Structure, Dynamics and Complex Formation of Eukaryotic Transcriptional Regulators

Jordi Medina Vives

Barcelona, 2017

INSTITUT
DE RECERCA
BIOMÈDICA

IRB
BARCELONA

UNIVERSITAT DE
BARCELONA

Structure, Dynamics and Complex Formation of Eukaryotic Transcriptional Regulators

Jordi Medina

Structure, Dynamics and Complex Formation of Eukaryotic Transcriptional Regulators

Jordi Medina

PhD Thesis

INSTITUT
DE RECERCA
BIOMÈDICA
IRB
BARCELONA

Structural and computational
biology programme

UNIVERSITAT DE
BARCELONA

Facultat de farmàcia i ciències
de l'alimentació

Programa de doctorat en Biomedicina

Structure, Dynamics and Complex Formation of Eukaryotic Transcriptional Regulators

PhD Thesis by Jordi Medina Vives

Submitted in fulfilment of the requirements for
the degree of

Doctor per la Universitat de Barcelona

Jordi Medina Vives PhD Candidate	Maria J. Macias Hernandez, PhD Thesis Director ICREA research Prof.	Antonio Zorzano Olarte, PhD Thesis Tutor Prof. Faculty of Biology

Research performed at the
Structural Characterization of Macromolecular Complexes Laboratory,
Institute for Research in Biomedicine – IRB Barcelona

Barcelona, November 2017

Declaration

1. This work has been carried out in the Structural Characterisation of Macromolecular Assemblies laboratory, part of the Structural and Computational Biology program at the Institute for Research in Biomedicine – IRB Barcelona, under the supervision of its Principal Investigator Maria J. Macias (ICREA research professor).

2. No portion of the work referred to in this thesis as original has been submitted in support of an application for a degree qualification to any University.

3. Part of the work exposed in this thesis has been published in one peer-reviewed journal.

4. Financial support for the research included in this thesis has been provided by the IRB Barcelona International PhD Programme Fellowship and the Spanish National Research Program, Ministry of Economy and Competitiveness Grants SAF2011-25119 and BFU2014-53787-P.

Acknowledgements

First, I would like to thank my thesis director and principal supervisor, Prof. Maria J. Macias, for giving me the opportunity to carry out this research in her laboratory, for her support through these years and encouragement to persist on my goals.

I am very grateful to the members of my thesis advisory committee: Prof. Antonio Zorzano, who is also the tutor of this thesis, Dr. Jesus Garcia, Dr. Joan Pous and Dr. Montserrat Soler-Lopez. To all of them, I thank their advice during these years and their useful suggestions to improve the quality of the research presented in this thesis. I also want to thank Prof. Harold A. Scheraga and Dr. Gia G. Maisuradze for our fruitful collaboration. I also want to show my gratitude to all the researchers at the IRB Barcelona and the University of Barcelona that at some point have assisted me technically and provided their advice, with a special mention to Dr. Irene Fernandez and Dr. Gerardo Acosta.

The main support during this stage of my life have been my colleagues, many of which I consider my friends now. To begin with, I want to show my gratitude to Pau and Lidia, to you I owe you more than I can write here, both as a scientist and as a person. To Àngela, David, Jorge, Marta, Oriol and Tiago, I thank you the good moments and coffee times between EMSA and EMSA and the silly tunes in the lab. To Toni and Alessia, respectively my mentor and mentee in the world of peptide synthesis: thanks for your patience and the language lessons. To my Riera colleagues, Álvaro and Anna, for the funny times assigning and purifying. To Albert, Constanze, Eric and James: I am very grateful for your guidance through

the laboratory when I arrived. To Ewelina, Marco and Regina, thanks all the knowledge you shared with me.

I also want to thank my dear friends, Guille and Valèria, whom even if the distance separates us, I know I can rely on you. To Neus, Alba and Anna, whom I hope to see more often from now on. To my reunited friends from high school, Alícia, Ernest and Laura, let's do another room escape soon. To my favourite Spanish learners, Elisabetta, Lorenzo, Michal and Tomer, thanks for the great times (and meals!) together.

Finally, I am infinitely thankful to my family for all their endless support and warmth. To my sister, Laura: you know you are my favourite. To my special one, Antonio, I cannot imagine how these years would have been without you by my side. To my nephews, Marc and Clàudia, which bring me back the innocence and happiness of childhood. But most of all, to my parents, Joan and Rosa, thanks to be always by my side, to keep watching over me even if I am continuously complaining and for loving me.

To all of them I dedicate this thesis.

Table of Contents

The wise know their weakness too well to
assume infallibility; and he who knows
best knows how little he knows

Thomas Jefferson

Chapter 1

Introduction

1.1. Transcriptional Regulation via Smad Proteins in TGF-β Signalling

Cytokines belonging to the transforming growth factor β (TGF-β) superfamily are composed by about 35 different proteins in humans, including variant forms, which perform crucial regulatory roles related to the stringent control of cell growth, pluripotency and differentiation in metazoans. These are major biological processes involved in a multitude of systemic events, ranging from embryo development to immunity and tissue homeostasis in adult organisms. Due to the relevance of these events, a defective functioning of their key components is linked to severe conditions, namely cancer and metastasis (Massagué 2008; Chen and Dijke 2016). These disorders are triggered by anomalous epithelial-mesenchymal transitions (EMT) and mesenchymal-epithelial transitions (MET) that mobilise tumourigenic cells and spread the disease (Xu et al. 2009).

Members of the TGF-β superfamily determine cell fate by triggering specific mechanisms for the regulation of gene expression, which enhance or repress specific target genes in a highly context-dependent manner. The identity and number of affected genes is dependent on activating cytokine, cell tissue and development stage. The Smad family of transcription factors (TFs) is at the core of the intracellular effects of TGF-β signalling. According to the specific activated pathway (the precise membrane receptor and the Smad proteins involved in signal transduction) and sequence similarity, TGF-β cytokines are classified into two subfamilies, namely the BMP / AMH / GDF (bone morphogenic protein / anti-Müllerian hormone / growth and differentiation factor) and the TGF-β / Activin / Nodal (de Caestecker 2004). In short, these families will be referred to as the TGF-β- and BMP-activated subfamilies hereafter.

1.1.1. The Canonical Signalling Pathway of the TGF-β Superfamily of Cytokines

Despite some particularities, signalling by different members of the TGF-β superfamily presents a common transduction chain. Non-canonical forms differing from the one exposed below have been described, which all the same involve certain elements from the canonical pathway (Derynck and Zhang 2003).

The members of the TGF-β superfamily are secreted in a dimeric pro-peptide form hold together by a disulphide bridge. Upon cleavage, cytokines dimers can trigger the signalling cascade by binding to two different pairs of serine/threonine kinase membrane receptors (classified as type I and type II receptors), which will consequently interact with one another and lead to the formation of a hetero-tetrameric receptor complex **(Fig. 1.1A)**. While cytokines of the TGF-β subfamily directly bind to the type II receptor exclusively, in members of the BMP subfamily the preference depends mainly on their specific affinity for each type of receptors. The cytokine-driven oligomerisation of the receptor subunits leads to the activation of its cytosolic part by phosphorylation of the type I subunit by the constitutively active type II subunit. As a result, a repeated pS-X-pS (phosphoserine- any amino acid -phosphoserine) motif is created at the GS region

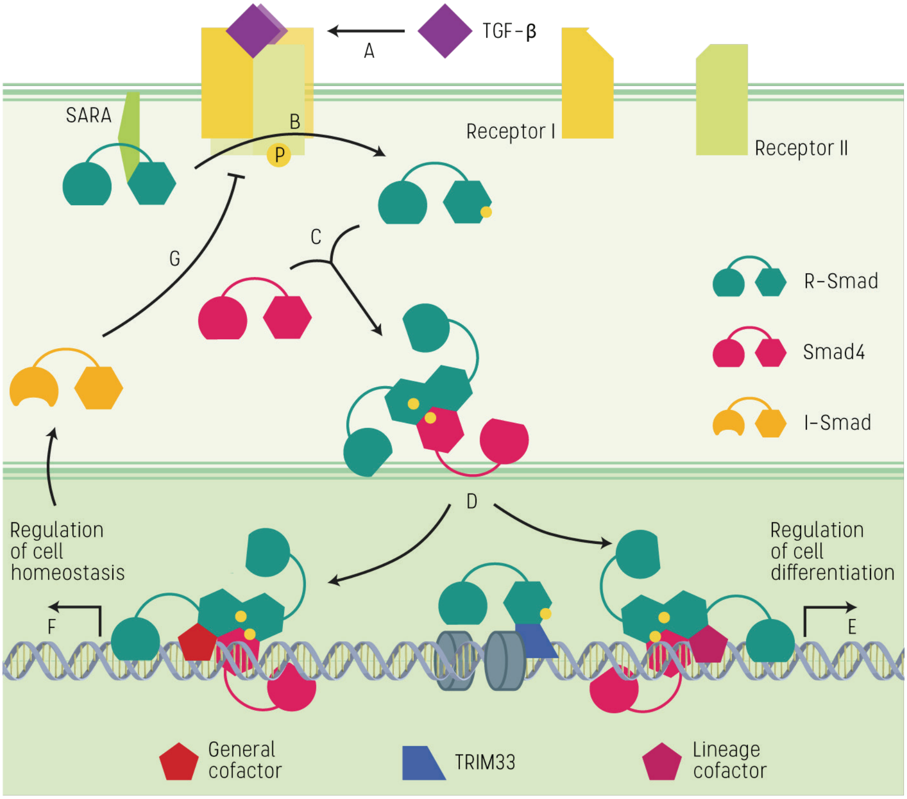


Figure 1.1. General Scheme of the Canonical TGF-β Pathway and its Potential Transcriptional Effects

The interaction of a TGF-β cytokine with its corresponding receptor subunit triggers the formation of the hetero-tetrameric complex and its activation by self-phosphorylation (**A**). SARA presents an R-Smad moiety to the active receptor, activating in turn the R-Smad by phosphorylation at the C-terminal (**B**). Active R-Smads are recruited by other R-Smad molecules and Smad4 (**C**). The hetero-trimer formed by two R-Smad and one Smad4 was described to be the main species to enter the nucleus. In the nucleus, the trimer acts as a transcription factor that conforms a platform for gene expression, recruiting other factors, cofactors and transcriptional machinery (**D**). TGF-β signalling regulates both differentiation (**E**) and homeostasis genes (**F**). As part of its transcriptional activity, Smad complexes activate a negative feedback response by inducing the transcription of I-Smads, which promote ubiquitin priming of membrane receptors (**G**).

found in the cytosolic N-terminal of type I subunits (Wrana et al. 1994; Shi and Massagué 2003; Ikushima and Miyazono 2010).

The main intracellular elements of the signalling pathway are the members of the Smad family of transcription factors. The Smad family is composed of eight different proteins in mammalians (from Smad1 to Smad8), and each is highly conserved in distantly related metazoans. Nevertheless, the overall number of family members is dependent on the species: four Smad homologs exist in *Drosophila* and only three in *C. elegans.* Smad proteins are classified according to their role in signal transduction (Massagué et al. 2005):

- ❖ **Receptor-activated Smads** (**R-Smads**). The first group of Smad proteins involved in signal transduction is composed by those members of the Smad family which are activated by active membrane receptors. In the basal state, R-Smads are found in the cytosol bound to membrane anchor proteins. These membrane anchors present two features that are essential for the regulation of Smad signalling: a structurally-disordered Smad binding domain (SBD) and a phospholipid recognition FYVE domain. Consequently, they facilitate the interaction of inactive R-Smads with activated membrane receptors by localising the former to the cell membrane, preferentially to early endosomes. When encountering an active membrane receptor complex, an R-Smad molecule binds to the pS-X-pS motif in the type I subunit and presents its own Ser-X-Ser C-terminal motif to the type II subunit for phosphorylation **(Fig. 1.1B)**. Once phosphorylated, the activated R-Smad is released from the anchor. Mammalian R-Smads activated by BMP-subfamily receptors are Smad1, Smad5 and Smad8, while Smad2 and Smad3 are activated by TGF-β-subfamily receptors.

- ❖ **Co-Smad or Smad4.** Smad4 is a unique member of the Smad family and hence, it is classified separately. Monomeric Smad4 is localised in the cytoplasm thanks to a nuclear export sequence (NES) in its linker. In the cell, Smad4 binds to receptor-phosphorylated R-Smads in the cytoplasm, forming hetero-dimers and hetero-trimers **(Fig. 1.1C)**. This oligomerisation is required for most gene responses to cytokine signalling. The current opinion is that trimers are predominant and are conformed by one molecule of Smad4 and two (potentially different) R-Smad molecules. Smad4 contains a nuclear localisation sequence (NLS) which allows the R-Smad-Smad4 hetero-trimer to enter the nucleus **(Fig. 1.1D)**. Once inside the nucleus, the Smad complex directs gene transcription by interacting directly with DNA plus other transcription factors and regulatory proteins, acting as a platform for gene transcription regulation (**Fig. 1.1E-F**). Nevertheless, whether the presence of Smad4 in the transcription complex is an indispensable requisite for all gene responses remains unknown.

❖ **Inhibitory Smads** (**I-Smads**). Members of this peculiar group, composed by proteins Smad6 and Smad7, are reported not to act as transcription factors, since they lack DNA-binding capabilities. Yet, I-Smads participate in the response triggered by TGF-β and BMP cytokines, acting as a negative feedback mechanism. Once expressed, Smad6 and Smad7 recruit E3 HECT ubiquitin ligases such as Smurf1 and Smurf2 to mark the different TGF-β and BMP membrane receptor complexes for degradation via the ubiquitin-proteasome pathway (**Fig. 1.1G**). Moreover, Smad6 has been reported to sequester BMP-activated R-Smads and Smad4 in the cytoplasm and similarly prime them for degradation (Kavsak et al. 2000; Yan et al. 2016).

Ubiquitin-proteasome degradation is also the main way to regulate TGF-β signalling termination in the nucleus. Alternatively, phosphatase-mediated R-Smad deactivation terminates Smad signalling by favouring R-Smad export to cytosol.

1.1.2. General Features of the Smad Transcription Factors

Smad transcription factors constitute a family of highly-conserved proteins in distantly-related metazoans that share main structural features, including domain arrangement. This layout generally consists in two Mad-homology domains, MH1 and MH2, coupled by a linker region (**Fig. 1.2**). Nevertheless, each Smad protein presents functional specificities and differs from other family members in protein length (Massagué et al. 2005).

The N-terminal MH1 domain is the smaller of the two and a DNA-binding β-hairpin that can interact with the major groove of GC-rich sequences found in regulatory DNA regions. The structure of the domain is stabilised by the coordination of a zinc ion. The DNA-binding properties of the MH1 are conserved

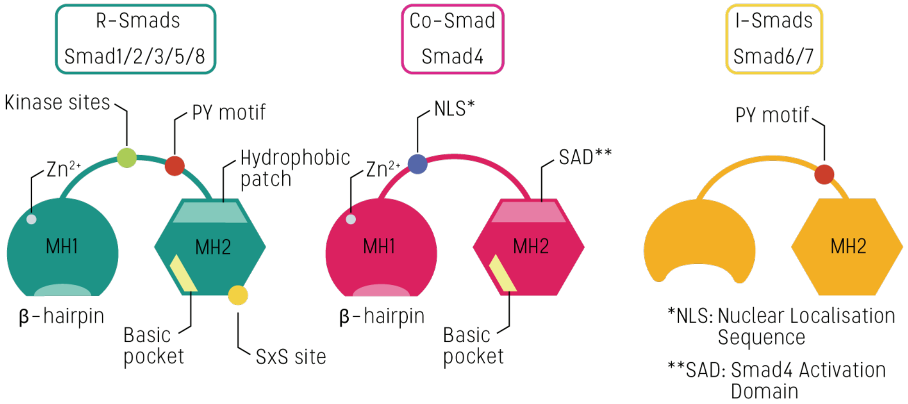


Figure 1.2. Domain Layout and Features of the Smad Transcription Factors

Smad proteins are conformed by two globular domains hold together by a linker region. In R-Smads and the Co-Smad, the MH1 domain presents a basic-rich β-hairpin responsible for DNA interaction. The MH2 domain of Smad proteins features several protein-protein interfaces participating in the interactions membrane receptors, other Smads or transcriptional machinery. The largely-unstructured linker presents regulatory motifs, including nuclear kinase phosphorylation sites and a WW-domain-recruiting PY motif that regulate Smad turnover.

in all R-Smads and Smad4, but not in the cytosol-located I-Smads. In the case of I-Smads, their MH1 domain is smaller than in the rest of the members of the Smad family, causing the loss of DNA-binding capabilities.

The C-terminal MH2 domain, on the other hand, acts as a hub for protein-protein interactions. It contains a region on its surface presenting a high proportion of hydrophobic residues, thus named the "hydrophobic patch" and first described in the structure of the Smad2 MH2 with the membrane anchor SARA (Wu et al. 2000). This hydrophobic patch is also predicted to be responsible for the interaction with other membrane anchors and a broad range of regulatory proteins and transcription factors described to partner with the MH2 domain. Moreover, in R-Smads, it contains the Ser-X-Ser activation motif that, once phosphorylated, is crucial in the formation of the hetero-trimeric complex with Smad4 by docking to the basic pocket present in both kinds of domains.

Connecting the MH1 and MH2 domains, a largely unstructured linker region, variable in length, acts as a regulatory target by presenting several motifs that determine Smad function and turnover through a series of phosphorylation *switches.* In the case of Smad4, it contains, as well, a nuclear export signal that participates in the nucleo-cytoplasmatic dynamics of the protein.

1.1.3. The MH1 Domain Binds to DNA Regulatory Sequences

The Smad MH1 is a highly-conserved domain across DNA-interacting Smads **(Fig. 1.3)** and metazoan species, presenting full sequence identity in its principal characteristic: the DNA-binding β-hairpin (with the exception of Smad2 in *D. melanogaster*). Oligonucleotide binding screens and the X-ray crystal structure of the MH1 domains of Smad3 and Smad4 showed that the MH1 domains of Smad 1/3/4/5 recognize with high affinity a GTCT DNA motif (or its complementary CAGAC) comprised in the palindromic sequence 5'–GTCTAGAC–3' (Shi et al. 1998; BabuRajendran et al. 2010). This motif is subsequently known as the Smad binding element (SBE). Later, Smad1 and Smad5 were shown to additionally recognise GC-rich motifs in certain BMP-responsive genes (Kusanagi et al. 2000; Morikawa et al. 2011), defining the regions as BMP response elements (BRE). This long-established difference in DNA motif recognition, striking given the sequence identity of the β-hairpin and the high conservation of the MH1 domain in R-Smad proteins and Smad4, has been recently refuted. A novel study has shown that Smad3 and Smad4 can bind to GC-rich DNA sequences *in vitro* with a similar affinity to the CAGAC SBEs with data support from ChIP-Seq results on the *Goosecoid* oncogene promoter (Martin-Malpartida et al. 2017).

Evidence suggests that the elements determining the exact response to BMP and TGF-β signals could be a set of master transcription factors poised on sites throughout the genome in a context-dependent manner, including differentiation stage and cell lineage (Mullen et al. 2011). Therefore, the regulation of distinct gene collections by Smad1/5/8 and Smad2/3 would be conditioned by the precise composition of each R-Smad-Smad4 heterotrimer and the ability of the transcription factors poised on the genome to recruit the Smad complex to distinct loci and determine gene expression.

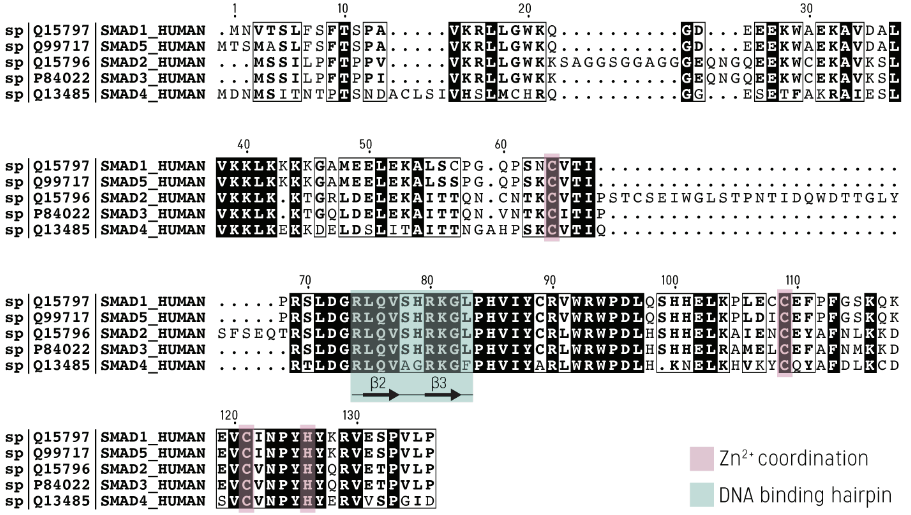

1
10
20
30
sp|Q15797|SMAD1_HUMAN .MNVTSLFSFTSPA.....VKRLLGWKQ..........GD...EEEKWAEKAVDAL
sp|Q99717|SMAD5_HUMAN MTSMASLFSFTSPA.....VKRLLGWKQ..........GD...EEEKWAEKAVDAL
sp|Q15796|SMAD2_HUMAN ...MSSILPFTPPV.....VKRLLGWKKSAGGSGGAGGGEQNGQEEKWCEKAVKSL
sp|P84022|SMAD3_HUMAN ...MSSILPFTPPI.....VKRLLGWKK..........GEQNGQEEKWCEKAVKSL
sp|Q13485|SMAD4_HUMAN MDNMSITNTPTSNDACLSIVHSLMCHRQ..........GG...ESETFAKRAIESL
40
50
60
sp|Q15797|SMAD1_HUMAN VKKLKKKKGAMEELEKALSCPG.QPSNCVTI.......................
sp|Q99717|SMAD5_HUMAN VKKLKKKKGAMEELEKALSSPG.QPSKCVTI.......................
sp|Q15796|SMAD2_HUMAN VKKLK.KTGRLDELEKAITTQN.CNTKCVTIPSTCSEIWGLSTPNTIDQWDTTGLY
sp|P84022|SMAD3_HUMAN VKKLK.KTGQLDELEKAITTQN.VNTKCITIP......................
sp|Q13485|SMAD4_HUMAN VKKLKEKKDELDSLITAITTNGAHPSKCVTIQ......................
70
80
90
100
110
sp|Q15797|SMAD1_HUMAN .....PRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLECCEFPFGSKQK
sp|Q99717|SMAD5_HUMAN .....PRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLDICEFPFGSKQK
sp|Q15796|SMAD2_HUMAN SFSEQTRSLDGRLQVSHRKGLPHVIYCRLWRWPDLHSHHELKAIENCEYAFNLKKD
sp|P84022|SMAD3_HUMAN ......RSLDGRLQVSHRKGLPHVIYCRLWRWPDLHSHHELRAMELCEFAFNMKKD
sp|Q13485|SMAD4_HUMAN ......RTLDGRLQVAGRKGFPHVIYARLWRWPDLH.KNELKHVKYCQYAFDLKCD
β2
β3
120
130
sp|Q15797|SMAD1_HUMAN EVCINPYHYKRVESPVLP
sp|Q99717|SMAD5_HUMAN EVCINPYHYKRVESPVLP
sp|Q15796|SMAD2_HUMAN EVCVNPYHYQRVETPVLP
sp|P84022|SMAD3_HUMAN EVCVNPYHYQRVETPVLP
sp|Q13485|SMAD4_HUMAN SVCVNPYHYERVVSPGID
Zn²⁺ coordination
DNA binding hairpin

Figure 1.3. Sequence Alignment of the DNA-Binding Smad MH1 domains

The overall structural ensemble and the two main features of the MH1 domain, its DNA-binding hairpin and the cysteine and histidine residues required to coordinate a zinc ion, are maintained across all DNA-binding Smad proteins.

Post-translational modifications (PTMs) have also been shown to modulate the binding activity of the MH1 domain. The best described of these PTMs is acetylation of Lys19 in Smad3 and the Smad2(ΔE3) isoform, which has been shown to promote transcriptional activation by favouring DNA binding as seen both *in vitro* and *in vivo* (Simonsson et al. 2006).

Structurally, the Smad MH1 domain consists of four α-helixes and six β-strands plus a zinc ion coordinated by a His-Cys(3x) tetrad. The β-strands conform three anti-parallel pairs: β_1-β_5, β_2-β_3, and β_4-β_6, of which the β_2-β_3 hairpin is responsible for DNA binding. Specifically, three strictly conserved residues (Arg74 and Gln76 in β_2 and Lys81 in β_3) participate in a network of hydrogen bonds with the dsDNA. Other interactions, such as those of residues around the β_2-β_3 hairpin with the DNA phosphates or other water-mediated interactions help to stabilise the complex (Shi et al. 1998).

1.1.4. The Smad Linker Region Is a Ubiquitination Target that Regulates Smad Turnover

The inter-domain linker region has a fundamental role on the regulation of transcriptional activity and turnover in Smad proteins. A concatenation of (Ser/Thr)-Pro residues placed along the linker act as phosphorylation-dependent switches that generate docking sites for nuclear coactivators and repressors of the BMP and TGF-β response. The Smad linker also contains a PPxY sequence (or PY motif) that is recognised by partner proteins containing WW domains (see subsection 1.2.2 for more information on WW domains).

Once in the nucleus, the R-Smad linker undergoes two rounds of phosphorylation events. First, cyclin-dependent kinases CDK8 and CDK9 (components of the Mediator and Elongation complexes, respectively) recognise and phosphorylate specific serine and threonine residues in the linker, favouring the interaction of R-Smads with transcriptional coactivators and hence promoting transcriptional activity (Matsuura et al. 2004). For instance, CDK8/9 phosphorylation of Smad1 at S206 and S214 creates binding sites that are recognised by the WW domains of the Yes-associated protein 1 (YAP1), a versatile transcriptional effector downregulated by the Hippo pathway (Alarcón et al. 2009).

Besides, CDK8/9-mediated phosphorylation is a priming event for a second phosphorylation round directed by glycogen synthase kinase 3 (GSK3). GSK3 phosphorylation causes a switch in the interacting properties of the linker region, converting it into a binding site for HECT E3 ubiquitin ligases that mark R-Smads for proteasome-mediated degradation (Gao et al. 2009; Aragón et al. 2011). R-Smad-interacting E3 Ubiquitin ligases present at least one WW domain which is able bind to the PY motif in the vicinity of the linker phosphorylation sites. In the case of CDK8/9-phosphorylated Smad1, additional phosphorylation by GSK3 of amino acids T202 and S210 switches the Smad1 binding preference from YAP1 to the HECT E3 ubiquitin ligase Smurf1 owing to a change in the binding affinity.

On the other hand, the CDK8/9 phosphorylation sites in the Smad linker region also function as sites for mitogen-activated protein kinases (MAPKs) in response to stress-related signalling pathways (Fuentealba et al. 2007).

1.1.5. The MH2 Domain Is the Smad Hub for Protein–Protein Interactions

The MH2 domain mediates the interaction of R-Smads and Smad4 with a large number of proteins in the cell. In the case of R-Smads, these include, non-exclusively, the BMP and TGF-β membrane receptors, other R-Smads after phosphorylation at the SxS motif, the SARA cytoplasmic anchor, chromatin methylation readers such as TRIM33 and multiple DNA-binding cofactors (Massagué et al. 2005; He et al. 2006). Regarding Smad4, apart from the trimer formation with R-Smads, the best-described interaction is that with c-SKI/SnoN (Luo 2004).

In general, the overall Smad2 MH2 structure contains a central β-sandwich formed by two sets of five and six anti-parallel β-strands flanked by a triple helix bundle (H3, H4 and H5) at one end and a loop–helix region (L1, L2, L3 and H1) at the other (Shi and Massagué 2003). The main structural difference respect to the MH2 domain in the Co-Smad is a large flexible linker between α-helixes H3 and H4 in the triple-helix bundle (Shi et al. 1997).

As introduced previously, membrane anchors (such as SARA in TGF-β signalling or endofin in BMP signalling) are responsible to present R-Smads to the corresponding membrane receptor and potentiate signalling transmission by means of a direct interaction involving the MH2 domain (Tsukazaki et al. 1998; Shi et al. 2007). This role is defined by two elements: a lipid recognition FYVE domain, that enables SARA to bind to phosphatidylinositol triphosphate (PIP_3) lipid molecules and localise R-Smads to the cell membrane, and the structurally disordered Smad binding domain (SBD), which is described to be about 80 residues long and is responsible for the interaction between the anchor and the R-Smads through the hydrophobic patch (Wu et al. 2000; **Fig. 1.4**).

A region adjacent to the L3 loop in the MH2 domain of R-Smads and rich in arginine and lysine residues, the basic pocket, has been suggested to be the binding site for type I receptor subunits. This positively-charged patch is also predicted in the Smad7 MH2 domain, potentially explaining the basis for I-Smad-dependent receptor proteolysis. The basic pocket region in R-Smads is also responsible for the binding to the pS-X-pS activated motif on a neighbouring MH2 domain for trimer formation with Smad4. Moreover, additional contacts include an interface formed by a set of large loops and the first α-helix on one side of a MH2 domain interact with a triple-helical bundle of a neighbouring MH2 domain, following a head-to-tail arrangement with threefold symmetry (Chacko et al. 2004).

