2 - The structure of monomeric UBA1 (brown), covalently bound to ubiquitin (green ribbon graphic) with the recruited E2 enzyme, Ubc4 (pink) (PDB: 4II3) (Olsen & Lima 2013). Image generated using PyMol.

A relatively mild 25-40 degree rotation around the flexible hinge between the AAD and UFD bring the catalytic E1 and E2 cysteines to within 8 Å (Lee & Schindelin 2008; Olsen & Lima 2013). The thioester bound ubiquitin is then transferred to the active site cysteine of the E2 by transthioesterification (Haas & Rose 1982).

1.4 E2 Ubiquitin Conjugators

The ubiquitin E1 sits at the top of the cascade, distributing ubiquitin to the rest of the cascade with downstream protein regulation limited to localisation to the nucleus or cytoplasm depending on which isoform is preferentially expressed. The second group of enzymes are more numerous, comprised of roughly 35 active members in humans - although there are also inactive ubiquitin E2 variant (UEV) members that lack an active site cysteine, but have a regulatory role (Sancho et al. 1998; Ye & Rape 2009). E2 family members are characterised by the presence of a ubiquitin-conjugating domain (UBC)

containing the active site cysteine. A 150 residue domain, UBCs across the E2 family have a high level of sequence and structural homology consisting of four α-helices and a short 3_{10}-helix on one face, and a four stranded antiparallel β-sheet on the opposite face (Hamilton et al. 2001). Most of the E2 is comprised of the UBC which carries out the main function of the conjugator, whilst the presence of N-terminal and C-terminal tails, and insertions in to the UBC often have regulatory roles (Ye & Rape 2009). The catalytic cysteine is found in a groove surrounded by conserved residues important for thioester and isopeptide bond formation (Wenzel et al. 2011). Figure 1.4.1 shows the structure of an E2 enzyme bound to the C-terminal glycine of ubiquitin (Middleton & Day 2015).

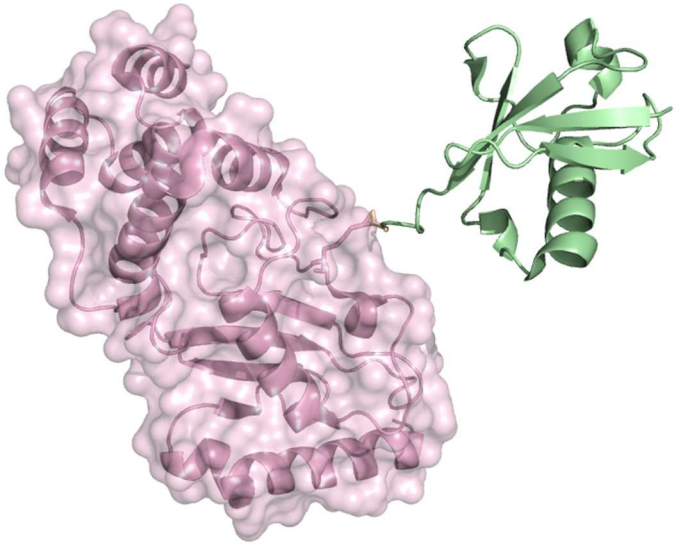


Figure 1.4.1 - The structure of the E2, Ube2K (in pink), bound to ubiquitin (green ribbon), solved using X-ray diffraction (PDB: 5DFL) (Middleton & Day 2015). Ube2K has the UBC domain and a C-terminal bundle of three helices. Image generated using PyMol.

As well as interacting with the E1 in order to be loaded with ubiquitin, each E2 can interact with a number of E3 enzymes. The E2 enzymes represent a second layer of regulation in the UPS with each E2 restricting ubiquitination of downstream substrates because of their limited choice of E3 binding partners. E2 regulation ranges from changes in expression level, to phosphorylation and autoubiquitination, and E2s are central in determining the type and length of ubiquitin chain formed on the substrate (Pickart & Eddins 2004; Banka et al. 2015; Ye & Rape 2009; van Wijk & Timmers 2010).

Once a substrate is ubiquitinated the switch must be made between initiation to elongation. There is evidence to suggest this switch depends on the E2 associated with the E3, since some E2s lack the ability to initiate ubiquitination and others lack the ability

to elongate from an initiating ubiquitin monomer (Rodrigo-Brenni & Morgan 2007; Ye & Rape 2009; Christensen et al. 2007). E2s also have the ability to enhance the rate of ubiquitin chain formation by forming complexes of pre-loaded E2s, increasing the availability of ubiquitin at the site of polymerisation and lowering the dependency on recharging by the E1 (Brzovic et al. 2006; Ye & Rape 2009). Other E2s pre-assemble ubiquitin chains on an E2-bound ubiquitin before transferring the chain to its protein target (Li et al. 2007; Ye & Rape 2009).

E2s can form chains composed of specific ubiquitin lysine linkages by orientating the target ubiquitin in such a way so as to expose the appropriate lysine to the active site, in order to be polymerised with the incoming ubiquitin. An inactive UEV protein has been shown to form a heterodimer with an active E2 in order to orient the acceptor ubiquitin in such a way as to only allow Lys63 polyubiquitination by the active conjugator (Eddins et al. 2006; Ye & Rape 2009). While the canonical proteasomal degradation lysine linkage Lys48 has been shown to be governed by the interaction of acceptor ubiquitin with an acidic loop of the E2, and also by an extension at the UBC domain C-terminal (Haldeman et al. 1997; Ye & Rape 2009).

The role E2s play in chain assembly is particularly relevant for the substrates of one particular group of E3 enzymes called RING (Really Interesting New Gene) finger domain E3s, whereby the E3 has a substrate selection role and facilitates the handover of the ubiquitin molecule from the E2 active site to the target lysine, and may be of less significance in HECT domain E3s (Ye & Rape 2009; Kim & Huibregtse 2009). Unlike RING finger E3 ligases, HECT domain E3 ligases contain an active site cysteine, accept activated ubiquitin directly, and can assemble ubiquitin chains independent of E2s.

1.5 E3 Ubiquitin Ligases

While the E1 puts energy in to the ubiquitin-transfer system, and the E2s regulate the type of ubiquitin chain formed, the E3 ligases are responsible for specificity in the system by selecting which substrates are ubiquitinated. Since there are such a large number of diverse targets that require ubiquitination, the number of genes encoding E3 ligases is much larger than the rest of the cascade, with estimates currently in the six hundreds. There are two main families of E3 ligase, the RING finger E3 ligase family and

the HECT E3 ligase family, and one smaller group of E3s called U-box E3 ligases that are related to the RING E3s. HECT and RING ligases have a distinct mechanism of action governed by their structure, but both specifically select their substrates through a number of different types of protein interaction domain.

1.5.1 RING Domain E3 Ligases

The RING domain family E3 ligases are named after the presence of the RING motif in their amino acid sequence. This motif is found in the genome over 600 times, which makes this the largest family of E3s. The RING motif consists of a series of seven cysteines, one histidine and conserved hydrophobic residues found at specific intervals in a sequence about 40-60 amino acids long. These residues bind two zinc ions at a ratio of four to one, with the first and third pair of residues in the sequence engaged in binding one ion, and the second and fourth pair engaged in binding another, in what is called a 'cross-brace' formation, shown in Figure 1.5.1 (Borden & Freemont 1996). Subgroups of the RING family have other residues in place of cysteines at some positions, but maintain the same pattern of zinc binding (Jackson et al. 2000; Borden & Freemont 1996). The related U-box domain family hold the same conformation as the zinc finger of the RING domain but they lack the ability to coordinate zinc ions and instead form salt bridges to stabilise the structure. Despite the abundance of cysteines, RING E3 ligases lack a catalytic cysteine, do not covalently bind ubiquitin, and are dependent on E2 heterodimerisation in order to transfer ubiquitin to target proteins, as such they can be considered to be a type of adapter protein.

Both the RING motif and the surrounding domains of RING E3s are structurally diverse (Borden & Freemont 1996), a feature which is necessary for the broad range of substrate targets and E2 interactions. It is common for RING E3s to form homodimers and heterodimers in order to carry out their activity, and the formation of heterodimers is an important step for some E3s to carry out their function, particularly for those that lack intrinsic E2-binding capacity (Metzger et al. 2014).

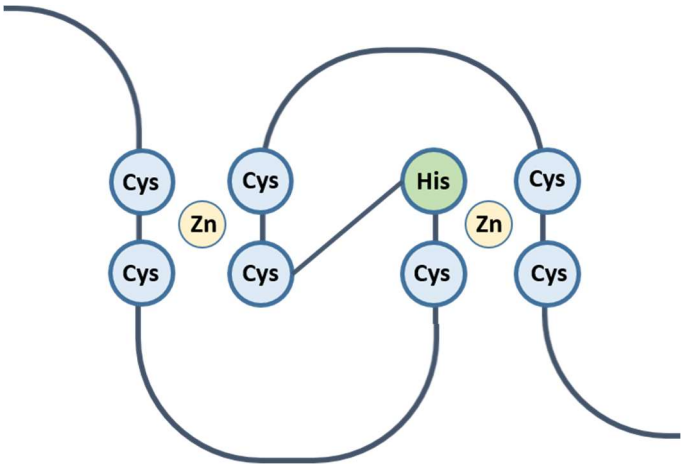


Figure 1.5.1 - A graphical representation of the RING finger domain cross-brace formation, with the first, second and fifth, sixth ion-coordinating residues (cysteines) bound to one Zn^{2+} atom, and the third, fourth and seventh, eighth ion-coordinating residues (three cysteines and a histidine) bound to a second Zn^{2+} atom. Adapted from (Jackson et al. 2000; Borden & Freemont 1996).

Interactions between RING E3s and their E2 enzymes are governed largely by contacts between a shallow groove formed by the RING motif and two loops of the E2 UBC domain, although interactions outside of the RING and UBC domains do occur and can cooperate to enhance specificity and affinity (van Wijk & Timmers 2010; Pickart 2000; Zheng et al. 2000). Certain residues of the E2 and E3 interaction interface have varying importance between different E2-E3 pairs. The interaction of C-terminus of Hsc70 interacting protein (CHIP) E3 ligase with UbcH5a requires phenylalanine 62 of the E2, whereas its alternative E2 binding partner Ubc13 has a methionine 64 in the analogous position which is not significant in the interaction mechanism, but the same residue is important in interactions with TRAF6 and Rad5 E3s in yeast (Wenzel et al. 2011).

A structure of the E2 UbcH5a, ubiquitin and the RING E3 ligase RNF4 in complex shows a heterotrimeric dimer (Figure 1.5.2), in which two RING E3 ligases are dimerised, each interacting with one E2 which is covalently bound to a ubiquitin molecule (Plechanovová et al. 2012). The complex shows minimal conformational change to the overall structures but a rearrangement of the residues in the active site, and the ubiquitin is pinned back on to the E2. Both of these changes prime the C-terminal tail of ubiquitin for attack by the lysine of a target substrate. The ubiquitin interacts with both RING domains of the dimer (Figure 1.5.2), and explains the need for homodimerisation of this E3 to allow for ubiquitinating activity (Plechanovová et al. 2012).

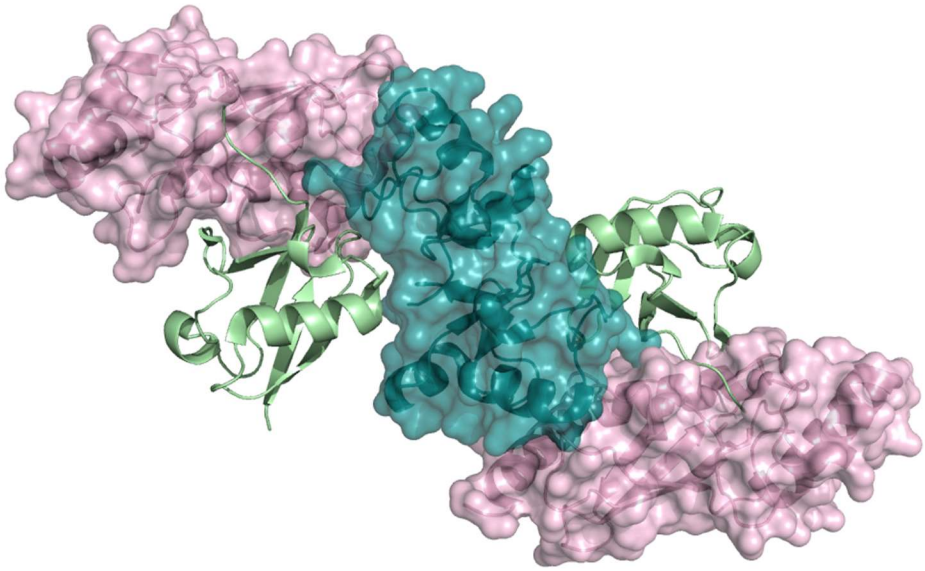


Figure 1.5.2 - The structure of the RNF4 dimer (teal), in complex with two UbcH5a E2 conjugators (pink), each bound to a ubiquitin monomer (green ribbon). Solved by X-ray diffraction (PDB: 4AP4) (Plechanovová et al. 2012). Image generated using PyMol.

In order to allow for transfer of the ubiquitin molecule to its target protein, the E2-E3 complex must recruit the substrate to the complex. To do this, the E3 must contain a protein interaction motif. The RING E3 c-Cbl interacts with its target substrate via an Src homology 2 (SH2) domain, which binds to phosphorylated tyrosine motifs. This allows c-Cbl to degrade activated receptor tyrosine kinases. Crystallisation of the c-Cbl SH2 and RING domain in complex with its E2 (UbcH7) and a peptide corresponding to the binding motif of the SH2 domain shows that the binding site of the substrate is some 60 Å from the active site cysteine, but that a deep channel is formed by the SH2 domain, RING domain and E2 that would allow for further interaction with the substrate in order to position it for ubiquitin ligation (Zheng et al. 2000).

1.5.2 HECT Domain E3 Ligases

The HECT domain E3 ligases are a smaller family of E3 enzymes, at around 30 members, and are named as such because of the presence of a conserved HECT domain in their amino acid sequence. The eponymous HECT family member, the E6AP E3 ligase, is implicated in the development of cervical carcinomas caused by the human papillomavirus (HPV). The viral protein E6 binds E6AP and causes polyubiquitination of

the tumour suppressor p53, which is not the physiological target of E6AP, causing its degradation (Huibregtse et al. 1994). Mutations that prevent expression of E6AP or render the HECT domain catalytically inactive are associated with the genetic disorder Angelman syndrome (Tomaić & Banks 2015).

The HECT domain is much larger than the RING motif at around 350 amino acids, which form two lobes and contains an E2 interaction interface (Figure 1.5.3) (Huibregtse et al. 1995). Unlike RING E3 ligases, HECT domain E3 ligases contain a catalytic cysteine and bind ubiquitin directly through a thioester bond (Scheffner 1995). HECT E3s are able to bind E2s, accept ubiquitin, bind substrates through protein interaction domains and catalyse ubiquitin transfer independent of E2 enzymes. They are generally large proteins with a C-terminal HECT domain that carries out ubiquitin transfer, and protein interaction and cell localisation domains found N-terminal to the HECT domain.

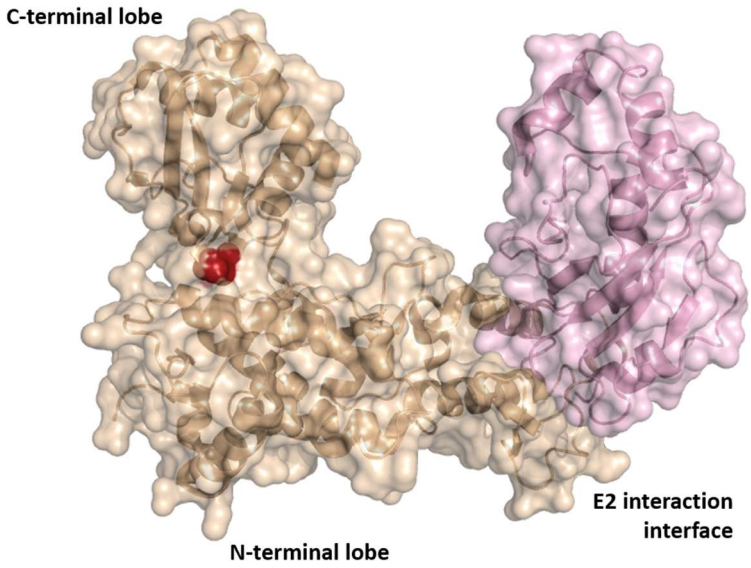


Figure 1.5.3 - The structure of the E6AP HECT domain (light brown), in complex with the E2, UbcH7 (pink). The HECT domain catalytic cysteine is shown in red. Solved using X-ray diffraction (PDB: 1C4Z) (Huang et al. 1999). Image generated using PyMol.

The HECT domain has two lobes, with the larger N-terminal lobe responsible for E2 interaction whilst the smaller C-terminal lobe harbours the ubiquitin binding cysteine and is responsible for catalysis (Huang et al. 1999). The crystal structure of E6AP in complex with one of its E2s UbcH7 (also a partner for the RING E3 c-Cbl), shows that the E2 interaction interface is found on a subdomain portion of the N-terminal lobe (Figure 1.5.3). A hydrophobic groove is created by two antiparallel helices on one side and a two

stranded antiparallel β-sheet on the other (Huang et al. 1999). Like RING motifs, this groove interacts with two loops and the N-terminal α-helix of the UBC domain (Huang et al. 1999). Phenylalanine 63, which is found at the first loop of the E2, binds the centre of the groove, making hydrophobic interactions with 6 HECT residue side chains (Huang et al. 1999). Phenylalanine at this position is conserved amongst HECT-binding E2 conjugators and directs specificity (Huang et al. 1999). The catalytic cysteine is found at the centre of a loop of the C-terminal lobe (shown in red in Figure 1.5.3) and in this crystal structure is 41 Å away from the catalytic cysteine of the E2 (Huang et al. 1999). Despite this distance, the cysteine must be able to come in to close proximity with the thioester bond between the E2 cysteine and the ubiquitin - which is missing in this model.

Crystallisation of another HECT domain, this time from WWP1, shows the familiar bilobal structure with some helical and β-sheet insertions into each domain. The C-terminal lobe however, is in a dramatically different position (Figure 1.5.4) (Verdecia et al. 2003). A horizontal tilt of 30°, and a rotation of the C-terminal lobe by 100° around a 4 residue hinge loop that joins the two lobes together, means that this structure puts the catalytic cysteine 16 Å away from a theoretical E2 cysteine (when modelled based on the E6AP costructure) (Verdecia et al. 2003). While this modelled distance is closer to a realistic minimum distance required for an energetically favourable reaction to occur, it is still further than it should be. Another HECT crystal structure, from SMURF2 (Smad Specific E3 Ubiquitin Protein Ligase 2) shows the C-terminal lobe to be in a different orientation at 50 Å away from a modelled binding site (Ogunjimi et al. 2005). Analysis of the flexible hinge that links the two lobes shows that a rotation is permitted by the dihedral angle restraints which brings the active site cysteine to within 5 Å of the modelled E2 cysteine (Verdecia et al. 2003). Mutational analysis aimed at either removing the hinge, or reducing its rotational freedom, resulted in a significantly reduced ability of WWP1 to ubiquitinate substrates (Verdecia et al. 2003).

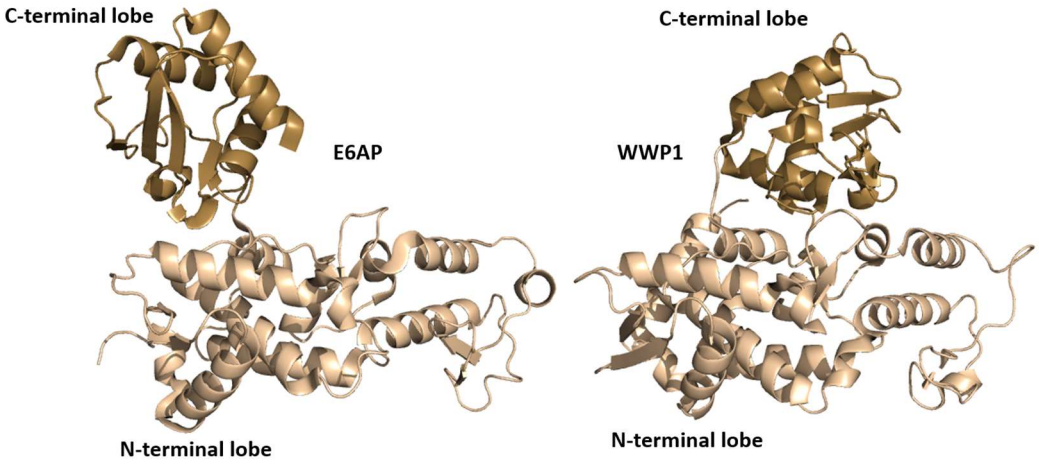


Figure 1.5.4 - A comparison of the position of the C-terminal lobe (dark brown) in the E6AP (PDB: 1C4Z) and WWP1 (PDB: 1ND7) HECT domain structures. Solved by X-ray diffraction (Huang et al. 1999; Verdecia et al. 2003). Image generated using PyMol.

Another HECT domain from NEDD4L (NEDD4-Like, also known as NEDD4-2) was crystallised with one of its E2s, UbcH5B, bound to ubiquitin (Kamadurai et al. 2009). The E3 carries a mutation of the catalytic cysteine, substituting in a serine or an alanine, which prevents the HECT domain from attacking the thioester bond of the E2. This allowed observation of the HECT domain in a conformational position ready to accept the ubiquitin group. In this structure, the E2-bound ubiquitin C-terminal is sandwiched between the active sites of the E2 and E3. The position of the residue substituted for the E3 active site cysteine is closer to the ubiquitin C-terminal, but still slightly too far at 8 Å. This position is achieved by rotation of the C-terminal lobe, which interacts directly with the ubiquitin through hydrophobic and electrostatic interactions (Kamadurai et al. 2009). Mutational analysis shows these interactions to be important for the transfer of ubiquitin by allowing correct orientation of the C-terminal lobe. Although the distance here is larger than required distance for transthioesterification, a 4° rotation of the C-terminal lobe brings the catalytic residue within range, a rotation which would be complemented by interactions between residues of the C-terminal lobe and the E2 (Kamadurai et al. 2009). This is supported by mutational analysis of these residues, which result in reduced ubiquitin transfer activity of NEDD4L (Kamadurai et al. 2009).

The type of lysine linkage formed when HECT E3s ubiquitinate substrates has been shown to be entirely dependent on the HECT E3 and, unlike RING E3s, the E2 plays no part (Kim & Huibregtse 2009). A series of assays using chimeric HECT domains showed that chain specificity of HECT E3 ligases depends entirely on the last 60 amino acids of the

HECT domain (Kim & Huibregtse 2009). This region encompasses the C-terminal portion of the C-terminal lobe, and includes the catalytic loop, three β-strands and an α-helix.

A crystal structure of the truncated HECT domain yeast E3 ligase Rsp5 in complex with ubiquitin, and a ligand peptide, showed that a three-way interaction between the two lobes of the HECT domain and ubiquitin creates a catalytic surface where ubiquitin is sandwiched between both lobes (Kamadurai et al. 2013). In conjunction with the crystal structure, alanine scanning mutagenesis and ubiquitin ligation assays showed that this interaction between the two lobes is essential for ubiquitin transfer by orientating the E3-ubiquitin thioester bond towards the substrate lysine.

In order to best understand the ubiquitin-proteasome system it has been necessary to determine the mechanisms by which ubiquitin is transferred along its cascade. While there are still gaps in our knowledge, great strides have been made. The greater the knowledge of the mechanisms, the better therapeutics can be designed and refined to target the UPS in disease states. Each step of the pathway represents a putative target, the earlier in the cascade the target, the broader the effects. Bortezomib, discussed above, targets the proteasome and subsequently has adverse side effects through off target effects in neural cells. Cancer therapies, and proteasome inhibitors are no exception, often rely on increased sensitivity of a cancerous cell to treatment that will also kill normal cells, just more slowly. Established and emerging therapeutics exploit qualities such as increased protein expression, increased metabolism, increased demand for vasculature or overexpression of growth factor receptors. In order to increase drug tolerance, reduce off target effects and therefore increase the likelihood of a favourable outcome, therapeutics need to be as targeted as possible.

In the context of the UPS this means targeting the sharp end of the cascade, the E3s, which select which proteins are degraded. Discussed above are the mechanism by which ubiquitin is transferred to the HECT domain from its E2 or from the E2 to the substrate by RING motif E3s. Targeting the E2 binding interfaces or the catalytic domains of E3s has potential, but because of the conserved nature of the RING and HECT domains, and also the redundancy and plurality amongst E2-E3 partners, designing an effective therapeutic that affects one E3 and not another becomes a challenge. In order to achieve this, the focus should be on the way in which E3s interact with their substrates. The protein interaction domains, outside of the HECT and RING domains, are the key to the

specific interactions with the diverse array of substrates in the UPS. Protein interaction domains in E3s range from phosphotyrosine-binding SH2 domains, leucine-rich repeat and WD-40 motifs, to tryptophan-tryptophan (WW) domains (VanDemark & Hill 2002). The E3 studied here, WWP2, belongs to one subset of HECT E3 ligases called the NEDD4 family, which all contain one type of protein-protein interaction domain; the WW domain. Many of these ligases have been linked to regulation of the TGFβ signalling pathway, which will be explored in further detail in the following section.

1.6 TGFβ signalling

1.6.1 Latent TGFβ

The transforming growth factor β (TGFβ) peptides are responsible for triggering the TGFβ signalling cascade. They constitute one subfamily of the TGFβ superfamily, a closely related group of highly conserved cell regulatory proteins that are important in regulating cell division, differentiation, homeostasis and a variety of different cellular characteristics. The TGFβ signalling peptides, of which there are three known variants (TGFβ1, TGFβ2, TGFβ3), are produced as latent propeptide homodimers with N-terminal latency-associated peptide (LAP) domains (Gentry & Nash 1990). The LAP domain homodimer is cleaved in the Golgi apparatus by furin proprotein convertase, but remains tightly bound to the TGFβ homodimer through non-covalent interactions, which renders the cytokine inactive (Gentry & Nash 1990; Dubois et al. 1995; Leitlein et al. 2001; Annes et al. 2003). The LAP domain forms disulphide bridges with latent TGFβ-binding proteins (LTBP), which forms a complex called the large latent complex (LLC) (Miyazono et al. 1991; Saharinen et al. 1996). The LLC is secreted from the cell and localises to components of the extracellular matrix (ECM), limiting the bioavailability of the cytokine (Nunes et al. 1997; Zilberberg et al. 2012). Here the LLC complex senses the extracellular environment, releasing the mature TGFβ signalling molecule in response to a variety of different stimuli, including: proteolytic cleavage by plasmin and matrix metalloproteinases (MMP) 2 and 9, reactive oxygen species (ROS), protein-protein interactions with integrins and thrombospondin-1, and environmental factors such as pH and temperature (Lawrence et al. 1985; Sato & Rifkin 1989; Lyons et al. 1990; Schultz-Cherry & Murphy-Ullrich 1993). These stimuli act on the LAP domain, diminishing the tight non-covalent interaction with

TGFβ, causing the release of the mature signalling molecule, which diffuses to the cell surface and binds to its receptors to initiate the signalling cascade.

1.6.2 TGFβ receptors

TGFβ induces intracellular signalling by binding to a complex of two different types of serine/threonine kinase receptors. These are TGFβ receptor type I (TβR-I) and TGFβ receptor type II (TβR-II) receptors. TGFβ binds homodimeric pairs of type II receptors, which then recruit homodimeric pairs of type I receptors to form a heterotetrameric complex (Lin et al. 1995; Lin et al. 1992; Groppe et al. 2008). The serine/threonine kinase activity of the intracellular domain of TβR-II is constitutively active, and, upon dimerisation, transphosphorylates several TβR-I glycine-serine motifs (Wrana et al. 1994; Chen & Weinberg 1995). This activates the serine/threonine kinase activity of the type-I receptor and causes the recruitment of the Smad intracellular signalling effectors.

1.6.3 Smads 2/3

There are two Smads that convey signalling from the receptor complex to the nucleus, these are Smad2 and Smad3, also known as the receptor Smads (r-Smads). A schematic of the signalling cascade is shown in Figure 1.6.1. The r-Smads share a high level of sequence identity, and each have two Mad-homology domains (MH1 and MH2) separated by a linker region that contains a proline rich motif that bind WW domains. The MH domains are autoinhibitory, ensuring that r-Smad activity is restricted in the absence of TGFβ stimulation. The Smad anchor for receptor activation (SARA) protein shuttles Smad2/3 to the activated receptor, where they are phosphorylated at their C-terminal SSxS motifs by the activated TGFβ receptor complexes (Tsukazaki et al. 1998). This relinquishes the autoinhibition of the two MH domains, and allows Smad2/3 to form heterotrimers with Smad4 (otherwise known as the common Smad). Smads translocate to the nucleus and bind to DNA CAGA boxes called Smad binding elements (SBE). Transcriptional upregulation of genes local to these CAGA boxes is achieved by association

with numerous transcriptional cofactors and activators, while association of co-repressors repress the gene of interest.

Smad-dependent gene regulation results in a culmination of events that drives the TGFβ response. The cellular responses to TGFβ are diverse, and include growth inhibition, angiogenesis and epithelial to mesenchymal transition (EMT). These TGFβ signalling gene programs have obvious significance in the growth and metastasis of cancer cells. In some cancers the signalling pathway is switched off altogether, in order to overcome the tumour suppressive properties, while in others, specific mutations or overexpression of binding partners is used to circumvent these properties, while capitalising on the tumour promoting properties.

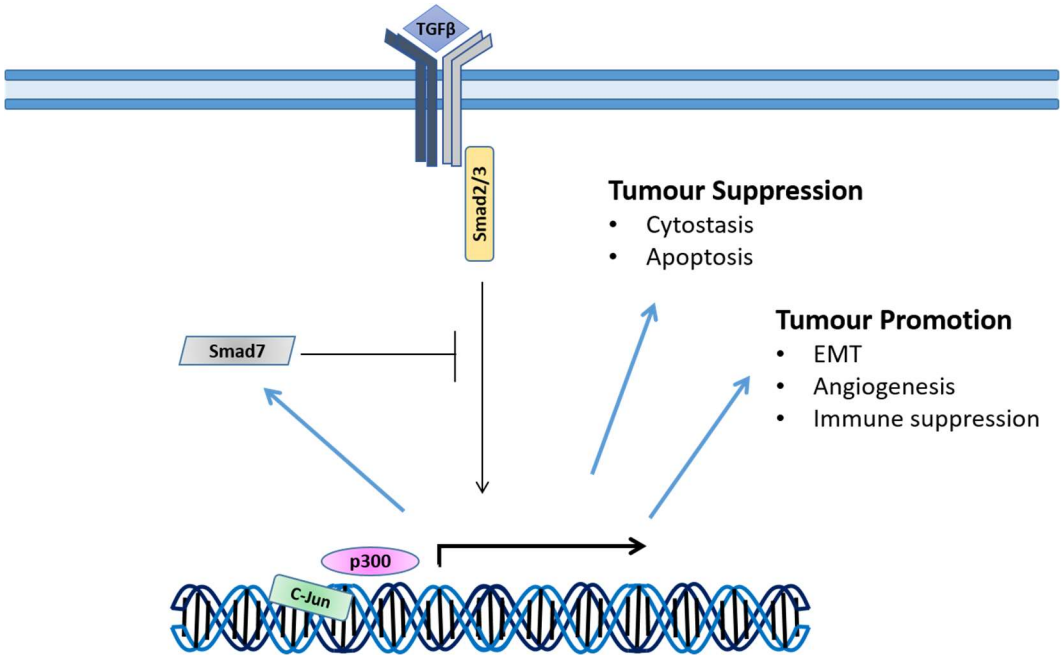


Figure 1.6.1 - A schematic showing elements of the TGFβ signalling cascade. TGFβ is bound to its receptors at the cell surface, the recruited Smad2/3 is activated and binds transcriptional co-activators at gene promoters (p300 and c-Jun are used as examples), and TGFβ-dependent gene programs are upregulated. The Smad7 inhibitor is upregulated and acts as a negative feedback mechanism.

1.6.4 TGFβ gene programs and Cancer

TGFβ-mediated cytostasis is driven by the upregulation of cell cycle inhibitors $p15^{ink4B}$, $p21^{cip1}$ and $p57^{kip2}$ (Hannon & Beach 1994; Datto et al. 1995; Scandura et al. 2004; Heldin et al. 2009). Upregulation of these proteins induces cell cycle arrest at the G1 cell cycle check point by inhibiting cyclin/CDK (cyclin dependent kinase) activity. Smads also exert inhibitory control over the promoter of the cell growth enhancer, and oncogene, c-Myc (Gomis et al. 2006). Repression of this promoter is enabled by association of Smads with the CCAAT/enhancer-binding protein-β (C/EPBβ), and this transcription factor is also indispensable in the upregulation of p15 (Gomis et al. 2006). However, a variable transcript of the C/EBPβ gene generates the dominant-negative inhibitory isoform called liver-enriched inhibitory protein (LIP), which reduces the repression of c-Myc and inhibits the upregulation of p15. LIP has been shown to be overactive in metastatic breast cancer cells and is a potent inhibitor of the TGFβ cytostatic response (Gomis et al. 2006). Overexpression of LIP selectively overcomes the growth inhibitory response to TGFβ, and allows the cancer cell to maintain an active TGFβ pathway so as to benefit from its tumour promoting properties.

TGFβ aids in the process of angiogenesis by upregulating levels of vascular endothelial growth factor (VEGF) and connective-tissue growth factor (CTGF) which cooperate with hypoxia-inducible factor 1 (HIF-1) to cause the formation of new vasculature (Sánchez-Elsner et al. 2001). This process is significant in the rapidly growing tumour, where cells quickly become oxygen starved. This causes the upregulation of HIF-1 and when combined with TGFβ stimulation, results in the secretion of high levels of angiogenic factors (Sánchez-Elsner et al. 2001; Padua & Massagué 2009). This causes the upregulation and secretion of MMPs to reorganise the ECM, and endothelial cells migrate along the chemotactic gradient, which divide and vascularise the area.

In EMT, epithelial cells, that are polar, immotile and have high levels of cell-cell and cell-matrix adhesion junctions, transition in to mesenchymal cells, which lack polarity, intercellular junctions, are motile and exhibit stem cell-like properties. EMT involves the downregulation of cell-cell contacts, the reorganisation of the cellular cytoskeleton and motility components, and the secretion of cytokines, growth factors and ECM molecules. This results in the detachment of the cell from the epithelium, migration in to the mesenchyme, and the adoption of the fibroblast phenotype. The fibroblastic phenotype

is characterised by high levels of N-cadherin, vimentin, fibronectin and smooth muscle actin. TGFβ induces EMT by implementing a genetic program that involves a network of transcription factors. These transcription factors include bHLH (basic-helix-loop-helix) proteins Twist and E47, zinc-finger proteins Snail1 and Snail2, zinc-finger/homeobox domain proteins ZEB1, ZEB2 and forkhead protein FOXC2 (Peinado et al. 2003; Yang et al. 2004; Mani et al. 2007; Heldin et al. 2012). These act by regulating the E-cadherin promoter, epithelial splicing regulatory proteins (ESRPs), MMPs and platelet-derived growth factor receptor-α (PDGFRα), to induce the mesenchymal phenotype. In normal epithelial cells, EMT occurs at the G1 cell cycle arrest and is an essential process in embryonic development for the formation of different tissues and organs, and in wound repair for the migration of epithelia at wound margins. However, epithelial cell cancers (carcinomas) can use this mechanism for the purpose of tumour cell invasion and metastasis, and overcome G1 cell cycle arrest, allowing cancerous cells to continue proliferating as they do so (Heldin et al. 2012; Valcourt et al. 2005).

TGFβ also creates a microenvironment that is conducive to EMT by acting on cells in the surrounding stroma (Bierie & Moses 2006). The microenvironment is important in EMT and cancer, and the coordination of other signalling pathways including Wnt, Notch and growth factor mediated receptor tyrosine kinase signalling also have a role to play (Moustakas & Heldin 2007). TGFβ primes the microenvironment to allow the growth and metastasis of the tumour, not only by encouraging angiogenesis but by inducing paracrine signalling by stimulating surrounding stromal cells to secrete various cytokines, ECM components and proteases to stimulate and aid in the transition of the tumour cells through the stroma. Because of the potent oncogenic effects that TGFβ has, it is tightly controlled in healthy cells. The central mechanism of control is through the Smad7 inhibitory component.

1.6.5 Smad7

Smad7 is upregulated by Smad2 and Smad3 in response to TGFβ stimulation (Figure 1.6.1); this is achieved by direct upregulation of expression by Smad2/3 at the Smad7 promoter (Stopa et al. 2000). Smad7 forms a negative feedback loop by inhibiting TGFβ signalling, it does this by stably associating with the activated type I TGFβ receptor, and blocks the activating phosphorylation of Smad2/3 (Hayashi et al. 1997). Because of

its inhibitory role in this pathway, Smad7 expression is altered in a variety of cancers, and like TGFβ it can be either overactive or underactive, depending on context (Yan et al. 2009). Smad7 is related to Smad2/3, and like these two signalling molecules, also has a proline rich linker region that recruits WW domain-containing binding partners. Specifically, Smad7 associates with members of the NEDD4 family of E3 ubiquitin ligases. These proteins interact with Smad7 through their WW domains, and are recruited to the TGFβ receptor along with Smad7. Here, they use their HECT domains to polyubiquitinate Smad7 and the TGFβ receptors. Smad7 and the receptors are subsequently degraded and the pathway is switched off.

1.7 NEDD4 Ubiquitin E3 Ligases

There are nine members of the NEDD4 family of E3 ubiquitin ligases: NEDD4, NEDD4L, SMURF1, SMURF2, HECW1, HECW2, ITCH, WWP1 and WWP2 - although these proteins have been given a number of different names throughout the literature. NEDD4, NEDD4L, SMURF1, SMURF2, WWP1 and WWP2 are among these proteins confirmed to have altered expression or splicing in several malignancies including colorectal, breast, gastric, bladder, pancreatic, ovarian, melanoma and prostate cancers (Tanksley et al. 2013; Chen et al. 2007; Sun et al. 2014; Wang et al. 2007; Kwei et al. 2011; Kwon et al. 2013; Jung et al. 2014; Soond et al. 2013; Chen & Matesic 2007). The importance of these proteins in malignancies is down to the proteins that they bind and degrade. The tumour suppressor PTEN (phosphatase and tensin homolog), the oncogene OCT4 (Octamer-binding transcription factor 4), genome stability components, the Wnt pathway, the EGFR (epidermal growth factor receptor) pathway and the TGFβ pathway are among just a few of the targets of these E3 ligases (Tanksley et al. 2013; Kwon et al. 2013; Wang et al. 2007; Li et al. 2009; Blank et al. 2012; Maddika et al. 2011; Xu et al. 2004).

The structures of the NEDD4 family E3 ligases determine their function. Figure 1.7.1 shows the domain composition of each of the NEDD4 family members. They all share the same domain arrangement with a calcium-dependent phospholipid-localisation C2 domain found at the N-terminus, the ubiquitin ligating HECT domain at the C-terminus, and sandwiched in between is a sequence of between 2 and 4 WW protein interaction domains that determine which substrates they bind. SMURF1, HECW1 and HECW2 each

contain two WW domains, while SMURF2 contains three and NEDD4, NEDD4L, ITCH, WWP1, and WWP2 each contain four WW domains.

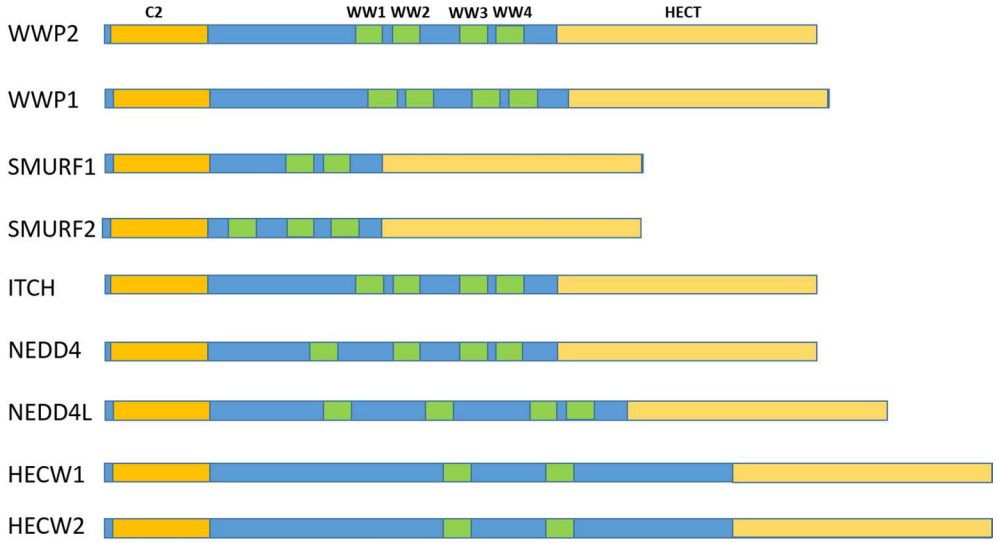


Figure 1.7.1 - A schematic representation of the NEDD4 E3 ligase family members, showing their domain composition. The N-terminal C2 domain is shown in orange, the WW domains are shown in green and the HECT domain is shown in yellow.

1.7.1 WW domains

WW domains are protein-protein interaction modules that bind proline-rich motifs and are found across many different unrelated proteins with many different functions. The classic WW domain motif is about 40 amino acids long, forming a three stranded antiparallel β-sheet with two characteristic tryptophans found 20-22 amino acids apart - although some WW domains only have one tryptophan and instead have an alternative hydrophobic residue at the same position. A pair of hydrophobic residues that can be either tyrosines or phenylanines or a combination, are highly conserved between the two tryptophans as part of the second β-strand, and a proline is found three residues C-terminal to the second tryptophan (Staub & Rotin 1996). The WW domain β-sheet tends to be twisted and curves inwards on the surface upon which proline-rich ligands bind. A group of hydrophobic residues that includes the N-terminal tryptophan, the conserved proline and the second residue of the hydrophobic pair on the second β-strand form a structural hydrophobic cluster on the opposite surface.

Four groups of WW domains have been defined based on their proline-rich binding motifs. Group I domains bind a PPxY motif, where x is any amino acid, Group II domains bind PPLP sequences, group III domains bind polyproline-arginine motifs, and group IV domains bind phosphorylated serines or threonines followed by a proline (pSP or pTP) (Bedford et al. 2000; Lu et al. 1999; Bedford et al. 1997; Espanel & Sudol 1999; Ingham et al. 2005). Although, a 3 group system has been suggested based on the similarities between the proteins they interact with, these groups also show a preference for the motifs described above, however in this system group II and group III merge (Ingham et al. 2005). Two WW domains of the same protein can both interact with a common binding partner but are also able to be selective, distinguishing between and preferentially binding other proteins (Ingham et al. 2005). Proline-rich motifs can be recognised by different WW domain groups through different sequence combinations (Ingham et al. 2005).

Ligands with proline repeats form polyproline II (PPII) helical conformations and are bound by a hydrophobic patch on the surface of the WW domain which consists of the first residue of the conserved hydrophobic pair, found on the second β-strand as discussed above, and the C-terminal tryptophan (Macias et al. 1996; Zarrinpar & Lim 2000). The conserved tyrosine/phenylalanine and the C-terminal conserved tryptophan pack together in an almost parallel conformation forming two ridges and a groove between the aromatic side chains called the XP binding groove, so called because the ridge allows for neat insertion of the PPII helix proline side chain in to the hydrophobic pocket, plus another residue X, which in PPxY motifs is the first proline (Zarrinpar & Lim 2000). This conformation actually allows for proline rich motifs to bind in two different orientations, due to conformational and hydrogen bond symmetry in the ligand and the grooves used to recognise it, and may allow for alternative motif recognition (Zarrinpar & Lim 2000).

A WW domain of the Smad-binding YAP1 (Yes kinase-associated protein 1) binds the PPxY motif. The WW domain residues involved in binding are the conserved hydrophobic tyrosine at position 188, a leucine at position 190, a histidine at position 192 and residues 194-199 which are one aspartic acid, one glutamine, three threonines and the final tryptophan (Macias et al. 1996). The YAP1 binding residues are shown in Figure 1.7.2.

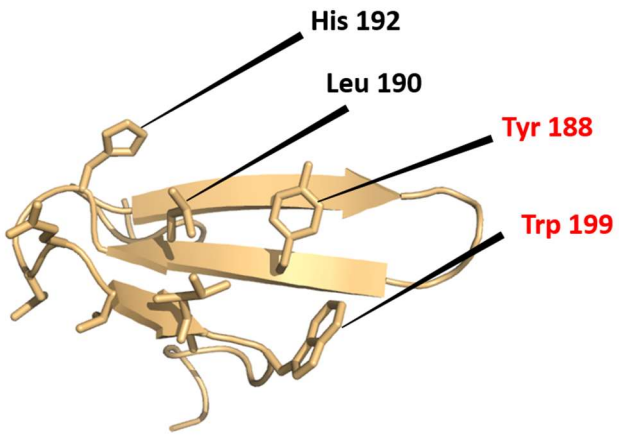


Figure 1.7.2 - The structure of the YAP1 WW domain, with the side chains of the residues involved in binding the PPxY motif shown. Tyrosine 188 and tryptophan 199 constitute the XP binding pocket and are labelled in red. Note, this structure was solved as a co-structure, bound to a Smad7 peptide, which has been removed for the purpose of this illustration. Structure solved by solution NMR (PDB: 2LTW) (Aragón et al. 2012).

The C-terminal tryptophan at position 199 forms the XP groove with tyrosine 188, and interacts with the first and second proline of the PPxY motif. The x residue in this instance is another proline, the side chain is facing outwards from the binding interface and the backbone carbonyl hydrogen bonds the hydroxyl group of tyrosine 188. The Y residue of the PPxY motif is facing towards the interaction interface and interacts with leucine 190 and histidine 192 and may form a hydrogen bond with either histidine 192 or glutamine 195 (Macias et al. 1996).

Leucine 190 and histidine 192 constitute a specificity pocket, conferring selectivity for the tyrosine of the PPxY motif (Zarrinpar & Lim 2000). Mutation of the tyrosine-binding leucine 190 to a tryptophan is sufficient to switch the affinity of this group I WW domain to a group II recognition motif PPLP, which is further enhanced by substitution of the second tyrosine binding residue, histidine 192 for a glycine (Espanel & Sudol 1999). Whereas mutation of the glutamine at position 195 was insufficient to change motif preference, placing leucine 190 as the most significant determinant of motif preference between group I and group II WW domains (Espanel & Sudol 1999).

In a Pin1 (Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1) group IV pSP or pTP binding WW domain, again the C-terminal tryptophan (34) and the tyrosine (23) on the second β-strand (shown in Figure 1.7.3), corresponding to tyrosine 188 above, accommodate the proline residue (Verdecia et al. 2000). The tyrosine 23 hydroxyl group

hydrogen bonds the ligand phosphoserine phosphate group through a water molecule. The phosphate group also hydrogen bonds serine 16 and arginine 17 side chains, as well as the arginine backbone amino group, which form part of the first β-strand and the start of the loop to the second β-strand (Verdecia et al. 2000).

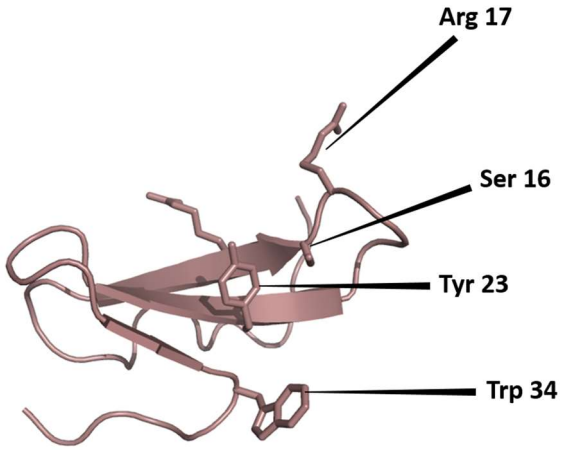


Figure 1.7.3 - The Pin1 WW domain structure (PDB: 2M8I). The residues involved in binding its phospho-ligand are labelled. Different residues are significant in binding the group IV Pin1 substrates, when compared to group I binding by YAP1 above, although the XP pocket residues are still involved in binding. Solved by solution NMR (Luh et al. 2013).

1.7.2 NEDD4 family WW domains

NEDD4, the founding member of the NEDD4 E3 ligases, binds to polyproline sequences of the epithelial sodium channel (ENaC) via WW domain interactions, causing its polyubiquitination and degradation (Staub et al. 1997; Staub et al. 1996). The genetic hypertensive disorder Liddle's syndrome is caused by mutations of the polyproline sequence in ENaC, which reduce or abolish the affinity of NEDD4 for the receptor (Staub et al. 1996; Staub et al. 1997). Degradation of the sodium channel is inadequate and this results in poor regulation of blood sodium. The solution structure of the second and third WW domains of rat NEDD4 expressed in tandem shows that each domain holds a very similar structure to YAP1 (Kanelis et al. 1998). When analysing binding of an ENaC PPxY motif, different affinities across the NEDD4 WW domains were observed, with tightest binding by the third WW domain (Kanelis et al. 2001). The XP groove (phenylalanine 476 and tryptophan 487) accommodates the first two prolines of the motif, shown in Figure

1.7.4. As above, the x residue, asparagine 617, faces outwards and the tyrosine (618) residue is accommodated by an isoleucine (478) and histidine (480) that correspond to the leucine and histidine in YAP1 (Kanelis et al. 2001). Unlike some ligand interactions with WW domains, amino acids C-terminal to the PPxY motif of the ligand are also implicated in binding the WW domain. The peptide makes a helical turn after the canonical motif and binds to residues of the first β-strand, this contributes to the tight affinity of this domain for its ligand (Kanelis et al. 2006; Kanelis et al. 2001).

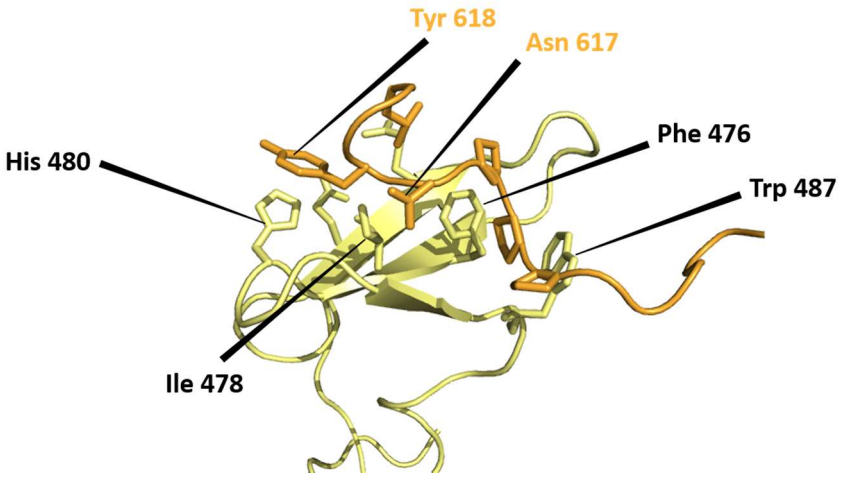


Figure 1.7.4 - The structure of rNEDD4 WW3 (yellow) bound to a PPxY motif-containing ENaC peptide (orange). Ligand residues are labelled in orange. Solved by solution NMR (PDB: 1I5H) (Kanelis et al. 2001).

SMURF2 targets receptors of the TGFβ signalling pathway for degradation. It achieves this by binding and remaining in complex with the negative regulator of the pathway Smad7, which allows for its nuclear export and recruitment to the active cell surface TGFβ receptors (Ogunjimi et al. 2005). Once there, polyubiquitination of the receptors causes their internalisation and degradation (Ogunjimi et al. 2005). Smad7 contains a PPPY motif in the middle of its sequence, to which the third WW domain of SMURF2 binds with 40μM affinity (Chong et al. 2006). The usually conserved C-terminal tryptophan is substituted for a phenylalanine at postion 325 (Figure 1.7.5), which reduces the affinity for the peptide when compared to the canonical tryptophan. The phenylalanine in this position maintains the XP binding pocket, along with tyrosine 314, but with a slightly altered conformation. The carbonyl group of the second proline (209) of the PPxY motif hydrogen bonds threonine 323 of the third β-strand (Chong et al. 2006). The third proline (210) faces away from the interaction interface and the tyrosine (211) sits in the binding pocket formed by a histidine (318), as seen above, a valine (316) (at the

position of the leucine and isoleucine of YAP1 and NEDD4 respectively), and an arginine (321) in the same position as the glutamine described in YAP1 (Figure 1.7.5) (Chong et al. 2006). As with the NEDD4/ENaC interaction, the WW3 structure in complex with the motif peptide shows a further interaction between the WW3 domain and a stretch of sequence C-terminal to the PPxY motif that contributes to the binding affinity (Chong et al. 2006). The peptide turns in the binding site and makes backbone and side chain contacts with the first and second WW domain β-strands. The peptide used in this structure has a longer projection from the C-terminal of the PPxY motif, and contacts become evident with the loop between the first two strands, loop 1 (Chong et al. 2006). NEDD4L WW2 and SMURF1 WW2 domains also bind the same motif in Smad7 (Aragón et al. 2012). The structure of SMURF1 WW2/Smad7 peptide is similar to that of SMURF2 WW3 as described above, however when bound to NEDD4L WW2, the peptide adopts a long hairpin structure (Aragón et al. 2012).

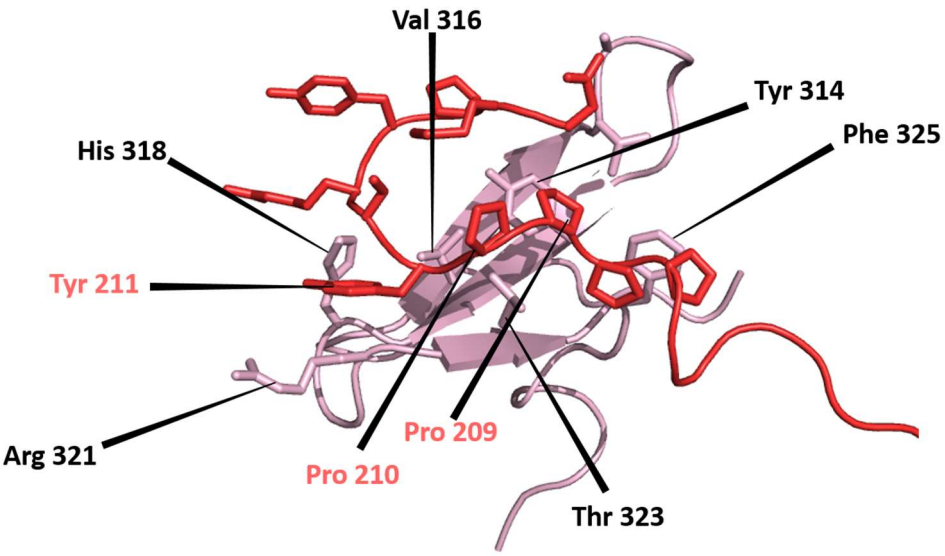


Figure 1.7.5 - The structure of the SMURF2 WW3 domain bound to a Smad7 PPxY motif-containing peptide, with significant residues labelled in black (WW3) and light red (Smad7 ligand). Solved by solution NMR (PDB: 2DJY) (Chong et al. 2006).

1.7.3 WWP2 HECT E3 ligase

The structure of the WWP2 E3 ligase has not been studied as extensively as some of its family members. WWP2 counts among its targets: the PTEN tumour suppressor, the

OCT 4 tumour promoter and pluripotency marker, and like other NEDD4 family ligases, multiple components of the TGFβ signalling pathway (Maddika et al. 2011; Soond & Chantry 2011; Xu et al. 2009). Three human WWP2 isoforms have been identified that have different substrate preferences for various substrates of the TGFβ signalling pathway (Soond & Chantry 2011). The three isoforms are generated by variations in mRNA splicing and different transcription start sites at the *WWP2* gene, and are called WWP2-FL, WWP2-N and WWP2-C, shown in Figure 1.7.6.

The full-length isoform WWP2-FL contains the full complement of domains, with the same domain architecture as other NEDD4 E3s - a C2 domain at the C-terminus followed by four WW domains and a HECT domain at the C-terminus. The N-terminal isoform WWP2-N is generated by retention of intron 9-10 which causes an early stop codon in the transcript. This creates a protein that contains only the C2 domain and the first WW domain but significantly, lacks the HECT domain. The C-terminal isoform WWP2-C is thought to arise from intronic promoter activity at intron 10-11, creating a transcript with a start codon from exon 13 which runs through to the canonical stop codon. Subsequently, WWP2-C has only the fourth WW domain and the HECT domain, and so remains catalytically active.

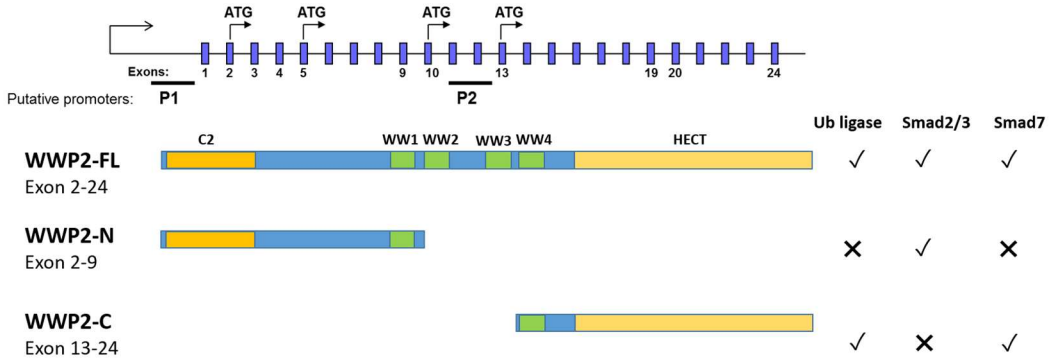


Figure 1.7.6 - The WWP2 isoforms and their domain composition, showing the WWP2 gene (not to scale), the position of the putative promoters, the ubiquitin ligase activity, and the substrate preference.

WWP2-FL interacts with the receptor Smads 2 and 3 that positively propagate the TGFβ signal from the cell surface receptor to the nucleus, and the inhibitory Smad7 (i-Smad) which negatively regulates the pathway (Soond & Chantry 2011). Degradation of Smad7 by WWP2-FL is quite pronounced, while some degradation of Smad3 is also evident. WWP2-N appears not to interact with Smad7, but does interact with Smad2 and Smad3, although this isoform is unable to induce substrate degradation because it lacks

the HECT ubiquitin ligase domain. Because of the ability of WWP2-N to bind r-Smads and not Smad7, it is thought that the WW domain present, WW1, has a preference for r-Smad polyproline motifs over the Smad7 motif. WWP2-FL and WWP2-N interaction with the r-Smads seems to be TGFβ-dependent, suggesting there may be a regulatory phosphorylation event around the PPxY motif that enhances their interaction, something that is true of other NEDD4 E3/Smad interactions (Aragón et al. 2011). A truncated Smad3 with the PPxY motif intact shows constitutive interaction with WWP2-N and FL (Soond & Chantry 2011). The truncation may allow uninhibited access to phosphorylation sites by GSK/CDKs (glycogen synthase kinases/cyclin-dependent kinases) which can switch affinities from WW domains of transcriptional coactivators to E3 ligases, and destruction (Aragón et al. 2011). WWP2-N also interacts with the C-terminal of full length WWP2, and it is thought that this relieves autoinhibition and upregulates WWP2-FL activity against r-Smads. This upregulation of activity against r-Smads may be because of the increased concentration of Smad-binding WW1 around the HECT domain, or may orientate r-Smads in a more favourable position for ubiquitination. WWP2-C appears to only interact with Smad7 but not r-Smads, and causes its degradation at the proteasome. It is thought the WW domain present in WWP2-C, WW4, preferentially binds Smad7 over r-Smads.

1.8 Aims of the thesis

WWP2 represents a system in which a single gene, by different promoter and splice factor activity, can fine-tune the TGFβ signalling pathway in one self-contained module. This is achieved through the activity of alternative isoforms with different substrate specificity. This has implications for the role of this E3 ligase in cancer because of the Jekyll and Hyde role TGFβ plays in tumour progression, sometimes pro and sometimes anti. Perhaps the same could be said now regarding WWP2, depending on which isoform is expressed and, therefore, which part of the pathway is degraded - the activating Smad2/3 or the inhibitory Smad7.

This system merits further investigation to determine exactly what is happening in the various cell-based assays (Soond & Chantry 2011). In order to understand the experimental information, it is important now to look more closely at the relationship between WWP2 and the Smad proteins. And in particular, to look at the structures of the WWP2 domains and their interaction with Smads, and relate this to the activity seen *in*

vivo. To achieve this, crystallography will be employed to attempt to probe the structure of the HECT domain and the WWP2-C isoform. The solution structure of the WW4 domain, which seems to play an important role in Smad7 turnover, will be explored here. The interactions between the WW4 domain and the Smad PPxY motifs will be assessed by using nuclear magnetic resonance (NMR) spectroscopy. The potential for a novel isoform will be examined, which could have a unique contribution to the WWP2 regulatory module. The aim of this thesis is to not only contribute to the understanding of WWP2 and its isoforms, but also to the understanding of the WW protein-protein interaction domains, the HECT domain ligase function and E3 ligases as a whole.

2. Materials and Methods

2.1 Recipes

2.1.1 Lysogeny Broth (LB)

10 g Tryptone

10 g Yeast extract

5 g NaCl

1 ml 100 $mg.ml^{-1}$ ampicillin or 50 $mg.ml^{-1}$ kanamycin

Distilled water to 1 L

2.1.2 LB Agar

10 g Tryptone

10 g Yeast extract

5 g NaCl

30 g Agar

1 ml 100 $mg.ml^{-1}$ ampicillin or 50 $mg.ml^{-1}$ kanamycin

Distilled water to 1 L

2.1.3 Minimal Essential Medium (MEM)

100 ml 10x M9 salts

10 ml 100x MEM Vitamin mix (Sigma Aldrich)

1 ml 1000x Micronutrient mix

2 mM $MgSO_4$

10 µM $CaCl_2$

10 µM $FeSO_4$ (fresh)

4 g Glucose

1 ml 100 $mg.ml^{-1}$ ampicillin or 50 $mg.ml^{-1}$ kanamycin

Distilled water to 1 L

2.1.4 10x M9 salts

6 g Na_2HPO_4

3 g KH_2PO_4

0.5 g NaCl

1 g NH_4Cl

Distilled water to 1 L

pH 7.4

2.1.5 1000x Micronutrient mix

3 µM $(NH_4)_2MoO_4$

4 µM H_3BO_3

30 µM $CaCl_2$

10 µM $CuSO_4$

80 µM $MnCl_2$

10 µM $ZnSO_4$

2.1.6 Immobilised metal ion affinity (IMAC) buffers

	Binding buffer	**Elution buffer**
NaCl	0.5 M	0.5 M
Na_2HPO_4	20 mM	20 mM
Imidazole	30 mM	300 mM
pH	7.4	7.4

Table 2.1.1 Nickel NTA buffers

2.1.7 Gel filtration buffer

50 mM Tris HCl

150 mM NaCl

5 mM DTT (where appropriate)

pH 7.5

2.1.8 PBS

8 g NaCl

0.2 g KCl

0.24 g KH_2PO_4

0.72 g Na_2HPO_4

Distilled water to 1 L

pH 7.4

2.1.9 NMR sample buffer

20 mM Na_2HPO_4

50mM NaCl

pH 6.8

2.1.10 Polyacrylamide gels

	10% Resolving gel	15% Resolving gel	6% Stacking gel
Distilled water	7.2 ml	5.3 ml	2.9 ml
40% (w/v) Acrylamide	3.75 ml	5.63 ml	0.75 ml
1.5 M Tris HCl pH 8.8	3.75 ml	3.75 ml	-
0.5 M Tris HCl pH 6.8	-	-	1.25 ml
10% (w/v) SDS	150 µl	150 µl	50 µl
10% (w/v) APS	150 µl	150 µl	50 µl
TEMED	15 µl	15 µl	5 µl

Table 2.1.2 Polyacrylamide gel recipes

2.1.11 SDS-PAGE buffer

3 g Tris base

14.4 g Glycine

1 g SDS

Distilled water to 1 L

2.1.12 Tricine gels

	Resolving gel	Stacking gel
Distilled water	3.1 ml	6.7 ml
40% (w/v) Acrylamide 19:1	5 ml	1.05 ml
Gel buffer	5 ml	2.5 ml
70% (v/v) Glycerol	2 ml	-
10% (w/v) APS	66 µl	80 µl
TEMED	6.6 µl	8 µl

Table 2.1.3 Tricine gel recipes

2.1.13 Tricine gel buffers

	Gel buffer	Cathode buffer	Anode buffer
Tris HCl	18.15 g	-	-
Tris base	-	24.2 g	48.4 g
Tricine	-	35.84 g	-
SDS	150 mg	2 g	-
Distilled water	50 ml	200 ml	200 ml
pH	8.45	8.25	8.9

Table 2.1.4 Tricine gel buffers

2.1.14 Silver stain solutions

50% (v/v) Methanol

5% (v/v) Methanol

2 µM DTT

0.1% $AgNO_3$

Developing solution

2.1.15 Silver stain developing solution

7.5 g Na_2CO_3

125 µl 35% (v/v) Formaldehyde

Distilled water to 250 ml

2.1.16 2x Laemmli buffer

125 mM Tris

4% (w/v) SDS

20% (v/v) Glycerol

0.01% (w/v) Bromophenol blue

100 mM DTT

pH 6.8

2.1.17 1% Agarose-TAE gel

1 g Agarose

100 ml 1x Tris-Acetate-EDTA (TAE)

10 µl 10,000x SYBR Safe (Fisher Scientific)

2.1.18 50x TAE

242 g Tris base

57.1 ml Glacial acetic acid

100 ml 0.5 M EDTA pH 8.0

Distilled water to 1 L

pH 7.4

2.2 In-Fusion cloning

WWP2 isoforms and the HECT domain were cloned in to the pOPINF vector (Addgene plasmid # 26042) (Berrow et al. 2007) to generate N-terminal hexa-his tagged recombinants. The WW domains used here were cloned in to the pSKDuet01 vector (Addgene plasmid # 12172) (Iwai et al. 2006) to generate a GB1-WW domain recombinant with an N-terminal hexa-his tag. Boundaries used here are outlined in Table 2.2.1.

Forward and reverse primers shown in Table 2.2.2 were designed to clone from a parent pET28a WWP2-FL plasmid. As per the In-Fusion (Clontech) cloning technique, an extension roughly 15bp long was added to the 5' region of each primer that corresponds to the points of insert in to the pOPINF or pSKDuet01 vectors. Primer stocks were made by resuspending lyophilised oligos in nuclease free water to a concentration of 100 µM. Primers were diluted further to 10 µM for use in PCR reactions.

	Genbank accession code	**N-terminal residue number**	**C-terminal residue number**
WWP2-FL	NM_007014.4	1	870
WWP2-N	NM_001270455.1	1	335
WWP2-C	NM_199424.2	440	870
WWP2-HECT	-	495	865
WWP2-WW3	-	402	438
WWP2-WW4	-	438	480
WWP2-WW3-4	-	402	480

Table 2.2.1 WWP2 isoforms, HECT and WW domain boundaries in relation to the WWP2-FL amino acid sequence

	Forward primer 5'-3'	Reverse primer 5'-3'
WWP2-FL	ATGGCATCTGCCAGCTCTAG	CTCCTGTCCAAAGCCCT
WWP2-N	ATGGCATCTGCCAGCTCTAG	GCCTGGAGGAAGGGGC
WWP2-C	ATGATCCAGGAACCAGCTCTGC	CTCCTGTCCAAAGCCCT
WWP2-HECT	TTTCGGTGGAAGTATCACCAGTTCC	CTCGGTCTCCTCAATGGCATAC
WWP2-WW3	GATCCCCTGGGCCCCCT	CTGGGTCCGGGGATCC
WWP2-WW4	GGTGGTGCTGGTGGTCAGGGGATG ATCCAGGA	CTCAAACCCCGGGCGA
WWP2-WW3-4	GATCCCCTGGGCCCCCT	CTCAAACCCCGGGCGA
pOPINF vector specific	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pSKDuet01 vector specific	CGTAACGGAAGGATCC	ATGCGGCCGCAAGCTTTTA

Table 2.2.2 WWP2 isoforms, HECT and WW domain, pOPINF and pSKDuet01 vector-specific forward and reverse primer pairs for In-Fusion cloning

The high fidelity DNA polymerase Phusion (Fisher Scientific) was used to amplify DNA using the 20 µl reaction described in Table 2.2.3 and the cycles described in Table 2.2.4 and Table 2.2.5

	20 µl Reaction	100 µl Reaction
Nuclease free H_2O	12.5 µl	62.5 µl
5x Phusion buffer	4 µl	20 µl
10 mM dNTPs	0.4 µl	2 µl
Forward primer 10 µM	1 µl	5 µl
Reverse primer 10 µM	1 µl	5 µl
Template DNA	1 µl	5 µl
Phusion polymerase	0.2 µl	1 µl

Table 2.2.3 Components and volumes of the 20 µl and 100 µl Phusion polymerase reactions

	WWP2-FL	WWP2-N	WWP2-C	HECT
Initial denaturation	98°C for 30 sec	98°C for 30 sec	98°C for 30 sec	98°C for 30 sec
Denaturation	98°C for 5 sec	98°C for 5 sec	98°C for 5 sec	98°C for 5 sec
Annealing	58°C for 10 sec	62°C for 10 sec	58°C for 10 sec	65°C for 10 sec
Extension	72°C for 40 sec	72°C for 20 sec	72°C for 20 sec	72°C for 20 sec
Cycle to step 2	x29	x29	x29	x29
Final extension	72°C for 5 min	72°C for 5 min	72°C for 5 min	72°C for 5 min

Table 2.2.4 The PCR reaction steps for the pOPINF inserts

	WWP2-WW3	WWP2-WW4	WWP2-WW3-4
Initial denaturation	98°C for 30 sec	98°C for 30 sec	98°C for 30 sec
Denaturation	98°C for 5 sec	98°C for 5 sec	98°C for 5 sec
Annealing	58°C for 10 sec	58°C for 10 sec	58°C for 10 sec
Extension	72°C for 5 sec	72°C for 5 sec	72°C for 5 sec
Cycle to step 2	x34	x34	x34
Final extension	72°C for 5 min	72°C for 5 min	72°C for 5 min

Table 2.2.5 The PCR reaction steps for the pSKDuet01 inserts

Successful amplification was confirmed by mixing the reactions with 4 µl of 6x loading dye (Promega) and running them on a 1% TAE-agarose gel with SYBR Safe at 100 V for 60 min, and analysing the gel with an ultraviolet light. Large scale reactions (100 µl) were run using the same cycle and, using a clean scalpel, the DNA bands corresponding to the amplified region were excised from the agarose gel. DNA was purified from the agarose gel using the QIAquick gel extraction kit (QIAGEN) according to the manufacturer's instructions. DNA concentrations were calculated by UV absorbance at 260 nm, using a nanodrop.