The hydrophobic patch and adjacent regions have also been described to be involved in most interactions of the MH2 domain with nucleoporins and DNA-binding partners in the nucleus (Xu et al. 2002; Randall et al. 2002). These

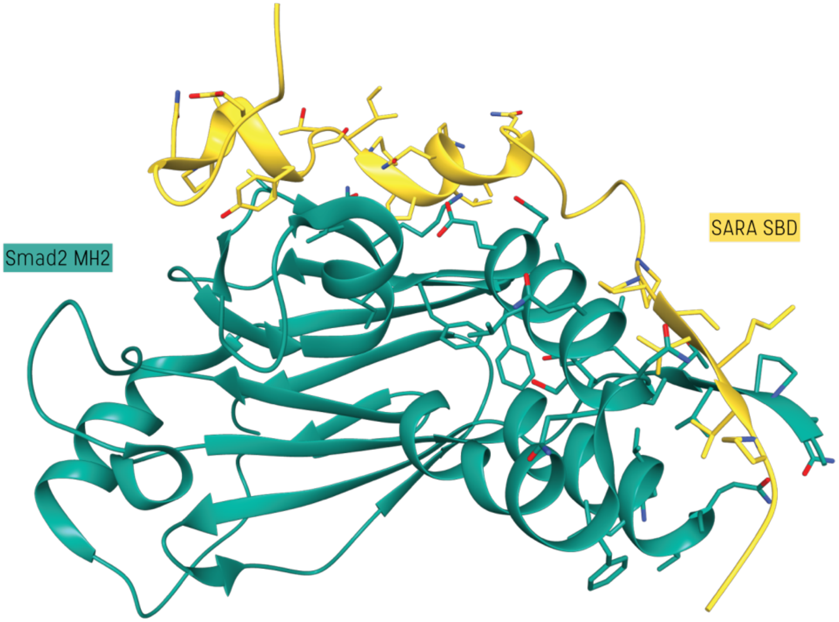


Figure 1.4. Smad2 MH2 domain interaction with the SARA Smad-binding domain (SBD)

A region of the membrane anchor SARA interacts with the MH2 domain of the TGF-β Smad2 protein by means of two structural features: an N-terminal loop, a short α-helix and a β-strand. Interestingly, a structure-upon-binding model has been proposed for the SARA SBD. Surprisingly, the N-terminal β-strand of the Smad2 MH2 does not appear in other structures of the same domain, yet here it is forming an antiparallel β-sheet with SARA.

transcription factors can help the trimeric Smad complex to bind to specific promoters and enhancers on the DNA. Examples of sequence-specific transcription factors reported to interact with an R-Smad MH2 domain are FoxH1, FoxO proteins, and Sp1. In addition, a high number of transcription co-regulators (co-factors and co-repressors, which act as auxiliary transcription modulators) participate in the regulation of Smad-directed transcription via interaction with the MH2 domain, including co-factors NCOA6, CBP/p300 and ARC105, and co-repressors c-Myc, TGIF and the Smad4 specific c-SKI/SnoN. The members of former group act by recruiting transcriptional machinery to the Smad-bound promoter to start transcription, while those proteins from the latter group either repress transcription in the basal state (and are removed by the Smad hetero-trimer) or actively repress transcription by affecting specific steps of the Smad signalling pathway (Macias et al. 2015; Hill 2016).

The aforementioned protein-protein interactions are determinants of the specificity of TGF-β signalling at several levels and thus, they are of great relevance to better understand the functioning of the Smad-related regulation activity. Yet, many are still poorly characterised, particularly at the structural level. The structure of only three of these interactions has been solved: the domain organisation upon formation of the R-Smad-Smad4 heterotrimer (Chacko et al. 2004), the structure of the Smad2 MH2 interaction with the SARA SBD fragment (Wu et al. 2000) and the Smad4-Ski complex (Wu et al. 2002).

Additionally, a greater number of potential Smad-interacting proteins have been described through ChIP-seq studies. Master transcription factors OCT4, SOX2, MYOD and PU.1 have been co-localised with Smad3 in regulating sequences in the genome (Mullen et al. 2011). Still, proof of direct binding for many of these factors to Smad proteins is yet to be found. Another explanation for the co-occupation of genome regions is a potential role of these master transcription factors in the creation of favourable chromatin conditions for Smad DNA-binding (Leichsenring et al. 2013).

The Smad MH2 domain is highly conserved across species and human SNPs are scarce and highly conservative. Thanks to modern medical procedures, including routine sequencing of tumour tissue samples, a large number of disease-related mutations have been described, particularly in the Smad4 MH2 (Shi et al. 1997). Many of such mutations are localised on trimer interfaces and in residues that are crucial for the formation of secondary structure elements. Other mutations in R-Smads include protein truncation, preventing the active trimer formation since it relies on the C-terminal SxS motif.

1.1.6. FoxH1 Regulates the Expression of Cell-Lineage-Defining Genes in Cooperation with Smad Proteins

FoxH1, also known as Fast-1, is a member of the winged-helix/forkhead family of transcription factors. As such, it presents a forkhead (FH) DNA-binding domain, a roughly-110-residues-long sequence that binds to DNA in a manner similar to histone H5 **(Fig 1.5)**. Structurally, the forkhead domain presents a particular winged-helix structure, very similar to a helix-turn-helix structural motif, that owes its name to the resemblance of its tri-dimensional structure to a butterfly (Brennan 1993). The human consensus binding sequence for the forkhead domain is (C/G)AAT(A/C)CACA(A/T), found in all identified FoxH1-regulated promoters (Yeo et al. 1999; Attisano et al. 2001). Even if not all bases directly participate in the interaction with FoxH1, the consensus sequence includes all those nucleotides which are critical for maximum transcriptional activity.

FoxH1 has a fundamental role in the early stages of embryo development by directing cell differentiation in endoderm and mesoderm progenitors. The expression of FoxH1 is ubiquitous during the gastrulation phase and is progressively restricted until day 8.5 in mouse embryos (Weisberg et al. 1998). The ability of FoxH1 to modulate gene transcription is not intrinsic. Instead, transcriptional regulation is dependent on the recruitment of Smad2/3-Smad4 complexes by FoxH1 moieties bound to promoter and enhancer elements in the context of a nodal/activin signalling response (Chen et al. 1997; Wotton et al.

1999). Nevertheless, the greater severity of a nodal depletion phenotype when compared to FoxH1 depletion suggests that additional Smad2/3-recruiting transcription factors may exist (Slagle et al. 2011).

Genes regulated in early embryo cells by the Smad-FoxH1 complexes include *goosecoid* (*Gsc*) and *Xbra* (the *Xenopus T* gene homolog) in mesendoderm progenitors (Labbé et al. 1998; Watanabe and Whitman 1999). The TF-assisted Smad regulation of gene activation and expression defines a paradigm as supported by studies on other key development and differentiation gene targets such as *GATA5* in erythroid progenitors and *SPI1* and *CEBPA* in myeloid progenitors (Morikawa et al. 2013). These observations provide an explanation on how cells go through distinct differentiation transition events under the control of BMP and TGF-ꞵ cytokines.

The different members of the winged-helix/forkhead family and the different homologs of FoxH1 present low sequence identity. In the latter group, the two main conserved elements are the forkhead domain itself and a C-terminal proline-rich region, which was found to be crucial for binding to the MH2 domains of Smad2 and Smad3 (Chen et al. 1997). Together with the SBD in SARA, the SIM (Smad interaction motif) in FoxH1 was one of the first sequences discovered to interact with the MH2 domain of R-Smads. Even though both interacting motifs have been compared in the literature owing to their relatively high ratio of prolines, the sequences share low identity and greatly differ in the length described as necessary for the interaction (Randall et al. 2002). Studies on the SIM-MH2 interaction have indicated the presence of a PPNK motif within SIM which is responsible for the direct interaction with the MH2 domain (Germain et al. 2000). Moreover, later work suggests the presence of an additional Smad-binding motif in FoxH1 named FM (FoxH1 motif), which is described to bind to the Smad2 MH2 on a surface area that partially overlaps with the SIM binding site. Consequently, it was proposed that the presence of two Smad-interacting areas would allow a single FoxH1 moiety to interact with the two R-Smads conforming the heterotrimeric Smad complex (Randall et al. 2004).

Additionally, Smad-independent roles for FoxH1 have been described. A Smad-independent indirect interaction of FoxH1 with the androgen receptor (AR) was described to repress the trans-activation of AR-regulated genes (Chen et al. 2005). On the other hand, the activation of the nodal gene Xnr3 in *Xenopus* embryo cells was shown to require the presence of FoxH1 and XTcf3 (homolog to the human transcription factor β-catenin) in a Smad2-independent manner (Kofron et al. 2004).

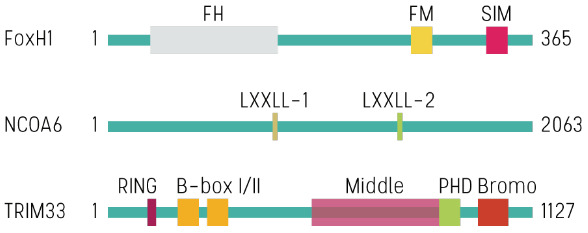


Figure 1.5. Layout of different Smad2 MH2 nuclear partners

The domain and motif layout of the MH2-binding proteins transcription factor FoxH1, transcriptional co-activator NCOA6 and ubiquitin-ligase and chromatin reader TRIM33. The undefined TRIM33 middle shown as shadowed.

1.1.7. General Transcriptional Co-activator NCOA6 Mediates TGF-β Regulation of Cholesterol Metabolism

NCOA6 (Nuclear receptor coactivator 6, also referred to as RAP250, ASC-2 or PRIP) is a member of the NCOA1 family of co-activators, a well-studied group of proteins that acts concertedly with nuclear receptors (NRs) and, potentially, other transcription factors for the promotion of gene transcription (Caira et al. 2000).

In general, co-activators participate in the organisation of multiprotein complexes in the nucleus and they can be involved in the different layers of gene regulation, including chromatin remodelling by ATP-dependent, SWI/SNF-containing complexes through co-activators with intrinsic histone acetyltransferase activity (Roberts and Orkin 2004), general co-activators such as CBP2 and p300 (Hermanson et al. 2002) and the mediator complex that recruits RNA polymerase II transcription machinery (Malik and Roeder 2005). In the specific case of

NCOA6, despite its lack of any intrinsic enzymatic activities, it can potentially recruit proteins with histone acetyltransferase, methylase and helicase activity (Ko et al. 2000; Froimchuk et al. 2017; Zhu et al. 2001; Ju et al. 2006).

NCOA6 contains of two LXXLL motifs (or NR-box motifs), which are leucine-rich sequences that interact with nuclear receptors, and a glutamine-rich activation motif required to perform as a transcriptional co-activator **(Fig 1.5)**. Remarkably, the two LXXLL motifs present different interacting properties. The first LXXLL motif (LXXLL-1, close to the activation motif) is reported to bind to most nuclear receptors, as is the case for most NR-boxes in NCOA1 family members. NCOA6 has been reported to participate in several biological processes through its partnership with nuclear receptors such as hormone receptors. For example, it has been shown that NCOA6 is crucial for insulin secretion in pancreatic β-cells through sterol receptor SREBP-1c (Yoon et al. 2017). Besides, several studies have shown NR-independent transcription promotion by NCOA6 through unclear mechanisms of action.

On the other hand, the C-terminal terminal LXXLL-2 motif is highly selective regarding its recognition of nuclear receptors. It has been described to interact only with liver X receptors (LXRs) and crucially determining LXR signalling (Lee et al. 2001; Kim et al. 2003). Later on, a distinct role for a LXLL-2-containing sequence was identified: it was seen to interact with the MH2 domain in Smad2 and Smad3 (Antonson et al. 2008), thus revealing the participation of NCOA6 in TGF-β signalling.

A relevant example of this dual role of NCOA6 is cholesterol metabolism. LXRs are responsible for the promotion of reverse cholesterol transport, a role that is key in in the evolution of atherosclerotic lesions involving macrophages. Other studies have shown that TGF-β participates as well in atherosclerosis-related processes, including the accumulation of lipids in the vessel wall (Steffensen and Gustafsson 2006).

1.1.8. TRIM33 Recognises Histone Modifications and Regulates Smad-Driven Gene Transcription

E3 ubiquitin-protein ligase TRIM33 (known also as TIF1γ and Rfg7) has been proposed to be of great importance in the access of Smad proteins to poised promoters of master regulators such as lineage transcription factors under the command of TGF-β signals. The most well-studied role of TRIM33 is in the activation of genes driving differentiation of haematopoietic stem cells (HSC) as seen in both human and zebrafish cells (He et al. 2006; Bai et al. 2010; Quéré et al. 2014). It is also essential in the differentiation of mammary alveolar epithelial cells (Hesling et al. 2013). On the other hand, TRIM33 is not required in homeostatic responses.

TRIM33 is part of the TRIM family of proteins, characterised by an N-terminal tripartite motif consisting of a RING E3-ubiquitin ligase domain and two B-box domains of unknown function. In addition, TRIM33 presents a C-terminal plant homeodomain - bromodomain tandem (PHD-Bromo), responsible for the recognition of epigenetic marks in histone H3 tails, and a ~350-residues-long, mostly unstructured middle region between both domain clusters **(Fig 1.5)**. It is within this middle region where the interaction motif with Smad2/3 MH2 domain has been described (He et al. 2006).

Two different mechanisms of action have been suggested to describe the role of TRIM33 in TGF-β-directed gene regulation. First, a study in mouse embryo stem cells (ESCs) described the recognition by TRIM33 of histone H3 modification Lys9 tri-methylation (K9me3) and Lys18 acetylation (K18ac), which would then cause the displacement of the heterochromatin-defining protein HP1 and subsequently, a remodelling of the chromatin favouring gene expression (Xi et al. 2011). According to this model, TRIM33 would be poised on transcriptionally-repressed chromatin to participate in a potential acute activation of gene expression by specific developmental signals, similarly to other master regulators of differentiation such as Oct4, Sox2, and Nanog (Young 2011). Poised TRIM33 molecules would recruit activated Smad2/3 to upstream enhancer regions of *Gsc*, *Mixl1* and other mesendoderm genes by means of the MH2-middle interaction. The reported R-Smad-TRIM33 complex would regulate gene transcription in a Smad4-independent manner.

In a different study, TRIM33 was shown to recognise histone H3 acetylation marks at Lys18 and Lys23. TRIM33 has been shown to monoubiquitinate the Smad4 MH2 domain by its E3 ubiquitin ligase activity (Dupont et al. 2005), disrupting the R-Smad-Smad4 heterotrimer and potentially priming Smad4 for proteasomal degradation. It is suggested as well that TRIM33 presents an inactive state where intramolecular interactions between its N-terminal RING domain and the PHD finger-bromodomain self-inhibit the E3 ubiquitin ligase activity, which is therefore activated upon chromatin binding (Agricola et al. 2011). In this model, TRIM33 regulates the residence time of Smad complexes on the chromatin.

The two exposed models have conflicting points, especially on the role of Smad4, and coincidences like the incompatibility of TRIM33 activity and Lys4 modification. Even if both models are contradictory, since Smad complexes recruit not only transcription factors but other regulators like histone acetyl-transferases (HATs) and histone deacetylases (HDACs), a potential explanation for the different reported mechanisms could be that the precise TRIM33 activity is dependent on histone H3 acetylation and methylation marks. Therefore, it is plausible that K9me3 and K23ac epigenetic marks would somehow determine the role of TRIM33 in TGF-β-mediated cell differentiation.

1.2. WW Domains: Folding and Aggregation

1.2.1. Introduction to Protein Folding

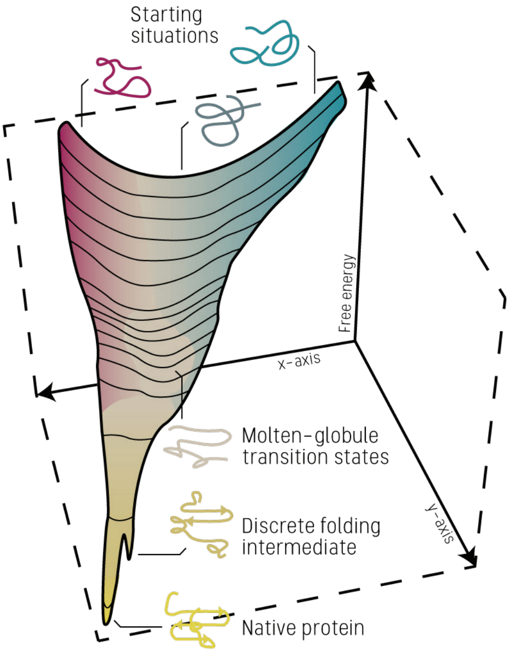
Proteins are linear polymers of concatenated amino acids which are produced by ribosomes either in the cytosol or in the endoplasmic reticulum (ER). Newly synthesised chains have the ability to self-assemble with high precision into a defined, reproducible tri-dimensional structure that is known as the protein's native state. Nonetheless, many details on the folding process remain elusive and it is not fully understood yet how a disordered chain can achieve a functional native state only with the information in its amino acid sequence (that is, its primary structure) nor how folding-related diseases arise and propagate.

It is widely accepted that polypeptide chains undergo a stochastic exploration of the many conformations available to each sequence (Wolynes et al. 1995). Provided that native-like interactions between residues are more persistent than randomised non-native contacts, the former are able to direct folding to the most thermodynamically-stable structure by trial and error. A stochastic formation of contacts provides an explanation on why folding occurs in the nanosecond-to-second timescale range despite the enormous number of potential folds and regardless of the high variability in amino acid composition and length between protein chains (Mayor et al. 2003).

The free energy landscape (FEL) of a protein **(Fig. 1.6)** provides a graphical representation of the different folding trajectories it can follow typically by plotting structural variables against free energy (Leopold et al. 1992). Due to their most common form, FELs are usually named *protein funnels*. Stochastic contacts in the peptide chain start to lead the folding, generating a series of starting structures. New structures are formed from selected native contacts and non-native transient contacts that reduce the overall free energy until a transition quasi-globular state (known as a molten globule state) is formed from key native-like contacts (Baldwin and Rose 2013). This state presents secondary structure elements conforming the overall topology of the native fold but with a high degree of disorder and lacking a condensed, packed protein core. Mechanisms such as hydrophobic collapse or interdigitating side chains help to conform the protein core and exclude water molecules, forming the final native structure (Cheung et al. 2002). However, intermediate folds may exist in local energy minima. Some of these intermediate states have been described to have relevant roles in biological processes (Yamada et al. 2013).

Figure 1.6. Simplified representation of the folding free energy landscape of a protein

Linear polypeptides are not the most optimum species thermodynamically. For this reason, proteins generally start to fold immediately after synthesis, in some cases even co-translationally (as they leave the ribosome). Free energy landscapes (FELs) are multidimensional plots representing the folding *funnel* (the section of a cartoon version is represented on the right), and describe the characteristics of the different structures a peptide chain can take along a series of variables. These generally include free energy and other not-so-established variables such as the number of inter-residual contacts and native-like interactions. At the top of the funnel, starting structures formed by quasi-aleatory contacts can be found. By selection of native-like contacts, the free energy of the polypeptide chain will decrease and the range of total residue contacts will gradually shrink. The graphic description provided by FELs help to describe folding trajectories and the possible transition and intermediate states of a protein.


Starting situations
Free energy
x-axis
y-axis
Molten-globule transition states
Discrete folding intermediate
Native protein

A derivative from the exposed vision on folding is that the FEL of a protein is in fact encoded in its primary structure and hence it is responsible for the greatest part of its folding determinants. Consequently, protein sequences have been under evolutionary pressure to ensure that proteins fold rapidly and efficiently as well as to avoid amino acid sequences (those alternating hydrophobic and polar residues, for instance) that favour structures seen in malfunctioning proteins such as amyloid-like β-sheets (Broome and Hecht 2000). An evolutionary standpoint on protein sequences is further supported by the existence of hereditary diseases caused by misfolded mutational variants (Horwich 2002).

Biologically, cells keep a tight control on the correct folding of proteins. These mechanisms are crucial to maintain homeostasis, since an undesirable folding may cause incorrect interactions with other biomolecules deriving in harmful effects. Molecular mechanisms assisting proper protein folding are commonly classified in two groups: chaperones, which are a highly heterogeneous series of ATPconsuming protein complexes that scaffold and provide newly-synthesised chains with favourable conditions to fold correctly (Hartl et al. 2011), and folding catalysts, mostly isomerases such as prolyl isomerases and disulphide isomerases, which correct the cis-trans isomerisation of prolines and enhance the formation and reorganisation of disulphide bridges, respectively (Schiene and Fischer 2000).

1.2.2. Structure and Function of WW Domains

WW domains, a name derived from two characteristic tryptophan amino acid residues present in their sequence, consist on triple-stranded anti-parallel β-sheets flanked by two unstructured regions that tend to interact and stabilise the folding, a characteristic observed in other signalling domains as SH2 and PH (**Fig. 1.7**). WW domains are comprised of approximately 40 amino acids, which makes them the smallest type of independently-folding globular domain found in living organisms without the assistance of disulphide bridges. The structure formed by this short sequence is hold together thanks to a hydrophobic core extended along the sequence and composed by a conserved tryptophan / tyrosine / proline cluster (Macias et al. 1996; Ranganathan et al. 1997).

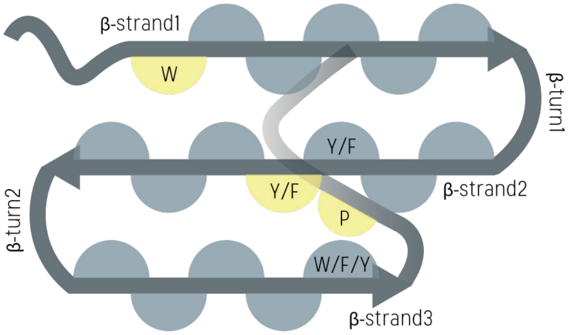


Figure 1.7. Layout of a WW domain

Arrows picture the characteristic anti-parallel β-sheet folding of WW domains. Amino acid residues along the sequence are represented as semicircles pointing to one side of the β-sheet plane according to side-chain orientation. Residues conforming the hydrophobic core (in yellow) and conserved aromatic positions are labelled.

Functionally, WW domains have a significant role in intracellular signalling through their participation in protein-protein interactions. These domains are involved in the functional regulation of protein partners by recognising proline-containing short peptide motifs[1] (Macias et al. 2002). The short length and a certain flexibility in the precise motif sequence make these domains highly versatile (Kato et al. 2002, 2004). Accordingly, WW domains are classified by the motifs that they are capable to recognise **(Table 1.1)**.

WW-domain-containing proteins have been described to be involved in a miscellany of biological processes, from RNA transcription regulation to protein localisation (Sudol et al. 2001; Ingham et al. 2005). Moreover, a myriad of partners has been identified, defining an expanded network of contacts and associated biological processes. As a result, a relevant number of human diseases are related to alterations in proteins containing WW domains.

Table 1.1. Classification of the WW domains by binding motif

WW domain groups	Binding surface	Recognised motifs	
Group I	xP and Tyr grooves	PPxY	PY motif
Group II/III	xP and xP2 grooves	PP(L/R/P)Px	PL /PR /PP motifs
Group IV	xP groove and P patch	(pS/pT)P	pSP /pTP motifs

1.2.3. WW Domains and Amyloid-Like Aggregation: The Second WW Domain of Formin Binding Protein 28

Formin binding protein 28 (FBP28) is the murine homologous to Transcription elongation regulator 1 (TCERG 1; also known CA150), a transcription factor that represses transcription by interacting with RNA polymerase II through its FF repeat domains and additionally by interacting with the splicing-transcription factor SF1 through its WW1 and WW2 domains. Both interactions are needed for an efficient transcription repression (Goldstrohm et al.

[1] The functionality of the Smad linker region, as described in subsection 1.1.4, is one of many examples where the role of WW domains and their specificity in the recognition of proline-rich motifs participate on the regulation of cellular events.

2001). It was proposed that both the WW1 and WW2 bind somewhere in a proline-rich, 80-amino-acid-long region within SF1, with a novel PPPxQ motif being suggested to be responsible for the interaction. Binding to peptides containing the aforementioned sequence was independently corroborated by NMR-monitored titrations (unpublished results). Later work on FBP28 demonstrated that these domains were also able to bind PPLPx motifs (Ramirez-Espain et al. 2007), indicating that FBP28 WW1 and WW2 domains belong to the group II/III.

Structurally, the FBP28 WW2 domain is part of a subgroup of WW domains lacking an N-terminal proline. Moreover, even though WW domains are considered to be very flexible sequence-wise, FBP28 WW2 presents highly uncommon amino acid residues in positions 452, 453 and 455 – located in the domain's β-turn2. Nevertheless, the domain presents the WW canonical triple anti-parallel β-sheet tertiary structure (Macias et al. 2000; **Fig. 1.8**).

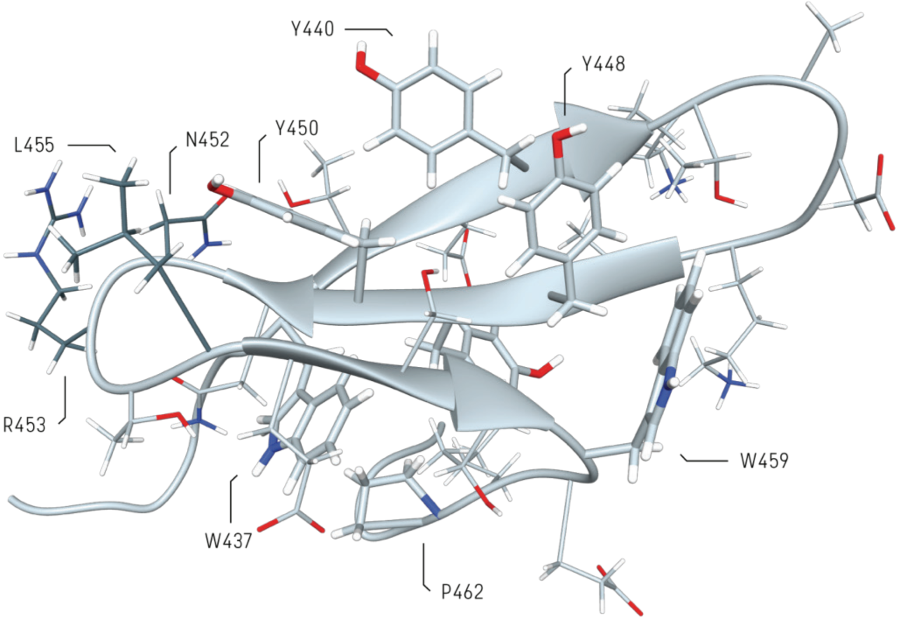


Figure 1.8. Structure of the FBP28 WW2 domain

Lowest-energy structure from the FBP28 WW2 domain structure ensemble (PDB ID: 1E0L). Thicker bond lines represent the most conserved and structurally crucial amino acid residues and a contrasting colour is used to display uncommon residues in positions in the β-turn2.

Due to their small size and secondary structure composition as well as their ability to generally fold correctly after mutagenesis in most positions along their sequence, WW domains have been vastly studied in both experimental and computational research on β-sheet folding. Particularly, the FBP28 WW2 domain has been studied in detail mostly owing to two particularities.

First, a vivid debate arose in the early 2000 about the folding kinetics of the FBP28 WW2 domain as seen experimentally and in different molecular dynamics (MD) simulation studies. A first report (Ferguson et al. 2001) studied the unfolding of single-point and truncation mutants by chaotropic agents using near-UV and far-UV circular dichroism (CD) and concluded that a single exponential kinetics (via a two-state mechanism) fast folding occurred. On the other hand, a later report (Nguyen et al. 2003) using an extended number of mutant domains, folding properties were characterised using temperature denaturation instead and CD data was complimented with nanosecond-resolution laser temperature-jump. The results indicated the presence of a folding intermediate at temperatures below T_m. This intermediate state could be tuned to an apparent two-state folding at higher temperatures and diminished in certain mutant constructs. Succeeding works leant towards the three-state hypothesis, with computational biologists taking an interest on the topic since a small domain like FBP28 WW2 provided good testing grounds for MD simulations. The most concurred explanation for the emergence of the interim state is the differential timescale for the register of the domain's β-hairpins (Karanicolas and Brooks 2003; Mu et al. 2006; Zhou et al. 2014; Davis and Dyer 2014).

The second factor of interest was the particularity of FBP28 WW2 to form fibrils similar to those found in amyloid-related diseases such as Alzheimer's under physiological conditions. These fibrils present nucleation kinetics and histochemical properties – Congo red birefringence and thioflavin-T binding – identical to those of pathological amyloids. X-ray and electron microscope structural characterisation of the fibrils revealed the characteristic amyloid diffraction pattern and intertwined amyloid fibres, even though the plates as seen by cryo-EM are wider than in the canonical amyloid forms (Ferguson et al. 2003).

MD simulation studies on FBP28 WW2 fibril formation have proposed the formation of aggregation-inducing dimers, generating a 10-unit protofibril model (Mu et al. 2006).

Folding intermediates have been implicated in amyloid fibril formation (Jahn et al. 2006; Neudecker et al. 2012). Eventually, both phenomena were found to be related also in the case of FBP28 WW2. It has been proposed that the slow register of β-hairpin2 that gives rise to the intermediate is caused by the intrinsic tendency of the β-turn2 sequence to be structured otherwise, most likely due to contacts between R453 and E456 that would prevent the native structure. The resulting misfolded domain would be more prone to dimerization and, potentially, aggregation (Mu et al. 2006).

1.2.4. Amyloid Fibrils and Disease

Amyloid fibrils (or simply, amyloids) are formed by the aggregation of misfolded proteins which in turn become capable to induce the amyloidogenic conformation on other natively-folded molecules of the same protein. While some kinds of amyloid fibrils are benign or even biologically functional in a few cases, there is a series of diseases caused by extracellular deposits of amyloids under the generic name of amyloidosis (Pepys 2006). Moreover, amyloids are also closely related to neurodegenerative disorders such as Alzheimer's or prion conditions; however, the actual degree of correlation is still unclear (Cleary et al. 2005; Holmes et al. 2008). Even in some amyloidosis, it is not clear whether toxicity is caused by the amyloid *per se* or it is collateral to fibril formation (for instance, by mislocation or sequestering of key proteins; Olzscha et al. 2011).

There are about 40 known proteins which can form amyloid fibrils and are associated with human disease (Sipe et al. 2016). Usually, a disease is specifically related to one amyloid-forming protein (including certain variants), even though a number of different amyloids can be present in certain processes. The heart and the kidney are the most affected organs by amyloidosis. Amyloid-related disorders (including neurodegenerative conditions) affect over 100 million people every year and related medical expenditure is increasing annually in developed countries (Johns 2013).

Amyloid structures follow a cross-β fibre pattern, named after their X-ray diffraction pattern (Eanes and Glenner 1968). The architecture of short amyloid fragments is based on a motif known as the *dry steric zipper*, where tight interactions are formed due to interdigitating side chains. This conformation has two effects: first, the exclusion of water molecules from the interface between β-sheets (thus, *dry*) and second, the alignment of hydrogen bonds, forming numerous aligned electrostatic interaction sites and hence creating intense dipoles. The latter is responsible of high stability and persistence, characteristic of amyloid fibrils (Eisenberg and Jucker 2012).

The loss of the native structure of proteins is at the basis of amyloid fibril formation. *In vitro* studies have shown that causing a partial unfolding by exposing proteins to harsh conditions, such as high pH or surface denaturation from agitation, can end in fibril formation, even in the case of proteins not involved in amyloid-related conditions (Fändrich et al. 2003). *In vivo*, on the other hand, amyloids have been linked to abnormal overexpression and reducing cell environment (Balch et al. 2008; Ferreira et al. 2015). Biologically, two separate phases of fibril formation have been defined. First, a lag phase, where around 3-4 molecules form an amyloid-starting fragment, that will act as a nucleation point for the second phase, in which the growth of the fibrils will be driven by aggregation of monomers and association of fibrils (Nelson et al. 2005; Lee et al. 2011). Since this behaviour is typical of nucleation processes, the lag phase can be practically abolished by addition of pre-fibrillary aggregates acting as precast nucleation points in a similar fashion to seeding in crystallography.

Folding intermediates and transient oligomers have been postulated to play a role in fibril initiation. In the case of the intrinsically disordered protein (IDP) iAAP, which forms pancreatic islets in diabetes mellitus type-II, the formation of a transient parallel intermolecular β-sheet in the unfolded protein has been postulated as the driving event towards the formation of fibril-like intermediate oligomers. Once a critical number of molecules is reached, the protofibril's structure changes to the final β ensemble, with the transient β-region being part of a loop (Buchanan et al. 2013). Besides, TTR amyloids, responsible of transthyretin amyloidosis, is described to start fibrils by dimerisation, similarly to iAAP. In this case, possible driving events could be destabilisation of *edge strands* (that is, those β-strands at the edge of a β-sheet in the native structure) by a mutation or refolding from mature fibrils (Yang et al. 2003; Olofsson et al. 2001). Another example of an intermediate involved in amyloid formation is β_2-microglobulin, which is the cause of haemodialysis-associated amyloidosis. In this case, the effect of *edge strands* is also postulated to be the main cause for aggregation, generating a folding intermediate highly similar to the native structure with a few distorted β-strands that induces self-assembly (Jahn et al. 2006; Feige et al. 2008).

1.3. Principles of Solution Nuclear Magnetic Resonance Spectroscopy

The tri-dimensional structure of biologic macromolecules and their conformational dynamics are at the core of biologic function and therefore data obtained in the field is of great interest for the scientific community. Structural biologists rely on several techniques to describe macromolecules, but only a few can yield structural information with atomic resolution. These are Macromolecular Crystallography (MX), Nuclear Magnetic Resonance (NMR) Spectroscopy and, more recently, Cryo-Electron Microscopy (Cryo-EM).

NMR spectroscopy is a well-stablished technique that can provide atomic information about molecules, materials and complex samples by interacting with atoms nuclei (which act like small magnets) using radiofrequency. Since it is exceptionally versatile, it is used in a miscellany of industrial and research applications. In the specific case of organic chemistry, biochemistry and structural biology, NMR spectroscopy can provide valuable information on sample purity (of a chemically-synthesised product, for instance), protein structure both in solution and in solid (as for membrane proteins), protein dynamics (for example in function-related open and closed conformations) and molecular recognition (by monitoring protein - ligand interactions).

1.3.1. Quantum Properties of Nuclei Are Exploited in NMR

NMR is a physical phenomenon based on the properties derived from the spin (*s*) of atoms' nuclei. The spin is a property described in quantum mechanics for elementary particles and atomic nuclei. The spin can be described as a form of angular momentum which is not caused by rotation but is intrinsic to the particle itself. From the spin arise the gyromagnetic ratio (γ), which is dependent on the number of dipoles in the nucleus (in turn, dependent on both mass and charge), and the spin quantum number (*I*) which is dependent of atomic mass (total sum of protons and neutrons). These variables compose the nuclear magnetic moment (μ), which is at the fundaments of NMR **(Eq. 1.1)**:

$$\mu = \gamma I \hbar \tag{1.1}$$

Accordingly, for a nucleus to have a magnetic moment it is required that its spin quantum number (*I*) is different to zero. This is the reason why atoms such as ^{12}C and ^{16}O are *NMR-inactive*. On the other hand, while atoms with an $I > ½$ are certainly altered by radiofrequency, they are not used in standard solution NMR. The magnetic quantum number (*m*) represents the aggregate of possible spin states a given nucleus can gain (-I, -I+1, ..., I-1, I). The circumstance of presenting more than two (four or eight) spin states gives rise to a quadrupole moment affecting NMR measurements. Therefore, solution NMR essentially uses atoms presenting an I=½ (nuclear magnetic dipoles) including isotopes of the most common elements in biomolecules: ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P –. Since the natural abundance of some NMR-suitable isotopes is considerably low (^{13}C=0.0111, ^{15}N=0.0037), isotopic enrichment is required during recombinant protein expression for certain

NMR experiments. Additionally, the magnitude of γ is experimentally relevant, since it defines the strength of the interaction with a magnetic field and thus determine experimental sensitivity. Owing to its particular combination of a suitable I value, a natural predominance and a high γ, ^{1}H is the central atom in NMR spectroscopy.

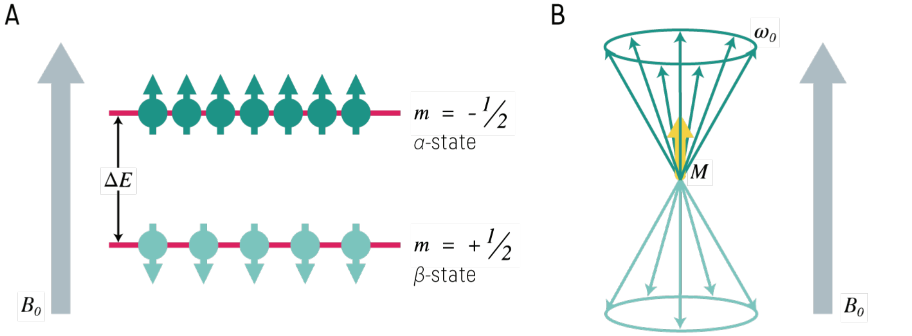


Figure 1.9. Effects of a static magnetic field on nuclear dipoles

A) Zeeman splitting occurs when nuclear dipoles orient in a magnetic field. A small nuclei excess is oriented parallel to the field. **B)** According to the Planck-Einstein relation, energy and frequency are tightly related. Therefore, spins (in green) precess at a frequency dependent on the nucleus and the magnetic field. The small excess of nuclei in α-state generates a net magnetisation, M (in yellow).

1.3.2. Atoms Are Separated in Energy Populations Under a Static Magnetic Field

When a magnetic dipole (I =½; m =$^{\pm}$½) is placed into a magnetic field B_0, the active nuclei are oriented conforming two states (α-state and β-state) in a phenomenon known as Zeeman splitting. States α and β are energetically equal in magnitude but opposed in sign and direction. That means that while α-nuclei are in a low energy state (aligned with the magnetic field), β-nuclei are aligned against the field, and hence in a higher energy state **(Fig. 1.9A)**. For this reason, α- and β-states are also commonly referred to as parallel and anti-parallel states respectively. The energy gap between states is dependent on B_0 intensity and the quantum properties of each type of nucleus as defined by the Zeeman equation **(Eq. 1.2)**.

$$\Delta E = -\gamma \hbar B_0 \qquad (1.2)$$

Nonetheless, oriented dipoles are not static. Since they have an angular momentum, they precess about B_0. The precession defines the resonance frequency of a nucleus, which is known as Larmor Frequency (ω_0, **Eq. 1.3**). As described by the Boltzmann equation (**Eq. 1.4**), there is a slightly higher amount of spins (nuclei) in the parallel state. This population difference results in a net magnetisation (M) in the same direction as B_0, the z'-axis, but with a much smaller module since it is dependent on ΔE (**Fig. 1.9B**). For this reason, NMR is considered a low-sensitivity technique. Practical consequences are the requirement of high concentrations of molecules in solution and an accumulative acquisition of spectra dependent on the type of experiment and the atoms involved.

$$\omega_0 = -\gamma B_0 \qquad (1.3)$$

$$\frac{N_\beta}{N_\alpha} = e^{-\frac{\Delta E}{kT}} \qquad (1.4)$$

1.3.3. The Total Magnetic Field Experienced by a Nucleus Is Particular

For a given nucleus in the context of a macromolecule, its actual Larmor frequency may be slightly different to equal nuclei in the 10^{-6} order of magnitude. This phenomenon is caused by surrounding electrons causing small local magnetic fields that affect the total magnetic field acting on each nucleus. Specific causes are electron density, electronegativity of neighbouring groups, anisotropic magnetic fields and the scalar coupling (or J-coupling), which occurs between nuclear spins connected through covalent bonds. The reference-normalised variation of the

Larmor frequency of a given nucleus respect to that presented by the same kind of nucleus is called the *chemical shift* (δ, **Eq. 1.5**) and it is normally expressed in parts per million (ppm). The particular chemical shifts of nuclei in a sample are key to identify the individual nucleus correlating to each NMR signal.

$$\delta = \frac{\omega_0 - \omega_{0,ref}}{\omega_{0,ref}} \cdot 10^6 \qquad \textbf{(1.5)}$$

1.3.4. The Vector Model Provides an Intelligible Representation of NMR Spectroscopy

Despite its limitations to describe NMR in comparison to quantum mechanics, the vector model provides a more simple and intuitive picture on how an NMR Spectroscopy experiment is operated.

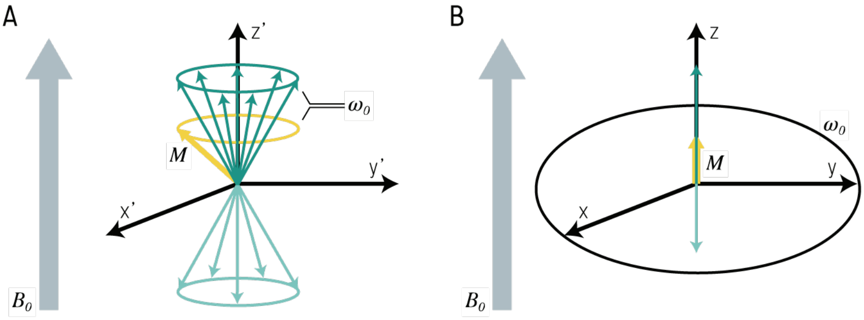


Figure 1.10. Applying the rotating frame

A) Nuclei under a magnetic field represented in the *laboratory frame*. Precession of nuclei (in green) caused by their angular momentum implies the procession of net magnetisation (*M*, in yellow) about B_0 at the Larmor frequency (ω_0). **B)** Nuclei under a magnetic field represented in the *rotating frame*. To simplify the representation of magnetisation in NMR, the axis are permanently rotating on the x,y-plane at ω_0. Thus, magnetisation precession is compensated with frame rotation and both factors can be cancelled out.

As introduced in subsection 1.3.2., *M* determines the z'-axis. Starting from the z'-axis, perpendicular x'- and y'-axis are also stablished, defining the *laboratory frame* (**Figure 1.10A**). When a radiofrequency pulse is applied (a pulse can be defined as a linearly oscillating magnetic field), a fraction of the nuclei excess in the parallel state is forced into the antiparallel state. Using the vector model, this event can be represented as *M* being tilted from the z'-axis. Yet, nuclei conforming *M* are not static, as they precess about the z'-axis at the Larmor frequency as well. To simplify the representation of simultaneous motions, a new reference frame, the *rotating frame*, is applied, where x-axis and y-axis will rotate around z'-axis (now z-axis) at the Larmor frequency (**Figure 1.10B**). This *rotating frame* is the reference of the vector model.

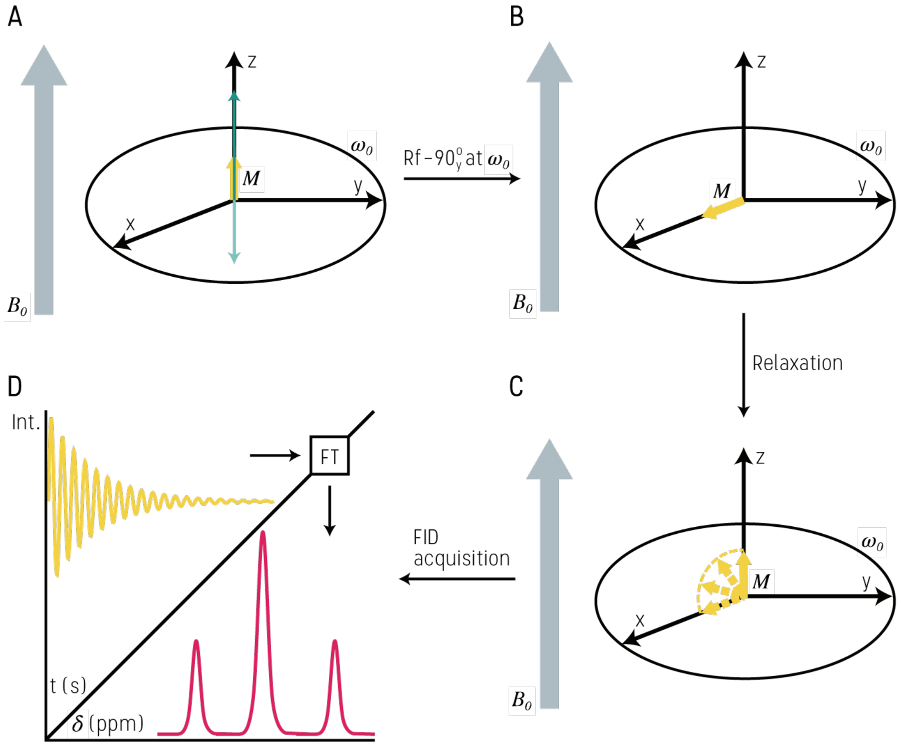


Figure 1.11. A ^{1}H-monodimensional NMR experiment

A) ^{1}H nuclei in equilibrium state under the main magnetic field (B_0). **B)** A -90^o_y radiofrequency pulse at the ^{1}H-Larmor frequency (ω_0) brings *M* to the x-axis. **C)** After the pulse application, due to the effect of B_0, *M* gradually returns to the z-axis in a process known as relaxation. **D)** The relaxation phenomenon emits radiofrequency, which can be detected in the form of Free-induction decay (FID). Applying the Fourier transform to the FID to change the time dimension to frequency results in the intensity-frequency spectrum, where frequency is usually expressed as chemical shift (δ).

1.3.5. Applying a -90^{o}_{y} Pulse Using the Vector Model

A radiofrequency pulse of 90° on the y-axis (-90^{o}_{y}) is a radiofrequency signal applied during a precisely defined period of time and with sufficient power to move *M* to the x,y-plane (in this case, to the x-axis; **Fig. 1.11A-B**). As explained, the actual frequency is dependent on the nuclei, ^{1}H in this case. After the pulse transmission at ω_0, the signal evolves back to the z-axis due to precession about B_0 **(Fig. 1.11C)**. This process is called relaxation and the emitted radiofrequency can be detected, providing a relation of variation of magnetisation intensity over time known as free induction decay (FID). The intensity-time information provided by the FID is not interpretable as it is, it requires to be transformed to the intensity-frequency domain. This is achieved by applying the Fourier transform (FT) to the FID **(Fig. 1.11D)**.

The experiment described corresponds to a ^{1}H-monodimensional, one of the most common routine experiments in NMR spectroscopy.

Chapter 2

Aims of the Thesis

The general scope of this thesis is the study of the atomic and biophysical particularities of protein complexes and folding, particularly those of nuclear proteins of high interest in both basic and applied biomedical research. Two different projects sharing these characteristics and goals are presented hereafter and their main objectives have been grouped accordingly.

Aim 1. To Establish Structural Determinants Defining Protein-Protein Interactions in the Smad2 MH2 Domain

Transforming growth factor ß (TGF-ß) signalling is a key pathway in early differentiation stages and cell homeostasis. Cytokines belonging to the TGF-ß superfamily regulate the epithelial-mesenchymal transition (EMT) that is abnormally promoted in tumourigenic cells, deriving in cell pluripotency, mobility and metastasis. The central elements of TGF-ß signalling are Smad proteins. Receptor-activated Smad proteins (R-Smads) are activated upon cytokine binding and enter the nucleus, where they act as transcription factors. The general consensus considers that the regulatory form consists on a hetero-trimer conformed by two R-Smad molecules and one Smad4. This hetero-trimer coordinates with other proteins participating in chromatin packing and transcriptional machinery and regulators. This activity can be performed thanks to a series of protein-protein interactions with the different Smad partners by means of its linker region, and mostly, its MH2 domain. Understanding how these interactions occur at the atomic label and achieving related biophysical data can provide information that is relevant for researchers in the structural biology, oncology and drug design areas.

The MH2 domain from Smad2, an R-Smad belonging to the TGF-β-activated subfamily, was the system selected for this study. This decision was taken after analysing the information available on protein-protein interactions involving R-Smads. From a large pool of reported Smad2 partners, the transcription factor FoxH1, the transcriptional coactivator NCOA6 and the E3 ubiquitin-ligase/chromatin reader TRIM33 were selected.

The hypothesis defining the framework of this project is that two sets of binding sites exist in the surface of an MH2 domain that regulate protein-protein interactions. A first set would expose generic, low-affinity binding sites the second confer ligand specificity in each Smad subtype. This theory would explain how a domain that is highly conserved across Smad proteins recognise other proteins in a subtype-specific manner. It is also conceivable that yet undefined motifs in Smad-binding sequences exist that would classify Smad partners in binding groups.

The specific aims of this thesis were:

a) The production of the different protein elements under study. This includes the Smad2 MH2 domain and the different nuclear partners selected for the study and requires the evaluation of different molecular biology and organic chemistry strategies and, on the other hand, the definition of sequence boundaries.

b) Detection of the binding events between Smad2 MH2 and its partners by means of a native assay that reports differences in the physico-chemical properties of the elements.

c) Determination of the biophysical characteristics of the interaction between the Smad2 MH2 domain and its different partners, foremost, the constant of dissociation of the complex.

d) Characterisation of the structural properties of the Smad partners in the unbound state.

e) Resolution of the tri-dimensional structure of the Smad2 MH2 – partner complexes.

Aim 2. To Reveal New Aspects of the Folding Mechanisms of the FBP28 WW2 Domain Regarding the Formation of a Folding Intermediate

The particular second WW domain present in Formin binding protein 28 (FBP28) has been object of debate because of its unclear folding kinetics and its tendency to form fibrils that are very similar to those found in amyloidoses and degenerative diseases like Alzheimer's. Specifically, the controversy on the folding kinetics of the FBP28 WW2 is related to the disputed existence of an intermediate state (and therefore, the existence of single or double kinetics) due to differing experimental results. Interestingly, folding intermediates have been related with amyloid aggregation. Folding simulations enriched the debate by providing some support to the existence of an intermediate form, although inconclusively.

This project was carried out as part of a collaboration with the Baker Laboratory of Chemistry and Chemical Biology at Cornell University, New York. The group, led by Harold A. Scheraga, has been working to adapt the physics-based united-residue (UNRES) force field for molecular modelling of the FBP28 WW2 system. Previously, the role of those amino acids involved the strand-crossing hydrophobic contacts that define the WW structure had been addressed. Interestingly, it was revealed that these were not the limiting residues, but that those residues responsible for the greatest fluctuations in folding modelling were placed on the vicinity of β-turn2 and β-strand3.

The work presented in this thesis is based on the hypothesis that a few residues in the β-turn2 and the β-strand3 in FBP28 WW2 hinder the fast, single kinetics folding that is typical for WW domains and cause an intermediate step that has a tendency to form amyloid-like fibrils. The strategy to identify the residues consists on the generation of single-point mutants at certain positions based on previous experiments and sequence alignments.

The goals presented here include both those that are the object of this thesis and those carried out by collaborators, in which case it is indicated. The aims of the work presented in this thesis were:

a) Design of single-point mutant FBP28 WW2 domains in positions around the β-turn2 and the β-strand3 that can dramatically effect folding dynamics (work performed by collaborators).

b) Cloning, expression and purification of the FBP28 WW2 single-mutant domains in a quality and concentration suitable for structural studies.

c) Determination of the tri-dimensional structures of the FPB28 WW2 mutant domains and comparison of their particularities.

d) Study of the effect of the different point mutations on the domain's thermal stability.

e) Application of the experimental data to UNRES-powered folding simulations and data analysis to detect differences in folding trajectories (work performed by collaborators).

Chapter 3

Materials and Methods

3.1. Molecular Biology Methods to Obtain Recombinant Proteins

Structural biology studies demand protein samples at high concentration, purity and homodispersity. As a consequence of these requirements, the first studied proteins were those that could be purified directly from its source in sufficient amounts using classical purification methods (Kendrew 1962).

The development of recombinant DNA methodologies and the progress in competent cell strains from multiple organisms, such as *E. coli, S. Cerevisiae* or *Baculovirus*-transfected insect cells, made heterologous protein overexpression the method of choice in structural biology (Higgins and Hames 1999). In addition, these methodologies allowed for more sophisticated purification strategies, particularly the addition of affinity tags.

Affinity tags facilitate the enrichment of a sample with protein of interest versus host endogenous proteins. Later on, the introduction of sites specifically recognised by proteases allowed to cleave the tag which could be removed using an extra purification step (Nilsson et al. 1997).

Nowadays, numerous cell strains variants are available as well as a vast catalogue of plasmid vectors containing one or several affinity tags. Moreover, technological breakthroughs in nucleic acid chemistry provide the possibility to have a large custom DNA synthesised at low cost, thus saving time and several cloning steps.

3.1.1. Cloning of DNA Sequences into Plasmids

Either a mammalian vector containing cDNA or a bacterial plasmid codifying the target protein was used as a template to amplify a region of interest using the polymerase chain reaction (PCR) method (Mullis 1997). In general, amplification of DNA fragments was verified using agarose-gel electrophoresis, comparing the size of the amplified fragments to a DNA ladder standard. The PCR product was then purified using the PureLink™ PCR purification kit (Invitrogen Inc., USA) and incorporated into the plasmid of choice by either ligation (T4 ligase[1]) or recombination (RecA recombinase[1]). The collection of expression vectors is indicated in **Table 3.1**.

All the vectors selected for the works presented in this thesis present the following characteristics: a bacterial origin of replication, antibiotic resistance, a multi-cloning site (a region containing multiple restriction sites), the *lac* promoter and at least one (often removable) affinity tag **(Fig. 3.1A)**.

3.1.2. Site-Directed Mutagenesis

Different constructs containing single changes in the FBP28 WW domain and the truncated C-terminal versions of the Smad2 MH2 (241/268-457) were obtained by site-directed mutagenesis. This procedure uses a proof-reading

[1] Restriction enzymes and other cloning tools were bought from New England Biolabs, USA.

Table 3.1 Description of vectors used in this thesis

Vector name	N-terminal tag	Tag-cleaving protease
pETM10	His6	None
pETM11	His6	TEV protease
pETM30	His6-GST	TEV protease
pCoofy18	His10	HRV-3C protease
pCoofy35	His6-GST	HRV-3C protease
pOPINS	SUMO	HRV-3C protease
pOPINF	His6	HRV-3C protease
pGAT2	His6-GST	Thrombin

polymerase (Pfu polymerase1) to ensure that the plasmid sequence is correctly amplified and therefore the designed mutation is specifically introduced in the desired position. DpnI[1] restriction was next used to eliminate template DNA as it contains N6-methyladenine. Finally, the PCR product (those plasmids containing the specific mutation) was purified **(Fig. 3.1B)**.

3.1.3. Transformation of *E. coli* Competent Cells

Transformation of *E. coli* cloning strains (DH5*α* and OmniMAX™2) and expression strains (BL21, BL21-Rosetta, B834) was achieved by heat shock followed by growth in SOC medium. Last, cells were plated over LB-agar plates containing the antibiotic for which the plasmid confers resistance and grown overnight at 37°C. Only successfully-transformed bacteria incorporated the plasmid vector and are able to survive antibiotic selection.

3.1.4. Plasmid Preparation

After cloning a DNA fragment into a plasmid or after introducing mutations to a plasmid using Site-directed mutagenesis (methods in subsections 3.1.1 and 3.1.2, respectively), the purified product was transformed (methods in subsection

3.1.3). In each case, a few colonies were selected from the antibiotic-supplemented LB-agar plate and grown in liquid LB medium. After overnight growth, cells were pelleted and plasmids. Cell pellets were resuspended, lysed and purified using the GeneJet™ plasmid miniprep kit (ThermoFisher Scientific, USA) and were sequenced for sequence validation (GATC Biotech, Germany; **Fig. 3.1C**).

3.1.5. Protein Expression and Solubilisation Trials

Sequence-verified clones were transformed into *E. coli* expression strains, usually BL21 and BL21-Rosetta (methods in subsection 3.1.3). Pre-cultures were grown in antibiotic-supplemented LB overnight at 37°C. Expression trials were performed using 24-well plates covered by a permeable membrane allowing high-oxygen consumption to favour and increased bacterial growth. Tested conditions included different media (LB and TB), expression-inducing lactose analogue (IPTG) concentrations (150μM, 300μM, 500μM) and temperature (20°C, 30°C and 37°C). Inoculated conditions were grown at 37°C until the culture reached a OD_{600} of 0.6-0.8AU, equilibrated to the expression temperature for 30min and induced using the designated IPTG concentration. Samples were taken after 4h and after overnight expression. These samples were analysed by SDS-PAGE.

Best expression conditions were defined as the amount of the soluble protein obtained after lysis. To test final soluble yield, cells were lysed by sonication (see 3.1.8) and the soluble and insoluble fractions were separated by centrifugation. Both fractions were analysed by SDS-PAGE. In those cases where results were not satisfactory enough lysis was repeated using alternative lysis buffers. The best overall condition was selected for large-volume expression, since structural biology methods normally require proteins in milligram amounts.