The pOPINF plasmid (32 µl at 85 ng. μl^{-1}) was mixed with 2 µl KpnI (Promega), 2 µl HindIII (Promega) restriction enzymes and 4 µl 10x reaction buffer and incubated at 37°C for 3 hr to linearize the plasmid. The pSKDuet01 vector (32 µl at 130 ng.μl^{-1}) was mixed with 2 µl BamHI (Promega), 2 µl HindIII (Promega) restriction enzymes and 4 µl 10x reaction buffer and incubated at 37°C for 3 hr to linearize the plasmid. The digest was run on a 1% agarose-TAE gel, the band corresponding to the linearized plasmid was excised and the DNA purified using the QIAquick kit. DNA concentration was calculated by UV absorbance at 260 nm.

The 5x In-Fusion HD enzyme premix (1 µl) was mixed with the PCR product and linearized vector (in a ratio of 2:1, respectively) to make a 5 µl reaction. This was incubated for 15 min at 50°C. The reaction mix was transformed in to Stellar competent cells (Clontech) using the heat shock protocol outlined in Section 2.3.

Five colonies for each construct were used to inoculate 5x 10 ml LB-ampicillin which were grown overnight at 37°C with agitation at 200 rpm. PCR was performed on each culture by mixing a 20 µl reaction without the polymerase, as in Table 2.2.3, but replacing the plasmid DNA template with 1 µl of culture, and heating at 98°C for 2 min. Polymerase was added and PCR was performed using the cycle from Table 2.2.4.

Loading dye was added to the reactions and they were run on a 1% agarose-TAE gel and analysed under ultraviolet light. One culture from each construct that generated a band of the correct size was selected for mini prep using the QIAprep kit (QIAGEN).

2.3 Heat shock transformation

Competent bacterial cells were thawed on ice and 50 µl was mixed with 1 µl of plasmid DNA and incubated on ice for 30 min. The mixture was heat shocked in a 42°C water bath for 45 sec and immediately placed back on ice for 10 min, 500 µl LB was added to the mix and the transformation was incubated at 37°C for 1 hr with agitation at 200 rpm. LB-agar-ampicillin (pOPINF) or LB-agar-kanamycin (pSKDuet01) plates were inoculated with 100 µl of the transformation mixture under sterile conditions. Agar plates were incubated at 37°C overnight.

2.4 Overexpression of recombinant proteins - LB

Plasmids were used to transform competent bacteria: BL21 (DE3) pLysS (Promega) for pOPINF constructs or BL21 Star (DE3) (Fisher Scientific) for pSKDuet01 constructs. These were spread on agar plates as in Section 2.3. A single colony from each plate was used to inoculate 20 ml LB-ampicillin or LB-kanamycin which was grown overnight at 37°C with agitation (200 rpm). The overnight culture was used to inoculate 1 L of LB-ampicillin or LB-kanamycin at a ratio of 1 in 50, which was grown at 37°C until the optical density

(O.D.) at 600 nm reached between 0.6-0.8. A 1 M solution of IPTG (Isopropyl β-D-1-thiogalactopyranoside) was used to induce the expression of recombinant proteins at the concentrations shown in Table 2.4.1. Cultures were incubated at the temperatures shown in Table 2.4.1 overnight with agitation.

Overexpression was confirmed with SDS-PAGE by mixing culture aliquots with laemmli buffer, running the samples on 10% acrylamide gels at 180 V for 60 min, and staining with InstantBlue (Expedeon). Scaling up was achieved by increasing the initial overnight culture volume to maintain the 1 in 50 inoculation ratio for larger volumes of final culture.

	IPTG concentration	Expression temperature	Agitation	Expression time
WWP2-FL	0.5 mM	25°C	200 rpm	4 hours
WWP2-N	0.5 mM	25°C	200 rpm	4 hours
WWP2-C	0.5 mM	20°C	200 rpm	Overnight
WWP2-HECT	0.1 mM	20°C	200 rpm	Overnight
WWP2-WW3	0.8 mM	30°C	200 rpm	Overnight
WWP2-WW4	0.8 mM	30°C	200 rpm	Overnight
WWP2-WW3-4	0.8 mM	30°C	200 rpm	Overnight

Table 2.4.1 Expression conditions for the recombinant proteins

2.5 Immobilised metal ion affinity chromatography

Cultures were centrifuged at 4000 rpm for 10 min at 4°C to harvest the bacterial cells. Cell pellets were resuspended in 30 ml IMAC binding buffer or until viscosity became reasonable so that the suspension was easily manipulated.

The suspension was passed through a French pressure cell (Glen Mills) twice at 10,000 psi. The lysate was clarified by ultracentrifugation at 60,000 rpm for 30 min in an ultracentrifuge (Beckman Coulter).

The lysate was passed through a nickel charged 1 ml or 5 ml HisTrap FF column (GE Healthcare) at 1 $ml.min^{-1}$ or 5 $ml.min^{-1}$ respectively, using a Peristaltic Pump P-1 (GE Healthcare). The column was attached to an ÄKTA-FPLC (GE Healthcare) and the column was flushed with the buffers and volumes shown in Table 2.5.1. Purity was analysed by

running aliquots of alternate fractions on 10% acrylamide gels. Fractions of sufficient purity were pooled.

	1 ml HisTrap column		5 ml HisTrap column	
	Volume	**Fraction size**	**Volume**	**Fraction size**
Binding buffer wash	20 ml	2 ml	50 ml	5 ml
Gradient	10 ml	1 ml	50 ml	2 ml
Elution buffer	10 ml	2 ml	50 ml	5 ml

Table 2.5.1 The volumes and fraction sizes of the ÄKTAFPLC HisTrap elution programs

2.6 Size-exclusion chromatography

2.6.1 Column calibration

The HiLoad 16/600 Superdex 75 prep grade column (GE Healthcare) column calibration was performed using the following standard column calibrants: cytochrome c, carbonic anhydrase, bovine serum albumin, alcohol dehydrogenase and β-amylase. The relationship between elution volume and molecular was used to produce the following equations (for two different S75 columns) that were used to predict the molecular weight of bacterially expressed recombinant proteins:

$$y = -37.721x + 122.67$$

$$y = -1.3341x + 6.3953$$

$y = \text{elution volume}$

$x = \log_{10} \text{molecular weight}$

2.6.2 Recombinant protein size-exclusion chromatography

The HiLoad 16/600 Superdex 75 prep grade column (GE Healthcare) was connected to the ÄKTA-FPLC and washed with 2 column volumes of water and equilibrated with 2 column volumes of gel filtration buffer at 1 ml.min^{-1}. The sample was concentrated to 2 ml using

a 5 kDa MWCO Vivaspin centrifugal concentrator (Sartorius Stedim Biotech), and injected on to the column. The sample was gel filtered at 1 $ml.min^{-1}$ and 2 ml fractions were collected. Aliquots of alternate fractions were analysed for purity by SDS-PAGE. Fractions of sufficient purity were pooled.

2.7 3C protease digest

The pOPINF recombinants contained an N-terminal 3C protease-cleavable hexa-his tag. The tag was cleaved after gel filtration by adding 1 unit of Human Rhinovirus (HRV) 3C protease (Novagen) per 100 µg of protein, as determined by absorbance at 280 nm using a nanaodrop. The digest was incubated at 4°C until sufficient digestion was achieved, as determined by SDS-PAGE.

2.8 Crystallisation trials

Samples were concentrated using Vivaspin 5 kDa MWCO centrifugal concentrators until a concentration of 10 $mg.ml^{-1}$ or a minimum volume of 48 µl per crystallisation plate was reached. The MRC Crystallisation Plate (Molecular Dimensions) was used to set crystallisation trials (96-wells held 2 sitting drops per condition). Two plates were set for each of the following crystallisation trials: Structure Screen I (conditions 1-48), Structure Screen II (conditions 49-96), JCSG-plus, PEG/Ion and PACT premier (Molecular Dimensions).

Crystallisation trial solutions (100 µl) were aliquoted in to each of the wells and 0.25 µl of the protein was mixed with either 0.25 µl (upper drop) or 0.5 µl (lower drop) of the well condition. Plates were covered with adhesive film and incubated at 16°C and 4°C. Plates were monitored for the formation of crystals over the following 8 weeks.

2.8 Overexpression of isotopically enriched recombinant proteins

For isotopically labelling proteins, plasmids were transformed into competent bacteria and spread on agar plates containing appropriate antibiotics. After an overnight incubation at 37°C, 5 ml of LB was inoculated with a single colony and grown overnight at 37°C with agitation at 200 rpm. The culture was used to inoculate 50 ml of MEM at 1 ml in 50 ml respectively. The MEM culture was grown overnight at 37°C and cells were harvested the following morning by centrifugation at 4000 rpm, 25°C for 10 min

The cell pellet was resuspended in 1 L of isotopically enriched MEM made using ^{15}N NH_4Cl for nitrogen labelled proteins, or ^{15}N NH_4Cl and ^{13}C glucose for nitrogen and carbon labelled proteins. The culture was grown at 37°C until it reached an O.D. at 600 nm of 0.8-1.0, at which point expression was induced with IPTG using the conditions outlined in Table 2.4.1.

The culture was then centrifuged at 4000 rpm and the recombinant protein was purified from the cell pellet as outlined in Section 2.5.

2.9 Thrombin digest

The GB1 fusion proteins used here contained a thrombin-cleavable His-tag at the N-terminus. The His-tag was removed after nickel-affinity purification. Thrombin was added at 2 units per mg of purified protein. Using 7 kDa MWCO Snakeskin dialysis tubing (Fisher Scientific), the protein/thrombin mix was dialysed in to PBS at room temperature until a sufficient amount of protein was digested, as determined by SDS-PAGE.

A HisTrap FF column was equilibrated with PBS and the digest was passed through. The column was washed with PBS. The flow-through and wash were pooled. Undigested protein was eluted from the column by flushing with 10 column volumes of elution buffer, and discarded.

The cleaved protein was concentrated and gel filtered as outlined in Section 2.6.

2.10 NMR sample preparation

After size-exclusion chromatography the protein was concentrated using 5 kDa MWCO centrifugal concentrators until the sample volume was roughly 430 µL. The sample was made up to 500 µL by the addition of D_2O to a final concentration of 10% v/v, DSS to 200 µM and NaN_3 to 0.03%. The sample was filtered using a 0.22 µm spin filter (Corning), and transferred to a 5 mm PP535 NMR tube (Wilmad).

2.11 NMR spectroscopy

NMR data were collected at 298 K from either a Bruker Avance III 800 MHz spectrometer or a Bruker Avance I 500 MHz spectrometer. The spectra used for the backbone and side-chain assignment were: ^{1}H-^{15}N-HSQC, ^{1}H-^{13}C-HSQC, CBCA(CO)NH, HNCACB, CC(CO)NH, H(CCO)NH TOCSY, an aromatic ^{13}C TOCSY (hnCBcgcdceHE) and an aromatic TROSY HSQC. Through-space distance restraints for the structural calculation were determined by NOESY (Nuclear Overhauser Effect Spectroscopy) spectra, a ^{15}N-NOESY-HSQC (100 ms mixing time) and a ^{13}C-NOESY-HSQC (100 ms mixing time). Table 2.11.1 shows the parameters for each of the spectra.

		Increments			**Spectral width**		
	Scans	**^{1}H**	**^{15}N**	**^{13}C**	**^{1}H**	**^{15}N**	**^{13}C**
^{1}H-^{15}N-HSQC	8	1024	256	-	12019	2595	
^{1}H-^{13}C-HSQC	64	1024	-	256	12019	-	16077
CBCA(CO)NH	24	1024	64	130	12019	2433	15105
HNCACB	24	1024	64	132	12019	2433	15105
CC(CO)NH	32	1024	42	128	12019	2433	15105
H(CCO)NH	48	1024	48	160(^{1}H)	12019	2433	11990(^{1}H)
Aromatic ^{13}C TOCSY	32	1024	-	48	11160	-	2823
Aromatic TROSY HSQC	1024	1024	-	256	7212	-	8052
^{15}N-NOESY-HSQC	32	1024	48	152(^{1}H)	7508	1519	6997(^{1}H)
^{13}C-NOESY-HSQC	16	1024	192(^{1}H)	64	11161	10395(^{1}H)	15105

Table 2.11.1 NMR acquisition parameters

2.12 Spectral processing

Linear prediction (non-directly detected dimensions), zero-filling, phasing and Fourier transformation were applied to all spectra, which were processed using the NMRPipe software suite (Delaglio et al. 1995). The DSS standard peak frequency was used to directly calibrate ^{1}H chemical shifts and indirectly calibrate, via conversion with the appropriate gyromagnetic ratio, heteronuclear dimensions.

2.13 Resonance assignment

The CCPN Analysis software package (Vranken et al. 2005) was used to analyse all spectra. To assign the backbone, the CBCA(CO)NH and HNCACB spectra were used in conjunction with the ^{1}H-^{15}N-HSQC spectrum. Resonances corresponding to unknown α and β-carbons in the HNCACB were assigned to their amide group resonance in the ^{1}H-^{15}N-HSQC. Amide resonances were put in sequence using the CBCA(CO)NH and HNCACB to match *i* and *i-1* resonances. Comparison of the α and β-carbon resonances with the Biological Magnetic Resonance Bank (BMRB) standard values (Ulrich et al. 2008) allowed assignment of residue type.

Side chain hydrogen resonances were assigned in the same fashion using the H(CCCO)NH, aromatic TROSY and the ^{1}H-^{13}C-HSQC. The remaining unassigned side chain carbon resonances were assigned using the CC(CO)NH, aromatic TOCSY and the ^{1}H-^{13}C-HSQC spectra.

2.14 Structure calculation

NOE (Nuclear Overhauser Effect) peak picking and assignment was performed on the carbon and nitrogen NOESY spectra by UNIO (Volk et al. 2008; Fiorito et al. 2008) and the ATNOS/CANDID algorithms therein (Herrmann et al. 2002a; Herrmann et al. 2002b). Input files were: resonance assignments in the XEASY format (Bartels et al. 1995), CARA formatted (Keller 2004) NOESY spectra, the protein amino acid sequence in CYANA format

and chemical shift-based phi and psi dihedral angles from the TALOS+ server (Shen et al. 2009).

The ATNOS algorithm performed automated peak picking and gave upper limit restraints to each peak based on its intensity. The CANDID algorithm assigned the peaks against the assignments determined in Section 2.15. Peaks with multiple possibilities were considered ambiguous and given different possible assignments. The ATNOS/CANDID algorithms worked in conjunction with the CYANA torsion angle dynamics algorithm (Güntert 2004).

The UNIO algorithms went through seven cycles whereby the ambiguous assignments were evaluated against, at first, an initial covalent structure (cycle 1) and then a structural ensemble iteration calculated by CYANA in the previous cycle by simulated annealing calculations (cycle 2-7). After each cycle, the peak selections and assignments were further refined by the ATNOS/CANDID algorithms and ambiguities were progressively removed if found to be incompatible with each tentative ensemble, until a list of unambiguous restraints were determined by cycle 7. Models incorporated NOE restraints, and were optimised against dihedral restraints with simultaneous energetic minimisation.

The final set of optimised, ambiguous, restraints (cycle 6) were then incorporated into a more computationally involved, atomic-level calculation using CNS 2.3 (Brünger et al. 1998) with the publicly available EBI RECOORD scripts (Nederveen et al. 2005). Restraints from cycle 6 of the UNIO calculation were used to maintain ambiguity. CNS calculated models were then subject to a final round of CNS-based calculations to account for effects contributed by the H_2O solvent. The calculation generated 100 models and the 20 with the lowest overall energy, and no violations of the dihedral and NOE restraints, were selected for the final ensemble.

2.15 Bacterially expressed Smad7 peptide

The Smad7 peptide sequence expressed here is based on a PPxY motif-containing Smad7 peptide sequence conventionally used in WW domain ligand binding experiments, but which has been extended slightly at the N-terminus to increase the molecular weight. Buffer exchange is an essential step in the preparation of this peptide from bacteria so this increase in molecular weight is to allow the peptide to be dialysed using one of the

smallest MWCO dialysis membranes available without comprehensive loss of the peptide. Conventional and extended peptide sequences are shown in Table 2.15.1

	Peptide amino acid sequence
Conventional Smad7 peptide	**203 -** ELESPPPPYSRYPMD **- 217**
Extended Smad7 peptide	**199 -** SRLCELESPPPPYSRYPMD **- 217**

Table 2.15.1 Conventional and extended Smad7 peptide amino acid sequences

The peptide was expressed as a SUMO conjugate using a modified pET21d(+) vector, a gift from Dr Robin Maytum, University of Bedfordshire. The vector carries an N-terminal His-tag sequence followed by a sequence coding for the ubiquitin-like protein SUMO. Downstream of the SUMO tag are two restriction sites, BsaI followed by BamHI. The BsaI restriction site generates a blunt end after the sequence that corresponds to the glycine-glycine of SUMO, and BamHI cuts just 3' to the BsaI site. An expression tag was used to avoid proteolytic destruction of the small peptide.

The SUMO tag was used because, unlike commonly used proteases, it allows for the removal of the tag without leaving residual amino acids from the cleavage site which might make a significant contribution to the properties of the peptide, due to its small size. The digest was achieved using ubiquitin-like-specific protease-1 (ULP-1) which recognises the tertiary structure of SUMO and cleaves at the C-terminus of the second glycine in the glycine-glycine motif. Creating this seamless protease site also necessitates the use of BsaI to create a blunt end, so as to eliminate restriction site artefacts.

2.15.1 Cloning

The Smad7 peptide was cloned in to the pET21d(+)SUMO vector using the In-Fusion cloning system and the primer pairs outlined in Table 2.15.2. The Smad7 peptide sequence was amplified from a parent pRK5-HA Smad7 plasmid following the PCR protocol outlined in Section 2.2, with an annealing temperature of 61°C and an extension time of 3 sec, all remaining parameters were kept the same.

Restriction digest of the pET21d(+)SUMO vector was performed by mixing 20 µl vector (162 ng.µl^{-1}), 2 µl BamHI, 2 µl BsaI (Fisher Scientific), 4 µl 10x buffer and 12 µl nuclease-free water. This was incubated at 37°C for 3 hr to linearize the plasmid. PCR product and

linearized plasmid were gel extracted using the QIAquick kit, as in Section 2.2. The In-Fusion reaction, competent bacteria transformation and plasmid mini-prep were performed as in Section 2.2.

	Forward primer 5'-3'	**Reverse primer 5'-3'**
Smad7(199-217)	AGCCGACTCTGCGAACTAGA	ATCCATCGGGTATCTGGAGTA
pET21d(+)SUMO vector-specific	GAACAGATTGGAGGT	GCTCGAATTCGGATCCTTA

Table 2.15.2 Smad7 peptide and pET21d(+)SUMO vector-specific forward and reverse primer pairs for In-Fusion cloning

2.15.2 Overexpression and purification

The heat-shock transformation protocol in Section 2.3 was used to transform competent BL21-CodonPlus(DE3) cells (Agilent Technologies) with the pET21d(+)SUMO-Smad7 plasmid for large scale expression. The LB expression or MEM expression protocols, in Section 2.4 and 2.8 respectively, were followed, using 0.8 mM IPTG for induction and a temperature of 32°C for expression overnight. The recombinant protein was purified following the protocol for IMAC purification in Section 2.5.

2.15.3 ULP-1 protease digest

The pooled protein was dialysed in to gel filtration buffer overnight at 4°C using 7 kDa MWCO Snakeskin dialysis tubing. His-tagged ULP-1 from the lab of Dr Tharin Blumenschein, School of Chemistry, UEA (expressed and purified by Danielle De Bourcier) was added in a ratio of 0.5:100 (volume of ULP-1 to volume of SUMO Smad7) and incubated at room temperature until sufficient digest was achieved, as determined by SDS-PAGE. A 1 ml or 5 ml HisTrap column was equilibrated with gel filtration buffer and the digest was passed through, the column was washed with 10 column volumes of gel filtration buffer and then elution buffer. The flow-through and the gel filtration buffer wash were pooled and the elution buffer wash was discarded.

2.15.4 Concentrating the peptide

Conventional techniques such as centrifugal concentrators and stirred ultrafiltration cells present an ineffective way to concentrate small peptides because of the limited range of molecular weight cut-off concentrators available. Lyophilisation was trialled but when attempting to resuspend the peptide under a variety of conditions, resulted in aggregates that ran as large molecular-weight smears when analysed by SDS-PAGE. Concentration by centrifugal evaporation proved to be the most successful method.

One of the challenges presented by this approach is the potential to concentrate buffer components along with the peptide, and since the target volumes were so low, dialysis after concentrating was impractical and had the potential to cause buffer in-flux which would dilute the peptide. To prevent this, and using Spectra/Por 1 kDa MWCO membrane (Spectrum Laboratories), the peptide was dialysed using 5 L distilled water for 6 hr, after which the water was replaced, this was repeated 5 times.

The peptide was aliquoted in to 2 ml tubes and covered with Parafilm M (Bemis) which was pierced multiple times. This was placed in to a miVac centrifugal vacuum evaporator (Genevac) and centrifuged at room temperature under a vacuum until sufficiently concentrated. NMR sample buffer components and DTT were added from a 10x stock.

2.16 NMR ligand titration

2.16.1 Bacterially expressed Smad7

For titration of Smad7 peptide purified from bacteria, two 600 µl NMR samples per titration were made according to Table 2.16.1; one at the start concentration of Smad7 (titration point 1) and one at the final concentration (titration point 11). WW domain concentrations were limited by peptide concentrations and were therefore well below optimum.

	WW4		WW3-4		WW3	
	1:0	**1:10**	**1:0**	**1:10**	**1:0**	**1:10**
GB1:WW4^{15N}	0.08 mM	0.08 mM	-	-	-	-
GB1:WW3-4^{15N}	-	-	0.08 mM	0.08 mM	-	-
GB1:WW3^{15N}	-	-	-	-	0.08 mM	0.08 mM
Smad7 peptide	-	0.8 mM	-	0.8 mM	-	0.8 mM
NaCl	50 mM	50 mM	50 mM	50 mM	50 mM	50 mM
Na_2HPO_4	20 mM	20 mM	20 mM	20 mM	20 mM	20 mM
DTT	5 mM	5 mM	5 mM	5 mM	5 mM	5 mM
D_2O	10%	10%	10%	10%	10%	10%
NaN_3	0.03%	0.03%	0.03%	0.03%	0.03%	0.03%
DSS	200 µM	200 µM	200 µM	200 µM	200 µM	200 µM
pH	6.8	6.8	6.8	6.8	6.8	6.8

Table 2.16.1 NMR samples used to titrate Smad7 against ^{15}N enriched WW4 or WW3-4

A ^{1}H-^{15}N-HSQC was performed on both samples (titration points 1 and 11). A series of titration points were performed whereby an aliquot of the 1:0 sample was removed and replaced with an aliquot of equal volume from the 1:10 sample. Titration points are outlined in Table 2.16.2. After each titration point a ^{1}H-^{15}N-HSQC was collected. These were processed using the method outlined in Section 2.12 and analysed using the CCPN Analysis software.

Titration point	WW4:ligand ratio	Volume of 1:0 sample removed	Volume of 1:10 sample added	WW domain concentration	Ligand concentration
1	1:0	0 µl	0 µl	0.08 mM	0 mM
2	1:0.1	5 µl	5 µl	0.08 mM	0.008 mM
3	1:0.199	5 µl	5 µl	0.08 mM	0.01592 mM
4	1:0.494	15 µl	15 µl	0.08 mM	0.03952 mM
5	1:0.975	25 µl	25 µl	0.08 mM	0.078 mM
6	1:1.9	50 µl	50 µl	0.08 mM	0.152 mM
7	1:3.6	100 µl	100 µl	0.08 mM	0.288 mM
8	1:5.2	100 µl	100 µl	0.08 mM	0.416 mM
9	1:6.8	100 µl	100 µl	0.08 mM	0.544 mM
10	1:8.4	100 µl	100 µl	0.08 mM	0.672 mM
11	1:10	*n/a*	*n/a*	0.08 mM	0.8 mM

Table 2.16.2 Titration points of the bacterially expressed Smad7

2.16.2 Synthetic peptide

Synthetic peptides were either bought from Proteogenix or were synthesised by Richard Steel, School of Pharmacy, UEA. Sequences are shown in Table 2.16.3

	Peptide amino acid sequence
Smad7 peptide	**203 -** ELESPPPPYSRYPMD **- 217**
Smad7 phosphopeptide	**203 -** ELEpSPPPPYSRYPMD **- 217**
Smad2 peptide	**217 -** IPETPPPGYISEDGE **- 231**
Smad3 peptide	**176 -** IPETPPPGYLSEDGE **- 190**

Table 2.16.3 Amino acid sequences of the synthetic peptides

Based on manufacturer-given weight, and the molecular weight of the peptides as trifluoracetic acid (TFA) salts, peptides were resuspended in NMR sample buffer to a concentration of 100 mM. A single 500 µl NMR sample was made for each of the titrations, and for each titration point a small volume of peptide was added as shown in Table 2.16.4. As in Section 2.16.1, ^{1}H-^{15}N-HSQC spectra were acquired after each titration point.

		WW4 0.78 mM	**WW4 0.78 mM**	**WW4 0.336 mM**	**WW3-4 0.543 mM**
Titration point	**Protein:Ligand ratio**	**Volume of Smad7**	**Volume of pSmad7**	**Volume of Smad2**	**Volume of Smad7**
1	1:0	0.0 µl	0.0 µl	0.0 µl	0.0 µl
2	1:0.1	0.4 µl	0.4 µl	0.2 µl	0.25 µl
3	1:0.2	0.4 µl	0.4 µl	0.2 µl	0.25 µl
4	1:0.5	1.2 µl	1.2 µl	0.5 µl	0.75 µl
5	1:1	2.0 µl	2.0 µl	0.8 µl	1.25 µl
6	1:2	4.0 µl	4.0 µl	1.7 µl	2.5 µl
7	1:4	7.8 µl	7.8 µl	3.4 µl	4.0 µl *(1:3.6)*
8	1:6	7.8 µl	7.8 µl	3.4 µl	6.0 µl
9	1:8	7.8 µl	7.8 µl	3.4 µl	-
10	1:10	7.8 µl	7.8 µl	3.4 µl	13 µl

Table 2.16.4 Titration points of the synthetic peptides, showing ligand increment volumes

2.17 Dissociation constant calculation

Spectra were processed as in Section 2.12. Changes in shift of the amide peaks in the ^{1}H-^{15}N-HSQC spectra were weighted according to gyromagnetic ratio and fitted to a Protein-Ligand fast-exchange calculation in CCPN Analysis, as follows:

$$y = A\left(B + x - \sqrt{(B+x)^2 - 4x}\right)$$

$A = \Delta\delta_{\infty}/2$

$B = 1 + K_d/a$

$x = b/a$

$y = \Delta\delta_{obs}$

$a =$ total protein concentration

$b =$ total ligand concentration

$\Delta\delta_{obs} =$ change in chemical shift

$\Delta\delta_{\infty} =$ difference between start chemical shift and chemical shift at saturation

Peaks with shift changes that fit poorly with the equation were ignored and the binding site was identified by the presence of a significant peak trajectory upon ligand titration. Movement of amide resonances from outside of the binding site were considered the result of allosteric changes or non-specific binding. An average of the binding site dissociation constants was taken to give the final K_d value.

2.18 GB1 recombinant protein sequence labelling

Typically throughout this text, the amino acids of the GB1 recombinant proteins used here are referred to by their amino acid number in the native WWP2 amino acid sequence, and not by their position in the recombinant protein. For the purpose of clarity, Table 2.18.1 outlines the amino acid sequence numbering in the recombinant protein and the wild-type protein.

		WW3		WW4	
	GB1 tag	**Recombinant**	**Wild-type**	**Recombinant**	**Wild-type**
GB1:WW3	1-61	62-98	402-438	-	-
GB1:WW4	1-66	-	-	67-109	438-480
GB1:WW3-4	1-66	67-145^{WW3-4}	402-480^{WW3-4}		

Table 2.18.1 GB1 recombinant protein sequence numbering relative to the wild-type WWP2 sequence and the recombinant protein sequence. Note, unlike GB1:WW4 and GB1:WW3-4, the GB1:WW3 construct does not contain an artificial linker region between the tag and WW domain.

2.19 Semi-quantitative PCR

2.19.1 Mammalian tissue culture

Adherent mammalian tissue cell lines outlined in Table 2.19.1 were maintained in 5% CO_2 at 37°C. Cells were grown in T75 flasks (Nunclon) with media described in Table 2.19.1, supplemented with 10% foetal bovine serum (Fisher Scientific), 1% penicillin streptomycin (Fisher Scientific) and 1x Glutamax (Fisher Scientific).

Cell line	Tissue	Cell type	Media
A375	Skin	Melanoma	DMEM (Fisher Scientific)
COLO357	Pancreas	Adenocarcinoma	RPMI 1640 (Fisher Scientific)
SK-MEL28	Skin	Melanoma	DMEM
VCaP	Prostate	Vertebral metastasis	DMEM

Table 2.19.1 Mammalian tissue cell lines

When cells reached 80-90% confluency the media was aspirated, cells were rinsed with 5 ml Dulbecco's Phosphate Buffered Saline (Fisher Scientific) and 5 ml TrypLE was added. This was incubated at 37°C for 5 min, the cells were spun at 1200 rpm for 5 min and resuspended in 1 ml warm media. Cells were seeded in 6-well plates (Fisher Scientific) and 2 ml media was added. Cells were grown overnight to allow adherence.

2.19.2 TGFβ stimulation

Cells were serum starved in 0.5% foetal bovine serum for 16 hr, media was replaced and TGFβ (R&D Systems) was added at 5 $ng.ml^{-1}$.

2.19.3 Reverse transcription

RNA was harvested at a series of time points using the SV Total RNA Isolation kit (Promega), according to the manufacturer's instructions. Concentrations were measured by absorbance at 260 nm. Routinely, 1-0.5 µg of RNA was mixed with 0.5 µg random primers (Promega) and nuclease-free water to a volume of 5 µl. This was heated to 70°C for 5 min and chilled on ice for 5 min. The mixture was centrifuged for 10 sec and the GoScript Reverse Transcriptase components (Promega) from Table 2.19.2 were added, and the heat cycle from Table 2.19.3 was run.

Kit component	Volume
GoScript 5x Reaction buffer	4 µl
$MgCl_2$ (25 mM)	3.2 µl
PCR Nucleotide mix (10 mM)	1 µl
Recombinant RNasin	20 units
GoScript reverse transcriptase	1 µl
Nuclease-free water	5.3 µl

Table 2.19.2 GoScript reaction components

	Temperature	Time
Annealing	25°C	5 min
Extension	42°C	1 hr
Transcriptase inactivation	70°C	15 min

Table 2.19.2 Reverse transcription heat cycle

2.19.4 GoTaq PCR

Primers outlined in Table 2.19.1 were used in conjunction with the GoScript kit (Promega). The WWP2C-ΔHECT primers were designed to amplify between exon 17 and just inside intron 19-20 of WWP2. A 20 µl reaction was mixed for each of the time points for GAPDH following the recipe in Table 2.19.2. PCR was run on the reaction mix following the heat-cycle in Table 2.19.3, these were run on a 1% Agarose-TAE gel and observed under an ultraviolet light.

	Forward 5'-3'	**Reverse 5'-3'**
WWP2C-ΔHECT	GCTGGGAAGAACAATTACTG	TTCCTCTGTAACATGCTCCCT
GAPDH	ACCACAGTCCATGCCATCAC	TCCACCACCCTGTTGCTGTA

Table 2.19.1 Semi-quantitative PCR primers

A 20 µl reaction was mixed for each of the time points for WWP2C-ΔHECT and GAPDH as in Table 2.19.2 but the amount of DNA was adjusted based on the intensities of GAPDH. PCR was run on the sample as in Table 2.19.3 and these were run on a 1% Agarose-TAE gel, and observed under an ultraviolet light.

Kit component	**Volume**
5x Green GoTaq Flexi buffer	4 µl
$MgCl_2$ (25 mM)	1.5 µl
PCR Nucleotide mix (10mM)	0.4 µl
Forward primer	2 µl
Reverse primer	2 µl
cDNA	0.4 µl
GoTaq polymerase	0.1 µl
Nuclease-free water	to 20 µl

Table 2.19.2 GoTaq reaction components

	Temperature	Time
Preheat	95°C	-
Initial denaturation	95°C	2 min
Denaturation	95°C	30 sec
Annealing	54.4°C	30 sec
Extension	72°C	45°C
Cycle to step 3	x34	-
Final extension	72°C	5 min
Soak	4°C	-

Table 2.19.3 GoTaq PCR heat-cycle

2.20 Mass spectrometry

Proteins were run on SDS-PAGE gels and individual bands were excised using a scalpel. Proteins bands were send to Dr Gerhard Saalbach at the John Innes Centre, Norwich, to be analysed by MALDI-TOF (Matrix-assisted desorption/ionization time-of-flight) mass spectrometry after having been digested in to fragments with trypsin.

3. Approaches Towards Purifying the WWP2 Protein

3.1 Introduction

Research in to NEDD4 E3 ligase HECT domain structures has recently provided some revelations in the understanding of how ubiquitin is accepted from E2 conjugators. The mechanism by which HECT E3 ligases bind substrates and position them for ubiquitination is, so far, poorly understood, and knowledge is limited to observations gleaned from HECT crystal structures in the absence of their substrate-interaction domains. The structure of the WWP2 HECT domain at the time of carrying out lab work for this thesis had not been elucidated, although a paper has recently been released that reports to show the crystal structure (PDB ID: 4Y07) (Gong et al. 2015). This will be explored further in the discussion section of this chapter.

It is apparent that the WW domains of WWP2 are not only responsible for substrate binding and positioning, but also play a role in auto-inhibition of the HECT domain by blocking ubiquitin charging by its E2 ubiquitin conjugator (Soond & Chantry 2011; Riling et al. 2015). In the NEDD4 E3 ligase ITCH, the central WW domains WW2 and WW3 are responsible for this auto-inhibition (Riling et al. 2015), and it has been suggested that WWP2 WW domains may be antagonistic towards each other in their ability to bind substrates (Jiang, Wang, et al. 2015). A particular curiosity in the WWP2-C isoform structure and function is due to its potential to act as an oncogene via the TGFβ pathway. If the C-terminal isoform preferentially binds the inhibitory Smad7 over its propagator cousins Smad2 and Smad3, as previously suggested (Soond & Chantry 2011), it could facilitate overactivity of the TGFβ pathway. Overactivity of the TGFβ pathway can drive the transdifferentiation of epithelial cells to mesenchymal cells (EMT), which is an important step in the progression of tumours to an aggressive invasive phenotype (Katsuno et al. 2013). This is particularly pertinent when considering isoform expression, due the intimate relationship between alternate splicing programs, TGFβ and EMT (Shapiro et al. 2011; Horiguchi et al. 2012).

To fully understand the differences between WWP2 isoform activities, a comprehensive structural analysis of the different domains is necessary in order to elucidate the features governing their activity. The lack of information regarding, firstly, the WWP2 HECT domain structure and differences or similarities with HECT domains of other ligases; secondly, how the different WW domain structures contribute to their activity; and thirdly, how the WW domains and HECT domain work together in the

different isoforms, drives an ambitious interest in obtaining structures of WWP2 isoforms and their domains.

There are two approaches most commonly used to determine protein structures. These are, X-ray diffraction of protein crystals, obtained using the technique of crystallography, and NMR spectroscopy, in which the property of atomic resonance is exploited. One of the most challenging aspects of crystallography is that the protein of interest must be amenable to crystallisation. This might mean that instead of being able to crystallise and solve the structure of the protein of interest, only a single domain might crystallise, or an ortholog of the protein of interest. Proteins that are most successfully crystallised typically hold a compact conformation with minimal disorder. Protein samples must be of high purity and stability, because the formation of crystals requires regular contacts to be made between homogenous protein units. Contaminants or partially degraded proteins can disrupt the regular contacts and hinder crystal formation, or lower the diffraction resolution of a crystal that does form.

Using NMR spectroscopy, disorder can be tolerated, as long as peaks are dispersed sufficiently in the HSQC experiments so as to allow individual peaks to be assigned. This approach is limited to proteins below a certain size (roughly 30 kDa), firstly because of peak crowding in the spectra which challenges peak assignment, and secondly, as protein size increases, the tumbling of the molecule slows which causes faster decay of the NMR signal, resulting in peak broadening. NMR spectroscopy requires proteins of high purity because of the potential for contaminant proteins to produce signals in the spectra that confuse or invalidate the results. Proteins need to be stable enough at biological temperatures during the acquisition of NMR data over periods of days, and ideally longer for practical purposes. The emphasis on the need for stable proteins with high concentrations and high purity require certain approaches to be taken to fulfil these requirements, it also excludes some proteins from being studied. The use of NMR spectroscopy to study protein structure will be discussed in further detail in Chapter 4.

3.1.1 The T7 expression system

In order to achieve the high protein concentrations required for structural biology investigations, an *E. coli* bacterial host under the control of an IPTG-inducible T7 RNA

polymerase/promoter expression system was used. This system employs the presence of the T7 RNA polymerase in the host genome which has been inserted in to the host chromosome. T7 RNA polymerase is under the control of the *lac* promoter and operator which allows for induction of expression by the non-degradable lactose analogue IPTG. IPTG-inducible expression is necessary to prevent over-expression of proteins during early stage growth when bacterial density is low, growth would be inhibited and yield would be low. IPTG binds to and releases a repressor at the *lac* operator which allows upregulation of T7 RNA polymerase expression. This system is used in conjunction with a plasmid containing the gene to be overexpressed. This gene is under the control of the T7 promoter as well as the *lac* operator. The presence of the *lac* operator means that expression is also under the control of IPTG. The T7 promoter means that T7 RNA polymerase is responsible for transcription. T7 RNA polymerase synthesises RNA at a rate many times that of the host RNA polymerase and can saturate the host ribosomes with RNA from the target gene, making it ideal for over-expression (Tabor 2001; Studier & Moffatt 1986).

3.1.2 Immobilised metal affinity chromatography

In order to purify the WWP2 proteins from bacterial lysate, a cleavable polyhistidine tag was introduced at the N-terminus. His-tag purification relies on the affinity of the histidine imidazole side chain for transition metal ions (Block et al. 2009). To enhance the affinity of tagged proteins for metal ions, the His-tag is composed of a six histidine repeat. The column used to purify the protein is composed of nickel immobilised by an NTA-agarose matrix, the nickel binds the tag with high affinity and allows untagged proteins to flow through the matrix with limited binding. The bound protein is eluted from the column by imidazole in solution, which displaces the imidazole histidine side chains and allows the protein to flow through the matrix.

3.1.3 Size-exclusion chromatography

To enhance the purity of proteins purified by immobilised metal affinity chromatography, a further purification step is performed whereby the proteins are

subjected to size-exclusion chromatography. The protein is passed under pressure through a column that is packed with a porous matrix. Large proteins or protein aggregates that are too large to enter the pores elute first in the void volume having experienced no diversion through the matrix. The void volume is equal to the mobile phase volume and constitutes the buffer outside of the column matrix. Proteins that travel through the matrix are eluted largest first and smallest last, because smaller proteins have access to a greater column volume by travelling further in to the pores, and experience a longer transit. The volume of buffer within the matrix is known as the stationary phase volume. Larger proteins have access to some of the stationary phase volume, but not all of it, by travelling through some pores but not the smaller pores. By using this approach, proteins can be separated across a relatively large volume, according to their size. This allows impurities to be removed from the protein sample. Column resolution is a limiting factor in the ability to produce the purest protein sample using size-exclusion chromatography, as impurities of a similar molecular weight will elute at a similar volume.

3.1.4 Protein crystallisation

In protein crystals, a lattice of protein molecules is maintained by weak contacts between neighbouring molecules. On average 50% of their mass is solvent, which forms channels and cavities throughout the crystal (McPherson 2004). Because of this, protein crystals are fragile and diffract X-rays poorly. The extent to which the crystal diffracts X-rays, and ultimately the structural resolution achievable, is determined by the internal order of the crystal. Internal disorder arises because of the weak interaction between the loosely packed neighbour molecules. Lattice units can therefore occupy slightly different orientations, while protein flexibility means the backbone and side chains can occupy different conformations (McPherson 2004).

Approaches towards the crystallisation of proteins and macromolecules are based on a diverse set of principles, experiences, and ideas which have historically had no unifying theory (McPherson 2004). To form protein crystals the protein must be supersaturated in solution, after which a slow decrease in volume is intended to allow the protein to nucleate and crystallise by forming regular stacked units. Protein crystallisation trials contain many different conditions which attempt to find a combination of pH, salts, additives and precipitants that suit the individual requirements of different proteins to

crystallise. The purpose of the precipitant is to sequester water molecules from the protein solution so as to mimic the loss of solvent, and subsequently concentrate and supersaturate the protein. In vapour diffusion approaches, the protein is mixed with the precipitant, the drop then equilibrates with the precipitant and increases the protein and precipitant concentrations. This lowers the availability of water to the protein, decreasing its ability to remain in solution and increasing the possibility of crystal formation.

Nucleation is the start point of crystallisation and is the point at which the protein transitions from a disordered system to an ordered one (Manuel García-Ruiz 2003). At supersaturation the protein exceeds the saturation limit of the system, but faces a free energy barrier to transition in to the solid state, either crystal nucleus or precipitate, and so supersaturation is further enhanced as solute availability continues to decrease during vapour diffusion (McPherson 2004). As supersaturation increases, the energy barrier to nucleation decreases (Manuel García-Ruiz 2003). In order to visualise the formation of stable nuclei in solution one can imagine protein particles, which represent the lattice units, colliding in solution. Some of these units will interact to form nascent nuclei whose stability is determined by the forces holding them together and the forces pulling them apart (Manuel García-Ruiz 2003). The forces holding them together relate to the number and stability of the inter-unit bonds while the forces pulling them apart relate to the surface exposed to the solvent, and therefore stability is dependent on the volume to surface area ratio (Manuel García-Ruiz 2003). As the cluster of units reaches a critical volume, the forces pulling them apart become equal to the forces holding them together and the probability of the cluster falling apart becomes equal to the probability of the cluster remaining intact (Manuel García-Ruiz 2003). At this point the nucleus has become stable and the change in free energy becomes more favourable as the emergent crystal lattice grows (Manuel García-Ruiz 2003). Once nucleation occurs, the saturated protein continues to transition to the solid state, driving growth of the crystal until the protein is no longer supersaturated and equilibrium in the system is reached (McPherson 2004).

3.1.5 X-ray diffraction

An electron density map is deduced from the crystal through the diffraction of X-rays by electrons of the protein molecule. When an X-ray beam of a single wavelength is directed towards a crystalline protein molecule, the X-ray photons are scattered by the

electron cloud. When the scattering is recorded, a pattern of spots of different positions and intensities is seen. The different intensities of the spots relate to the amplitude of the reflected X-rays and therefore the amount of matter encountered (or electron density). The different positions of the spots relate to different sets of planes of electrons in the crystal scattering the X-rays at different angles. Bragg's law is as follows (Pecharsky & Zavalij 2008):

$$n\lambda = 2d_{hkl}sin\theta_{hkl}$$

$\lambda =$ the wavelength of light

$d_{hkl} =$ distance between planes of diffracting atoms

$\theta_{hkl} =$ the angle between the incident light and the crystal surface

Constructive interference occurs when n is an integer value.

Bragg's law dictates that the reflection of the X-ray wave by planes of electrons can result in constructive and destructive interference; this is due to the relationship between the phase at which the electromagnetic wave is scattered from an electron, when compared to the phase of scattering from another (Ilari & Savino 2008). If the wave is in exactly the same phase then constructive interference occurs and the amplitudes sum. Electrons emit X-rays in the same phase when they are in the same plane as each other, at which point the scattering will be equivalent to reflection. If electron density on a particular plane is high then there are more emitted X-rays that sum with one another and more intense spots are seen. Electrons in parallel planes also interfere with each other, so that only electrons in planes separated by a path-length equal to an integral number of wavelengths produce a signal (Ilari & Savino 2008). If the phase is even slightly out, then pairs of scattered waves which have opposite phases cancel each other out and no scattering is observed. This explains why the diffraction pattern presents as spots, since only diffracted X-rays with constructive interference can be seen.

3.1.6 Experimental aims

By using the approaches described above, the expression and purification of WWP2 proteins and some of its domains is to be explored. The purpose of this is to

produce samples of sufficient purity, concentration and stability that they may be used in crystallographic and NMR approaches. Considering the number of challenges along the way to obtaining structures, a broad approach will be taken initially, exploring the potential of each of the WWP2 isoforms and the HECT domain, and progressing with the most likely candidates. At first, proteins of interest are to be overexpressed in an *E. coli* host, and then, through the process of bacterial cell lysis and immobilised metal affinity chromatography, the proteins are to be extracted from the crude cell materials, after which size-exclusion chromatography is to be used to enhance purity even further. SDS-PAGE is to be used to analyse protein quality at each step. The use of crystallisation trials is to be applied to these samples in an attempt to produce crystals suitable for X-ray diffraction, so that a structure might be obtained. The hope is that this will provide new insights in to the conformation of WWP2 domains, providing information about the function of the WWP2 protein and its isoforms.

3.2 Results

Isoforms WWP2-FL, WWP2-N and WWP2-C were cloned in to the T7 bacterial expression vector pOPINF. The Rosetta 2 (DE3) pLysS *E. coli* strain was selected as a bacterial host to express the protein. This strain has a plasmid coding for tRNAs compatible with seven rare codons. These are not commonly found in the host DNA and their expression levels are therefore low, resulting in an increased likelihood of early termination of protein translation or incorrect codon pairing (Kane 1995). An abundance of rare codons can decrease the yield of protein on purification and increase the chance of incorrect residue incorporation. The WWP2-FL DNA sequence has a total of 77 of these codons as shown in Table 3.2.1, and therefore the use of the Rosetta 2 strain was considered suitable.

	Occurrence in WWP2 DNA	**Residue**
AGA	12	Arginine
AGG	3	Arginine
CGG	16	Arginine
CUA	2	Isoleucine
GGA	17	Leucine
CCC	24	Glycine
AUA	3	Proline

Table 3.2.1 Rare codons in the WWP2 DNA sequence

3.2.1 WWP2 isoform expression and solubility

Figure 3.2.1 shows the expression and solubility of WWP2 isoforms in the Rosetta 2 *E. coli* strain. Intense overexpression bands can be seen in each of the gels when comparing the post-induction lysate to the pre-induction lysates.

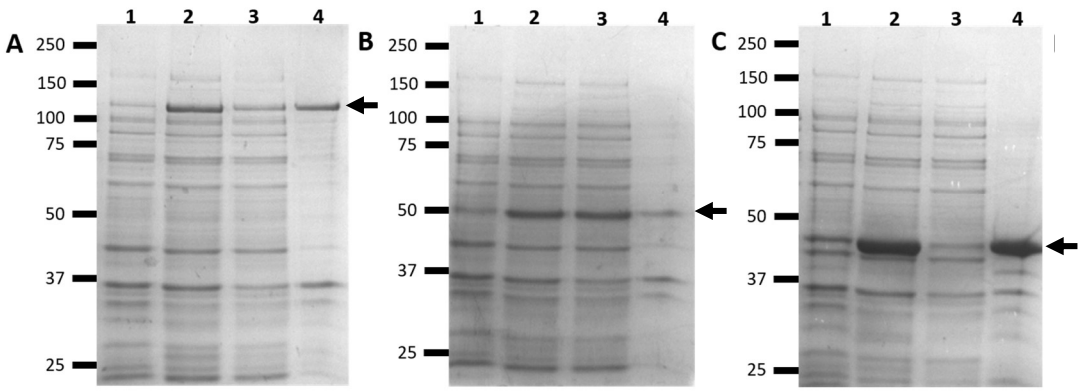


Figure 3.2.1 - SDS-PAGE analysis of WWP2 isoform protein expression and solubility after induction with 0.5 mM IPTG at 25°C for 4 hours in Rosetta 2 cells. Relevant protein bands are highlighted with arrows. A: WWP2-FL expected molecular weight 101 kDa. B: WWP2-N expected molecular weight 37.4 kDa. C: WWP2-C expected molecular weight 53 kDa. *1 - pre-induction whole cell lysate, 2 - post-induction whole cell lysate, 3 - soluble lysate, 4 - insoluble lysate*.

The WWP2-FL band can be seen just above the 100 kDa marker, running approximately at the expected molecular weight. The WWP2-N and WWP2-C bands ran at approximately 50 kDa and 45 kDa respectively. Despite the inconsistency between the apparent molecular weights and the expected molecular weights (WWP2-N 37.4 kDa and WWP2-C 53 kDa), plasmid sequencing confirmed the presence of the correct insert. In addition, MALDI-TOF mass spec analysis (performed by Dr Gerhard Saalbach at the John Innes Centre, Norwich) of the excised SDS-PAGE isoform bands from preliminary metal affinity purifications, showed peptide mass fingerprints covering the correct regions of WWP2 (Table 3.2.2, Table 3.2.3, Table 3.2.4).

WWP2-FL and WWP2-N expression levels were comparatively low, and attempts at purifying them resulted in low yields of low purity. WWP2-C shows significantly higher expression levels but the soluble and insoluble fractions in Figure 3.2.1C showed that the majority of the overexpressed protein was insoluble.

MASASSSR**AG**	**VALPFEK**SQL	TLKVVSAKPK	VHNRQPR**INS**	**YVEVAVDGLP**
SETKKTGKR**I**	**GSSELLWNEI**	**IILNVTAQSH**	**LDLK**VWSCHT	LR**NELLGTAS**
VNLSNVLKNN	GGK**MENMQLT**	**LNLQTENK**GS	VVSGGELTIF	LDGPTVDLGN
VPNGSALTDG	SQLPSRDSSG	TAVAPENR**HQ**	**PPSTNCFGGR**	SRTHRHSGAS
ARTTPATGEQ	SPGARSRHRQ	PVKNSGHSGL	ANGTVNDEPT	TATDPEEPSV
VGVTSPPAAP	LSVTPNPNTT	SLPAPATPAE	GEEPSTSGTQ	QLPAAAQAPD
ALPAGWEQRE	LPNGRVYYVD	HNTK**TTTWER**	**PLPPGWEK**RT	DPRGR**FYYVD**
HNTRTTTWQR	**PTAEYVRNYE**	**QWQSQRNQLQ**	**GAMQHFSQRF**	**LYQSSSASTD**
HDPLGPLPPG	**WEK**RQDNGRV	YYVNHNTRTT	QWEDPRTQGM	IQEPALPPGW
EMKYTSEGVR	**YFVDHNTR**TT	TFKDPRPGFE	SGTKQGSPGA	YDRSFRWK**YH**
QFRFLCHSNA	**LPSHVK**ISVS	R**QTLFEDSFQ**	**QIMNMKPYDL**	**R**RR**LYIIMR**G
EEGLDYGGIA	R**EWFFLLSHE**	**VLNPMYCLFE**	**YAGKNNYCLQ**	**INPASSINPD**
HLTYFRFIGR	**FIAMALYHGK**	**FIDTGFTLPF**	**YK**RMLNKRPT	LK**DLESIDPE**
FYNSIVWIKE	**NNLEECGLEL**	**YFIQDMEILG**	**K**VTTHELKEG	GESIR**VTEEN**
KEEYIMLLTD	**WR**FTRGVEEQ	TK**AFLDGFNE**	**VAPLEWLR**YF	DEKELELMLC
GMQEIDMSDW	QKSTIYRHYT	KNSK**QIQWFW**	**QVVK**EMDNEK	RIR**LLQFVTG**
TCRLPVGGFA	**ELIGSNGPQK**	FCIDKVGKET	WLPRSHTCFN	RLDLPPYKSY
EQLREKLLYA	IEETEGFGQE			

Table 3.2.2 WWP2-FL MALDI-TOF tryptic digest peptide matches against the WWP2-FL amino acid sequence. The WWP2-FL protein band was excised from an SDS-PAGE gel. Matches are shown in bold.

MASASSSR**AG**	**VALPFEK**SQL	TLKVVSAKPK	VHNRQPR**INS**	**YVEVAVDGLP**
SETKKTGKR**I**	**GSSELLWNEI**	**IILNVTAQSH**	**LDLKVWSCHT**	**LRNELLGTAS**
VNLSNVLKNN	GGK**MENMQLT**	**LNLQTENK**GS	VVSGGELTIF	LDGPTVDLGN
VPNGSALTDG	SQLPSR**DSSG**	**TAVAPENRHQ**	**PPSTNCFGGR**	SRTHRHSGAS
AR**TTPATGEQ**	**SPGAR**SRHRQ	PVKNSGHSGL	ANGTVNDEPT	TATDPEEPSV
VGVTSPPAAP	LSVTPNPNTT	SLPAPATPAE	GEEPSTSGTQ	QLPAAAQAPD
ALPAGWEQRE	LPNGRVYYVD	HNTKTTTWER	PLPPGWEKRT	DPRGRFYYVD
HNTRTTTWQR	PTAEYVRNYE	QWQSQRNQLQ	GAMQHFSQRF	LYQSSSASTD
HDPLGPLPPG	WEKRQDNGRV	YYVNHNTRTT	QWEDPRTQGM	IQEPALPPGW
EMKYTSEGVR	YFVDHNTRTT	TFKDPRPGFE	SGTKQGSPGA	YDRSFRWKYH
QFRFLCHSNA	LPSHVKISVS	RQTLFEDSFQ	QIMNMKPYDL	RRRLYIIMRG
EEGLDYGGIA	REWFFLLSHE	VLNPMYCLFE	YAGKNNYCLQ	INPASSINPD
HLTYFRFIGR	FIAMALYHGK	FIDTGFTLPF	YKRMLNKRPT	LKDLESIDPE
FYNSIVWIKE	NNLEECGLEL	YFIQDMEILG	KVTTHELKEG	GESIRVTEEN
KEEYIMLLTD	WRFTRGVEEQ	TKAFLDGFNE	VAPLEWLRYF	DEKELELMLC
GMQEIDMSDW	QKSTIYRHYT	KNSKQIQWFW	QVVKEMDNEK	RIRLLQFVTG
TCRLPVGGFA	ELIGSNGPQK	FCIDKVGKET	WLPRSHTCFN	RLDLPPYKSY
EQLREKLLYA	IEETEGFGQE			

Table 3.2.3 WWP2-N MALDI-TOF tryptic digest peptide matches against the WWP2-FL amino acid sequence. WWP2-N is shown in black font, and had several hits, the remainder of the WWP2-FL sequence is in grey, and had no hits. The whole sequence is shown because the server used to identify the protein from the mass spec profile did not discriminate against different WWP2 isoforms, and identified the full length isoform as the most likely candidate.

MASASSSRAG	VALPFEKSQL	TLKVVSAKPK	VHNRQPRINS	YVEVAVDGLP
SETKKTGKRI	GSSELLWNEI	IILNVTAQSH	LDLKVWSCHT	LRNELLGTAS
VNLSNVLKNN	GGKMENMQLT	LNLQTENKGS	VVSGGELTIF	LDGPTVDLGN
VPNGSALTDG	SQLPSRDSSG	TAVAPENRHQ	PPSTNCFGGR	SRTHRHSGAS
ARTTPATGEQ	SPGARSRHRQ	PVKNSGHSGL	ANGTVNDEPT	TATDPEEPSV
VGVTSPPAAP	LSVTPNPNTT	SLPAPATPAE	GEEPSTSGTQ	QLPAAAQAPD
ALPAGWEQRE	LPNGRVYYVD	HNTKTTTWER	PLPPGWEKRT	DPRGRFYYVD
HNTRTTTWQR	PTAEYVRNYE	QWQSQRNQLQ	GAMQHFSQRF	LYQSSSASTD
HDPLGPLPPG	WEKRQDNGRV	YYVNHNTRTT	QWEDPRTQGM	IQEPALPPGW
EMK**YTSEGVR**	**YFVDHNTR**TT	TFKDPRPGFE	SGTKQGSPGA	YDRSFRWK**YH**
QFRFLCHSNA	**LPSHVK**ISVS	RQTLFEDSFQ	QIMNMKPYDL	RRR**LYIIMRG**
EEGLDYGGIA	**R**EWFFLLSHE	VLNPMYCLFE	YAGK**NNYCLQ**	**INPASSINPD**
HLTYFRFIGR	**FIAMALYHGK**	**FIDTGFTLPF**	**YK**RMLNKRPT	LK**DLESIDPE**
FYNSIVWIKE	**NNLEECGLEL**	**YFIQDMEILG**	**K**VTTHELKEG	GESIRVTEEN
KEEYIMLLTD	WRFTRGVEEQ	TK**AFLDGFNE**	**VAPLEWLR**YF	DEKELELMLC
GMQEIDMSDW	QKSTIYRHYT	KNSK**QIQWFW**	**QVVK**EMDNEK	RIR**LLQFVTG**
TCRLPVGGFA	ELIGSNGPQK	FCIDKVGKET	WLPR**SHTCFN**	**R**LDLPPYK**SY**
EQLREK**LLYA**	**IEETEGFGQE**			

Table 3.2.4 WWP2-C MALDI-TOF tryptic digest peptide matches against the WWP2-FL amino acid sequence. WWP2-C is shown in black font, and had several hits, the remainder of the WWP2-FL sequence is in grey, and had no hits. The whole sequence is shown because the server used to identify the protein from the mass spec profile did not discriminate against different WWP2 isoforms, and identified the full length isoform as the most likely candidate.

3.2.2 WWP2-C purification and crystallisation trials

In an attempt to increase the solubility of WWP2-C during synthesis, the rate of expression was reduced by lowering the temperature after induction to 20°C. Figure 3.2.2 lanes 1-3 shows the expression and solubility of WWP2-C using the altered conditions. Despite the abundance of protein in the insoluble fraction, it appears that at least some of the protein was present in the soluble fraction, although it is hard to distinguish because of the high concentration of bacterial proteins. When the soluble lysate was

passed through a nickel column, the WWP2-C protein eluted on an imidazole gradient (lanes 1-15).

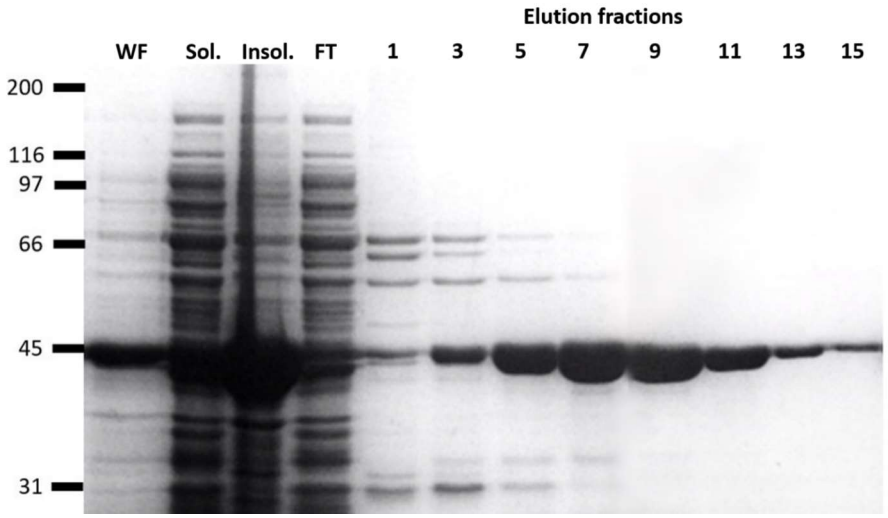


Figure 3.2.2 - SDS-PAGE analysis of WWP2-C expression and solubility at 20°C and Ni-NTA purification. Whole fraction (WF), soluble fraction (sol.), insoluble fraction (insol.), column flow through (FT) and aliquots of 5 ml elution fractions (1-15) collected around the 280 nm absorbance peak.

The imidazole gradient separated some of the non-specifically bound protein, which can be seen in the early fractions, from the bulk of the WWP2-C protein, which, by observation of the gel, was pure and of high concentration. When the protein was gel filtered using a HiLoad 16/600 Superdex 75 prep grade column the protein eluted as a symmetrical minor peak early in the run at 46 ml, and a symmetrical major peak, with between 3-6x greater peak absorbance, later in the run at 58 ml. Figure 3.2.3A shows the absorbance trace of the column eluate at 280 nm. Figure 3.2.3B shows an SDS-PAGE gel of the gel filtration elution fractions of WWP2-C with the first peak visible in fraction 5 and the second peak from fraction 9-17.

The column had been calibrated previously by using a set of protein standards to give the following straight line equation:

$$y = -37.721x + 122.67$$

$y = \text{elution volume}$

$x = \log_{10} \text{molecular weight}$

Using this equation, the apparent molecular weight of the first peak is 107.8 kDa and the apparent molecular weight of the second peak is 51.8 kDa which seems to correspond well with a WWP2-C monomer (53 kDa).

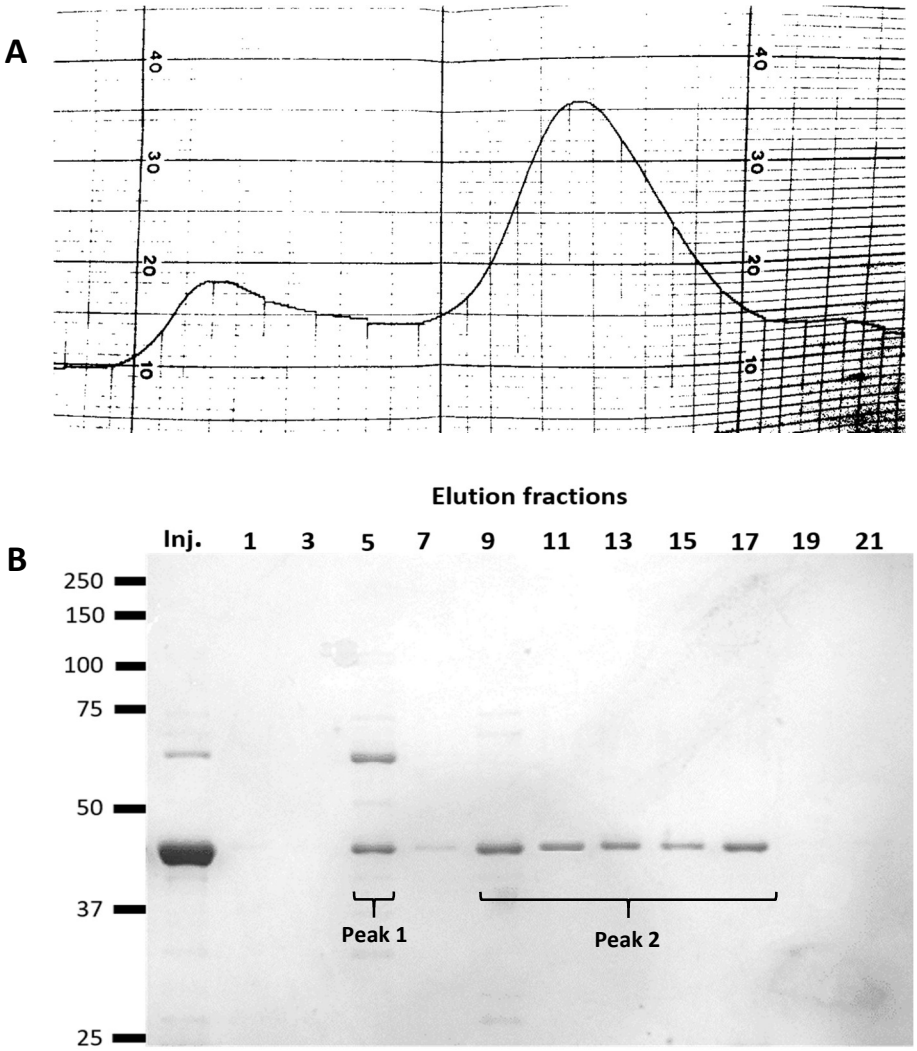


Figure 3.2.3 - A: A scanned paper trace 280 nm absorbance profile of the WWP2-C gel filtration run (acquired using an old ÄKTA). B: SDS-PAGE analysis of the WWP2-C gel filtration. An aliquot of the Injected sample (inj.), and aliquots of alternate fractions of the column eluate. The first peak spans fractions 5-6 (an elution volume of between 44-48 ml), and the second peak spans fractions 9-17 (an elution volume of between 54-70 ml). The corresponding fractions were pooled as separate peaks.

While it is tempting to speculate that the first peak might correspond to a WWP2-C dimer, especially as there is evidence of dimeric and oligomeric NEDD4 E3 ligase activity (Aragón et al. 2012; Todaro et al. 2015), it is also possible that WWP2-C at this apparent molecular weight is in complex with a bacterial protein, which could be responsible for

the band in fraction 5 at roughly 65 kDa. It is also important to note that the first peak eluted very close to the void volume (45 ml) and that molecular weight prediction at this volume becomes increasingly inaccurate, since resolution close to the void volume is poor, also raising the possibility that the peak represents partially aggregated protein. The major peak was trialled for crystallisation at 15 $mg.ml^{-1}$ (as measured using absorbance at 280nm with an extinction coefficient of 88280 $M^{-1}.cm^{-1}$, as determined using the ProtParam ExPasy online server (Wilkins et al. 1999)) using 1:1 and 1:2 ratios of protein to trial condition at 4°C and 16°C. After over 8 weeks of observation the crystallisation trials yielded no suitable crystalline material.

In an attempt to improve the crystallisation of WWP2-C, His-tagged 3C protease was used to cleave the purification tag of WWP2-C using the cleavage site c-terminal to the six histidines. Figure 3.2.4 shows the successful full cleavage of the His-tag in lanes 1 and 2, however when the sample was passed through a nickel column so as to remove the 3C protease and the tag peptide, no WWP2-C protein flowed through the column (lane 3). Instead, when an elution gradient was performed on the column, the WWP2-C protein elutes, seemingly interacting non-specifically with the column matrix (lanes 4-14).

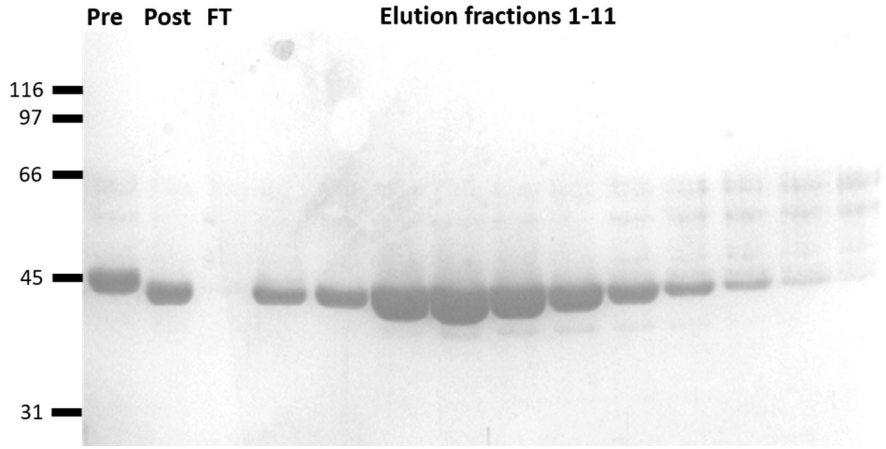


Figure 3.2.4 - SDS-PAGE analysis of the digestion of WWP2-C, performed using HRV 3C protease and subsequently passed through a Ni-NTA column to separate undigested and digested protein. Pre-digest sample (lane 1), post-digest sample (lane 2) showing the appropriate shift in molecular weight, the column flow through (lane 3) and the imidazole gradient elution fractions 1-11 (lanes 4-14). Fractions 1-7 were pooled.

To remove the His-tag peptide and 3C-protease, the protein was purified using gel filtration. Since the molecular weights are much smaller than the WWP2-C protein (approximately 1 kDa and 20 kDa respectively), the column resolved the proteins

sufficiently. WWP2-C without the His-tag was trialled for crystallisation at 2 $mg.ml^{-1}$ and 4 $mg.ml^{-1}$ (as measured using absorbance at 280nm with an extinction coefficient of 82780 $M^{-1}.cm^{-1}$, as determined using the ProtParam ExPasy online server (Wilkins et al. 1999)). After over 8 weeks of observation these crystallisation trials also failed to yield suitable crystalline material. A further two trials were set at 6.8 $mg.ml^{-1}$, one of which contained a short Smad7 peptide containing the PPxY motif in a molar ratio of 1:1, based on previous experiments that suggest Smad7 is a binding partner of WWP2-C (Soond & Chantry 2011). Condition 19 of Structure Screen I (0.2 M zinc acetate dihydrate, 0.1 M sodium cacodylate, pH 6.5, 18% w/v PEG 8000) with the WWP2-C/Smad7 mix produced a small crystal after 3 weeks of observation. The condition was optimised by varying the pH from 5.9-7.1 in increments of 0.2, and the percentage of PEG 8000 from 14%-24% in increments of 2%, zinc acetate and sodium cacodylate were kept at the same concentration. After 2 weeks, crystals of varying sizes formed in 11 of the conditions. The most promising candidates were at well B2: 16% PEG 8000 pH 6.1, well D3: 20% PEG 8000 pH 6.3 and well F6: 24% PEG 8000 pH 6.9. Glycerol at 20% was used as a cryopreservative and one crystal from B2 and D3, and two crystals from F6 were snap frozen in liquid nitrogen. The crystals were sent to the Diamond Light Source facility, but it was found that these crystals gave no detectable diffraction.

It was thought that perhaps the stretch of amino acids linking the WW4 domain and the HECT domain might be preventing crystallisation due to a lack of intrinsic order. To identify parts of the sequence that might be disordered, the WWP2-C amino acid sequence was submitted to the online server DISOPRED. DISOPRED predicts the probability of residue disorder by identifying sequence patterns associated with disorder. DISOPRED defines disorder as residues that lack coordinates in the electron density maps of high-resolution X-ray crystal structures (Available at: http://bioinf.cs.ucl.ac.uk/psipred) (Ward et al. 2004). The server returned the disorder profile shown in Figure 3.2.5A. Part of the linking sequence between the two domains is indeed predicted to have some level of intrinsic disorder, as well as the N and C-termini. In addition, the secondary structure prediction web server PSIPRED also predicted that this region lacks any distinctive secondary structure (Figure 3.2.5B); although the server also failed to predict the third strand of the WW4 β-sheet (Jones 1999). To address this issue, a new construct was designed that looked exclusively at the HECT domain, excluding the WW4 domain and the linking region.