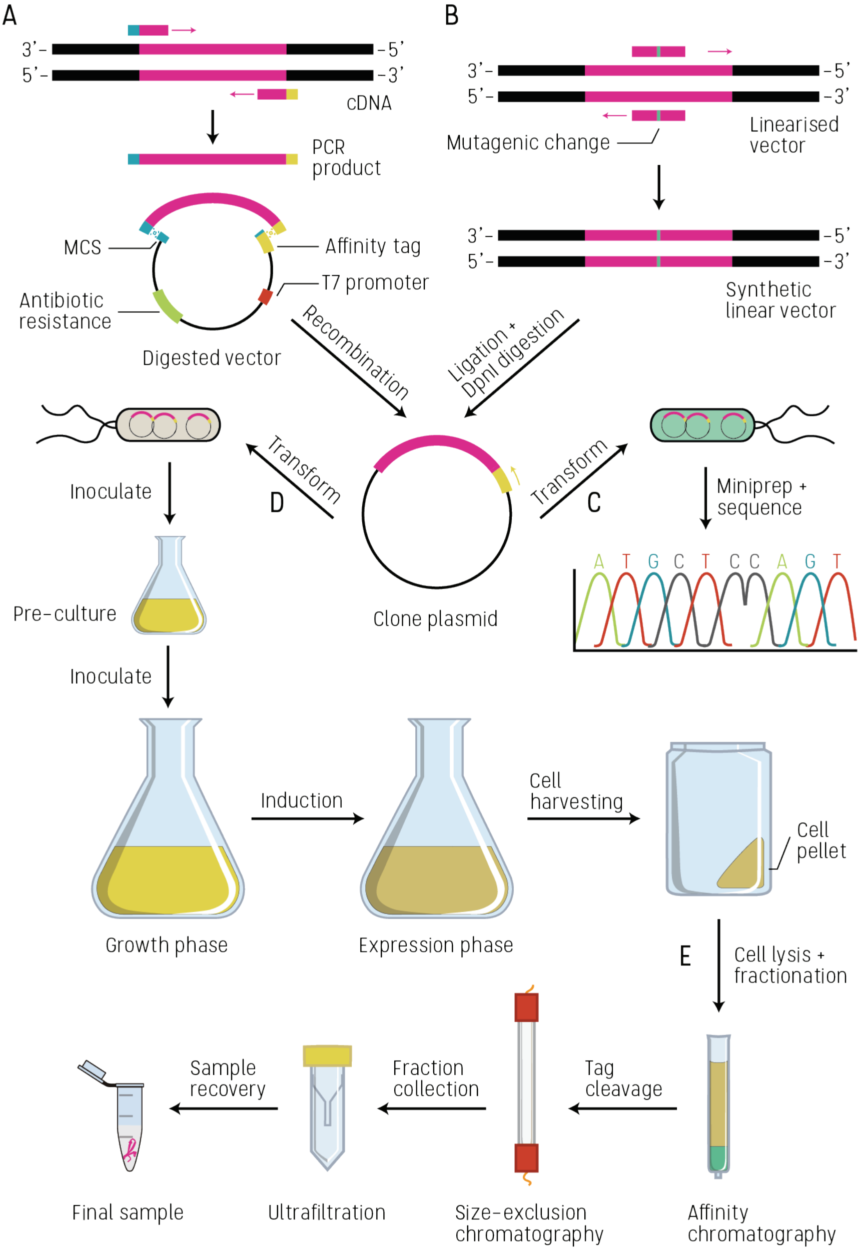


Figure 3.1. General workflow of recombinant protein cloning, expression and purification

Protein sequences were cloned by **A**) template gene amplification and DNA recombination or **B**) site-directed mutagenesis on a cloned protein. **C**) New DNA constructs were transformed into an *E. coli* cloning strain, mini-prepped and sequenced. **D**) Correct clones were transformed into an *E. coli* expression strain and grown to saturation. Expression media was inoculated and protein expression was triggered by addition of an inducer. **E**) Harvested cells were lysed and recombinant protein was purified by affinity chromatography and size-exclusion chromatography. Pure protein was concentrated by ultrafiltration.

3.1.6. Ultra-Frozen Bacterial Cell Stocks

To make repeated expression of a specific protein more convenient, a stock of transformed *E. coli* cells in glycerol was prepared. A fraction of a saturated culture was brought to a final glycerol concentration of 15%(v/v) and transferred into a cryogenic vial for long-term storage at -80°C. When used, the frozen surface was scratched and the peeled off fragments were used to start a pre-culture.

3.1.7. Scaled-Up Expression of Unlabelled Proteins

Pre-cultures were set by inoculation of antibiotic-supplemented LB medium starting from either LB-agar plated colonies or a frozen cell stock. Pre-cultures were grown overnight at 37°C with agitation at a ratio of 15mL per 1L expression volume.

1L of sterilised media was supplemented with antibiotic and inoculated by adding 10mL of the pre-culture in a baffled 2L Erlenmeyer flask. Cultures were grown at 37°C with agitation for approximately 3-4h, until an OD_{600} of 0.6-0.8AU was reached. When needed, cultures were equilibrated to the expression temperature for 30min prior to induction. Expression was induced by adding IPTG at the optimum concentration and carried out overnight. Bacterial cells were harvested by centrifugation and were either purified immediately or stored at -20°C after resuspension in lysis buffer (**Fig 3.1D**).

3.1.8. Scaled-Up Expression of Isotopically-Labelled Proteins

Heteronuclear NMR experiments require samples that have incorporated active nuclei sensitive to the effects of magnetic fields. In the case of a protein sample, this usually means replacing ^{12}C and ^{14}N with ^{13}C and ^{15}N respectively. To achieve this, standard carbon and nitrogen sources must be replaced by isotopically enriched molecules (^{13}C-glucose and ^{15}N-ammonium chloride).

Following methods described in the literature, bacterial cells were separated by gentle centrifugation after reaching an OD_{600} of 0.6-0.8AU. The obtained pellet was washed using PBS solution to eliminate remaining non-labelled substrates and cells were pelleted again. Next, cells were resuspended using M9 minimal medium and the required combination of isotopically-enriched sources ($^{15}NH_4Cl$ and ^{13}C-glucose) was added. Cells were incubated in the fresh medium for 1h before induction and protein expression was carried out overnight.

3.1.9. Cell Lysis

Harvested cells were resuspended in lysis buffer, typically Buffer A or PBS**.** Lysozyme and DNaseI were added to the suspension to facilitate lysis. Lysis was carried out by either liquid homogenisation, using a temperature-monitored EmulsiFlex-C3 homogeniser (Avestin), or sonication, using a Vibra-cell™ VCX 750 (Sonics & Materials). Soluble protein in the lysis extract was separated by centrifugation.

3.1.10. Recombinant Protein Purification

Protein purification using commercially pre-packed tag-affinity columns was carried out in either an HPLC Äkta™ Purifier10 system (GE Healthcare Life Sciences, Sweden) or an NGC Chromatography System (Bio-Rad Laboratories, USA). Otherwise, purification could be carried out using a resin suspension and gravity flow in a reusable plastic column **(Fig. 3.1E)**.

The first purification step in a general purification procedure was affinity chromatography. Two affinity tags were used depending on the protein: a poly-histidine N-terminal tag (His-tag) or a Glutathione S-Transferase N-terminal tag (GST-tag), as well as a third option which combined both. Proteins containing a His-tag were purified by Ni^{2+}-charged IMAC using either 1mL HisTrap™ HP columns (GE Healthcare Life Sciences, Sweden) or Nickel NTA Agarose Resin (Agarose Bead Technologies, Spain). GST-tagged proteins were purified by

Glutathione-affinity chromatography using Glutathione Sepharose™ 4B (GE Healthcare Life Sciences, Sweden). Buffers containing imidazole or reduced glutathione were used, respectively, for protein elution from the column.

In the case of the insoluble FBP28 WW2 mutant, a denaturant Ni^{2+}-charged IMAC was performed to recover protein from the insoluble fraction. Cell pellets were resuspended in a phosphate buffer containing 6M guanidine hydrochloride and refolding was performed *in* column by extensive wash with PBS 1x.

In all cases, the affinity tag was removed overnight by addition of the specific protease – thrombin, TEV or HRV-3C proteases, depending on the expression vector.

After cleavage, a Size Exclusion chromatography (SEC) step was used to separate the protein of interest from remaining *E. coli* endogenous protein impurities, the cleaved affinity tag and – or, in some cases – from aggregated or partially unfolded protein. For this step, HiLoad™ 16/60 Superdex™ 30, 75 and 200 preparative-grade columns were selected according to the size of the protein of interest.

In the case of N-terminally optimised Smad2 MH2 clones, an Ion Exchange chromatography step was added prior to Size Exclusion chromatography as it improved sample monodispersity as seen by Dynamic Light Scattering (DLS). Ion Exchange chromatography exploits the intrinsic physico-chemical properties of the protein, separating proteins according to the charge in a particular pH. Both positively charged (Anion-Exchange) and negatively charged (Cation-Exchange) resins are commercially available. In this specific case, Anion-Exchange chromatography at pH 6.6 was used to successfully separate Smad2 MH2.

The resulting products from each purification step were analysed by SDS-PAGE. Final fractions containing pure protein were mixed together and proteins were concentrated by ultrafiltration using Vivaspin™ 15R centrifugal concentrators (Sartorius Stedim Biotech GmbH, Germany). Protein concentrations were measured using a NanoDrop™ 1000 spectrophotometer (ThermoFisher Scientific, USA). Proteins were stored at 4°C for short-term use or flash-frozen in liquid nitrogen for mid-term storage at -80°C.

3.1.11. Protein N-Terminal Fluorescence Labelling in Solution

N-terminal labelling of the recombinant protein TRIM33 656-715 with Fluorescein isothiocyanate (FITC) was performed in carbonate buffer overnight. FITC excess was neutralised using ammonium chloride. Resulting FITC - protein conjugate was separated using a PD-10 desalting column (GE Healthcare Life Sciences, Sweden) and the fluorescence/protein ratio was measured using a UV-Visible spectrophotometer (Shimadzu Scientific Instruments, Japan).

3.2. Solid-Phase Peptide Synthesis

Biologically-relevant short linear motifs in proteins are often found in very flexible regions. Thus, recombinant expression and purification of such constructs a difficult task. Many proteins contain short linear motifs, which are biologically-relevant as the recognition sites of other proteins. These sites are often present in flexible regions of the protein sequence, facilitating access to interacting partners. Preparing protein samples from motifs that are excessively short or that lack tertiary structure using recombinant techniques is not always straightforward, since they can be easily degraded by proteases. Moreover, the active form of these motifs may contain post-translational modifications, which are difficult and costly to introduce homogeneously into a recombinant protein as would be required in structural studies. Therefore, using well-stablished organic chemistry protocols is a sensible alternative to protein expression for the large-scale synthesis of a natural peptide which contains a linear motif of interest and additionally provides the flexibility to incorporate whichever number of motifs and biologically-relevant modifications.

The Solid-Phase Peptide Synthesis (SPPS) approach was developed in the 1960s by R. B. Merrifield (Merrifield 1965; Kent 1988) and it consists on the chemical synthesis of peptides on a solid support (commonly, low cross-linked polystyrene beads) by the sequential incorporation of amino acids by their α-carboxyl ($-COO^-$) group to the growing chain by successiveness of reaction cycles. A crucial aspect for a successful and ordered coupling of the amino acids is the protection of the α-amine ($-NH_2$) and reactive side chains using chemical groups with differing, pH-dependent lability. After coupling, the amide group of the first amino acid is de-protected and the coupling reaction is repeated with the carbonyl group of next residue. The repetition of this coupling reaction yields the desired peptide chain, which can include phosphorylated amino acids or other modifications. Side-chains protecting groups are maintained thorough the synthesis until the final cleavage step.

The technique became especially popular after the introduction of a SPPS variant based on the base-labile Fmoc protective group (Carpino and Han 1972), since it overcomes several limitations of Boc-based SPPS and circumvents the use of particularly hazardous hydrogen fluoride (HF) for peptide cleavage from the solid support and side chain deprotection.

Over the past several decades, many of the key elements in SPPS have been continuously improved and the approach has become the standard method for synthesising peptides, largely replacing liquid-phase synthesis in most laboratories (Chan and White 2000).

Especially remarkable is the development of automated microwave-assisted SPPS (Vanier 2013), which highly reduces coupling times by applying microwave radiation. For instance, a peptide about 20-residues long can be synthesised in a few hours by microwave-assisted SPPS while obtaining the same peptide following a manual approach would require a minimum of one week of work.

Due to its many advantages, peptide synthesis is becoming a common practice to substitute protein heterologous expression of proteins <40 residues by means of microwave-assisted SPPS.

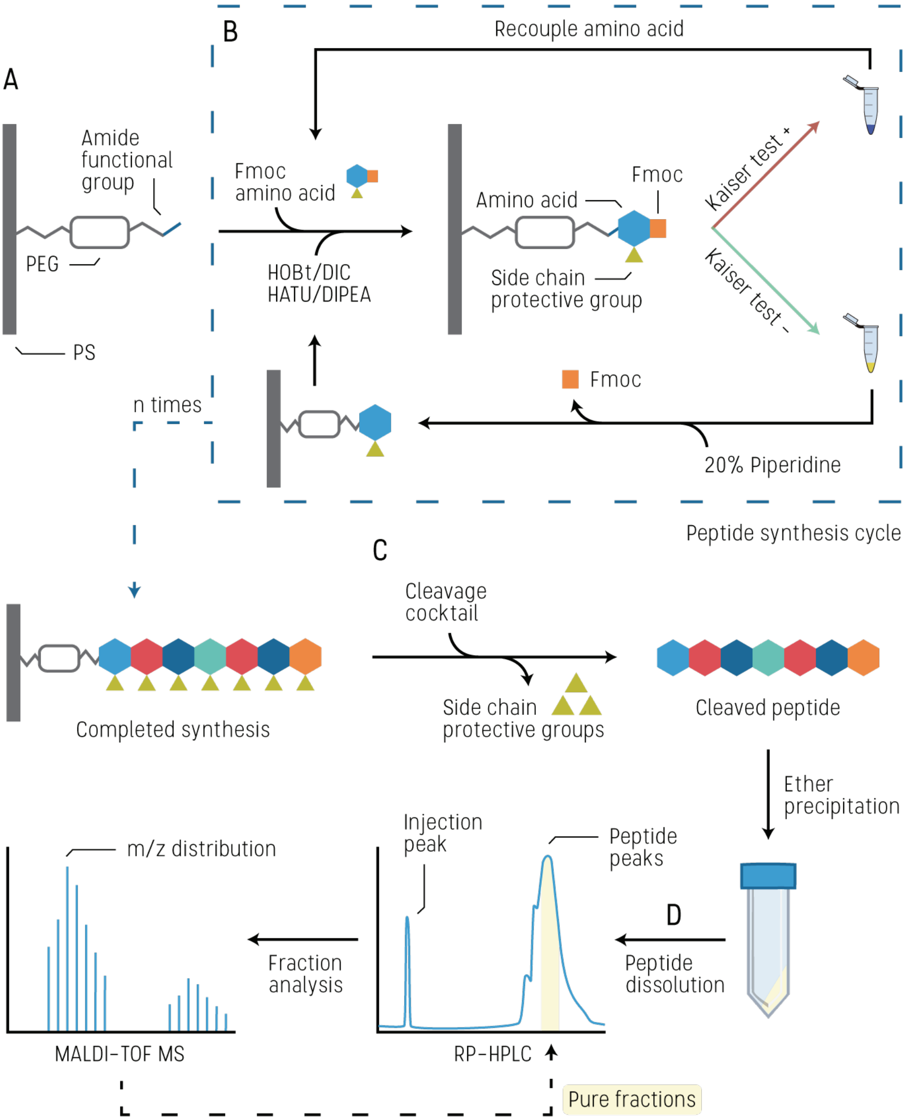


Figure 3.2. General workflow of solid-phase peptide synthesis

A) Solid-phase peptide synthesis is performed on a solid support (or matrix). This matrix contains an active linker where the peptide chain grows. **B)** Fmoc-protected amino acid *i* was coupled by means of carbodiimide or aciluronium/ aminium ester activation. Free amines were detected using the Kaiser test – also the chloranil test in the case of a proline residue. A positive test result indicates an incomplete coupling of amino acid *i*. A negative test indicates a complete reaction. When a negative test result was achieved, amino acid n-terminal deprotection was performed using piperidine. The cycle is restarted with coupling of amino acid *i*=*i*+1 and up to *i*=n. **C)** Once completed, the peptide was cleaved using TFA supplemented with scavengers and precipitated in cold diethyl ether. **D)** The peptide is purified by RP-HPLC and analysed by MS. Fractions containing the pure peptide are mixed and lyophilised.

3.2.1. Amino Acid Couplings in Manual Solid-Phase Peptide Synthesis

Peptides were synthesised in a 0.1-0.5mmol scale using H-Rink Amide ChemMatrix® resin as the solid support. The ChemMatrix® consists on a polyethylene glycol (PEG) support, with the latter presenting a rink amide functional group **(Fig. 3.2A)**. Syntheses were carried out in a polypropylene reaction syringe with polyethylene ring filters coupled to a vacuum line for liquid draining. Prior to the first amino acid coupling, the resin was swelled first in DMF and then in DCM.

Carbodiimide-mediated ester activation using DIC and HOBt was used as the default coupling procedure (Sarantakis et al. 1976). In each reaction, 3.5 equivalents of the Fmoc-protected amino acid and HOBt in DMF were used[2]. The solution was poured into the reaction vessel and DIC was added. The coupling proceeded for 90min under agitation. After the coupling reaction, the resin was thoroughly washed with DMF and DCM

Difficult (incomplete) amino acid couplings were repeated using the stronger acyluronium / aminium ester activation strategy instead (Carpino and El-Faham 1994). 2.5 equivalents of reactives DIPEA and HATU were used, and the coupling of the amino acid was terminated after 20min.

3.2.2. Tests for the Detection of Free Amines

Following each coupling reaction, the resin was tested to determine whether all peptide chains had incorporated an amino acid molecule **(Fig. 3.2B)**. The Kaiser test (Kaiser et al. 1970), based on the reaction of ninhydrin with free primary amides, was used by default. A small number of resin beads were transferred into a boil-proof microcentrifuge tube (Axygen - Corning Life Sciences, USA), 5-10μL of

[2] An HTML5 website was developed to automatically perform the required calculations to prepare the amino acid and reactives mixture for each coupling reaction, including a switch between HOBt/DIC and HATU/DIPEA-mediated ester formation. More information available in the results section.

each Kaiser reactive were added and the mixture was heated at 95°C for 5min. The beads were examined, with development of a blue or brownish colour indicating the presence of free primary amines.

Due to proline presenting a secondary amine in the α-amino group, the acetaldehyde / chloranil test (Vojkovsky 1995) was used in any coupling following a proline residue. As in the Kaiser test, a few beads were separated to a test tube. 10µL of each solution were added and the mixture was stirred. The beads were inspected after 3min. Dark green-coloured beads indicate incomplete reaction.

Nevertheless, both methods can potentially yield false positive and false negative results. Possible causes for failure can be reagent interference (for example, traces of DMF solvent result in a false positive Kaiser test), or the hindered accessibility of amino groups due to secondary structure. Therefore, it is a common good practice to perform periodic test cleavages and mass spectrometry analysis to monitor the quality of the peptide being synthesised.

3.2.3. Fmoc α-Amine Deprotection

The Fmoc group protecting the α-amino group of the latest incorporated amino acid was removed by a two-step piperidine treatment. 0.2M HOBt was added to the deprotection solutions to reduce aspartamide formation, a common side reaction in SPPS. Once finished, the peptide is ready for the next coupling reaction (**Fig. 3.2B**).

3.2.4. Determination of the Resin Load

In cases where an accurate determination of the resin load was required (for instance, after the acquisition of a new resin batch or in the event of an intentional load reduction), the dibenzofulvene-piperidine assay was used after the first amino acid coupling (Dryland and Sheppard 1986). Approximately 2mg of resin were treated with 50% piperidine for 1h with stirring. A fraction of the supernatant was diluted by 1/50. Absorbance of the reaction supernatant and a blank was measured at 301nm in a UV-Visible spectrophotometer (Shimadzu Scientific Instruments, Japan) for load calculation **(Eq. 3.1).**

$$load\ \ (mmol/g)=\frac{(A_s\text{-}A_b)D}{\varepsilon dm} \tag{3.1}$$

3.2.5. Automated Microwave-Assisted Solid-Phase Peptide Synthesis

Depending on availability and sequence, some syntheses were carried out using a Liberty Blue™ automated microwave peptide synthesiser (CEM, USA). This automated synthesiser makes use of microwave radiation to reduce coupling and deprotection times. Microwaves act by rotating ions and dipoles in solution under the alternating electric field. The generated movement increases molecule collisions and thus allows for easier access of reactive molecules to the growing end of the peptide chain.

For the automated syntheses, the carbodiimide ester activation strategy was used. In this case HOBt was substituted by the safer (non-explosive) alternative reactive OxymaPure. Also, piperazine in DMF/NMP/Methanol was used as the Fmoc-removing agent.

3.2.6. N-Terminal Modification of Peptides

Systematically, a fraction of each peptide synthesis product was modified at the N-terminal prior to cleavage from the solid support.

Peptides to be used in electrophoretic mobility shift assays (see section 3.6) were covalently-labelled at the N-terminal with the fluorescent dye 6-Carboxyfluorescein (Fischer et al. 2003). 2.5 equivalents of 6-Carboxyfluorescein in DMF were used together with HOBt and DIC in a 18h-coupling. Since 6-

Carboxifluorescein is not F-moc protected, repeated 20% piperidine treatment steps were required to remove fluorophore repetitions in the peptide chain.

Table 3.2. Reactives and solutions used in solid-phase peptide synthesis

Fmoc amino acids			
Alanine	Fmoc-Ala-OH·H2O	Leucine	Fmoc-Leu-OH
Arginine	Fmoc-Arg(Pbf)-OH	Lysine	Fmoc-Lys(Boc)-OH
Asparagine	Fmoc-Asn(Trt)-OH	Methionine	Fmoc-Met-OH
Aspartic acid	Fmoc-Asp(tBu)-OH	Phenylalanine	Fmoc-Phe-OH
Cysteine	Fmoc-Cys(Trt)-OH	Proline	Fmoc-Pro-OH·H2O
Glutamic acid	Fmoc-Glu(tBu)-OH·H2O	Serine	Fmoc-Ser(tBu)-OH
Glutamine	Fmoc-Gln(Trt)-OH	Threonine	Fmoc-Thr(tBu)-OH
Glycine	Fmoc-Gly-OH	Tryptophan	Fmoc-Trp(Boc)-OH
Histidine	Fmoc-His(Trt)-OH	Tyrosine	Fmoc-Tyr(tBu)-OH
Isoleucine	Fmoc-Ile-OH	Valine	Fmoc-Val-OH
Solutions			
Manual coupling HOBt/DIC	3.5 eq. Fmoc amino acid 3.45 eq. HOBt 3.45 eq. DIC	Kaiser A	50g/L Ninhydrin in EtOH
		Kaiser B	80% Phenol; $20\%_{v/v}$ EtOH;
Manual coupling HATU/DIPEA	2.5 eq. Fmoc amino acid 2.45 eq. HATU 2.45 eq. DIPEA	Kaiser C	0.02M KCN (aqueous) in Pyridine
		Chloranil A	2% Acetaldehyde in DMF
Manual deprotection	30% Piperidine; 0.2M HOBt in DMF	Chloranil B	2% Chloranil in DMF
Automated amino acid	0.2M Amino acid in DMF	Cleavage (default)	95% TFA; 2.5% H_2O; 2.5% TIS
Automated coupling A	0.5M DIC in DMF	Cleavage Reagent B	88% TFA; 5% Phenol; 5% H_2O; 2% TIS
Automated coupling B	1M OxymaPure in DMF	HPLC A	0.1% FA in H_2O
Automated deprotection	10% Piperazine; $90\%_{v/v}$ NMP; $10\%_{v/v}$ EtOH; 0.1M HOBt	HPLC B	95% acetonitrile; 5% H_2O; 0.1% FA

3.2.7. Peptide Release and Side-Chain Deprotection

Peptide cleavage from the resin and side-chain deprotection were carried out at the same time by the addition of highly-concentrated TFA supplemented with nucleophilic scavengers, which protect from carbocations by-products that can otherwise irreversibly react with functional groups in the peptide. The standard cleavage cocktail consisted of TFA/ H_2O/ TIS 95/ 2.5/ 2.5. For those peptides containing tyrosine or tryptophan residues the Reagent B cocktail (TFA/ phenol/ H_2O/ TIS) was used instead (Sole and Barany 1992). The cleavage time was between 1-4h.

The liquid phase was carefully collected and the peptide was precipitated in 5 volumes of ice-cold diethyl ether and left in ice for at least 30min, followed by centrifugation and decantation of the ether phase. The process was repeated 3 times in total to completely remove the protective groups and the peptide pellet was dried in the fume hood. The peptide was dissolved in H_2O/acetonitrile and lyophilised.

3.2.8. Peptide Purification and Characterisation

Prior to purification, a preliminary characterisation of the crude was done by MALDI-TOF mass spectrometry (4700 proteomics analyzer, AB Sciex, Canada) in α-cyano matrix.

The lyophilised peptide was dissolved in H_2O/acetonitrile/0.1% FA, by gradually increasing the acetonitrile concentration. Solubility was facilitated by sonication in a water bath. For very insoluble peptides, traces of DMF were added.

The solubilised peptide was purified by reverse phase HPLC in an Äkta™ Purifier10 system (GE Healthcare Life Sciences, Sweden) or a 1200 Infinity series LC System (Agilent Technologies, USA). Reverse phase separation columns used were either Symmetry® C8 or C18 in both analytical and preparative sizes (Waters Corp., USA). Peptides were eluted using a linear acetonitrile gradient optimised for each peptide and were collected in 0.5mL fractions.

Fractions related to peaks in the chromatogram were analysed routinely by MALDI-TOF MS and by LC-MS in an Acquity UPLC (Waters Corp., USA). According to the mass results, pure fractions were combined. Impure fractions were saved for a potential later purification step. Eventually, both types of fractions were lyophilised and stored at -20°C.

Whenever a peptide was required in an experiment, a fraction of the pure peptide was dissolved in the appropriate buffer (depending on the peptide and the experiment) and quantified by amino acid analysis.

3.3. Biomolecular Solution NMR Spectroscopy

An in-depth understanding of the function of biomolecules requires insights into their structures. Moreover, this information can be used as a starting point for drug design. Solution NMR spectroscopy is the most informative method to obtain this information since it data can be acquired in biologically relevant conditions where proteins are functional.

Structure-solving using NMR data relies on the assignment of individual resonances in the experiment to a nuclear spin present in the molecule. While the structure of small molecules, peptides and small proteins (approximately up to 6kDa) can be solved using only two-dimensional proton correlation experiments, bigger molecules require a more complex workflow including protein isotopic labelling and the assignment of an NMR experiments set to obtain a structure. Since these experiments are acquired in solution, macromolecular dynamics also influence the experimental results. This is reflected in the fact that the result of structure solution by NMR is an ensemble of structures representing the macromolecule instead of a single static structure.

NMR spectroscopy is a method that is also used to study macromolecular interactions. In this case, the main limitation is the final size of the complex. Solution NMR has a practical limit around 50kDa, even though state-of-the-art equipment and highly optimised pulse sequences can overcome this limit and work with macromolecular complexes about 150kDa.

3.3.1. Sample Preparation and Equipment

Protein samples were prepared in a fitting aqueous buffer and supplemented with 10% D_2O. NMR data was recorded at 285-298K in a Avance III 600MHz Spectrometer (Bruker, Germany) equipped with a z-pulse field gradient unit and a triple resonance (1H, 13C, and 15N) probe head. NMRPipe was used for spectra processing.

3.3.1. ^{1}H monodimensional experiment

The experimental procedure to obtain a ^{1}H monodimensional spectrum was described in subsection 1.3.5. A ^{1}H monodimensional spectrum reveals one (or more signals if coupling is allowed to develop) for each proton present in the sample. In the case of proteins, it provides a quick means to check sample quality, including a general assessment of protein folding and determining the presence of contaminants.

3.3.2. Total Correlation Spectroscopy and Nuclear Overhauser Effect Spectroscopy

Total Correlation Spectroscopy (TOCSY) and Nuclear Overhauser Effect Spectroscopy (NOESY) are two-dimensional proton correlation experiments used in the latest stage of resonance assignment for protein structure calculation.

TOCSY (Bax and Davis 1985) correlates all protons within a spin system, that is, all the protons coupled by successive J-coupling. Consequently, correlation peaks

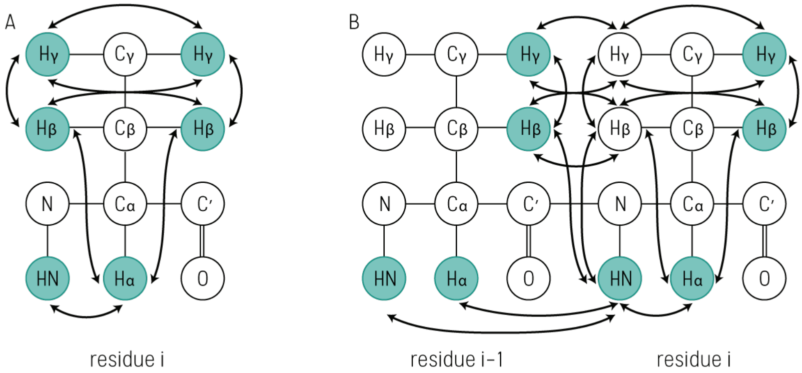


Figure 3.3. Magnetisation transfer scheme of TOCSY and NOESY 2D NMR experiments

A) Schematic representation of magnetisation transfers in a TOCSY experiment. Magnetisation is dispersed over the spin system protons by successive scalar coupling. **B**) Schematic representation of magnetisation transfers in a NOESY experiment. Magnetisation is exchanged between protons by NOE.

between certain protons in TOCSY implies that these form part of the same spin system, that is, the same amino acid **(Fig. 3.3A)**.

On the other hand, NOESY (Jeener et al. 1979) is based on the effect of magnetic dipole-dipole interactions on cross-relaxation between two spins. This phenomenon is called Nuclear Overhauser Effect and occurs between spins close in space, independently of the existence of a covalent bond between them. Hence, proton correlation in a NOESY experiment indicates that those protons are spatially near, specifically in a distance <6Å and the correlation intensity is distance-dependent **(Fig. 3.3B)**.

Usually, the main restrains in structure calculation are the distances derived from NOE and torsion angles derived from J-coupling.

3.3.3. Heteronuclear Single Quantum Spectroscopy

^{1}H,^{15}N and ^{1}H,^{13}C Heteronuclear Single Quantum Coherence Spectroscopy (HSQC) are two-dimensional experiments where a nitrogen atom **(Fig 3.4A)** or a carbon atom **(Fig 3.4B)** correlates with through-bond protons (Bodenhausen and

Ruben 1980). Often, the ^{1}H,^{15}N HSQC is used to obtain a “fingerprint” spectrum of a protein since a single signal is obtained for each backbone amide (additional signals coming from nitrogen-containing side chains appear). During titrations with an interacting partner, changes in the chemical shift for each residue can be traced to establish which residues are affected upon ligand binding.

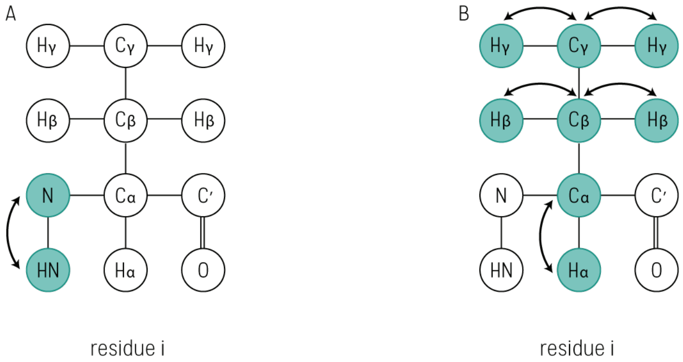


Figure 3.4. Magnetisation transfer scheme of ^{1}H,^{15}N and ^{1}H,^{13}C HSQC experiments

A) Schematic representation of magnetisation transfers in a ^{1}H,^{15}N HSQC experiment. Magnetisation is transferred from hydrogen to attached ^{15}N nuclei via the J-coupling. **B)** Schematic representation of magnetisation transfers in a ^{1}H,^{13}C HSQC experiment. Magnetisation is transferred from hydrogen to attached ^{13}C nuclei via the J-coupling.

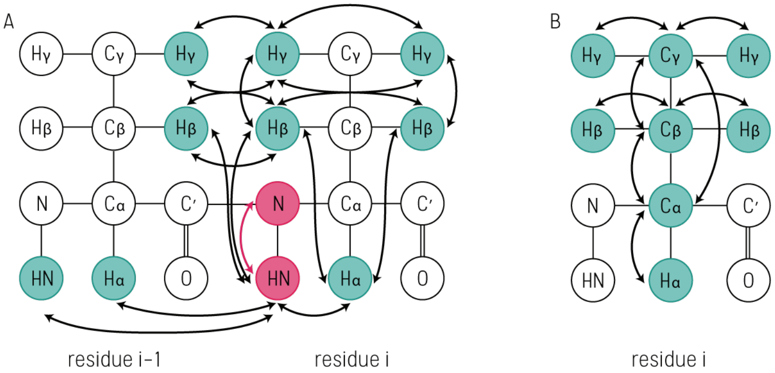


Figure 3.5. Magnetisation transfer scheme of 3D double resonance NMR experiments

A) Schematic representation of magnetisation transfers in a ^{15}N HSQC-NOESY experiment. Magnetisation is exchanged between protons by NOE, transferred to ^{15}N nuclei and back to protons. **B)** Schematic representation of magnetisation transfers in a HCCH-TOCSY experiment. Magnetisation is transferred from protons to ^{13}C via J-coupling and back to protons.

3.3.4. Tri-Dimensional NMR Experiments of proteins labelled with either ^{15}N or ^{13}C

Tri-dimensional experiments are based on magnetisation transfer that involves three nuclei. Experiments using the quantum properties of one labelled atom additionally to ^{1}H can be understood as two-dimensional homonuclear experiments where a third dimension is added to individualise signals and filter information by using the magnetic properties of a different atom.

^{15}N HSQC-NOESY (Marion et al. 1989b) is a ^{15}N-edited ^{1}H NOESY spectra. Magnetisation is exchanged between protons via NOEs, then is transferred to via J-coupling to the ^{15}N-labelled amide nitrogen and back to the proton for detection **(Fig 3.5A)**. This allows to correlate each amide proton to other protons in the same residue and with neighbour residues via NOEs.

^{13}C NOESY-HSQC (Marion et al. 1989a) is an experiment where magnetisation is exchanged between protons using the NOEs and then is transferred to neighbouring ^{13}C nuclei and back to ^{1}H for detection. Transfer is selected either from aliphatic or aromatic side chains depending on the carrier selected. Since present a similar frequency to water, this experiment is normally performed in D_2O in order to maximize the number of identifiable Hα protons.

In a ^{13}C HMQC-NOESY (Fesik and Zuiderweg 1988), magnetisation is first transferred from ^{1}H to ^{13}C and back, and then, it is transferred to other protons close in the structure via NOE. An advantage over the ^{13}C-NOESY-HSQC is that the NOESY dimension is directly detected, therefore increasing the resolution in the NOESY dimension significantly. It can be also very useful to correctly assign the aromatic protons of proteins rich in aromatic residues and to obtain valuable restraints for the structure calculation.

HCCH-TOCSY (Bax et al. 1990) correlates side chain aliphatic protons with ^{13}C via scalar couplings. This results in spectra containing individualised signals for each carbon with the aliphatic and aromatic protons present within the same spin system. With this information, protons can be identified more straightforwardly compared to homonuclear TOCSY and NOESY (**Fig 3.5B**).

3.3.5. Tri-Dimensional NMR Experiments of proteins labelled with both ^{15}N and ^{13}C

^{15}N, ^{13}C-triple-resonance experiments use three different types of nuclei to further select how magnetisation is transferred.

CBCANH and CBCA(CO)NH constitute the standard set of experiments used for protein backbone assignment. In CBCANH (Grzesiek and Bax 1992a), magnetisation is transferred from protons bound to carbons α and β from residues *i* and *i+1* through carbon nuclei to the backbone amide proton of the latter(**Fig 3.6A**). On the other hand, magnetisation in CBCA(CO)NH (Grzesiek and Bax 1992b) from the same aliphatic protons is transferred to the respective carbons and finally to the backbone *i+1* amide proton, providing information on a spin system – *i* – individually (**Fig 3.6B**).

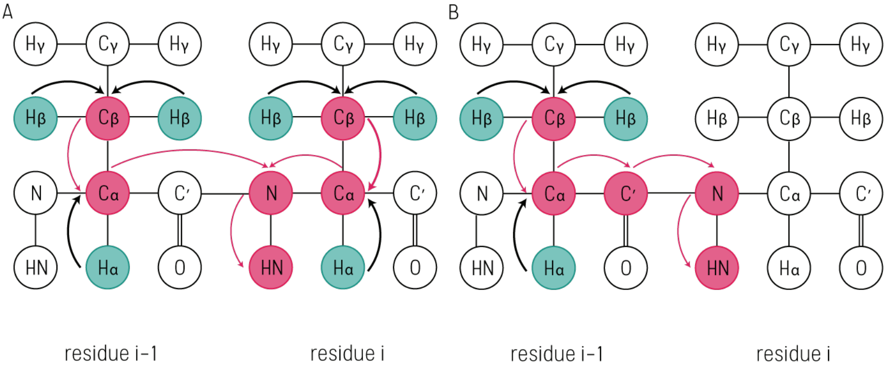


Figure 3.6. Magnetisation transfer scheme of backbone 3D NMR experiments

A) Schematic representation of magnetisation transfers in a CBCANH experiment. **B**) Schematic representation of magnetisation transfers in a CBCA(CO)NH experiment.

The combination of these two experiments allows for the sequential assignment of backbone residues. In turn, backbone assignment provides the assignment for ^{15}N HSQC experiments and valuable information for HCCH-TOCSY, ^{15}N HSQC-NOESY and ^{13}C HMQC-NOESY assignment.

3.3.6. Protein Structure Calculation

Intramolecular proton distances and side-chain torsion angles restraints were obtained from assigned chemical shifts and NOE signals in 2D homonuclear TOCSY and NOESY experiments. CARA software was used for spectra analysis, assignment and integration using the batch integration method of the XEASY package (Bartels et al. 1995). These restrains were used in molecular dynamics simulations using the StructCalc in-house-build server based on CNS 1.1 (Brünger et al. 1998). Structure quality analysis for the calculated models was performed using PROCHECK-NMR (Laskowski et al. 1996). The final structure is represented by an ensemble usually containing between 20 and 50 lowest-energy model structures. Molecular representations were generated using UCSF Chimera (Pettersen et al. 2004).

3.4. Macromolecular Crystallography

Macromolecular Crystallography (MX) is based on the analysis of crystals formed by biomolecules, mainly proteins and complexes, using X-ray radiation to obtain a diffraction pattern from which structural information can be obtained. It accounts for approximately 90% of structures published in the Protein Data Bank (PDB) and together with NMR currently are the two only methods able to provide structural information with atomic resolution (Rhodes 2006).

Obtaining a protein crystal is the first step in MX and often the most time-consuming part of the whole process. Crystallisation occurs in supersaturated conditions where protein molecules tend to form an ordered lattice, as opposed to precipitation. For a protein to crystallise it is usually necessary to be highly pure and homogeneous. Factors influencing the formation of crystals include the pH of the condition, precipitants and additives, temperature and protein flexibility and stability.

Following X-ray diffraction, Fourier transformed to the obtained diffraction pattern to generate an electron density map. This map is used to fit atomic volumes to form protein structure.

3.4.1. Protein Crystallisation Experiments

Protein was purified as described in 3.1.9. Homogeneity of the sample was defined by monodispersion as seen by Dynamic Light Scattering (DLS; Malvern Instruments, UK). The Pre-Crystallisation Test kit (Hampton Research, UK) was used to determine the initial protein concentration. Condition screening was performed in 96-well plates using a Cartesian high-throughput liquid dispenser (Genomic Solutions, UK) at 4°C and 20°C. Favourable conditions were optimised in 24-well plates. Crystals were harvested using nylon loops and stored in liquid nitrogen until data acquisition.

X-Ray crystal diffraction was carried out at the European Synchrotron Radiation Facility (ESRF; Grenoble, France).

3.5. Calorimetric Methods in Structural Biology

Intramolecular and intermolecular biological processes involve a change in the energetic state of a system. Calorimetry measures the heat exchange produced by such physico-chemical changes and provides thermodynamic parameters which are informative of the biological events occurred within the system.

Direct quantification of heat capacity variation ($\Delta Q/\Delta t$) and enthalpy variation ($\Delta H/\Delta t$) of a system can be performed using modern calorimeters **(Fig 3.7)**. The effect of events of interest, namely unfolding or ligand interaction, can be monitored by measuring heat exchange of the sample with a reference cell at a constant volume and pressure (O'Neill 1964).

Calorimetric methods in biophysics and structural biology have gained relevance thanks to quantifiable, high-precision results and diminishing sample quantity requirements thanks to miniaturisation. The two main calorimetric methods are Differential Scanning Calorimetry (DSC) and Isothermal Titration Calorimetry (ITC). The former is used to measure changes in the heat capacity of a sample under a gradual temperature increase and is used to monitor protein folding stability (Freire 1995); the latter, on the other hand, determines the necessary heat power to maintain the sample and reference cells at the same temperature and is used to define protein - ligand binding parameters (Ladbury and Chowdhry 1996).

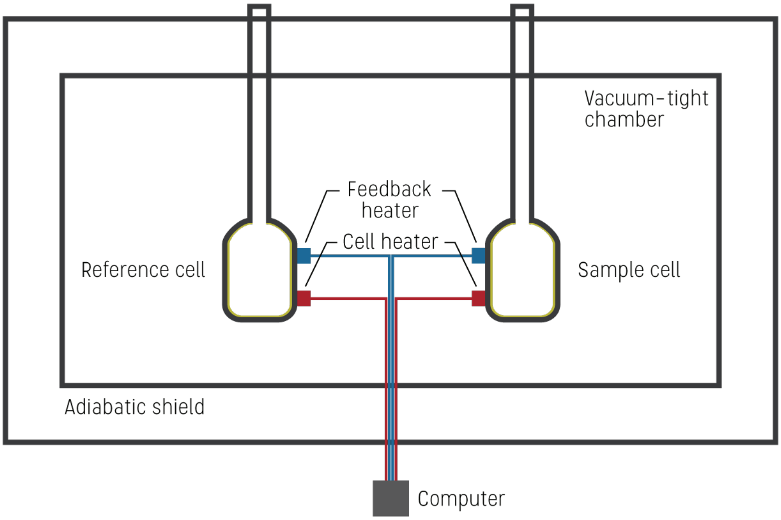


Figure 3.7. A schematic diagram of a calorimeter

Gold-coated reference and sample cells are enclosed by an adiabatic shield in a vacuum-tight chamber. This prevents an alteration of experimental data due to room temperature fluctuations. The system relies on two heating devices: the cell heater, which is in charge of maintaining a stable temperature (or a specified gradient in the case of DSC), and the feedback heater, which reports cell temperature to the computer.

3.5.1. Differential Scanning Calorimetry

DSC measurements were performed using a Nano DSC (TA Instruments, USA). Protein unfolding was measured against a reference cell with the same buffer as in the protein sample. The use of a thermally-stable buffer is required to ensure data quality. A temperature gradient from 20°C to 95°C was applied, increasing the temperature at a 1°C per minute rate. Data was analysed using the NanoAnalyzer software.

3.5.2. Isothermal Titration Calorimetry

ITC measurements were performed using a Nano ITC (TA Instruments, USA). Protein and peptide samples were prepared using the same buffer to avoid noise caused by buffer mismatch. Due to constraints related to peptide solubility at high concentrations, the protein was titrated to the peptide. Experiments were carried out at 20°C following a 35-step titration method. Experimental conditions were adjusted approximately to an estimated C-value of 20. Data was analysed using the NanoAnalyzer software.

3.6. Intrinsic Fluorescence Spectroscopy

In brief, fluorescence consists on the photon radiation by a molecule in an excited energy level when returning to a lower excitation state. Due to energy loss during non-radiative transitions, fluorescence emission is always at a longer wavelength (λ) than the excitation wavelength. This phenomenon is known as Stokes shift. Fluorescence has a broad range of practical applications and has been vastly exploited in physics and chemistry.

Intrinsic fluorescence spectroscopy is a method that relies on the natural fluorescence of (a part of) the molecule under study. In the case of proteins, the fluorescence comes from the aromatic residues, particularly tryptophan, in the sequence. This technique allows to monitor events affecting the fluorescence source, namely protein folding or protein-ligand interactions by a shift on emission wavelength (Hofmann 2010).

3.6.1. Experimental procedure

Buffer exchange to an acetate buffer was performed on samples purified in a Tris buffer to avoid temperature-driven pH changes. The emission λ_{max} shift phenomena caused by the sensitivity of the tryptophan ring to local environment (Vivian and Callis 2001) was used to follow protein melting caused by temperature. Protein samples were analysed using a QuantaMaster™ 300 Fluorescence / Phosphorescence Spectrofluorimeter (Photon Technology International, USA). A temperature gradient between 20°C and 90°C was applied in 2°C increments. At each temperature step, the sample was excited with a laser at λ=295nm and the emission spectra was acquired in the λ=325-370nm range. Data was analysed using GraphPad Prism v.7. The λ value corresponding to the emission maximum intensity at each temperature step was selected and then correlated to temperature to determine the melting temperature (T_m, **Fig. 3.8**).

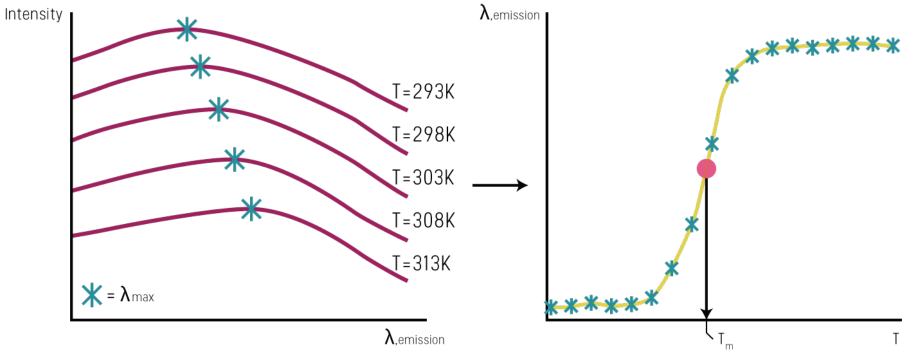


Figure 3.8. Scheme of a protein melting studied using Intrinsic fluorescence spectroscopy

For each temperature, a range of emission wavelengths is scanned. Each λ value corresponding to the maximum intensity is isolated. The collection of λ values is then plotted against temperature. The obtained curve can be modelled as a sigmoidal, from which T_{50} corresponds to T_m.

3.7. Thermofluor Shift Assay

Identification of working conditions that improve the solubility of difficult proteins was carried out by means of a Thermofluor screening protocol (Ericsson et al. 2006). Proteins were screened against the Solubility & Stability Screen (Hampton Research Corp., USA) in a 96-well plate containing the protein, the corresponding screen condition and the SPYRO orange fluorescent dye. This dye is able to interact with the hydrophobic core residues that become solvent-exposed upon unfolding. As a consequence of the interaction, the fluorescence intensity is highly increased. A QuantStudio® 3 real-time PCR thermocycler (Applied Biosystems, USA) was used to perform protein unfolding and fluorescence detection.

3.8. Electrophoretic Mobility Shift Assay

Numerous techniques exist to validate the interaction of proteins with other biomolecules. Many of them, though, have complex specificities that need to be attended for optimal, unambiguous results. Electrophoretic mobility shift assays (EMSA) provide a fast means to probe the ability of a labelled biomolecule (usually a DNA or an RNA, but also a labelled peptide or a protein) to bind to a protein of interest by the appreciation of a band shift essentially due to the change in molecular weight and isoelectric point (Park et al. 1999; Park and Raines 2004). Furthermore, EMSAs can provide insights on the stoichiometry (or stoichiometries) of a certain interaction (Hendrickson 1985). Notwithstanding, affinity results are qualitative and therefore it is required to complement those results with data from quantitative techniques such as ITC.

3.8.1. Experimental procedure

Native 12% polyacrylamide gels were prepared in Tris-Glycine 1x buffer using a 40% acrylamide / bisacrylamide 19:1 stock solution. A dilution series of the protein was mixed with a constant amount of 6-Carboxyfluorescein-labelled peptide. The appropriate peptide concentration was determined empirically to optimise fluorescence signal. Electrophoreses ran for approximately 1h at 100V in Tris-Glycine 1x. Gels were scanned using a Typhoon 9200 imaging system (GE Healthcare Life Sciences, Sweden).

3.9. Dynamic Light Scattering

Monodispersity is a highly desirable characteristic for a protein sample to be used in crystallisation trials. Dynamic light scattering (DLS) is a fast and convenient method to characterise the size of the particles present in a solution or suspension, including proteins and other macromolecules. Most interestingly, it can detect soluble aggregates potentially hindering crystallisation.

DLS is based on the measurement of the scattering of a beam caused by diffusing particles. When a photon hits a particle, the former is scattered (it *bounces* against the particle, changing its angle) in a phenomenon known as Rayleight scattering. In a solution, particles present Brownian motion, therefore scattered light intensity will fluctuate. Measurements of the fluctuation at increasing intervals follows an exponential decay, revealing information about the movement timescale that can be related to particle rotation diameter.

Chapter 4

Results

4.1. Characterisation of Protein–Protein Interfaces on the Smad2 MH2 Domain

In the BMP and TGF-β signalling pathways, membrane anchor SARA presents R-Smads to the cytokine-activated receptor. Receptor-activated R-Smads associate with Smad4, generally in a 2:1 ratio, and then enter the nucleus. The Smad hetero-trimers act as transcription factors, participating in the transcription of many differentiation genes in a context-dependent manner. All the members of the Smad family present a similar domain layout, consisting on two Mad-homology domains: MH1 and MH2. The former is able to bind to DNA whereas the latter acts as an interaction hub for many protein partners. As an exception, inhibitory Smads Smad6 and Smad7 present a degenerated MH1 domain unable to bind to DNA. A flexible linker region connects the MH1 and MH2 domains and regulates protein activity and turnover by means of a serine / threonine phosphorylation code and a WW-domain-binding PY motif. Overall, the sequence of the linker is less conserved than the domains.

The protein-protein interactions occurring in the MH2 domain of the Smads are key determinants of the specific transcriptional function of the active Smad hetero-trimer in the nucleus. Many Smad partners have been identified *in vivo* or using biochemical methods. Nevertheless, few of these have been thoroughly studied, and little information has been reported about the biophysics governing the interaction or the structure of the complex. In the case of TGF-β-regulated R-Smads, it is remarkable that only the complex formation of the Smad2 MH2 domain with the membrane anchor SARA has been successfully characterised at the structural level (Wu et al. 2000). As previously described in subsection 1.1.5, SARA potentiates the effects of TGF-β cytokines by presenting inactive R-Smads to the Ser/Thr-kinase cytokine receptor, resulting in signal transmission by C-terminal phosphorylation of the Ser-X-Ser R-Smad activation sequence.

Regarding the interaction between the Smad2 MH2 domain and protein ligands, the first interaction characterised was, precisely, that with the proteolysis-resistant Smad-binding domain (SBD) in SARA. The structure of the complex defines the SARA SBD interacting sequence to be about 80 residues long and reveals the existence of three secondary structure features: a rigid coil, an amphipathic α-helix and a β-strand. Out of these, the rigid coil and the β-strand contribute mainly to bury hydrophobic residues on the MH2 surface. A comparative analysis of the Smad2 MH2 sequence with the rest of Smad proteins shows the existence of five residues exclusive of Smad2 and Smad3 MH2 domains that play an important role in the interaction with SARA and could explain the specificity of SARA for TGF-β R-Smad. Two of these residues (N381 and W368) form hydrogen bonds that could determine specificity, while the rest (Y366, I341 and F346) contribute to the stabilisation of the complex by means of Van-der-Waals forces.

From the information obtained from the Smad2 MH2 - SARA complex structure and available biochemical data, it was postulated that the highly hydrophobic SARA-interacting surface, consequently known as the hydrophobic patch (or corridor), would be the surface commanding interactions with protein partners other than Smads, although non-exclusively. This surface is overall conserved in the other R-Smad proteins (Shi et al. 1997; Randall et al. 2002).

The purpose of this project is to better understand how the protein-protein interactions between the Smad2 MH2 domain and nuclear partners occur. With this aim, a group of partner proteins was selected based their biological interest and the amount of evidence in literature. An approximation to their interacting sequences was made and the physico-chemical details of each interaction were investigated. The ultimate goal of the framework project would be to obtain tri-dimensional structures of the different complexes.

FoxH1 is a lineage-defining transcription factor that is essential in early embryo development stages. FoxH1 has been reported to recruit active Smad complexes to promoters to initiate the transcription of master mesendoderm differentiation genes like *Gsc* and *T*. The interaction of FoxH1 with the Smad hetero-trimer occurs in the nucleus through the MH2 domain of TGF-β-activated Smad2 and Smad3 (Zhou et al. 1998). Although no structure of the complex has been determined, it has been proposed that two regions in FoxH1, the Smad interaction motif (SIM) and the FoxH1 motif (FM), are responsible for Smad recruitment in an independent manner (Germain et al. 2000; Randall et al. 2004).

The general transcriptional coactivator NCOA6 is a participant on the regulation of cholesterol metabolism that recognises nuclear receptors by means of its two LXXLL motifs and subsequently promotes transcription by recruiting transcription machinery. While the first LXXLL motif recognises most nuclear receptors, the second one has been shown to recognise liver X receptors and TGF-β-activated R-Smads (Lee et al. 2001; Antonson et al. 2008). Nevertheless, for the latter interaction to occur, the presence of a proline-rich region C-terminal to the LXXLL-2 motif is required. This finding supports a previously-described relationship between NCOA6 and TGF-β-dependent cholesterol metabolism regulation as seen in atherosclerosis-related macrophages.

Another R-Smad protein partner is TRIM33, an E3 ubiquitin-protein ligase that, in turn, participates in the regulation of chromatin structure through the reading of histone methylation and acetylation marks. It has also been reported to have a relevant role in TGF-ß regulation, although it is still controversial what this is specifically. Two mechanisms have been proposed: in one, TRIM33 regulates the residence time of Smad complexes on chromatin by ubiquitination of Smad4 (Agricola et al. 2011); on the other, TRIM33 is proposed to be key in the recruitment of R-Smad proteins in unwound chromatin by means of a Smad4-independent R-Smad-TRIM33 complex. The Smad2 MH2 interaction site in TRIM33 was roughly localised within a ~300-residues-long disordered linker between the B-box-2 and PHD domains (He et al. 2006).

The working hypothesis of this thesis is that the biophysical properties defining protein – partner recognition could be explained by the presence of two general features:

- A limited number of surfaces exposing generic, low-affinity binding sites that anchor ligands and govern the space-time compatibility of regulatory elements;

- A larger number of secondary binding sites that confer ligand specificity and total affinity in each Smad subtype.

This hypothesis is based on the increasing variety of Smad MH2 partners being described in the literature and the dramatic effects of certain single-point mutations on the MH2 domain surface. Revealing the structures of the MH2-ligand complexes might show evidence backing this hypothesis and provide data relevant for drug design targeting Smad proteins. In fact, for years, different elements of the TGF-ß pathway have been considered a potential therapeutic target, but a very small number of drugs have reached Phase III trials, reflecting the complexity of the system. The identification of generic binding sites in the Smad proteins and its cofactors and relating the obtained information with MH2 oncogenic mutations may be relevant to further unveil the contextual dependence of the signalling processes triggered by TGF-ß.

The work here presented here consisted on the synthesis of several peptides related to the Smad partners previously introduced: FoxH1, NCOA6 and TRIM33. In order to establish putative binding motifs in the protein targets, the interacting sequences were defined and compared with the cellular and biochemical information available. The peptides were synthesised using solid-phase peptide synthesis (SPPS), incorporating a fluorescent 6-Carboxyfluorescein molecule at the N-terminal for detection. On the other hand, the Smad2 MH2 domain was obtained by recombinant expression in *E. coli*, being necessary the use of several clones to finally obtain high-quality protein in milligram quantities. Since the MH2 domain is located at the C-terminal of Smad2, the different clones mainly differed in N-terminal length, since retaining part of the linker region conferred better solubility, and expression vector, including several options for affinity tag and cleavage sequence.

4.1.1. Cloning and Expression of Smad2 MH2 Domain Constructs Described in Literature

Structural studies related to the R-Smad MH2 domain were conducted at the beginning of the 2000s. Many were focused on the formation of the active trimer, a priming question at that time (Chacko et al. 2001; Wu et al. 2001; Chacko et al. 2004). Current knowledge on the matter describes at the atomic level the contacts defining R-Smad homotrimers and R-Smad-Smad4 heterotrimers, although some details about the formation of Smad complexes *in vivo* remain elusive. As described in subsection 1.1.1, Smad oligomerisation is dependent on two features located in the MH2 domain:

❖ **R-Smad activation sequence.** Situated at the very C-terminal of the R-Smad MH2 domain, the Ser-X-Ser sequence is the target for R-Smad activation. After the oligomerisation and activation of the corresponding TGF-β- or BMP-family membrane receptor caused by the extracellular presence of cytokine molecules, R-Smad molecules presented by SARA are phosphorylated at the two serine residues at the C-terminal by the kinase activity of the receptor. As a consequence, R-Smads become active: that is, able to form oligomers and enter the nucleus.

❖ **Basic pocket.** This second feature, conserved through R-Smads and Smad4, consists on a loop-strand pocket surface formed by the L3 loop and the B8 strand in the MH2 that constitutes the binding site for the activated (phosphorylated) R-Smad C-terminal S-X-S sequence. The intermolecular interactions between the loop-strand pocket and the phosphorylated C-terminal are dominated by the two phosphate groups in the latter and a group of lysine and arginine residues (hence, the *basic* pocket) plus a tyrosine residue in the former that together coordinate a series of electrostatic interactions, hydrogen bonds and well-ordered water molecule bridges conforming and stabilising the interaction.

In the aforementioned publications, as well as in the solution of the Smad2 MH2 - SARA complex, the Smad2 MH2 domain was obtained by recombinant expression in *E. coli.* One of the starting points of this project consisted on the compilation and analysis of these sequences to define the Smad2 MH2 constructs to be used. Works from the Shi laboratory (Wu et al. 2000, 2001) were carried out using a construct starting at M241, leaving approximately 25 residues of the flexible linker at the N-terminal of the MH2 domain. However, in the crystal diffraction structure, the first ordered residue was E270. On the other hand, structures from the Lin group (Chacko et al. 2004) were obtained using a construct with practically all the linker which was later removed using limited chymotrypsin digestion, resulting in a construct likely starting at residue S268 as discerned from digestion simulation results.