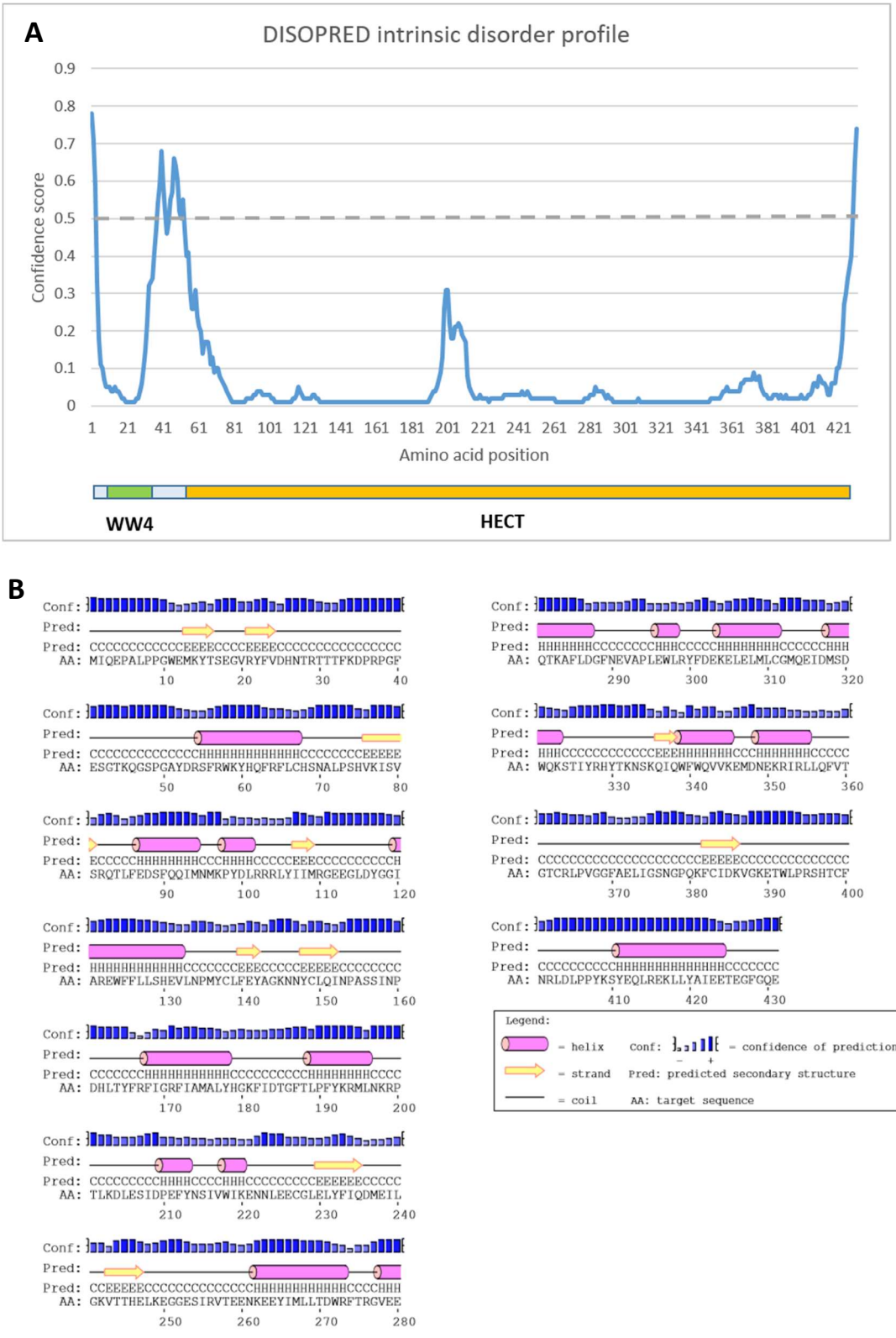


Figure 3.2.5 - A: DISOPRED intrinsic disorder profile of WWP2-C. Residues with a confidence score over 0.5 (grey line) are considered disordered. Disordered amino acids: 1-3, 38-42, 45-50, 52 and 429-431. WW4 domain residues 4-38, HECT domain residues 56-431. B: PSIPRED secondary structure prediction for the WWP2-C sequence.

3.2.3 WWP2-HECT construct design and expression

The HECT domain of the closely related WWP1 has previously been crystallised (Verdecia et al. 2003) and the sequence shares 83% identity with that of WWP2 HECT. Because of this similarity, the construct used in WWP1 HECT crystallisation was used to inform the design of the WWP2-HECT construct in conjunction with the secondary structure prediction from PSIPRED. The start of the construct was positioned towards the beginning of the first secondary structure feature after the WW4 domain, a helix, at residue position 495 and like the WWP1 construct the last five residues at the C-terminus were excluded, terminating at residue position 865 (both relative to the WWP2-FL amino acid sequence, residues 56-426 relative to WWP2-C). Figure 3.2.6A shows the expression and solubility of WWP2-HECT.

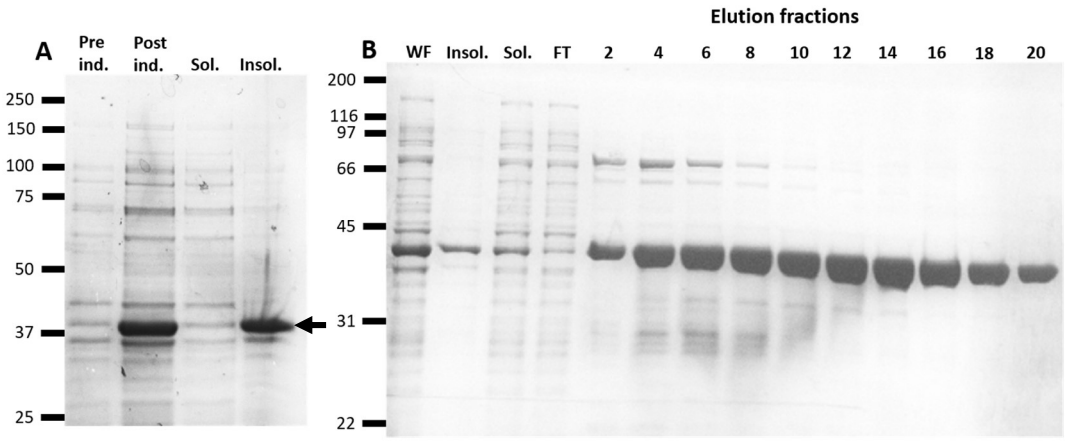


Figure 3.2.6 - A: SDS-PAGE analysis of WWP2 HECT protein expression and solubility after induction with 0.5 mM IPTG at 25°C for 4 hours in Rosetta 2 cells, pre-induction, post-induction, soluble and insoluble fractions. Expected molecular weight: 46.5 kDa B: SDS-PAGE analysis of WWP2 HECT protein expression, solubility and purification by Ni-NTA after induction with 0.1 mM IPTG at 20°C overnight. Whole fraction, insoluble fraction, soluble fraction, flow through from the nickel column and aliquots of the fractions collected around the 280 nm absorbance peak during elution with an Imidazole gradient. Fractions 10-20 were pooled. Note, different markers and different % acrylamide are used for each gel, A: 10% and B: 15%.

The molecular weight of the overexpression band appears at approximately 38 kDa. As with the isoform expression, DNA sequencing confirmed the presence of the correct insert and MALDI-TOF mass spec analysis of the SDS-PAGE band

showed peptide mass fingerprints covering the correct region of WWP2. Similar to WWP2-C, expression is reasonably high but solubility is poor and there is no visible HECT band in the soluble fraction. As with WWP2-C, the expression temperature was reduced to 20°C and since WWP2-C still showed a high level of insoluble protein at this temperature, the IPTG concentration was also reduced to 0.1 mM to help encourage more protein to remain soluble. The first three lanes in Figure 3.2.6B show the increased solubility at this temperature and IPTG concentration, with around 50% remaining soluble. When the protein was purified using metal affinity chromatography, a protein of high concentration and reasonable purity was recovered as seen in the elution fractions of Figure 3.2.6B. When the protein was gel filtered using a HiLoad 16/600 Superdex 75 prep grade column HECT also eluted as two peaks: one sharp symmetrical peak at 47.7 ml and one broad peak eluting at 59 ml, as shown in Figure 3.2.7A. The two peaks are merged in this profile because of the high concentration of the injected protein. The calibration equation for this column, which is shown below, gives apparent molecular weights of 97 kDa and 45 kDa for the first and second peaks respectively.

$$y = -1.3341x + 6.3953$$

$y = \text{elution volume}$

$x = \log_{10} \text{molecular weight}$

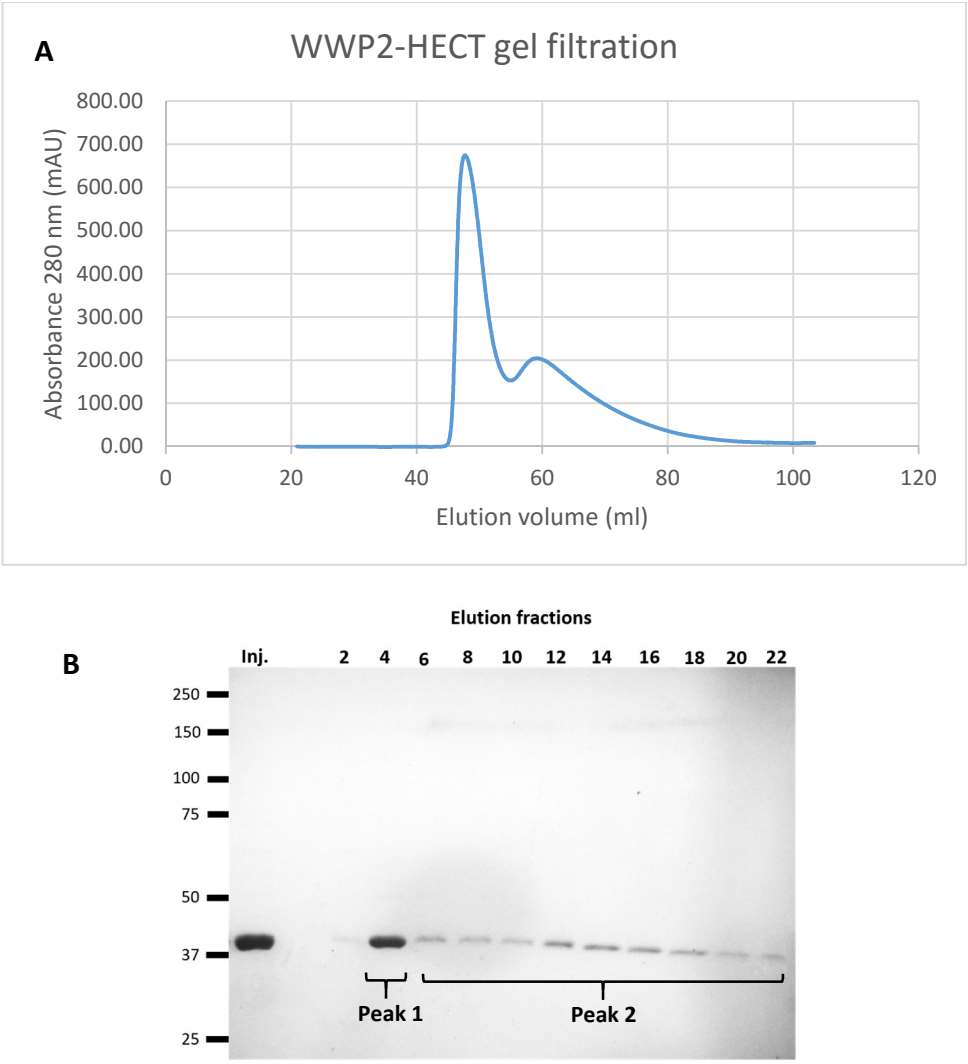


Figure 3.2.7 - A: WWP2 HECT gel filtration 280 nm absorbance profile showing two peaks. B: SDS-PAGE analysis of the WWP2-HECT gel filtration elution fractions, showing the first peak in fraction 4 (elution volume 48 ml), and the second peak in fractions 6-22 (elution volume 52-84 ml). Fractions 2-5 were pooled as peak 1 and fractions 6-22 were pooled as peak 2.

WWP2-C, the first peak is close to the void volume (45.2 ml), so the same difficulties predicting molecular weight also apply, and the possibility that the peak represents partially aggregated protein also stands. The two peak elution can be seen in the SDS-PAGE analysis of elution fractions (Figure 3.2.7B), with the first peak visible in fractions 2-6 and the second peak visible in fraction 8-22. Unlike WWP2-C, the elution fractions of the first peak contain exclusively HECT protein and it is possible that these two peaks could represent HECT in the monomeric and dimeric forms, the molecular weights of which would be 46.5 kDa and 93 kDa, respectively. To rule out the possibility of a cystine-mediated dimer, the protein was run on an SDS-PAGE gel in the absence of

the reducing agent DTT and in these conditions HECT still migrated as a monomer (data not shown). However, a native SDS-PAGE gel was not performed, which might have observed weaker non-covalent interactions under non-denaturing conditions.

The peaks were pooled separately and HRV 3C protease was added to each, so as to remove the His-tag. HECT proved to be a difficult protein to cleave. Several attempts were made to cleave the protein directly after Ni-NTA purification with no success and it was only after the protein was gel filtered that cleavage became possible, and even then only with a partial digest. Typically this would not be a problem as the digest is passed through the nickel column again, removing undigested protein, free His-tag peptide, and the His-tagged protease, allowing the cleaved protein to flow through the column. Figure 3.2.8 shows the partial cleavage of both peaks, the column flow through for both peaks, which should contain the cleaved protein, and the eluate which should exclusively show the undigested protein. However, as with WWP2-C, HECT binds non-specifically to the column and the digested and undigested protein is not separated.

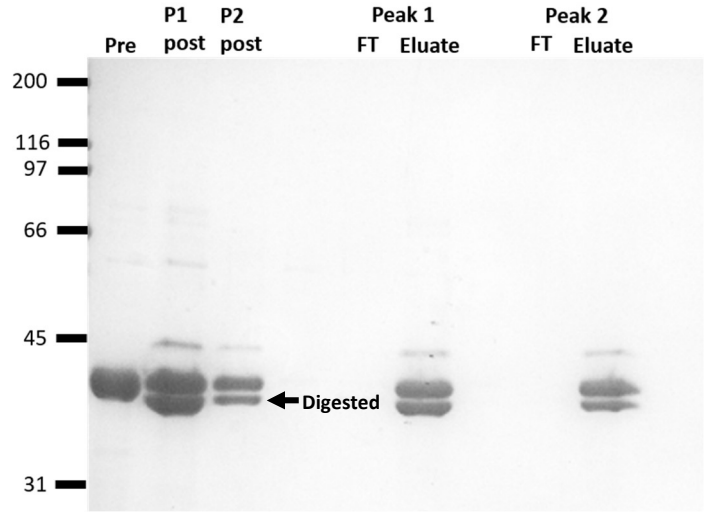


Figure 3.2.8 - SDS-PAGE analysis of the digestion of WWP2 HECT protein, performed using HRV 3C protease. The digest was performed on both of the peaks that eluted from the gel filtration run. Shown is the pre-digestion sample (Pre) and the post-digestion sample of peaks 1 and 2 (P1 and P2 post), and the Nickel column flow through and eluate after washing with high imidazole for both peaks.

Imidazole and salt gradients also failed to separate the digested and undigested proteins and the S75 gel filtration column lacked the resolution to separate proteins of such a similar molecular weight. Crystallisation trials were therefore performed on HECT with the His-tag still intact, at 2.7 $mg.ml^{-1}$ and 10 $mg.ml^{-1}$ (as measured using absorbance

at 280 nm with an extinction coefficient of 78310 $M^{-1}.cm^{-1}$, as determined using the ProtParam ExPasy online server (Wilkins et al. 1999)). In both screens PACT premier returned one crystal in well F1 which contained 0.2 M Sodium fluoride, 0.1 M Bis-Tris propane, 20 % w/v PEG 3350, pH 6.5. When the crystal was harvested it was robust and very resistant to physical manipulation. Since protein crystals are very sensitive to external forces, and disintegrate easily due to extensive solvent channels throughout the crystal, it was determined that the crystal was inorganic. The general lack of crystalline material led to the conclusion that without cleaving the His-tag, a comprehensive crystallisation trial would be unachievable. Some effort was made to express the protein as an untagged variant, since the protein seemed to bind the NiNTA matrix even without the six-histidine repeat. However, the protein was insoluble even at the low temperature and low IPTG concentration used for the tagged variant. Instead of investing more time in the crystallisation of HECT and the WWP2 isoforms, focus was changed to the protein interaction domains of WWP2.

3.2.4 Expression and purification of WW4

Previous attempts by this lab to express all four WW domains as a His-tagged recombinant gave only insoluble protein, and attempts at expressing the individual domains as His-tagged recombinants gave poor yields, particularly for the first and fourth WW domains. WW4 is of particular interest because of its presence in the putative tumour promoter WWP2-C, and therefore a new approach was taken to express this domain. Since this domain is small enough for structural analysis by NMR, it was decided that this approach would be taken, so as to avoid the potentially terminal issue of protein crystallisation. The B1 domain of the streptococcal protein G (GB1) has been found to significantly enhance solubility, expression and stability (Zhou & Wagner 2010; Hammarström et al. 2006; Huth et al. 1997; Hammarström et al. 2002). All three of these qualities lend themselves to successful preparation of a highly concentrated protein sample that has to remain stable at room temperature for long periods of time while NMR experiments are acquired. The small size of GB1, at 7.5 kDa, means that the tag can remain attached during NMR experiments, avoiding the challenges that can occur during tag cleavage. Furthermore, the tag is reported to be passive, whereby the tag does not

interact with the fused protein or protein-protein complexes, and so is amenable to protein interaction studies (Zhou & Wagner 2010).

WW4 was cloned in to pSKDuet01 to produce a GB1 recombinant with a thrombin-cleavable His-tag (purification by IgG is possible, but His-tag nickel affinity is cheaper). The competent *E. coli* strain BL21 Star (DE3) was chosen to express WW4 because of the need for high protein yields for NMR experiments. BL21 Star is designed to enhance protein expression because it carries an inactive mutant RNase and mRNA is subsequently stabilised, enhancing protein yield.

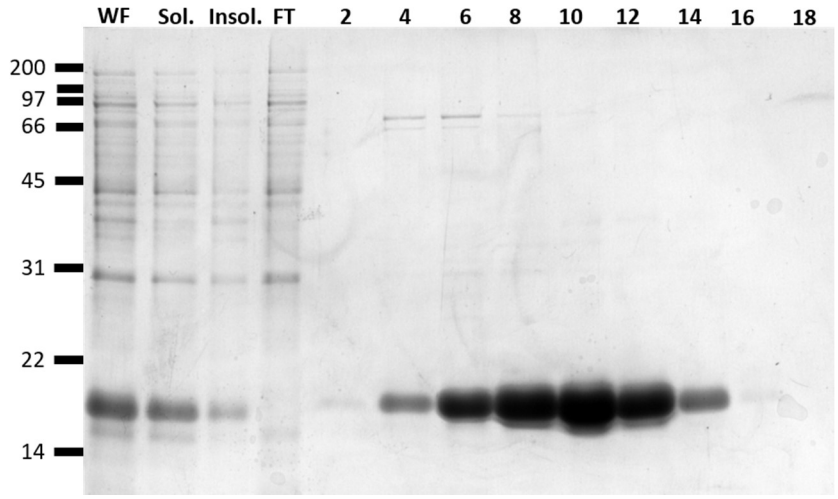


Figure 3.2.9 - SDS-PAGE analysis of GB1:WW4 expression and solubility at 30°C, and Ni-NTA purification. Shown are the whole fraction (WF), soluble fraction (Sol.), insoluble fraction (Insol.), nickel column flow through (FT) and nickel column elution fraction samples. Expected molecular weight 13.8 kDa. Fractions 4-14 were pooled.

Figure 3.2.9 shows the expression, solubility and Ni-NTA purification of GB1:WW4. Solubility is high and the purity and yield of the Ni-NTA purification are also very good. The band runs heavier than expected, but experience with the GB1 tag in this lab and our collaborators, indicates that the GB1 tag causes proteins to run at heavier weights than expected. DNA sequencing was performed to confirm the presence of the correct insert.

Figure 3.2.10A shows the digest of GB1:WW4. The post digestion sample shows two bands that correspond to the digested and undigested protein. The flow through contains only the digested protein, while the eluate contains only the undigested protein which had remained bound to the column. Figure 3.2.10B and C shows the gel filtration profile. The protein elutes as a symmetrical peak at 74.17 ml with an apparent molecular

weight of 16 kDa, compared to the expected molecular weight of 12 kDa for GB1:WW4 with the His-tag cleaved. However, the molecular weight prediction is based on a calibration calculated from gel filtration of a series of globular proteins. The calibration equation can therefore be inaccurate for the molecular weight prediction of non-globular proteins, which most likely applies to the GB1:WW4 recombinant protein because of its two non-interacting domains. The WW4 domain is suitable for further experiments and possesses the qualities required for structural analysis by NMR.

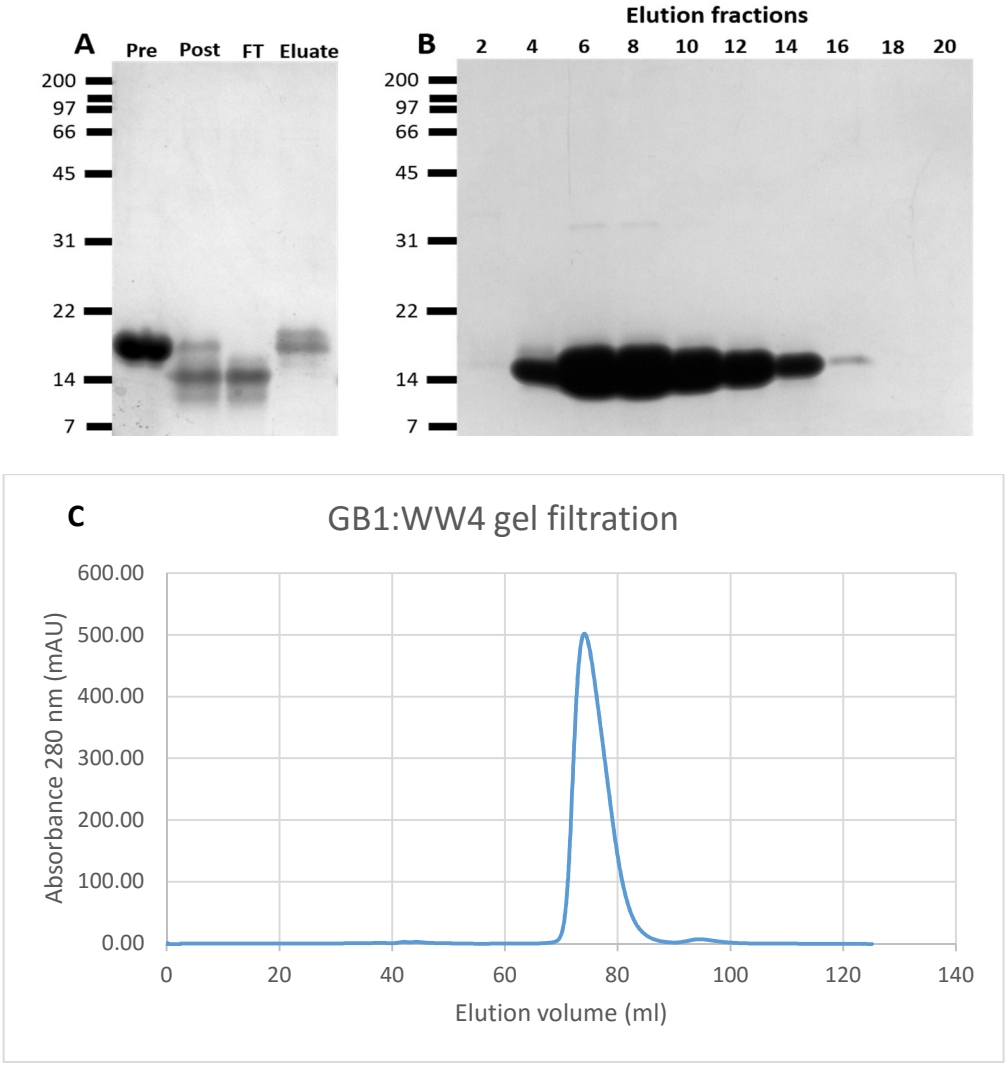


Figure 3.2.10 - A: SDS-PAGE analysis of the GB1:WW4 His-tag thrombin cleavage. After digestion with thrombin, the protein was passed through a nickel column to separate digested and undigested protein. Pre-digestion, post-digestion, nickel column flow through (digested) and nickel column immidazole eluate (undigested) fractions are shown. B: SDS-PAGE analysis of the GB1:WW4 gel filtration peak, fractions 4-14 (elution volume 68-88 ml) were pooled. C: The GB1:WW4 gel filtration 280 nm absorbance profile.

3.3 Discussion

The purification of WWP2 has been met with difficulty amongst at least one other lab group (Jiang, Zheng, et al. 2015). The approach used here is to explore avenues of research that are promising, while avoiding prolonged troubleshooting of approaches that are problematic and can be very time consuming. This method is appropriate for WWP2 as we have a broad interest in the whole of the structure, since there has been little structural information about this E3 ligase. It is for this reason that perhaps our lab, which is new to the field of structural analysis, has struggled with crystallisation of proteins. Many crystallisation labs use a blanket approach to protein crystallisation whereby many constructs are screened, each with slightly differing terminal regions. It is analogous to the multitude of screening conditions used to tempt the protein into crystallising, as it is not known exactly what characteristics of the protein cause a protein to crystallise in one condition over another. The high-throughput approach to construct design is an alternative to the informed approach used here, where an attempt is made to take characteristics such as disorder and secondary structure into account. The attempted crystallisation of the whole WWP2-C protein was perhaps a little ambitious, but the desire to see the structure of the entire intact isoform was too tempting to ignore. Reflecting on this, if more time were available a new recombinant, or several, with shortened or extended N and C termini, but still retaining WW4 and HECT, might present a greater hope of yielding a crystal.

In a recent Protein expression and purification paper, the authors publish findings on the expression and purification of WWP2 HECT (Jiang, Zheng, et al. 2015). The authors screened 96 constructs designed to find a soluble HECT domain, including the exact construct used here from residues 495-865. During systematic screening, however, they found that this construct, amongst others, resulted in inclusion body formation. They encountered problems with low yield and aggregation after purification. This insolubility is a characteristic consistent with the expression of WWP2 HECT here, however with a slight alteration of expression conditions it was possible to obtain soluble protein that purified with a significant yield. Similar to the results here, this group settled for a low expression temperature (23°C), and low IPTG concentrations (0.2 mM) to express a construct that covered the last two WW domains of WWP2, and the HECT domain (Jiang,

Zheng, et al. 2015). The result of this paper is a statement of intent to crystallise this protein.

Another paper was released recently from a group in which the authors report the crystal structure of WWP2 HECT (Gong et al. 2015). The paper reports that they initially trialled a construct stretching from residues 486-870, compared to 495-865 used here (residue numbers relative to the WWP2-FL protein), but found aggregation at high concentrations. After trialling various N and C-terminal modifications in crystallisation trials, they found a recombinant stretching from residues 486-865 gave the most suitable crystals in 0.1 M HEPES pH 8.4, 0.2 M MgCl2, 15% ethanol at 4°C. Interestingly, the recombinant stretches in to a region predicted to have no secondary structure characteristics by PSIPRED, through the region of disorder predicted by DISOPRED (Figure 3.2.5), but copies the C-terminal 5 residue truncation from the WWP1 HECT crystallisation publication by Verdecia *et al.,* in 2003 (Verdecia et al. 2003). Whether this reflects a flaw in the attempt to follow certain principles in the design of constructs for protein crystallisation, improved crystallisation conditions, or a result of the acknowledged flaws in the computer prediction algorithms is a matter up for debate. Figure 3.3.1 shows the ribbon cartoon of the WWP2-HECT structure.

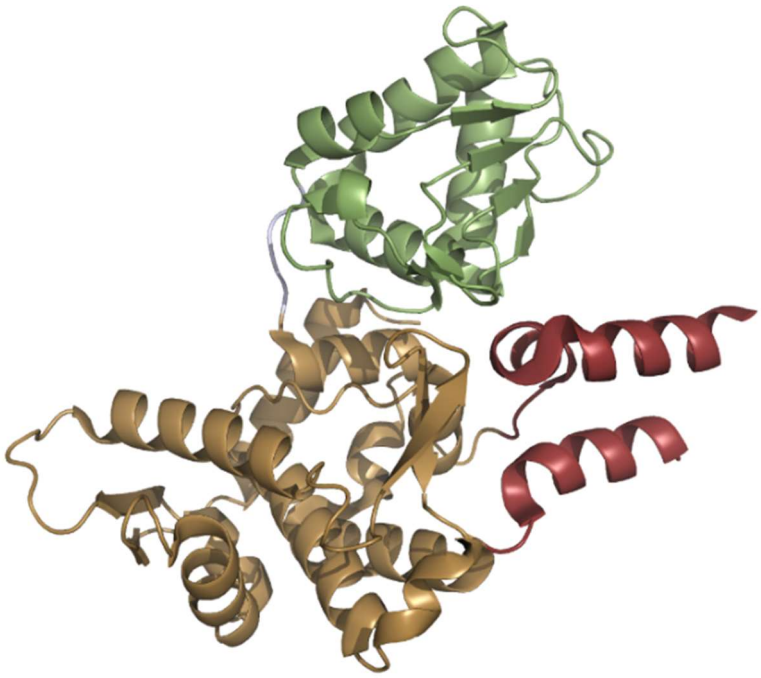


Figure 3.3.1 - The structure of the HECT domain of WWP2-HECT, with the C-terminal lobe in green, the N-terminal lobe in brown, and the E2 interaction interface in red (PDB: 4Y07) (Gong et al. 2015).

The structure is very similar to other HECT domains, and retains the bi-lobal structure found in the other HECT structures. In this crystal structure the C-terminal lobe is in the 'closed' formation, with the catalytic cysteine in relatively close proximity to the E2 binding site, creating an inverted T shape. A large portion of the E2 interaction interface is missing because of disorder in the protein crystal. The author comments on the different conformational states presented by each HECT domain despite the conservation among key residues, but fails to emphasise that these structures are from crystals, and therefore the conformation that each HECT domain structure holds is the conformation most amenable to crystallisation; and that these are potentially a range of conformations about which the C-terminal lobe rotates during the course of its activity.

The use of the GB1 solubility tag to express the WW4 domain has solved a problem that has hindered the progression of the structural work on the WWP2 WW domains. The GB1 tag has also been used by this group to enhance the expression and stability of the WW1, WW2 and WW3 domains, with differing levels of effectiveness. The GB1:WW4 recombinant works very well and is stable at room temperature for prolonged periods, and as well as being pure and highly concentrated, the protein can be expressed in minimal essential medium (MEM). Rich growth media often contain a source of amino acids, which the cultured bacteria can use to synthesise proteins. NMR requires protein samples to be uniformly labelled with isotopes, so if this approach were to be used to grow labelled proteins, the amino acids added to the media must also be uniformly labelled. It is possible to do this, but it is also very expensive. MEM has no amino acids added to the medium, and the nitrogen and carbon sources available to the bacteria are restricted. This forces the bacteria to synthesise amino acids from the isotopically labelled nitrogen and carbon sources added to the medium. The labelled amino acids are incorporated in to the overexpressed protein, which is consequently uniformly labelled. This method of protein expression is necessary for isotopic labelling, which is needed for the multi-dimensional NMR experiments required for structural elucidation. The GB1:WW4 protein is therefore suitable for the application of NMR experiments designed for the purpose of probing protein structures in solution. This will be explored further in the next chapter.

4. WW4 Domain Structure

4.1 Introduction

In Chapter 3, the use of the GB1 solubility tag allowed the expression and purification of a protein that possesses the qualities required for analysis by NMR spectroscopy. In this chapter the GB1:WW4 recombinant protein will be subject to structural analysis using solution state NMR. The process involves a series of steps that starts with the acquisition of the appropriate spectra, progress to resonance assignment and then finish with two rounds of structural computation using two different software packages. The results section of this chapter will include a discription of the progression through these steps for the GB1:WW4 protein, followed by an evaluation of the structure. The introductory section of this chapter will include a brief theoretical explanation behind the technique of NMR spectroscopy, from the understanding of a biologist.

4.1.1 Principles of NMR spectroscopy

NMR spectroscopy exploits a property of atomic nuclei called spin and to explain it we have to delve in to the physics of atoms and electromagnetic energy. This is best visualised using the vector model which will be described below, but it should be noted that this model does not go so far as to fully explain all of the quantum mechanical phenomena exploited in NMR spectroscopy.

It is common to hear about the difficulty in conceptualising spin, as there is no macroscopic equivalent. Spin is not rotation, but a physical property intrinsic to subatomic particles that possesses angular momentum and is described as a vector (Levitt 2001). The subatomic particles that compose a nucleus are neutrons and protons and each of these have a value of spin ½ (Levitt 2001). The overall quantum spin of atomic nuclei is the result of the coupling between proton and neutron, leading to a combination of its spin values (Levitt 2001). Hydrogen (^{1}H) for example contains one proton and has spin ½, while deuterium (^{2}H) has one proton and one neutron and has spin 1. Table 4.1.1 shows the general rules outlining the atomic spin of different elements or isotopes. Nuclei with spin ½ or greater have a magnetic moment associated with them. Nuclei with spin ½ generate a small dipolar magnetic field along the axis perpendicular to the direction of their angular momentum (Levitt 2001). In NMR spectroscopy, spectra generated using atomic nuclei with spin ½ are

the easiest to interpret, and it is for this reason that proteins are isotopically labelled with ^{15}N and ^{13}C during expression.

Number of Protons	Number of Neutrons	Spin
Even	Even	0
Odd	Even	Half-integer (½, 1½…)
Even	Odd	Half-integer (½, 1½…)
Odd	Odd	Integer value (1, 2...)

Table 4.1.1 The rules outlining the atomic spin of elements and isotopes with different numbers of neutrons and protons (Levitt 2001).

Nuclei with spin ½ have two opposing orientations, represented by *m* and depend on whether the spin is up or down (or precessing clockwise and anticlockwise, respectively). These two positions have the same energy in the absence of an external magnetic field, and nuclear spin vectors are therefore randomly oriented across a given sample. However, when the spin is placed in a magnetic field (B_0), the energy levels of the two spin orientations split in to a low energy (α) and a high energy (β) state as the spin aligns with the field or against the field, respectively (Figure 4.1.1A and B) (Keeler 2011; Levitt 2001). The difference in energy between these two orientations is directly proportional to the Larmor frequency (Keeler 2011). The Larmor frequency is calculated by the following equation:

$$\omega_0 = -\gamma B_0$$

Where ω_0 is the Larmor frequency, γ is the gyromagnetic ratio and B_0 is the applied magnetic field. This means that the energy required for a spin to transition from being aligned with the field to being aligned against the field, is related to the strength of the magnet being used and the nucleus being observed, since different elements and isotopes have different gyromagnetic ratios (Levitt 2001).

In NMR spectroscopy we observe a large number of atoms in a given sample. Once the magnetic field is applied, and after a period of equilibration, there will be a slight preference for nuclear spins across this large number of atoms to align with the field and assume the low energy state, as determined by Boltzmann distribution. In reality, the nuclei will be aligned in a variety of orientations that vary in their proximity to the field alignments, but when we consider an average of these orientations we find a bulk magnetisation vector aligning with the B_0 field in the low energy state (Keeler 2011). The bulk magnetic moment

aligns to the z-axis and has no x or y component because there is no energetic preference for distribution about the x and y axes.

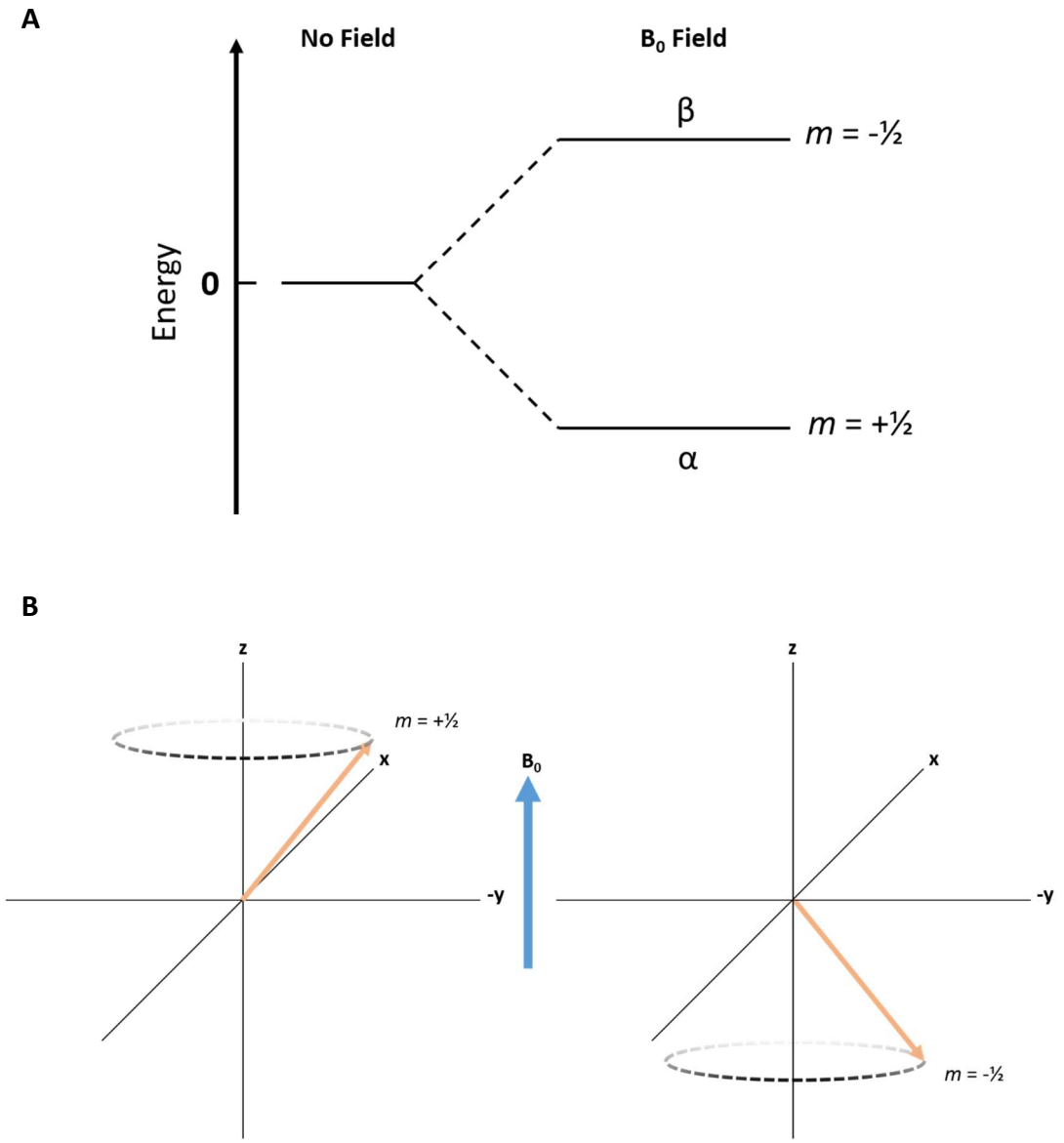


Figure 4.1.1 - A: An energy diagram showing the split in spin energy states for a nucleus with spin ½ upon the application of an external magnetic field. B: The magnetisation vector alignment and precession of the two energy states about the axis of the applied magnetic field, represented in Cartesian space. Adapted from (Keeler 2011).

As the magnetic field is applied to the sample, it interacts with the small magnetic moments, and the magnetic moment begins to precess around the axis of the B_0 field (Figure 4.1.1B). The frequency of this precession is the same as the Larmor frequency. When a radio frequency field (the B_1 field) is applied along the x axis, which is on resonance with the frequency of the precession (i.e. the Larmor frequency), the correct energy is applied to

cause individual spins to transition from the low energy state to the high energy state (Keeler 2011). When this pulse is applied for the appropriate period of time, the distribution of spins across the two energy levels becomes equal. At this point the z component of the bulk magnetisation vector diminishes, until there is no bulk magnetisation along the z axis. The magnetic moment precessions become coherent, that is they are in the same phase, and produce an observable component in the xy plane. This has the effect of rotating the bulk magnetisation 90° to point along -y (assuming a rotating frame), where it rotates about the z axis in the xy plane (Figure 4.1.2A and B) (Keeler 2011). The magnetisation oscillates in the x and y axes producing sine and cosine waves, respectively (Figure 4.1.2B).

The basic pulse-acquire experiment has three steps (Figure 4.1.2C), an initial period during which the sample, having been newly introduced to the B_0 field, is allowed to equilibrate. It is during this period that the bulk magnetisation vector emerges in the z axis. During the second period an RF field is applied along the x axis to induce a 90° rotation of the bulk magnetisation vector. During the third period, the free induction decay (FID) signal is recorded. The FID is generated by the individual magnetic moments which are precessing coherently. However as time goes on, the coherence diminishes and the strength of the signal is lost, until the individual magnetic moment vectors are completely out of phase. At this point there is no net magnetisation along the x and y axes. This loss of magnetic coherence perpendicular to the B_0 magnetic field is called spin-spin relaxation or T_2 relaxation, and explains the decay of the free induction signal that we observe, represented in Figure 4.1.2C (Keeler 2011). The bulk magnetisation vector eventually returns to the z axis as energy is transferred through the lattice of spins in the sample, and the Boltzmann energy distribution is reassumed. During this period, the bias of spins assuming the low energy state returns and the z axis component grows. This is known as spin-lattice or T_1 relaxation (Keeler 2011). The FID signal is converted in to the classical NMR spectrum by the application of a mathematical trick called the Fourier transform, which converts the time-dependent signal in to the frequencies from which it is composed.

In an NMR spectrum we observe nuclei that are resonating at a range of frequencies, not just one. The reason for this is the dependency of the Larmor frequency on the magnetic field as described in the equation above. However, the Larmor frequency is not solely dependent on the B_0 field, if you remove that dependency you get the following equation:

$$\omega = -\gamma B$$

This means that the frequency of the precession relies on the general field experienced by the nucleus, which not only depends on the B_0 field but also the local magnetic environment. The local magnetic environment is influenced by electrons which circulate in such a way that they generate a magnetic field that either opposes the B_0 magnetic field, in a process called shielding, or aligns with the B_0 field in a process called deshielding (Levitt 2001). This means that the magnetic moment of a nucleus will precess at a frequency dependent on its local chemical environment, such as electronegativity, and will produce a signal that is somewhat distinctive from other nuclei of the same molecule. This is called the chemical shift. Taking this into consideration, it should be noted that a nucleus in the same position in two different molecules in a sample will experience a different chemical environment. This effect is neutralised by tumbling of the molecule through the solution, which averages out the effects of the macroscopic chemical environment. For this reason, solution viscosity is a consideration for an NMR sample. If molecular tumbling is too slow, the observed frequency of precession for a nucleus will be spread over a range of frequencies, coherence is low, T_2 relaxation times are short and resolution will be low (Levitt 2001). Another factor affecting the rate of tumbling is the size of the molecule being studied, as molecular size increases, tumbling times increase and resolution decreases. For this reason, as well as spectral crowding considerations, molecular size needs to be considered when attempting to analyse large macromolecular systems such as proteins.

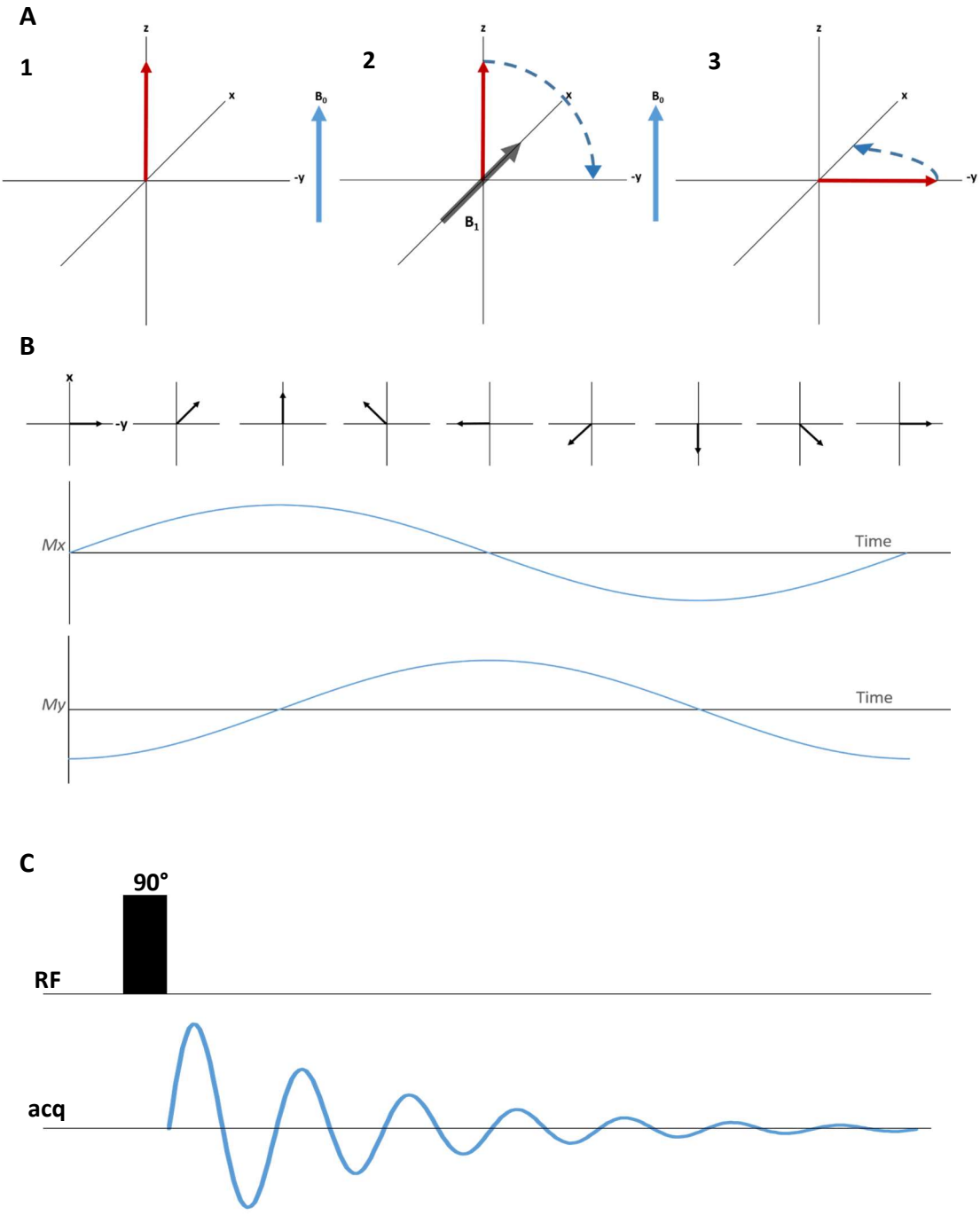


Figure 4.1.2 - A: Position of the bulk magnetisation vector before the radio frequency (RF) pulse (1), transition of the bulk magnetisation vector to the -y axis after the 90° B_1 RF pulse along the x axis (2), rotation of the bulk magnetisation vector in the xy plane (3). B: Points along the rotating path of the bulk magnetisation vector looking down the z axis, and the subsequent magnetisation in the x and y axes as a function of time, showing sine and cosine waves respectively. C: Pulse sequence for the pulse-acquire experiment, showing a period of equilibration before the application of the 90° B_1 field and then the acquired free-induction decay (FID) signal. Adapted from (Keeler 2011).

4.1.2 The ppm scale

Another point to note with the Larmor frequency equation is that, because the frequency of precession is dependent on the magnetic field strength, the resonance will be different between spectrometers with different field strengths. Resonance frequency increases linearly with field strength (Keeler 2011). This means that a resonance at 500 Hz (relative to the resonance of a reference compound) in a 500 MHz spectrometer will resonate at 800 Hz in an 800 MHz spectrometer, and this causes a problem when trying to compare spectra. To resolve this issue, the parts per million (ppm) scale is used. The ppm scale (δ) expresses the chemical shift frequency (ν) in relation to the field strength in which it was acquired and is calculated using the following equation:

$$\delta\ (ppm) = \frac{\nu\ (\mathrm{Hz})}{B_0\ (\mathrm{MHz})}$$

Using the ppm scale, both the 500 Hz resonance and the 800 Hz resonance in 500 MHz and 800 MHz magnetic fields respectively, will appear at 1 ppm. The higher field improves spectral resolution by increasing the spread of frequencies. If two resonances in the 500 MHz spectrometer are 50 Hz apart, in the 800 MHz spectrometer they will be 80 Hz apart. This means that 0.1 ppm corresponds to 50 Hz in a spectrum acquired using a 500 MHz spectrometer, but corresponds to 80 Hz in a spectrum acquired using an 800 MHz spectrometer. Another advantage of a spectrometer with a higher field is an increase in sensitivity. Since the Larmor frequency is directly proportional to the energy difference between the α and β energy states (Figure 4.1.1A), an increase in the B_0 field increases the difference between these energy states. This results in a greater preference of the magnetic moments to align with the field, and produces a stronger bulk magnetisation vector.

4.1.3 The HSQC spectrum

In terms of NMR spectroscopy, proteins are large. They have many nuclei to observe which means that single dimensional experiments are not sufficient to resolve and identify the source of individual resonances. The problem of spectrum crowding means that, at the very least, analysis of proteins requires spectra to be spread into a second dimension. The staple two dimensional experiment in protein NMR spectroscopy is the HSQC (Heteronuclear

single quantum coherence). The HSQC shows the frequency correlation of directly bound heteronuclei, such as ^{1}H and ^{15}N, or ^{1}H ^{13}C, by exploiting a phenomenon called J-coupling, whereby magnetic dipoles interact through the bonds that separate them (Keeler 2011). In the more common ^{1}H-^{15}N correlation HSQC, one dimension corresponds to hydrogen frequencies and the other corresponds to nitrogen frequencies. This means that if two hydrogens resonate at the same or similar frequencies that cannot be resolved in a single dimension, the resonance has a greater chance of being resolved in a second dimension when correlated with its nitrogen resonance. This is useful for protein NMR as it enables us to observe the amide moiety of each amino acid and each peak should correspond to one amino acid. Although, using this experiment we also observe the side chain amide group resonances of asparagine, glutamine and tryptophan. The arginine NεHε correlation is also visible, but typically outside of the spectral range, and at low pH arginine NηHη and lysine NζHζ correlations are also visible (Cavanagh et al. 2010).

As a preliminary experiment, the HSQC spectrum can provide several useful pieces of information about the protein sample. The pattern of peaks can act as a fingerprint as each protein typically has a distinctive dispersion of resonances. The signal to noise ratio can determine whether the protein needs to be of a higher concentration. The dispersion of peaks can be analysed to determine the feasibility of structural experiments, and some information can be gleaned regarding the ease with which the spectrum can be assigned. Amino acids that have no secondary structure have distinctive random coil chemical shifts in the HSQC, it is therefore possible to approximate what proportion of the protein is unfolded (Wishart et al. 1995). For these reasons the HSQC is one of the first experiments performed, and is used to inform the next steps taken with the sample. The specifics of the experiments used to acquire NMR spectra is beyond the scope of this thesis, but a brief outline will be given below.

The pulse sequence used for the HSQC is more complicated than the pulse acquire described above and involves pulses of radio frequencies within the hydrogen frequency and nitrogen frequency of Larmor precession (Keeler 2011). These two are different because of the different gyromagnetic ratios between the two atoms, with nitrogen one tenth of the hydrogen precession. Because of this, the signal from nitrogen is also weaker, since the bulk magnetisation vector is smaller. To compensate for some of this signal loss, a trick called magnetisation transfer is used, where magnetisation is transferred from the ^{1}H nucleus to the ^{15}N nucleus, where it is allowed to evolve under the chemical shift of the ^{15}N nucleus and then magnetisation is transferred back to the ^{1}H nucleus for detection (Levitt 2001; Keeler

2011). Figure 4.1.3 shows the basic principles behind a two-dimensional pulse program. The initial preparation step involves a series of pulses that transfer magnetisation from the ^{1}H spin to the ^{15}N spin. This is typically an insensitive nuclei enhanced by polarisation transfer (INEPT) pulse sequence (Keeler 2011; Cavanagh et al. 2010). The evolution time is variable, and is the step during which the magnetisation is 'labelled' with the nitrogen resonance. The mixing period involves the transfer of magnetisation back to the ^{1}H nucleus by a reverse INEPT pulse sequence and then the FID is collected (Keeler 2011; Cavanagh et al. 2010). This pulse sequence is repeated multiple times for one experiment, with increasing periods of time for the evolution period. By doing this, a series of data points are collected that reveal points along the changes in phase of the nitrogen resonance. As such, the data points can be assembled in to a matrix with points in time along the FID in one dimension (t_2, also known as F1), and points in time along the evolution period in another (t_1, also known as F2) (Keeler 2011). Both of these dimensions are Fourier transformed to give frequencies and these are shown as contours in two-dimensional spectra.

Figure 4.1.3 - The principles behind a two-dimensional NMR pulse sequence. The pulse sequence includes a preparation period, an evolution period which is varied incrementally, a mixing period and a detection period during which the FID signal is collected. Adapted from (Keeler 2011).

4.1.4 Three dimensional NMR and resonance assignment

Three dimensional NMR experiments spread peaks over a third dimension and are used to correlate three distinct resonances. For example, carbon side chain resonances can be correlated to the nitrogen and hydrogen resonances of the amide. The principles behind the three dimensional experiment are the same as two-dimensional NMR. Instead of one evolution time in the indirect dimension, a second variable evolution time, as well as a further mixing step, is incorporated (Figure 4.1.4) (Keeler 2011). Now, for every variable of t_1 a full complement of t_2 variables are acquired, as such this data can be assembled in to a three

dimensional matrix and each dimension is Fourier transformed. Practically, this gives multiple planes of sequential two-dimensional NMR spectra.

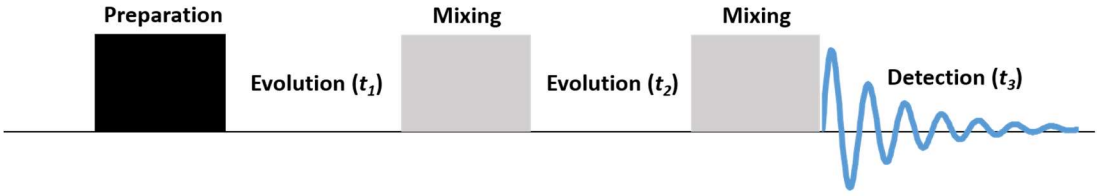


Figure 4.1.4 - The principles behind a three-dimensional NMR pulse sequence. The pulse sequence includes a preparation period, a first evolution period (t_1) which is varied incrementally, a first mixing period, a second evolution period (t_2) which is also varied incrementally, and a detection period during which the FID signal is collected. Adapted from (Keeler 2011).

Once an HSQC has been acquired, it is useful to assign the largely nondescript peaks to a residue in the sequence of the protein being studied. This is usually performed using a few key three-dimensional experiments that correlate the resonances of the α and β carbons of the residue with their amide resonances. Essentially, a HSQC with hydrogen along the 'x axis' and nitrogen along the 'y axis' is laid flat on its face and the carbon dimension projects upwards. Each peak in the HSQC has two peaks that relate to the α and β carbon resonances spaced directly above it in the carbon dimension. When analysing the spectrum, we see planes of nitrogen projecting in to the carbon dimension.

The HNCACB is a three dimensional spectrum that shows the α and β carbon resonances correlated to their amide resonances (the *i* residue), but importantly it also shows the α and β carbon resonances of the previous residue, also called the *i-1* residue. The α and β carbon resonances of the *i-1* residue show as less intense peaks when compared to *i* residue peaks. By aligning *i-1* peaks in one plane, with *i* peaks in another, the amide peaks in the HSQC are put in to a sequence, as demonstrated in Figure 4.1.5. In order to assign residue type, the chemical shifts of the α and β carbon resonances are compared to the BMRB resonance database that shows the most common range of shifts for a given residue (Ulrich et al. 2008). Many residues have distinctive shifts, for example the *i-2* Cβ peak (large 'peak' in red) in Figure 4.1.5 is lower than its Cα peak (in blue), which is characteristic of serine and threonine.

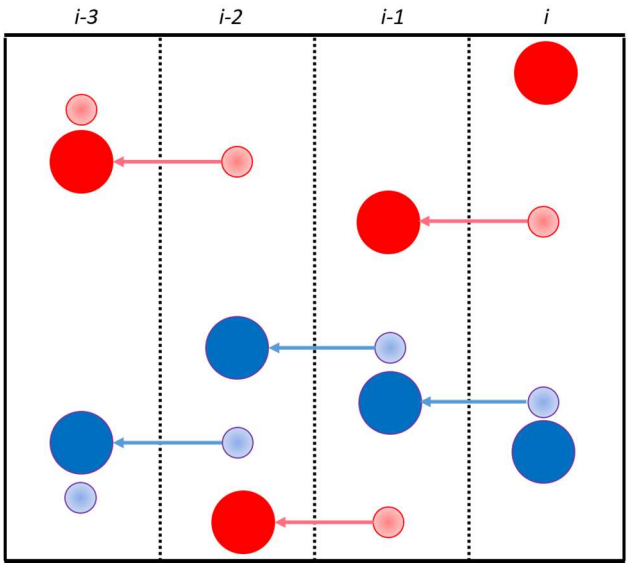


Figure 4.1.5 - A graphical representation of the backbone sequential assignment using the HNCACB spectrum. Each strip represents a different plane, the larger more intense peaks represent the *i* residue carbons, the smaller less intense peaks represent the *i-1* residues. Blue and red peaks represent positive and negative phase peaks, for the Cα and Cβ resonances, respectively. Adapted from (Roberts 2013).

A series of three dimensional and two dimensional experiments are used to assign as much of the backbone and side chains as possible. A set of NOESY experiments are then used to calculate through space restraints. The NOESY experiments produce spectra that correlate nearby (within 5 Å) proton nuclei to either the amide proton resonances (^{15}N-NOESY) or the CH_x group protons of the sidechain (^{13}C-NOESY). The nearby protons produce NOE peaks in the NOESY spectrum that are assignable by comparison to the resonance assignments determined from the experiments and techniques described above. The intensity of these peaks is proportional to the distance of the observed proton from the CH or amide group protons. From the NOESY experiments, it is therefore possible to obtain a series of distance restraints between atoms of a protein, which make the structural calculation possible. These distance restraints are used in conjunction with dihedral restraints which describe the rotational freedom of the amide and carboxyl groups around the amino acid α-carbon. The dihedral restraints used here are generated using the Talos+ web server (Shen et al. 2009). Talos+ backbone torsion angles are generated using the amino acid sequence and backbone chemical shifts as inputs. Chemical shift is closely related to secondary structure, and Talos+ matches the secondary chemical shift (the difference of the chemical shift of the residues in the given sequence and the chemical shift in the random coil

conformation) to a database of high resolution structures (Shen et al. 2009). Database matches are used to predict backbone torsion angles.

4.1.5 Experimental aims

By applying the use of NMR spectroscopy, the principles of which are described above, the aim of this chapter is to acquire the spectra required for resonance assignment of GB1:WW4, the preparation of which is described in Chapter 3. Following the sequential backbone assignment methodology described above, the aim is to then assign residue type to the amide peaks and the carbon resonances. After acquiring the spectra necessary for side chain assignment, these resonances will be assigned as thoroughly as possible. The assignments will then be used in structural calculation using the NOESY spectra required for the calculation of the GB1:WW4 solution structure.

4.2 Results

4.2.1 ^{1}H-^{15}N-HSQC

Purification of GB1:WW4 from 2 litres of ^{15}N ^{13}C labelled MEM gave a yield sufficient to make a 1.2 mM NMR sample. The ^{1}H-^{15}N-HSQC spectrum shown in Figure 4.2.1 was acquired using a Bruker Avance III spectrometer at 800 MHz, with a triple resonance probe. The HSQC showed peaks of good intensity, as would be expected with a sample at 1.2 mM, and good dispersion and resolution, indicating a folded protein with a spectrum that is not overly crowded. Certain features could be discerned, such as the amide group side chain peaks that appear as pairs of teardrop-shaped peaks on the right hand side (due to the two hydrogens bound to one nitrogen, producing one nitrogen resonance with two hydrogen resonances). The tryptophan side chain amine group peaks could be seen at their distinctive position at the very bottom left - a position occupied due to deshielding by the aromatic ring electrons, which generate a magnetic field in the same direction as the B_0 field.

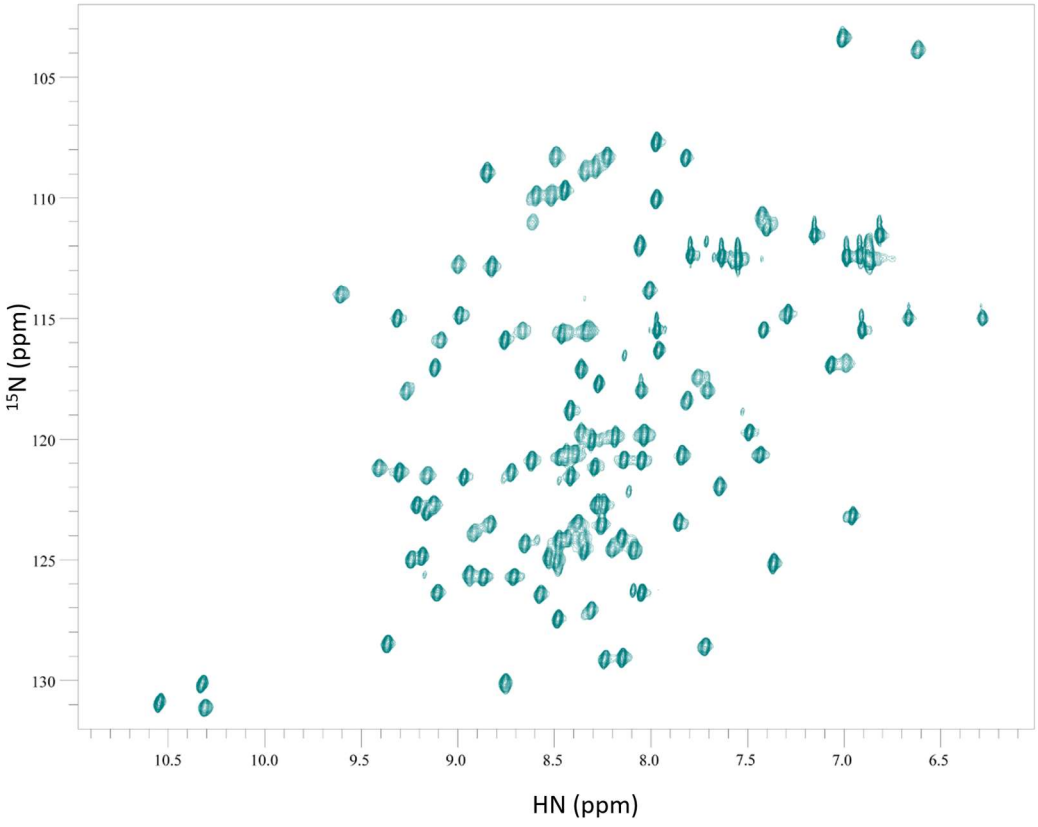


Figure 4.2.1 - 1H-15N-HSQC of GB1:WW4 at a concentration of 1.2 mM, acquired at 800 MHz, 298 K. The sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8

4.2.3 Resonance assignment

Using the backbone assignment method described above, the NH, Cα and Cβ resonances were sequentially assigned using the HNCACB and CBCA(CO)NH experiments. Figure 4.2.2 shows the assigned ^{1}H-^{15}N-HSQC spectrum. All of the peaks were assigned, except for two folded peaks in the middle of the spectrum, and the two tryptophan side chain peaks at the bottom left. This equated to a 92% assignment of the backbone NH groups. The unassigned residues included 5 prolines, three residues at the N-terminus and a serine at position 456 (in the amino acid sequence of WWP2). Prolines are not seen in the HSQC due to the absence of an NH group when they are part of a polypeptide chain. The signal from the N-terminal residues is typically weak in NMR experiments, because of, amongst other things, high levels of proton exchange due to protonation of the N-terminal NH_2 group - and so the absence of peaks for these residues was somewhat expected. The serine showed no resonance in the HSQC and it was thought that the peak might be hidden behind another. However, detailed examination of the HNCACB and CBCA(CO)NH experiments showed no unassigned resonances appropriate for this position, and it may be that this particular residue is in conformational exchange, which might diminish the intensity of its signal.

Having assigned the amide resonances, the CC(CO)NH and H(CCO)NH spectra were used in conjunction with the ^{1}H-^{13}C-HSQC to assign the side chain hydrogen and carbon resonances of the aliphatic side chains. The Aromatic ^{13}C-TOCSY and aromatic ^{13}C-TROSY were used to assign as much of the aromatic side chain nuclei resonances as possible.

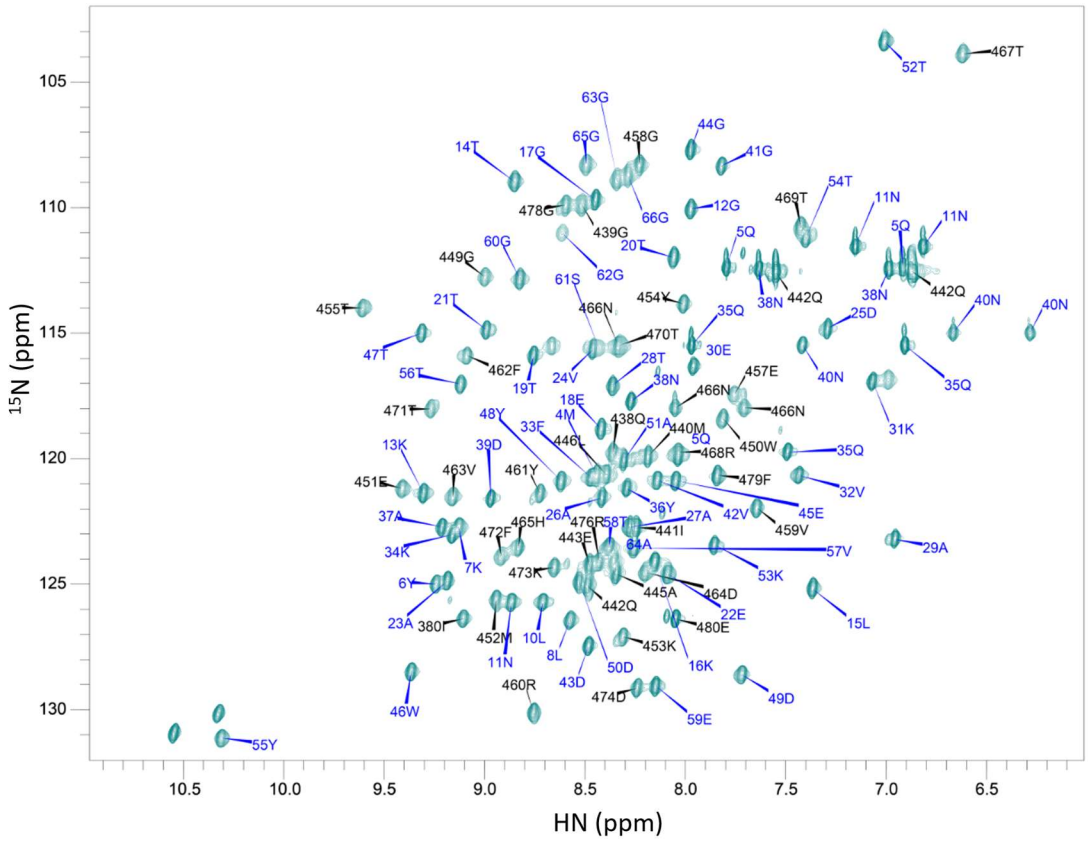


Figure 4.2.2 - The 1H-15N-HSQC of GB1:WW4 at a concentration of 1.2 mM acquired at 800 MHz, 298 K, as shown above but with labels. Sequentially assigned using the HNCACB and CBCA(CO)NH experiments. GB1 (1-66) peaks are labelled in blue, WW4 (438-480, corresponding to the position in the WWP2 amino acid sequence) peaks are labelled in black.

4.2.4 UNIO automated NOESY peak picking

To calculate structures from NMR data, experiments are required that give information on the distances between nuclei. The three dimensional nuclear Overhauser effect (NOESY) experiments give the spatial restraints required. The ^{15}N-NOESY-HSQC is used to correlate all proton resonances within an approximate distance of 5 Å, with NH group resonances that appear in the ^{1}H-^{15}N-HSQC. And so, correlated with each backbone NH group are a series of 'cross-peaks' that relate to nearby hydrogen resonances from side chains and other backbone amides. The ^{13}C-NOESY-HSQC is also used to correlate hydrogen resonances within an approximate distance of 5 Å, but instead with CH group resonances that appear in the ^{1}H-^{13}C-HSQC. Practically, this means that correlated with each side chain CH group resonance are a series of peaks that relate to nearby hydrogen resonances from side chain

methyl groups. The chemical shifts of the correlated peaks in the NOESY experiments will be identical to the chemical shifts from the experiments used to assign the protein. It is here that the purpose of the assignment becomes apparent, as this information can be used to identify the cross-peaks (and therefore the parts of the protein that are within a close proximity). The intensity of each peak has a direct relationship with the distance from the NH or CH group to which it is correlated, and from this information a series of distance restraints can be calculated.

The NOESY spectra hold a large amount of data and assigning the peaks can often present a problem. Manual assignment is not uncommon but is very time consuming because of the sheer number of cross-peaks. The spectra can often be crowded and because of the similarities between some resonances, objective assignment can often be challenging. In order to overcome these challenges, the automated peak picking and assignment program UNIO (Volk et al. 2008; Fiorito et al. 2008) was used to pick and assign the cross-peaks from the GB1:WW4 NOESY spectra. UNIO adds another criterion to the assignment of NOESY peaks and the distance restraints they produce, by using iterative preliminary structures to inform their refinement. Because this process is automated, it is best to have as much of the resonances assigned as possible so as to avoid misassignment of peaks that belong to unassigned spins. A value of 90% or over is ideal for the most accurate assignment, but not necessary. After thorough assessment of the spectra, approximately 81% of the sequence was assigned. Most of the unassigned regions were the result of prolines, the ends of long aliphatic chains where recovered signal is weak, and some unassigned aromatic side chain atoms because of some irreconcilable ambiguities in the spectra.

A list of the manual assignments and the referenced NOESY spectra were used as input files for the UNIO program. These were complemented with a set of dihedral angle restraints from the TALOS plus web server, which returned 156 Φ and Ψ backbone restraints that were in good agreement with the TALOS plus angle database (Shen et al. 2009). The algorithms picked and assigned 1224 cross-peaks in the ^{15}N-NOESY and 2244 cross-peaks in the ^{13}C-NOESY and this gave 1452 restraints.

4.2.5 Initial GB1:WW4 ensemble

After seven cycles of peak picking and refinement, UNIO produced 20 preliminary models that represent the lowest energy structures which fulfil the restraints as closely as possible. The GB1 and WW4 domain are well represented in terms of inter-residue restraints, while the linking residues have very few restraints (Figure 4.2.3). The GB1 domain of this recombinant protein holds a similar conformation to other GB1 domain structures found in the Protein Data Bank (PDB), a four stranded β-sheet running parallel to an α-helix (more precisely, two parallel double-stranded antiparallel β-sheets that have an alpha helix in the middle) (Figure 4.2.4). Some of the models lack the third β-strand and instead have an organised loop region. The WW4 domain holds a three-stranded antiparallel β-sheet as expected (Figure 4.2.5), and it has a slight twist, producing convex and concave surfaces which is typical of other WW domain structures. Some of the ensemble models have helical turns between the second and third β-strands. The two domains have no NOE distance restraints between them, indicating that the GB1 domain does not interact with or distort the WW domain, and that the model can be used as a reasonable representation of the native WW4 domain structure.

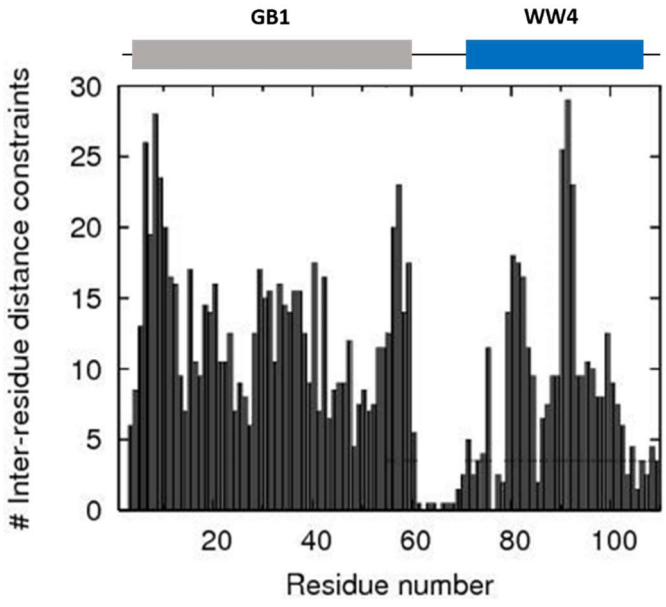


Figure 4.2.3 - The number of inter-residue distance restraints per residue for the GB1:WW4 UNIO calculation, as determined using the Protein Structure Validation Suite (PSVS) (Bhattacharya et al. 2007). Residue numbering is relative to the recombinant protein sequence. The GB1 and WW4 domain boundaries have been aligned in the diagram above.

The root mean-square deviation (RMSD) average of the ensemble for the fusion protein as a whole is 8.360 Å as determined using PyMol (Delano 2002). However, since the two domains are separated by a long stretch of residues that have very few NOE restraints, this linker region is flexible and the domains are orientated at different angles in the different models, as seen in Figure 4.2.4A and Figure 4.2.5A. The RMSD average of the molecule is therefore not an accurate representation of the resolution of the model. When the individual domains are aligned instead, the RMSD average of the GB1 domain is 0.568 Å while the RMSD average of the WW4 domain is 0.586 Å. The selected WW4 region incorporates the rigid structure stretching from the tryptophan at position 450 to the phenylalanine at position 472. These represent the residue positions of the typical tryptophan-tryptophan motif (although the WW4 domain is atypical as it has a phenylalanine instead of the C-terminal tryptophan). One of the concerning features of the WW4 domain structures is the twist seen in the turn between the second and third β-strands. When shown by PyMol as a helical section, it looks like an abnormal twist in the protein backbone. However, when the turn is aligned with the backbone of the recently resolved crystal structure of the first WW domain of YAP2 (Figure 4.2.5D), they align well (Martinez-Rodriguez et al. 2015). This indicates that, although there are some issues, the initial structure is a reasonable representation of a WW domain fold, and that progressing to the next stage of structure refinement is not inappropriate.

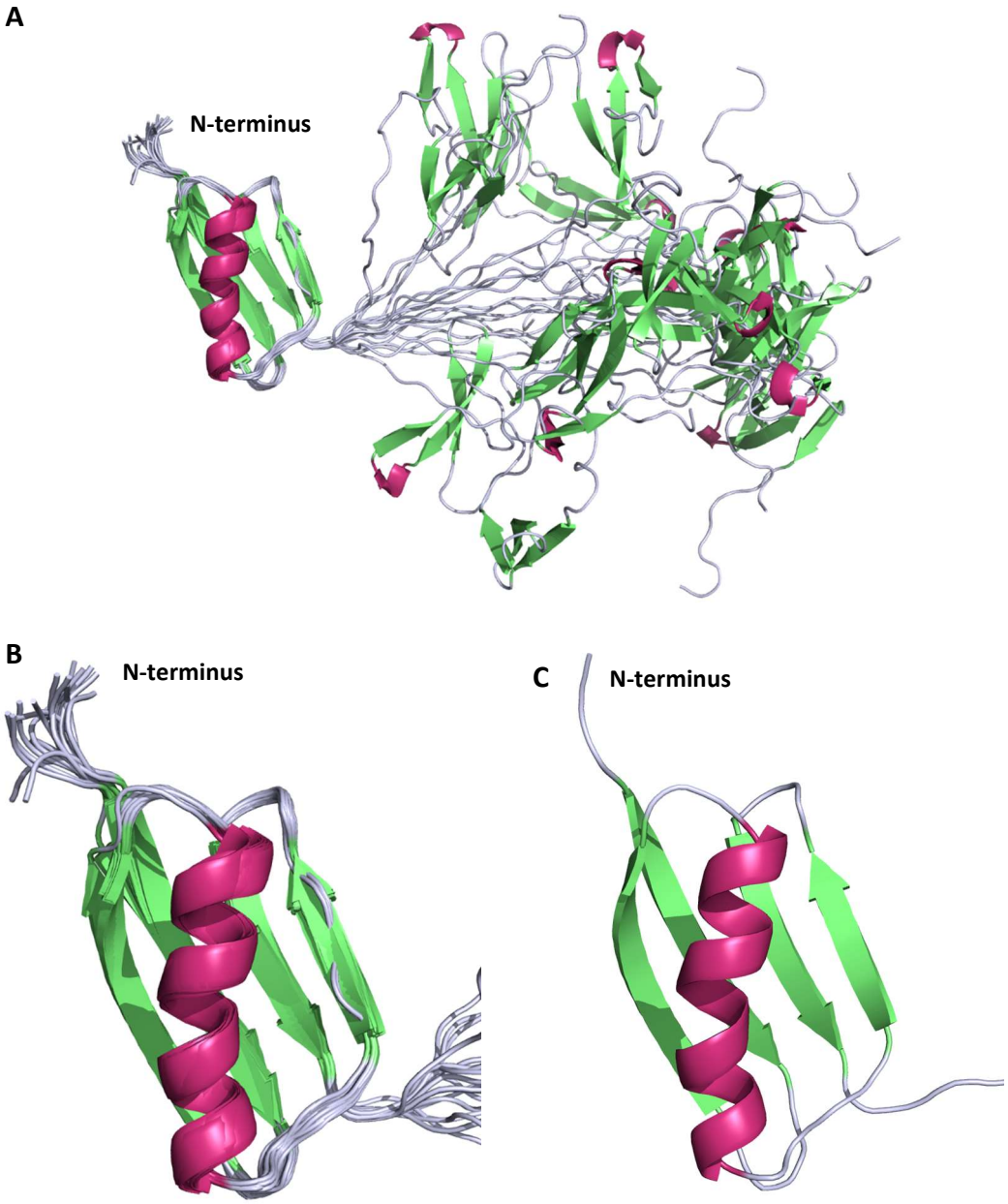


Figure 4.2.4 - A: The unrefined, 20 model UNIO GB1:WW4 ensemble structure with the GB1 domain aligned. The WW4 domain, connected by a stretch of flexible residues, can be seen in several different orientations in relation to the GB1 domain. B: The GB1 ensemble in closer detail and C: One model from the unrefined ensemble.

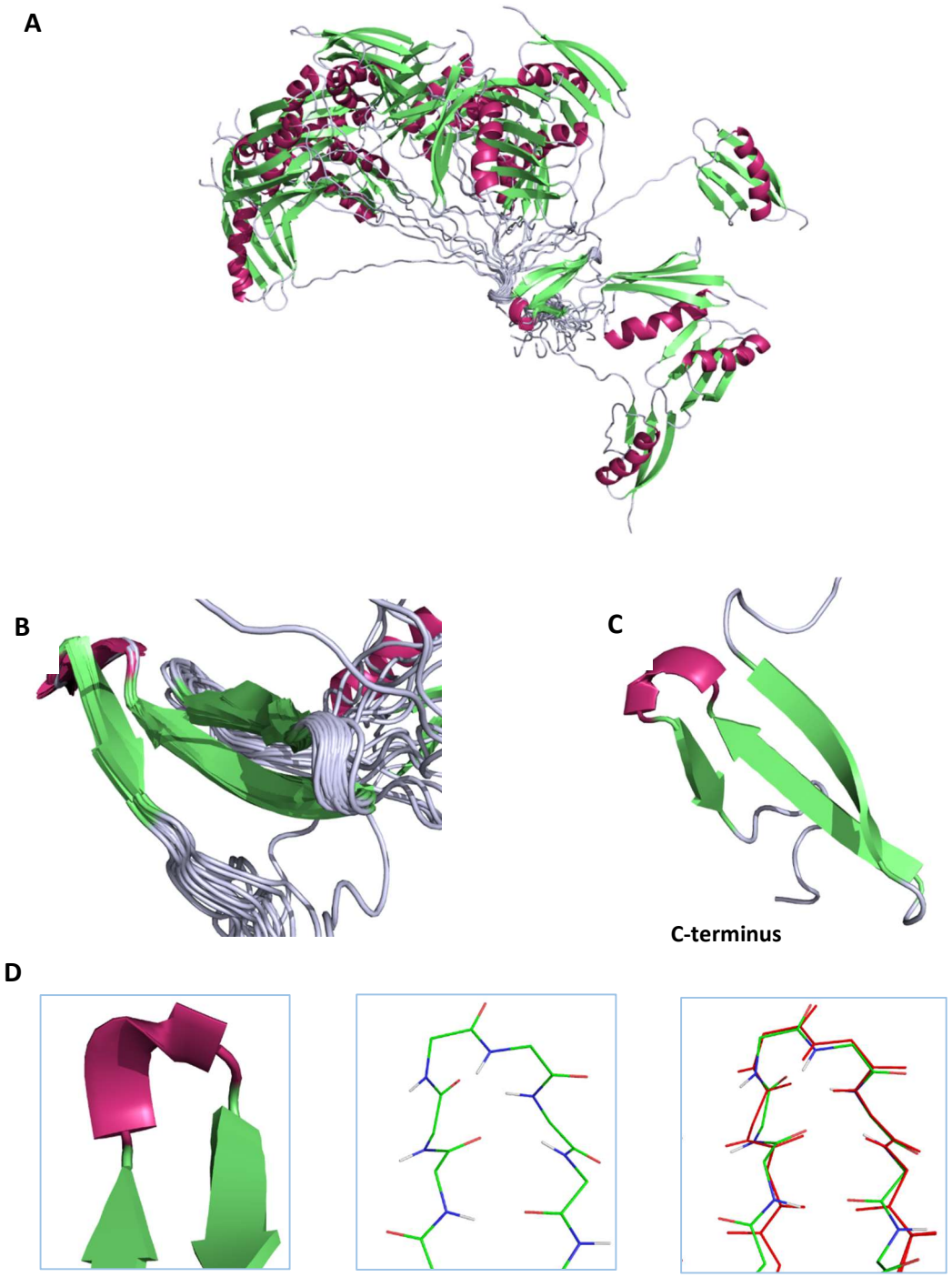


Figure 4.2.5 - A: The unrefined, 20 model UNIO GB1:WW4 ensemble structure with the WW4 domain aligned. The WW4 domain is at the centre of the ensemble, and the GB1 domain can be seen in several different orientations. B: The WW4 ensemble in closer detail and C: One model from the unrefined ensemble. D: The cartoon and stick representations of the loop between β-strands 1 and 2 of WW4, and the alignment with the crystal structure of the first WW domain of YAP2 in red (PDB code: 4REX).

Several criteria are used to determine the validity of the UNIO calculation (Herrmann et al. 2002a).

- The average Target Function of the cycle 1 ensemble should be less than 250 $Å^2$, and the average Target Function of the cycle 7 ensemble should be less than 10 $Å^2$.
- The average RMSD of cycle 1 should be less than 3 Å, and the RMSD between the mean structures from the first and seventh cycle should also be less than 3 Å.
- Over 80% of the long range NOE restraints in cycle 1 should be retained in the cycle 7 ensemble.
- Over 80% of the NOE cross peaks should be assigned in the cycle 7 ensemble.

Target Function is a measure of the quality of the structure and takes in to consideration violations of the distance and dihedral restraints, and any energy violations arising from these restraints. Small Target Function values are more favourable than larger values. The values for each cycle were as follows:

Cycle No.	Target Function ($Å^2$)	RMSD (Å)	RMSD Drift (Å)	Restraints
1	260.24 ±20.35	8.81	22.83	1593
2	130.42 ±6.08	8.21	17.82	1756
3	48.67 ±1.97	9.80	6.25	1690
4	17.78 ±1.06	8.54	3.78	1633
5	9.2 ±0.9	10.32	4.18	1594
6	5.07 ±0.3	9.83	2.54	1519
7	4.53 ±0.26	9.21	0	1452

Table 4.2.1 Target Function, RMSD values, RMSD drift values and number of restraints for the UNIO calculation of GB1:WW4

The Target Function of cycle 1 was within range of the 250 $Å^2$ limit, when taking the error in to account, while the cycle 7 Target Function was within the 10 $Å^2$ boundary. As discussed above, the presence of relatively long stretches of flexible linker means that the structure is not amenable to broad RMSD analysis, and therefore as expected, the RMSD values were out of range. However, when comparing the RMSD between the GB1 helical stretch of amino acids of cycle 1 and cycle 7 mean structures, a value of 2.507 Å was obtained, and the RMSD values of the cycle 7 ensemble GB1 and WW4 domains were well under the 3 Å boundary, as described above. The types of NOE restraint and the numbers in each cycle can be seen in Table 4.2.2. Long-range restraints (5Å and over) actually increased from 375

to 512 between cycles 1 and 7, as the program better defined the two individual domain folds. The number of assigned peaks in cycle 7 (2776) was also higher than the number of picked peaks in cycle 1 (2306), and therefore the fourth criteria is also satisfied. From this it can be concluded that the restraints generated by the UNIO ATNOS/CANDID cycles have been optimised and are a reasonable representation of the data.

	NOE restraints			
Cycle no.	**Intra-residue**	**Sequential**	**Medium-range (≤4 Å)**	**Long-range (≥5 Å)**
1	877	503	231	375
2	1251	824	374	391
3	1242	815	344	480
4	1273	812	330	471
5	1260	801	323	482
6	1244	760	304	468
7	1226	734	303	512

Table 4.2.2 Breakdown of the types of NOE restraints generated from the 7 cycles of simulated annealing

4.2.6 Refined GB1:WW4 ensemble

The UNIO calculation described above employs a simplified modelling system to identify the protein fold and to create a list of NOE restraints from the raw data in a relatively short period of time. This is achieved by employing a computationally-light calculation, whereby only the degrees of freedom about the proteins dihedral angles are taken in to account. This method of conformation sampling is called torsion angle molecular dynamics. While this method is highly efficient, it also does not specifically take into account atomic interactions or the implications of solvent on the surface of the protein, which is particularly relevant, since the data is collected from a protein in solution. For these reasons, a further step is taken in refining the structure using a computationally-demanding method called Cartesian dynamics. Cartesian dynamics modelling is used to model the individual atoms in Cartesian space and takes in to account more molecular dynamics parameters during simulated annealing, such as bond length oscillations, which are frozen in torsion angle sampling (Lian & Roberts 2011).

The refinement calculation was performed using the CNS software package (Brünger et al. 1998), the dihedral restraints and the NOE restraints list from the sixth cycle of the initial UNIO calculation. The restraints list from the sixth cycle was used to allow freedom in the calculation by maintaining a level of ambiguity in the atom assignment, while maintaining the restraint distance. However, a further level of freedom is given by allowing a 0.2 Å relaxation of the restraint distance. The RECOORD scripts, along with their parameter set (Nederveen et al. 2005) were used to run the CNS calculation. The scripts were used to calculate 200 models, of which 26 models had no violations of the angular and distance restraints which were used as inputs for the calculation. When the models were organised by their CNS-calculated total energy, which is a measure of how energetically favourable they are, 20 of the models with no violations were among 29 of the lowest energy structures. These 20 models were used in the final ensemble. The ensemble structures were aligned to the GB1 domain, and can be seen in Figure 4.2.6.

For this ensemble the average RMSD for the GB1 domain is 1.533 Å, while the average RMSD of WW4 is 1.639. Since this method incorporates more variables in to the calculation, it is expected that the RMSD between the models will increase as factors such as solvent interaction and interatomic force are taken in to account. Going solely by the RMSD values, this is an acceptable ensemble. However, when observing the different models, some variability is evident, in particular with the fifteenth model, shown in red in Figure 4.2.6C. The calculation seems to have found an energetically favourable conformation in which the second (outer) β-strand is folded inwards, across the first β-strand. Despite the GB1 domain not being of particular interest, it can be considered to be an internal control that measures the quality of the calculation, since the GB1 structure has been published previously. Similar to some of the WW4 domain loop structures in the initial UNIO structure, two of the refined GB1 models have loop regions that are depicted as twisted helical-type turns (Figure 4.2.6D), while the loop in the WW4 structure itself is no longer depicted as helical when displaying the PyMol ribbon graphic (Figure 4.2.7).

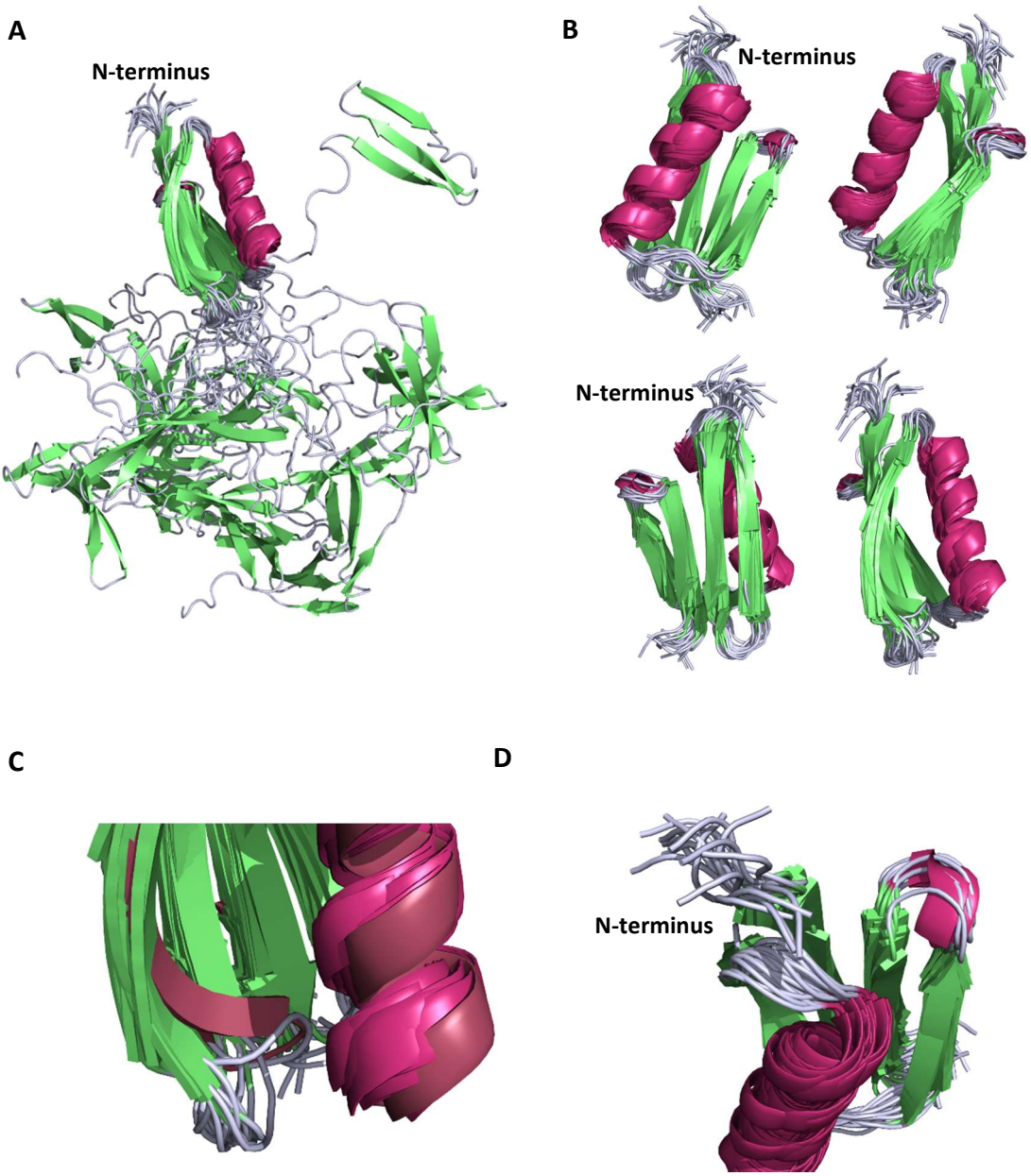


Figure 4.2.6 - The GB1:WW4 ribbon diagram refined CNS models (20 models) with the GB1 domain aligned. A: The GB1:WW4 ensemble aligned to the GB1 domain. The disordered loop region between the two domains ensures that the WW4 domain is oriented at a variety of angles when compared to the GB1 domain. B: The GB1 domain ensemble. C: One of the models (in red) has a distorted β-strand, when compared to the other ensemble models. D: Two of the ensembles have loop regions with helical geometries.

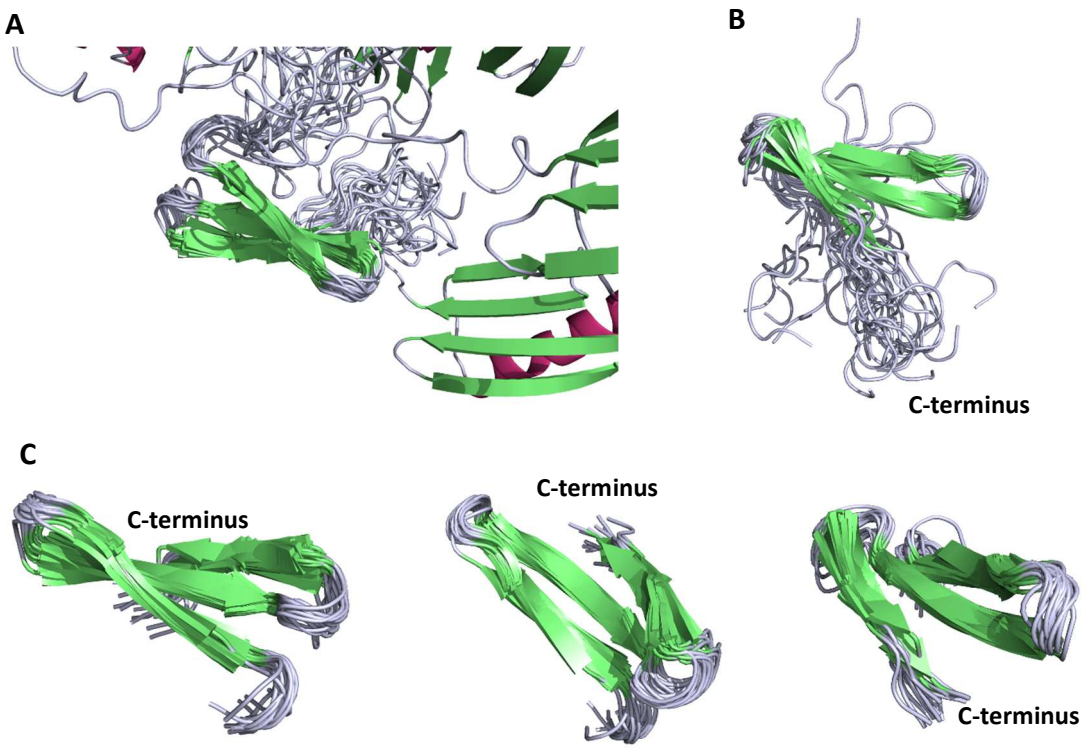


Figure 4.2.7 - The GB1:WW4 ribbon diagram refined CNS models (20 models) with the WW domain aligned. A: The WW4 aligned ensemble. B: The WW4 domain ensemble without the GB1 tag. C: The ordered residues of the WW4 domain ensemble from three angles.

The N-terminal and C-terminal tails are highly disordered, which is to be expected with relatively few NOE restraints, and is likely to be representative of a disordered coil region that assumes many different conformations in solution. Visually there seems to be greater homogeneity in the β-strand region of the WW4 domain ensemble, when compared to the GB1 alignment, with the majority of the variation in the backbone centred around the length and angle of the third β-strand.

4.2.7 Ensemble validation

Evaluation of the solution structure ensemble of GB1:WW4 has thus far mainly focused on the accurate reflection of the resonance data and its assignment. In order to thoroughly evaluate the quality of the ensemble, certain properties must be compared to conventional values so that the structure can be considered a plausible representation of the proteins true conformation. The web server iCING is a structure validation tool optimised to

perform comprehensive analyses of NMR structure ensembles, and provide an integrated evaluation of the quality of the structure (Doreleijers et al. 2012). The scores for the GB1:WW4 ensemble can be seen for a variety of parameters in Table 4.2.3. The server has an overall score ranking system that marks residues as either red (problems), orange (potential problems) or green (no problems found) depending on how favourably they perform, and a score of below 20% green and over 50% red is considered to be indicative of a poorly modelled structure (Doreleijers et al. 2012). With this in mind, the GB1:WW4 ensemble performed reasonably well, with 15% of the residues with a red score, and 68% of the residues with a green score. When considering the scores of each domain separately, 72% of the GB1 domain, and 75% of WW4 domain residues were given a green ranking (Table 4.2.3). Unsurprisingly, four out of five of the prolines ranked as problematic, likely due to the difficulty in proline resonance assignment and therefore the absence of NOE restraints. Seven of the residues in the flexible linking region between the two domains also ranked as problematic, which is most likely also related to the lack of NOE distance restraints for this region.

The server also incorporates scores from the WHAT IF evaluation tool, which compares a variety of geometry criteria to a database of high resolution crystal structures that are used to define 'ideal values' (Vriend 1990). WHATIF scores these criteria using Z-scores, which is the number of standard deviations away from the mean value, *i.e.*:

$$Z = \frac{x - mean}{SD}$$

Outliers are considered to be more or less than 4 standard deviations from the mean. The majority of the scores are a rank of quality, and positive values indicate results that are better than average. RMS-Z-scores are also used, which give an indication of how rare outliers are. RMS-Z-scores should be close to 1, and are used to determine the extent to which the parameters are constrained. If, for example, bond lengths are too variable then this will give an RMS-Z-score of over 1. If bond length distribution is poor then a score below 1 is given, and the parameter is considered to be overly constrained.

The packing quality is a measure of how favourable the conformation suits the amino acid sequence, or how 'comfortable' each residue is in its local environment. The refined ensemble of GB1:WW4 and the individual domains perform better than average in this respect. The ensemble performed worst on Ramachandran plot appearance and χ^1/χ^2 rotamer normality which are a measure of backbone and side chain torsion angles,

respectively - although the values are within the -4 range that would classify them as true outliers. The WW4 domain performed much better than the rest of the structure in Ramachandran plot appearance and backbone conformation, but still suffered when comparing χ^1/χ^2 rotamer normality to the standard data set. The structure performs reasonably well in bond length, side chain planarity, dihedral distribution, and inside/outside distribution (a measure of how well the residues are located, hydrophilic 'outside', hydrophobic 'inside'). However, the distribution of bond angles and omega angles (peptide bond torsion) is more tightly constrained when compared to the standard data set.

	GB1:WW4 1-109	**GB1 residues 1-61**	**WW4 residues 77-104 (448-475)**
CING ROG Score			
Red	16 (15%)	3 (5%)	3 (11%)
Orange	19 (17%)	14 (23%)	4 (14%)
Green	74 (68%)	44 (72%)	21 (75%)
Average RMSD			
Backbone	8.65 Å	1.26 Å	1.06 Å
Heavy atoms (C^{13}, N^{15})	9.01 Å	1.71 Å	1.79 Å
Model closest to mean	10	9	3
WHAT IF (Z-score)			
(Positive values rank better than average)			
1st generation packing quality	0.867	2.717	2.570
2nd generation packing quality	4.681	4.152	5.854
Ramachandran plot appearance	-2.351	-2.245	0.388
χ^1/χ^2 rotamer normality	-2.927	-3.506	-2.943
Backbone conformation	-0.006	-0.213	1.030
RMS-Z-score *(Ideal value: 1.0)*			
Bond lengths	1.149	1.157	1.133
Bond angles	0.520	0.508	0.528
Omega angle restraints	0.689	0.691	0.718
Side chain planarity	1.201	1.256	1.181
Improper dihedral distribution	0.940	0.952	0.897
Inside/Outside distribution	1.133	1.061	1.146
PROCHECK Ramachandran plot			
Favoured	89.3%	93.2%	92.7%
Allowed	9.6%	6.2%	7.3%
Generously allowed	0.6%	0.3%	0.0%
Disallowed	0.6%	0.4%	0.0%

Table 4.2.3 Outputs from analysis of the 20 model GB1:WW4 CNS ensemble by the online structure validation server iCING. Three regions were submitted, the entire structure, the GB1 domain from residues 1-61, and the rigid portion of the WW4 domain from residues 77-104 (448-475 with regards to the WWP2-FL sequence) (Doreleijers et al. 2012).

Ramachandran analysis involves plotting each residues Φ and Ψ angles against each other, and is a good indicator of the quality of the protein structure. The Ramachandran plot has allowed and disallowed regions that are shown as 'islands' of favoured and allowed regions, where the residue Φ and Ψ angles most commonly sit. Outside of those islands are disallowed regions, where the dihedral angles are considered highly unfavourable. The iCING server integrates a PROCHECK Ramachandran analysis of the ensemble dihedral angles (Laskowski et al. 1993). The results are shown in Table 4.2.3. The majority of the ensemble dihedral angles sit within the favoured and allowed regions, while 0.6% sit within the disallowed regions. This percentage reduces to 0% when only considering the WW4 domain structure. MolProbity is an alternative structure evaluation suite, with slightly stricter areas for allowed and disallowed Ramachandran islands (Chen et al. 2010). When the ensemble was analysed using MolProbity, 91.2% of the dihedrals were within the most favoured regions, 7.1% were within the allowed regions and 1.7% were within disallowed regions (the disallowed angles are shown in Table 4.2.4). The Ramachandran plot can be seen in Figure 4.2.8.

Model no.	Disallowed angles (Φ and Ψ)
1	44 GLY (-163.8, -70.0) 70 ILE (74.9, 16.9) 106 PRO (-46.5, -75.2)
2	70 ILE (74.0, 61.8) 106 PRO (-86.6, -148.9)
3	72 GLU (-110.3, -52.3) 104 PRO (-60.7, -75.2) 106 PRO (-93.5, -31.2)
4	69 MET (179.8, 107.6) 104 PRO (-82.0, -93.8)
5	44 GLY (177.2, -82.2) 73 PRO (-104.5, 31.3)
6	104 PRO (-87.7, -35.9)
7	3 HIS (70.5, -90.2) 73 PRO (-41.4, 93.9)
8	104 PRO (-95.9, 85.6) 106 PRO (-64.2, -65.1) 108 PHE (-167.7, -41.9)
9	63 GLY (177.8, -34.8) 67 GLN (51.0, 84.3)
10	104 PRO (-99.2, 54.2)
11	75 LEU (-130.7, -60.7)
12	71 GLN (178.4, 38.5) 104 PRO (-45.0, -72.6)
13	73 PRO (-20.8, 91.8) 104 PRO (-102.7, 41.4) 108 PHE (-170.9, -31.6)
14	2 SER (63.4, -82.0) 60 GLY (-147.2, -59.7)
15	12 GLY (59.7, -78.8) 63 GLY (143.9, 78.9) 73 PRO (-18.7, 96.1) 75 LEU (70.8, 132.6)
16	44 GLY (176.8, -90.2) 60 GLY (26.4, 88.7) 73 PRO (-115.9, 39.9)
17	64 ALA (-173.0, -64.0) 75 LEU (65.6, 96.9) 104 PRO (-73.5, -55.3)
18	
19	
20	60 GLY (165.2, 42.1) 72 GLU (-105.4, -66.3) 73 PRO (-7.0, 79.7) 104 PRO (-68.0, -63.9)

Table 4.2.4 The disallowed dihedral angles for each model of the 20 model GB1:WW4 CNS ensemble, from MolProbity (Chen et al. 2010).

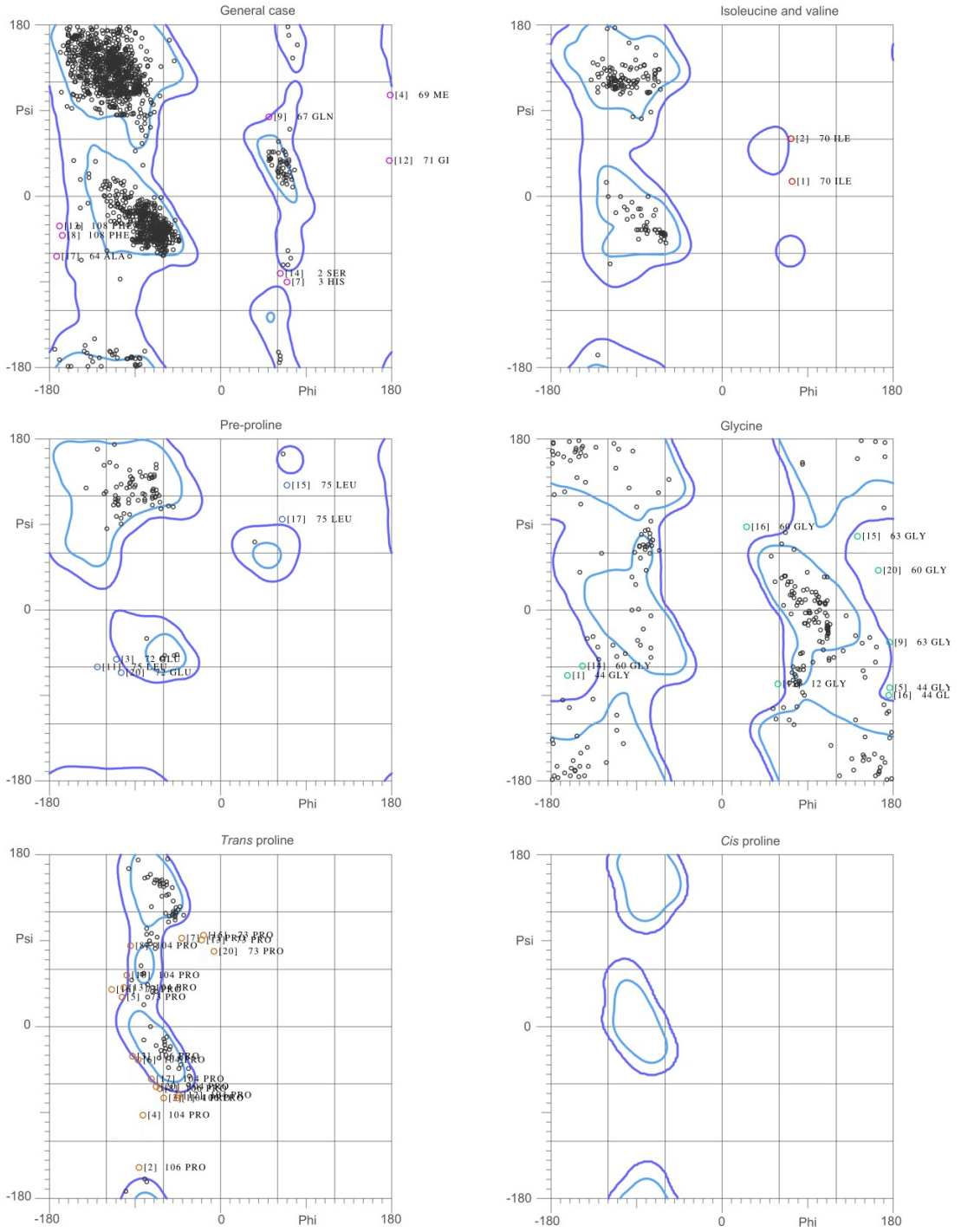


Figure 4.2.8 - The Ramachandran plot of the GB1:WW4 ensemble Φ and Ψ angles, from MolProbity (Chen et al. 2010). Favourable regions are shown in turquoise and allowed regions are shown in blue. Outliers are shown as coloured circles with the model number in square brackets preceding the residue number and the residue type.

4.3 Discussion

Using the NMR data acquired on the isotopically labelled, bacterially expressed and solubility enhanced WWP2 WW4 domain, an ensemble of structures has been calculated that accurately represents the NMR data and ranks reasonably well in a variety of structure validation criteria. There are, however, some quirks to the ensemble calculated. Some of the models have a helical turn between the third and fourth β-strands of the GB1 domain. This is immediately noticeable using the PyMol molecular graphics software, which uses backbone geometry and hydrogen bonding patterns to determine secondary structure visualisation. The amino acids 50-52 in models 2 and 19 are classified as helical in PyMol. A PROCHECK analysis of the ensemble confirmed that some of the models have helical geometries, but disagreed on which models (models 5, 6, 10, 15, 16, 17). PROCHECK uses geometry and hydrogen bonding patterns outlined by Kabsch and Sander to determine secondary structure characteristics (Laskowski et al. 1993; Kabsch & Sander 1983). The CSI 2.0 web server (Hafsa & Wishart 2014) uses backbone chemical shift data to predict protein secondary structures. CSI 2.0 predicts that this region should form a loop as opposed to a helical conformation, the output from this tool can be seen in Figure 4.3.1.

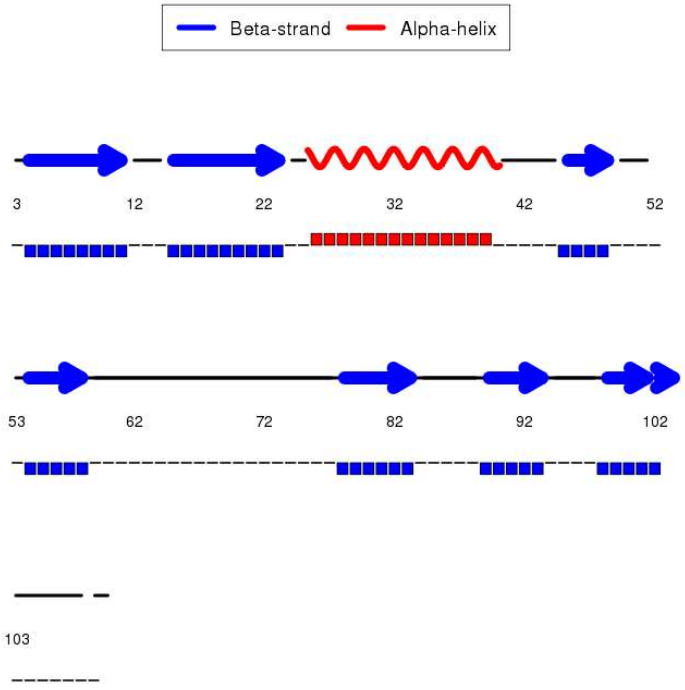


Figure 4.3.1 - The output from the CSI 2.0 web server, using the GB1:WW4 chemical shifts as the input (Hafsa & Wishart 2014). Helical regions are shown in red, β-strands are shown in blue and coil regions are shown in black.

The second β-strand is a source of some variation. The 19th model has a shortened strand consisting of only three residues (19-21), where the output CSI 2.0 indicates this region should be a nine residue β-strand. While the second β-strand of the 15th model has a distorted conformation from residues 13-15, as discussed above, which is likely to be related to the dihedral angle of glycine 12, which occupies a disallowed position in the Ramachandran plot for this model. The GB1 domain model that best described the mean of the GB1 domain ensemble, as determined by iCING, is model 9. The alignment of this model with a GB1 solution structure that has been deposited in the PDB is shown in Figure 4.3.2 (PDB ID: 3GB1) (Kuszewski et al. 1999). The structures align to 1.850 Å, there are variations in the position of some components of the structure, particularly the helix, and the positioning of the side chains.

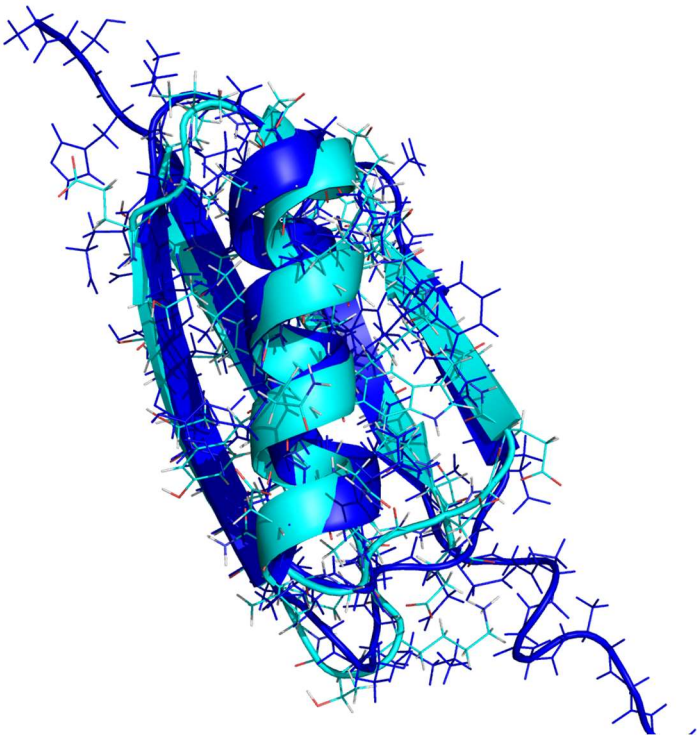


Figure 4.3.2 - Alignment of the GB1 domain of the most representative model from the GB1:WW4 ensemble, model 9 in blue, with a solution structure of GB1 that has been deposited in the PDB, in turquoise (PDB ID: 3GB1) (Kuszewski et al. 1999).

The WW4 domain itself performs particularly well in the validation step, and appears to have very little that is obviously wrong with the backbone geometries. The typical WW domain three stranded β-sheet is present and holds the expected twist described in the introductory chapter. WW domains typically have a hydrophobic cluster that reinforces the domain conformation on one side of the β-sheet, this typically incorporates the N-terminal

tryptophan, a C-terminal proline and the second residue of a hydrophobic pair on the second β-strand. For WW4 this is tryptophan 450, phenylalanine 462 and proline 475. Figure 4.3.3A shows the position of these residues in the WW4 ensemble. The proline has a highly variable position in all of the models, which could be representative of the side chain movement in a solution structure, but is likely to be related to the lack of NOE restraints, since proline assignment is very limited. The tryptophan and phenylalanine positions are much more consistent.

The binding site is present on the opposite side of the β-sheet. In WW4, the XP binding groove consists of phenylalanine 472 (which substitutes for the second tryptophan), and tyrosine 461 which can be seen in red in Figure 4.3.3B. They are stacked almost parallel to each other, creating the groove in which the ligand proline sits. The phenylalanine shows some variation in the side chain positioning, which could again represent variable positioning in solution. However, not all of the aromatic ring is assigned, and as a result this variation could be due to a lack of restraints. The second 'specificity' pocket is shown in orange in Figure 4.3.3B and consists of valine 463 and histidine 465. These residues are important in determining WW domain 'type'. Valine 463 is at the position of the most important residue for determining specificity (Zarrinpar & Lim 2000). A leucine at this position typically binds tyrosine, and indicates that the WW domain will bind group I ligands, the PPxY motif. The similarities between leucine and valine indicate that WW4 is most likely to bind with type I specificity.

While the ensemble calculated here is not without its problems, the structure performs reasonably well in a series of structure validation criteria, particularly when considering only the WW domain. Some of the errors arise from the lack of NOE restraints, particularly the proline residues and the flexible region between the two domains. However, the presence of the flexible region ensures the two domains operate independently of each other, which means it is possible to use this structure as a reasonably accurate representation of the native WW4 domain. This structure is close to completion and, despite much time and effort being invested in to investigating errors thus far, requires further refinement before being deposited in to the PDB and being published.

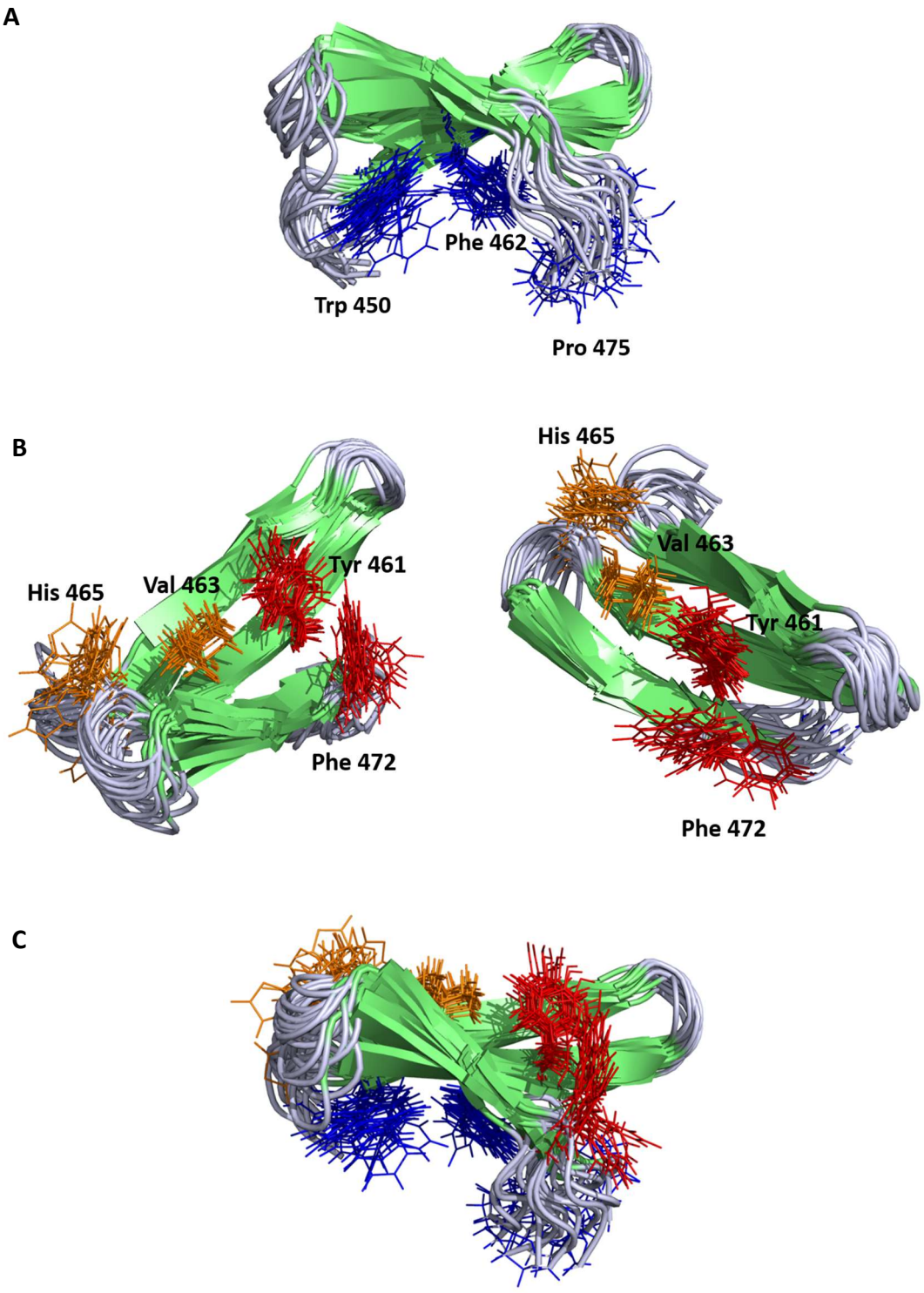


Figure 4.3.3 - A: The WW4 refined, 20 model CNS ensemble with the hydrophobic underside shown in blue, consisting of tryptophan 450, phenylalanine 462 and proline 475. B: The binding site residues from two angles, the XP groove is shown in red and consists of tyrosine 461 and phenylalanine 472. The specificity pocket is shown in orange and consists of valine 463 and histidine 465. C: The overall architecture of the WW4 domain with the binding site and hydrophobic underside shown.

5. WW domain substrate interactions and a new WWP2 isoform

5.1 Introduction

5.1.1 NEDD4 E3 ligase activity in the TGFβ pathway

Many of the NEDD4 family of E3 ligases have been linked to regulation of the TGFβ signalling pathway and interaction with its signalling components. SMURF1 interacts with Smad7 but does not interact with Smad2 and Smad3 (Ebisawa et al. 2001; Zhu et al. 1999). When SMURF1 binds to Smad7, the complex translocates to the cytoplasm and is recruited to the TβR-I via Smad7 interaction with the receptor, facilitated by the C2 phospholipid localisation domain of SMURF1 (Ebisawa et al. 2001; Suzuki et al. 2002). Whereupon the TβR-I and Smad7 are polyubiquitinated, causing their degradation at the proteasome (Ebisawa et al. 2001). SMURF1 serves to inhibit the TGFβ signalling pathway by preventing the propagation of the cytokine signal across the cell surface, thereby preventing the phosphorylation and activation of r-Smads 2 and 3.

SMURF2 interacts with r-Smads through its second and third WW domains, with a preference for Smad2 over Smad3 (Lin et al. 2000; Zhang et al. 2001). The interaction between Smad2 and Smurf2 is dependent upon phosphorylation of Smad2 during TGFβ stimulation, and it is thought that SMURF2 serves to selectively degrade Smad2 upon activation (Lin et al. 2000; Bonni et al. 2001). There is conflicting evidence as to whether SMURF2 polyubiquitinates Smad2; rather, it is thought that Smad2 serves as an adapter, causing the ubiquitination and degradation of its cofactors (Lin et al. 2000; Bonni et al. 2001; Zhang et al. 2001). SMURF2 also binds to Smad7 through both the second and third WW domain, but shows little ubiquitinating activity in the absence of TGFβ stimulation (Kavsak et al. 2000). As with SMURF1, upon binding Smad7 the complex is exported from the nucleus and recruited to the activated TβR-I via receptor interaction with Smad7. The type-I receptor and Smad7 are ubiquitinated and degraded through the action of SMURF2 (Kavsak et al. 2000). Through the action of Smad7 interaction and receptor degradation, SMURF2 acts to negatively regulate the TGFβ signalling pathway.

The r-Smads have several phosphorylation sites; during activation by their receptors, the r-Smad C-terminal tail is phosphorylated, which relinquishes Smad autoinhibition and allows Smads to oligomerise with their transcription cofactors. The disinhibition of Smad2/3 also allows regulatory kinases to access the linker region

between their two MH domains, where the PPxY motif resides. CDK8 and CDK9 are two such kinases that are responsible for phosphorylating Smad2 at threonine 220 and Smad3 at threonine 179, which in both instances are two residues N-terminal to the PPxY motif (Alarcón et al. 2009). NEDD4L binds to the PPxY motifs of Smad2 and Smad3 exclusively through its second WW domain, but only once they have been phosphorylated by CDK8/9 (Gao et al. 2009). This interaction causes activated Smad2/3 to be polyubiquitinated and degraded, and serves to limit the intensity of TGFβ signalling (Gao et al. 2009). NEDD4L has also been shown to bind Smad7 (Aragón et al. 2012; Yan et al. 2016).

WWP1 has been shown to negatively regulate the TGFβ pathway (Komuro et al. 2004). Functional assays show that transcriptional activity by Smad2/3 is reduced, and that WWP1 interacts with Smad2/3 and Smad7 with equal affinity. WWP1 does not ubiquitinate Smad2/3 but instead, upon binding to Smad7, is recruited as part of the WWP1/Smad7 complex to the activated TβR-I. WWP1 then ubiquitinates and downregulates TβR-I (as well as Smad7), which prevents the phosphorylation and activation of Smad2/3. Through the same action as SMURF1 and SMURF2, WWP1 serves to negatively regulate and limit the duration of TGFβ signalling.

Smad7 plays a central role in the ubiquitin-mediated regulation of the TGFβ pathway, and because of this the interaction between Smad7 and NEDD4 E3 ligases have been studied in some detail. A number of NEDD4 family WW domains have had their dissociation constant (K_d) values measured with respect to the Smad7 PPxY ligand, and many of them have had their structures resolved in the bound conformation. The K_d values for several different WW domains and the Smad7 ligand are shown in Table 5.1.1. Generally WW domain dissociation constants lie within the low mM to high nM range for proline-rich motifs, and the low mM range for phosphoserine-proline and phosphothreonine-proline motifs (Macias et al. 2002). The majority of the Smad7-specific WW domains of the NEDD4 family have dissociation constants that lie within the low μM range.

Protein	WW domain	Smad7 PY peptide K_d (µM)
Nedd4L	WW1	23.6 ±3.6
	WW2	4.2 ±0.1
	WW3	8.0 ±0.3
	WW4	12.4 ±1.8
	WW1-2	18.3 ±7.4
	WW3-4	16.9 ±1.0
SMURF1	WW1	>100
	WW2	4.1 ±0.1
	WW1-2	1.7 ±0.5
SMURF2	WW1	>100
	WW2	>100
	WW3	4.5 ±0.2
	WW2-3	1.7 ±0.4

Table 5.1.1 The dissociation constants for the interaction between various NEDD4 family WW domains and the Smad7 PPxY ligand, including three NEDD4 E3 ligase family members, reproduced from (Aragón et al. 2012). K_d values were obtained using isothermal titration calorimetry (ITC) at 15°C.

5.1.2 WWP2 isoform activity

As discussed in the first chapter of this thesis, WWP2 has three isoforms that are believed to have distinct activities in regulating the TGFβ signalling pathway (Figure 5.1.1A and B). In immunoprecipitation assays the full length isoform WWP2-FL interacted with Smad2, Smad3 and Smad7, the WWP2-N isoform interacted with Smad2 and Smad3 but not Smad7 while WWP2-C interacted with Smad7 only (Soond & Chantry 2011). The Smad2 and Smad3 interactions were TGFβ-dependent. WWP2-FL showed minimal ubiquitinating activity against Smad2 and Smad3 (a presumed monoubiquitination looks more like a phosphorylation), but polyubiquitinated and caused the degradation of Smad7. This, coupled with a decrease in Smad2/3-dependent promoter activity during WWP2-FL overexpression, indicates that WWP2-FL most likely operates through the same mechanism as SMURF1/2 and WWP1 by polyubiquitinating and degrading the TGFβ receptors and Smad7. The N-terminal isoform WWP2-N has no ubiquitin ligase activity

because it is missing the HECT domain, but is believed to play a role in stimulating the activity of the full-length isoform by preventing its autoinhibition, and has been shown to upregulate WWP2-FL activiy against Smad2/3 (Soond & Chantry 2011). Autoinhibition has been observed in the NEDD4 family members of WWP2, SMURF2 has been shown to display autoinhibitory activity by binding of its C2 domain to its HECT domain near the catalytic cysteine, and preventing ubiquitin charging by its E2 (Wiesner et al. 2007). This autoinhibitory binding is displaced by SMURF2 substrates which activates its ubiquitinating activity, it is therefore feasible that WWP2-N might displace WWP2-FL autoinhibitory activity by acting as a binding partner. WWP2-N expression is thought to be the result of an alternative splicing event and raises the prospect of integrating alternative splicing programs into the control of the pathway. In the TGFβ pathway, the C-terminal isoform WWP2-C is thought to exclusively bind and ubiquitinate Smad7, a process which is enhanced by TGFβ stimulation (Soond & Chantry 2011). Unlike WWP2-FL, this ubiquitination of Smad7 causes the activity of the Smad2/3 promoter activity to increase, indicating that the activating arm of the pathway remains active and that, potentially, the TGFβ receptors are preserved. WWP2-C represents an alternative mode of TGFβ regulation by the exclusive degradation of the Smad7 inhibitory component, increasing the impact of TGFβ signalling by prolonging the duration and intensity of activation.

The different apparent activities of the WWP2 isoforms (Figure 5.1.1B) are thought to be linked to their different domain architectures, and raises questions about the roll the different WW domains play in selecting substrates. Each isoform has a different combination of domains, these are shown in Figure 5.1.1A. It is thought that the first WW domain might confer Smad2/3 binding specificity, while it is expected that the fourth WW domain, the structure of which has been solved in chapter 4 of this thesis, should confer selectivity for Smad7 by binding the Smad7 PPxY motif.

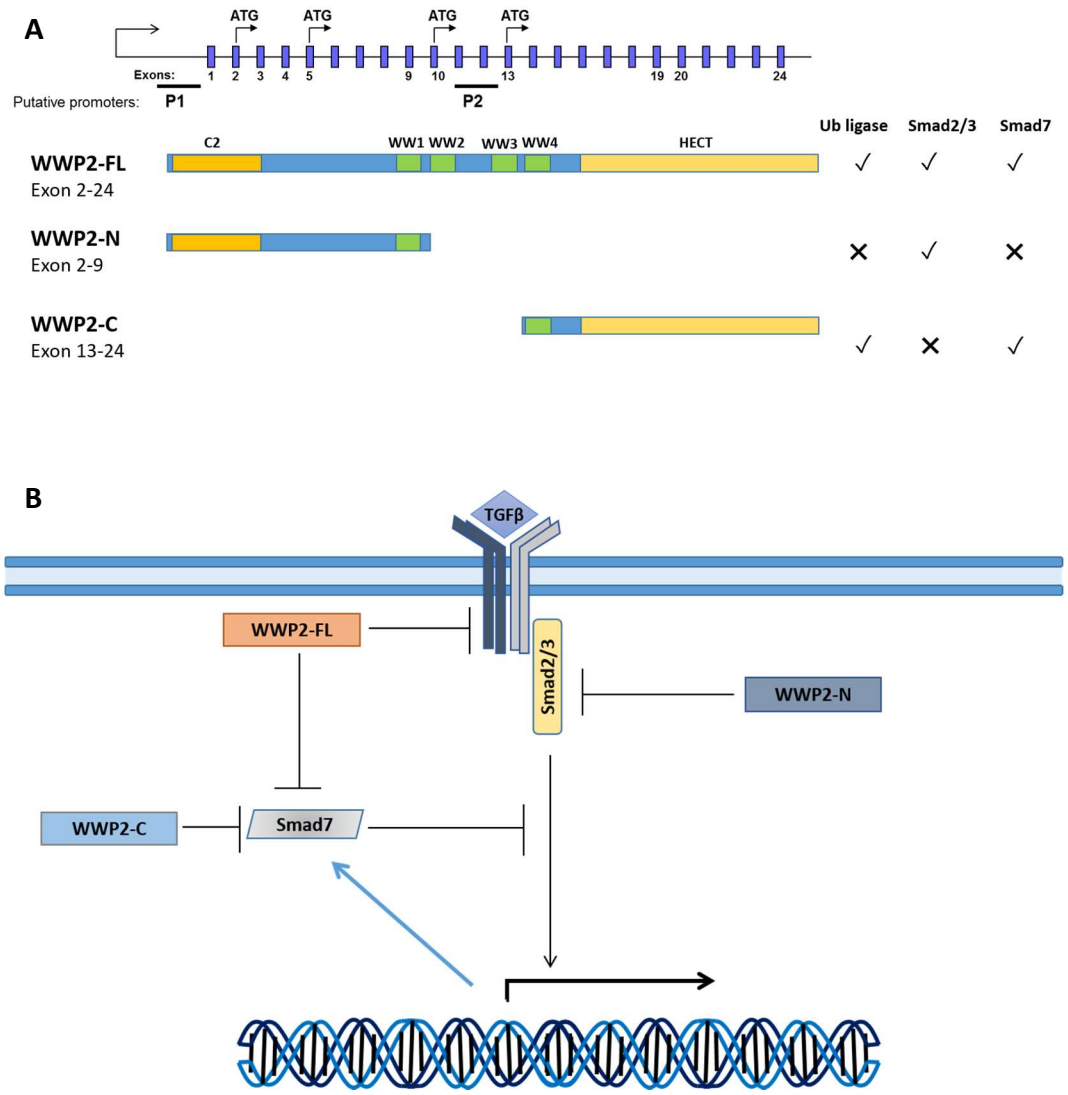


Figure 5.1.1 - A: The WWP2 gene locus (not to scale) and the different WWP2 isoforms and their domain architecture (also shown in Figure 1.7.6). WWP2-FL and WWP2-N are under the control of the same promoter. The latter is the result of alternative splicing which causes the retention of intron 9/10, introducing a premature stop codon. WWP2-C is believed to be under the control of a secondary promoter, regulating a start codon at exon 13. WWP2-FL has a C2 domain (orange), a full complement of WW domains (green) and a HECT domain (yellow). The N-terminal isoform retains a C2 domain and the first WW domain. WWP2-C has the fourth WW domain and the HECT domain. Ubiquitin ligase activity and Smad binding partners are also shown. B: The different apparent activities of the WWP2 isoforms in the TGFβ signalling pathway. WWP2-FL is predicted to be active against Smad7 and the TGFβ receptors, WWP2-N is believed to upregulate degradation of Smad2/3 and WWP2-C is believed to be active against Smad7.

5.1.3 NEDD4 family Smad7 interactions

Since WWP2 WW4 has been shown to bind Smad7 in immunoprecipitation assays, and that several other WW domains of the NEDD4 family also bind Smad7, it is thought that sequences of these domains might share some similarities. Figure 5.1.2 shows the sequence alignment and sequence identity between WWP2 WW4 and the other WW domains of WWP2, SMURF1, SMURF2, WWP1, and NEDD4L. Unsurprisingly, the WW4 domain of the closely related WWP1 E3 ligase shares the highest identity with WWP2 WW4, at 78.38%. WWP2 WW4 shares 56.67% identity with the Smad7 binding SMURF1 WW2 and SMURF2 WW3. There are three other WW domains besides WWP2 WW4 that share a phenylalanine at the position of the canonical second tryptophan, labelled 4*, two of which (SMURF1 WW2 and SMURF2 WW3) have been shown to have high affinities for the Smad7 PPxY ligand. Therefore, phenylalanine at this position does not preclude WWP2 WW4 from also binding Smad7, and likewise valine at 2*. Histidine at position 3* is largely conserved across all of the WW domains, besides SMURF1 WW1 and SMURF2 WW2 which have a threonine instead. As discussed in Chapter 1 of this thesis, this histidine is important in conferring selectivity for the tyrosine of the PPxY motif. These two domains show low affinity for the Smad7 ligand, and it may be that this position in the secondary specificity pocket has some influence. SMURF2 WW1, which exhibits low affinity for Smad7, has a histidine at this position, but unusually, has a glutamine at position 1* of the XP binding pocket which may be responsible for its low affinity.

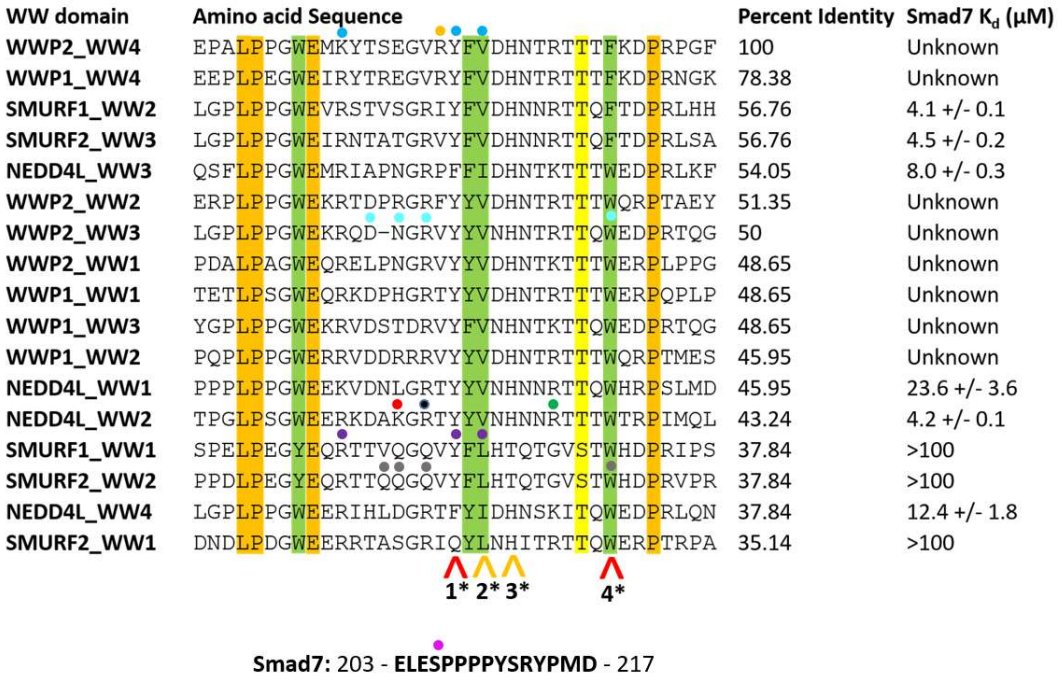


Figure 5.1.2 - Sequence alignment between WW4 and the WW domains of the Smad7-binding NEDD4 family members, generated using the Clustal Omega online server (Goujon et al. 2010; Sievers et al. 2011). Highlighted in orange are conserved residues, in green are residues with strongly related properties and in yellow are residues with weakly related properties, according to the Clustal Omega server. Red arrows indicate the residues of the XP binding pocket and orange arrows indicate residues of the secondary specificity pocket. Coloured dots indicate important residues involved in ligand binding, as discussed in the text. Sequences are organised in relation to percent identity with the WWP2 WW4 domain. Affinities for the Smad7 ligand are shown, as determined by ITC (Aragón et al. 2012). The Smad7 peptide is also shown.

Key residues in NEDD4L WW2 interaction with Smad7 are the 2* valine, 3* histidine, the green dot arginine and the yellow highlighted threonine (Figure 5.1.3, and Figure 5.1.2), all of which interact with the tyrosine of the PPxY motif.

Figure 5.1.3 - The NEDD4L WW2 domain amino acid sequence. Percent identitiy with the WWP2 WW4 domain is shown, as well as Smad7 affinity. The XP binding site is marked with red arrows, the secondary specificity pocket is marked with orange arrows and important binding residues are marked with coloured dots.

The 4* tryptophan, 1* tyrosine, and yellow highlighted threonine interact with the PPxY motif prolines. These residues are largely conserved in the WWP2 WW4 domain. An arginine at -2 from the 1* residue (black dot) is involved in binding Smad7 glutamic acid 205 (N-terminal to the PPxY motif). The lysine at -4 from the 1* residue (red dot) as well as the arginine at -2 (black dot) are involved in binding aspartic acid 217 of the ligand (C-terminal to the PPxY motif). These residues serve to substitute the electrostatic phospho-threonine interaction that NEDD4L typically experiences when it interacts with its phosphorylated Smad2/Smad3 ligands (Aragón et al. 2012). Arginine at this position is conserved amongst the WW domains that bind Smad7 with a high affinity, and is substituted in those that do not. WWP2 WW4 has a valine substituted at the site of the arginine, and a glutamic acid substituted at the site of the lysine. In fact, when these positions were substituted to glutamic acid in the study of this interaction, the affinity was reduced 4-5 fold (Aragón et al. 2012). It is likely that the double substitution in WWP2 WW4 will lower the affinity of WW4 for Smad7. The arginine N-terminal to the substituted valine of WWP2 WW4 (Figure 5.1.2, orange dot) should have no recuperative effect, as it faces the opposite side of the β-sheet in the structure from Chapter 4.

During the study of the SMURF1 WW1 domain and a phosphorylated Smad1 ligand, it was discovered that phosphorylation was central to the activity of receptor Smad proteins (Aragón et al. 2011). First, phosphorylation of Smad1 enhanced the affinity for its transcriptional activator, and then a further phosphorylation switched its affinity, so that the SMURF1 WW1 domain preferentially bound, leading to its degradation (Aragón et al. 2011). The residues involved in coordinating the phosphate group are a tyrosine (presumably through its hydroxyl group), arginine and leucine (purple dots, Figure 5.1.4 and Figure 5.1.2).

Figure 5.1.4 - The SMURF1 WW1 domain amino acid sequence. Percent identitiy with the WWP2 WW4 domain is shown, as well as Smad7 affinity. The XP binding site is marked with red arrows, the secondary specificity pocket is marked with orange arrows and important binding residues are marked with coloured dots.

The corresponding residues in WWP2 WW4 are a lysine, tyrosine and valine (Figure 5.1.2, blue dots), which are either identical or related, indicating that WWP2 WW4 may have some group IV pS/T-P affinity. Polyubiquitination and subsequent degradation

of Smad7 by WWP2-C is enhanced during stimulation by TGFβ (Soond & Chantry 2011). Three possible, and equally feasible, causes include: the modification of WWP2-C, the introduction or modification of a scaffold molecule, or the modification of Smad7 to enhance its affinity for WWP2-C. Given the precedent of a "phophoserine code" determining the outcome of r-Smad signalling, whereby phosphorylation enhances Smad/WW domain affinity (Aragón et al. 2011), it is possible that phosphorylation at a Smad7 S/T-P motif might enhance its affinity for WWP2-C. NEDD4L-mediated degradation of r-Smads is dependent on CDK-mediated phosphorylation of a threonine 2 residues N-terminal to the PPxY motif (Gao et al. 2009). The corresponding position in Smad7 is a serine at position 206 (Smad7 peptide pink dot Figure 5.1.2). A paper released in 2001 explored Smad7 phosphorylation, and during the expression of a C-terminal portion of Smad7 it was found that serine 206 was phosphorylated (Pulaski et al. 2001). Phosphorylation prediction servers rank the probability of this site being phosphorylated as low (results not shown), but the server also misses the major phosphorylation site discussed in this paper, serine 249. This leads us to the conclusion that a phosphorylation event at serine 206 of Smad7 might have some influence on the affinity of the WW4 domain of WWP2 for the Smad7 PPxY ligand, and merits further exploration.