Therefore, Smad2 MH2 constructs comprising residues 241-468 and 268-468 were cloned from a full-length Smad2 template[1] **(Fig. 4.1**, in magenta**)** and were cloned into pETM vectors pETM11 (His6) and pETM30 (His6-GST) as well as in modern pCoofy vectors pCoofy18 (His10) and pCoofy35 (His6-GST).

[1] The Smad2 full length DNA Template was a gift from Joan Massagué, Memorial Sloan Kettering Cancer Center, New York, USA.

Three variants of each construct were cloned, varying on the C-terminal sequence: the wild-type SMS sequence (inactive Smad), a DMD mutant mimicking the effect of C-terminal phosphorylation (active, trimer-forming Smad) and a last truncated construct missing the last eleven residues (241/268-457) to abolish any basal trimerisation.

The expression of the constructs was performed overnight and at 20°C using BL21-Rosetta cells, which compensate for tRNAs that are rare in *E. coli*. Expression analysis using SDS-PAGE revealed that expression in pETM11 and pETM30 is inconsistent and generally lower when compared to expression in pCoofy vectors. Therefore, the latter were preferred.

Cells were disrupted either by sonication or pressure homogenisation. Buffer optimisation required extensive additive trials since lysate separation repeatedly resulted in the accumulation of the expressed protein in the insoluble fraction. Finally, only the incorporation of a very high concentration (20%) of dimethyl sulfoxide (DMSO) into the buffer could rescue the protein from the lysis pellet. Nevertheless, a brusque buffer exchange to lower DMSO concentrations during or after purification largely resulted in protein precipitation. An inconvenient of this

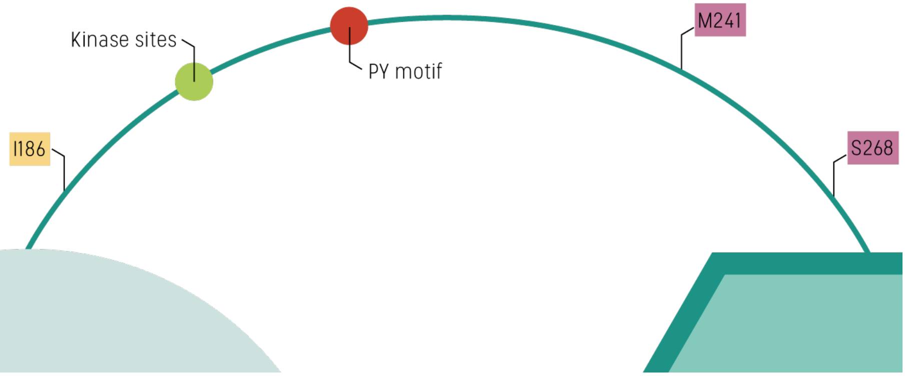


Figure 4.1. Starting residue of Smad2 MH2 constructs as derived from publications

Three constructs of the Smad2 MH2 domain were cloned starting from data available in the literature. First, constructs following the boundaries used in previous of structural studies were cloned from cDNA (in pink) in several vector systems. These constructs present only a fragment of the linker between the MH1 and MH2 domains. Later, an additional construct was obtained that presents practically the totality of the linker (in yellow).

purification methodology is that it prevents tag cleavage. For this reason, only pCoofy18 (His10) constructs were used.

Consequently, a two-step protocol for buffer exchange by dialysis to achieve soluble protein at lower DMSO concentrations was developed by trial and error. After size exclusion chromatography, protein eluted in 20% DMSO was dialysed to 15%, aggregated protein was removed, and then it was further dialysed to a final concentration of 10% DMSO, being the lowest tested concentration preventing total protein precipitation. This process yields its highest amount of protein when using constructs starting at residue 241.

The protein obtained through this method was analysed by size-exclusion chromatography and SDS-PAGE. The results from both methods reveal the presence of soluble aggregates as a result of dialysis process **(Fig. 4.2A-B).** In the case of the Smad2 241-467(DMD) construct (the construct presenting the highest recovery yield) SDS-PAGE shows a major band corresponding to a molecular weight higher than that of the monomeric form. Native gel electrophoresis reveals

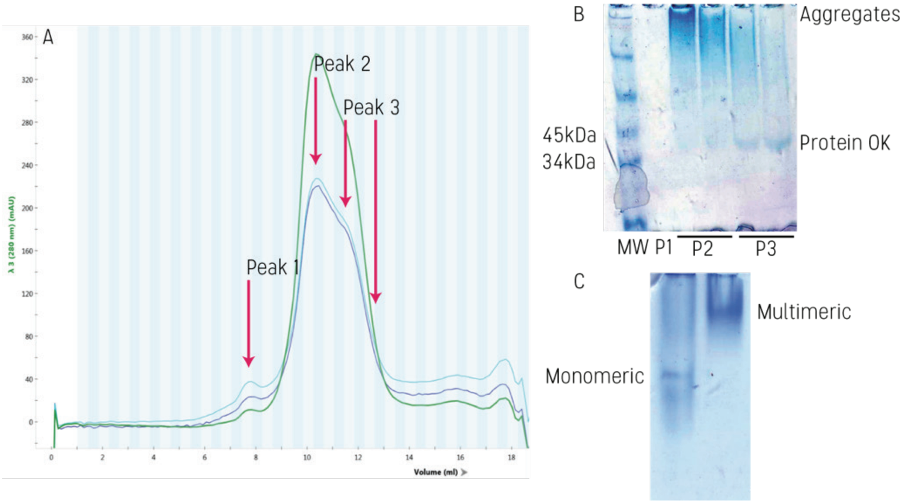


Figure 4.2. Effects of DMSO on Smad2 MH2 aggregation and oligomeric state

A) Analytical gelfiltration of DMSO-purified Smad2 241-467(DMD) protein after dialysis. **B)** SDS-PAGE of the fractions obtained by analytical gelfiltration showing aggregates and soluble protein. **C)** Native gel of protein samples prior (left) and after (right) dialysis indicates a different oligomeric state in each.

that oligomerisation of the phosphomimetic construct is favoured at lower DMSO concentrations, while the protein prior to dialysis presents a lower molecular weight, most likely a monomer (**Fig. 4.2C**). Nevertheless, completely eliminating the presence of soluble aggregates is not possible, as a sort of equilibrium between soluble protein and aggregation occurs when high concentrations required for structural experiments are reached. These observations were further supported by low-quality SAXS data.

Alternative constructs were designed to overcome these handicaps. One of these new constructs made use of the SUMO-tag-containing pOPINS vector. The expression of SUMO fusion proteins has been described to present two particular benefits: first, additionally to favouring overall protein solubility, a few cases have been reported where protein folding is facilitated in particularly unstable proteins; second, the ULP1-derived SUMO protease recognises the C-terminal Gly-Gly motif in SUMO, which allows to leave no N-terminal tag residues in the protein of interest (Peroutka Iii et al. 2011). The new SUMO-Smad2 MH2 241-468 and 268-468 clones present good levels of protein expression and affinity purification can be performed in typical conditions. Unfortunately, SUMO cleavage, as seen by SDS-PAGE, had a greater preference for a Gly-Gly sequence present in a loop inside the MH2 domain, fragmenting the latter instead. Consequently, this clone was discarded.

Besides, we were provided with a highly soluble clone previously used in biophysical studies (Kondé et al. 2010)[2]. This Smad2 MH2 clone, named Smad2 186-467(EEME) hereafter, is cloned into vector pETM10 (non-cleavable His-tag) presents at the N-terminal an extension consisting on practically the entire linker region and phosphomimetic serine to glutamic acid C-terminal substitutions (**Fig. 4.1**, in yellow). This clone presents high expression levels and solubility during lysis and purification.

[2] The Smad2 186-467(EEME) clone was a gift from Sophie Zinn-Justin et Jean-Baptiste Charbonnier, Centre CEA de Saclay, Gif-sur-Yvette, France.

The exposed data and empirical observations illustrate that maintaining a large fragment of the linker region at the N-terminal of the Smad2 construct seems to favour solubility and stability, both being requirements for biophysical and structural studies.

4.1.2. General Approach to the Chemical Synthesis of Smad-Interacting Protein Fragments

To study the characteristics of the interactions between the Smad2 MH2 domain and nuclear proteins, a large group of Smad partners described in the literature was analysed. Due to their biological and structural interest as well as practical considerations, three proteins were selected out of this group. As explained previously, these are the transcription factor FoxH1, the co-activator NCOA6 and the chromatin reader ubiquitin ligase TRIM33. Interacting motifs in FoxH1 and NCOA6 have been described biochemically with a sufficient level of sequence detail. TRIM33, on the other hand, is described to interact with the MH2 domain with an unspecific motif contained in a fragment comprising ~300 residues. Some of these sequences were cloned and expressed in early stages of the project. Nevertheless, the approach was unsuccessful mostly due to problems that arise in to too short (or too large), largely unstructured proteins. Therefore, the approach was changed to chemically synthesise, by means of Solid-phase peptide synthesis (SPPS), the Smad-interacting motifs in FoxH1 and NCOA6.

Peptide syntheses followed, in general, the proceedings described in section 3.2. As a summary, peptides are synthesised on a solid support by sequential coupling of Fmoc-protected amino acids, generally by means of carbodiimide-mediated ester activation and piperidine-mediated N-terminal deprotection cycles. For all peptides in this thesis, Chemmatrix® Rink-amide resin was used as the solid support. Coupling and Fmoc de-protection reactions are monitored using colorimetric tests for the detection of primary and secondary amines. Once completed, peptides are released from the solid support by acidic cleavage, precipitated and purified using RP-HPLC. Chromatography fractions are analysed by UPLC-MS and MALDI-TOF MS. Fractions containing highly pure peptide are freeze-dried and stored at -20ºC. Unless stated otherwise, two variants of each peptide were synthesised: one presenting a free N-terminal amine and the other as a modified version including a 6-Carboxifluorescein molecule at the N-terminal for fluorescence detection.

Besides, to facilitate the synthesis workflow and the customisation of each coupling reaction, an interactive .html file was developed using bootstrap[3]. The website allows to introduce the particular variables of each synthesis and modify those affecting a particular coupling reaction thanks to the incorporated database that contains all chemical reagents. The output of the program provides the exact amounts of each reactive required in a coupling step. A screenshot of the landing page of the website is included in the Appendix section (**Fig. S1**).

4.1.3. Syntheses of Peptides Containing the FoxH1 SIM, FoxH1 FM and NCOA6 LXXLL-2 Motifs

A priori, the most straightforward approach to study the interactions between the Smad2 MH2 domain and nuclear partners is to start with short, defined sequences reported to drive the interaction. According to published data, this prerequisite was best accomplished by FoxH1 and NCOA6 motifs described to directly participate in the interaction.

In the case of FoxH1, two different regions have been described to be responsible for Smad recruitment in an independent manner: the Smad interaction motif (SIM) and the FoxH1 motif (FM). The boundaries of both motifs were defined by ChIP and in the case of the SIM motif, its interaction with the Smad2 MH2 has been studied by means of biophysical studies. Nevertheless, neither the tri-dimensional structure of the complexes nor the individual motifs are available.

[3] Bootstrap (getbootstrap.com) is an open source toolkit for developing with HTML, CSS, and JavaScript. It includes an extensive library of prebuilt components that allows to quickly prototype websites and mobile applications.

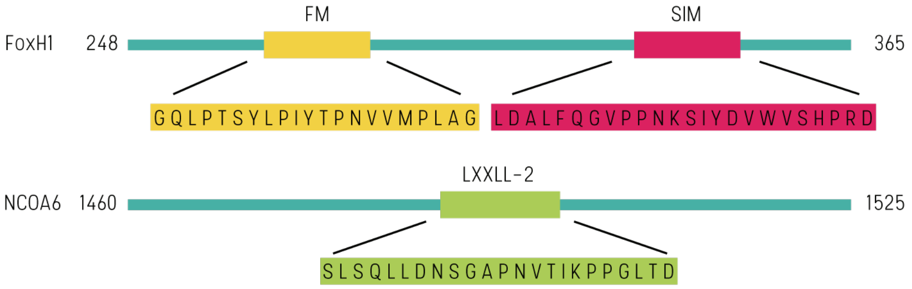


Figure 4.3. Peptide sequences from reported Smad-binding fragments in FoxH1 and NCOA6

Fragments in FoxH1 and NCOA6 reported to bind to the Smad2 MH2 domain are coloured. The sequences of the peptides containing these motifs are shown correspondingly.

The boundaries of the SIM and FM motifs used in this study **(Fig. 4.3)** are the ones indicated in literature (Germain et al. 2000; Randall et al. 2004).

The synthesis of the FoxH1 SIM peptide was performed using microwave-assisted SPPS while in the case of the FoxH1 FM motif, it was carried out manually. Coupling of the N-terminal 6-Carboxifluorescein was carried out manually using a fraction of the synthesis product before cleavage. Purification after cleavage was carried out by RP-HPLC using an acetonitrile (AcN) gradient. Mass spectrometry pointed which fractions were the purest. After freeze-drying the samples, the total peptide weight was obtained to calculate the synthesis yield **(Table 4.1).**

On the other hand, previous studies on the association of NCOA6 to TGF-β-activated R-Smads limit the interacting sequence to the second NR-box (LXXLL-2) in the protein. Nevertheless, yeast-two-hybrid experiments on the same studies showed that the presence of the LXXLL-2 motif alone could not explain the interaction, but that a fragment C-terminal to the motif was required. Therefore, the peptide synthesised in this study was designed to contain both the LXXLL-2 and the necessary adjacent fragment **(Fig 4.3)**.

The synthesis of peptide NCOA6 LXXLL-2 was performed manually. 6-Carboxifluorescein was coupled to a fraction of the synthesis product. Purification after cleavage was carried out using RP-HPLC using an acetonitrile gradient. Mass

Table 4.1. Results of the SPPS of FoxH1 and NCOA6 Smad-interacting fragments in the literature

Peptide name	Predicted monoisotopic mass	Elution in AcN gradient	Experimental monoisotopic mass*	Final synthesis yield
FoxH1 FM	2230.6	33-35%	2231.8	92%
FoxH1 SIM	2753.1	27-31%	2752.8	82%
NCOA6 LXXLL-2	2336.6	21-25%	2337.5	89%

*Major species

spectrometry pointed which fractions were the purest. After freeze-drying the samples, the total peptide weight was obtained to calculate the synthesis yield (**Table 4.1**).

4.1.4. General Considerations on the Observation of Smad2 MH2 Domain – Peptide Interactions by Electrophoretic Mobility Shift Assay

The interaction between two biomolecules (in the systems presented in this thesis, a protein and a peptide) is dependent on the affinity constant of the system, commonly expressed as the dissociation constant (K_d), which represents the concentration where 50% of the potential complexes are formed. This translates in the fact that in a protein - peptide mixture where the protein concentration is below K_d, the greater protein fraction is not forming a complex. Alternatively, at protein concentrations higher than the K_d, the greater fraction of protein molecules is forming a complex with the peptide.

Electrophoretic mobility shift assay (EMSA) was the method selected to identify the interaction between purified recombinant Smad2 MH2 domains and 6-Carboxifluorescein-labelled synthetic peptides and tentatively determine a K_d range for the interaction. In an EMSA experiment, a native acrylamide gel is used to separate pre-incubated mixtures of protein and peptide ligand conforming a gradient of increasing protein concentration. The fluorescence emitted by 6-Carboxifluorecein is used to report gel bands containing the peptide, either free or in complex with the protein. Consequently, in a typical EMSA gel reporting protein - peptide interaction, fluorescence imaging reveals free peptide bands at protein concentrations below K_d, while protein - peptide complexes (which migrate differently due to the differences in size and charge) are reported at protein concentrations above K_d there. The band shift indicates the interaction event and those concentrations where the transition from free to bound can be seen determine a K_d range.

EMSA were designed assuming that the interactions addressed in this thesis would present an affinity in the same order of magnitude as similar MH2 domain - peptide interactions described previously (Bourgeois et al. 2013). Unless stated otherwise, the protein construct Smad2 186-467(EEME) described in subsection 4.1.1 was employed. The concentration range generally used in EMSA experiments hereafter goes from 0μM to 300μM, with an expected K_d in the 2μM - 20μM range. On the other hand, 6-Carboxifluorescein peptides were used at a final concentration of 0,75μM (kept constant across all samples) as empirically optimised for fluorescence detection.

4.1.5. EMSA Can Specifically Reproduce the Interactions of FoxH1 SIM and NCOA6 LXXLL-2 with the Smad2 MH2 Domain

As indicated, EMSA were carried out to assess the binding capabilities of the FoxH1 SIM, FoxH1 FM and NCOA6 LXXLL-2 peptides as described in subsection 4.1.4. The interactions of the Smad2 MH2 domain with the FoxH1 SIM and NCOA6 LXXLL-2 peptides can be reproduced in EMSA (**Fig. 4.4B-C**). Nevertheless, the interaction of the FoxH1 FM motifs reported in the literature is not reproduced, neither in the same buffer conditions as the other two interactions nor changing the buffer pH to higher or lower values (**Fig. 4.4A**).

The same experiments were conducted as well using the Smad4 MH2 domain as a control for the selectivity/specificity of the interactions between different MH2 domains and protein partners and, additionally, to clarify which component of the Smad2-Smad4 heterotrimer is the predominant binding site for these

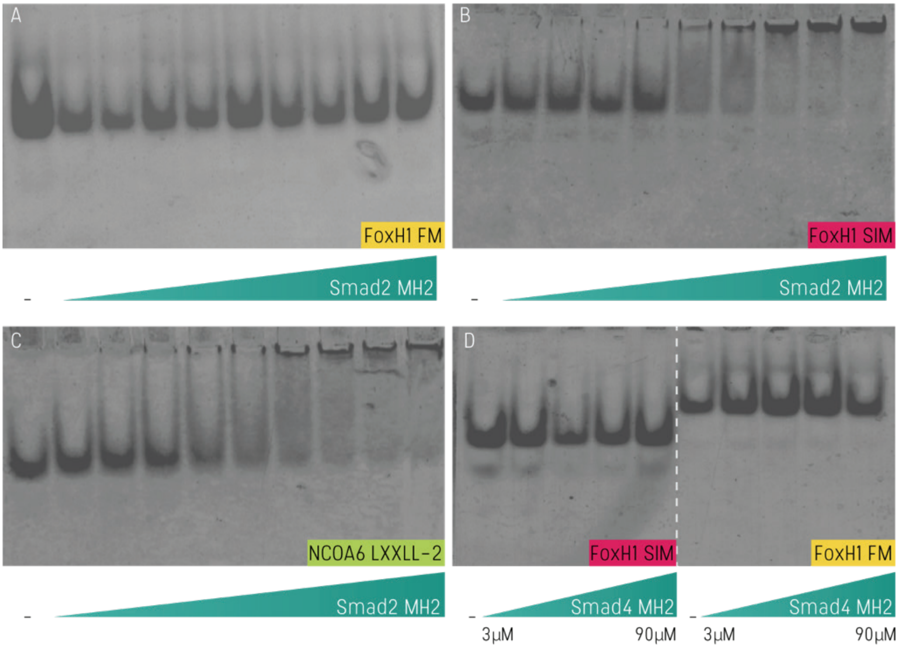


Figure 4.4. Interaction detection of FoxH1 and NCOA6 peptides with Smad2 MH2 as seen by EMSA

EMSA experiments using the Smad2 186-467(EEME) protein and peptides FoxH1 FM (A), FoxH1 SIM (B) and NCOA6 LXXLL-2 (C). Binding specificity of positive peptides for the Smad2 was tested in EMSA using the MH2 domain of Smad4 (D).

protein partners. The Smad4 MH1 domain was prepared by recombinant expression and purified as described (Wu et al. 2002). The EMSA carried out using Smad4 permit to investigate two characteristics. The results show that none of the peptides can interact with the Smad4 MH2 (**Fig. 4.4D**), corroborating the specificity of the FoxH1 SIM and NCOA6 peptides for the Smad2 MH2 domain. They also indicate that FoxH1 FM motifs neither bind to Smad2 nor to Smad4 MH2 domains under the experimental conditions used in our study.

4.1.6. Sequence Comparison of the MH2-Interacting Peptides and TRIM33 Middle Shows a Potential Leucine-Proline Pattern

A detailed comparison of the peptides standing out in the EMSA experiments revealed some sequence similarities that could potentially explain why all these proteins interact with the MH2 domain of Smad2. In the literature describing the FoxH1 SIM - Smad2 MH2 interaction (Germain et al. 2000; Randall et al. 2002), it is speculated that four residues in the SIM motif, PPNK, are responsible for most contacts with Smad2. Interestingly, while in the analogous publication about NCOA6 the focus was put on the LXXLL-2 motif itself (Antonson et al. 2008), the adjacent C-terminal sequence (equally necessary for Smad binding) contains a double proline (PP) segment, drawing similarities with FoxH1 SIM. Not only this, but at the immediate N-terminal of FoxH1 SIM there is a leucine rich LLCDL that resembles the LXXLL motif.

The same premises were applied to TRIM33 in order to examine the potential presence of similar patterns within its Smad2-interacting middle region (He et al. 2006). Promisingly, a sequence similar to those present in FoxH1 and NCOA6 was found in the form of an overlapped tandem consisting an LXXLL-like motif flanked by two proline pairs at a distance similar to interacting sequences in FoxH1 and NCOA6. **(Fig. 4.5)**.

From this analysis, the next working hypothesis was that a specific type of MH2 domain binding motif would require two elements: a leucine-rich sequence (likely L-L/X-X-L/X-L) and, at least, two consecutive proline residues at a distance between 7 and 10 residues. The presence of this pattern, called hereafter LP pattern, would define the capacity of a peptide to bind to an R-Smad MH2 domain.

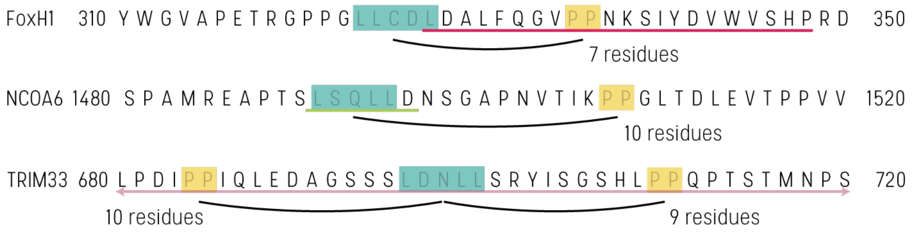

FoxH1 310 Y W G V A P E T R G P P G L L C D L D A L F Q G V P P N K S I Y D V W V S H P R D 350
7 residues
NCOA6 1480 S P A M R E A P T S L S Q L L D N S G A P N V T I K P P G L T D L E V T P P V V 1520
10 residues
TRIM33 680 L P D I P P I Q L E D A G S S S L D N L L S R Y I S G S H L P P Q P T S T M N P S 720
10 residues
9 residues

Figure 4.5. Sequences of Smad partners displaying a feature pattern

Sequence analysis reveals that the Smad-binding fragments in the transcriptional regulators under present a common leucine-rich fragment in the form of L-X/L-X-X/L-L (in green) followed by two prolines (PP) less than 10 residues apart (in yellow). This pattern suggests a potential new motif in the sequences.

4.1.7. Syntheses of FoxH1 LP, TRIM33 Nter-LP and TRIM33 Cter-LP Peptides

Despite the ability of the previously synthesised FoxH1 SIM peptide to bind to the Smad2 MH2 domain, potential differences in binding mode or binding affinity caused by the presence of the full LP pattern present in FoxH1 were evaluated. With this aim, a new FoxH1 peptide was synthesised. This peptide, named FoxH1 LP, contains the full SIM sequence plus the proline pair at its N-terminal (**Fig. 4.6**).

In the case of TRIM33, two peptides were synthesised containing the two parts conforming an LP pattern in the discovered LP tandem sequence (**Fig. 4.6**). This was done in order to address the binding capabilities of both parts of the sequence individually and also to avoid potential difficulties that arise during the synthesis of lengthy sequences by SPPS.

The synthesis of the FoxH1 LP peptide was performed manually starting from previous FoxH1 SIM uncleaved synthesis. The syntheses of TRIM33 Nter-LP and TRIM33 Cter-LP were carried out manually as well. A fraction of the synthesis product was N-terminally labelled with 6-Carboxifluorescein. Purification after cleavage was carried out by RP-HPLC using an acetonitrile gradient. After freeze-

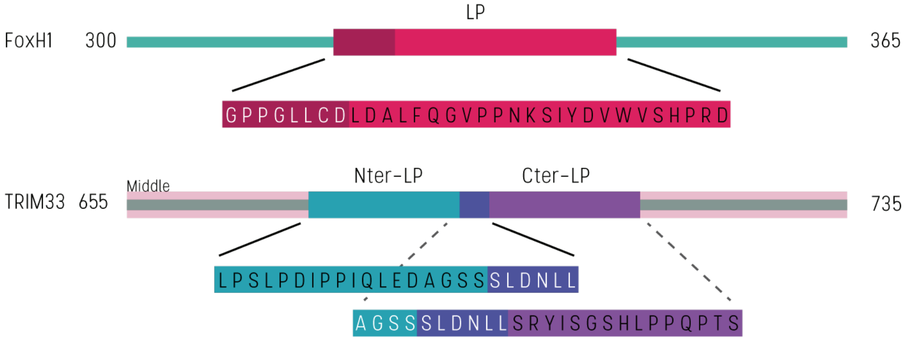


Figure 4.6. Sequences of peptides following the suggested LP pattern

Sequences containing the full LP pattern are highlighted. In the case of TRIM33, two peptides containing one side of the LP tandem each – and sharing the leucine-rich part – are synthesised with the aim to assess their binding capabilities independently.

Table 4.2. Solid-phase peptide synthesis masses and yields

Peptide name	Predicted monoisotopic mass	Elution in AcN gradient	Experimental monoisotopic mass*	Final synthesis yield
TRIM33 Nter-LP	2503.8	32-35%	2504.7	80%
TRIM33 Cter-LP	2583.8	36-38%	2582.9	69%
FoxH1 LP	3238.6	37-40%	3239.5	72%

*Major species

drying the samples, the total peptide weight was obtained to calculate the synthesis yield (**Table 4.2**).

4.1.8. The Tandem LP Sequence in TRIM33 Does Not Interact with the Smad2/4 MH2 Domains in EMSA Experiments

Once more, EMSA were employed to assess the binding capabilities of the new peptides. These were carried out using the different elements at the concentrations exposed in subsection 4.1.4. FoxH1 LP, as expected from the FoxH1 SIM assays, interacts with the Smad2 MH2 domain. Unfortunately, the particularities of gel shift assays make it difficult to discern small-magnitude changes in affinity. As a consequence, a discrete determination of the K_d value is required to determine a potential better affinity thanks to the inclusion of the full leucine-rich part.

In the case of the two TRIM33 tandem LP, no interaction was detected by EMSA for any of the two peptides covering the sequence (**Fig. 4.7A-B**). Again, as in the case of FoxH1 FM, two additional buffer conditions were tested with similar results. The experiments were also repeated using the Smad4 MH2, which does not result in any band shift that could correspond to a misattribution of the TRIM33 binding to Smad2 (**Fig. 4.7D**).

Provided that the peptides were designed by dividing the tandem LP pattern in TRIM33, the possibility that the whole sequence was required for the interaction with the MH2 domain was studied. Due to its size, the whole TRIM33

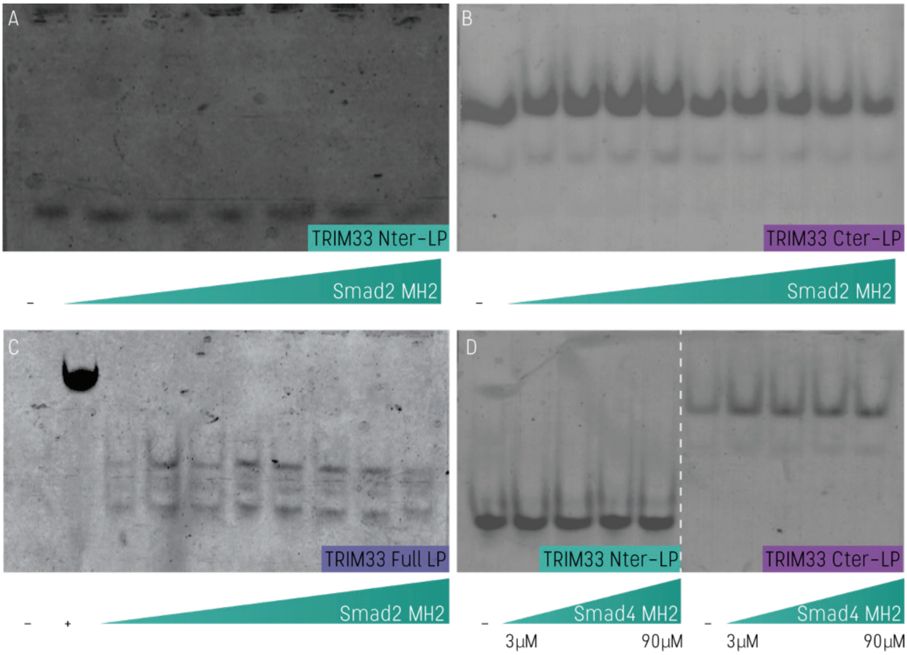


Figure 4.7. Interaction assays of TRIM33 LP peptides with Smad2 MH2

Binding of the LP tandem in TRIM33 with Smad2 MH2 cannot be detected, nor in its spliced N-terminal (**A**) and C-terminal (**B**) forms nor using the full recombinant fragment (**C**). EMSA with Smad4 MH2 discard a misattribution of the binding the other Smad member (**D**).