5.1.4 Tandem WW domains

Cooperation and communication amongst tandem WW domains is a well-documented phenomenon (Dodson et al. 2015; Aragón et al. 2011; Webb et al. 2011; Fedoroff et al. 2004; Chong et al. 2010). Five models have been proposed: 1 - Tandem WW domains bind target motifs independently but when present together, effective binding affinity is increased; 2 - The second domain does not participate in binding but stabilises the first; 3 - Binding of the first WW domain to one motif enhances the affinity of the second domain for another motif; 4 - Presence of a second WW domain alters the stability and dynamics of the first, changing the bound conformation and possibly changing ligand preference; 5 - Tandem WW domains bind different ligands independently, but if one WW domain is not in the bound conformation it may have a disruptive effect on the binding affinity of its tandem domain (Dodson et al. 2015).

The SMURF2 WW3 domain binds the PPxY motif of Smad7 with high affinity, as shown in Table 5.1.1, but the WW2 domain does not (Chong et al. 2010; Aragón et al. 2012). When the two domains are present in tandem, the WW2 domain makes contacts

with the WW3 domain, and binds to the ESP residues from the N-terminal tail of the Smad7 PPxY motif (Chong et al. 2010). This enhances the affinity of the interaction (Table 5.1.1) via a combination of the second/third models described above. The same is true of SMURF1 WW1 and WW2 domains (Table 5.1.1), since SMURF1 and SMURF2 are very similar (Chong et al. 2010; Aragón et al. 2012). An alternative isoform of SMURF1 exists that is the product of alternative splicing, and contains a 26 residue insert between the two domains. This insert reduces the inter-WW domain contacts; subsequently, the Smad7 ligand interaction is altered and affinity for the motif is reduced (Chong et al. 2010). However, it has also been proposed that the secondary contacts between SMURF2 WW2 domain and the N-terminal tail of Smad7 play a relatively minor role in ligand binding, and instead contributes to ligand affinity through dimerisation and oligomerisation (Aragón et al. 2012). The WW3 and WW4 domains of NEDD4L, on the other hand, seem to have a slightly reduced affinity for Smad7 (Table 5.1.1), and may be mutually disruptive when present in tandem (Aragón et al. 2012).

The WWP2 E3 ligase has multiple WW domain repeats, and there is an emerging layer of complexity in which WWP2 isoforms with fewer repeats have different regulatory roles. This raises questions about the role of the WW domain repeats, in particular the difference between the Smad7 affinity of WWP2-C, which only has the WW4 domain, and WWP2-FL, which has WW3 and WW4 in very close proximity (the C-terminal tryptophan of WW3 and N-terminal tryptophan of WW4 are only separated by 17 residues, the sequence separating SMURF2 domains is 23 residues). The residues involved in SMURF2 WW2 cooperative binding of Smad7 are the side chains of three glutamines and the C-terminal tryptophan (grey dots, Figure 5.1.5, and Figure 5.1.2) (Chong et al. 2010).

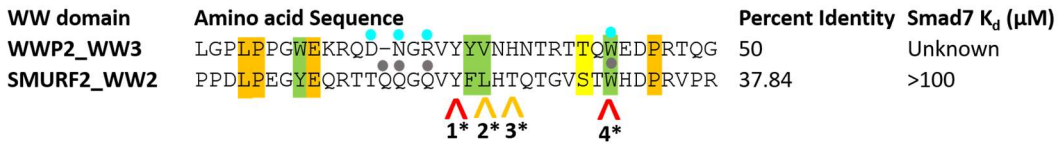


Figure 5.1.5 - The WWP2 WW3 and SMURF2 WW2 amino acid sequences. Percent identitiy with the WWP2 WW4 domain is shown, as well as Smad7 affinity. The XP binding site is marked with red arrows, the secondary specificity pocket is marked with orange arrows and important binding residues are marked with coloured dots.

The corresponding residues in WWP2 WW3 are an aspartic acid, asparagine, an arginine and the C-terminal tryptophan (turquoise dots), which are only broadly related. This indicates that WW3 may cooperate in the same fashion. Because there is evidence

of isoform-based substrate selection, outlined in the previous paragraph, it is plausible that the WW3 domain in tandem with the WW4 domain might have some influence over Smad7 affinity, and will be explored further in this chapter.

5.1.5 Evidence of a new WWP2 isoform

During the course of experiments directed at determining a potential link between EMT and the expression of different WWP2 isoforms, this lab has explored the effects of epithelial splicing regulatory proteins (ESRPs) on isoform expression. TGFβ has been shown to downregulate ESRPs and in so doing, drive the progression of EMT (Horiguchi et al. 2012). Figure 5.1.6A shows a western blot from one such experiment performed by a former post doc in our lab, Dr Surinder Soond, looking at the effects of ESRPs on isoform expression. The blot was performed on lysates from mammalian epithelial cells that were either unstimulated or stimulated with TGFβ, and that had been transfected with a different combination of ESRPs. The antibody used was raised to target only the WW4 domain portion of WWP2, as shown in Figure 5.1.6B, with the original purpose of detecting the WWP2-C isoform. The WWP2-C isoform, with a molecular weight of 51 kDa, has not been found to be widely expressed, and cannot be seen in this blot. The two bands that can be seen correspond to WWP2-FL (99 kDa) and an unknown TGFβ-inducible band at roughly 30 kDa. This band is not likely to be WWP2-N which, unlike WWP2-C, is not TGFβ-inducible and does not contain the WW4 domain (Soond & Chantry 2011). The band is also unlikely to correspond to any of the known isoforms at this molecular weight. We are therefore inclined to believe that this band might represent a novel WWP2 isoform, and that since its expression seems to be altered by ESRPs (Figure 5.1.6A), the isoform might arise from alternate splicing events at the WWP2 gene. Since this band cross-reacted with the antibody designed to detect the WW4 domain portion of WWP2, we assumed that this isoform contained the WW4 domain. If this protein were to arise from the putative WWP2-C promoter P2 (Figure 5.1.1), it would require the retention of intron 19-20, which would generate a premature stop codon and produce a protein product from exons 13-19 with a molecular weight of 31.5 kDa. This protein would also cross-react with the anti-WWP2-C antibody. This putative new isoform would contain the fourth WW domain but would terminate at residue 706, midway through the HECT domain, shown in Figure 5.1.6C. This isoform would lack the HECT C-terminal lobe, which

crucially contains the catalytic cysteine, and would therefore be unable to ubiquitinate substrates. The termination point is towards the C-terminal half of the E2 binding site, and it is unclear whether E2-binding would be preserved, or whether the domain would be able to fold correctly since two helices of the large subdomain are also missing. This new isoform will, from here on, be referred to as WWP2C-ΔHECT.

The expressed sequence tag (EST) database is a repository of short sequences of cDNA from numerous sources. These cDNA sequences correspond to small regions of mRNA transcripts and can be used to identify new genes. They are also useful for identifying transcript variants from the same gene. This is achieved by aligning the EST to the gene of interest and identifying intronic regions that are retained in the EST (and therefore the mRNA transcript). By doing this, transcription start sites and early stop codons can be identified, and therefore N and C-termini of potential new isoforms can be discerned. This approach will be used here to help identify the transcript responsible for the WWP2C-ΔHECT isoform.

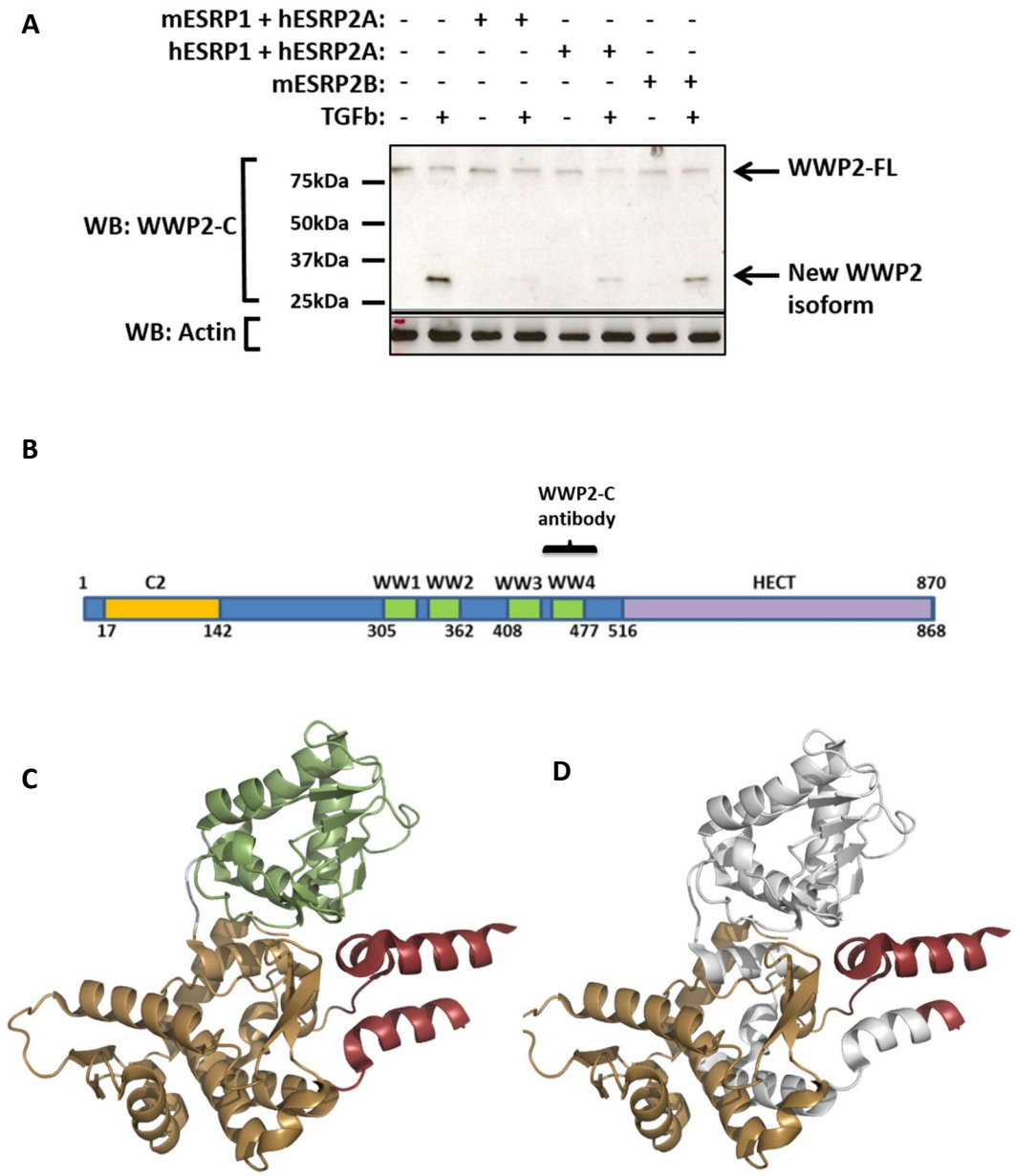


Figure 5.1.6 - A: Western blots of mammalian cell lysates probed with anti-WWP2-C antibody, samples shown are unstimulated and stimulated with TGFβ and transfected with a combination of ESRPs. The β-actin control is also shown. B: The region of WWP2-FL used as an epitope to raise the WWP2-C-specific antibody. C: The WWP2 HECT domain crystal structure (PDB: 4Y07) (also shown in Figure 3.3.1), with the C-terminal lobe in green, the E2 binding site in red (a lack of electron density means that only a partial structure of this region was determined) and the large subdomain in brown. D: The WWP2 HECT domain crystal structure (Gong et al. 2015) with the region missing in the putative new isoform in white - note, a small region from residues 661-702 is missing from the structure due to a lack of X-ray diffraction for this area, resulting in an incomplete electron density map.

Given that SDS-PAGE based molecular weight determination is notoriously inaccurate, it is not by any means certain that the new isoform is the product of exon 13-19. Molecular weight ambiguity in SDS-PAGE gels seems to be particularly relevant for WWP2 isoforms. For example, in Chapter 3 of this thesis the bacterial expression and purification of WWP2-C, a 53 kDa protein, produced a protein that had an apparent SDS-PAGE molecular weight of 45 kDa. While WWP2 HECT, a 46.5 kDa protein, had an apparent SDS-PAGE molecular weight of 40 kDa. This raises the possibility that WWP2C-ΔHECT is actually slightly larger than it appears. Therefore, another curious prospect is that this isoform might contain more WW domains, and thereby have an alternative substrate preference mediated by tandem WW domains, as outlined above. Should WWP2C-ΔHECT contain WW3 and WW4 it might have a greater affinity than WWP2-C for the Smad7 ligand and, in the context that both isoforms were expressed simultaneously, would effectively out compete WWP2-C for Smad7. Since it is almost certain WWP2C-ΔHECT has no ubiquitin ligase activity, it would act to preserve Smad7 by acting as a dominant negative and blocking ubiquitination.

5.1.6 Ligand interaction by NMR

In this project, NMR spectroscopy was used to explore the interactions between Smad7 and the following WW domains of WWP2: WW4, WW3 and WW3-4. Isothermal titration calorimetry is the conventional technique used to determine the dissociation constants between protein and ligand. However, NMR spectroscopy offers the ability to examine the interaction at an atomic level. The specific binding residues and their affinities can be determined by tracking the change in resonance of the backbone amide in hydrogen-nitrogen correlated HSQCs (Fielding 2003). The backbone amide resonance changes when it experiences alterations in its local magnetic field. There are two reasons that this might change during the addition of a ligand, one is allosteric changes as the protein conformation adjusts to the newly introduced ligand, and the second is direct interaction with the ligand, which introduces a significant perturbation to the local magnetic field.

During the study of ligand interaction by NMR there are three possible outcomes that are determined by the rate of exchange between bound and unbound conformations: fast exchange, intermediate exchange and slow exchange. During ligand

interaction each nucleus will have two resonance values, one in the unbound conformation and one in the bound conformation. In fast exchange interactions, the bound and unbound conformation exchange quicker than the difference between the resonant frequencies of the two different states (rate of exchange $> \nu_{unbound} - \nu_{bound}$), so that we observe an average of the two positions. This will be weighted towards which conformation is most abundant, so that when more ligand is added, the resonance moves further towards the bound conformation resonance position. In slow exchange, there is no chemical exchange during the detection period and we therefore observe two peaks, one corresponding to the unbound position and one corresponding to the bound position. The intensity of the peaks is dependent on which conformation is most abundant. The third timescale is intermediate exchange, during which we observe one peak that broadens, decreasing in intensity as the exchange process is at the same frequency as the difference between the resonant frequencies of the two different states (rate of exchange $\approx \nu_{unbound} - \nu_{bound}$), so that it interferes with signal detection. The resonance position is uncertain, so the linewidth of the peak is spread over a large area. The broadening of the linewidth holds information about the rate of exchange.

Upon titration of a ligand in fast-exchange, those amide resonances that do not bind the ligand should experience minimal magnetic field perturbation and any movement seen should be insignificant. Those resonances directly involved in ligand binding should move a substantial amount in a pattern consistent with a binding curve, and saturation should eventually be evident. The rate of changing chemical shift can be used to calculate the dissociation constant for each residue. This method is also important in the process of generating a co-structure, in which the WW domain structure is solved in complex with its ligand, and informs the design of sample conditions for these experiments.

5.1. Experimental aims

Given the apparent interaction in immunoprecipitation assays between Smad7 and the WW4 domain of WWP2, the aim of this chapter is to use NMR spectroscopy titrations to observe an interaction between the GB1:WW4 protein used in the previous chapters of this thesis, and a Smad7 PPxY-containing ligand (Soond & Chantry 2011). And, since in the same immunoprecipitation experiment it appeared as though there is no

interaction between r-Smads and WW4, it is expected that no interaction will be observable between the Smad2/3 ligands and WWP2 WW4. Given that Smad7 turnover by WWP2-C appears to be TGFβ-dependent, it is thought that some level of phospho-regulation might be evident. Using a phosphorylated Smad7 ligand in NMR titration experiments, this chapter will attempt to determine whether there is an interaction between GB1:WW4 and a phosphorylated Smad7 ligand. NMR will be used to determine the WW4 binding site that accepts the poly-proline ligand, and to determine preliminary dissociation constants that might help elucidate substrate preference. Preliminary steps will be taken towards obtaining a bound structure of WW4 and its ligand. Using the same method of NMR titration experiments, this chapter will determine the differences, if any, between tandem WW domain ligand affinity, and the affinity of the WW3 and WW4 domains individually. Using these experiments, the binding site will be mapped on to the WW4 structure solved in Chapter 4. This chapter will also explore the possibility that a novel TGFβ-inducible WWP2 isoform might exist, and using semi-quantitative PCR it is hoped that evidence will be found to identify the region of WWP2 to which it corresponds.

5.2 Results

5.2.1 WW4 and Smad7

The Smad7 ligand was designed around the stretch of peptide most commonly used in the various studies of WW/Smad7 interactions found in the literature. A synthetic peptide was used that started at residue 203 of Smad7 and ended at 217 and incorporated the PPxY motif as below:

203 - ELESPPPPYSRYPMD **- 217**

The peptide was titrated into a 0.78 mM ^{15}N labelled sample of GB1:WW4, and after each titration point a ^{1}H-^{15}N-HSQC was acquired. A total of 10 titration points were collected from a molar ratio of 1:0 protein to peptide, to a molar ratio of 1:10, where the concentration of Smad7 was 10-fold higher than GB1:WW4. Figure 5.2.1 shows the migration of resonances that were observed during the titration.

WW4 amide peaks can be seen migrating upon the addition of peptide to the sample, from which it can be deduced that WW4 does indeed interact with the Smad7 ligand, and this interaction is in fast-exchange. Figure 5.2.2 shows the shift trajectories plotted against residue number. Shift changes in the spectra are unevenly weighted towards hydrogen, because of the difference in gyromagnetic ratios between the two elements. To calculate the shift trajectories, the changes in shift in each dimension were weighted appropriately so as to adjust for this. From this plot it is clear that although some minor shift changes occurred with the GB1 domain amide peaks, likely due to non-specific effects, by far the greatest change in shift was observed with peaks from the WW4 domain. In particular the third strand of the β-sheet, although the first and second strands also experience a significant perturbation of the local magnetic field, and therefore it is these regions that correspond to the Smad7 binding site.

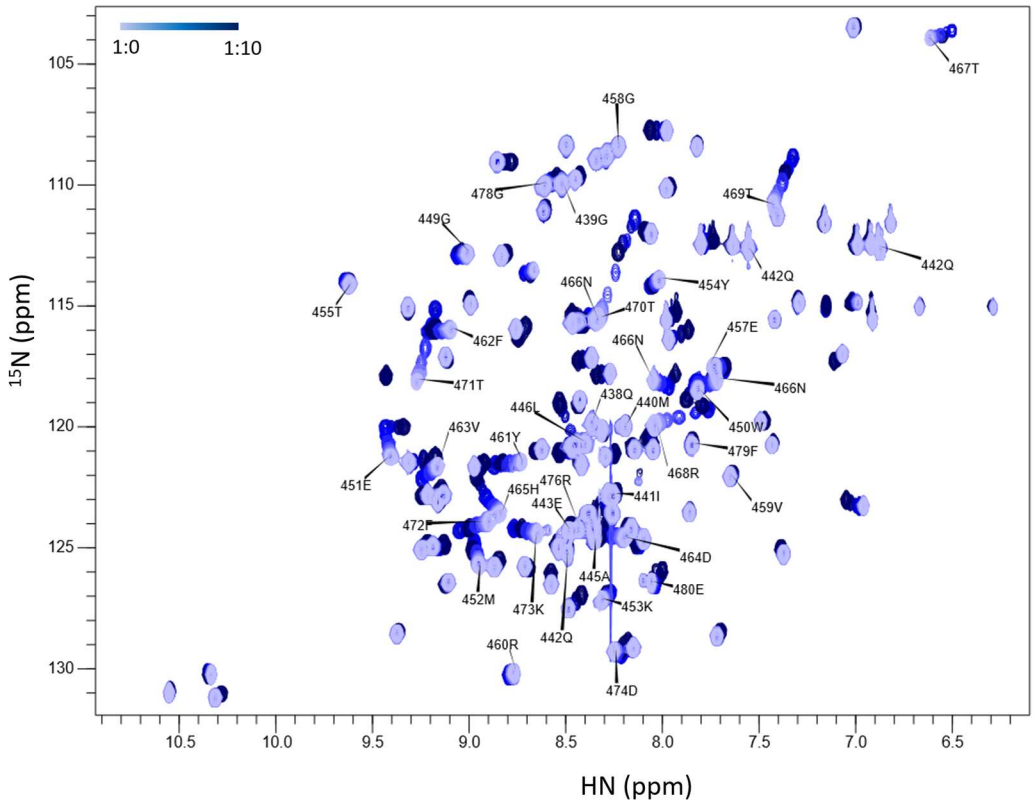


Figure 5.2.1 - An overlay of the HSQC spectra from the GB1:WW4 Smad7 titration with only the WW4 domain peaks (438-480) labelled, for clarity. Light blue peaks are from the first point along the titration at a molar ratio of 1:0 protein to Smad7 ligand, and the darkest blue peaks are from the final titration point at 1:10. The experiment was performed at 500 MHz, 298 K. The sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. The concentration of GB1:WW4 was 0.78 mM.

By observation of the graph in Figure 5.2.2A three broad levels of shift change have been defined. The three threonines (469-471) of the third β-strand experience the greatest change in shift between 0.851-0.408 ppm, these are shown in red in Figure 5.2.2. Arginine 468, glutamate 451 and histidine 465 all experience a change in shift between 0.324-0.256 ppm, these are shown in orange in Figure 5.2.2. While tryptophan 450, methionine 452, tyrosine 461, phenylalanine 462, valine 463, aspartate 464, threonine 467 and phenylalanine 472 all experience a shift in the range of 0.214-0.151 ppm and these are shown in yellow in Figure 5.2.2. Because of the extent of chemical shift change, these residues are believed to be the binding site. The remainder of the WW4 domain residues experienced a change in shift comparable to the minor changes experiences by the GB1 domain. The residues corresponding to the XP binding pocket are phenylalanine 472 and tyrosine 461, and the residues corresponding to the secondary specificity pocket

binding region are valine 463 and histidine 465. All of these experience some level of magnetic field perturbation consistent with the docking of a ligand. The trajectory provides information about the binding site of the peptide but does not in itself give any detail about the affinity of the interaction. To determine the K_d we need to consider the rate of change in shift and the point of saturation.

Figure 5.2.3 shows the change in shift plotted against molar ratio for a selection of the residues of the binding pocket, alongside the corresponding region in the HSQC. From observation of the HSQC, it became apparent that the migration of chemical shift seemed to have two components. During the first stages of titration the peaks move along one trajectory, but once a molar ratio of approximately 1:6 is reached, the peak trajectory changes and either doubles back on itself or moves in another direction altogether. This manifests itself in the ΔShift graphs as a two stage curve, the first part of which appears to be reaching saturation until the second phase continues. The pattern of a two stage binding curve is consistent across all of the binding site residues.

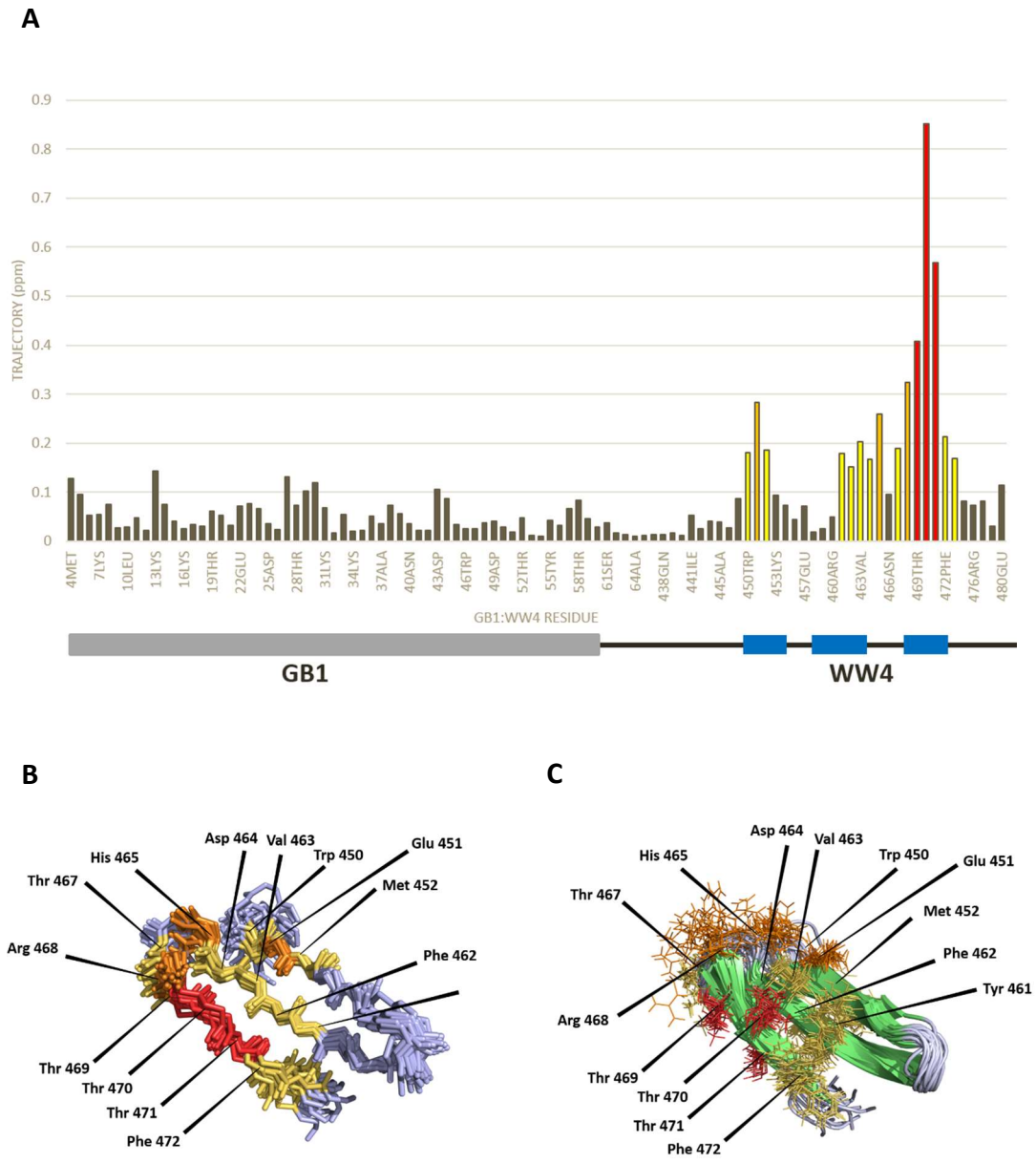


Figure 5.2.2 - A: The trajectory (in ppm) of each GB1:WW4 backbone amide assigned in the 1H-15N-HSQC (there is no information for prolines), upon titration of the Smad7 ligand. A schematic is aligned with the graph showing the GB1 domain in grey and the three strands of the WW4 domain β-sheet in blue. B: The WW4 domain structure backbone ensemble viewed from the binding surface with the trajectories visualised as a heat map. C: The WW4 domain structure ensemble viewed as the PyMol cartoon graphic with the side chains of the binding site showing, and colour coded according to the trajectory heat map. It is presumed that the surface which binds the ligand is the same as other WW domain, as described in Chapter 1, and this surface is orientated upwards.

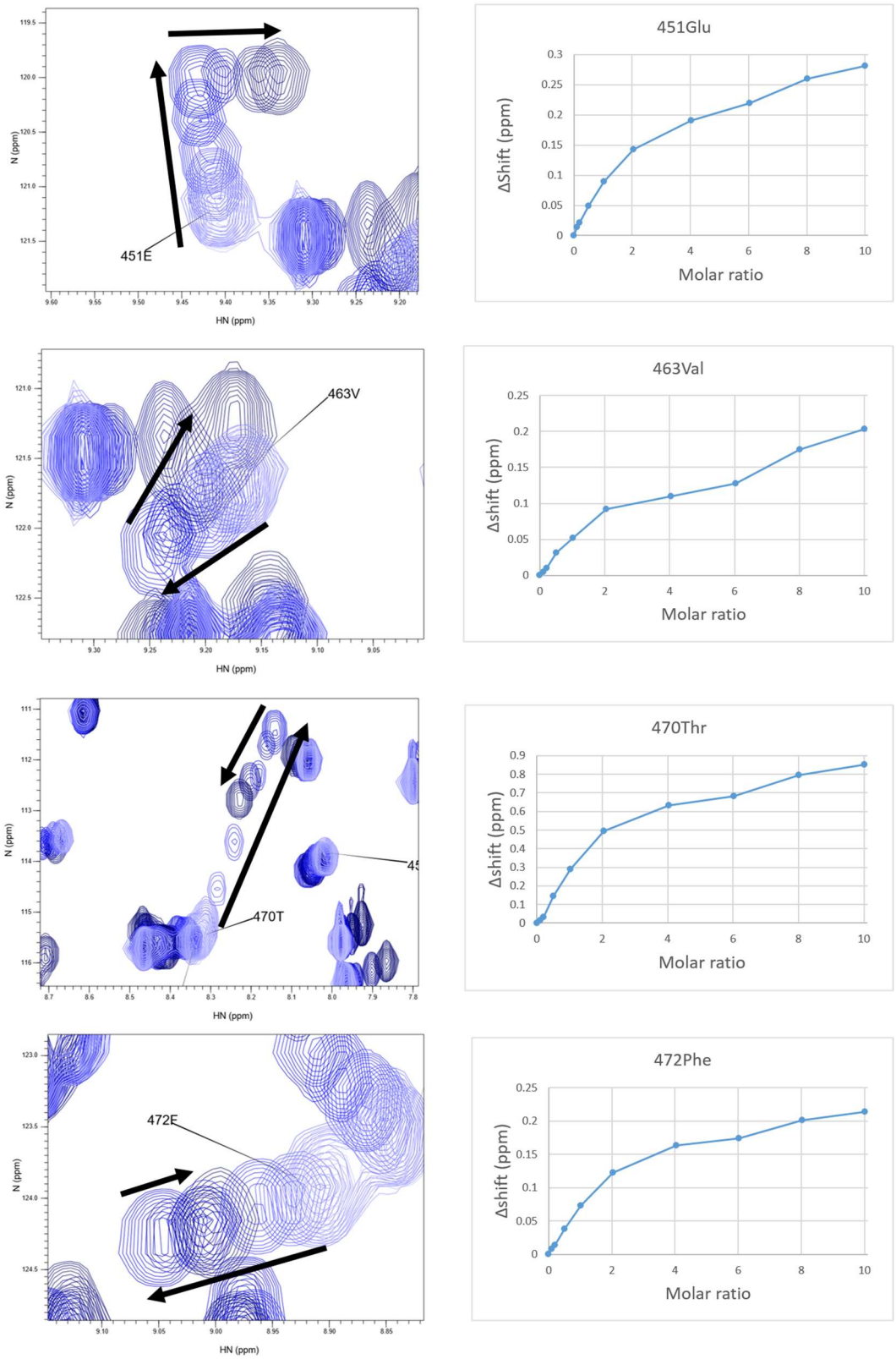


Figure 5.2.3 - A selection of GB1:WW4/Smad7 titration resonance migration patterns in detail, showing the appropriate region in the HSQC and their change in shift plotted against molar ratio. The black arrows show the direction of migration of the HSQC peaks. It is evident from the migration patterns and change in shifts that a second phase of migration occurs at a molar ratio of roughly 1:6.

Using the CCPN Analysis software, the change in shifts and saturation point for each residue were fit to the following equation, which is appropriate for a fast exchange interaction:

$$y = A\left(B + x - \sqrt{(B+x)^2 - 4x}\right)$$

$A = \Delta\delta_\infty/2$

$B = 1 + K_d/a$

$x = b/a$

$y = \Delta\delta_{obs}$

$a =$ total protein concentration

$b =$ total ligand concentration

$\Delta\delta_{obs} =$ change in chemical shift

$\Delta\delta_\infty =$ difference between start chemical shift and chemical shift at saturation

Figure 5.2.4 shows the CCPN Analysis software K_d fit for the same four residues of the binding site, which were shown in Figure 5.2.3. The K_d values for each of these are much higher than expected.

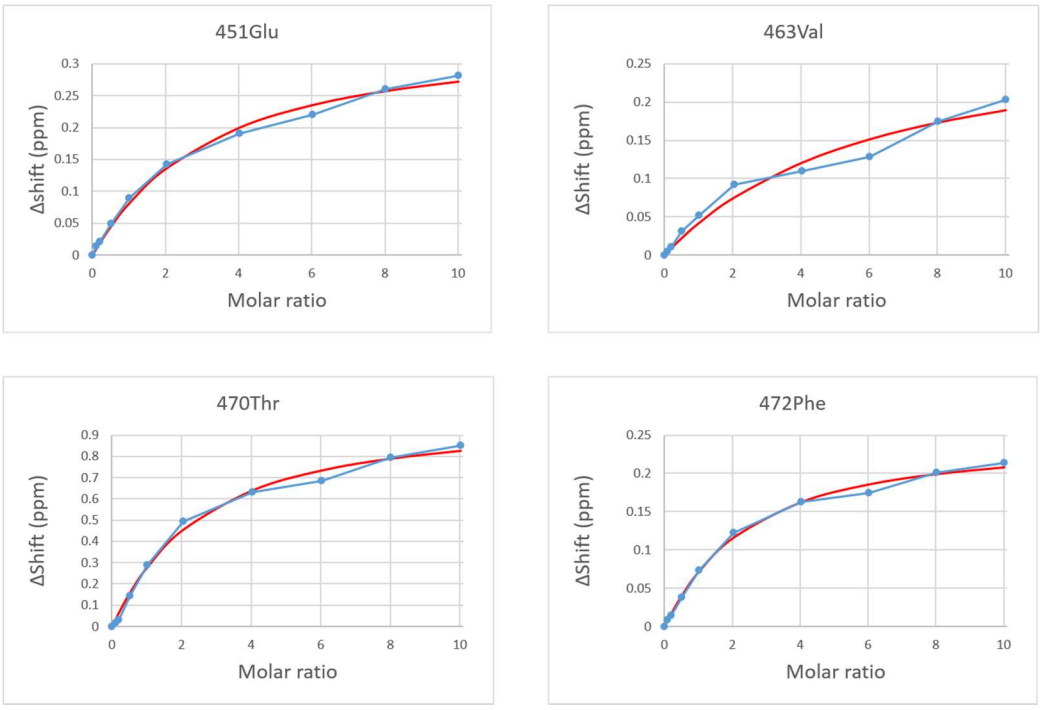


Figure 5.2.4 - The CCPN Analysis software K_d fit, shown in red, for a selection of residues from the GB1:WW4 Smad7 binding site. The dissociation constants for these residues are: 451Glu 1.98 mM, 463Val 4.08 mM, 470Thr 1.47 mM and 472Phe 1.39 mM.

None of the K_d fits acceptably satisfy the lineshapes of the ΔShift plots. From the graphs alone it might be easy to imagine an error occurred during the preparation of the sample at a molar ratio of 1:6. However, coupled with the observations made from the HSQC, it is more plausible that two separate fast-exchange binding events are producing the changes in shift being observed. Peak migration would, therefore, indicate three states: unbound, bound 1 and bound 2. Bound 1 would have a lower K_d and therefore higher affinity, reflecting the first stage of contribution to the migrating shift, while bound 2 would have a lower K_d, making the second stage contribution after bound 1 appears close to saturation. In order to obtain an accurate K_d for these binding events, a curve was fit to both stages. Initially, the equation was fit to the first stage of binding using CCPN Analysis, but it was felt that the prediction of the point of saturation was consistently slightly off, and subsequently either predicting a slightly tighter K_d or looser K_d. Therefore, the above equation was fit manually to the two curves apparent in the ΔShift plots, by optimising an RMSD error value for the closeness of fit. These are shown in Figure 5.2.5.

Two curves were fit for each residue of the binding site, each with a different K_d. The first curve was fit by optimising the RMSD value against the first 8 titration points up to a molar ratio of 1:6, while the second curve was fit by optimising the RMSD value of the curve fit to the last 3 titration points from where it intercepts the first phase, at a molar ratio of 1:6. The dissociation constants are listed in Table 5.2.1 alongside the K_d calculated by CCPN Analysis.

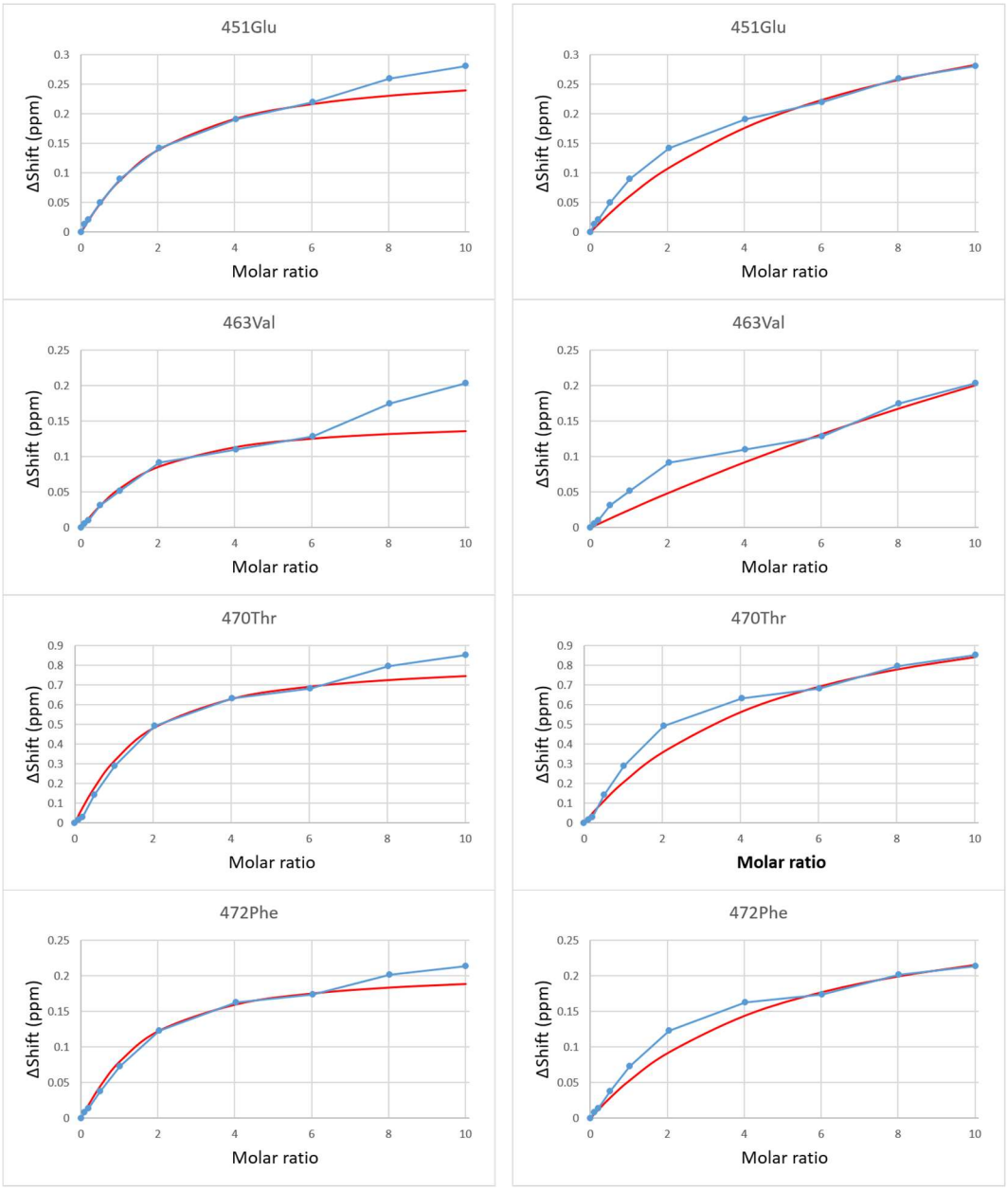


Figure 5.2.5 - The manual K_d fit in red for the same selection of residue shifts in the GB1:WW4/Smad7 titration. In each example the graph on the left is the fit for the first stage of migration and the graph on the right is the fit for the second stage of migration.

Residue	Analysis K_d (mM)[CCPN]	Bound1 K_d (mM)	Bound2 K_d (mM)
450Trp	*198.56 ±941.02*	*1.0*	*26.0*
451Glu	1.98 ±0.29	1.2	4.5
452Met	1.19 ±0.19	0.68	3.5
461Tyr	1.18 ±0.15	0.7	2.9
462Phe	1.13 ±0.14	0.85	2.2
463Val	4.08 ±1.56	0.9	27.5
464Asp	1.64 ±0.18	1.15	3.7
465His	*11.1 ±2.82*	*2.6*	*14.0*
467Thr	1.53 ±0.3	0.77	6.0
468Arg	0.931 ±0.12	0.56	1.4
469Thr	1.27 ±0.18	0.6	3.4
470Thr	1.47 ±0.25	0.82	3.2
471Thr	1.13 ±0.15	0.66	2.5
472Phe	1.39 ±0.17	0.8	3.2
473Lys	1.35 ±0.19	1.0	4.0
Average K_d	**1.56 ±0.8**	**0.82 ±0.2**	**5.23 ±6.78**

Table 5.2.1 Binding site K_d values for the GB1:WW4 Smad7 interaction as determined by CCPN Analysis, and a manual fit for bound1 and bound2 curves. Average K_d values were calculated excluding 450Trp and 465His. Individual Kd errors indicate the fit error. Standard deviation is given as the error for K_d averages.

Two residues had poor curve fits when trying to manually fit bound1. Those were 450 tryptophan and 465 histidine, which also had poor fits by CCPN Analysis. This was mainly because it seemed as though the bound2 migration was more apparent early on in the titration. Because of a level of uncertainty around how to fit these curves, and the poor fits made by Analysis, they were omitted from the average K_d. Using this approach, the average K_d of the main binding event 'bound1' is 0.82 mM across the binding site, as defined by the residues which undergo a significant perturbation of the local magnetic field consistent with the docking of a ligand. The second binding event 'bound2' appears to have a much lower affinity with a K_d at 5.23 mM. The Analysis K_d values appear to be the result of an average fit of two curves and can therefore be largely disregarded. The cause of the two stage binding curve is somewhat unclear, however it seems likely that the first stage binding is the result of the main WW4/Smad7 interaction, and the second

stage might be the result of a low affinity non-specific interaction. The possible causes will be explored later in this chapter.

Whilst the bound1 affinity of 0.82 mM is within the affinity range of a WW domain interaction, when compared to dissociation constants from other NEDD4 family WW domain Smad7 affinity experiments, which are typically in the low micromolar range, the interaction seems to be much weaker. Although other NEDD4/Smad7 interaction experiments are typically performed at a lower temperature (15°C compared to 25°C here) and using ITC. In theory K_d values calculated from these two different techniques should be comparable. It is possible that the second non-specific binding event affects the migration of the peaks even during the early phase of the titration, and that the two K_d values cannot be so easily deconvoluted. Some attempts were made to perform fluorescence based affinity assays using a fluorescein isothiocyanate (FITC) labelled peptide so as to observe fluorescence anisotropy during binding. However, no anisotropy was observed, possibly due to the FITC tag interfering with the interaction.

5.2.2 WW4 and monophosphorylated Smad7

The ligand used for these experiments was a synthetic Smad7 peptide that corresponds to the exact same region used in Section 5.2.1, but with a phosphorylated serine at position 206 as below:

203 - ELEpSPPPPYSRYPMD **- 217**

The phospho-peptide was titrated into a 0.78 mM ^{15}N labelled sample of GB1:WW4, and after each titration point a ^{1}H-^{15}N-HSQC was acquired. To ensure the results of the phosphorylated and non-phosphorylated ligand titrations were as comparable as possible, the GB1:WW4 sample was from exactly the same batch. A total of 10 titration points were collected from a molar ratio of 1:0 protein to peptide, to a molar ratio of 1:10, where the concentration of Smad7 was 10-fold higher than GB1:WW4. Figure 5.2.6 shows the migration of resonances that were observed during the titration.

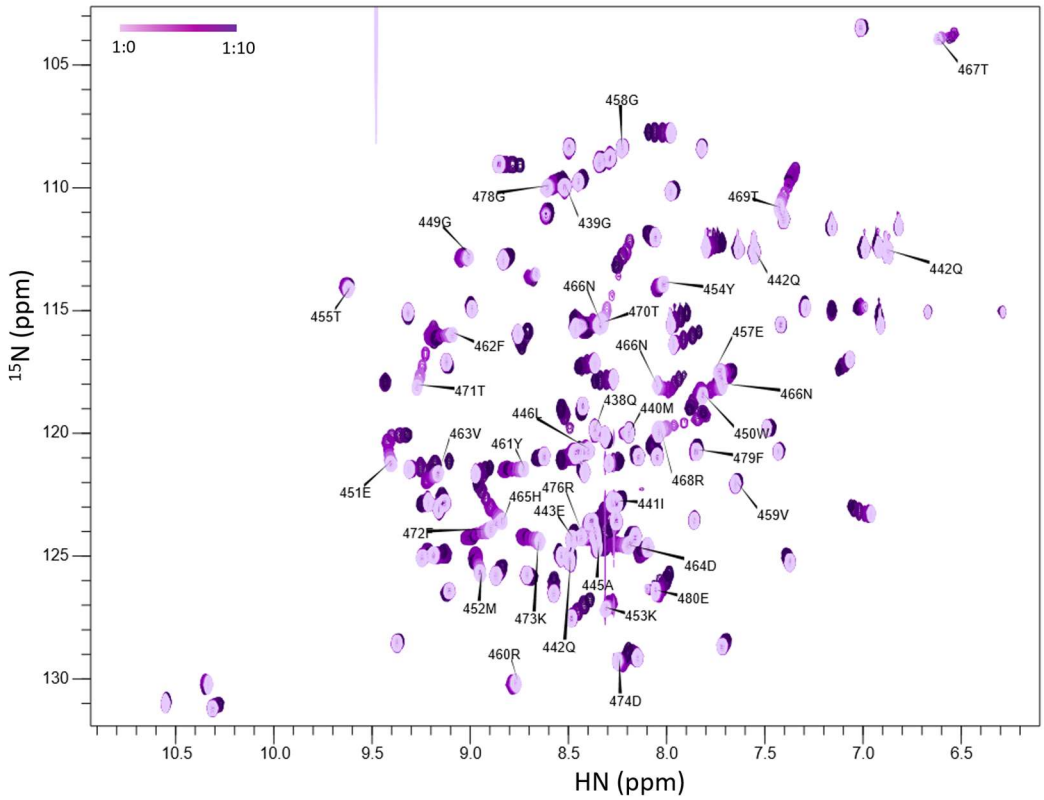


Figure 5.2.6 - An overlay of the 1H-15N-HSQC spectra from the GB1:WW4 phosphoSmad7 titration with only the WW4 domain peaks (438-480) labelled, for clarity. Light purple peaks are from the first point along the titration at a molar ratio of 1:0 protein to ligand, and the darkest purple peaks are from the final titration point at 1:10. The experiment was performed at 500 MHz, 298 K. The sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. The concentration of GB1:WW4 was 0.78 mM.

As with the non-phosphorylated peptide, peak migration consistent with a fast-exchange interaction is evident. The trajectories were plotted against residue number and these are shown in Figure 5.2.7. The general pattern of binding is preserved, although there are some small variations. The same residues are seemingly involved in the binding site. When compared to the Smad7 titration, the trajectories were somewhat smaller, particularly in the binding site. This might have been related to the differences in the HSQC, where migration of the peak along the direction of the second binding event seemed to occur at an earlier point in the titration. The peaks retained the second phase of peak migration, the 'bound2' phase (Figure 5.2.8), but they seemed to start migrating along the second stage at a molar ratio of 1:4 instead of 1:6. This is best shown in the ΔShift plots in Figure 5.2.8. From these plots it seems as though the bound1 stage of migration occurred at a faster rate and seems to begin to saturate sooner, before the

bound2 phase takes over, giving the inclination that the phosphorylated Smad7 ligand bound with a tighter K_d.

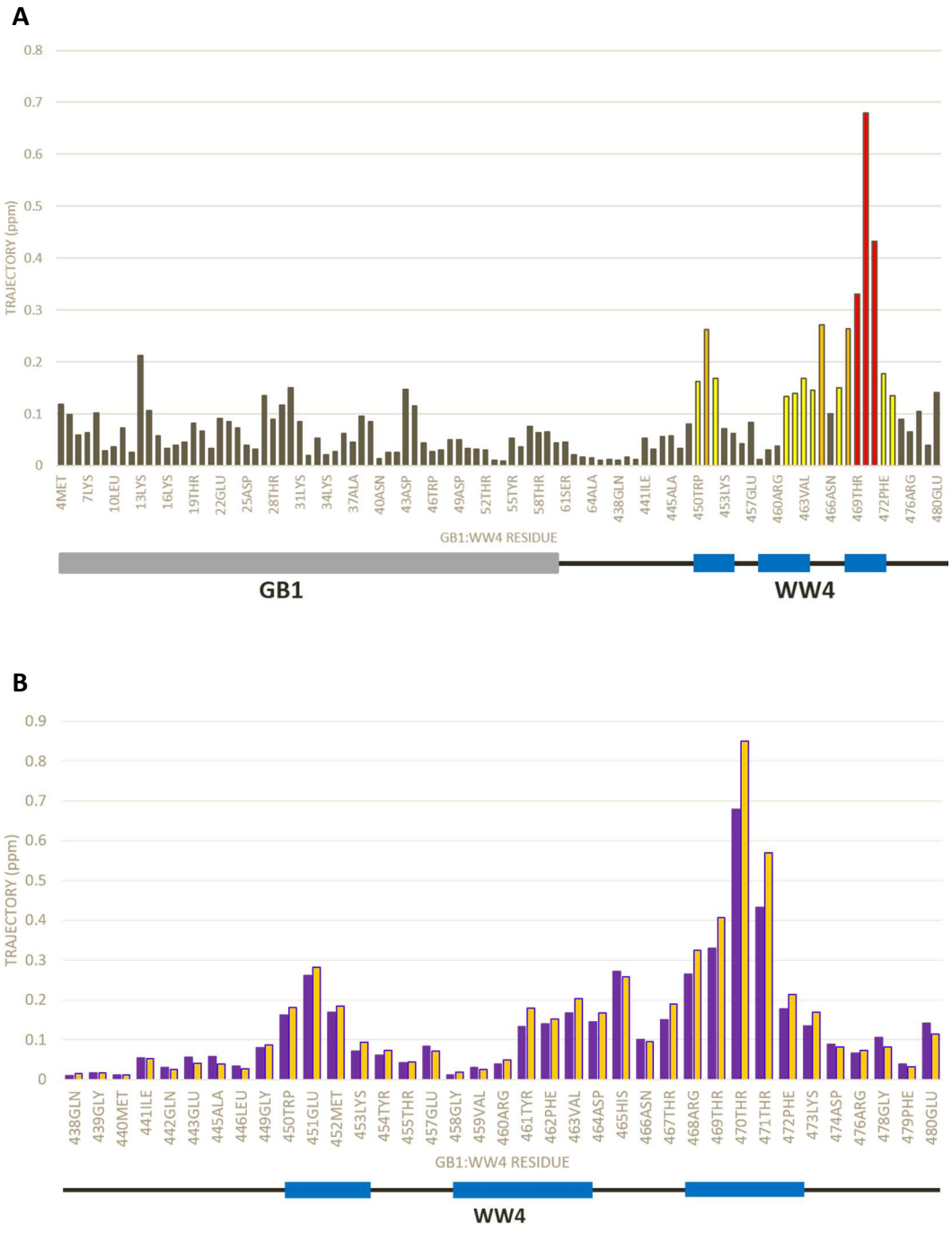


Figure 5.2.7 - A: The trajectories of each residue in the 10 point GB1:WW4 pSmad7 titration, with the same residues from the non-phosphorylated ligand titration highlighted in red, orange and yellow, indicating the extent to which the chemical shift migrates. B: A comparison of the trajectories between the Smad7 titration, in orange, and the pSmad7 titration, in purple. Only the WW4 residues are shown.

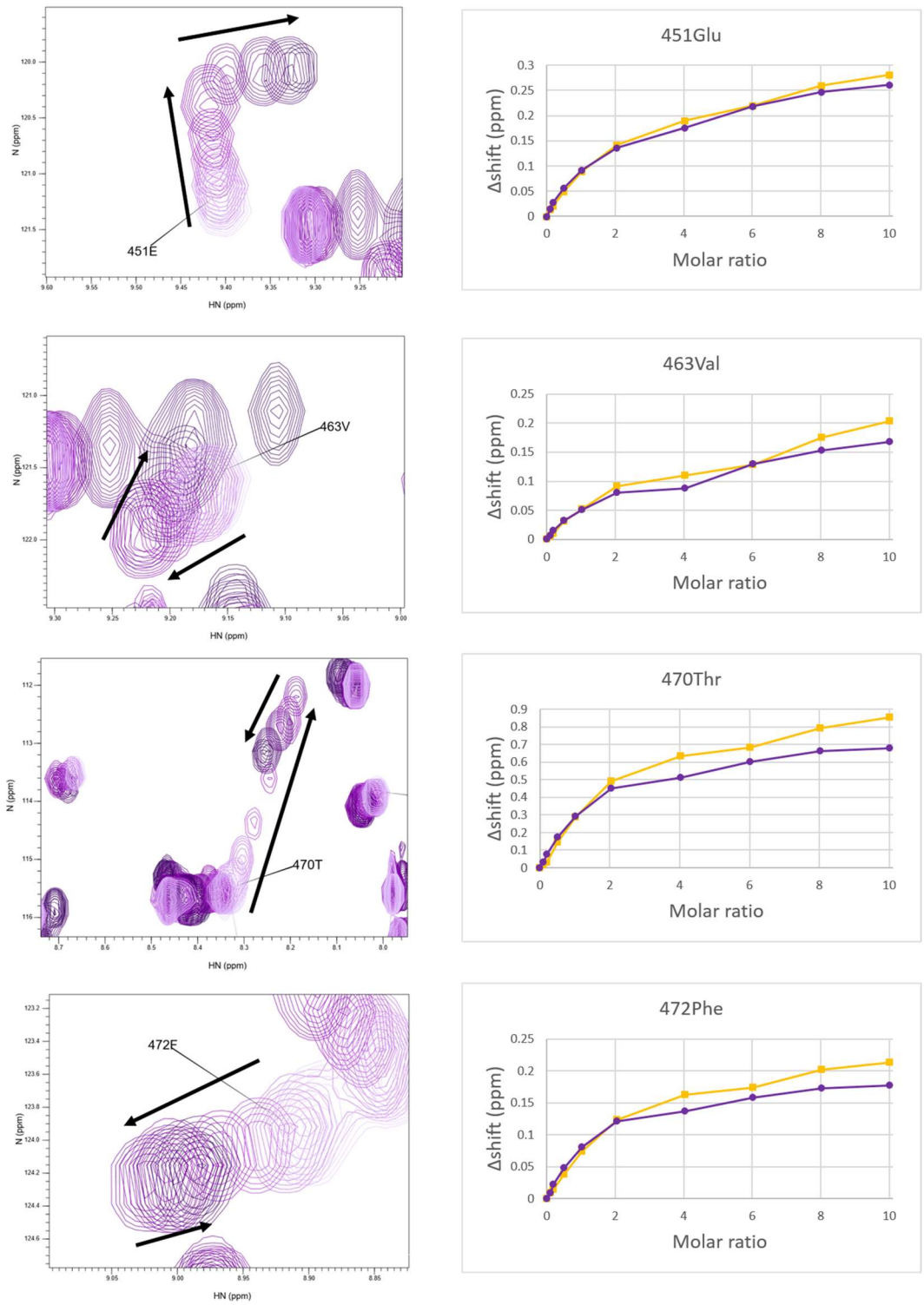


Figure 5.2.8 - The peak migration of 451Glu, 463Val, 470Thr and 472Phe from the GB1:WW4/pSmad7 titration ^{1}H-^{15}N-HSQCs. The black arrows show the direction of migration of the HSQC peaks. The ΔShift plots on the right show the pSmad7 shift change in purple and the non-phosphorylated Smad7 shift change in orange.

As with the Smad7 titration, the CCPN Analysis K_d fit for the curves was poor, so a K_d was manually fit to the two binding events for each of the residues of the binding site, using the same equation, but optimising the RMSD error of the bound1 curve against the first 7 titration points instead of the first 8 titration points. The bound2 curve RMSD error was optimised against the last 4 titration points, instead of the last 3. A few examples of these can be seen in Figure 5.2.9.

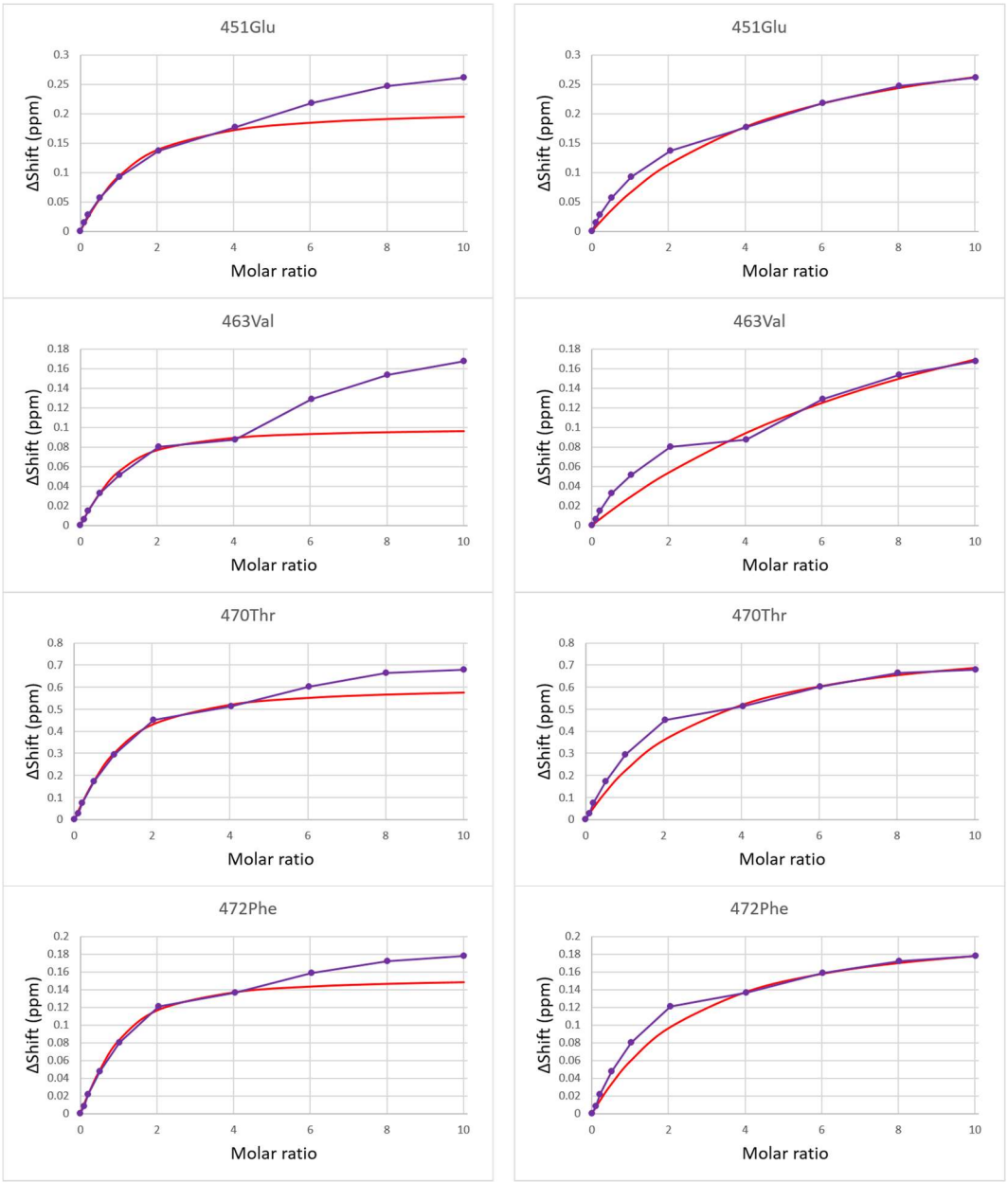


Figure 5.2.9 - The manual K_d fit shown in red for the same four residues of the GB1:WW4/pSmad7 titration. For each example the bound1 curve fit is on the left and the bound2 curve fit is on the right.

The dissociation constants are shown in Table 5.2.2.

Residue	Analysis K_d (mM)^{CCPN}	Bound1 K_d (mM)	Bound2 K_d (mM)
450Trp	*40.33 ±46.49*	*0.8*	*10*
451Glu	1.62 ±0.28	0.55	3
452Met	0.92 ±0.15	0.46	1.85
461Tyr	0.65 ±0.07	0.39	1.15
462Phe	0.88 ±0.13	0.43	1.5
463Val	2.90 ±1.00	0.3	8
464Asp	0.97 ±0.08	0.69	1.37
465His	*4.71 ±0.73*	*1.7*	*4.8*
467Thr	1.26 ±0.38	0.3	4.4
468Arg	0.63 ±0.07	0.62	0.62
469Thr	0.74 ±0.11	0.36	1.5
470Thr	0.80 ±0.13	0.43	1.7
471Thr	0.53 ±0.05	0.44	0.75
472Phe	0.68 ±0.10	0.33	1.5
473Lys	0.75 ±0.11	0.4	1.5
Average K_d	**1.03 ±0.63**	**0.44 ±0.12**	**2.64 ±2.0**

Table 5.2.2 Binding site K_d values for the GB1:WW4 phospho-Smad7 interaction as determined by CCPN Analysis, and a manual fit for bound1 and bound2 curves. Average K_d values were calculated excluding 450Trp and 465His. Individual K_d errors indicate the fit error. Standard deviation is given as the error for K_d averages.

As with the Smad7 titration, 450 tryptophan and 465 histidine had poor fits and were excluded from the average K_d. Only one curve could be fit to 468 arginine, so the same K_d is given for both binding events. While the affinity of the interaction appears to still be looser than other NEDD4 family Smad7 interactions, the average K_d of both binding events was tighter than that of the unphosphorylated Smad7 interaction, for almost every residue. The average dissociation constant of the main binding event 'bound1' was nearly halved when titrating the phosphorylated ligand, giving the indication that phosphorylation near the PPxY motif of Smad7 might have a significant impact on this interaction *in vivo*, and that of other WW domains.

5.2.3 WW4 and Smad2/3

In immunoprecipitation assays, WWP2-C, which only contains the WW4 domain and the HECT domain, did not appear to interact with Smad2 or Smad3. It was therefore expected that the WW4 domain would show no interaction with the PPxY motif of these proteins in ligand titrations, performed in the same manner as above. The PPxY motifs of Smad2 and Smad3 are very similar, the only difference being the first residue C-terminal to the PPxY motif, which is an isoleucine in Smad2 compared to a leucine in Smad3. Their sequences are shown below, along with the region of Smad7 used in Sections 5.2.1 and 5.2.2.

Smad7: 203 - ELESPPPPYSRYPMD **- 217**

Smad2: 217 - IPETPPPGYISEDGE **- 231**

Smad3: 176 - IPETPPPGYLSEDGE **- 190**

Synthetic peptides corresponding to these regions were titrated in to ^{15}N labelled GB1:WW4 samples. After each titration point ^{1}H-^{15}N-HSQC spectra were acquired as before, and these are shown in Figure 5.2.10A and B. Unexpectedly, both Smad2 and Smad3 ligands seemed to interact with the WW4 protein and peak migration is evident in both titrations. The trajectories were plot against residue number (Figure 5.2.11). The Smad3 titration exhibited the same non-specific secondary peak migration but, significantly, the peaks in the Smad2 titration did not. This can be seen in Figure 5.2.10 but is shown more clearly in Figure 5.2.12 with a comparison between peak migrations of four residues in both spectra. This pattern is consistent with almost all residues in the Smad2 titration, although a small change in direction is apparent with glutamic acid 451 (Figure 5.2.12).

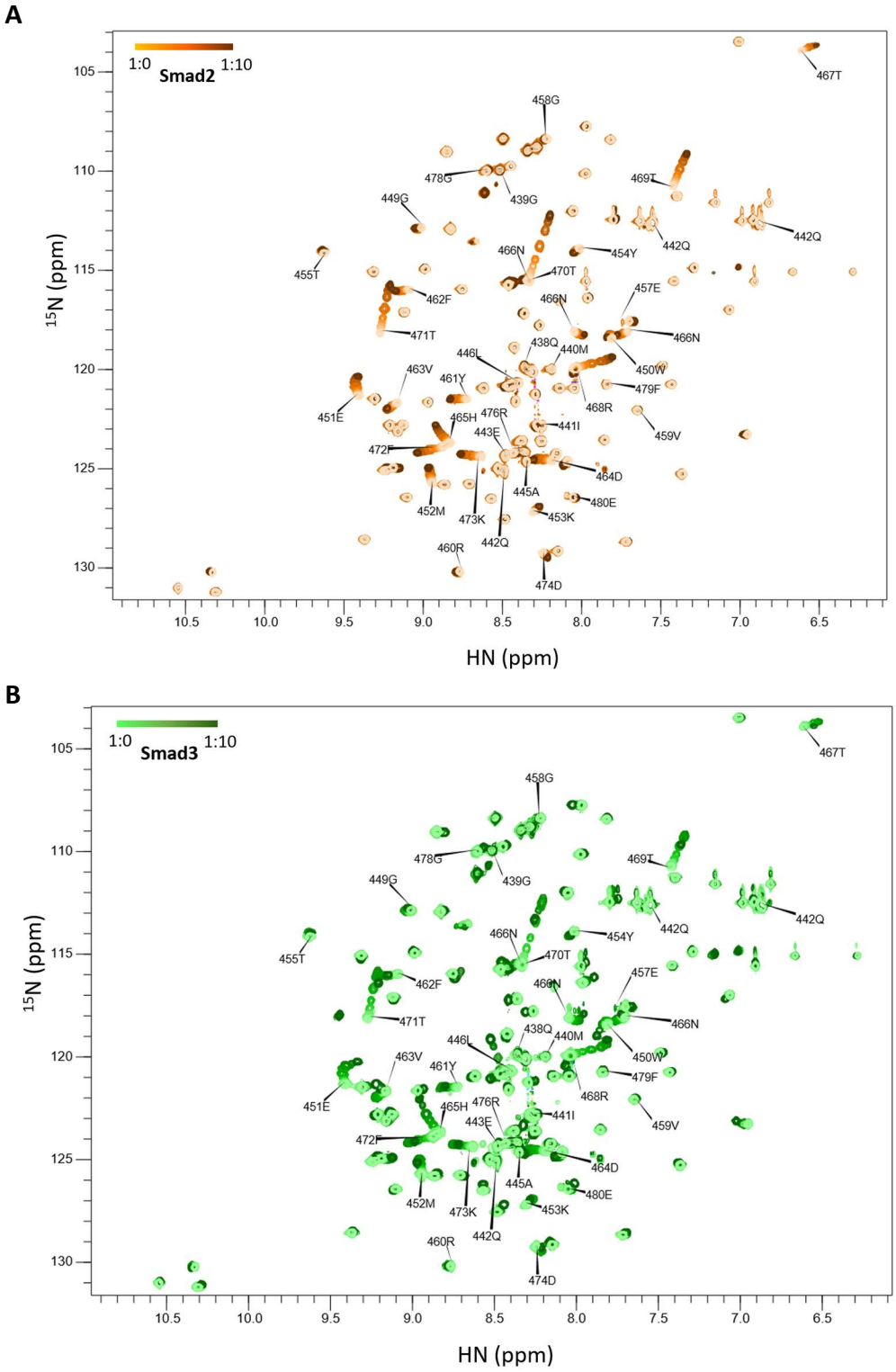


Figure 5.2.10 - A: The GB1:WW4 Smad2 titration 1H-15N-HSQC overlay; the starting titration point is in light orange and the end titration point is in dark orange. GB1:WW4 is at 0.38 mM B: The GB1:WW4 Smad3 titration HSQC overlay; the starting titration point is in light green and the end titration point is in dark green. GB1:WW4 is at 0.33 mM. The experiments were performed at 500 MHz, 298 K. The samples were prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8

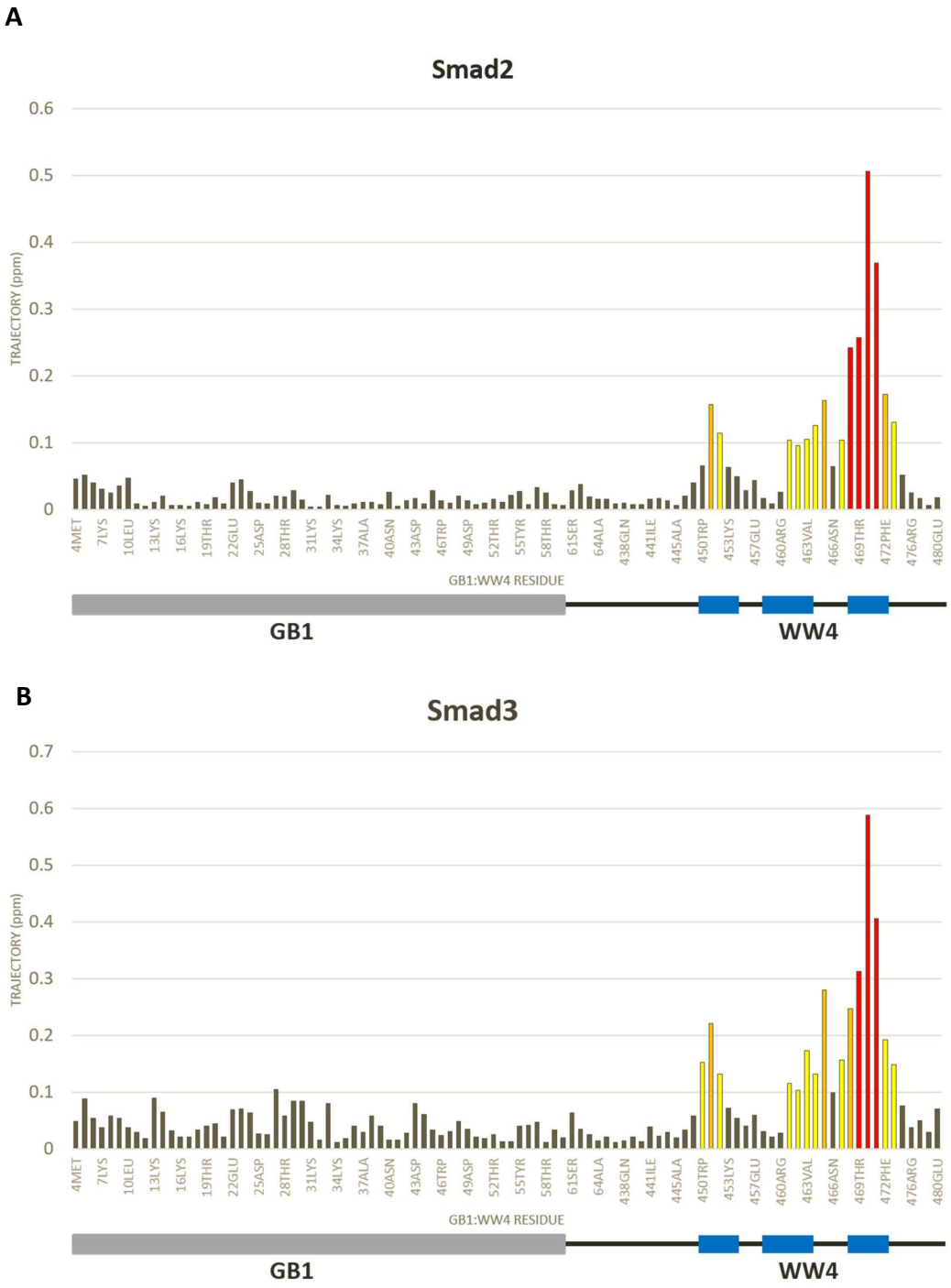


Figure 5.2.11 - A: The trajectory (in ppm) of each GB1:WW4 backbone amide assigned in the 1H-15N-HSQC (there is no information for prolines), upon titration of the Smad2 ligand. Residues are colour coded to indicate extent of peak migration. B: The trajectory (in ppm) of each GB1:WW4 backbone amide assigned in the ^{1}H-^{15}N-HSQC (there is no information for prolines), upon titration of the Smad3 ligand. Residues are colour coded to indicate extent of peak migration. A schematic is aligned with the graphs showing the GB1 domain in grey and the three strands of the WW4 domain β-sheet in blue.

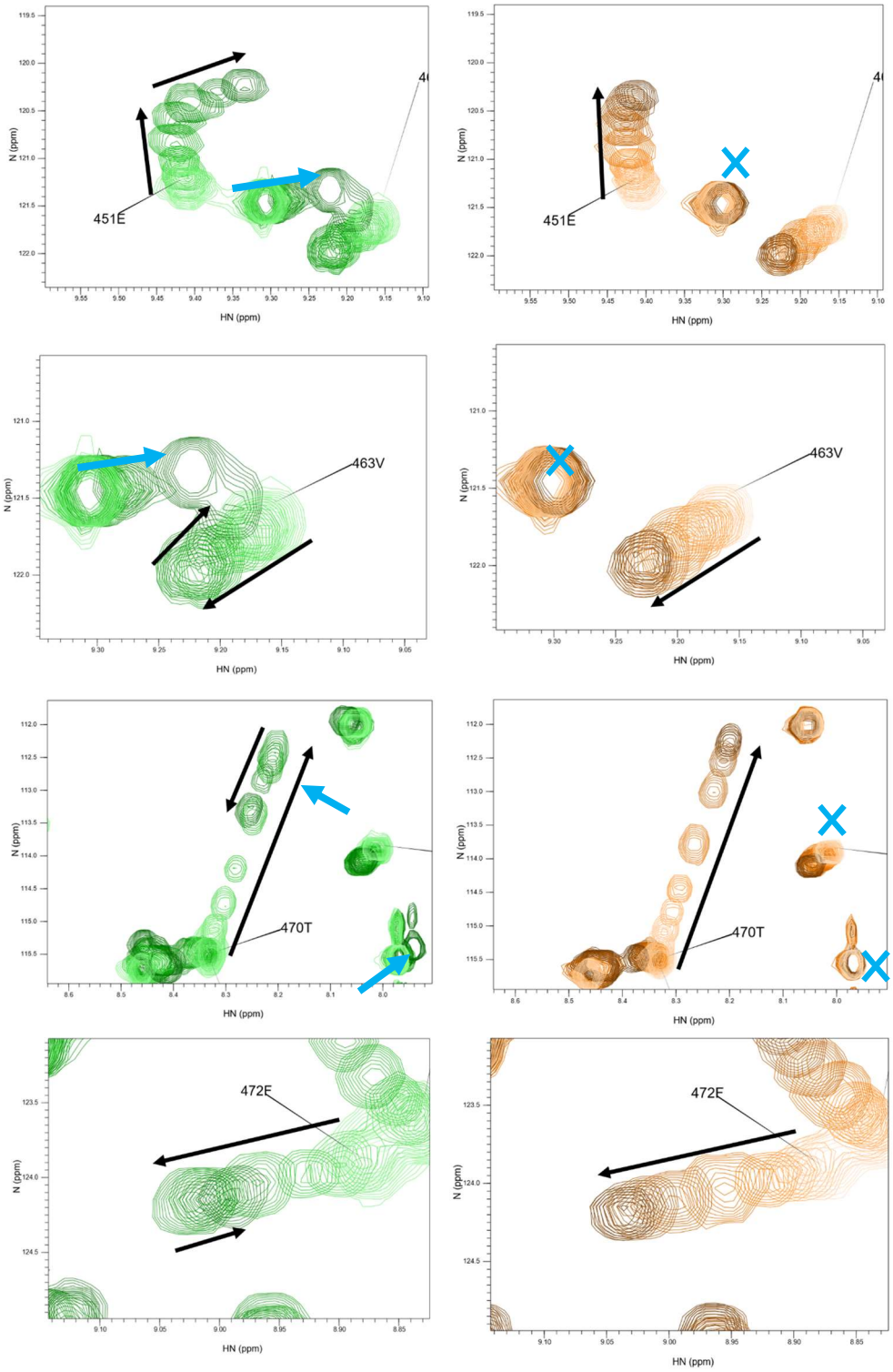


Figure 5.2.12 - Peak migration of four GB1:WW4 residue resonances in the Smad3 titration 1H-15N-HSQCs in green, and the Smad2 titration 1H-15N-HSQCs in orange, shown side by side. The black arrows show the direction of migration of the binding site HSQC peaks, the blue arrows show the direction of migration of the other peaks at high Smad3 ligand concentrations, the blue crosses indicate the same peaks that do not move in the Smad2 titration.

The trajectories in the Smad2 titration were lower than those in the Smad3 titration. This might be because the trajectories also include the secondary migration, when the peak movement increases again. The general pattern of significant peak migration changed slightly and a few of the peaks transcend the arbitrary red, orange, yellow levels, which is ranked relative to the other peak trajectories and is somewhat subjective. The 450 tryptophan peak, which had large errors in the other titration K_d calculations, and has been disregarded in the final K_d, migrated quite insignificantly in the absence of the secondary migration. However, for the purpose of comparison, 450 tryptophan has still been included in the rest of this section. The GB1 region also moved less in the Smad2 titration and it seems that this is also related to the secondary event, as in the Smad7, phospho-Smad7 and Smad3 titrations, the point at which the secondary peak migration occurred, coincided with a movement of all residue peaks. In Figure 5.2.12, examples of GB1 peaks moving at high concentrations of Smad3 ligand have been highlighted with blue arrows, whereas the same peaks remain stationary in the Smad2 titration, and have been highlighted with blue crosses (note, the peaks highlighted in 451E and 463V are the same). The result of this is that the Smad2 titration ΔShift plots, whilst not perfect, generally exhibited only one phase of binding, whereas the Smad3 titration ΔShift plots showed two. This can be seen in Figure 5.2.13 and Figure 5.2.14.

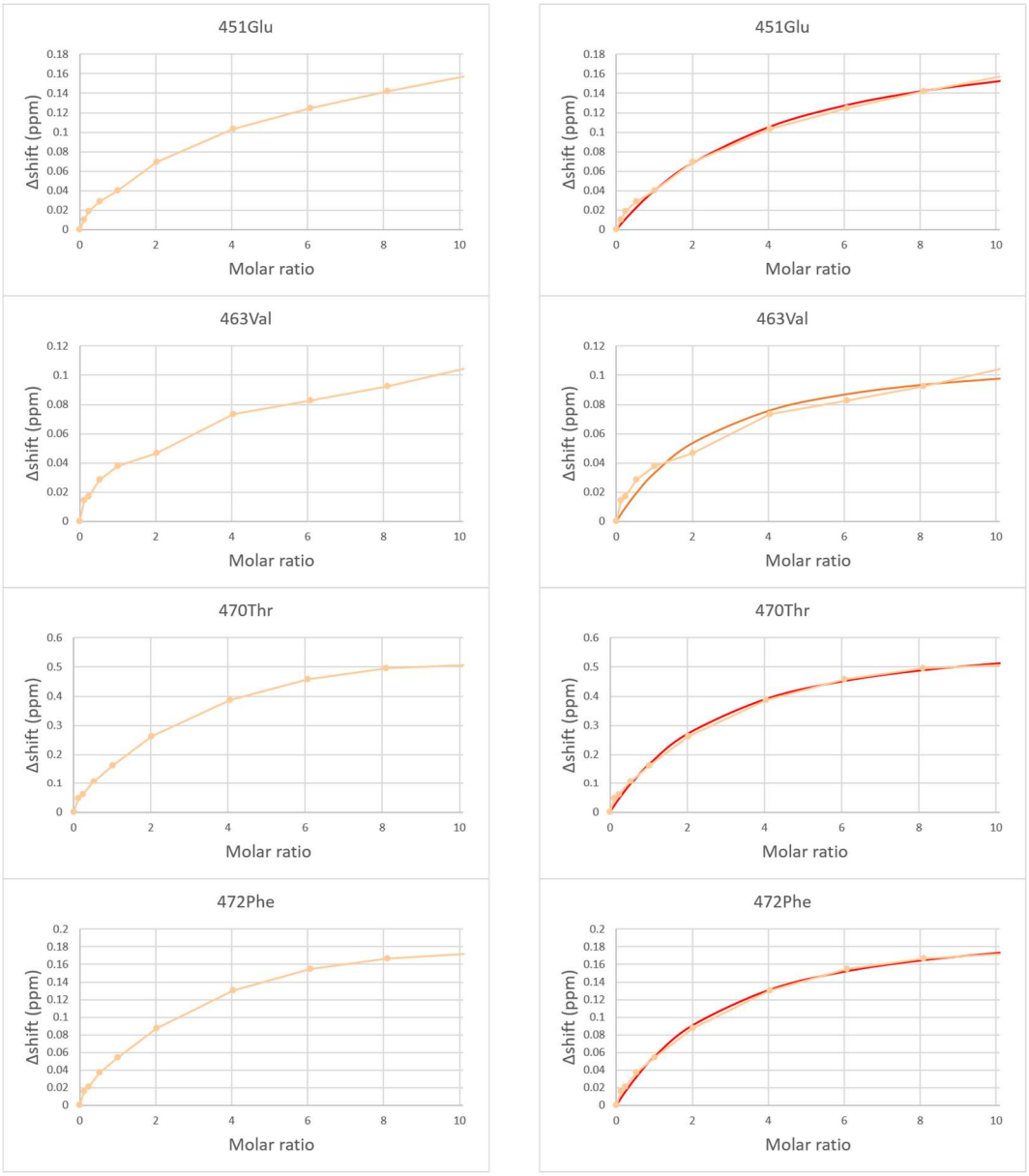


Figure 5.2.13 - The ΔShift plots for four residue amide resonances of the GB1:WW4/Smad2 titration on the left, and the CCPN Analysis K_d fits in red on the right.

The Smad2 titration K_d fits by Analysis were generally good across the binding site. While there were some points that are slightly outside of the trend for some of these residues, the fact that there was only one migration direction in the HSQC, and the fact that more data points could be incorporated, adds confidence to the K_d prediction for these residues. The K_d values for each of these residues are shown in Table 5.2.3.

Residue	Analysis K_d (mM)[CCPN]
450Trp	*0.69 ±0.33*
451Glu	1.31 ±0.19
452Met	0.86 ±0.08
461Tyr	0.85 ±0.09
462Phe	0.91 ±0.18
463Val	0.70 ±0.23
464Asp	1.53 ±0.23
465His	*3.93 ±0.82*
467Thr	0.79 ±0.13
468Arg	1.63 ±0.06
469Thr	0.79 ±0.10
470Thr	0.79 ±0.09
471Thr	0.86 ±0.09
472Phe	0.81 ±0.09
473Lys	0.81 ±0.08
Average K_d	**0.973 ±0.31**

Table 5.2.3 Binding site K_d values for the GB1:WW4 Smad2 interaction as determined by CCPN Analysis. The average K_d was calculated excluding 450Trp and 465His. Individual K_d errors indicate the fit error. Standard deviation is given as the error for the K_d average.

Tryptophan 450 was excluded from the average K_d as the peak only had a small trajectory and the curve fit poorly. Histidine 465 was excluded from the average K_d as the peak migration did not reach close to saturation, making the K_d prediction unreliable.

The ΔShift plots for the Smad3 titration are shown in Figure 5.2.14. Because of the two-stage migration for this titration, curves were fit manually for the residues of the binding site using the same approach as with Smad7, optimising an RMSD value for the first 8 points for the bound1 curve and an RMSD value for the last 3 points for the bound2 curve. The dissociation constants are listed in Table 5.2.4.

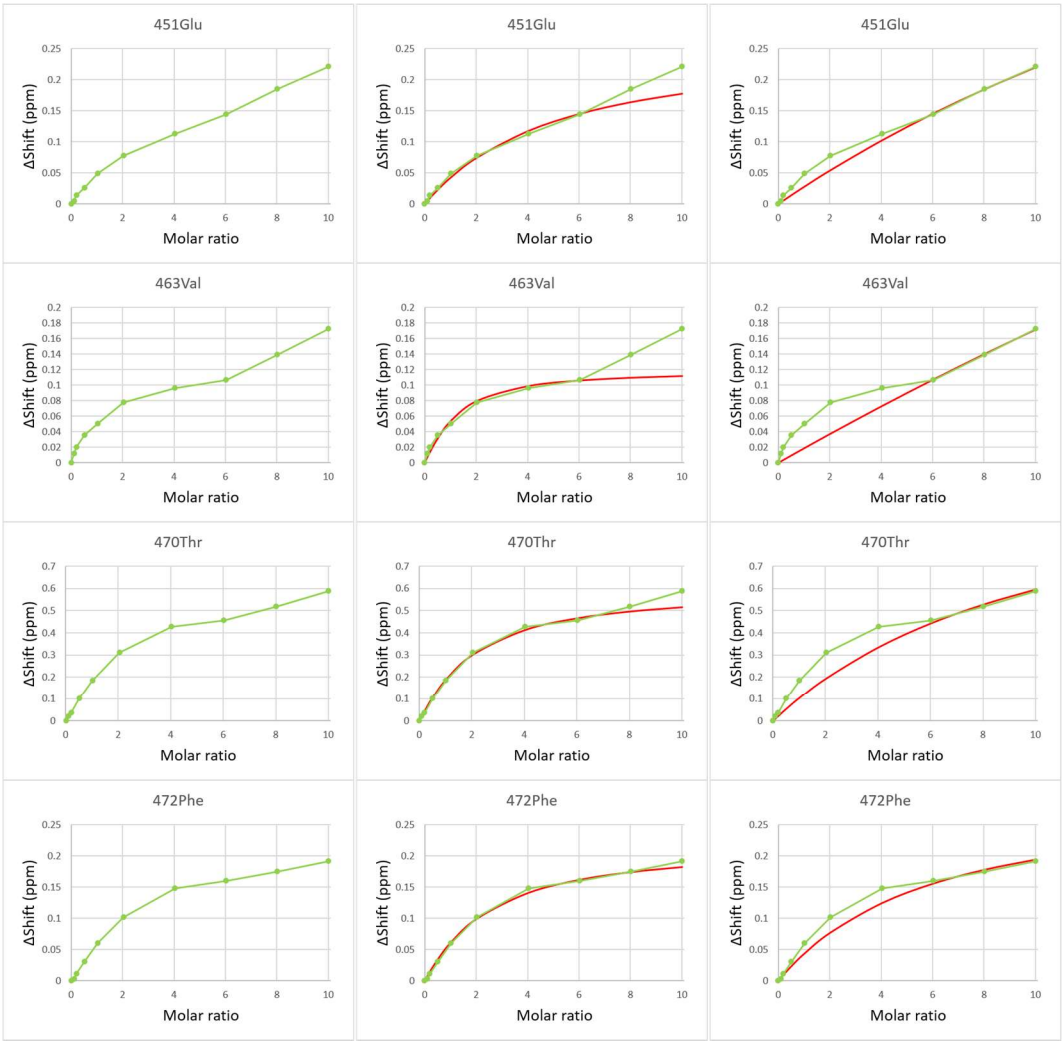


Figure 5.2.14 - The ΔShift plots for four of the residue amide resonances in the GB1:WW4/Smad3 titration. The bound1 migration K_d fit is the centre curve for each residue and the bound2 migration K_d fit is the curve on the right for each of the residues.

Residue	Analysis K_d (mM)[CCPN]	Bound1 K_d (mM)	Bound2 K_d (mM)
450Trp	*213.86 ±4410.00*	*1.00*	*220.0*
451Glu	3.27 ±0.91	1.55	10.70
452Met	0.67 ±0.06	0.55	1.30
461Tyr	0.61 ±0.06	0.45	1.20
462Phe	0.49 ±0.13	0.35	4.00
463Val	1.09 ±0.47	0.25	32.0
464Asp	1.1 ±0.1	1.10	1.10
465His	*31.89 ±23.96*	*3.40*	*85.0*
467Thr	1.18 ±0.31	0.41	6.40
468Arg	1.32 ±0.23	1.30	1.30
469Thr	0.51 ±0.08	0.30	2.90
470Thr	0.73 ±0.12	0.50	3.20
471Thr	0.56 ±0.06	0.42	1.50
472Phe	0.72 ±0.09	0.60	1.70
473Lys	0.68 ±0.1	0.62	1.70
Average K_d	**0.994 ±0.74**	**0.65 ±0.41**	**5.31 ±8.48**

Table 5.2.4 Binding site K_d values for the GB1:WW4 Smad3 interaction as determined by CCPN Analysis, and a manual fit for bound1 and bound2 curves. Average K_d values were calculated excluding 450Trp and 465His. Individual K_d errors indicate the fit error. Standard deviation is given as the error for K_d averages.

Tryptophan 450 and histidine 465 were excluded from the final K_d again because of poor fits. Only one curve could be fit to 464 aspartate and 468 arginine and the fits were similar to the Analysis output, which confirms that the optimisation approach used is valid. The K_d is tighter than the Smad2 dissociation constant; when the ΔShift graphs are compared, the peaks in the Smad3 titration had a faster rate of migration and appeared to saturate sooner. This can be seen in Figure 5.2.15.

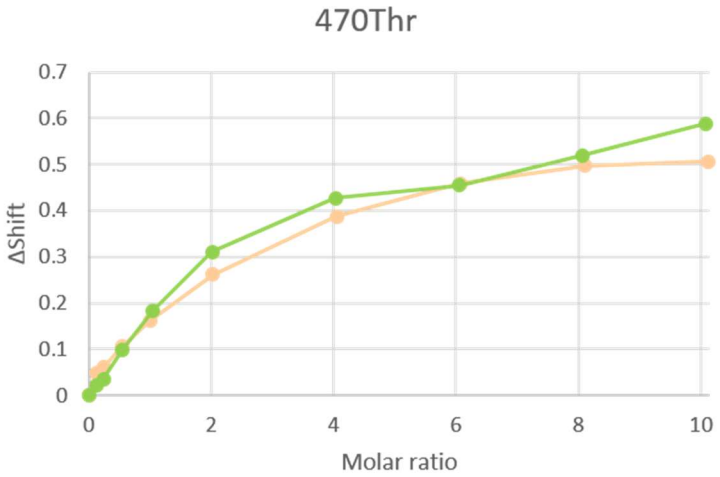


Figure 5.2.15 - Change in shift of the GB1:WW4 470 threonine amide peak upon Smad2 titration (orange) and Smad3 titration (green).

This is quite surprising considering the similarities between the two peptides. When compared to the Smad7 dissociation constant, Smad2 appears to bind with the lowest affinity out of all of them, whereas Smad2 binds with a higher affinity than both Smad3 and Smad7. Both Smad2 and Smad3 appear to have a lower affinity for WW4 than the phosphorylated Smad7 ligand. The preference of Smad3 over Smad2 would seem to correlate with functional assays that observed a small amount of turnover of Smad3, but not Smad2, by WWP2-C (Soond & Chantry 2011).

One of the caveats when discussing the affinities described here is that there is a level of uncertainty as to firstly, the cause of the secondary peak migration and secondly, whether the secondary migration is affecting the observed K_d during the primary stage of migration. This is particularly pertinent when comparing titrations that do have the secondary event and those that do not, such as the Smad2 and Smad3 titrations. While it might be more realistic to make comparisons between titrations that do have the secondary migration, the uncertainty around the influence of the secondary migration over the primary migration means that the accuracy of the dissociation constant is in doubt. The natural question to pose at this point is, what exactly is causing the secondary migration, where the peaks either travel in a completely different trajectory, or double back on themselves. Initially thoughts were towards perhaps a ligand-binding-induced dimerisation, or perhaps aggregation at high ligand concentration, which cannot be entirely ruled out. WW domain dimers and higher orders of oligomerisation have been observed by NMR with SMURF WW domains that form a 'β-clam' conformation, mediated by interaction of the hydrophobic underside surface (Aragón et al. 2012). Almost all of the

peaks, including those from the GB1 domain, start to migrate at higher concentrations of ligand; suggesting that their local magnetic field changes at higher ligand concentrations. Dimerisation or aggregation would explain this, but we would also expect to see peak broadening as the tumbling time of the larger molecule increased.

While there are a variety of possible explanations, the fact that the Smad2 titration did not exhibit the same pattern of migration gives a significant indication as to a possible cause. The Smad3, Smad7 and pSmad7 ligands were all sourced from Proteogenix, a commercial laboratory, while the Smad2 ligand was sourced from a different laboratory. One of the simplest explanations is, therefore, that a problem with the Proteogenix peptides could be contributing to the secondary peak migration at high peptide concentrations. This problem could be related to impurities in the peptide or, since there are polyproline repeats in these peptides, cis/trans isomerisation of the proline peptide bonds that result in more than one species of peptide. To test this theory and to hopefully acquire a more reliable dissociation constant, a different approach was taken to source the peptide, which also has the advantage of being cheap. Bacterial expression of the peptide allows it to be isotopically labelled, which is generally necessary for structure calculation, and this is also the first step in generating a WW4/Smad7 bound structure.

5.2.4 SUMO Smad7 and WW4

Bacterial expression of peptides is restricted because of the high level of protease activity against short sequences. There is also the problem of affinity tagging the peptides, as residual amino acids from the cleavage site would constitute a significant proportion of the sequence, and might interfere with ligand affinity assays. ULP-1 is a protease that recognises the tertiary structure of SUMO, a ubiquitin-like protein, and cleaves immediately C-terminal to a di-glycine motif, in the same manner as deubiquitinating enzymes in the ubiquitin pathway (Mossessova & Lima 2000). Using a SUMO tagged peptide positioned C-terminal to a di-glycine motif with a His-tag at the N-terminus, it is possible to express and purify a short sequence of amino acids from bacteria. Using the ULP-1 protease, the affinity and SUMO tags can be cleaved so as to leave the native peptide without excess amino acids. Once the peptide is purified, the challenge is then preparing it for experimental application, which typically requires dialysis and

concentration steps. The molecular weight needs to be of a certain size so as to prevent significant losses during dialysis, since dialysis tubing molecular weight cut-offs are not infinitely small. Visualising the peptide after cleavage is also problematic, as there is only a small number of residues which can bind the stain used. Conventional concentration approaches are also difficult; proteins used in the rest of this thesis were concentrated with centrifugal concentrators, however these are also limited by molecular weight cut-off and the same problem is encountered when considering pressurised stirred-cells. Molecular weight of the peptide should be a consideration when undertaking peptide purification using this approach.

A short sequence corresponding to the PPxY region of Smad7 was cloned in to a SUMO expression vector. To overcome the molecular weight problems, the Smad7 sequence cloned in to the expression vector was extended slightly, so as to ensure the peptide was heavier than 2 kDa. The peptide included the sequence used in the previous sections of this chapter, but incorporated four extra residues at the N-terminus. The SUMO:Smad7 recombinant protein expressed well and gave high yields during purification, shown in Figure 5.2.16A. The digestion was successful but the peptide could not be visualised using Coomassie staining; instead, silver staining was used on tricine gels. The peptide can be seen in Figure 5.2.16B. Once the cleaved SUMO tag was removed by nickel affinity, initially the approach to concentrating the peptide was to dialyse into ammonium bicarbonate using 0.5-1 kDa MWCO dialysis tubing and freeze drying. However, when attempting to reconstitute the freeze dried peptide, the peptide formed high molecular weight aggregates even after attempting reconstitution at a variety of pHs, buffers and additives (Figure 5.2.16C). Freeze drying in water and low concentrations of TFA were also attempted, but with no success, so instead the peptide was dialysed into water and centrifugal evaporation was successfully used to concentrate the peptide. However, precipitation often occurred during concentration, and only limited concentrations were attained, typically around 5 mM.

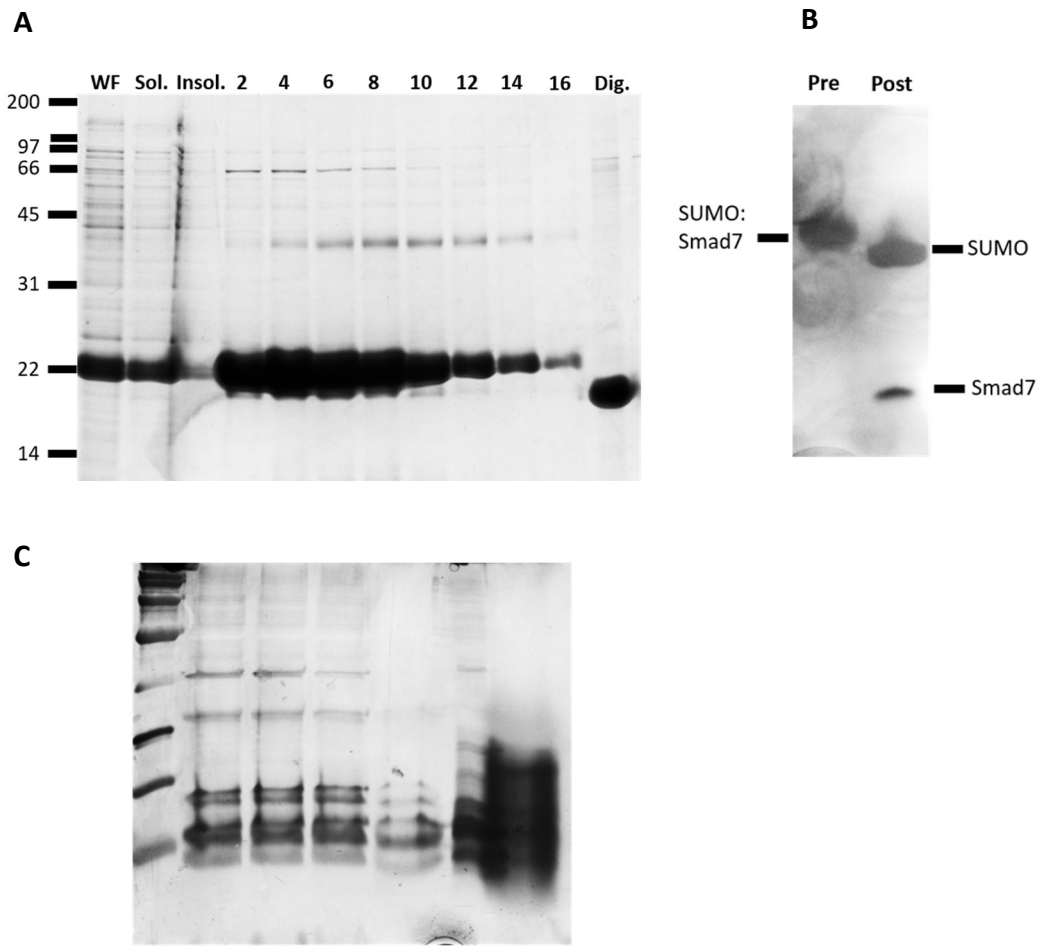


Figure 5.2.16 - A: SDS-PAGE analysis of the nickel affinity purification of SUMO:Smad7 recombinant protein, showing the whole fraction, soluble and insoluble fractions, alternate elution fractions and the post digestion sample. The end lane is the post digestion sample and shows a clear shift in molecular weight. B: The silver stain SDS-PAGE tricine gel for the SUMO:Smad7 digest, showing the pre and post digestion samples. C: Attempted resuspension of the freeze-dried peptide using several different conditions, these included buffers of acidic and basic pH, addition of DMSO and addition of TFA.

Since the concentration of the peptide is restricted and we needed to reach a molar ratio of 1:10, the approach taken with the synthetic peptide, whereby volume is added to the NMR sample by adding small amounts of highly concentrated peptide, is not possible. Instead a two sample approach was taken, where two NMR samples were made, one at the start titration point of 1:0 molar ratio and one at the end titration point at 1:10 WW4 to Smad7, each with exactly the same concentration of ^{15}N labelled GB1:WW4. Because of the relatively low concentration of Smad7, and the volume limitations of

making the NMR sample from two protein samples, is was necessary for the concentration of GB1:WW4 to be very low, at 0.08 mM.

^{1}H-^{15}N-HSQC spectra were taken for both samples, and then aliquots of the 1:0 NMR sample were removed and replaced with aliquots of the 1:10 NMR sample so as to titrate in increasing concentrations of the Smad7 ligand, but to ensure the GB1:WW4 concentration remained the same. The HSQCs from this titration can be seen in Figure 5.2.17A and the trajectories have been plotted in Figure 5.2.17B. Smad7 from the SUMO:Smad7 recombinant will be referred to as Smad7(SUMO).

The Smad7(SUMO) peptide binds GB1:WW4 and the HSQC showed peak migration as with the synthetic peptide titration. The general pattern of binding was the same, indicating that the extra residues do not alter the binding site or bind to any extra residues. However, histidine 465 and tryptophan 450, which have consistently been excluded from K_d averages, did not migrate significantly. Histidine 465 and tryptophan 450 will be included in the titration analysis, for the purpose of comparison. The titration did not show the same secondary migration that was present in the synthetic Smad7 titration. The ΔShift plots for the same four residues analysed in the previous sections of this chapter are shown in Figure 5.2.18A.

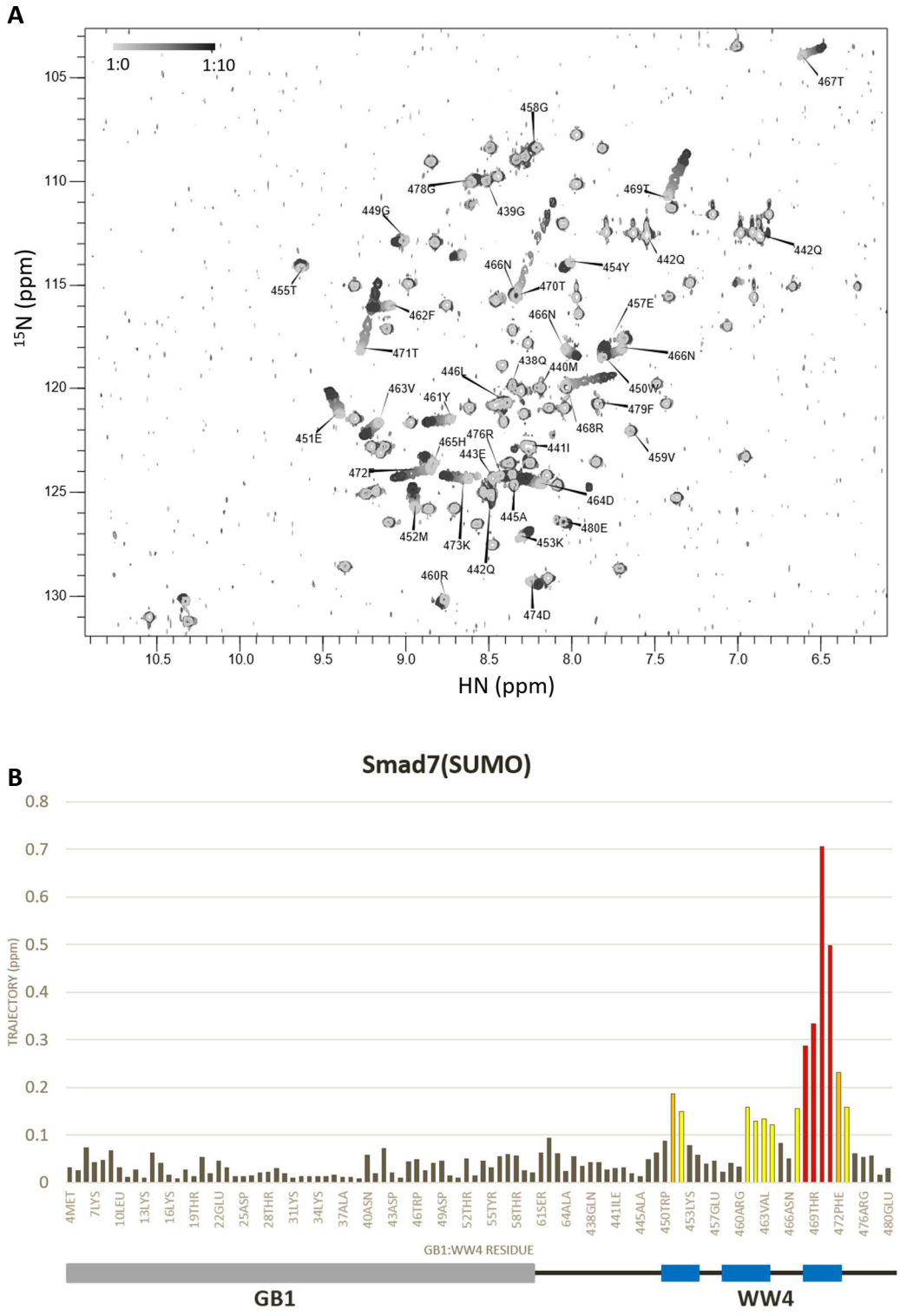


Figure 5.2.17 - A: The GB1:WW4 Smad7(SUMO) titration ^{1}H-^{15}N-HSQCs, lower ligand concentrations are in lighter grey and higher concentrations are in darker grey. The experiment was performed at 500 MHz, 298 K. The sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. The protein concentration was 0.08 mM. B: The trajectory (in ppm) of each GB1:WW4 backbone amide upon titration of the Smad7(SUMO) ligand. Residues are colour coded to indicate extent of peak migration.

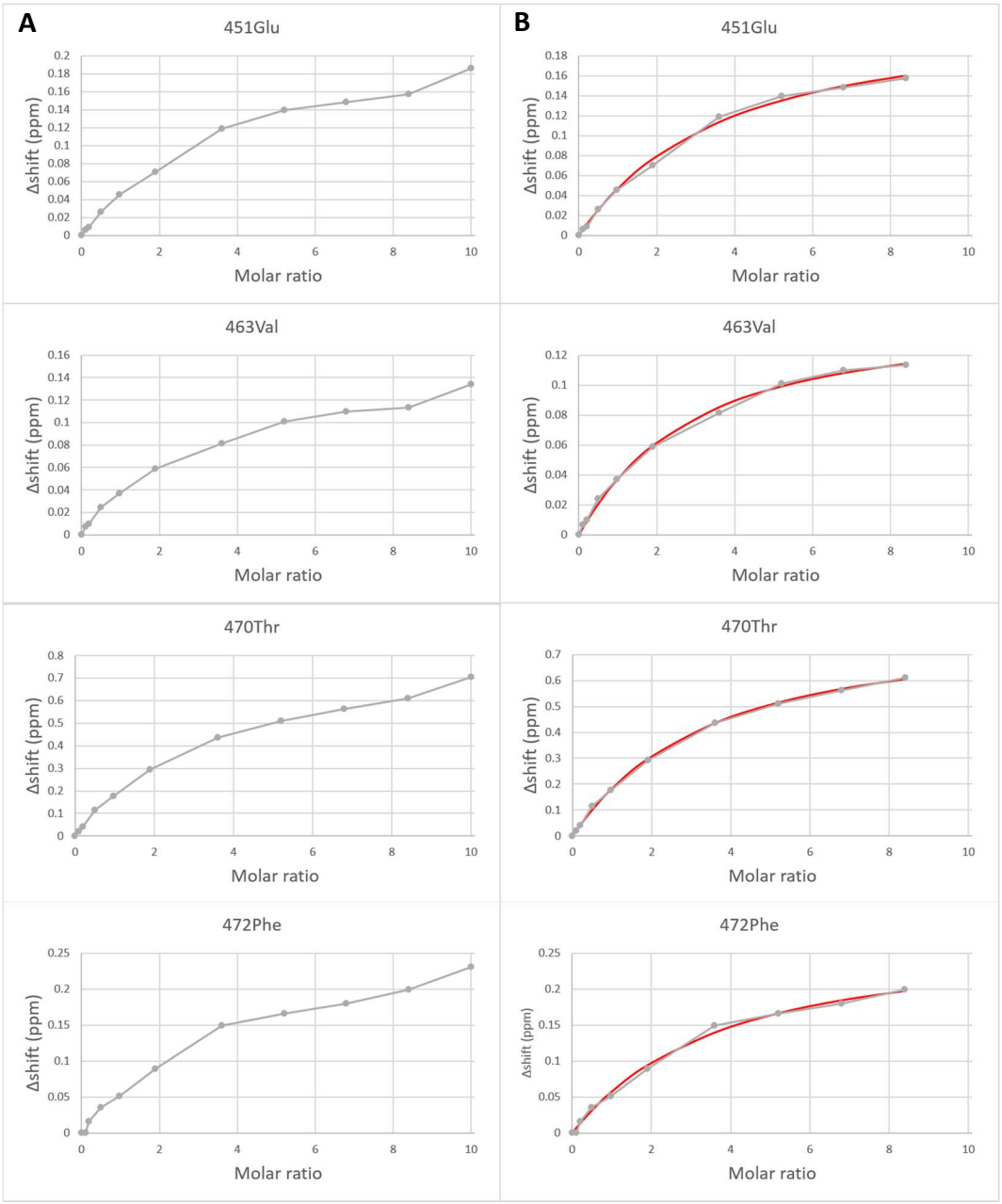


Figure 5.2.18 - A: The ΔShift plots for four binding site residue amide resonances of the GB1:WW4/Smad7(SUMO) titration. B: The CCPN Analysis K_d fits (in red) for the first 10 titration points for the same residues.

The last titration point in each example saw the peak migrate further than expected, and does not fit with the general trend of the titration. This jump in ΔShift is consistent across the majority of the GB1:WW4 peaks. Unlike the synthetic peptide titration, the anomalous migration is in the same direction as the previous titration points. Given the previous propensity for perturbation of peak migration, it was presumed that this jump in peak migration was the result of non-specific binding at high molar ratios.

Since point 11 did not fit with the general trend, Analysis was used to fit the K_d to the first 10 titration points only, Figure 5.2.18B. The binding site K_d values are listed in Table 5.2.5.

Residue	Analysis K_d (mM)[CCPN]
450Trp	*0.21 ±0.05*
451Glu	0.24 ±0.03
452Met	0.23 ±0.02
461Tyr	0.18 ±0.02
462Phe	0.20 ±0.02
463Val	0.18 ±0.02
464Asp	0.26 ±0.01
465His	*0.15 ±0.03*
467Thr	0.22 ±0.05
468Arg	0.23 ±0.02
469Thr	0.23 ±0.02
470Thr	0.22 ±0.01
471Thr	0.22 ±0.01
472Phe	0.24 ±0.04
473Lys	0.27 ±0.03
Average K_d	**0.23 ±0.03**

Table 5.2.5 Binding site K_d values for the GB1:WW4 Smad7 (SUMO) interaction as determined by CCPN Analysis. Average K_d was calculated excluding 450Trp and 465His, which showed minor peak migration. Individual K_d errors indicate the fit error. Standard deviation is given as the error for the K_d average.

The average dissociation constant for the binding site of WW4 and the Smad7 ligand, produced by means of bacterial expression, is much tighter than the K_d calculated from the synthetic peptide titration at 0.82 mM. It is likely that a large contributing factor to the difference in K_d is the secondary peak migration evident at higher concentrations of peptide which seems to be influencing the primary peak migration. It should be noted however, that because the concentration of WW4 is so low, at 0.08 mM, binding saturation by the Smad7 ligand is not reached, this means that K_d prediction is unreliable, and should be treated with caution. With a K_d of 0.23 mM, the estimated percentage of bound WW4 is only 72.7% at the 10^{th} titration point, compared to 83.2% at the 8^{th} titration point with the synthetic Smad7, and 85% at the 7^{th} titration point of the phosphoSmad7 titration (the final points along the first migration phase). Since the secondary migration

occurred at an earlier titration point in the tighter binding phospho-peptide, it is possible that the secondary migration is related to the percentage bound, seemingly at around 80-85%, and the SUMO Smad7 titration might not be saturated enough to show the second phase of migration. Smad3, however, only reached 72% saturation before the secondary migration became apparent, whereas Smad2 reached 78.6% saturation without any unusual peak migration patterns.

The advantage of using a bacterially expressed peptide is that the ligand can be isotopically labelled with heavy carbon and nitrogen isotopes, which are necessary for structure elucidation. These experiments have confirmed that the synthetic peptide can be purified effectively, and bind to the WW4 domain efficiently. The preparations have been completed in order to set in motion the acquisition of the data required for a WW4/Smad7 bound structure, as has been published for several of the NEDD4 family members.

5.2.5 Tandem WW domains

Tandem WW domains have been shown to cooperate in other NEDD4 family members, in order to enhance the affinity for their substrates. WWP2 has four sequential WW domains, and the third and fourth WW domains of WWP2 are separated by a small number of amino acids. This led us to believe that there might be some level of communication between these two domains. To explore the potential for cooperation between the WW3 and WW4 domains of WWP2, resulting in altered binding affinities, a tandem domain GB1:WW3-4 construct was made.

Figure 5.2.19A shows the expression and nickel-affinity purification of GB1:WW3-4. The protein was expressed in ^{15}N and ^{13}C labelled minimal media so as to allow sequential resonance assignment, using the same approach used to assign the GB1:WW4 construct. The His-tag was cleaved, GB1:WW3-4 was gel filtered to enhance purity (Figure 5.2.19B) and the relevant spectra were acquired. It was possible to assign the entire GB1 and WW4 domains in the GB1:WW3-4 ^{1}H-^{15}N-HSQC, however, almost all of the WW3 domain amide resonances, besides five residues, lacked HNCACB/CBCACONH peaks, or could not be fit into the sequence as a result, and were therefore unassignable.

In order to assign the WW3 peaks from the GB1:WW3-4 spectrum, a GB1:WW3 construct was made, and this was expressed in isotopically labelled minimal media and purified by nickel affinity purification (Figure 5.2.20A). The His-tag was cleaved by thrombin, and GB1:WW3 was gel filtered (Figure 5.2.20B). The appropriate spectra were acquired for sequential assignment of GB1:WW3 (^{1}H-^{15}N-HSQC, CBCACONH, HNCACB). The resonance assignment was performed by Jack Dwyer, a project student in Dr Tharin Blumenschein's laboratory at the University of East Anglia. These assignments were used to inform the assignment of the GB1:WW3-4 HSQC spectrum, as the majority of the peaks were in roughly the same position. Using this approach, the majority of the GB1:WW3-4 HSQC peaks were assigned. The assigned GB1:WW3 and GB1:WW3-4 spectra are shown in Figure 5.2.21. A Smad7(SUMO) titration was performed on ^{15}N labelled GB1:WW3-4 (Figure 5.2.22A), and the trajectories were plotted against residue number as above (Figure 5.2.22B).

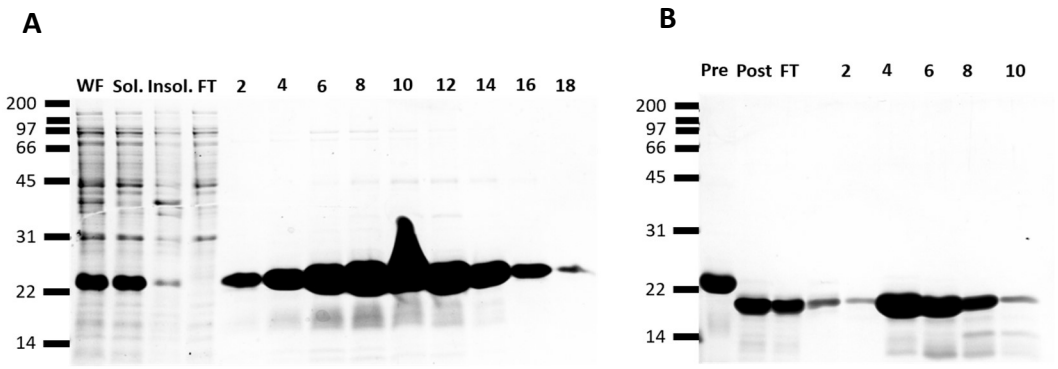


Figure 5.2.19 - A: Expression and nickel affinity purification of GB1:WW3-4, showing the whole fraction, soluble and insoluble fractions, the nickel column flow through and alternate elution fractions. B: Thrombin digest of GB1:WW3-4, pre and post digestion fractions, and the nickel column flow through which shows only the cleaved protein without the His-tag attached. Alternate fractions are shown from the gel filtration peak of the GB1:WW3-4 protein.

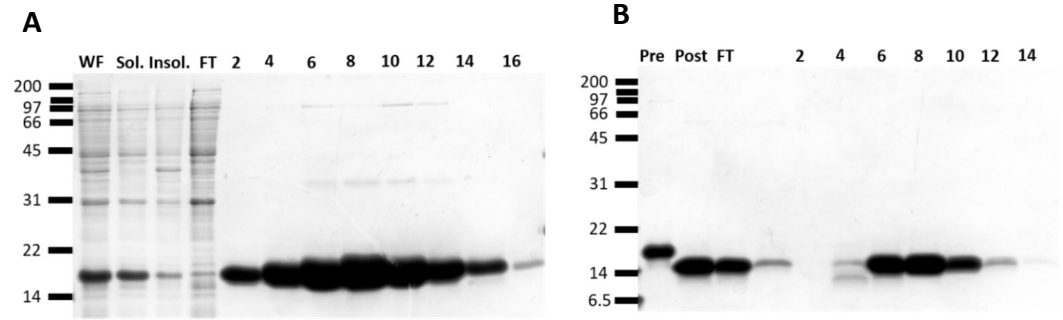


Figure 5.2.20 - A: Expression and nickel affinity purification of GB1:WW3, showing the whole fraction, soluble and insoluble fractions, the nickel column flow through and alternate elution fractions. B: Thrombin digest of GB1:WW3, pre and post digestion fractions, and the nickel column flow through which shows only the cleaved protein without the His-tag attached. Alternate fractions are shown from the gel filtration peak of the GB1:WW3 protein.

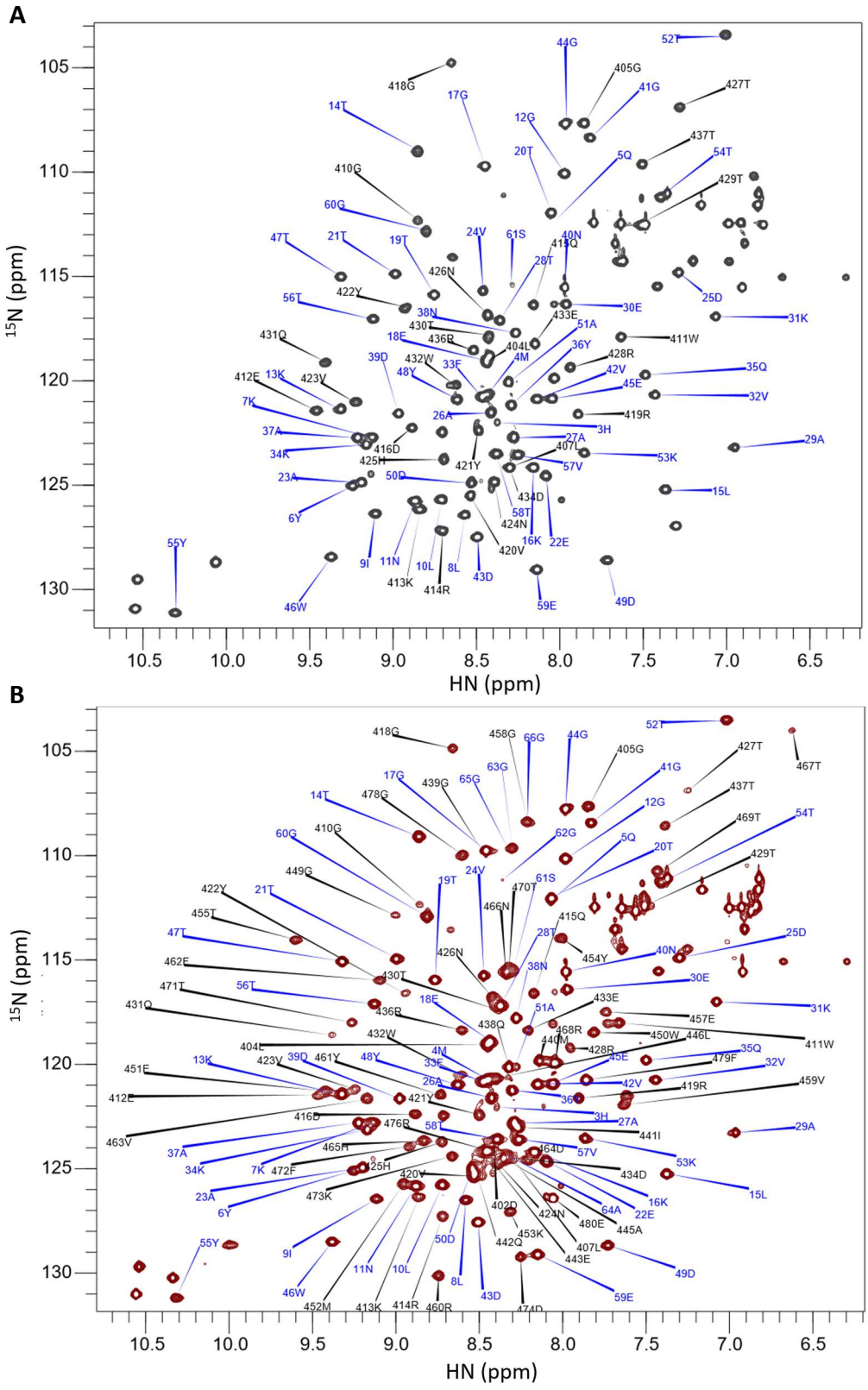


Figure 5.2.21 - A: The assigned GB1:WW3 1H-15N-HSQC spectrum. B: The assigned GB1:WW3-4 1H-15N-HSQC spectrum. Spectra were acquired at 500 MHz at 298 K. The samples were prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. The GB1 domain labels are shown in blue and the WW domain labels are shown in black.

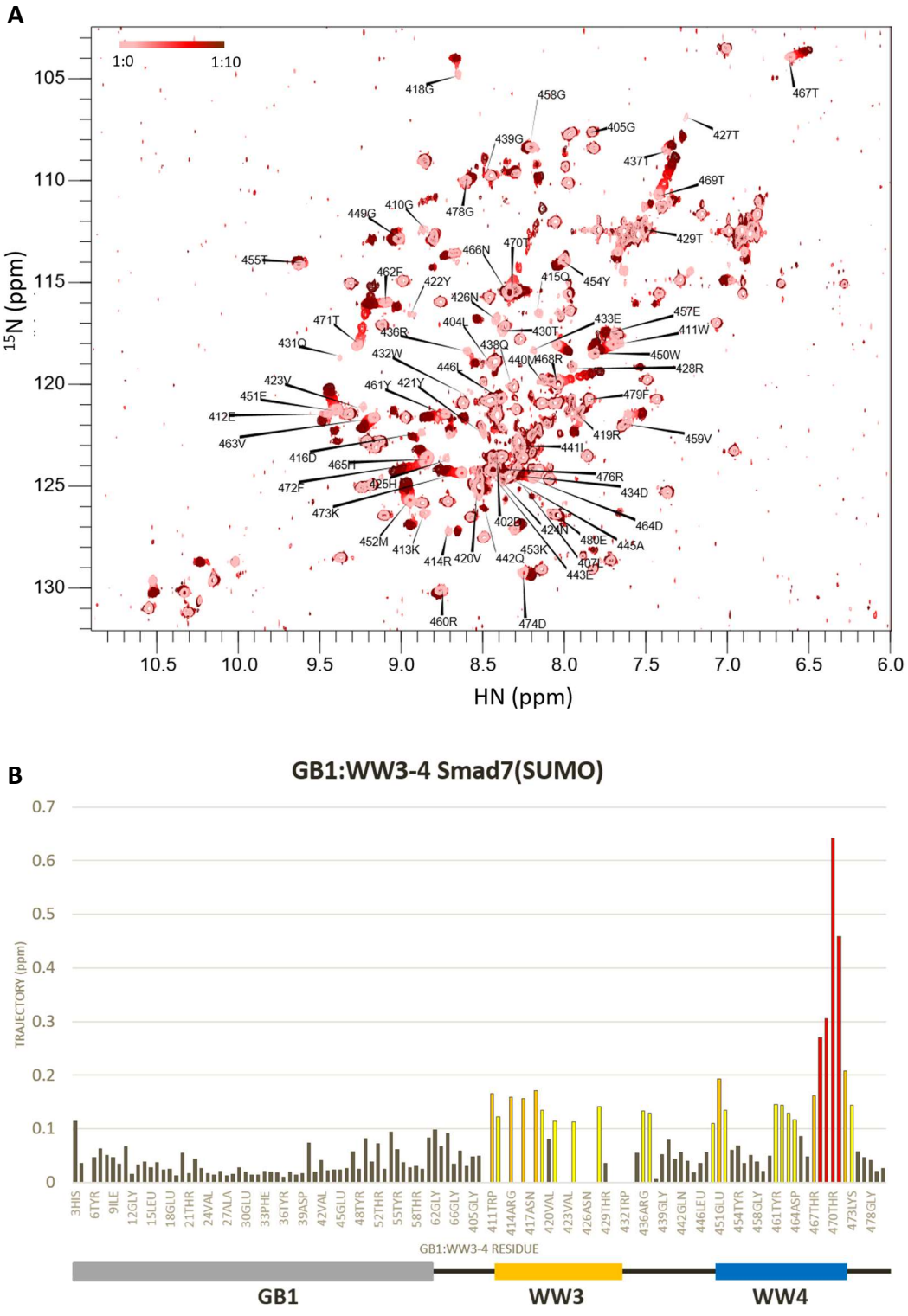


Figure 5.2.22 - A: The GB1:WW3-4/Smad7(SUMO) titration ^{1}H-^{15}N-HSQCs, lower ligand concentrations are shown in lighter red and higher ligand concentrations are shown in darker red. Only the WW3-4 residues are labelled. The experiment was performed at 500 MHz, 298 K. The sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. The GB1:WW3-4 concentration was 0.08 mM. B: The trajectories in ppm for each peak plotted against their residue numbers.

The same residues of WW4 appear to be involved in coordinating the Smad7 ligand. From the trajectory plot, it was clear that some of the peaks assigned to the WW3 domain were also migrating in a Smad7-dependent fashion. Some of the peaks appeared to be in intermediate exchange, which caused the peaks to broaden and decrease in intensity as the resolution decreases. Since the protein concentration is so low, the peak broadening meant that the peaks disappeared, and it was not possible to track the migration or the change in intensity. This is the reason for the missing trajectories throughout the WW3 domain residues in Figure 5.2.22B. Because the WW3 domain appears to be involved in the binding of Smad7, we wondered whether WW3 domain might bind the Smad7 ligand independently of WW4. To test this, a Smad7(SUMO) titration was also performed on the GB1:WW3 recombinant (Figure 5.2.23A and B).

The Smad7 ligand bound to the WW3 domain in the absence of the WW4 domain, and there was a mixture of fast exchange and intermediate exchange processes. The pattern is subtly different from the WW3 domain of GB1:WW3-4 titration, and fewer residues were in intermediate exchange. As with the GB1:WW3-4 titration, the peaks in intermediate exchange broadened and did not re-emerge, indicating that binding did not reach saturation. The binding regions of WW3 and WW4 domains from the Smad7(SUMO) titrations have been aligned in Figure 5.2.24 below. The residues have been colour coded according to the extent of their trajectories (yellow-orange-red), and the residues that appear to be in intermediate exchange have been highlighted in blue.

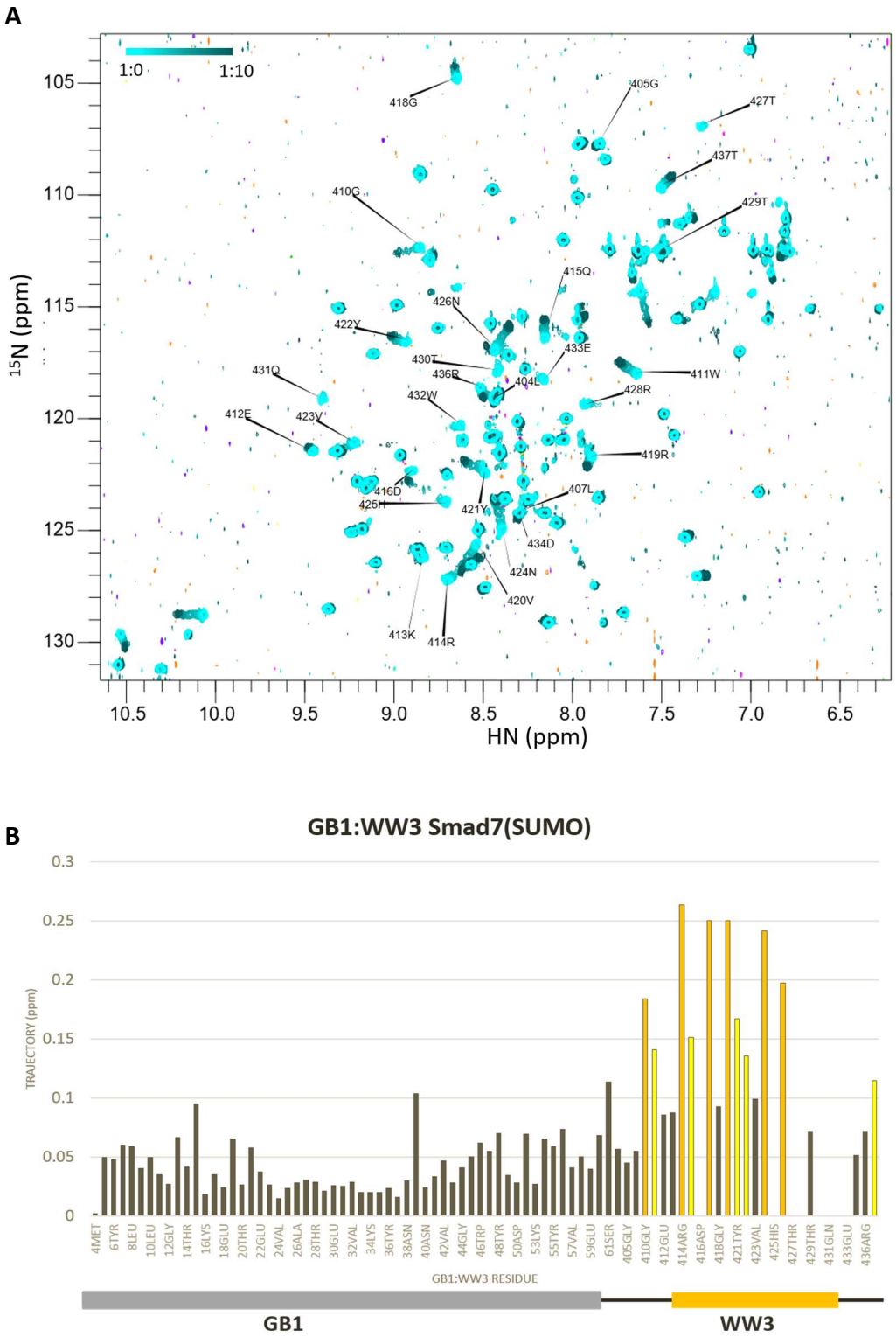


Figure 5.2.23 - A: The GB1:WW3/Smad7(SUMO) titration, lower ligand concentrations are shown in lighter turquoise and higher ligand concentrations are shown in darker turquoise. The experiment was performed at 500 MHz, 298 K. The 0.08 mM protein sample was prepared in 20 mM Sodium phosphate buffer, 150 mM NaCl, pH 6.8. B: The trajectories for each peak plotted against their residue numbers (note the scale is smaller than in other titrations because of the absence of the large WW4 domain trajectories).

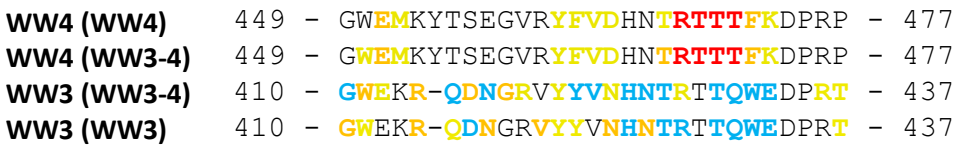


Figure 5.2.24 - Alignment of the amino acids of WW4, WW3-4 and WW3 of WWP2, with residues colour coded according to the extent of their HSQC peak trajectories in the Smad7(SUMO) titrations. Red indicates the most extreme peak movement, orange indicates medium peak movement, yellow indicates smaller peak movement, blue indicates peaks that appear to be in intermediate exchange and black indicates residues that show only minor movement.

If the WW3 domain binds in the same fashion as the WW4 domain, it is likely that we are unable to calculate the majority of the WW3 domain binding site affinities. Particularly as the part of WW3 corresponding to the red threonines of WW4 are mostly not visible (Figure 5.2.24 and Figure 5.2.23). In fact, the threonine that we are able to see (429Thr) is in a very crowded region of the HSQC, and it is not entirely certain whether the peak migration can be trusted. There are, therefore, limits to the conclusions that can be drawn from the data. Since the WW3 domain appears to bind the Smad7 peptide, this effectively doubles the binding site concentration and halves the ligand ratio, which was taken into account when calculating the K_d values. Using the CCPN Analysis software, dissociation constants were fit to the peaks of the binding site for each of the titrations.

For the GB1:WW3 titration, approximately half of the binding site residues showed a dramatic change of shift at the last titration point, as with the GB1:WW4 Smad7(SUMO) titration, while the other half showed a less dramatic change. The K_d fits for four of the GB1:WW3 residues are shown in Figure 5.2.25A. For the GB1:WW3-4 titration, virtually all of the binding site residues from both domains showed a dramatic change, the last titration point was therefore ignored for the dissociation constant calculation. The K_d fits were much better for the WW3 domain binding residues in the GB1:WW3 titration (Figure 5.2.25A), than for the WW3 domain binding residues in the GB1:WW3-4 titration (Figure 5.2.25B), where peak migration appeared to be much less consistent. In fact some of the peak migrations seemed to have a secondary phase of migration similar to the synthetic peptide titration. This second phase might be related to a certain point in the saturation curve, as these residues appear to reach saturation relatively soon in the titration. The early saturation is reflected in the affinities, which are

high for the WW3 domain of GB1:WW3-4 (K_d is low). The dissociation constants are shown in Table 5.2.6.

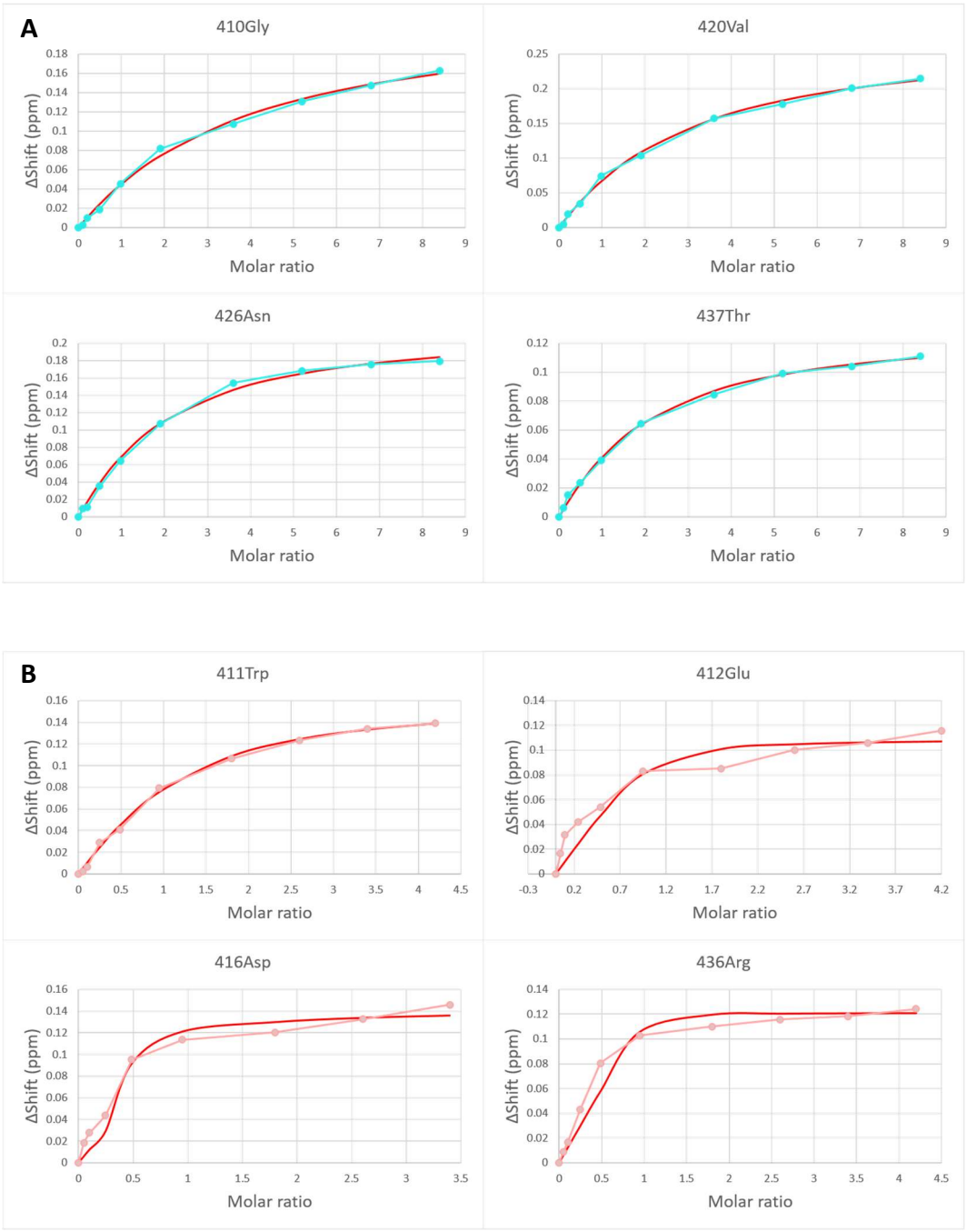


Figure 5.2.25 - A: The ΔShift plots for four of the binding site residues of GB1:WW3 from the Smad7(SUMO) titration (cyan) with the K_d curve (red). B: The ΔShift plots for four of the binding site residues of GB1:WW3-4 from the Smad7(SUMO) titration (pink) with the K_d curve (red).

Residue	WW3-4 Smad7(SUMO)	WW3 Smad7(SUMO)	WW4 Smad7(SUMO)
410Gly		0.26 ±0.04	
411Trp	0.09 ±0.01	0.16 ±0.03	
412Glu	0.01 ±0.02	-	
414Arg	0.02 ±0.03	0.12 ±0.01	
415Gln		0.16 ±0.02	
416Asp	0.03 ±0.02		
417Asn		0.26 ±0.03	
418Gly	0.005 ±0.01		
419Arg	0.04 ±0.05		
420Val		0.19 ±0.02	
421Tyr	0.01 ±0.03	0.10 ±0.02	
422Tyr		0.09 ±0.02	
424Asn	0.03 ±0.05	0.15 ±0.01	
428Arg	0.02 ±0.03		
426Asn		0.12 ±0.01	
436Arg	0.002 ±0.004	-	
437Thr	0.01 ±0.02	0.12 ±0.01	
450Trp	0.07 ±0.04		*0.21 +/-0.05*
451Glu	0.21 ±0.05		0.24 +/-0.03
452Met	0.43 ±0.13		0.23 +/-0.02
461Tyr	0.44 ±0.07		0.18 +/-0.02
462Phe	0.08 ±0.04		0.20 +/-0.02
463Val	0.46 ±0.09		0.18 +/-0.02
464Asp	0.30 ±0.05		0.26 +/-0.01
467Thr	0.14 ±0.04		0.22 +/-0.05
468Arg	0.39 ±0.09		0.23 +/-0.02
469Thr	0.42 ±0.08		0.23 +/-0.02
470Thr	0.59 ±0.19		0.22 +/-0.01
471Thr	0.51 ±0.15		0.22 +/-0.01
472Phe	0.59 ±0.16		0.24 +/-0.04
473Lys	0.44 ±0.11		0.27 +/-0.03
Average K_d (mM)	**0.21 ±0.22**	**0.16 ±0.06**	**0.23 ±0.03**

Table 5.2.6 Binding site K_d values for the GB1:WW3-4, GB1:WW3 and GB1:WW4 interactions with the Smad7(SUMO) ligand, as determined by CCPN Analysis. Individual K_d errors indicate the fit error. Standard deviation is given as the error for the K_d average.

The WW3 domain by itself appears to have a higher affinity than the WW4 domain for the Smad7 ligand, although many of the residue dissociation constants could not be defined. The GB1:WW3-4 affinity is tighter than the GB1:WW4 domain by itself, but looser than the GB1:WW3 affinity. When the two domains are expressed in tandem,

the affinity of the WW3 domain for the Smad7 ligand increases substantially (when only comparing the residues of the WW3 domain), and the dissociation constant is in the low micromolar range, at 23 ±25 μM (only taking in to account residues from glycine 410 to threonine 437 from the WW3-4 titration). Although, again this is in the absence of a lot of the residue dissociation constants. This is offset in the average K_d by a decrease (relative to the WW4 domain by itself) in the WW4 domain Smad7 affinity, which is at 362 μM. The GB1:WW3-4 WW3 domain K_d still incorporates the secondary stage of migration. Although there is no change in the direction of migration, it is possible that the increase in shift migration at this point might relate to some interference from the lower affinity WW4 domain binding, which might be causing a local magnetic field perturbation, although we do not observe the WW3 domain binding perturb the WW4 domain peak migrations.