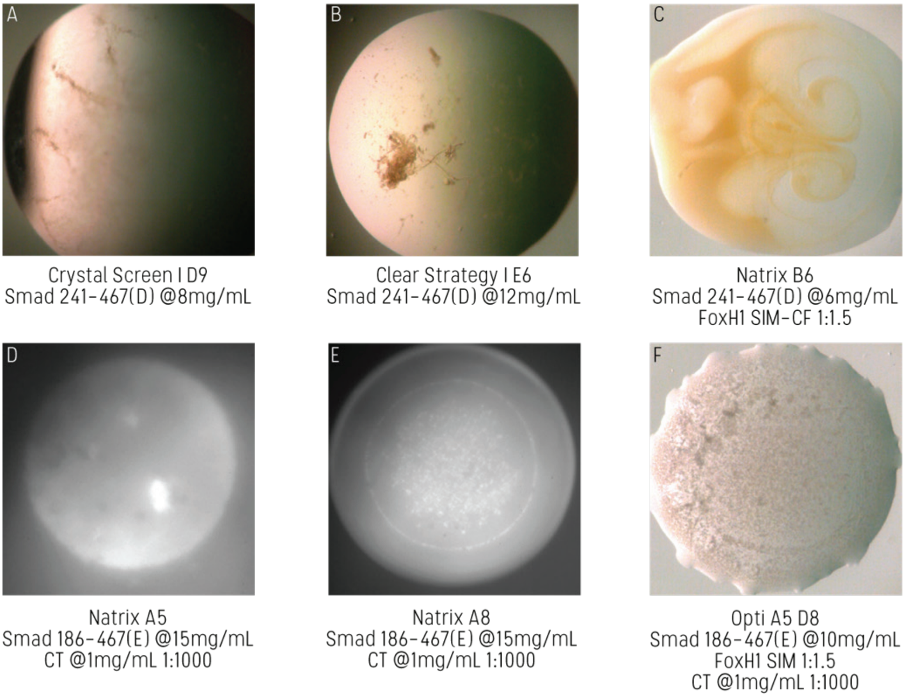
LP sequence was obtained by recombinant expression and purified under denaturing conditions. After purification, the protein was labelled with fluorescein isothiocyanate (FITC), which under specific conditions reacts mainly with the tertiary amine in the N-terminal. The labelled protein was purified and concentrated to be used in EMSA. The assay using the labelled tandem protein did not reveal any interaction with the Smad2 MH2 either (**Fig. 4.7C**).

4.1.9. Failure of First Crystallisation Screenings Suggests Problems Caused by Protein Stability and Peptide Solubility

Due to the size of the Smad2 MH2 - peptide complexes, especially in the case of the trimer-forming phosphomimetic constructs, macromolecular crystallisation was chosen as the best-fitting method to obtain tri-dimensional structures at atomic resolution.

Condition screenings were performed for the Smad2 241-468(DMD) and Smad2 186-468(EEME) constructs, either alone or in the presence of FoxH1 SIM or NCOA6 LXXLL-2 peptides. A series of commercial screenings was selected. These include empirical screenings, such as sparse matrix sampling based on statistical conditions (Jancarik and Kim 1991), and mathematical screenings such as Pi sampling (Gorrec et al. 2011).

Unfortunately, none of the screenings yielded crystals. Screenings using Smad2 241-468(DMD) consistently resulted in a majority of conditions presenting heavy, amorphous precipitation almost immediately **(Fig. 4.8A-C)**.



Crystal Screen I D9
Smad 241-467(D) @8mg/mL

Clear Strategy I E6
Smad 241-467(D) @12mg/mL

Natrix B6
Smad 241-467(D) @6mg/mL
FoxH1 SIM-CF 1:1.5

Natrix A5
Smad 186-467(E) @15mg/mL
CT @1mg/mL 1:1000

Natrix A8
Smad 186-467(E) @15mg/mL
CT @1mg/mL 1:1000

Opti A5 D8
Smad 186-467(E) @10mg/mL
FoxH1 SIM 1:1.5
CT @1mg/mL 1:1000

Figure 4.8. Visible and UV captures of crystallisation screening drops

Different kinds of precipitates can be seen in the drops using Smad2 241-467(DMD) protein with or without peptide (A-C). Using *In-situ* proteolysis with the larger Smad2 186-467(EEME) protein micro crystals appear in a few conditions (D-E) and different type of precipitates (G).

These results are consistent with the hypothesis on the existence of a soluble-aggregate equilibrium at high protein concentrations described in subsection 4.1.1. On the other hand, clear drops were the main outcome of Smad2 186-468(EEME) screenings, with precipitate appearing over time. Most likely, the large linker fragment that provides solubility to the protein construct is hindering crystallisation, since it is known that flexible, unstructured regions often interfere with nucleation. Consequently, none of the analysed conditions seemed promising enough for optimisation.

However, Smad2 186-468(EEME) is a construct that can be highly concentrated without relevant aggregation occurring thanks to its large N-terminal extension, which becomes a problem only in crystallisation screenings. Similar cases have been addressed using *in-situ* proteolysis (Dong et al. 2007; Wernimont and Edwards 2009), which consists on the progressive removal of flexible parts once in the screening drops by the addition of a protease, making crystal-yielding protein available progressively, as nucleation and crystal growth advance. Hence, adapting also methodology from the MH2 domain literature (Chacko et al. 2004), screening experiments were designed where a minimal amount of α-chymotrypsin (1:1000w/w) was added to the protein sample immediately before drop setting. A series of conditions produced protein microcrystals following this approach, although not in the presence of peptide **(Fig. 4.8D-F)**. Crystal-yielding conditions were optimised by single variable modification (pH and precipitant concentration) but crystallisation could not be reproduced.

Nevertheless, the screening results provided other relevant information as well. For instance, it revealed that the FoxH1 peptides presented a lower-than-expected solubility in aqueous buffer. This was identified after the formation of a solid layer in many condition drops of different screenings, which was attributed to the peptide after using the 6-carboxyfluorescein derivatives. Since conditions containing peptide could not form crystals, most likely these peptide layers were hindering protein nucleation.

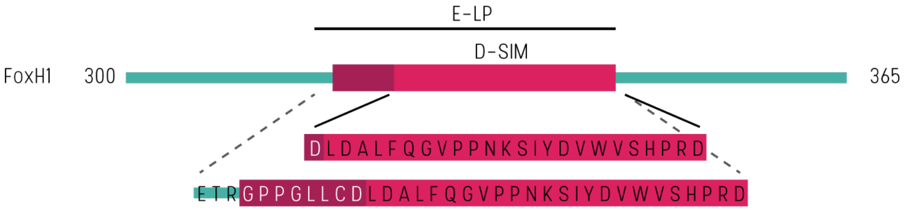

E-LP
D-SIM
FoxH1
300
365
DLDALFQGVPPNKSIYDVWVSHPRD
ETRGPPGLLCDLDALFQGVPPNKSIYDVWVSHPRD

Figure 4.9. Versions of FoxH1 SIM and FoxH1 LP peptides with improved solubility

New peptides incorporating an N-terminal extension to the original Smad-binding FoxH1 peptides incorporate charged residues to favour solubility.

4.1.10. Improvement of FoxH1 SIM peptides solubility: FoxH1 D-SIM and FoxH1 E-LP

In order to overcome the solubility problems detected in the crystallisation screenings, new versions of the FoxH1 SIM and FoxH1 LP peptides were designed. To avoid any alteration on their interaction with the Smad2 MH2 domain, straightforward options like the addition of an extension containing charged amino acids or mutating highly hydrophobic residues were not considered. Instead, solubility predictions were run and potential options were assessed by addition of adjacent residues in the FoxH1 sequence. By applying these results, new peptides consisting in the addition of just one amino acid to the FoxH1 SIM peptide and three to FoxH1 LP were synthesised (**Fig. 4.9**). These were named FoxH1 D-SIM and FoxH1 E-LP after the first residue in their N-terminal.

Since both peptides are N-terminal derivatives from the original FoxH1 SIM peptide, the synthesis was divided in two steps. A 0.5μmol scale synthesis of the FoxH1 D-SIM peptide was performed by microwave-assisted automatic SPPS. The resin was split in two, one half being directly cleaved to obtain the FoxH1 D-SIM peptide and the other half was used to manually extend the sequence to obtain FoxH1 E-LP. Cleavage and purification were carried out as usual. After freeze-drying the samples, the total peptide weight was obtained to calculate the synthesis yield (**Table 4.3**). EMSAs were conducted as a functionality check.

Table 4.3. Solid-phase peptide synthesis masses and yields

Peptide name	Predicted monoisotopic mass	Elution in AcN gradient	Experimental monoisotopic mass*	Final synthesis yield
FoxH1 D-SIM	2868.2	37-40%	2868.4	79%
FoxH1 E-LP	3876.3	41-45%	3877.5	55%

*Major species

4.1.11. Designing New Smad2 MH2 Constructs Balancing Protein Solubility and Linker Length

All the experimental data exposed in this thesis indicates that the length of the linker attached to the Smad2 MH2 at its N-terminal is the key determinant of the behaviour of the protein regarding its solubility and stability, particularly of its tendency to form aggregates. Therefore, a series of new constructs varying in linker length were designed.

According to the results exposed in subsection 4.1.1, the length-solubility balance would be at some point between residues 186 and 241 in the N-terminal. This deduction was the basis for the design of the new constructs. Five new constructs were designed and cloned[4], selecting the Smad2 186-467(EEME) construct as the basis. The constructs were therefore designed by discretely stablishing N-terminal boundaries, four of them starting at a residue within the aforementioned interval, and one slightly shorter to test for any potential benefits derived from the new vector chosen for the constructs (pOPINF vector, presenting an HRV-3C-protease-cleavable His6-tag; **Fig. 4.10**).

All constructs starting within the 186-241 range presented good solubility after lysis. On the other hand, the Smad2 248-467(EEME) was mostly insoluble. Soluble proteins were purified by IMAC and SEC in a phosphate buffer. From the soluble constructs, all but Smad2 248-467(EEME), the shortest, could be obtained

[4] New constructs were designed, cloned and tested for solubility after lysis with the assistance of the *Protein expression core facility* at the *IRB Barcelona*

at concentrations exceeding 1mM, similarly to Smad2 186-467(EEME). Smad2 238-467(EEME) could reach concentrations up to 400μM, still useful for structural studies, but it was more prone to temperature- and time-dependent aggregation.

Functionality of the new constructs was determined by EMSA. As a result, biophysical experiments were carried out using the Smad2 231-467(EEME) construct, which provides the best length/solubility balance. Both Smad2 231-467(EEME) and Smad2 238-467(EEME) proteins were selected to perform new crystallisation screenings, using the latter at a lower starting concentration.

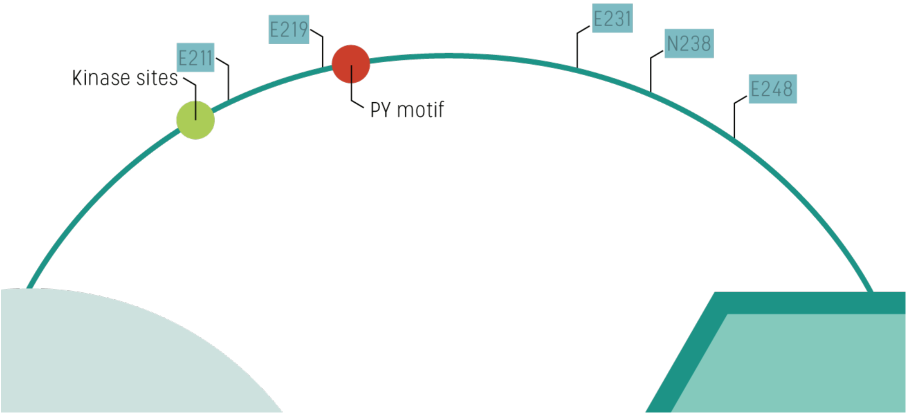
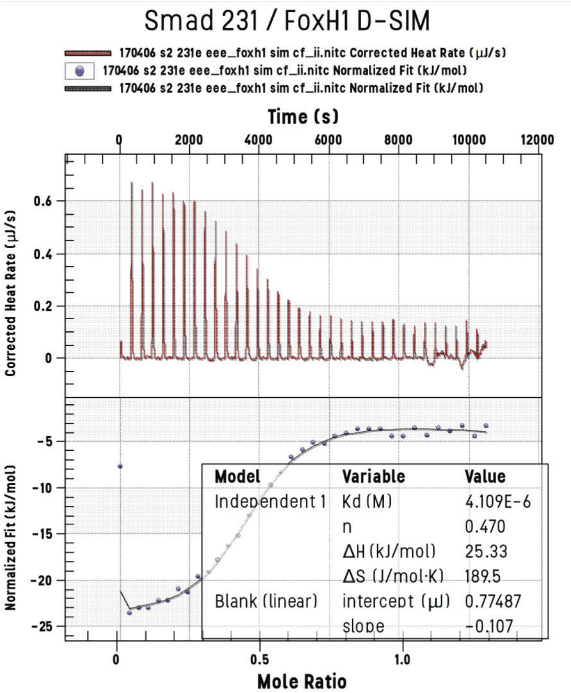


Figure 4.10. Smad2 MH2 constructs differing N-terminal extension for length optimisation

N-terminal boundaries of the new constructs were designed with the aim to optimise the balance between protein solubility provided by the presence of a large linker region and the requirements of macromolecular crystallography regarding the minimisation of flexible regions.

4.1.12. Affinity Constant Determination Indicates a Common Low-Molar Range Affinity for All Interacting Peptides

The determination of the affinity constant provides information on the dynamics of the interaction events between two (or more) elements. In this case, affinity was measured by isothermal titration calorimetry (ITC), which quantifies heat exchanges caused by the interaction in an adiabatic system. By applying a



Model	Variable	Value
Independent 1	Kd (M)	4.109E-6
	n	0.470
	ΔH (kJ/mol)	25.33
	ΔS (J/mol·K)	189.5
Blank (linear)	intercept (µJ)	0.77487
	slope	-0.107

Figure 4.11. Smad2 MH2 / FoxH1 D-SIM binding ITC curves and kinetics model

On the upper panel, heat exchange against time is represented. The graph follows the classical ITC profile caused by the binding of the titrated molecule. Peak integration after baseline subtraction results in normalised energy values. A mathematical model (in this case a sigmoidal curve with a linear blank correction) is fitted into the data to obtain the kinetic variables (lower panel). ITC curves for other peptides are available in the Appendix section **(Fig. S2).**

mathematic model to the integrated heat exchange for each titration step, affinity (or, more precisely, the value of the dissociation constant (K_d) governing the system) can be obtained.

To perform ITC experiments, the new Smad2 231-467(EEME) construct was selected. The probed peptides were FoxH1 D-SIM, FoxH1 E-LP and NCOA6 LXXLL-2. Due to the limited number of aromatic residues present in the peptides (or lack, in the case of theNCOA6 LXXLL-2 peptide) sample quantification was performed by HPLC-amino acid analysis[5].

Experiments were carried out placing the peptide in the cell and titrating small volumes of highly-concentrated protein. Appropriate concentration values for both elements were empirically determined with the assistance of the built-in curve

[5] HPLC amino acid analysis was conducted by *Unitat de tècniques separatives, Centres Científics i Tecnològics de la Universitat de Barcelona.*

Table 4.4. Dissociation constant values as obtained by ITC

Titrated protein	Peptide in cell	Dissociation constant (Kd)
Smad2 231-467(E)	FoxH1 D-SIM	4.41±0.8μM
TRIM33 Cter-LP	FoxH1 E-LP	3.33±0.4μM
FoxH1 LP	NCOA6 LXXLL-2	5.12±0.9μM

simulation. Experiments show an exothermic binding for FoxH1 peptides and endothermic for the NCOA6 LXXLL-2 peptide. Integrated data fits a sigmoidal model, allowing for the determination of kinetic variables (**Fig. 4.11; Fig. S2**).

Affinity values obtained by the aforementioned peptides show that all peptides present an affinity that is in the low-micromolar range (**Table 4.4**). Surprisingly, the affinity of the FoxH1 D-SIM peptide detected here is 3-fold higher than that previously reported using the FoxH1 SIM peptide (Bourgeois et al. 2013). Besides this observation, affinity differences between peptides are minimal, and despite the FoxH1 E-LP peptide presenting the lowest K_d value, the differences are not significant enough to neither proof or refute a critical role for the leucine-rich/ double proline motif.

4.1.13. Crystallisation Results Using Optimised Protein Constructs and Peptides

Prior to setting new crystallisation screenings using the new constructs and peptides, protein stability was assessed in buffers using lower salt concentrations (to avoid misleading salt crystals). With this goal, a thermofluor shift assay using a commercial buffer grid was performed. Of the 96 conditions tested, 18 provided a similar level of protein stability. For practical reasons, the bis-tris buffer system was chosen.

Dynamic light scattering (DLS) was used to analyse sample monodispersity in the low-salt bis-tris buffer. Results show that the sample is highly monodispersed.

Additionally, the predicted gyration radius suggests that – as expected – the protein is in an oligomeric state, likely a trimer.

To speed up the purification process and buffer exchange, an anion exchange step using bis-tris was used substituting the SEC step in protein samples for crystallography.

Again, several commercial screenings following different approaches were used as well as those optimisation screens designed in previous stages. Crystals were identified in three conditions, which were then harvested and stored in liquid nitrogen. Crystals were tested at the ESRF synchrotron. Out of these, two were salt crystals, and the other (the most promising one) was lost due to an operative failure of the sample changer. New experiments aiming to reproduce complex crystallisation may be conducted in the future.

4.1.14. Structure Analysis of the Smad2 MH2-Interacting Peptides

Peptides were dissolved in phosphate buffer, which is NMR-inactive, at a final concentration of 250μM. ^{1}H-TOCSY and ^{1}H-NOESY experiments were sequentially acquired to assign proton resonances and determine the tri-dimensional structures of the unbound peptides in solution. Additionally, natural abundance ^{13}C-HSQC experiments were acquired to determine the statistical probability of each residue to be part of a secondary structure element according to characteristic Cβ-Cα shifts.

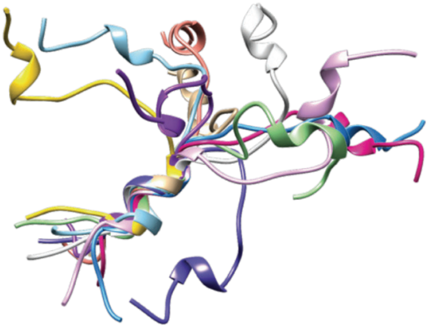


Figure 4.12. Preliminary structure of the FoxH1 D-SIM peptide

The solution NMR structure ensemble of the FoxH1 D-SIM peptide shows that it is an all-α protein with a disordered segment coinciding with the double proline fragment. The 10-lowest energy structures here represented have been aligned to the N-terminal helix.

The most remarkable case is that of FoxH1 SIM. By a great amount, the FoxH1 D-SIM peptide is the one displaying the largest number of intramolecular cross-peaks. A preliminary structure of the FoxH1 D-SIM peptide derived from NOE distances shows an all-alpha structure (**Fig. 4.12**) but, interestingly, the N-terminally extended FoxH1 E-LP does not present many of these cross-peaks despite containing FoxH1 D-SIM in its sequence. Previous reports on the matter indicate as well a lack of secondary structure of the FoxH1 SIM sequence (Randall et al. 2002). A similar characteristic has been described for some intrinsically disordered proteins (IDP), which present pre-structured, molten-globule-like motifs that are completely formed upon binding (Lee et al. 2012). On the other hand, NCOA6-LXXLL presents less NOE cross-peaks and its secondary structural propensities are less pronounced than that of FoxH1 peptides.

4.1.15. Looking for Helical Tendency to Design New TRIM33 Middle Peptides

After the results obtained from the FoxH1 D-SIM peptide regarding its secondary structure, a new question aroused regarding the possibility that some Smad2 MH2 binding sequences present a helical structure. This possibility was checked on the TRIM33 middle region. Since it had not been possible to discern what part of TRIM33 middle is the responsible for the interaction with the Smad2 MH2 domain, a secondary structure prediction software, PSIPRED, was used to look for regions with a potential helical tendency.

This analysis revealed two regions with a clear helical tendency. Surprisingly, these were the only two areas in the middle region with any probability for secondary structure at all according to the prediction results (**Fig. 4.13**).

The two predicted helical regions were obtained by manual SPPS. They were named TRIM33 PH-1 and TRIM33 PH-2. Synthesis, cleavage and purification were proceeded as usual. After freeze-drying the samples, the total peptide weight was obtained to calculate the synthesis yield (**Table 4.5**).

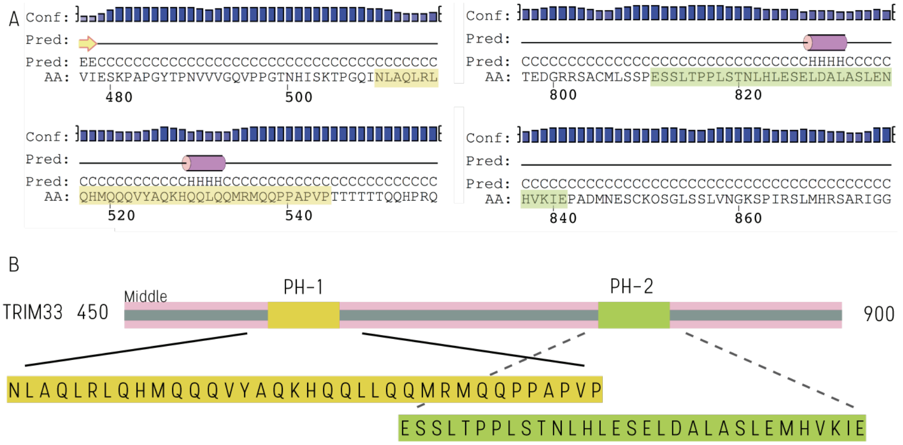

A
Conf:
Pred:
Pred: EECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
AA: VIESKPAPGYTPNVVVGQVPPGTNHISKTPGQINLAQLRL
480
500
Conf:
Pred:
Pred: CCCCCCCCCCCCHHHHCCCCCCCCCCCCCCCCCCCCCCCC
AA: QHMQQQVYAQKHQQLQQMRMQQPPAPVPTTTTTTQQHPRQ
520
540
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCC
AA: TEDGRRSACMLSSPESSLTPPLSTNLHLESELDALASLEN
800
820
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
AA: HVKIEPADMNESCKOSGLSSLVNGKSPIRSLMHRSARIGG
840
860
B
TRIM33 450
Middle
PH-1
PH-2
900
NLAQLRLQHMQQQVYAQKHQQLLQQMRMQQPPAPVP
ESSLTPPLSTNLHLESELDALASLEMHVKIE

Figure 4.13. Predicted helical regions within the TRIM33 middle region

A) Secondary structure prediction regions in the Smad-binding region (middle) in TRIM33. Two separated helical parts are predicted. **B)** Fragments of TRIM33 middle containing the predicted α-helixes are highlighted. Two sequences including the predicted α-helixes are obtained by SPPS.

The 6-Carboxyfluorescein variant of the peptides were used in EMSA experiments against the Smad2 231-467(EEME) construct. Unfortunately, no interaction could be detected in the case of TRIM33 PH-2 (**Fig. 4.14B**). On the other hand, the case of TRIM33 PH-1 was more peculiar, since its high isoelectric point required a modification of the EMSA protocol, changing the running buffer to TAPS-Imidazole and the use of pyronin as the loading and front dye. Moreover, the obtained EMSA results are debatable, since a dim fluorescent band is observed at the same height as the protein after coomassie staining (**Fig. 4.14A**). Nevertheless, the multiplicity of bands and the different intensities observed thorough the gel, most likely to the presence of impurities in the peptide sample,

Table 4.5. Solid-phase peptide synthesis masses and yields

Peptide name	Predicted monoisotopic mass	Elution in AcN gradient	Experimental monoisotopic mass*	Final synthesis yield
TRIM33 PH-1	4294.0	30-34%	4293.2	53%
TRIM33 PH-2	3386.7	41-45%	3398.5	70%

*Major species

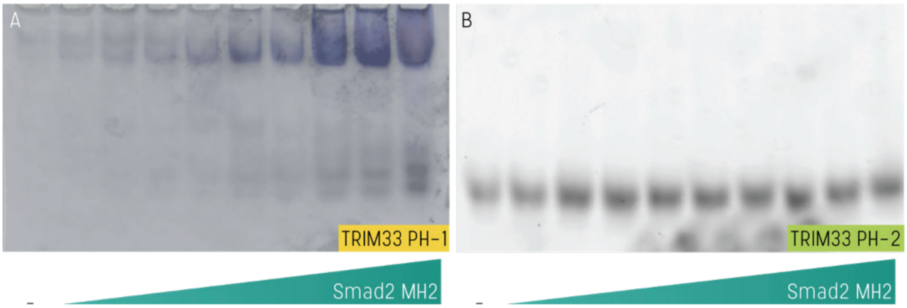


Figure 4.14. Interaction assays using predicted helical fragments in TRIM33 middle

A) EMSA to assess the interaction of the N-terminal predicted helix (PH-1) peptide and Smad2 186-467(EEME). Overlayed, a Coomassie staining of the same gel. **B)** EMSA using the C-terminal predicted helix (PH-2) instead.

make it difficult to attribute it to binding with certainty. Alternatively, for the TRIM33 PH-1 peptide or any other sequence within the TRIM33 middle region, a different approach to determine the specific Smad-binding sequence may be applied in the future.

4.2. Discerning Residues in the FBP28 WW2 Domain Defining Folding Kinetics[6]

Protein folding is a field in the biological sciences that aims to describe from a physico-chemical point of view how protein sequences reach their biologically active conformation(s). Protein folding consists on a series of events relying on dynamic conformational exchange that govern the equilibrium between folded and unfolded conformations, including irreversible steps towards misfolded states. Therefore, protein folding is a delicate cellular process that can be fatal in the case

[6] The work presented in this chapter was published in: Maisuradze GG, Medina J, Kachlishvili K, Krupa P, Mozolewska MA, Martin-Malpartida P, Maisuradze L, Macias MJ, Scheraga HA. 2015. **Preventing fibril formation of a protein by selective mutation**. *Proc Natl Acad Sci U S A* **112**: 13549–13554. A printout of the manuscript is available as an appended document in the Appendix section.

of protein misfolding and accumulation. Cellular mechanisms exist for large or damage- and aggregation-prone proteins in the form of fold-rescuing machinery such as chaperones and chaperonins. Nevertheless, a group of diseases exists that is caused by the cellular accumulation of self-propagating misfolded protein called amyloidosis.

Biophysical methods like NMR and *in-silico* simulations are commonly used to study the dynamics of the folding events. To describe the rules of folding and unfolding processes, model proteins and site directed mutagenesis have been employed (Fersht 2008). Among the model systems, WW domains have been selected to study β-sheet formation have used due to its small, compact fold and fast folding - unfolding rate as well as its biological relevance (Jäger et al. 2001; Ferguson et al. 2001; Liu et al. 2008).

In the particular case of the second WW domain present in the murine Formin Binding Protein 28 (FBP28; Uniprot entry: Q8CGF7), identical to human homologue Transcription Elongation Regulator 1, reports based on diverse experimental methods have proposed conflicting hypotheses regarding how the domain folds and unfolds. While early publications described FBP28 WW2 to fold following two-state kinetics (Ferguson et al. 2001), later works indicated the existence of an dry molten globule intermediate state whose nature and composition could be modulated by temperature, domain truncation and point mutations (Nguyen et al. 2003; Davis and Dyer 2014). This intermediate state was proposed to be involved in the formation of fibrils similar in structure to those found in amyloidosis (Ferguson et al. 2003; Mu et al. 2006).

Molecular dynamics (MD) simulations can comprehensively illustrate protein folding and unfolding folding mechanisms, including the existence of intermediates difficult to characterise otherwise. Many popular MD packages rely on all-atom force fields (Pronk et al. 2013; Salomon-Ferrer et al. 2013), which are computationally intensive. Since complex biological events, including protein folding, occur in the millisecond to second timescale, the required computing power required by all-atom approaches has long been inaccessible to most laboratories. Alternatives to all-atom simulations are the family of coarse-grained models, where molecules are represented by combinations of atoms and pseudo-atoms. The main advantage is that, by decreasing the degrees of freedom, much longer simulation times can be studied than using classical atomistic models. An important number of coarse grained models have been designed in order to predict protein structures and protein interactions as well as dynamics. The pioneer work of Michael Levitt, Ariel Warshel, and Martin Karplus on the development of multiscale models for complex chemical systems based on coarse-grained protein modelling was recognized with the Nobel Prize award in Chemistry in 2013 (André 2014).