5.2.5 WWP2C-ΔHECT

Western blots that probed mammalian epithelial cell lysates with an anti-WWP2-C antibody revealed the presence of a previously unobserved TGFβ-inducible isoform (Figure 5.1.6). To explore the evidence of a new TGFβ-responsive WWP2 isoform, we searched the EST database for intronic sequences of the WWP2 gene that might identify a tentative transcript responsible for the band seen in the western blot shown in Figure 5.1.6. Our initial search led to the identification of an EST database entry that starts in exon 17, contains exon 18 and 19, and terminates 351 nucleotides into intron 19/20 of WWP2. Figure 5.2.26 shows a comparison between the WWP2-FL and WWP2-C transcript start and stop sites, and this EST, which has the genbank accession code BX471495.1. A stop codon arises from the first in frame codon in intron 19/20. A gene product from this transcript would not contain the full HECT domain and it is highly likely that the HECT domain would have no ubiquitin ligase activity as the DNA sequence coding for the catalytic C-terminal lobe of HECT, which also contains the catalytic cysteine, is 3’ from intron 19/20.

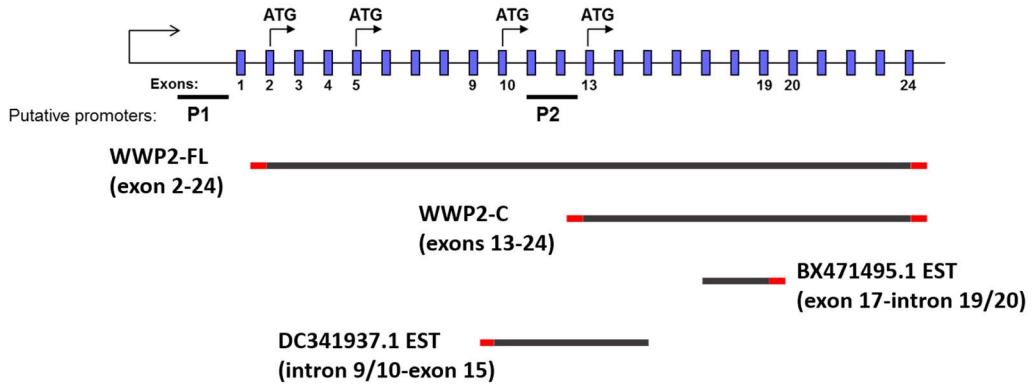


Figure 5.2.26 - The WWP2 gene locus (not to scale) depicted as exons in blue and introns depicted as thin black lines. A selection of start codons and putative promoters are labelled. The WWP2-FL and WWP2-C transcripts are shown as thick black lines that represent only the incorporated protein-coding exons, the red regions represent the intronic 5' and 3' untranslated regions. The ESTs discussed in the text are also shown, the thick black lines also represent the exons included in the ESTs and the red regions represent the intron regions that are also present, and which indicate either a new transcription start site (DC341937.1), or a new termination site created by the retention of intron 19/20 (BX471495.1).

A pair of primers were designed so that the forward primer was positioned at exon 17 and the reverse primer was positioned a short way in to intron 19/20. cDNA was synthesised by extracting RNA from the TGFβ-responsive COLO357 pancreatic adenocarcinoma cell line (Morgan et al. 1980), and performing reverse transcription using random primers. GoTaq DNA polymerase was used to perform PCR using the primers designed to detect the WWP2C-ΔHECT isoform. Figure 5.2.27A shows the expression levels of the ΔHECT transcript in COLO357, over the course of 8 hours TGFβ stimulation. There was also evidence of expression in the VCaP prostate cancer vertebral metastasis cell line (Korenchuk et al.) (Figure 5.2.27B), the melanoma cell line A375 (Figure 5.2.27C) and the melanoma cell line SK-MEL28 (Figure 5.2.27D).

There is evidence of expression of a ΔHECT isoform in all four cell lines, and that its expression is inducible by TGFβ, particularly in COLO357 and A375 although at different time scales (Figure 5.2.27). Some caution must be taken at this early stage when looking at patterns of expression because of a certain unreliability. It would be pertinent to perform further repeats to increase confidence in the observed expression patterns, although the TGFβ inducibility does align with the western blot data.

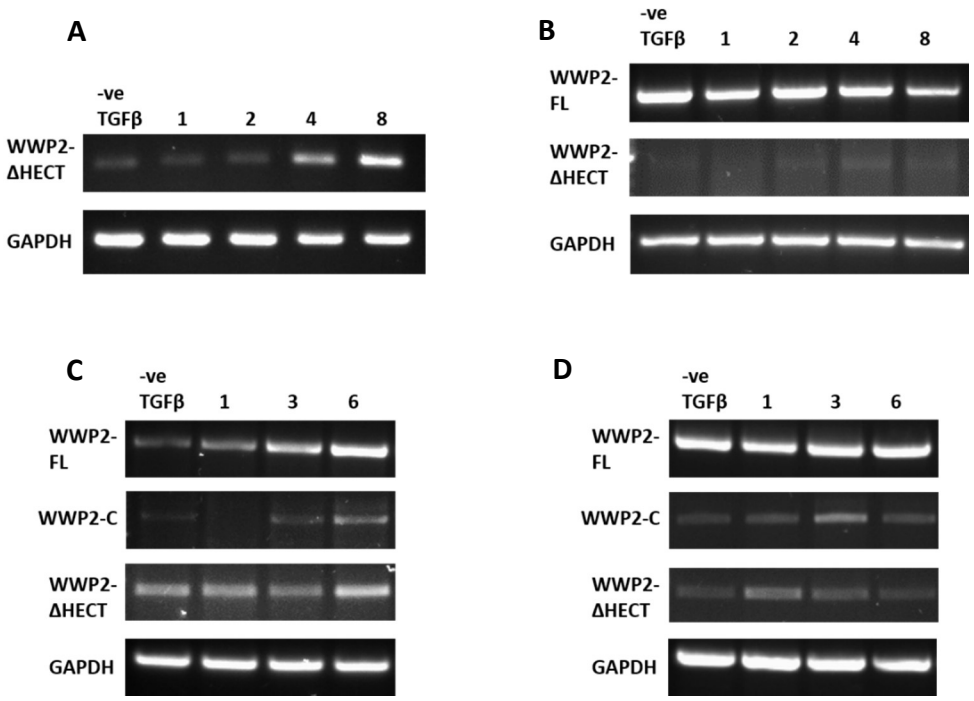


Figure 5.2.27 - A: Semiquantitative-PCR showing WWP2C-ΔHECT expression at 1, 2, 4 and 8 hours after TGFβ stimulation in COLO357 cells, and the GAPDH control. B: WWP2-FL and WWP2C-ΔHECT expression after TGFβ stimulation in VCaP cells. C: WWP2-FL, WWP2-C and WWP2C-ΔHECT expression after TGFβ stimulation in A375 cells. D: WWP2-FL, WWP2-C and WWP2C-ΔHECT expression after TGFβ stimulation in SK-MEL28 cells.

The main focus of these experiments was initially to confirm the presence of the ΔHECT isoform transcript, but the experiments were broadened to explore the potential interplay between WWP2 isoforms, using primers designed to detect the specific isoforms. WWP2-FL was expressed in VCaP and SK-MEL28 cell lines, and appeared to be induced by TGFβ in A375 (Figure 5.2.27), which is consistent with previous evidence of TGFβ-inducibility (Soond & Chantry 2011). WWP2-C which, up until now, has only been found to be expressed in the chondrogenic cell line ATDC5, is expressed in A375 and SK-MEL28 melanoma cell lines, but not the VCaP cell line. Expression of WWP2-C appears to be induced by TGFβ.

It was originally thought that the ΔHECT isoform was the product of a splice variant of the WWP2-C transcript, resulting in a 31.5 kDa protein. However, the consistent appearance of HECT-containing proteins at molecular weights smaller than expected when analysed by SDS-PAGE, gave the impression that the ΔHECT isoform might be larger than expected (WWP2C-ΔHECT appears at roughly 30 kDa in Figure 5.1.6). A search of the EST database with intronic regions 5' from the WWP2-C start codon at exon 13, produced

a hit at intron 9-10 with the genbank accession code DC341937.1. The sequence is from thalamus tissue and contains 126 nucleotides of intron 9-10, contains the entirety of exons 10-14, and terminates midway through exon 15 (Figure 5.2.26). There is a TATA box, an indicator of promoter activity, in intron 9-10, 5' from the start site of the EST.

If this were to be part of the same transcript as the ΔHECT EST described above, then this would result in a 38 kDa protein that contains both the WW3 domain and WW4 domain, and terminates midway through the HECT domain. Figure 5.2.28 shows a schematic of the domain composition of the two different possible WWP2C-ΔHECT isoforms that are discussed here. More time would be required to determine the full sequence of WWP2C-ΔHECT and define the N-terminus. Had this time been available, primer pairs would have been designed to target the intron regions predicted to produce the protein product seen in the western blots. Regardless of this, there is more evidence here of the existence of a novel WWP2 isoform, presumably with no ubiquitin ligase activity, and possibly containing tandem WW domains.

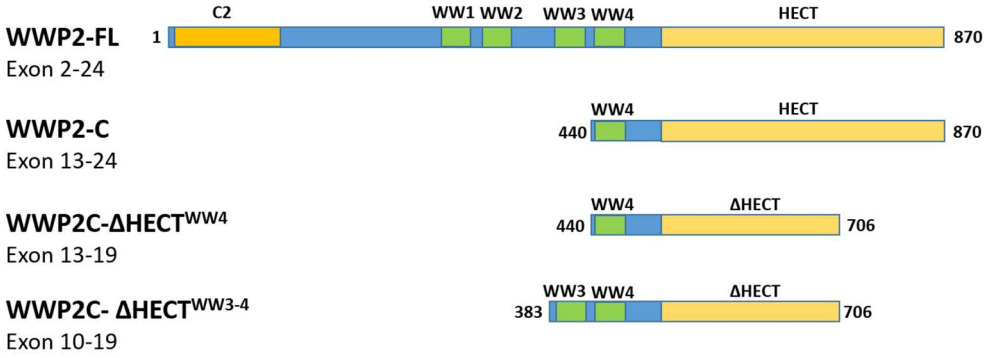


Figure 5.2.28 - A schematic showing the domain composition of the two potential WWP2C-ΔHECT isoforms compared to the domain composition of the WWP2-FL and WWP2-C isoforms. The sequence-coding exons are given, as well as amino acid N and C-termini numbers relative to the WWP2-FL protein sequence. The C2 domain is shown in orange, the WW domains in green and the HECT domain in yellow.

The EST database find suggesting the presence of a WW3 and WW4 domain-containing isoform, indicates that some tandem domain communication, such as enhanced substrate specificity, might be occurring. This would be significant in the context of multiple WWP2 isoforms being expressed in one cell system, as seen in Figure 5.2.27, particularly with the single WW domain-containing WWP2-C and the tandem domain containing WWP2-FL and, potentially, the WWP2C-ΔHECT isoform. The titration

data seem to affirm the notion that these domains cooperate, and it appears as though WW4 sacrifices substrate affinity to enhance the affinity of WW3 for the Smad7 substrate.

5.3 Discussion

A summary of the dissociation constants from the various titrations performed herein are shown in Table 5.3.1.

	Smad7	pSmad7	Smad2	Smad3	Smad7 (SUMO)
WW4	820±198 µM*	440±119 µM*	973±307 µM	650±408 µM*	227±27 µM
WW3	-	-	-	-	160±59 µM
WW3-4	-	-	-	-	214±214 µM
$WW3^{(WW3-4)}$	-	-	-	-	23±25 µM
$WW4^{(WW3-4)}$	-	-	-	-	362±175 µM

Table 5.3.1 The dissociation constant of WW3, WW4 and the tandem WW3-4 domains. Dissociation constants with an asterisk are the product of a curve which was fit manually to the first stage of a two stage migration. Standard deviation is given as the K_d error.

The expected interaction between Smad7 and the WW4 domain has been confirmed here, and the binding site has been determined. The introduction of a phosphate group at serine 206 of the Smad7 ligand enhanced the interaction nearly 1.9 fold over the non-phosphorylated ligand. It was thought that the introduction of this phosphate group might change the pattern of peak trajectories slightly, perhaps giving some indication as to the position of the residues that might be responsible for coordinating the phosphate group, but the pattern of trajectories is essentially the same. The biological relevance of the enhanced affinity between phosphorylated Smad7 and the WW4 domain of WWP2 needs to be further explored, but it can be speculated that this might be a means by which Smad7 turnover is enhanced by increasing affinity for its E3 ligase. Phospho-regulation between receptor Smads and their WW domains has been demonstrated before (Aragón et al. 2011; Alarcón et al. 2009; Gao et al. 2009), but has never been demonstrated between the inhibitory Smad7 and its binding WW domains.

Residues across the binding pocket of WW4 have different affinities for the ligand, which contribute to the global K_d. Figure 5.3.1A shows the K_d reciprocals, the K_a, for each of the residues of the binding pocket. The K_a values are used here because they are clearer to present when using graph plots. In this figure, the larger values are the tightest binding. Figure 5.3.1B shows a K_d heatmap, with the residues with the highest affinity for Smad7 highlighted in red. Each of the WW4 domain binding site residues had a higher affinity for the phosphorylated ligand when compared to the unphosphorylated ligand, besides

arginine 468, which had a slightly lower affinity (Figure 5.3.1A). In particular, valine 463, threonine 467 and phenylalanine 472 have significantly enhanced affinities. These residues have been coloured red in the phospho-Smad7 K_d heatmap (Figure 5.3.1C). Even the two tightest binding residues from the Smad7 titration, coloured in red in Figure 5.3.1B, have low affinities when compared to the phosphoSmad7 titration affinities, and would be classified in the low end of the yellow range in Figure 5.3.1C. Tyrosine 461 and phenylalanine 472 are the residues constituting the XP binding pocket, and valine 463 and histidine 465 constitute the secondary specificity pocket. Phenylalanine 472 and valine 463 are particularly tight binding in the phosphoSmad7 titration, but in both titrations histidine 465 had a low K_d and did not fit a binding curve very well. In the WW4 domain structure, several of the models showed the histidine 465 tilted away from the binding site, and this might be the reason that it does not participate in high-affinity binding. Interestingly, lysine 473 has a much higher affinity for the phosphorylated ligand and is one of the most enhanced. Given the predicted proximity of this residue to the N-terminal region of the ligand, and the positive charge of lysine at the pH used in these experiments, lysine 473 might represent the critical amino acid involved in coordinating the phosphate group and enhancing binding affinity.

Previously there was no evidence that the WW4 domain of WWP2 interacted with the receptor Smads; in immunoprecipitation experiments WWP2-C precipitated with Smad7 but not Smad2 or 3 (Soond & Chantry 2011). The interaction between Smad2 and Smad3 ligands and the WW4 domain was, therefore, unexpected. The Smad2 peptide gave the first titration that did not show a secondary migration, and the affinity was comparatively low. The Smad3 interaction, when the curve was manually fit, seemed to have a higher affinity, despite the ligand similarities. Many other NEDD4 E3 ligases bind to the receptor Smads, but do not ubiquitinate them, instead Smad binding partners, such as transcription cofactors, are ubiquitinated. This might explain the lack of ubiquitin ligase activity of WWP2-C against receptor Smads. The interaction between WW4 and Smad3 appears to be higher affinity than that of the native Smad7 ligand, but not the phosphorylated Smad7 ligand. Phosphorylation of Smad7 might therefore cause the preferential degradation of Smad7 over receptor Smads, or their binding partners (although there is currently no evidence of this kind of activity).

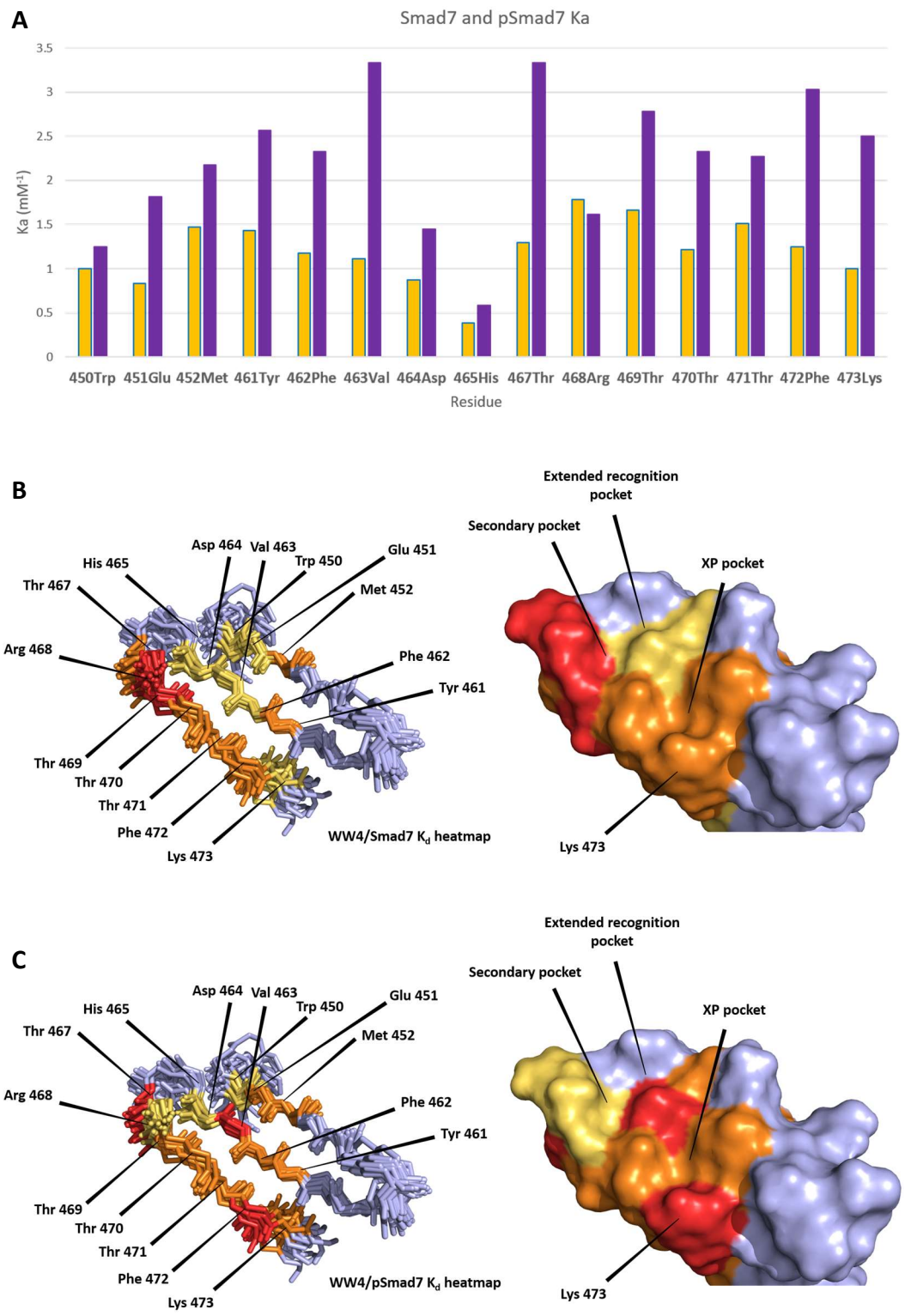


Figure 5.3.1 - A: Residue-specific Ka values of the WW4 domain binding site for the Smad7 ligand (orange) and the phosphorylated Smad7 ligand (purple). B: A heatmap indicating the tightest binding residues (red) in the Smad7 titration, relative to the rest of the binding site, shown on the WWP2 WW4 20 model CNS ensemble. C: A heatmap indicating the tightest binding residues in the pSmad7 titration, relative to the rest of the binding site.

Some of the titrations saw the appearance of a second migration phase, to which it was possible to fit two curves. Whilst it is interesting to observe differences between the fit curves of the second phase of migration, this type of analysis is not appropriate to determine any kind of affinity constant, since it seems as though the secondary migration does not occur at the same time as the initial migration. If this were the case, we would not expect to see the 'bump'that is apparent in each of the graphs that plot the change in shift, but instead we would expect to see a smooth curve that could not be fit to one K_d. As a result, only the analysis of the initial binding curve has been used to determine dissociation constants for the Smad/WW domain affinities. The cause of the secondary migration is not certain, but it is only evident at high ligand concentrations, and did not appear in the Smad2 titration. Since the source of the Smad2 peptide was different it is plausible that the cause was an impure or non-homogenous peptide, however, because the Smad2 affinity is lower than that of Smad3 and Smad7, the binding site did not reach the same point of saturation as the other titrations. WW domains are able to bind ligands in two different orientations (Zarrinpar & Lim 2000), the secondary migration might be the result of lower affinity ligand binding in the alternative orientation at high concentrations. Since there is evidence of WWP2-dimerisation from previous studies (Liao & Jin 2010; Soond & Chantry 2011), another possible reason could be ligand binding-induced dimerisation of the WW domains, which would cause a perturbation of the local magnetic field. Compared to the dissociation constants of the SUMO peptide titrations, the synthetic peptide titration dissociation constants are high and it is possible that the secondary migration event is interfering with the primary migration. Because of this the K_d is unreliable, although a certain level of internal comparison is appropriate.

The SUMO Smad7 titrations seem to have a higher affinity but because of the low concentrations involved, saturation was not reached. The caveat also applies then, that the K_d values are unreliable, but internal comparison is possible. The secondary migration was not seen, but again, because of the low saturation, it is possible that this was not observed. The WW3 domain affinity for the Smad7 ligand was higher than that of the WW4 domain when they were expressed as individual domains. However, several of the K_d values for the WW3 domain binding site were not included because the residues appeared to be in intermediate exchange. The K_d is, therefore, only tentative, but the fact that the residues are in intermediate exchange suggests that the rate of ligand diffusion away from the binding site is lower than for the WW4 domain, which in itself indicates that the affinity is higher. When the titration was performed on the tandem domains, it

was suspected that the WW4 domain affinity might be enhanced. However, it was the WW3 domain affinity that was significantly enhanced, and the WW4 domain affinity was reduced. As with other WW domains there seems to be cooperativity between them, and it appears that the WW4 domain sacrifices binding affinity to enhance the affinity of the WW3 domain for its ligand. Effectively this would mean the WW3 domain outcompetes the WW4 domain for the Smad7 substrate. The effect of different substrates binding at different WW domain positions along an E3 ligase is so far unknown, but it is possible to imagine the substrate positioning being optimised for ubiquitination at different sites, or if the WW domain is further away from the ubiquitin ligase domain, it might prevent the substrate from being ubiquitinated all together. NEDD4 family members involved in the regulation of the TGFβ signalling pathway have been shown to interact with Smads, but not ubiquitinate them. WWP2-FL has both WW3 and WW4 domains in tandem. If WWP2-FL operates by the same mechanism as other NEDD4 family members, the ligase would associate with Smad7 and translocate to the TGFβ receptor before ubiquitinating both the receptor and Smad7. It would appear that the interaction with Smad7 is mediated by the WW3 domain, which might position the Smad7 substrate away from the HECT active site and allow the WWP2/Smad7 complex to reach the receptor before a change in conformation, or perhaps a phosphorylation, decreased the affinity of Smad7 for WW3 and increased the affinity for WW4. Here the HECT domain might be active against the Smad7 substrate.

It has not gone unnoticed that the dissociation constants are generally much higher than those from the ITC experiments outlined in the introduction of this chapter, apart from that of the WW3 domain in tandem with the WW4 domain. In the context of the WW4 domain present in the WWP2-C isoform, the current model is that Smad7 is ubiquitinated by WWP2-C in the absence of translocation to the TGFβ receptor. Since the ligase does not need to remain in complex whilst traveling to the receptor, the lower affinity might be necessary to prevent receptor ubiquitination. In this context, the rate of ubiquitination be the HECT domain becomes more important. Questions are also raised here about the reliability of calculating K_d by NMR. A comparison of K_d by NMR and K_d by ITC showed that K_d by NMR was universally higher (typically over 4-fold) than that of ITC (Fielding et al. 2005) when observing interactions between bovine serum albumin and several ligands. The reason given for this is the propensity of non-specific binding to interfere with K_d by NMR. Some caution should therefore be taken when comparing the K_d values calculated here, with those calculated by ITC. While it is essential that ITC is

performed to determine a comparable K_d, this data has elucidated the binding site, which is essential if WW domain binding is to be disrupted by therapeutics. This data has also shown the difference between WW3 ligand affinity when expressed independently and when expressed as part of a tandem WW domain sequence.

The evidence of a new isoform here suggests a further level of complexity in the control of ligand affinity and substrate ubiquitination by WWP2. An isoform almost certainly exists with only a portion of the HECT domain. The null ligase activity might allow the isoform to stabilise Smad7 levels by competing with WWP2-C and WWP2-FL, thereby preventing its ubiquitination. The EST database search provides evidence of an isoform that contains the WW3 and WW4 domains. Given the enhanced affinity of WW3 for Smad7 when expressed with WW4, this has interesting implications on the substrate selectivity of this isoform. If this new WW3-4 tandem domain isoform is part of the WWP2C-ΔHECT isoform, it would outcompete WWP2-C for Smad7 and prevent its degradation. Given the TGFβ-inducibility of this isoform, we postulate that WWP2C-ΔHECT is playing a role in the negative feedback loop of Smad7 (which is also TGFβ-inducible), by prolonging the survival of Smad7. This might allow full length E3 ligases, with equal or higher affinity, to associate and Smad7 and degrade TGFβ signalling components. If this is indeed the method of action of this WWP2 isoform, then the expression of this isoform in several different cancerous cell lines indicates that WWP2C-ΔHECT could play a role in overcoming the pro-apoptotic and cytostatic gene programs implemented by TGFβ stimulation. This system merits further exploration in the future.

6. Discussion

6.1 Discussion

The aim of this thesis was to probe the structure of WWP2 and provide an insight in to the function of this E3 ligase in the TGFβ signalling pathway. The initial attempts at elucidating the structure were focused around crystallisation of WWP2 isoforms (Chapter 3). WWP2-C was examined in particular, because the expression and purification gave high yields. When WWP2-C failed to crystallise, a new construct tailored to the specific demands of crystallisation was designed around the HECT domain, whereby a minimal HECT sequence excluding disordered regions was used. When this returned negative results in the crystallisation trials, a choice was made between spending more time on an approach that might eventually fail to yield a structure, or to try and tackle the expression problems with the protein interaction WW domains, and employ NMR to determine the structure. Using the GB1 solubility enhancement tag, high yields of the WW4 domain were successfully expressed and purified. This was expressed in isotopically enriched minimal media so as to label the protein with ^{15}N and ^{13}C isotopes. Using multidimensional NMR, the structure of WW4 was determined (Chapter 4). The WW4 domain was shown to form a three stranded β-sheet, the canonical WW domain formation.

6.1.1 WWP2 WW4 and other NEDD4 family member WW domains

Of the current NEDD4 family WW domain structures deposited in the PDB, the WW2 domain of HECW1 aligns with the WWP2 WW4 domain the best, with an RMSD of 1.031 Å (Table 6.1.1). The most striking feature of Table 6.1.1 is the similarity between WW domains, and suggests that their method of substrate selection is subtle. Surprisingly, the WW4 domain of WWP1, which is closely related and shares 78.38% identity, aligned with the lowest RMSD at 3.448 Å (Qin et al. 2007). Looking more closely at this structure; it is unpublished and appears to be a poor model. It required the calculation of 1000 models before submission of only 10 conformers (100-300 models are typically calculated, and 20-30 conformers are commonly submitted). When the structure was submitted to the iCING structural validation server, the red-orange-green ranking, which gives a broad overview of the health of the structure, scored the 39 residue model as 67% red, 28% orange and only 5% of the residues as green (only 2 residues). The packing quality was ranked as poor and the Ramachandran plot appearance was ranked as bad. From these

results, it was concluded that the WWP1 WW4 structure is unreliable, and despite its high sequence similarity with WWP2 WW4, should not be compared with the WWP2 WW4 structure.

	PDB ID	Method	RMSD (Å)
WWP1 WW4[1]	2OP7	Solution NMR	3.448
SMURF2 WW2-3[2] **(WW2)**	2KXQ	Solution NMR	2.775
SMURF1 WW1[3]	2LAZ	Solution NMR	2.433
SMURF1 WW2[4]	2LB1	Solution NMR	2.226
ITCH WW3	1YIU	Solution NMR	2.170
SMURF1 WW1[5]	2LB0	Solution NMR	2.132
NEDD4-1 WW3[6]	2M3O	Solution NMR	1.904
NEDD4L WW2[2]	2LTY	Solution NMR	1.803
NEDD4L WW3[7]	2LAJ	Solution NMR	1.805
NEDD4L WW1	1WR3	Solution NMR	1.899
NEDD4L WW2[8]	2LB2	Solution NMR	1.799
ITCH WW4	2YSF	Solution NMR	1.792
NEDD4L WW3[9]	2MPT	Solution NMR	1.763
NEDD4L WW2	1WR4	Solution NMR	1.712
NEDD4-1 WW3[10]	4N7H	X-ray diffraction	1.637
SMURF1 WW2[2]	2LTX	Solution NMR	1.513
ITCH WW2	2DMV	Solution NMR	1.448
SMURF2 WW2-3[2] **(WW3)**	2KXQ	Solution NMR	1.447
NEDD4-1 WW3	5AHT	Solution NMR	1.444
SMURF2 WW3[2]	2LTZ	Solution NMR	1.356
NEDD4L WW3	1WR7	Solution NMR	1.260
NEDD4-1 WW3	4N7F	X-ray diffraction	1.169
HECW1 WW2	3L4H	X-ray diffraction	1.031

Table 6.1.1 NEDD4 family WW domain structures (homo sapiens) deposited in the PDB, and the RMSD scores from their alignment with the most representative WW4 structure (model 3), performed in PyMol (across all WW domain atoms). 1 1000 conformers calculated, 10 submitted. 2 In complex with a Smad7 peptide. 3 In complex with a mono-phosphorylated Smad1 peptide. 4 In complex with a Smad1 peptide. 5 In complex with a di-phosphorylated Smad1 peptide. 6 In complex with an α-ENaC peptide. 7 In complex with a di-phosphorylated Smad3 peptide. 8 In complex with a mono-phosphorylated Smad3 peptide. 9 In complex with a HECT domain peptide. 10 In complex with an ARRDC3 peptide.

The HECW1 WW2 domain (Walker et al. 2010) shares 51.35% sequence identity with WW4 (Figure 6.1.1A). Interestingly, both HECW1 and WWP2 interact with the Wnt signalling molecule, Dishevelled (Dsh) (Mund et al. 2015; Miyazaki et al. 2004). The PPxY motifs of Dsh and Smad7 are as follows:

```
Smad7 203 ELESPPPPYSRYPMD
Dsh   520 PPPCFPPAYQDPGFS
```

Dsh has a proline rich region N-terminal to the PPxY motif, and is dissimilar throughout the rest of the sequence, besides the PPxY region. This suggests that the mode of binding, in terms of preferred residue contacts, is different between these two substrates

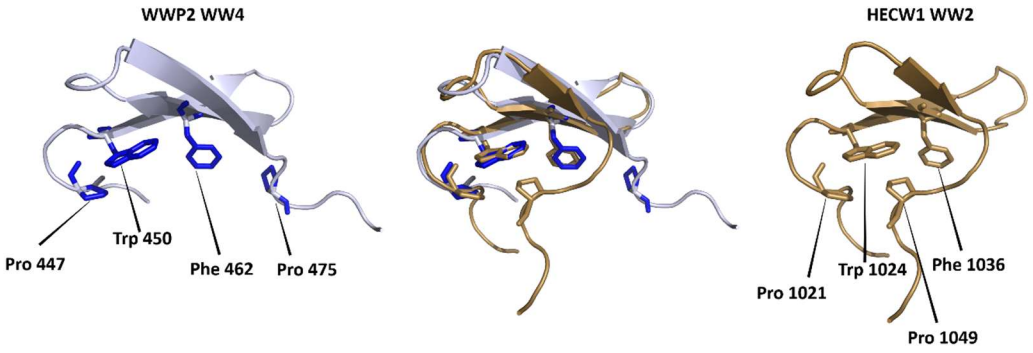
The HECW1 WW2 domain structure is from a crystal of a portion of HECW1 that contains the WW2 domain and a stretch of residues N-terminal to the WW domain that form a helical-box. The third WWP2 WW4 model was used for the alignments because it was ranked as the model closest to the mean structure by iCING. When the structure of this model is compared to the HECW1 structure, the most significant difference is the assembly of the hydrophobic underside. Four residues of HECW1 WW2 stabilise the fold by forming a hydrophobic core (proline 1021, tryptophan 1024, phenylalanine 1036 and proline 1049), shown in Figure 6.1.1B. The third β-strand is short, at three residues, and the loop that follows folds underneath the sheet, allowing proline 1049 to make hydrophobic contacts with the hydrophobic core. The equivalent residues in WW4 are proline 447, tryptophan 450, phenylalanine 462 and proline 475. Proline 447, tryptophan 450 and phenylalanine 462 are in virtually identical orientations, but proline 475 does not fold underneath the β-sheet after the third strand, and the loop which holds the proline extends away from the fold (Figure 6.1.1B). The third β-strand of WW4 is longer, at 6 residues in model 3, but consistently between 6-4 residues across the different models and always starting one residue earlier than the HECW1 β-strand (Figure 6.1.1A). This is probably the result of a stabilised β-sheet caused by the extra threonine at position 469 of WW4.

As a result of the extended β-strand, the position of phenylalanine 472 of the XP binding pocket is orientated differently and is almost parallel to the second residue of the XP binding pocket, tyrosine 461 (Figure 6.1.1C). HECW1 WW2 also has the atypical phenylalanine at this position (phenylalanine 1046), but it is rotated anticlockwise in a more open configuration to become almost perpendicular to the second residue of the XP binding pocket, which is another phenylalanine in this case (phenylalanine 1035), as shown in Figure 6.1.1C. The positioning of phenylalanine 472 of WW4 is closer to the orientation of the tryptophan side chain of WW domains that have a canonical tryptophan at this position. However, this is a source of some variation amongst the ensemble structures, and some of the models hold a conformation with phenylalanine 472 in a similar position to that of HECW1, notably models 14, 16 and 19.

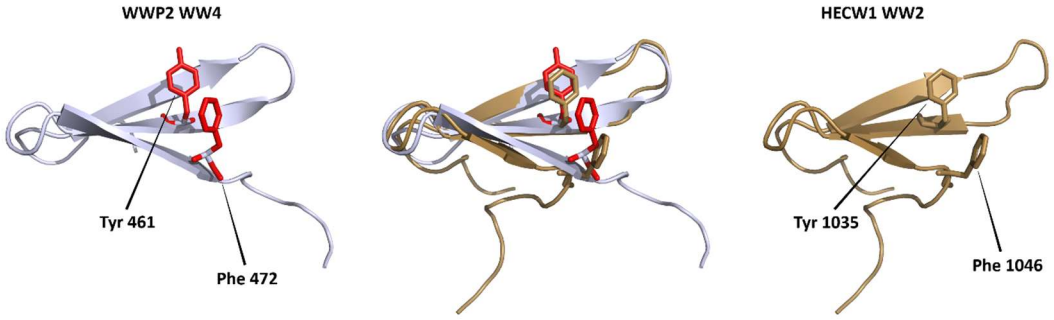
A

```
WWP2 WW4     443    EPALPPGWEMKYTSEGVRYFVDHNTRTTTFKDPRPGF
HECW1 WW2    1017   RLELPRGWEIKTDQQGKSFFVDHNSRATTFIDPRIPL
                    .  ** ***:*  .:*  :*****:*:*** ***  :
```

B



C



D

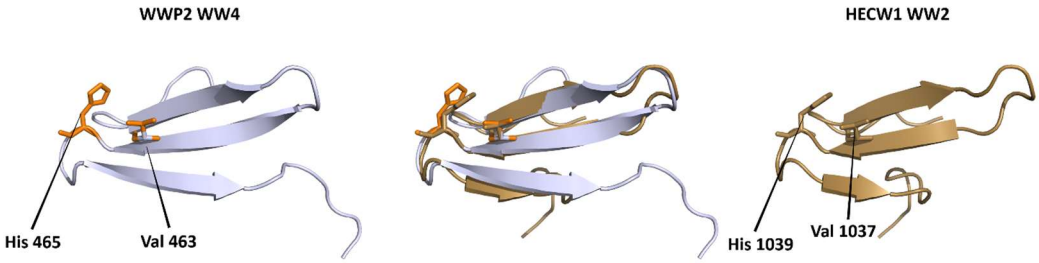


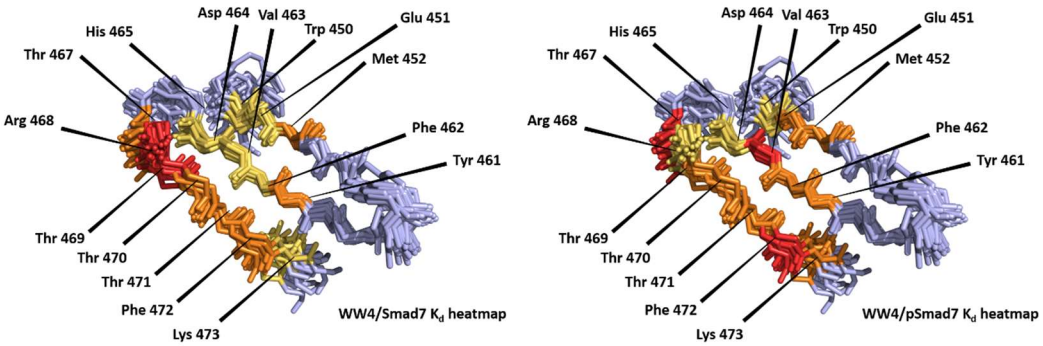
Figure 6.1.1 - A: WWP2 WW4 and HECW1 WW2 sequence alignment, with β-strands highlighted in green, and the variable region of the third WW4 β-strand highlighted in blue. B: Structural alignment of WWP2 WW4 in white and HECW1 WW2 (PDB: 3L4H) in brown, showing the residues that form the hydrophobic underside (Walker et al. 2010). C: Structural alignment showing the XP binding pocket residues of WWP2 WW4 and HECW1 WW2. D: Structural alignment showing the secondary specificity pocket residues of WWP2 WW4 and HECW2 WW2.

Valine 463 of the secondary specificity pocket is tightly restrained across the ensemble and it is in the same orientation as valine 1037 of the HECW1 WW2 domain structure (Figure 6.1.1D). Histidine 465, the second residue of this binding pocket, is tilted back slightly in model 3 when compared to histidine 1039 of HECW1 WW2. This position is highly variable across the different WW4 conformers which might be indicative of its diminished role in substrate binding observed during the titration experiments.

6.1.2 WWP2 WW4 Smad7 ligand interaction

In the introductory section of Chapter 5, several WW domain binding interactions were discussed, and key residues were outlined. The WW2 NEDD4L interaction with Smad7 involves the coordination of the ligand tyrosine from the PPxY motif by the valine and histidine of the secondary specificity pocket, and a threonine and arginine located a few residues C-terminal from the histidine (Aragón et al. 2012). These are all residues involved in the binding of Smad7 by the WW4 domain of WWP2 (valine 463, histidine 465, threonine 470 and arginine 467), in particular arginine 468, which has the tightest K_d of all WW4 residues when binding the Smad7 peptide (Figure 6.1.2A). Histidine 465 seems to play a minimal roll in binding, which might lower the overall K_d of the interaction. Similar to other WW domains, phenylalanine 472 and tyrosine 461 of the XP pocket, and threonine 470 are involved in binding, and most likely coordinate the N-terminal prolines of the PPxY motif, as in NEDD4L WW2 (Aragón et al. 2012). Threonine 470 had by far the most significant chemical shift perturbation in each titration, as was reflected by the trajectory of the amide peak in the HSQC. If the mode of binding is similar to NEDD4L WW2, as it appears to be, this might be the result of making contacts with the prolines of the PPxY motif, as well as the tyrosine. Glutamic acid 205 of the Smad7 peptide makes contact with an arginine of NEDD4L in loop1, at the same position as valine 459 (highlighted in yellow in Figure 6.1.2B), and it is this arginine and this interaction that is common amongst many WW domains that bind Smad7 with a high affinity. There is no indication of any chemical shift perturbation in this area in the WWP2 WW4 Smad7 titration, indicating that the ligand is not interacting with this area.

A



B

Smad7: 203 - **ELESPPPPYSRYPMD** - 217

```
WWP2 WW4     443    EPALPPGWEMKYTSEGVRYFVDHNTRTTTFKDPRPGF
NEDD4L WW2   364    TPGLPSGWEERKDAKGRTYYVNHNNRTTTWTRPIMQL
                     *.** *** :  ::*  *:*:**.****:. *   :
```

Figure 6.1.2 - A: Backbone of the WWP2 WW4 20 model CNS ensemble with the K_d heatmaps for the Smad7 and phosphoSmad7 titrations (also shown in Figure 5.3.1). B: The Smad7 peptide and the alignment between WWP2 WW4 and NEDD4L WW2.

The Smad7 peptide forms an extended hairpin in the NEDD4L WW2 binding pocket and the portion of sequence C-terminal to the PPxY motif makes extensive contacts with the first and second β-strands. WWP2 WW4 binding seems to differ here and the binding region terminates at methionine 452. Based on other WW domain binding orientations, and on the trajectory and K_d data, it is most likely that the Smad7 ligand binds WW4 with its N-terminus towards the end of the third WW4 β-strand, and then making contacts with residues along the third and second strands the ligand appears to turn as it reaches the second loop of WW4. When the peptide performs the turn, it most likely travels in the opposite direction back on itself, making contacts with residues of the first β-strand and probably making further contacts with residues of the second β-strand. Instead of forming an extended hairpin like NEDD4L WW2, it is likely that the C-terminal region of the peptide leaves the binding pocket when it reaches methionine 452. This configuration is similar to that of SMURF1 WW2 and Smad7, which is shown in Figure 6.1.3 (Aragón et al. 2012).

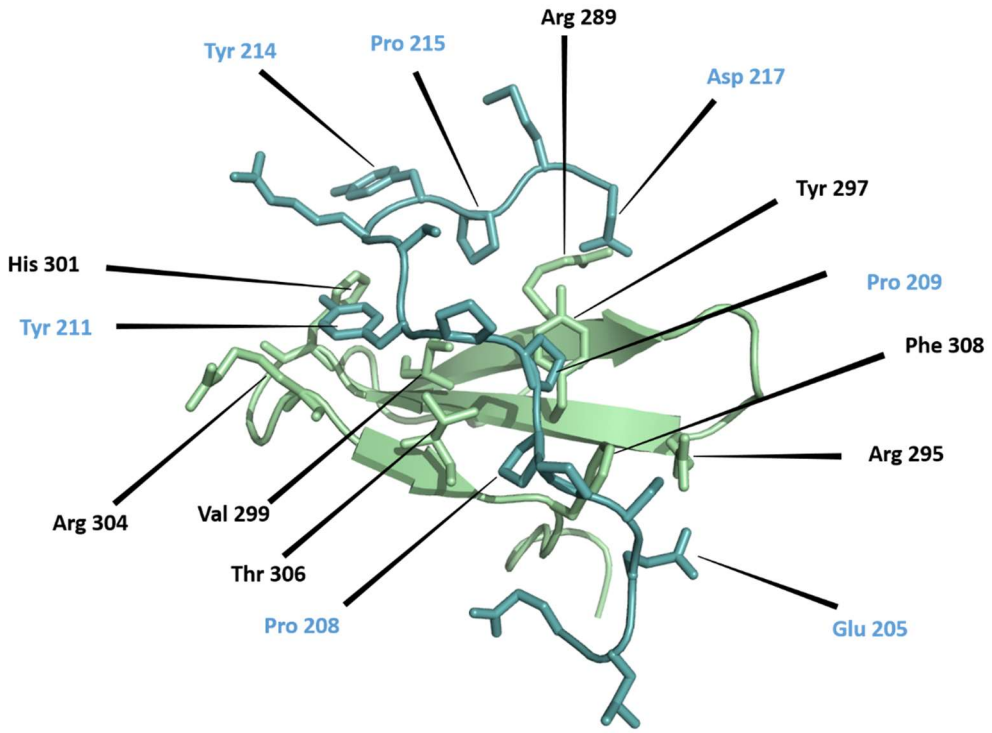
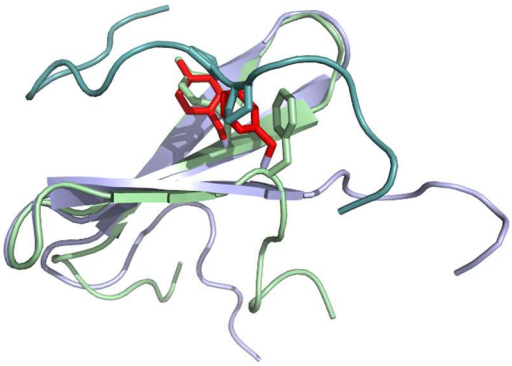


Figure 6.1.3 - The structure of the SMURF1 WW2 domain (PDB: 2LTX) in complex with the Smad7 ligand (Aragón et al. 2012). Key binding residues are labelled in blue for Smad7 and in black for the WW domain.

The interaction between SMURF1 WW2 arginine 295 and Smad7 glutamic acid 205, described above as being common amongst Smad7-binding domains, is maintained. Proline 208 of the ligand (the first proline of the PPxY motif) is stacked face to face with phenylalanine 308 of the XP binding pocket. Unfortunately, there is no unbound SMURF1 WW2 structure deposited in the PDB, so it is hard to tell how much the conformation changes upon ligand docking, but when compared to the side chain position of phenylalanine 472 in model 3 of WW4, it seems like the proline 208 conformation would clash (Figure 6.1.4A). However, when comparing the structures of the NEDD4L WW2 domain in the Smad7 bound and unbound configurations, ligand docking appears to cause tryptophan 393 at this position to rotate and tilt backwards (Figure 6.1.4B) (Aragón et al. 2012; Kowalski et al. 2005).

A



B

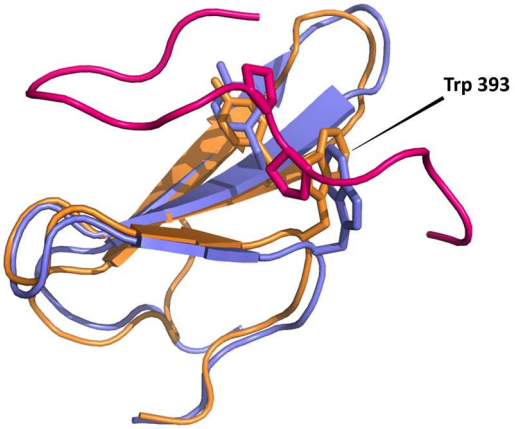


Figure 6.1.4 - A: Structural alignment of the Smad7-bound SMURF1 WW2 domain in green (PDB: 2LTX), with the WW4 domain in white, with residues of the XP binding pocket in red. B: Structural alignment of the NEDD4L WW2 domain in the unbound (orange) (PDB: 1WR4) and bound (purple) (PDB: 2LTY) conformations (Kowalski et al. 2005; Aragón et al. 2012).

Rotation of the proline helix in the SMURF1 WW2 binding pocket allows proline 209 to make hydrophobic contacts with the XP tyrosine 297 (Figure 6.1.3). Proline 210 is facing away from the binding surface, but the backbone makes contact with threonine 306. The equivalent position in WW4 is threonine 470, and the titration suggests it is significantly involved in binding. Tyrosine 211 of Smad7 sits within a hydrophobic pocket formed by valine 299, histidine 301 and arginine 304 in the SMURF1 WW2 complex (Figure 6.1.3). These residues are conserved in WW4 and occupy similar positions (Figure 6.1.5A).

The WW4 titration data indicate that there are further interactions along the first β-strand. This is consistent with many WW domain interactions, including the SMURF1 Smad7 complex, but the type of interaction appears to be different. The extended recognition motif of SMURF1 WW2 includes histidine 301 which contacts Smad7 tyrosine

214, arginine 289 which contacts Smad7 proline 215 and Smad7 aspartic acid 217 (Figure 6.1.3). Contact of Smad7 aspartic acid 217, either at the same position as arginine 289 in SMURF1 WW2, or further along the sequence in loop 1, is another feature of WW domains that bind Smad7 with high affinity. WWP2 WW4 has a lysine (453) at the position of arginine 289 (Figure 6.1.5A), so binding of Smad7 aspartic acid 217 might be expected, however, based on the trajectory data, this does not appear to be the case. In fact, none of the residues involved in the extended recognition motif of SMURF1 WW2 have significant changes in shift in the titration. Instead, the extended recognition region might resemble something similar to that of YAP WW2, in which Smad7 tyrosine 214 (which projects away from the binding site in SMURF1 WW2) is accommodated by contacts with tyrosine 247, isoleucine 249 and glutamic acid 237 (Figure 6.1.5B) (Aragón et al. 2012).

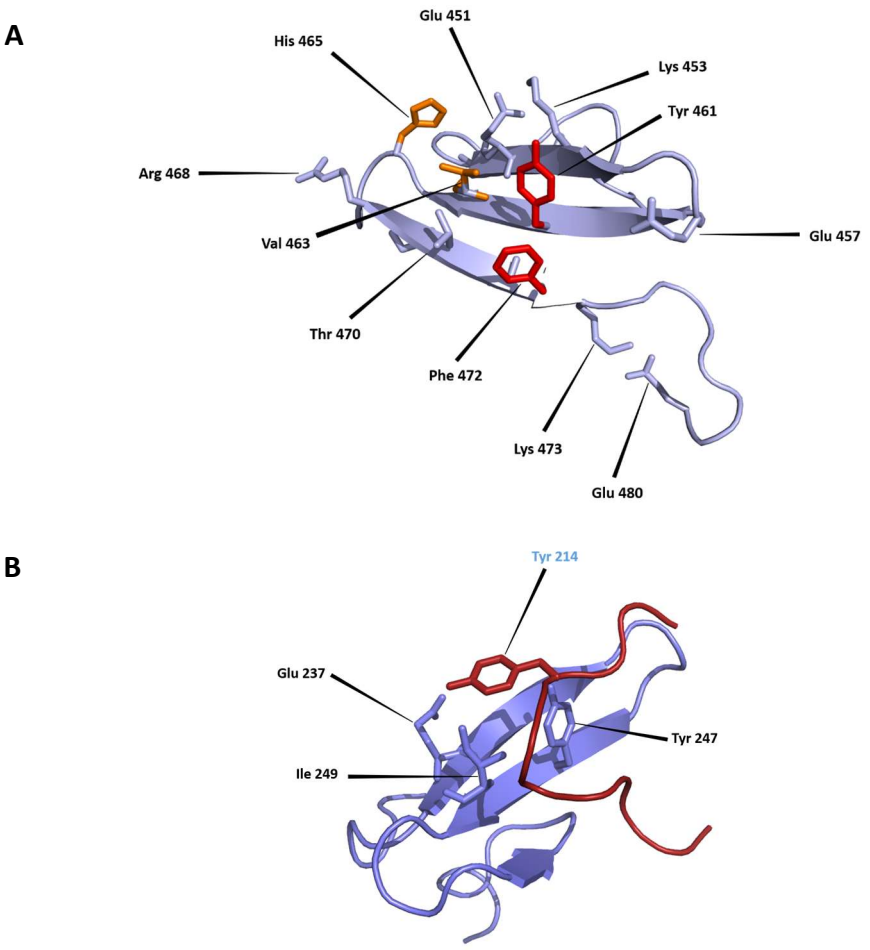


Figure 6.1.5 - A: The structure of WWP2 WW4 with key residues labelled. The XP binding pocket residues are shown in red and the secondary specificity pocket residues are shown in orange. B: The structure of the YAP WW2 domain (PDB: 2LTV) in complex with the Smad7 ligand, showing components of the extended recognition pocket (Aragón et al. 2012).

The equivalent positions in WW4 are tyrosine 461, valine 463 and glutamic acid 451 (Figure 6.1.5A), all three of which participate in ligand binding, as determined by the peak trajectory data. The apparent lack of interaction between lysine 453 of WW4 and Smad7 aspartic acid 217 is curious, given the complementary charges and likely proximity of these residues. A closer look at WW4 lysine 453 reveals that it seems to form a salt bridge with glutamic acid 451 (Figure 6.1.6A), and it appears as though Smad7 aspartic acid 217 is unable to displace this interaction. This interaction is present in all but four of the ensemble of WW4 structural models. The equivalent residues in NEDD4L WW2 are glutamic acid 372 and arginine 374. In the unbound structure, these residues do not form a salt bridge and are free to engage in electrostatic interactions with the ligand (Figure 6.1.6B).

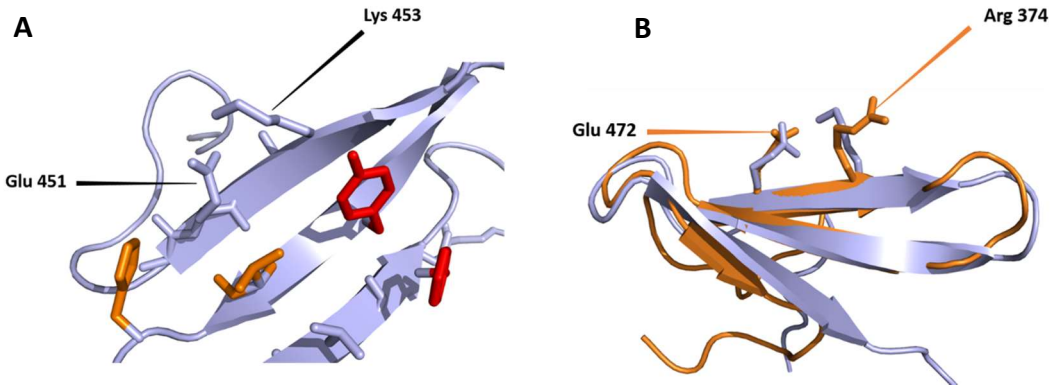


Figure 6.1.6 - A: The salt bridge between glutamic acid 451 and lysine 453 on the first β-strand of WWP2 WW4. The XP pocket is shown in red and the secondary specificity pocket is shown in orange. B: Structural alignment between NEDD4L WW2 in orange (PDB: 1WR4) and WW4 in white. The labelled residues are the equivalent residues of NEDD4L which do not form a salt bridge.

Given the lack of interactions with negative N and C-terminal residues of the Smad7 ligand with positive residues of the first β-strand and loop 1 of WW4, it would be unsurprising if a comparable K_d from ITC was higher (lower affinity) than those of the SMURF1 WW2, SMURF2 WW3, NEDD4L WW2 and other WW domains that exhibit a tight affinity. Glutamic acid 457 in loop 1 of WW4 (Figure 6.1.5) might serve to reduce the affinity further by actively repelling the acidic residues Smad7 glutamic acid 205 and aspartic acid 217, which are typically found to interact with an arginine in this position. Mutational analysis of SMURF1 WW2 and SMURF2 WW3 found that exchanging the arginine at this position (SMURF1 WW2 arginine 295, SMURF2 WW3 arginine 312) with a glutamic acid reduced the affinity of SMURF1 WW2 for Smad7 12-fold and SMURF2 WW3

8-fold (Aragón et al. 2012). Within the context of WWP2 activity, interaction and ubiquitination of Smad7, or if being used as a scaffold, its binding partners, is most likely restricted.

6.1.3 WWP2 WW4 phospho-Smad7 ligand interaction

During functional assays, WWP2-C ubiquitination of Smad7 was TGFβ-dependent, indicating that kinase activity might regulate Smad7 degradation. Accordingly, phosphorylation at serine 206 of the peptide enhanced the affinity between WW4 and its Smad7 ligand, and seemed to switch on the interaction. The exact cause of this increase in affinity is thus far unknown. The peak trajectories showed no significant differences that might be indicative of an interaction with the phosphate group, but the K_d data suggests that lysine 473 might be responsible, because of its significantly enhanced affinity. The co-structure of a phospho-Smad3 ligand and the second NEDD4L WW domain has been solved, and it shows that the phosphate group of Smad3 threonine 179 (the equivalent position of Smad7 serine 206) is coordinated by two basic residues (lysine 378 and arginine 380) of NEDD4L WW2 located in loop 1 of the β-sheet (Figure 6.1.7) (Aragón et al. 2011).

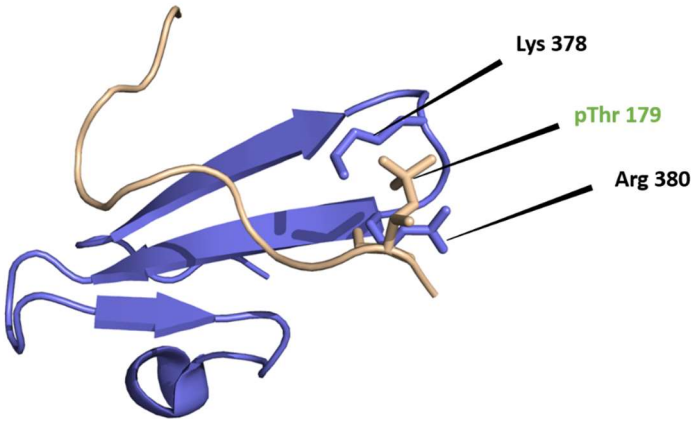


Figure 6.1.7 - The structure of NEDD4L WW2 bound to a mono-phosphorylated Smad3 ligand (PDB: 2LB2), showing residues involved in binding phospho-threonine 179 of the Smad3 ligand (labelled in green) (Aragón et al. 2011).

The equivalent residues in WW4 are glutamic acid 457 and valine 459, the former of which is in the exact position where the Smad7 serine 206 phosphate group would be orientated if the phospho-Smad7 ligand occupied an equivalent orientation. Figure 6.1.8A

shows the predicted surface charge distribution of WW4. Figure 6.1.8B shows the position of the phosphate group when the pSmad3 bound structure is aligned with the WW4 structure, this area has a negative patch due to glutamic acid 457. It is expected that the glutamic acid repels the electron dense phosphate group (as it would Smad7 glutamic acid 205), and in the absence of another interaction to compensate, decreases the affinity for the ligand. Glutamic acid at this position suggests that WW4 might have some specificity for PPxY ligands with basic residues in their N-terminal sequence.

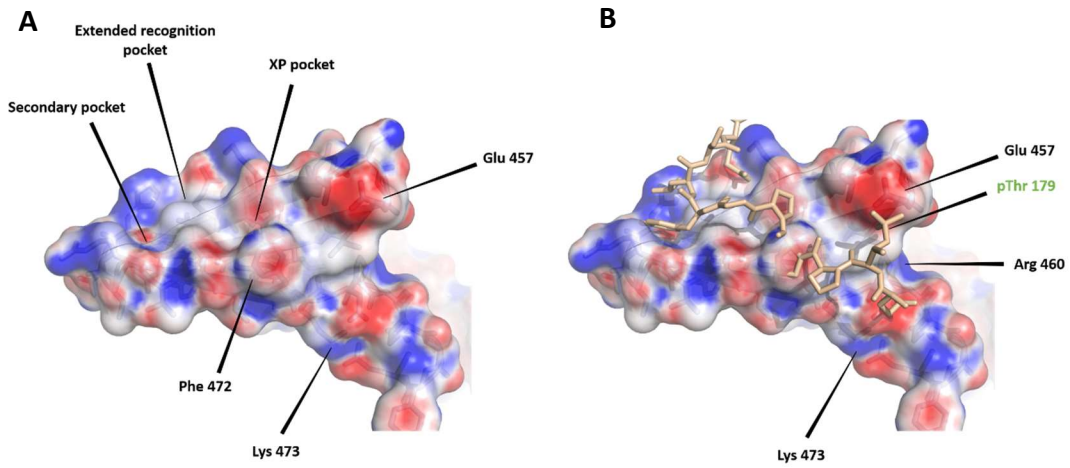


Figure 6.1.8 - A: The surface charge distribution of WWP2 WW4, predicted using the Adaptive Poisson-Boltzmann Solver (APBS) PyMol plugin (Dolinsky et al. 2004; Dolinsky et al. 2007). Negative patches are shown in red and positive patches are shown in blue. B: The alignment of phospho-Smad3 from the NEDD4L WW2 co-structure, with the WW4 structure showing the surface charge distribution.

There are some positive residues nearby that might accept the phosphate group, notably, arginine 460 and lysine 473. Arginine 460, although close to the position of arginines in other Smad7-binding WW domains, seems to have, through the course of evolutionary diversion, found itself one residue further toward the C-terminus, and therefore on the opposite side of the β-sheet. The distance from the phosphate group is reasonable and the angle seems unfavourable. The titration data also seems to discount arginine 460 as a binding site for the phosphate group. There are several indications that suggest lysine 473 is the key residue involved in binding the phosphorylated serine 206. These are: the K_d data, the proximity to serine 206 and the positive charge of this region predicted by APBS. However, the orientation does not seem to be ideal, and looking at lysine 473 in closer detail shows that in the unbound conformation it appears to form a salt bridge with the most C-terminal residue, glutamic acid 480 (Figure 6.1.9).

The criteria given for salt bridges are that the approximate centre of the charge in oppositely charged residues are within 4.0 Å of each other, and that at least one aspartic acid or glutamic acid side chain carboxyl oxygen atom is within 4.0 Å of the side chain nitrogen of the arginine, histidine or lysine (Kumar & Nussinov 1999). Both side chain oxygen atoms of glutamic acid 480 are within 4.0 Å of the lysine 473 side chain nitrogen (3.3 Å and 2.7 Å), indicating that this is indeed a salt bridge. In the other models of the ensemble, glutamic acid 480 is not always engaged in the same salt bridge, and instead, is also shown to be interacting with arginine 476, as well as nearby amide groups in different models. In some of the models, lysine 473 interacts with its neighbour, aspartic acid 474. Indeed, this region of the structure does not show much consensus. This is likely to be due to the lack of restraints, as there is not more than 4 inter-residue NOE restraints per residue for this loop region (Figure 4.2.3). The orientation of lysine 473 and glutamic acid 480 side chains in Figure 6.1.9 is therefore, not a result of NOE restraints, but instead, is a result of energetic minimisation by the modelling software. This does not necessarily preclude the formation of this salt bridge. The distance restraints are determined from NOESY spectra by observing the intensity of NOE cross peaks from methyl groups. So, although the acidic and basic functional groups of lysine 473 and glutamic acid 480 are within range for NOE detection, no such peaks would be visible, because of the lack of methyl groups. The nearest methyl groups are the lysine Cε group and the glutamic acid Cɣ group. The closest distance between protons of these groups in the model 3 structure is 5.2 Å. This is on the limit for NOE detection, which typically detects protons separated by distances of 5.0 Å and shorter (Wüthrich 1990). In addition, because of the way the experiment used to assign the proton resonance works, the methyl proton resonances of the residue at the C-terminus are not assignable, which means that no peaks are assignable to glutamic acid 480 in the carbon NOESY. As such, this salt bridge could be a feature of the WW4 structure, but the lack of restraints could be the reason it is not a consistent feature in the ensemble models.

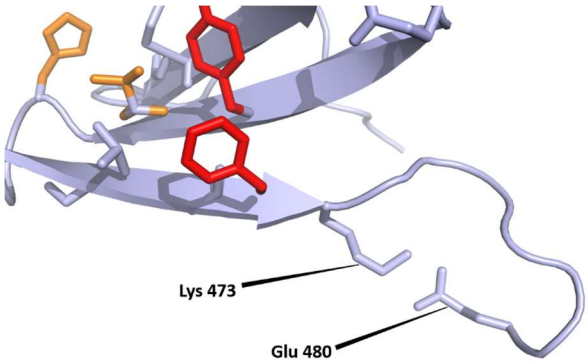


Figure 6.1.9 - The structure of WWP2 WW4 with the C-terminal salt bridge between lysine 473 and glutamic acid 480 shown. The XP pocket is shown in red and the secondary specificity pocket is shown in orange.

In the phospho-Smad7 titration, WWP2 WW4 glutamic acid 480 has a significantly large trajectory, and this might be the result of disruption of the salt bridge by the phosphate group. Whether or not this interaction is an artefact of the recombinant protein is uncertain, particularly as it involves the most C-terminal residue. However, if this orientation is maintained in the native protein, then some interesting possibilities are raised. Smad2, Smad3 and Smad7 all have glutamic acids N-terminal to the PPxY motif. Should the WWP2 WW4 domain have unquenched basic residues either in loop 1 like other WW domains, or C-terminal to the XP binding pocket, then WW4 might have Smad affinities similar to other WW domain family members. The acidic residue in loop 1 in WW4 ensures there is no interaction here, and the interaction between lysine 473 and glutamic acid 480 might prove too stable for efficient disruption by the Smad glutamic acids. But should a phosphate group be introduced to the N-terminal region of the PPxY motif, this might displace the salt bridge and activate the affinity of the ligand for the receptor. In this scenario, the acidic/basic residue salt bridge would act as an 'affinity gate' that has to be opened in order to enhance affinity across the binding site, essentially increasing the activation energy of the interaction. Coupled with the glutamic acid in loop 1, this could ensure selectivity of phosphorylated substrates over non-phosphorylated substrates. In the context of WWP2-C ubiquitin ligase activity against Smad7, non-phosphorylated Smad7 is preserved until the activation of a kinase (possibly as a result of TGFβ stimulation), phosphorylates serine 206, causing the recruitment of WWP2-C and subsequent degradation by the ubiquitin-proteasome system. As a result, TGFβ signal propagation would be enhanced in the absence or reduction of its Smad7-mediated

negative feedback loop, allowing activation of TGFβ-mediated gene programs such as cytostasis, apoptosis, angiogenesis and epithelial-mesenchymal transition.

Disruption of salt bridges by phosphorylation has been proposed to induce conformational changes in the retinoic acid receptor, and increase its affinity for its ligand (Chebaro et al. 2013). Phosphorylation at multiple sites of the cytoplasmic linker-associated protein 2 (CLASP2) has been proposed to decrease its affinity for its binding partner by the formation of stable intramolecular salt bridges between the phosphate groups and arginines of CLASP2 (Kumar et al. 2012). This prevents electrostatic interactions with its binding partner, and tunes the affinity of the interaction (Kumar et al. 2012). Both of these examples are from molecular dynamics simulations.

The function of the salt bridge between lysine 473 and glutamic acid 480 might also dictate the structure of the hydrophobic underside of WWP2 WW4, which appears to deviate from other WW domains. As described above, WW4 proline 475 does not appear to participate in hydrophobic contacts with the underside of the WW domain, as is common with proline of other WW domains at this position. As a result of their hydrophobic core configuration, residues at the corresponding position of lysine 473 in other WW domains (which are often threonines) are tucked further underneath the β-sheet, away from the binding ligand. The absence of proline 475 incorporation in to the hydrophobic core might, therefore, be necessary to position the salt bridge in order to carry out its function. Interestingly, the closely related WWP1 WW4 domain has a preserved lysine but has a serine at the same position of the glutamic acid, so this type of selectivity might be exclusive to WWP2 WW4 from the Smad7 binding NEDD4 family members. Since the Smad2/3 peptides also bound the WW4 domain in the titration experiments, this proposed mechanism has some importance, given the phosphorylation of the Smad3 threonine immediately preceding the PPxY motif (threonine 179, shown in Figure 6.1.7 and Figure 6.1.8). Phosphorylation at this site in Smad2 and Smad3 is dependent on CDK8/9 kinase activity driven by TGFβ pathway activation (Alarcón et al. 2009; Wang et al. 2009). In theory, this phosphorylation could also activate the affinity gate proposed above, and enhance the affinity of these ligands for WW4, which is low compared to the phosphorylated Smad7 ligand, despite the lack of interaction observed between Smad2/3 and WWP2-C in immunoprecipitation assays (Soond & Chantry 2011).

The presence of a similar salt bridge on the first strand of the β-sheet appears to prevent or reduce the interaction between lysine 453 of WW4 and Smad7 aspartic acid

217, as described above. This opens up the possibility that these residues operate with a similar 'affinity gate' mechanism. It is tempting to speculate that the serine directly C-terminal to the PPxY motif might accept a phosphate group that could disrupt this salt bridge. However, in the bound structures, serine 212 of Smad7 tends to be some distance from these residues. The most likely candidate is serine 186 of Smad3 which is a little closer or perhaps even Smad7 tyrosine 214, although there is, so far, no evidence of phosphorylation of these residues. The Smad2/3 titrations suggested that the affinity of Smad2 is lower for WW4 than Smad3. There are no bound Smad2/WW domain structures, but the bound phospho-Smad3/ NEDD4L WW2 domain structures suggests that the only residue that differs between these two peptides, leucine 185 in Smad3 (isoleucine 226 in Smad2), makes hydrophobic contacts with histidine 465 of the secondary specificity pocket. Histidine 465 has been troublesome in the majority of the titrations, with peak shift change patterns that do not fit a binding curve, the conclusion being that it does not participate significantly in binding. The structure of the side chain of leucine (Smad3) compared to isoleucine (Smad2) might allow a hydrophobic contact with a larger surface area and increase the affinity between WW4 and Smad3 compared to Smad2.

6.1.4 WWP2 WW3 Smad7 ligand interaction

During the titration of Smad7 with the third WW domain of WWP2, many of the peaks decreased in intensity upon titration of the ligand, indicating that the timescale of ligand exchange was such that it disrupted signal detection. This is known as intermediate exchange, and because the signal in the spectra were so weak to start with (due to low ligand concentrations), the decrease in signal intensity, which can be used to determine the K_d, could not be followed. Because of this, many of the residues that are involved in binding do not have dissociation constants. The remaining residues give some indication of the affinity of this domain for Smad7, and WW3 seemingly has a higher affinity than the WW4 domain. Out of the Smad7 binding NEDD4 family member WW domains that have had their structures solved, SMURF2 WW3 shares the highest identitiy at 62.16%, shown below in Figure 6.1.10 (second only to WWP1 WW3 of all of the Smad7 binding NEDD4 family member WW domains).

```
WWP2    WW3  404     LGPLPPGWEKRQD-NGRVYYVNHNTRTTQWEDPRTQGM
SMURF2  WW3  296     LGPLPPGWEIRNTATGRVYFVDHNNRTTQFTDPRLSAN
                     ********* *:  .****:*:**.****: *** ..
```

Figure 6.1.10 - Sequence alignment between the third WW domain of WWP2 and the third WW domain of SMURF2, with β-strands highlighted in green. The WWP2 WW3 domain has been colour coded to indicate the extent of residue trajectory distances, residues in blue appear to be in intermediate exchange.

WWP2 WW3 has the canonical XP binding pocket tryptophan at residue 432 whereas SMURF2 WW3 has a phenylalanine, but other than this, key residues in Smad7 binding are conserved between these two domains. In particular, the secondary specificity pocket is identical (WW3 histidine 425, valine 423), arginine 428 which participates in WW4/Smad7 binding is conserved. The residue involved in binding the acidic N-terminal residue of Smad7 is conserved; this is arginine 419 (SMURF2 WW3 arginine 212), which is at the position of the first residue of the second β-strand of SMURF2 WW3, and binds Smad7 glutamic acid 203. The arginine involved in binding the C-terminal aspartic acid (217) of Smad7 in SMURF1 WW2 is conserved; arginine 414 (SMURF2 WW3 arginine 306 midway through the first β-strand).

The indications are that, where WW4 has residues that might not engage in, or actively repel the binding of the Smad7 ligand, WW3 has residues that are conducive towards binding Smad ligands; Smad ligands are characterised by their PPxY motifs flanked by acidic residues that seem to be the determinates of WW domain ligand affinity. Arginine 414 is involved in binding, and could bind Smad7 aspartic acid 217 in the same fashion as SMURF1 WW2 (Figure 6.1.3). This is the position of lysine 453 in WW4, which forms a salt bridge with glutamic acid 451 and does not participate in ligand binding. WW3 also has a glutamic acid at the equivalent position (412), but the arginine/glutamic acid pair does not form a salt bridge in NEDD4L WW2 (Figure 6.1.6). This electrostatic interaction is likely to enhance the affinity of Smad7 for WW3, and might be responsible for the increase in affinity when compared to WW4. Surprisingly, valine 423 of the secondary specificity pocket does not experience a significant chemical shift perturbation, and neither does arginine 419 which binds the N-terminal Smad7 glutamic acid 203 in other WW domain structures.

	Smad7	pSmad7	Smad2	Smad3	Smad7 (SUMO)
WW4	820±198 µM*	440±119 µM*	973±307 µM	650±408 µM*	227±27 µM
WW3	-	-	-	-	160±59 µM
WW3-4	-	-	-	-	214±214 µM
WW3$^{(WW3-4)}$	-	-	-	-	23±25 µM
WW4$^{(WW3-4)}$	-	-	-	-	362±175 µM

Table 6.1.2 The dissociation constant of WW3, WW4 and the tandem WW3-4 domains, calculated in Chapter 5 (previously shown in Table 5.3.1). Dissociation constants with an asterisk are the product of a curve which was fit manually to the first stage of a two stage migration. Standard deviation is given as the K_d error.

6.1.5 Tandem WW domain Smad7 ligand interaction

When the titration was performed on the third and fourth WWP2 WW domains expressed in tandem, the global affinity was higher than that of the individual WW4 domain and lower than that of the individual WW3 domain (Table 6.1.2). It is expected that each WW domain binds one ligand each. When considering each individual WW domain K_d, the WW3 domain affinity seems to be dramatically increased (relative to the WW3 domain by itself), in the low micromolar range, similar to the WW domain affinities calculated by ITC (Aragón et al. 2012). The WW4 domain affinity, on the other hand, is slightly reduced. The pattern of WW3 domain peak trajectories is different (Figure 6.1.11). In particular, valine 423 of the secondary specificity pocket is now in intermediate exchange, and significantly, arginine 419 is now engaged in binding.

```
WW3    WW3   404    LGPLPPGWEKRQDNGRVYYVNHNTRTTQWEDPRTQGM
WW3-4  WW3   404    LGPLPPGWEKRQDNGRVYYVNHNTRTTQWEDPRTSAN
```

Figure 6.1.11 - The WWP2 WW3 domain sequence colour coded to indicate the extent of residue trajectory distances for the Smad7 titration for the individual WW3 domain and the WW3 domain expressed in tandem with WW4. Residues in blue appear to be in intermediate exchange.

WW4 domain affinities are universally reduced apart from tryptophan 450 and phenylalanine 462, which see their K_d reduced from 210 µM (poor fit) and 200 µM in the individual WW4 domain titration, to 70 µM and 80 µM respectively. These are residues

that form the hydrophobic core of the WW domain on the opposite side of the β-sheet. In the individual WW4 domain titrations, tryptophan 450 characteristically had a relatively uncoordinated pattern of peak migration and a satisfactory binding curve could not be fit. The fact that these residues have a relatively high affinity when compared to other residues of WW4, and that they form the hydrophobic core, indicates that they might be involved in dimerisation with the WW3 domain through hydrophobic contacts. This would appear to also explain some of the other changes in the trajectory patterns of WW3. For example, tyrosine 422 which is in the equivalent position of phenylalanine 462 and presumably forms the hydrophobic core of WW3, is in intermediate exchange in the tandem titration. The trajectory of the tryptophan 411 peak of WW3, which is in the equivalent position of tryptophan 450 of WW4, is greater. Based on these observations, the prediction is that the WW domains dimerise through hydrophobic interactions between their hydrophobic cores, similar to the SMURF WW domain 'β-clam' conformation, which are mediated by their hydrophobic underside surface, proposed here (Aragón et al. 2012). And that this interaction allows WW3 to bind Smad7 with greater affinity, possibly by allosteric alterations that optimise the position of arginine 419, to allow binding of Smad7 glutamic acid 203 at the N-terminal of the peptide, and as a result of this dimerisation, the affinity of WW4 for Smad7 is compromised.

6.1.6 Conclusions

There is an interesting picture emerging that relates the structures of WW3 and WW4 with their ligand affinities, which has significant implications on their function as part of their native proteins. The WWP2-C protein has previously been shown to selectively bind and degrade Smad7 (Soond & Chantry 2011). Smad7 is the inhibitory arm of the TGFβ signalling pathway that is upregulated by TGFβ stimulation and acts as a negative feedback loop by recruiting ubiquitin ligases to the TGFβ receptors. The TGFβ receptors are subsequently ubiquitinated, along with Smad7. These components are targeted to the proteasome by polyubiquitin chains that are assembled by the HECT domain of E3 ligases, which are supplied with ubiquitin monomers by the E2 conjugator enzymes. This prevents the activation of Smad2 and Smad3, which require phosphorylation of their C-termini by the TGFβ receptors in order to release their autoinhibition, and to carry out their gene regulatory activities. WWP2-C has been shown

in luciferase assays to enhance TGFβ signalling but, unlike other ubiquitin ligases that have been shown to associate with Smad7, WWP2-C does not cause the degradation of the receptors (Soond & Chantry 2011). Smad7 is targeted for degradation by WWP2-C in a TGFβ-dependent manner, and it seems now that Smad7 is earmarked for ubiquitination by a kinase that phosphorylates Smad7 near the PPxY motif. This kinase could either be under the control of TGFβ, or it could be the junction between two pathways, facilitating crosstalk. It could be related to the cell cycle, in the same way that Smad2/3 is phosphorylated by CDKs, and might allow TGFβ signalling to be prolonged in order to implement its cytostatic or apoptotic gene program without immediate intervention by the Smad7 negative feedback loop. Or perhaps the kinase is related to the EMT pathway, and facilitates the upregulation of genes involved in differentiation and migration. The evidence from tissue culture assays suggest that Smad7 ubiquitination by WWP2-C is significantly enhanced by TGFβ stimulation, and overexpression of WWP2-C increased the expression of Vimentin, a marker of EMT, during TGFβ stimulation (Soond & Chantry 2011). This suggests that Smad7 phosphorylation is performed by a kinase under the direct control of TGFβ. There are several alternative pathways that are activated by TGFβ, besides the canonical Smad pathway; these include MAP kinases, PI3K/AKT (phosphatidylinositol-3-kinase) and Rho-like GTPase signalling, and these could be a good place to start in the search of the responsible kinase (Zhang 2009). It should be noted, however, that the timescale used in the tissue culture assays was relatively long, at three hours, so protein translational effects cannot be ruled out, although Smad7 did co-immunoprecipitate with WWP2-C after 1 hour of stimulation with TGFβ.

It was previously assumed that the WWP2 WW4 domain was the principal domain involved in Smad7 binding by WWP2-FL, however, data here proves otherwise. The third WW domain now has to be considered the primary domain involved in recruiting Smad7. The WW4 domain is relegated to playing a supporting role by sacrificing its affinity for Smad7, in order to enhance the affinity of WW3 by interaction between the hydrophobic cores. The sacrifice of affinity by WW4 when it is in tandem in WWP2 might ensure phospho-Smad7 is not selected by tandem domain-containing isoforms. The WWP2C-ΔHECT isoform adds another layer of complexity to an already complex system. Exactly where the WWP2C-ΔHECT isoform fits in to this regulation system is dependent on where transcription starts. It is safe to say that the chances of this isoform being catalytically active are remote at best, and the function can be narrowed down to four possibilities; 1 - WWP2C-ΔHECT contains only one WW domain, WW4, and it preserves

Smad7 by competing with WWP2-C for phosphorylated Smad7; 2 - WWP2C-ΔHECT contains two WW domains, WW3 and WW4, and it competes with WWP2-FL for non-phosphorylated Smad7; 3 - WWP2C-ΔHECT activates other WWP2 isoforms; 4 - WWP2C-ΔHECT acts as a scaffold for other proteins. The expression pattern seems to suggest expression is induced by TGFβ, in some cell lines at least, and it might play a role in fine tuning the TGFβ response with inputs from the splice factor environment.

6.1.7 Future work

A lot of the work here requires further experimentation to clarify and confirm the results and theories presented in this thesis. There are three immediate priorities which are: Smad ligand titrations using the gold standard method, isothermal titration calorimetry; Confirmation of the N-terminus of WWP2C-ΔHECT and cloning of the full length isoform from RNA; WW domain structures in complex with Smad ligands to determine exact mode of binding.

ITC measures the energetics of a binding equilibrium by sensitively measuring the heat evolution from the specific binding of a ligand with its receptor. In order to calculate dissociation constants that can be confidently compared with those of other proteins commonly found in the literature, a comprehensive analysis of Smad/phospho-Smad affinities should be performed on not only WW4 and WW3, but also WW1 and WW2 and each of the combinations of tandem domains. By doing this, the ligand specificity and the function of each domain in the TGFβ pathway can be fully elucidated. Importantly, this will allow the affinities and function of WWP2 WW domains to be explained with full comparability with other NEDD4 family members and other Smad binding partners that use the WW domain binding interface. Ligand titration by NMR has been useful in exploring the WW domain binding site, which has revealed some interesting features. In particular, which residues are involved in binding and which residues are not, and which of those residues contribute most to the affinity of the interaction. To obtain similar data from ITC would require multiple rounds of mutational analysis and data collection, in order to identify which residues contribute to the interaction affinity. However, published affinities are almost exclusively acquired using ITC, and there is some hesitation here to directly compare dissociation constants acquired using different techniques. As such, this method should be used for the WWP2 WW domains as well.

When looking at possible termini for the WWP2C-ΔHECT isoform, evidence from the expressed sequence tag database suggested that a truncated HECT isoform could be the result of an early stop codon caused by the retention of intron 19/20. This was corroborated by the clear presence of a TGFβ-inducible transcript that retained intron 19-20 in several different cell lines. Molecular weight estimation from the western blots of this protein, and the cross-reactivity with the anti-WWP2-C antibody, indicated that this isoform was a truncated version of WWP2-C. However, during the course of the many purifications performed on WWP2 isoforms, it became clear that the molecular weight was anything but certain, and raised the possibility that the isoform was heavier than expected. Mass spectrometry would be useful to identify the region of WWP2 to which it corresponds, however, so far we have only observed the isoform on SDS-PAGE gels as a constituent of whole cell lysates. Immunoprecipitation could be employed to purify the isoform. Subsequent searches of the EST database revealed another transcript start site that would, if paired with the intron 19/20 retention, produce a protein within range of the estimated molecular weight and include the WW3 domain. The titrations revealed how critical this detail would be in determining the function of this isoform. So to begin to place this isoform in to the crowded room of TGFβ regulators, the first step must be to clone the full length transcript. Once the sequence is known, mammalian tissue based assays can be used to start to identify a role for this isoform. Luciferase assays using Smad reporters would be a good place to start, this would identify the effects of WWP2C-ΔHECT on the TGFβ gene program. Co-immunoprecipitation assays will help identify binding partners and supplement ligand affinity experiments.

There are several published NEDD4 family WW domain structures in complex with their ligands deposited in the PDB. These structures provide invaluable information about the interaction between two binding partners, particularly if the end purpose is to attempt to disrupt the interaction with a targeted therapeutic. In order to generate these kind of structures, both the WW domain receptor and the Smad ligand should be isotopically labelled so as to assign their respective resonances. NOE data can then be collected and used to determine ligand orientation. The challenge here is isotopically labelling the ligand, because the peptide sequence is so small, bacterial expression is often avoided. There are a few options besides expression. A synthetic peptide can be purchased which has been synthesised using ^{13}C and ^{15}N labelled amino acids. However, the cost of buying a synthetic peptide that is uniformly labelled with heavy carbon and nitrogen isotopes is prohibitively expensive. The second synthetic option is to buy

individual uniformly labelled amino acids and have the peptide synthesised using these amino acids. This option is still very expensive but there are economies of scale associated with buying the amino acids. This is, however, only beneficial if many labelled peptide experiments are planned. The Smad7 peptides used in this thesis include Smad7 from a bacterial source, having been expressed with a SUMO expression tag that cleaves without leaving excess residues from the cleavage site. The Smad7(SUMO) peptide was shown to successfully interact with WW3 and WW4. Because Smad7(SUMO) is from a bacterial source, it can be isotopically labelled with relatively minor expense. The WW4/Smad7 co-structure is poised to be solved, since the majority of WW4 resonance assignments can be transferred to the bound protein resonances, the only assignments that need to be performed thoroughly are for Smad7. Solving the bound structure will require an experiment not used here and that is the carbon-filtered NOESY. This experiment allows NOE peaks from ^{13}C labelled methyl group to be removed from the NOESY spectrum, and instead only ^{12}C methyl group NOEs are observed. Using this experiment, an unlabelled ligand can be observed from the perspective of the labelled receptor, giving distance restraints between the two. Determining the bound structure of WW4 and Smad7 can confirm some of the observations made from the titration data and provide a further insight in to ligand specificity.

There are several different approaches that can be taken to explore some of the other features observed here. The selectivity of the phosphorylated ligand over the non-phosphorylated ligand is a problem when considering bound structures, since incorporating a serine with a phosphate group is not possible with the bacterial expression system. There are bacterial expression systems that attempt to incorporate mammalian enzymes in to bacteria so as to create a protein with the correct post-translational modifications, but so far the kinase responsible for the tentative Smad7 phosphorylation is unknown. Firstly there should be further exploration of the viability of this site as a kinase target. It is possible to buy antibodies that have been raised against phosphorylated antigens, which would be a good starting point for analysis of Smad7 phosphorylation at serine 206. A mammalian cell system should be used to create cell lysates with TGFβ stimulation, and without it. These lysates should be probed with the phospho-specific antibody by western blot. Hopefully this will confirm phosphorylation of this site, and test the theory that TGFβ stimulation is responsible for the activity of this kinase. These experiments should be performed in the presence of proteasome inhibitors, so as to prevent the degradation of phosphorylated Smad7, seeing as the theory is that

turnover is increased by the significantly enhanced affinity of an E3 ubiquitin ligase for phosphorylated Smad7. This antibody could be used in conjunction with co-immunoprecipitation assays to determine if the specificity apparent in NMR titrations translates to a biological specificity. To explore the role of phosphorylation of serine 206 in the function of Smad7, this residue could be mutated to alanine and overexpressed in a mammalian cell system, alongside overexpression of WWP2-C. Using luciferase assays to monitor the effect on the TGFβ signalling. Based on the NMR data, this mutation should protect Smad7 from degradation by WWP2-C and decrease the amplitude of signalling, as compared to the non-mutated Smad7. The same mutation should reduce the amount of polyubiquitinated Smad7 species observed from western blot of cell lysates, from cells overexpressing Smad7 and WWP2-C, and treated with proteasome inhibitor. Likewise, Smad7 levels should be stabilised by this mutation in cells treated with Cycloheximide. In order to elucidate the kinase responsible for phosphorylation, there are several inhibitors available that target different groups of kinases, and can be used in conjunction with the phosphospecific antibody to narrow down the group of kinases responsible (no antibody cross-reactivity should be observed in western blots if the kinase is successfully inhibited). Once the group of kinases responsible has been determined, small interfering RNA can be used to knockdown expression and used in conjunction with the phosphospecific antibody to identify the specific kinase that phosphorylates Smad7 at this site.

To explore the 'affinity gate' mechanism proposed here, mutagenesis can be used again. Initially, alanine mutagenesis of WWP2-C itself at lysine 473 can be used in conjunction with luciferase assays, western blots and co-immunoprecipitation. If this mechanism functions as proposed, Smad2/3 reporter activity, Smad7 turnover and co-immunoprecipitation of Smad7 with WWP2-C should be reduced. Likewise, alanine mutagenesis of glutamic acid 480 should enhance Smad2/3 reporter activity, Smad7 turnover and co-immunoprecipitation of Smad7 with WWP2-C, by keeping the gate open. This will confirm the relevance of this mechanism in the context of the native protein, and exclude the possibility that the C-terminal salt bridge is an artefact of the recombinant protein. These mutations should also have a clearly observable affect in the Smad7/WW4 titration assays, and help to confirm the role of these residues in substrate selectivity.

References

Adams, J., 2003. The proteasome: structure, function, and role in the cell. *Cancer Treatment Reviews*, 29, pp.3–9.

Alarcón, C., Zaromytidou, A.-I., Xi, Q., Gao, S., Yu, J., Fujisawa, S., Barlas, A., Miller, A.N., Manova-Todorova, K., Macias, M.J., Sapkota, G., Pan, D. & Massagué, J., 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell*, 139(4), pp.757–69.

Annes, J.P., Munger, J.S. & Rifkin, D.B., 2003. Making sense of latent TGFbeta activation. *Journal of Cell Science*, 116(Pt 2), pp.217–24.

Aragón, E., Goerner, N., Xi, Q., Gomes, T., Gao, S., Massagué, J. & Macias, M.J., 2012. Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-β Pathways. *Structure (London, England : 1993)*, 20(10), pp.1726–36.

Aragón, E., Goerner, N., Zaromytidou, A.-I., Xi, Q., Escobedo, A., Massagué, J. & Macias, M.J., 2011. A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes & Development*, 25(12), pp.1275–88.

Banka, P.A., Behera, A.P., Sarkar, S. & Datta, A.B., 2015. RING E3-Catalyzed E2 Self-Ubiquitination Attenuates the Activity of Ube2E Ubiquitin-Conjugating Enzymes. *Journal of Molecular Biology*, 427(13), pp.2290–304.

Bartels, C., Xia, T.H., Billeter, M., Güntert, P. & Wüthrich, K., 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR*, 6(1), pp.1–10.

Beal, R., Deveraux, Q., Xia, G., Rechsteiner, M. & Pickart, C., 1996. Surface hydrophobic residues of multiubiquitin chains essential for proteolytic targeting. *Proceedings of the National Academy of Sciences of the United States of America*, 93(2), pp.861–6.

Bedford, M.T., Chan, D.C. & Leder, P., 1997. FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands. *The EMBO Journal*, 16(9), pp.2376–83.

Bedford, M.T., Sarbassova, D., Xu, J., Leder, P. & Yaffe, M.B., 2000. A novel pro-Arg motif recognized by WW domains. *The Journal of Biological Chemistry*, 275(14), pp.10359–69.

Berrow, N.S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., Stuart,

D.I. & Owens, R.J., 2007. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Research*, 35(6), p.e45.

Bhattacharya, A., Tejero, R. & Montelione, G.T., 2007. Evaluating protein structures determined by structural genomics consortia. *Proteins*, 66(4), pp.778–95.

Bierie, B. & Moses, H.L., 2006. TGF-beta and cancer. *Cytokine & Growth Factor Reviews*, 17(1-2), pp.29–40.

Blank, M., Tang, Y., Yamashita, M., Burkett, S.S., Cheng, S.Y. & Zhang, Y.E., 2012. A tumor suppressor function of Smurf2 associated with controlling chromatin landscape and genome stability through RNF20. *Nature Medicine*, 18(2), pp.227–34.

Block, H., Maertens, B., Spriestersbach, A., Brinker, N., Kubicek, J., Fabis, R., Labahn, J. & Schäfer, F., 2009. Immobilized-metal affinity chromatography (IMAC): a review. *Methods in Enzymology*, 463, pp.439–73.

Bonni, S., Wang, H.R., Causing, C.G., Kavsak, P., Stroschein, S.L., Luo, K. & Wrana, J.L., 2001. TGF-beta induces assembly of a Smad2-Smurf2 ubiquitin ligase complex that targets SnoN for degradation. *Nature Cell Biology*, 3(6), pp.587–95.

Borden, K.L. & Freemont, P.S., 1996. The RING finger domain: a recent example of a sequence—structure family. *Current Opinion in Structural Biology*, 6(3), pp.395–401.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. & Warren, G.L., 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica. Section D, Biological Crystallography*, 54(Pt 5), pp.905–21.

Brzovic, P.S., Lissounov, A., Christensen, D.E., Hoyt, D.W. & Klevit, R.E., 2006. A UbcH5/ubiquitin noncovalent complex is required for processive BRCA1-directed ubiquitination. *Molecular Cell*, 21(6), pp.873–80.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G., III, Skelton, N.J. & Rance, M., 2010. *Protein NMR Spectroscopy: Principles and Practice*, Academic Press.

Chebaro, Y., Amal, I., Rochel, N., Rochette-Egly, C., Stote, R.H. & Dejaegere, A., 2013. Phosphorylation of the retinoic acid receptor alpha induces a mechanical allosteric regulation and changes in internal dynamics. *PLoS Computational Biology*, 9(4),

p.e1003012.

Chen, C. & Matesic, L.E., 2007. The Nedd4-like family of E3 ubiquitin ligases and cancer. *Cancer Metastasis Reviews*, 26(3-4), pp.587–604.

Chen, C., Zhou, Z., Ross, J.S., Zhou, W. & Dong, J.-T., 2007. The amplified WWP1 gene is a potential molecular target in breast cancer. *International Journal of Cancer. Journal International Du Cancer*, 121(1), pp.80–87.

Chen, F. & Weinberg, R.A., 1995. Biochemical evidence for the autophosphorylation and transphosphorylation of transforming growth factor beta receptor kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 92(5), pp.1565–9.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. & Richardson, D.C., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Biological Crystallography*, 66(Pt 1), pp.12–21.

Chong, P.A., Lin, H., Wrana, J.L. & Forman-Kay, J.D., 2006. An expanded WW domain recognition motif revealed by the interaction between Smad7 and the E3 ubiquitin ligase Smurf2. *The Journal of Biological Chemistry*, 281(25), pp.17069–75.

Chong, P.A., Lin, H., Wrana, J.L. & Forman-Kay, J.D., 2010. Coupling of tandem Smad ubiquitination regulatory factor (Smurf) WW domains modulates target specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43), pp.18404–9.

Christensen, D.E., Brzovic, P.S. & Klevit, R.E., 2007. E2-BRCA1 RING interactions dictate synthesis of mono- or specific polyubiquitin chain linkages. *Nature Structural & Molecular Biology*, 14(10), pp.941–8.

Ciechanover, A. & Brundin, P., 2003. The Ubiquitin Proteasome System in Neurodegenerative Diseases. *Neuron*, 40(2), pp.427–446.

Ciechanover, A., Elias, S., Heller, H. & Hershko, A., 1982. “Covalent affinity” purification of ubiquitin-activating enzyme. *The Journal of Biological Chemistry*, 257(5), pp.2537–42.

Ciechanover, A., Heller, H., Elias, S., Haas, A.L. & Hershko, A., 1980. ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein

degradation. *Proceedings of the National Academy of Sciences of the United States of America*, 77(3), pp.1365–8.

Ciechanover, A., Heller, H., Katz-Etzion, R. & Hershko, A., 1981. Activation of the heat-stable polypeptide of the ATP-dependent proteolytic system. *Proceedings of the National Academy of Sciences of the United States of America*, 78(2), pp.761–5.

Ciechanover, A., Hod, Y. & Hershko, A., 1978. A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochemical and Biophysical Research Communications*, 81(4), pp.1100–1105.

Corn, P.G., 2007. Role of the ubiquitin proteasome system in renal cell carcinoma. *BMC Biochemistry*, 8 Suppl 1, p.S4.

Dantuma, N.P. & Bott, L.C., 2014. The ubiquitin-proteasome system in neurodegenerative diseases: precipitating factor, yet part of the solution. *Frontiers in Molecular Neuroscience*, 7, p.70.

Datto, M.B., Yu, Y. & Wang, X.F., 1995. Functional analysis of the transforming growth factor beta responsive elements in the WAF1/Cip1/p21 promoter. *The Journal of Biological Chemistry*, 270(48), pp.28623–8.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. & Bax, A., 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, 6(3), pp.277–93.

Delano, W., 2002. The PyMOL Molecular Graphics System.

Dodson, E.J., Fishbain-Yoskovitz, V., Rotem-Bamberger, S. & Schueler-Furman, O., 2015. Versatile communication strategies among tandem WW domain repeats. *Experimental Biology and Medicine (Maywood, N.J.)*, 240(3), pp.351–60.

Dodson, M., Darley-Usmar, V. & Zhang, J., 2013. Cellular metabolic and autophagic pathways: Traffic control by redox signaling. *Free Radical Biology and Medicine*, 63, pp.207–221.

Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G. & Baker, N.A., 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(Web Server issue), pp.W522–5.

Dolinsky, T.J., Nielsen, J.E., McCammon, J.A. & Baker, N.A., 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(Web Server issue), pp.W665–7.

Doreleijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A.E.M., Stevens, T.J., Vranken, W.F., Vriend, G. & Vuister, G.W., 2012. CING: an integrated residue-based structure validation program suite. *Journal of Biomolecular NMR*, 54(3), pp.267–83.

Dou, Q.P. & Zonder, J.A., 2014. Overview of proteasome inhibitor-based anti-cancer therapies: perspective on bortezomib and second generation proteasome inhibitors versus future generation inhibitors of ubiquitin-proteasome system. *Current Cancer Drug Targets*, 14(6), pp.517–36.

Dubois, C.M., Laprise, M.H., Blanchette, F., Gentry, L.E. & Leduc, R., 1995. Processing of transforming growth factor beta 1 precursor by human furin convertase. *The Journal of Biological Chemistry*, 270(18), pp.10618–24.

Ebisawa, T., Fukuchi, M., Murakami, G., Chiba, T., Tanaka, K., Imamura, T. & Miyazono, K., 2001. Smurf1 interacts with transforming growth factor-beta type I receptor through Smad7 and induces receptor degradation. *The Journal of Biological Chemistry*, 276(16), pp.12477–80.

Eddins, M.J., Carlile, C.M., Gomez, K.M., Pickart, C.M. & Wolberger, C., 2006. Mms2-Ubc13 covalently bound to ubiquitin reveals the structural basis of linkage-specific polyubiquitin chain formation. *Nature Structural & Molecular Biology*, 13(10), pp.915–20.

Espanel, X. & Sudol, M., 1999. A Single Point Mutation in a Group I WW Domain Shifts Its Specificity to That of Group II WW Domains. *Journal of Biological Chemistry*, 274(24), pp.17284–17289.

Fedoroff, O.Y., Townson, S.A., Golovanov, A.P., Baron, M. & Avis, J.M., 2004. The structure and dynamics of tandem WW domains in a negative regulator of notch signaling, Suppressor of deltex. *The Journal of Biological Chemistry*, 279(33), pp.34991–5000.

Fielding, L., 2003. NMR methods for the determination of protein-ligand dissociation constants. *Current Topics in Medicinal Chemistry*, 3(1), pp.39–53.

Fielding, L., Rutherford, S. & Fletcher, D., 2005. Determination of protein-ligand binding

affinity by NMR: observations from serum albumin model systems. *Magnetic Resonance in Chemistry : MRC*, 43(6), pp.463–70.

Fiorito, F., Herrmann, T., Damberger, F.F. & Wüthrich, K., 2008. Automated amino acid side-chain NMR assignment of proteins using (13)C- and (15)N-resolved 3D [(1)H, (1)H]-NOESY. *Journal of Biomolecular NMR*, 42(1), pp.23–33.

Fulda, S., Rajalingam, K. & Dikic, I., 2012. Ubiquitylation in immune disorders and cancer: from molecular mechanisms to therapeutic implications. *EMBO Molecular Medicine*, 4(7), pp.545–56.

Gao, S., Alarcón, C., Sapkota, G., Rahman, S., Chen, P.-Y., Goerner, N., Macias, M.J., Erdjument-Bromage, H., Tempst, P. & Massagué, J., 2009. Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. *Molecular Cell*, 36(3), pp.457–68.

Gentry, L.E. & Nash, B.W., 1990. The pro domain of pre-pro-transforming growth factor beta 1 when independently expressed is a functional binding protein for the mature growth factor. *Biochemistry*, 29(29), pp.6851–7.

Glick, D., Barth, S. & Macleod, K.F., 2010. Autophagy: cellular and molecular mechanisms. *The Journal of Pathology*, 221(1), pp.3–12.

Glickman, M.H. & Ciechanover, A., 2002. The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction. *Physiol Rev*, 82(2), pp.373–428.

Gomis, R.R., Alarcón, C., Nadal, C., Van Poznak, C. & Massagué, J., 2006. C/EBPbeta at the core of the TGFbeta cytostatic response and its evasion in metastatic breast cancer cells. *Cancer Cell*, 10(3), pp.203–14.

Gong, W., Zhang, X., Zhang, W., Li, J. & Li, Z., 2015. Structure of the HECT domain of human WWP2. *Acta Crystallographica. Section F, Structural Biology Communications*, 71(Pt 10), pp.1251–7.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. & Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(Web Server), pp.W695–W699.

Green, D.R. & Levine, B., 2014. To Be or Not to Be? How Selective Autophagy and Cell Death Govern Cell Fate. *Cell*, 157(1), pp.65–75.

Grenfell, S.J., Trausch-Azar, J.S., Handley-Gearhart, P.M., Ciechanover, A. & Schwartz, A.L., 1994. Nuclear localization of the ubiquitin-activating enzyme, E1, is cell-cycle-dependent. *The Biochemical Journal*, 300 (Pt 3, pp.701–8.

Groppe, J., Hinck, C.S., Samavarchi-Tehrani, P., Zubieta, C., Schuermann, J.P., Taylor, A.B., Schwarz, P.M., Wrana, J.L. & Hinck, A.P., 2008. Cooperative assembly of TGF-beta superfamily signaling complexes is mediated by two disparate mechanisms and distinct modes of receptor binding. *Molecular Cell*, 29(2), pp.157–68.

Güntert, P., 2004. Automated NMR structure calculation with CYANA. *Methods in Molecular Biology (Clifton, N.J.)*, 278, pp.353–78.

Haas, A.L. & Rose, I.A., 1982. The mechanism of ubiquitin activating enzyme. A kinetic and equilibrium analysis. *The Journal of Biological Chemistry*, 257(17), pp.10329–37.

Hafsa, N.E. & Wishart, D.S., 2014. CSI 2.0: a significantly improved version of the Chemical Shift Index. *Journal of Biomolecular NMR*, 60(2-3), pp.131–46.

Haldeman, M.T., Xia, G., Kasperek, E.M. & Pickart, C.M., 1997. Structure and function of ubiquitin conjugating enzyme E2-25K: the tail is a core-dependent activity element. *Biochemistry*, 36(34), pp.10526–37.

Hamilton, K.S., Ellison, M.J., Barber, K.R., Williams, R.S., Huzil, J.T., McKenna, S., Ptak, C., Glover, M. & Shaw, G.S., 2001. Structure of a Conjugating Enzyme-Ubiquitin Thiolester Intermediate Reveals a Novel Role for the Ubiquitin Tail. *Structure*, 9(10), pp.897–904.

Hammarström, M., Hellgren, N., van Den Berg, S., Berglund, H. & Härd, T., 2002. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein Science : A Publication of the Protein Society*, 11(2), pp.313–21.

Hammarström, M., Woestenenk, E.A., Hellgren, N., Härd, T. & Berglund, H., 2006. Effect of N-terminal solubility enhancing fusion proteins on yield of purified target protein. *Journal of Structural and Functional Genomics*, 7(1), pp.1–14.

Hannon, G.J. & Beach, D., 1994. p15INK4B is a potential effector of TGF-beta-induced cell cycle arrest. *Nature*, 371(6494), pp.257–61.

Hayashi, H., Abdollah, S., Qiu, Y., Cai, J., Xu, Y.Y., Grinnell, B.W., Richardson, M.A., Topper, J.N., Gimbrone, M.A., Wrana, J.L. & Falb, D., 1997. The MAD-related protein Smad7

associates with the TGFbeta receptor and functions as an antagonist of TGFbeta signaling. *Cell*, 89(7), pp.1165–73.

Heldin, C.-H., Landström, M. & Moustakas, A., 2009. Mechanism of TGF-beta signaling to growth arrest, apoptosis, and epithelial-mesenchymal transition. *Current Opinion in Cell Biology*, 21(2), pp.166–76.

Heldin, C.-H., Vanlandewijck, M. & Moustakas, A., 2012. Regulation of EMT by TGFβ in cancer. *FEBS Letters*, 586(14), pp.1959–70.

Herrmann, T., Güntert, P. & Wüthrich, K., 2002a. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319(1), pp.209–27.

Herrmann, T., Güntert, P. & Wüthrich, K., 2002b. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR*, 24(3), pp.171–89.

Hershko, A., Ciechanover, A., Heller, H., Haas, A.L. & Rose, I.A., 1980. Proposed role of ATP in protein breakdown: conjugation of protein with multiple chains of the polypeptide of ATP-dependent proteolysis. *Proceedings of the National Academy of Sciences of the United States of America*, 77(4), pp.1783–6.

Hershko, A., Ciechanover, A. & Rose, I.A., 1981. Identification of the active amino acid residue of the polypeptide of ATP-dependent protein breakdown. *The Journal of Biological Chemistry*, 256(4), pp.1525–8.

Hershko, A., Ciechanover, A. & Rose, I.A., 1979. Resolution of the ATP-dependent proteolytic system from reticulocytes: a component that interacts with ATP. *Proceedings of the National Academy of Sciences of the United States of America*, 76(7), pp.3107–10.

Hershko, A., Heller, H., Elias, S. & Ciechanover, A., 1983. Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown. *The Journal of Biological Chemistry*, 258(13), pp.8206–14.

Hirano, T., Serve, O., Yagi-Utsumi, M., Takemoto, E., Hiromoto, T., Satoh, T., Mizushima, T. & Kato, K., 2011. Conformational dynamics of wild-type Lys-48-linked diubiquitin in solution. *The Journal of Biological Chemistry*, 286(43), pp.37496–502.

Horiguchi, K., Sakamoto, K., Koinuma, D., Semba, K., Inoue, A., Inoue, S., Fujii, H.,

Yamaguchi, A., Miyazawa, K., Miyazono, K. & Saitoh, M., 2012. TGF-β drives epithelial-mesenchymal transition through δEF1-mediated downregulation of ESRP. *Oncogene*, 31(26), pp.3190–201.

Huang, L., Kinnucan, E., Wang, G., Beaudenon, S., Howley, P.M., Huibregtse, J.M. & Pavletich, N.P., 1999. Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the E2-E3 enzyme cascade. *Science (New York, N.Y.)*, 286(5443), pp.1321–6.

Huibregtse, J.M., Scheffner, M., Beaudenon, S. & Howley, P.M., 1995. A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase. *Proceedings of the National Academy of Sciences*, 92(7), pp.2563–2567.

Huibregtse, J.M., Scheffner, M. & Howley, P.M., 1994. E6-AP directs the HPV E6-dependent inactivation of p53 and is representative of a family of structurally and functionally related proteins. *Cold Spring Harbor Symposia on Quantitative Biology*, 59, pp.237–45.

Huth, J.R., Bewley, C.A., Jackson, B.M., Hinnebusch, A.G., Clore, G.M. & Gronenborn, A.M., 1997. Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. *Protein Science : A Publication of the Protein Society*, 6(11), pp.2359–64.

Ilari, A. & Savino, C., 2008. Protein structure determination by x-ray crystallography. *Methods in Molecular Biology (Clifton, N.J.)*, 452, pp.63–87.

Ingham, R.J., Colwill, K., Howard, C., Dettwiler, S., Lim, C.S.H., Yu, J., Hersi, K., Raaijmakers, J., Gish, G., Mbamalu, G., Taylor, L., Yeung, B., Vassilovski, G., Amin, M., Chen, F., Matskova, L., Winberg, G., Ernberg, I., Linding, R., O'donnell, P., Starostine, A., Keller, W., Metalnikov, P., Stark, C. & Pawson, T., 2005. WW domains provide a platform for the assembly of multiprotein networks. *Molecular and Cellular Biology*, 25(16), pp.7092–106.

Iwai, H., Züger, S., Jin, J. & Tam, P.-H., 2006. Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. *FEBS Letters*, 580(7), pp.1853–8.

Jackson, P.K., Eldridge, A.G., Freed, E., Furstenthal, L., Hsu, J.Y., Kaiser, B.K. & Reimann, J.D., 2000. The lore of the RINGs: substrate recognition and catalysis by ubiquitin

ligases. *Trends in Cell Biology*, 10(10), pp.429–39.

Jiang, J., Wang, N., Jiang, Y., Tan, H., Zheng, J., Chen, G. & Jia, Z., 2015. Characterization of substrate binding of the WW domains in human WWP2 protein. *FEBS Letters*, 589(15), pp.1935–42.

Jiang, J., Zheng, J., She, Y. & Jia, Z., 2015. Expression and purification of human WWP2 HECT domain in Escherichia coli. *Protein Expression and Purification*, 110, pp.95–101.

Johansen, T. & Lamark, T., 2011. Selective autophagy mediated by autophagic adapter proteins. *Autophagy*, 7(3), pp.279–96.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), pp.195–202.

Jung, J.-G., Stoeck, A., Guan, B., Wu, R.-C., Zhu, H., Blackshaw, S., Shih, I.-M. & Wang, T.-L., 2014. Notch3 Interactome Analysis Identified WWP2 as a Negative Regulator of Notch3 Signaling in Ovarian Cancer P. McKinnon, ed. *PLoS Genetics*, 10(10), p.e1004751.

Kabsch, W. & Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), pp.2577–637.

Kamadurai, H.B., Qiu, Y., Deng, A., Harrison, J.S., MacDonald, C., Actis, M., Rodrigues, P., Miller, D.J., Souphron, J., Lewis, S.M., Kurinov, I., Fujii, N., Hammel, M., Piper, R., Kuhlman, B. & Schulman, B.A., 2013. Mechanism of ubiquitin ligation and lysine prioritization by a HECT E3. *eLife*, 2, p.e00828.

Kamadurai, H.B., Souphron, J., Scott, D.C., Duda, D.M., Miller, D.J., Stringer, D., Piper, R.C. & Schulman, B.A., 2009. Insights into ubiquitin transfer cascades from a structure of a UbcH5B approximately ubiquitin-HECT(NEDD4L) complex. *Molecular Cell*, 36(6), pp.1095–102.

Kane, J.F., 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Current Opinion in Biotechnology*, 6(5), pp.494–500.

Kanelis, V., Bruce, M.C., Skrynnikov, N.R., Rotin, D. & Forman-Kay, J.D., 2006. Structural determinants for high-affinity binding in a Nedd4 WW3* domain-Comm PY motif complex. *Structure (London, England : 1993)*, 14(3), pp.543–53.

Kanelis, V., Farrow, N.A., Kay, L.E., Rotin, D. & Forman-Kay, J.D., 1998. NMR studies of tandem WW domains of Nedd4 in complex with a PY motif-containing region of the epithelial sodium channel. *Biochemistry and Cell Biology*, 76(2-3), pp.341–350.

Kanelis, V., Rotin, D. & Forman-Kay, J.D., 2001. Solution structure of a Nedd4 WW domain-ENaC peptide complex. *Nature Structural Biology*, 8(5), pp.407–12.

Katsuno, Y., Lamouille, S. & Derynck, R., 2013. TGF-β signaling and epithelial-mesenchymal transition in cancer progression. *Current Opinion in Oncology*, 25(1), pp.76–84.

Kavsak, P., Rasmussen, R.K., Causing, C.G., Bonni, S., Zhu, H., Thomsen, G.H. & Wrana, J.L., 2000. Smad7 Binds to Smurf2 to Form an E3 Ubiquitin Ligase that Targets the TGFβ Receptor for Degradation. *Molecular Cell*, 6(6), pp.1365–1375.

Keeler, J., 2011. *Understanding NMR Spectroscopy*, John Wiley & Sons.

Keller, R.L.J., 2004. The Computer Aided Resonance Assignment.

Kim, H.C. & Huibregtse, J.M., 2009. Polyubiquitination by HECT E3s and the determinants of chain type specificity. *Molecular and Cellular Biology*, 29(12), pp.3307–18.

Komander, D., 2009. The emerging complexity of protein ubiquitination. *Biochemical Society Transactions*, 37(Pt 5), pp.937–53.

Komuro, A., Imamura, T., Saitoh, M., Yoshida, Y., Yamori, T., Miyazono, K. & Miyazawa, K., 2004. Negative regulation of transforming growth factor-beta (TGF-beta) signaling by WW domain-containing protein 1 (WWP1). *Oncogene*, 23(41), pp.6914–23.

Korenchuk, S., Lehr, J.E., MClean, L., Lee, Y.G., Whitney, S., Vessella, R., Lin, D.L. & Pienta, K.J., VCaP, a cell-based model system of human prostate cancer. *In Vivo (Athens, Greece)*, 15(2), pp.163–8.

Kowalski, K., Merkel, A.. & Booker, G.W., 2005. Solution structures of the WW domains of Nedd4-2. *To Be Published*.

Kumar, P., Chimenti, M.S., Pemble, H., Schönichen, A., Thompson, O., Jacobson, M.P. & Wittmann, T., 2012. Multisite phosphorylation disrupts arginine-glutamate salt bridge networks required for binding of cytoplasmic linker-associated protein 2 (CLASP2) to end-binding protein 1 (EB1). *The Journal of Biological Chemistry*, 287(21), pp.17050–64.

Kumar, S. & Nussinov, R., 1999. Salt bridge stability in monomeric proteins. *Journal of*

Molecular Biology, 293(5), pp.1241–55.

Kuszewski, J., Gronenborn, A.M. & Clore, G.M., 1999. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *Journal of the American Chemical Society*, 121(10), pp.2337–2338.

Kwei, K.A., Shain, A.H., Bair, R., Montgomery, K., Karikari, C.A., van de Rijn, M., Hidalgo, M., Maitra, A., Bashyam, M.D. & Pollack, J.R., 2011. SMURF1 amplification promotes invasiveness in pancreatic cancer. *PloS One*, 6(8), p.e23924.

Kwon, A., Lee, H.-L., Woo, K.M., Ryoo, H.-M. & Baek, J.-H., 2013. SMURF1 plays a role in EGF-induced breast cancer cell migration and invasion. *Molecules and Cells*, 36(6), pp.548–55.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), pp.283–291.

Lawrence, D.A., Pircher, R. & Jullien, P., 1985. Conversion of a high molecular weight latent beta-TGF from chicken embryo fibroblasts into a low molecular weight active beta-TGF under acidic conditions. *Biochemical and Biophysical Research Communications*, 133(3), pp.1026–34.

Lee, I. & Schindelin, H., 2008. Structural insights into E1-catalyzed ubiquitin activation and transfer to conjugating enzymes. *Cell*, 134(2), pp.268–78.

Lee, M.J., Lee, B.-H., Hanna, J., King, R.W. & Finley, D., 2011. Trimming of ubiquitin chains by proteasome-associated deubiquitinating enzymes. *Molecular & Cellular Proteomics : MCP*, 10(5), p.R110.003871.

Leitlein, J., Aulwurm, S., Waltereit, R., Naumann, U., Wagenknecht, B., Garten, W., Weller, M. & Platten, M., 2001. Processing of immunosuppressive pro-TGF-beta 1,2 by human glioblastoma cells involves cytoplasmic and secreted furin-like proteases. *Journal of Immunology (Baltimore, Md. : 1950)*, 166(12), pp.7238–43.

Levitt, M.H., 2001. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*, Wiley.

Li, W., Tu, D., Brunger, A.T. & Ye, Y., 2007. A ubiquitin ligase transfers preformed polyubiquitin chains from a conjugating enzyme to a substrate. *Nature*, 446(7133), pp.333–7.

Li, Y., Zhou, Z., Alimandi, M. & Chen, C., 2009. WW domain containing E3 ubiquitin protein ligase 1 targets the full-length ErbB4 for ubiquitin-mediated degradation in breast cancer. *Oncogene*, 28(33), pp.2948–2958.

Lian, L.-Y. & Roberts, G. eds., 2011. *Protein NMR Spectroscopy: Practical Techniques and Applications*, Chichester, UK: John Wiley & Sons.

Liao, B. & Jin, Y., 2010. Wwp2 mediates Oct4 ubiquitination and its own auto-ubiquitination in a dosage-dependent manner. *Cell Research*, 20(3), pp.332–44.

Lin, H.Y., Moustakas, A., Knaus, P., Wells, R.G., Henis, Y.I. & Lodish, H.F., 1995. The soluble exoplasmic domain of the type II transforming growth factor (TGF)-beta receptor. A heterogeneously glycosylated protein with high affinity and selectivity for TGF-beta ligands. *The Journal of Biological Chemistry*, 270(6), pp.2747–54.

Lin, H.Y., Wang, X.F., Ng-Eaton, E., Weinberg, R.A. & Lodish, H.F., 1992. Expression cloning of the TGF-beta type II receptor, a functional transmembrane serine/threonine kinase. *Cell*, 68(4), pp.775–85.

Lin, X., Liang, M. & Feng, X.H., 2000. Smurf2 is a ubiquitin E3 ligase mediating proteasome-dependent degradation of Smad2 in transforming growth factor-beta signaling. *The Journal of Biological Chemistry*, 275(47), pp.36818–22.

Liu, H.-Y. & Pfleger, C.M., 2013. Mutation in E1, the ubiquitin activating enzyme, reduces Drosophila lifespan and results in motor impairment. *PloS One*, 8(1), p.e32835.

Löw, P., 2011. The role of ubiquitin-proteasome system in ageing. *General and Comparative Endocrinology*, 172(1), pp.39–43.

Lu, P.J., Zhou, X.Z., Shen, M. & Lu, K.P., 1999. Function of WW domains as phosphoserine- or phosphothreonine-binding modules. *Science (New York, N.Y.)*, 283(5406), pp.1325–8.

Luh, L.M., Hänsel, R., Löhr, F., Kirchner, D.K., Krauskopf, K., Pitzius, S., Schäfer, B., Tufar, P., Corbeski, I., Güntert, P. & Dötsch, V., 2013. Molecular crowding drives active Pin1 into nonspecific complexes with endogenous proteins prior to substrate recognition. *Journal of the American Chemical Society*, 135(37), pp.13796–803.

Lyons, R.M., Gentry, L.E., Purchio, A.F. & Moses, H.L., 1990. Mechanism of activation of latent recombinant transforming growth factor beta 1 by plasmin. *The Journal of Cell Biology*, 110(4), pp.1361–7.

Macias, M.J., Hyvönen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M. & Oschkinat, H., 1996. Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature*, 382(6592), pp.646–9.

Macias, M.J., Wiesner, S. & Sudol, M., 2002. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Letters*, 513(1), pp.30–37.

Maddika, S., Kavela, S., Rani, N., Palicharla, V.R., Pokorny, J.L., Sarkaria, J.N. & Chen, J., 2011. WWP2 is an E3 ubiquitin ligase for PTEN. *Nature Cell Biology*, 13(6), pp.730–735.

Mani, S.A., Yang, J., Brooks, M., Schwaninger, G., Zhou, A., Miura, N., Kutok, J.L., Hartwell, K., Richardson, A.L. & Weinberg, R.A., 2007. Mesenchyme Forkhead 1 (FOXC2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24), pp.10069–74.

Manuel García-Ruiz, J., 2003. Nucleation of protein crystals. *Journal of Structural Biology*, 142(1), pp.22–31.

Martinez-Rodriguez, S., Bacarizo, J., Luque, I. & Camara-Artigas, A., 2015. Crystal structure of the first WW domain of human YAP2 isoform. *J.Struct.Biol.*, 191, pp.381–387.

Martinez-Vicente, M., Sovak, G. & Cuervo, A.M., 2005. Protein degradation and aging. *Experimental Gerontology*, 40(8-9), pp.622–33.

McLean, J.R., Chaix, D., Ohi, M.D. & Gould, K.L., 2011. State of the APC/C: organization, function, and structure. *Critical Reviews in Biochemistry and Molecular Biology*, 46(2), pp.118–36.

McPherson, A., 2004. Introduction to protein crystallization. *Methods (San Diego, Calif.)*, 34(3), pp.254–65.

Metzger, M.B., Pruneda, J.N., Klevit, R.E. & Weissman, A.M., 2014. RING-type E3 ligases: master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. *Biochimica et Biophysica Acta*, 1843(1), pp.47–60.

Middleton, A.J. & Day, C.L., 2015. The molecular basis of lysine 48 ubiquitin chain synthesis by Ube2K. *Scientific Reports*, 5, p.16793.

Miyazaki, K., Fujita, T., Ozaki, T., Kato, C., Kurose, Y., Sakamoto, M., Kato, S., Goto, T.,

Itoyama, Y., Aoki, M. & Nakagawara, A., 2004. NEDL1, a novel ubiquitin-protein isopeptide ligase for dishevelled-1, targets mutant superoxide dismutase-1. *The Journal of Biological Chemistry*, 279(12), pp.11327–35.

Miyazono, K., Olofsson, A., Colosetti, P. & Heldin, C.H., 1991. A role of the latent TGF-beta 1-binding protein in the assembly and secretion of TGF-beta 1. *The EMBO Journal*, 10(5), pp.1091–101.

Mizushima, N. & Komatsu, M., 2011. Autophagy: renovation of cells and tissues. *Cell*, 147(4), pp.728–41.

Morgan, R.T., Woods, L.K., Moore, G.E., Quinn, L.A., McGavran, L. & Gordon, S.G., 1980. Human cell line (COLO 357) of metastatic pancreatic adenocarcinoma. *International Journal of Cancer. Journal International Du Cancer*, 25(5), pp.591–8.

Mossessova, E. & Lima, C.D., 2000. Ulp1-SUMO Crystal Structure and Genetic Analysis Reveal Conserved Interactions and a Regulatory Element Essential for Cell Growth in Yeast. *Molecular Cell*, 5(5), pp.865–876.

Moustakas, A. & Heldin, C.-H., 2007. Signaling networks guiding epithelial-mesenchymal transitions during embryogenesis and cancer progression. *Cancer Science*, 98(10), pp.1512–20.

Mund, T., Graeb, M., Mieszczanek, J., Gammons, M., Pelham, H.R.B. & Bienz, M., 2015. Disinhibition of the HECT E3 ubiquitin ligase WWP2 by polymerized Dishevelled. *Open Biology*, 5(12).

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A.E.M., Nabuurs, S.B., Güntert, P., Livny, M., Markley, J.L., Nilges, M., Ulrich, E.L., Kaptein, R. & Bonvin, A.M.J.J., 2005. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, 59(4), pp.662–72.

Nunes, I., Gleizes, P.E., Metz, C.N. & Rifkin, D.B., 1997. Latent transforming growth factor-beta binding protein domains involved in activation and transglutaminase-dependent cross-linking of latent transforming growth factor-beta. *The Journal of Cell Biology*, 136(5), pp.1151–63.

Nussbaum, A.K., Dick, T.P., Keilholz, W., Schirle, M., Stevanović, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D.H., Huber, R., Rammensee, H.G. & Schild, H., 1998. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase

1. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21), pp.12504–9.

Ogunjimi, A.A., Briant, D.J., Pece-Barbara, N., Le Roy, C., Di Guglielmo, G.M., Kavsak, P., Rasmussen, R.K., Seet, B.T., Sicheri, F. & Wrana, J.L., 2005. Regulation of Smurf2 ubiquitin ligase activity by anchoring the E2 to the HECT domain. *Molecular Cell*, 19(3), pp.297–308.

Olsen, S.K., Capili, A.D., Lu, X., Tan, D.S. & Lima, C.D., 2010. Active site remodelling accompanies thioester bond formation in the SUMO E1. *Nature*, 463(7283), pp.906–12.

Olsen, S.K. & Lima, C.D., 2013. Structure of a ubiquitin E1-E2 complex: insights to E1-E2 thioester transfer. *Molecular Cell*, 49(5), pp.884–96.

Padua, D. & Massagué, J., 2009. Roles of TGFbeta in metastasis. *Cell Research*, 19(1), pp.89–102.

Pecharsky, V. & Zavalij, P., 2008. *Fundamentals of Powder Diffraction and Structural Characterization of Materials, Second Edition*, Springer Science & Business Media.

Peinado, H., Quintanilla, M. & Cano, A., 2003. Transforming growth factor beta-1 induces snail transcription factor in epithelial cell lines: mechanisms for epithelial mesenchymal transitions. *The Journal of Biological Chemistry*, 278(23), pp.21113–23.

Pickart, C.M., 2000. Ubiquitin in chains. *Trends in Biochemical Sciences*, 25(11), pp.544–548.

Pickart, C.M. & Eddins, M.J., 2004. Ubiquitin: structures, functions, mechanisms. *Biochimica et Biophysica Acta*, 1695(1-3), pp.55–72.

Plechanovová, A., Jaffray, E.G., Tatham, M.H., Naismith, J.H. & Hay, R.T., 2012. Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. *Nature*, 489(7414), pp.115–20.

Pulaski, L., Landström, M., Heldin, C.H. & Souchelnytskyi, S., 2001. Phosphorylation of Smad7 at Ser-249 does not interfere with its inhibitory role in transforming growth factor-beta-dependent signaling but affects Smad7-dependent transcriptional activation. *The Journal of Biological Chemistry*, 276(17), pp.14344–9.

Qin, H.., Li, M.., Pu, H., Sankaran, S., Ahmed, S. & Song, J.X., 2007. NMR structure of the forth WW domain of WWP1. *To Be Published*.

Ramser, J., Ahearn, M.E., Lenski, C., Yariz, K.O., Hellebrand, H., von Rhein, M., Clark, R.D., Schmutzler, R.K., Lichtner, P., Hoffman, E.P., Meindl, A. & Baumbach-Reardon, L., 2008. Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy. *American Journal of Human Genetics*, 82(1), pp.188–93.

Riling, C., Kamadurai, H., Kumar, S., O'Leary, C.E., Wu, K.-P., Manion, E.E., Ying, M., Schulman, B.A. & Oliver, P.M., 2015. Itch WW Domains Inhibit Its E3 Ubiquitin Ligase Activity by Blocking E2-E3 Ligase Trans-thiolation. *The Journal of Biological Chemistry*, 290(39), pp.23875–87.

Roberts, G.C.K. ed., 2013. *Encyclopedia of Biophysics*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Rodrigo-Brenni, M.C. & Morgan, D.O., 2007. Sequential E2s drive polyubiquitin chain assembly on APC targets. *Cell*, 130(1), pp.127–39.

Saharinen, J., Taipale, J. & Keski-Oja, J., 1996. Association of the small latent transforming growth factor-beta with an eight cysteine repeat of its binding protein LTBP-1. *The EMBO Journal*, 15(2), pp.245–53.

Sánchez-Elsner, T., Botella, L.M., Velasco, B., Corbí, A., Attisano, L. & Bernabéu, C., 2001. Synergistic cooperation between hypoxia and transforming growth factor-beta pathways on human vascular endothelial growth factor gene expression. *The Journal of Biological Chemistry*, 276(42), pp.38527–35.

Sancho, E., Vilá, M.R., Sánchez-Pulido, L., Lozano, J.J., Paciucci, R., Nadal, M., Fox, M., Harvey, C., Bercovich, B., Loukili, N., Ciechanover, A., Lin, S.L., Sanz, F., Estivill, X., Valencia, A. & Thomson, T.M., 1998. Role of UEV-1, an inactive variant of the E2 ubiquitin-conjugating enzymes, in in vitro differentiation and cell cycle behavior of HT-29-M6 intestinal mucosecretory cells. *Molecular and Cellular Biology*, 18(1), pp.576–89.

Sato, Y. & Rifkin, D.B., 1989. Inhibition of endothelial cell movement by pericytes and smooth muscle cells: activation of a latent transforming growth factor-beta 1-like molecule by plasmin during co-culture. *The Journal of Cell Biology*, 109(1), pp.309–

15.

Scandura, J.M., Boccuni, P., Massagué, J. & Nimer, S.D., 2004. Transforming growth factor beta-induced cell cycle arrest of human hematopoietic cells requires p57KIP2 up-regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42), pp.15231–6.

Scheffner, M., 1995. Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. *Nature*, 373(6509), pp.81 – 83.

Schultz-Cherry, S. & Murphy-Ullrich, J.E., 1993. Thrombospondin causes activation of latent transforming growth factor-beta secreted by endothelial cells by a novel mechanism. *The Journal of Cell Biology*, 122(4), pp.923–32.

Shabbeer, S., Omer, D., Berneman, D., Weitzman, O., Alpaugh, A., Pietraszkiewicz, A., Metsuyanim, S., Shainskaya, A., Papa, M.Z. & Yarden, R.I., 2013. BRCA1 targets G2/M cell cycle proteins for ubiquitination and proteasomal degradation. *Oncogene*, 32(42), pp.5005–16.

Shang, F., Deng, G., Obin, M., Wu, C.C., Gong, X., Smith, D., Laursen, R.A., Andley, U.P., Reddan, J.R. & Taylor, A., 2001. Ubiquitin-activating enzyme (E1) isoforms in lens epithelial cells: origin of translation, E2 specificity and cellular localization determined with novel site-specific antibodies. *Experimental Eye Research*, 73(6), pp.827–36.

Shapiro, I.M., Cheng, A.W., Flytzanis, N.C., Balsamo, M., Condeelis, J.S., Oktay, M.H., Burge, C.B. & Gertler, F.B., 2011. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genetics*, 7(8), p.e1002218.

Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A., 2009. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR*, 44(4), pp.213–23.

Shi, D. & Grossman, S.R., 2010. Ubiquitin becomes ubiquitous in cancer: emerging roles of ubiquitin ligases and deubiquitinases in tumorigenesis and as therapeutic targets. *Cancer Biology & Therapy*, 10(8), pp.737–47.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D. & Higgins, D.G., 2011. Fast, scalable

generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, p.539.

Soond, S.M. & Chantry, A., 2011. Selective targeting of activating and inhibitory Smads by distinct WWP2 ubiquitin ligase isoforms differentially modulates TGFβ signalling and EMT. *Oncogene*, 30(21), pp.2451–2462.

Soond, S.M., Smith, P.G., Wahl, L., Swingler, T.E., Clark, I.M., Hemmings, A.M. & Chantry, A., 2013. Novel WWP2 ubiquitin ligase isoforms as potential prognostic markers and molecular targets in cancer. *Biochimica et Biophysica Acta*, 1832(12), pp.2127–35.

Staub, O., Dho, S., Henry, P., Correa, J., Ishikawa, T., McGlade, J. & Rotin, D., 1996. WW domains of Nedd4 bind to the proline-rich PY motifs in the epithelial Na+ channel deleted in Liddle's syndrome. *The EMBO Journal*, 15(10), pp.2371–80.

Staub, O., Gautschi, I., Ishikawa, T., Breitschopf, K., Ciechanover, A., Schild, L. & Rotin, D., 1997. Regulation of stability and function of the epithelial Na+ channel (ENaC) by ubiquitination. *The EMBO Journal*, 16(21), pp.6325–36.

Staub, O. & Rotin, D., 1996. WW domains. *Structure*, 4(5), pp.495–499.

Stopa, M., Anhuf, D., Terstegen, L., Gatsios, P., Gressner, A.M. & Dooley, S., 2000. Participation of Smad2, Smad3, and Smad4 in transforming growth factor beta (TGF-beta)-induced activation of Smad7. THE TGF-beta response element of the promoter requires functional Smad binding element and E-box sequences for transcriptional regulation. *The Journal of Biological Chemistry*, 275(38), pp.29308–17.

Studier, F.W. & Moffatt, B.A., 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology*, 189(1), pp.113–30.

Sullivan, J.A., Shirasu, K. & Deng, X.W., 2003. The diverse roles of ubiquitin and the 26S proteasome in the life of plants. *Nature Reviews. Genetics*, 4(12), pp.948–58.

Sun, A., Yu, G., Dou, X., Yan, X., Yang, W. & Lin, Q., 2014. Nedd4-1 is an exceptional prognostic biomarker for gastric cardia adenocarcinoma and functionally associated with metastasis. *Molecular Cancer*, 13, p.248.

Suzuki, C., Murakami, G., Fukuchi, M., Shimanuki, T., Shikauchi, Y., Imamura, T. & Miyazono, K., 2002. Smurf1 regulates the inhibitory activity of Smad7 by targeting Smad7 to the plasma membrane. *The Journal of Biological Chemistry*, 277(42),

pp.39919–25.

Tabor, S., 2001. Expression using the T7 RNA polymerase/promoter system. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, Chapter 16, p.Unit16.2.

Tanksley, J.P., Chen, X. & Coffey, R.J., 2013. NEDD4L is downregulated in colorectal cancer and inhibits canonical WNT signaling. *PloS One*, 8(11), p.e81514.

Thrower, J.S., Hoffman, L., Rechsteiner, M. & Pickart, C.M., 2000. Recognition of the polyubiquitin proteolytic signal. *The EMBO Journal*, 19(1), pp.94–102.

Todaro, D., Augustus-Wallace, A. & Haas, A., 2015. Mechanistic Characterization of the HECT Domain Ubiquitin Ligase Nedd4-2. *FASEB J*, 29(1_Supplement), p.883.10–.

Tomaić, V. & Banks, L., 2015. Angelman syndrome-associated ubiquitin ligase UBE3A/E6AP mutants interfere with the proteolytic activity of the proteasome. *Cell Death & Disease*, 6, p.e1625.

Tsukazaki, T., Chiang, T.A., Davison, A.F., Attisano, L. & Wrana, J.L., 1998. SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor. *Cell*, 95(6), pp.779–91.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Kent Wenger, R., Yao, H. & Markley, J.L., 2008. BioMagResBank. *Nucleic Acids Research*, 36(Database issue), pp.D402–8.

Valcourt, U., Kowanetz, M., Niimi, H., Heldin, C.-H. & Moustakas, A., 2005. TGF-beta and the Smad signaling pathway support transcriptomic reprogramming during epithelial-mesenchymal cell transition. *Molecular Biology of the Cell*, 16(4), pp.1987–2002.

VanDemark, A.P. & Hill, C.P., 2002. Structural basis of ubiquitylation. *Current Opinion in Structural Biology*, 12(6), pp.822–30.

Verdecia, M.A., Bowman, M.E., Lu, K.P., Hunter, T. & Noel, J.P., 2000. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nature Structural Biology*, 7(8), pp.639–43.

Verdecia, M.A., Joazeiro, C.A.., Wells, N.J., Ferrer, J.-L., Bowman, M.E., Hunter, T. & Noel, J.P., 2003. Conformational Flexibility Underlies Ubiquitin Ligation Mediated by the

WWP1 HECT Domain E3 Ligase. *Molecular Cell*, 11(1), pp.249–259.

Vijay-kumar, S., Bugg, C.E. & Cook, W.J., 1987. Structure of ubiquitin refined at 1.8 Å resolution. *Journal of Molecular Biology*, 194(3), pp.531–544.

Volk, J., Herrmann, T. & Wüthrich, K., 2008. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *Journal of Biomolecular NMR*, 41(3), pp.127–38.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. & Laue, E.D., 2005. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, 59(4), pp.687–96.

Vriend, G., 1990. WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics*, 8(1), pp.52–56.

Waelter, S., Boeddrich, A., Lurz, R., Scherzinger, E., Lueder, G., Lehrach, H. & Wanker, E.E., 2001. Accumulation of mutant huntingtin fragments in aggresome-like inclusion bodies as a result of insufficient protein degradation. *Molecular Biology of the Cell*, 12(5), pp.1393–407.

Walker, J.., Qiu, L., Li, Y., Weigelt, J., Bountra, C., Arrowsmith, C.., Edwards, A.., Bochkarev, A. & Dhe-Paganon, S., 2010. The tandem helical box and second WW domains of human HECW1. *To Be Published*.

Walz, J., Erdmann, A., Kania, M., Typke, D., Koster, A.J. & Baumeister, W., 1998. 26S proteasome structure revealed by three-dimensional electron microscopy. *Journal of Structural Biology*, 121(1), pp.19–29.

Wang, G., Matsuura, I., He, D. & Liu, F., 2009. Transforming growth factor-{beta}-inducible phosphorylation of Smad3. *The Journal of Biological Chemistry*, 284(15), pp.9663–73.

Wang, X., Trotman, L.C., Koppie, T., Alimonti, A., Chen, Z., Gao, Z., Wang, J., Erdjument-Bromage, H., Tempst, P., Cordon-Cardo, C., Pandolfi, P.P. & Jiang, X., 2007. NEDD4-1 Is a Proto-Oncogenic Ubiquitin Ligase for PTEN. *Cell*, 128(1), pp.129–139.

Ward, C.L., Omura, S. & Kopito, R.R., 1995. Degradation of CFTR by the ubiquitin-proteasome pathway. *Cell*, 83(1), pp.121–127.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T., 2004. The DISOPRED server

for the prediction of protein disorder. *Bioinformatics (Oxford, England)*, 20(13), pp.2138–9.

Webb, C., Upadhyay, A., Giuntini, F., Eggleston, I., Furutani-Seiki, M., Ishima, R. & Bagby, S., 2011. Structural features and ligand binding properties of tandem WW domains from YAP and TAZ, nuclear effectors of the Hippo pathway. *Biochemistry*, 50(16), pp.3300–9.

Wenzel, D.M., Stoll, K.E. & Klevit, R.E., 2011. E2s: structurally economical and functionally replete. *The Biochemical Journal*, 433(1), pp.31–42.

Wiesner, S., Ogunjimi, A.A., Wang, H.-R., Rotin, D., Sicheri, F., Wrana, J.L. & Forman-Kay, J.D., 2007. Autoinhibition of the HECT-type ubiquitin ligase Smurf2 through its C2 domain. *Cell*, 130(4), pp.651–62.

van Wijk, S.J.L. & Timmers, H.T.M., 2010. The family of ubiquitin-conjugating enzymes (E2s): deciding between life and death of proteins. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, 24(4), pp.981–93.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D. & Hochstrasser, D.F., 1999. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology (Clifton, N.J.)*, 112, pp.531–52.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. & Sykes, B.D., 1995. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of Biomolecular NMR*, 5(1), pp.67–81.

Wrana, J.L., Attisano, L., Wieser, R., Ventura, F. & Massagué, J., 1994. Mechanism of activation of the TGF-beta receptor. *Nature*, 370(6488), pp.341–7.

Wüthrich, K., 1990. Protein structure determination in solution by NMR spectroscopy. *The Journal of Biological Chemistry*, 265(36), pp.22059–62.

Xu, H., Wang, W., Li, C., Yu, H., Yang, A., Wang, B. & Jin, Y., 2009. WWP2 promotes degradation of transcription factor OCT4 in human embryonic stem cells. *Cell Research*, 19(5), pp.561–73.

Xu, H.M., Liao, B., Zhang, Q.J., Wang, B.B., Li, H., Zhong, X.M., Sheng, H.Z., Zhao, Y.X., Zhao, Y.M. & Jin, Y., 2004. Wwp2, an E3 Ubiquitin Ligase That Targets Transcription Factor Oct-4 for Ubiquitination. *Journal of Biological Chemistry*, 279(22), pp.23495–23503.

Yan, X., Liao, H., Cheng, M., Shi, X., Lin, X., Feng, X.-H. & Chen, Y.-G., 2016. Smad7 Protein Interacts with Receptor-regulated Smads (R-Smads) to Inhibit Transforming Growth Factor-β (TGF-β)/Smad Signaling. *The Journal of Biological Chemistry*, 291(1), pp.382–92.

Yan, X., Liu, Z. & Chen, Y., 2009. Regulation of TGF- signaling by Smad7. *Acta Biochimica et Biophysica Sinica*, 41(4), pp.263–272.

Yang, J., Mani, S.A., Donaher, J.L., Ramaswamy, S., Itzykson, R.A., Come, C., Savagner, P., Gitelman, I., Richardson, A. & Weinberg, R.A., 2004. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell*, 117(7), pp.927–39.

Ye, Y., Blaser, G., Horrocks, M.H., Ruedas-Rama, M.J., Ibrahim, S., Zhukov, A.A., Orte, A., Klenerman, D., Jackson, S.E. & Komander, D., 2012. Ubiquitin chain conformation regulates recognition and activity of interacting proteins. *Nature*, 492(7428), pp.266–70.

Ye, Y. & Rape, M., 2009. Building ubiquitin chains: E2 enzymes at work. *Nature Reviews. Molecular Cell Biology*, 10(11), pp.755–64.

Zarrinpar, A. & Lim, W.A., 2000. Converging on proline: the mechanism of WW domain peptide recognition. *Nature Structural Biology*, 7(8), pp.611–3.

Zhang, Y., Chang, C., Gehling, D.J., Hemmati-Brivanlou, A. & Derynck, R., 2001. Regulation of Smad degradation and activity by Smurf2, an E3 ubiquitin ligase. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3), pp.974–9.

Zhang, Y.E., 2009. Non-Smad pathways in TGF-beta signaling. *Cell Research*, 19(1), pp.128–39.

Zheng, N., Wang, P., Jeffrey, P.D. & Pavletich, N.P., 2000. Structure of a c-Cbl–UbcH7 Complex. *Cell*, 102(4), pp.533–539.

Zhou, P. & Wagner, G., 2010. Overcoming the solubility limit with solubility-enhancement tags: successful applications in biomolecular NMR studies. *Journal of Biomolecular NMR*, 46(1), pp.23–31.

Zhu, H., Kavsak, P., Abdollah, S., Wrana, J.L. & Thomsen, G.H., 1999. A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic pattern formation. *Nature*, 400(6745), pp.687–93.

Zilberberg, L., Todorovic, V., Dabovic, B., Horiguchi, M., Couroussé, T., Sakai, L.Y. & Rifkin, D.B., 2012. Specificity of latent TGF-β binding protein (LTBP) incorporation into matrix: role of fibrillins and fibronectin. *Journal of Cellular Physiology*, 227(12), pp.3828–36.