Prior work on the physics-based united-residue (UNRES) force field validated its ability to describe folding trajectories of small, fast-folding proteins (Liwo et al. 2005). UNRES models the peptide backbone by generating two coarse-grained beads (an interacting peptide-group and a non-interacting group) and the side-chains as a single ellipsoidal bead. The UNRES potential is a free energy function where all degrees of freedom are averaged out into effective potentials, except for those describing protein conformation. It distinguishes between bonded interactions, which include bonds, angles, and dihedrals and side-chain rotational potential, and non-bonded interactions, van der Waals and electrostatic forces between interacting and side-chain beads derived from ab initio or semi-empirical calculations. Additionally, the UNRES force field incorporates temperature-dependent correlation terms (Ingólfsson et al. 2014).

The folding properties of FBP28 WW2 wild-type (WT) and a group of mutants carrying single changes on its strand-crossing hydrophobic cluster (Y440/Y448/W459) were previously studied by UNRES-based Langevin MD and solution NMR as part of a collaboration between the Macias laboratory at IRB Barcelona and the Scheraga laboratory at Cornell University (Zhou et al. 2014). In this work, only a minor role of the mutated residues in the emergence of the intermediate states was observed, which predominantly consisted of β-strand1 and β-strand2, lacking the third strand characteristic of WW domains. This observation was in agreement with previous findings connecting intermediate state formation to slow register of β-hairpin2 (Karanicolas and Brooks 2003) and the particular propensity of the β-turn2 sequence to be intrinsically structured (Mu et al. 2006).

As the concluding phase of the collaborative project and with the aim to further investigate the formation of the β-strand2 and β-strand3 pair in FBP28 WW2, a series of new single-point mutants were designed to alter residues potentially involved in the key intermediate-to-native folding step. The effects of the mutations were studied combining high-resolution NMR and UNRES force field – Langevin MD. The end purpose of this project was to better understand how amyloid fibrils are initiated in neurodegenerative diseases by using FBP28 WW2 as a model system.

4.2.1. FBP28 WW2 Mutant Domains Present the Characteristic WW Domain Fold

Based on previously reported data based on molecular dynamics simulations (Maisuradze et al. 2013), three amino acids (L455, E456 and T458) were identified to be involved in the final folding step towards the native state. These three residues are placed along the third β-strand of the FBP28 WW2 domain, two of them very close to β-turn2 (**Fig. 4.15**).

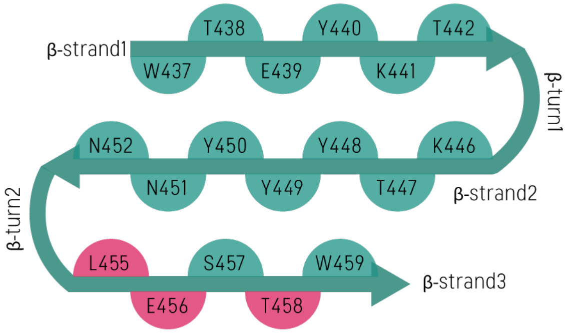
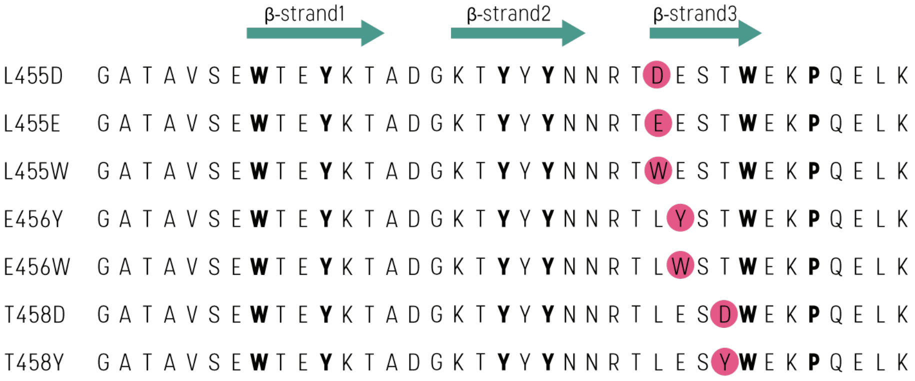


Figure 4.15. Layout of the FBP28 WW2 domain

Arrows picture the characteristic anti-parallel β-sheet folding of WW domains. Amino acid residues along the sequence are represented as semicircles pointing to one side of the β-sheet plane according to side-chain orientation. Residues described to potentially define the intermediate-to-native folding step (in magenta) were mutated with the aim to affect the folding trajectory.



L455D GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTDEST**W**EK**P**QELK

L455E GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTEEST**W**EK**P**QELK

L455W GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTWEST**W**EK**P**QELK

E456Y GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTLYST**W**EK**P**QELK

E456W GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTLWST**W**EK**P**QELK

T458D GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTLESD**W**EK**P**QELK

T458Y GATAVSE**W**TE**Y**KTADGKT**Y**Y**Y**NNRTLESY**W**EK**P**QELK

Figure 4.16. Sequences from the single-change FBP28 WW2 mutant domains

For all the mutant domains generated in this study, the respective sequences are displayed. Residues which are required to structure the domain are shown in bold. The new single-point mutations are highlighted in magenta.

To characterise the role of these positions in the folding-unfolding pathway, a total of seven new single-change mutants were designed affecting the aforementioned positions (**Fig 4.16**). L455 was substituted by either aspartic acid, glutamic acid or tryptophan. L455D and L455E mutations aimed to study the effect of changing a hydrophobic residue by a charged counterpart, since charged residues are often present at this position in the family of WW domain sequences. On the other hand, L455W was conceived to better understand the hydrophobic effect of wild-type L455 and previously studied mutant L455A on the formation of β-hairpin2 (Nguyen et al. 2003; Zhou et al. 2014). E456Y and E456W mutants were designed to participate in potential hydrophobic interactions with W437 and Y449 core residues, possibly reducing β-strand3 registry fluctuations and improving domain stability. Last, T458D and T458Y mutants were designed to observe changes in the formation of β-strand3 and the turn located immediately after.

The plasmid containing the FBP28 WW2 wild-type (Macias et al. 2000) was used as a template to obtain the new mutants by site-directed mutagenesis. 30-base-long primers that contained the designed mutation were designed. Mutant proteins were expressed in LB medium and purified using GST-affinity chromatography and SEC. Six out of seven mutants were successfully purified. However, E456W remained almost completely in the insoluble fraction after lysis and only marginal

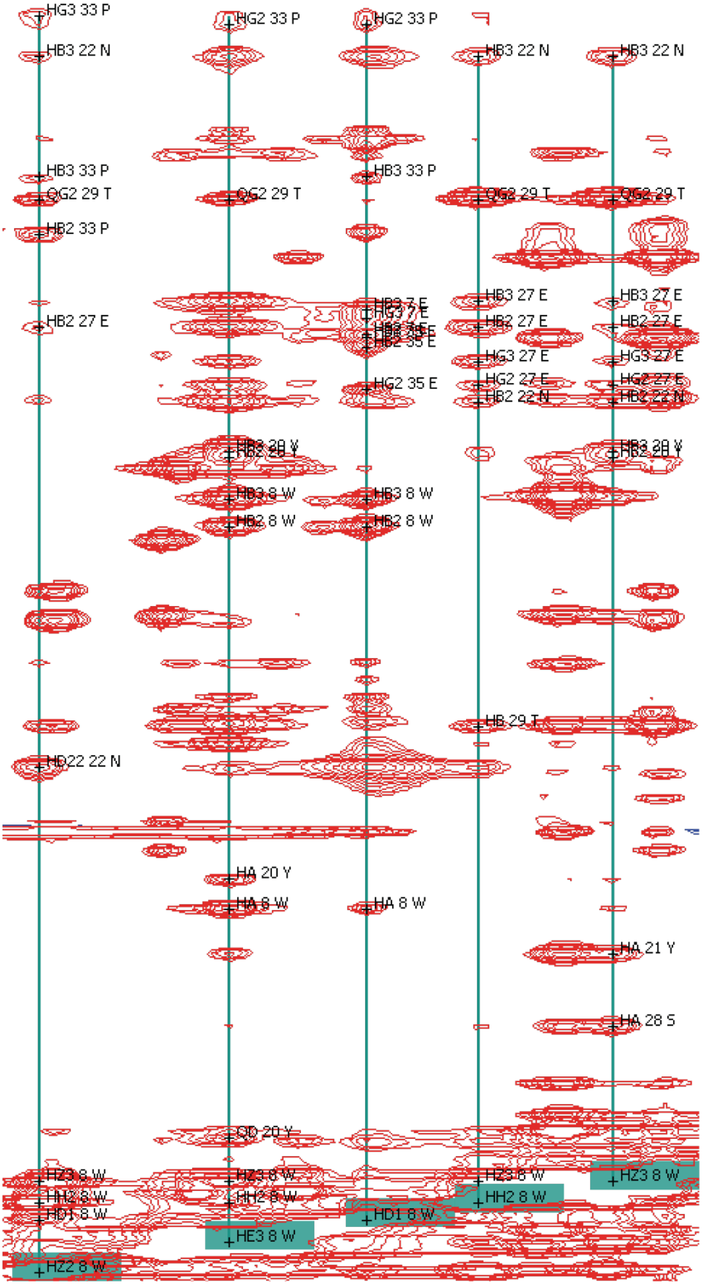


Figure 4.17. W437 aromatic protons in ^{1}H-NOESY from FBP28 WW2 L455W mutant domain shows strand-crossing contacts

Strand-crossing contacts are fundamental for the maintenance of the overall WW domain structure. Here, a ^{1}H-NOESY detail showing all contacts from aromatic protons HZ2, HE3, HD1, HH2 and HZ3 in W437. It can be observed that these protons are close in space with distance residues in the sequence belonging to the second and third ß-strands and the C-terminal *tail*.

quantities could be recovered. For this reason, a Ni^{2+}-charged IMAC purification protocol under denaturing conditions was carried out. However, the final recovery of E456W protein was limited, impeding the acquisition of high-quality NMR data required for structure determination.

For the remaining samples, ^{1}H-monodimensional, TOCSY and NOESY ^{1}H-bidimensional NMR experiments were acquired at 285K in aqueous buffer. Spin systems corresponding to all the residues in the sequence were identified and NOE peaks defining the pattern of contacts in the β-sheet formation were assigned **(Fig. 4.17)**. Experiments were also acquired in D_2O samples to simplify the assignment of alpha and aromatic protons' through-bond correlations. In the case of NMR data obtained from L455E and E456Y, the presence of two conformational states, which are slow in terms of NMR timescale, is revealed. This results in peak duplication for many proton correlations in equilibrium at 285K. An increase in the acquisition temperature up to 298K and the addition of 5% dimethyl sulfoxide-d6 to the sample were used to favour one of the states and thus facilitate peak assignment.

NMR-derived restraints were used to calculate the structures in solution of each mutant domain. The obtained structures confirm that all domains present the characteristic triple anti-parallel β-sheet folding of WW domains. During the refinement of the structure, restraints violated for the calculations were corrected on the assignment until a sufficient convergence was reached. For each structure, an ensemble containing the 20 lowest-energy structures (out of 300 calculated structures) is obtained. Methods for the analysis of dihedral angle distribution along the backbone and the side-chains (such as Ramachandran plots) were used as validation tools. Moreover, energy values were compared with the wild-type and previously-characterised mutants. Structures and the NMR assignments were deposited in the Protein Data Bank (PDB) and Biological Magnetic Resonance Bank (BMRB) public repositories, respectively.

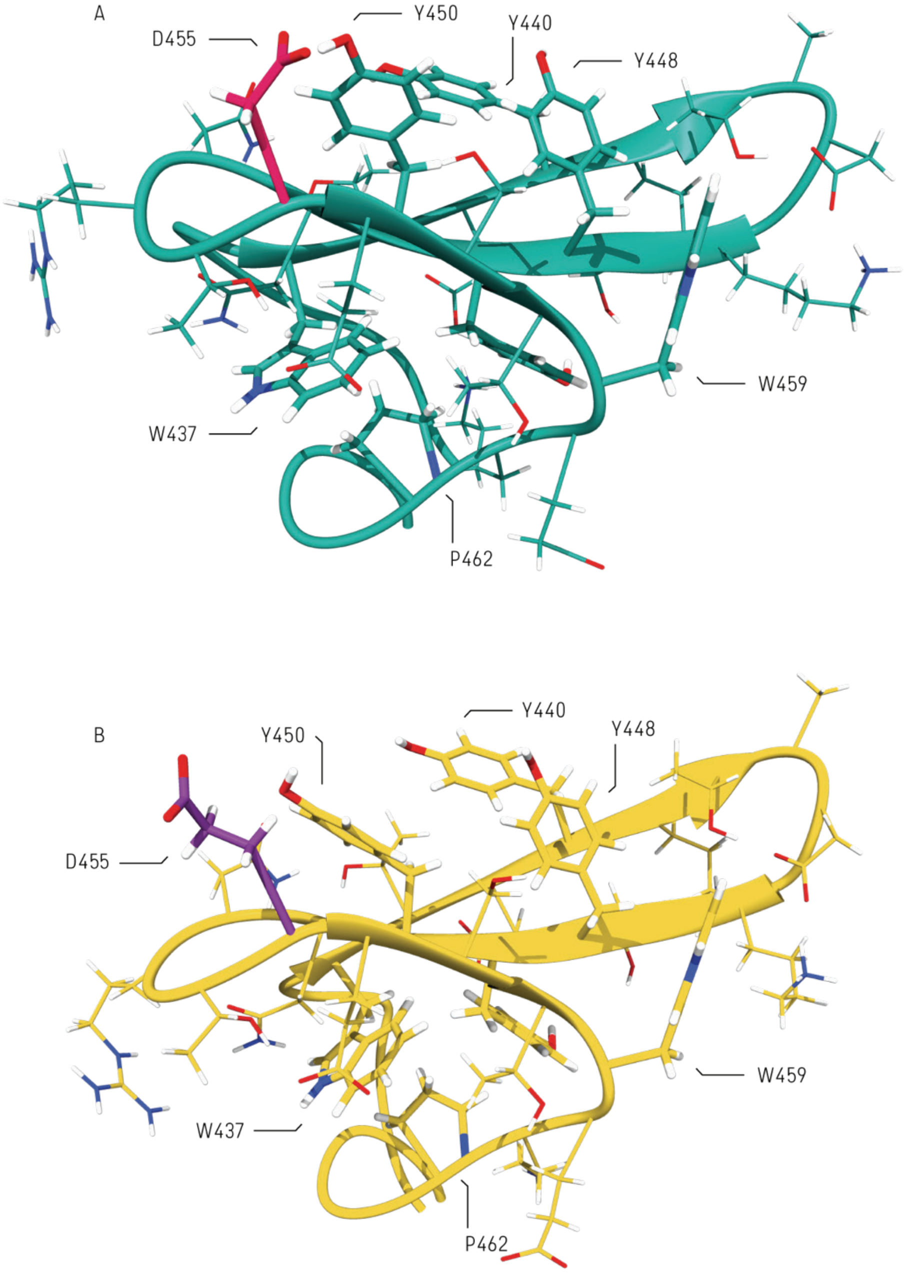


Figure 4.18. Structures of FBP28 L455D and L455E WW2 mutant domains

The lowest-energy structures in the ensemble calculated using NMR restraints from **A)** mutant L455D (PDB ID: 2n4r; BMRB ID: 25678) and **B)** mutant L455E (PDB ID: 2n4s; BMRB ID: 25679) are shown. A contrasting colour is used to display mutated residues.

4.2.2. Specific Structural Features of the FBP28 WW2 Mutant Domains

While all mutant domains presented the characteristic triple anti-parallel β-sheet folding of WW domains, a series of structural characteristics specific of each mutant protein were revealed by their structures and further explained qualitative analysis of NMR data.

Mutation of L455 to an acidic residue has singularly different consequences. In mutant L455D (PDB ID 2n4r; BMRB ID: 25678), residues Y450 and D455, key in the formation of β-turn2, diverge in dihedral angle χ_1 value respect to the wild-type L455. This results in a conformation where the carboxylic group in D455 is consistently oriented towards the hydroxyl in Y450 (**Fig. 4.18A**), indicating the potential presence of a water-mediated hydrogen bond that stabilises this specific orientation. This situation implies that this specific single change may have an effect on both the structure of β-turn2 and the correct registry of β-strand 2-3 pair. On the other hand, E455 in mutant L455E (PDB ID 2n4s; BMRB ID: 25679),

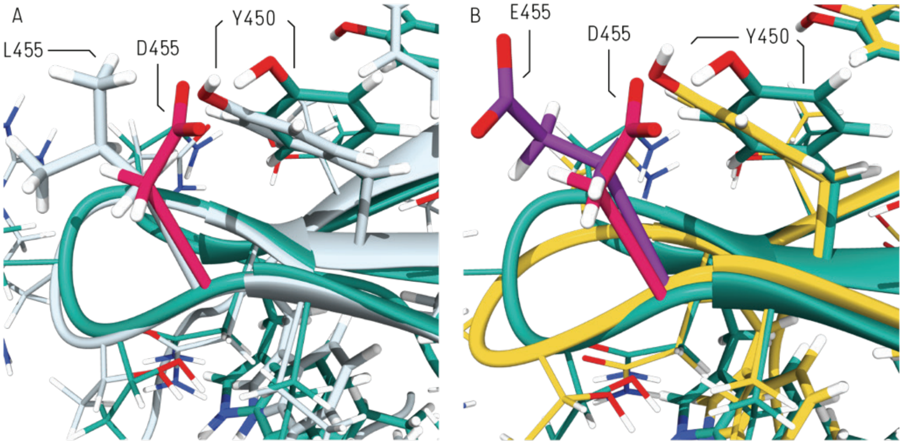


Figure 4.19. Detailed view of FBP28 L455D β-turn2 alignments

A) FBP28 L455D mutant domain aligned with FBP28 WW2 wild-type (PDB ID: 1E0L). This comparative reveals a similar turn structure despite the differences in the interactions of residue 455 with Y449. **B)** FBP28 L455D mutant domain aligned with FBP28 L455E. A lack of hydrophobic contacts between Y449 and E455 could explain the different turn structure in L455E.

which contains an extra CH_2 group, is arranged almost identically to the wild-type L455 (**Fig. 4.18B**).

Despite the previously described differences between L455D and the wild-type, their corresponding β-turns2 are structurally very similar (**Fig 4.19A**). Surprisingly, this is not the case for the L455E mutant, which presents a remarkably different turn structure. These dissimilarities could be explained by the differential degree of interaction between the residue in position 455 and Y450: whereas the aforementioned hydrogen bond in L445D would induce a turn structure equivalent to that induced by hydrophobic contacts between L445 and Y450 side chains in the wild-type, there is an apparent absence of contacts between E445 and Y450 in mutant L455E that results in a more flexible β-turn (**Fig. 4.19B**).

In the L455W mutant domain (PDB ID 2n4t; BMRB ID: 25680), the ensemble of structures reveals that several orientations of the W455 indole group

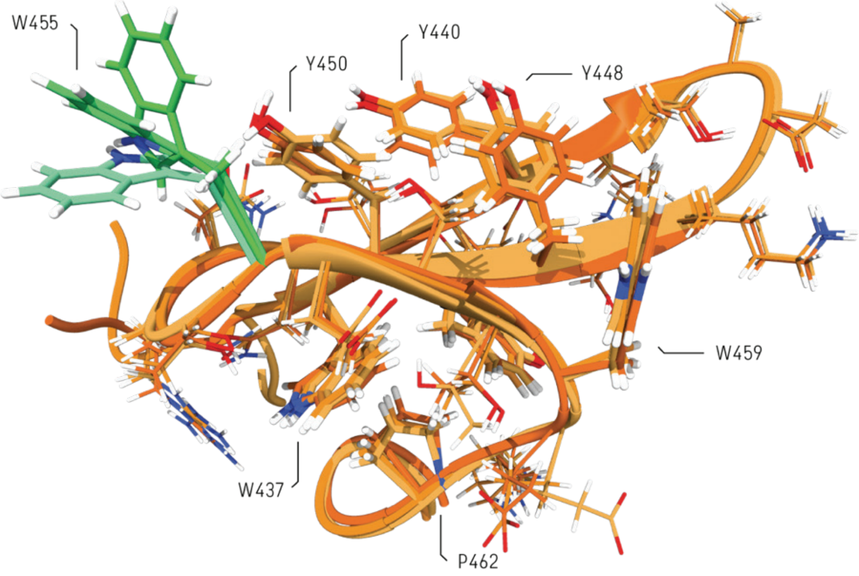


Figure 4.20. Structure of the FBP28 L455W WW2 mutant domain

A total of three structures included in the structure ensemble of mutant L455W (PDB ID: 2n4t; BMRB ID: 25680) are shown. Stronger colours indicate an overall lower energy and a contrasting colour is used to highlight the mutation. The different orientations of the W455 indole ring display a range of potential mobility for the aromatic ring that would comply with NMR restrictions.

are compatible with the experimental restraints even though NOEs correlating the aromatic rings of W455 and Y450 are observed in the NMR spectrum. This is due to the W455 χ_2 value that varies within a range of -60° to +60° in the twenty lowest-energy structures. In spite of that, the χ_1 values are consistently similar to the wild-type **(Fig. 4.20)**.

In the E456Y mutant, due to the significantly different steric and chemical properties presented by the wild-type glutamic acid and the mutant tyrosine residue, a reorganisation of side chains around the mutation site is forced as displayed by their structure (PDB ID 2n4u; BMRB ID: 25681). While the E456 side chain in the wild-type has almost no contacts and is exposed to the solvent, the voluminous and hydrophobic tyrosine requires to be solvent-protected and thus accommodated in the crowded "lower" face of the WW domain. As a result, the phenol ring of Y456 is tightly packed between the β-turn2 and the indole of W437 **(Fig. 4.21)**. The cross-strand contacts between W437 and Y456 rings are oriented

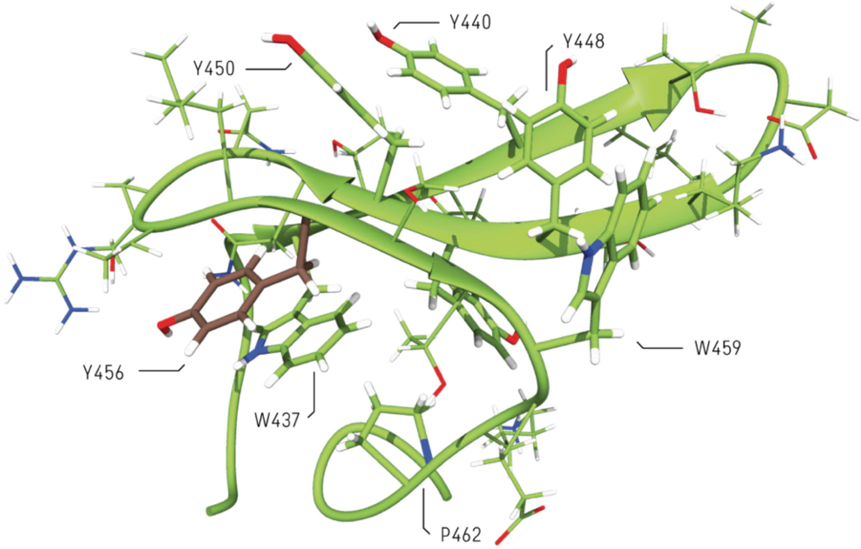


Figure 4.21. Structure of the FBP28 E456Y WW2 mutant domain

The lowest-energy structure in the ensemble calculated using NMR restraints from mutant E456Y (PDB ID: 2n4u; BMRB ID: 25681) is shown. A contrasting colour is used to highlight the mutation. Y456 is arranged towards the "back" of the domain (as opposed to solvent-oriented the wild-type glutamic acid) by means of cross-strand interactions with W437.

adopting a parallel-displaced pi stacking, a common geometry between aromatic rings together with the T-shaped (one ring is oriented perpendicular to the other).

The introduction of Y456 also affects the structure of β-turn2, which is less pronounced in comparison to the wild-type. Additionally, the tight packing of Y456 provides a feasible explanation to the low-solubility properties E456W protein, where an arrangement of the tryptophan side-chain similar to that observed for the Y456 is sterically disallowed, likely leading to protein misfolding and aggregation.

Mutations to T458 seem to have contrasting effects due to the different size and polarity of the substituting residues. From a structural point of view, the T458D (PDB ID 2n4v; BMRB ID: 25682) mutant is very similar to the wild-type domain. The dihedral angles formed by D458 are identical to the wild-type T458, which results in the carboxylic group from D458 being exposed to the solvent in a way comparable to the side chain hydroxyl in T458 **(Fig. 4.22A)**.

On the other hand, substitution of T458 by a voluminous, hydrophobic residue as is the case of mutant T458Y (PDB ID 2n4w; BMRB ID: 25683) generates a hydrophobic surface that leads to contacts between the Y458 phenol ring and P462. As a result, the turn at the end of β-strand3 (right after W459) is sharper compared to the wild-type domain. Besides, two possible orientations of Y458's aromatic ring are observed which are compatible with the experimental restraints **(Fig. 4.22B)**.

4.2.3. Single Change Mutations Affect the Thermal Stability of the FBP28 WW2 Domain

The effect on folding stability derived from the mutations introduced in the FBP28 WW2 domain was further studied by analysing the thermal stability of each protein by thermal denaturation. Thermal denaturation of a protein is a process where protein motion is induced by heating, leading to the complete elimination of its three-dimensional structure. Experiments based on denaturation can be used

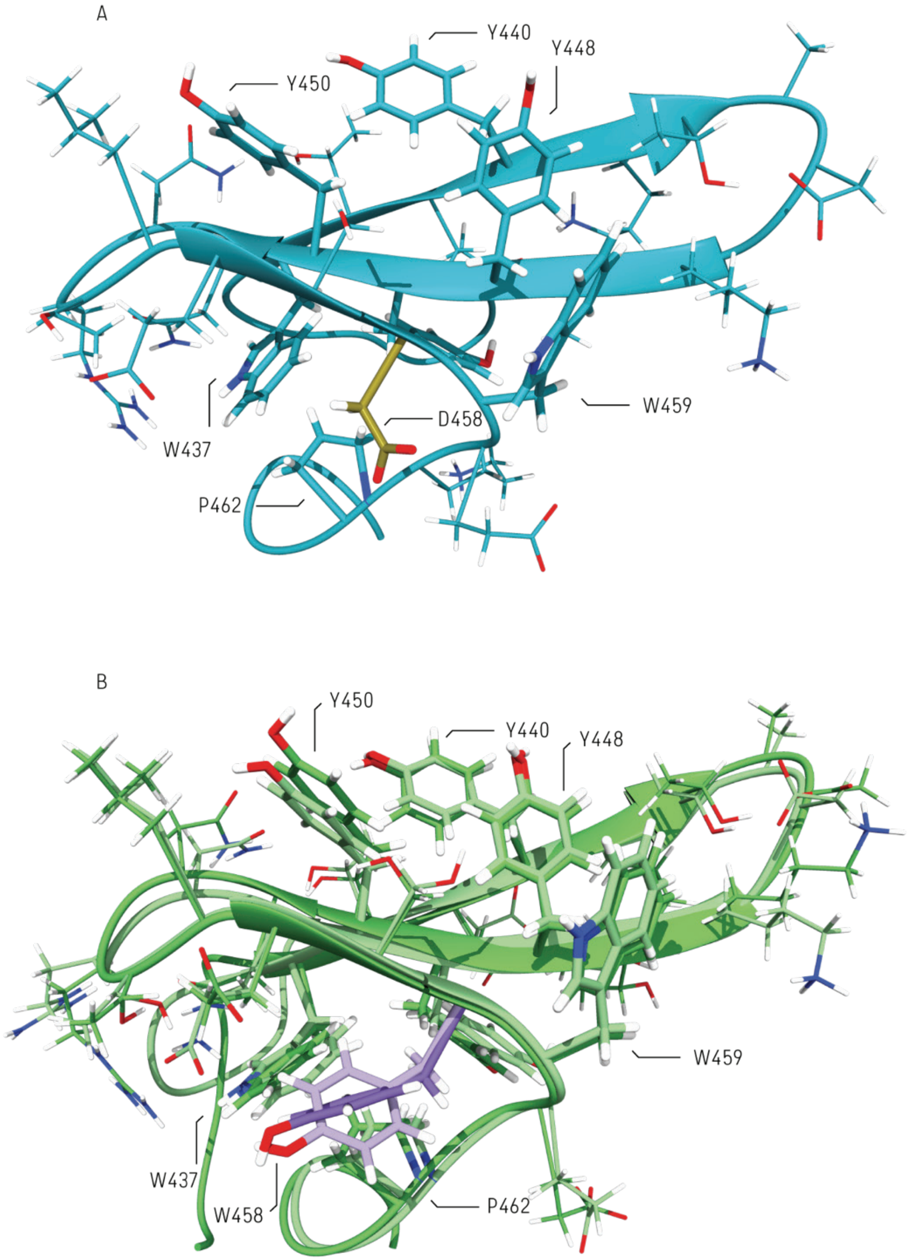


Figure 4.22. Structure of FBP28 T458D and T458Y WW mutant domains

The lowest-energy structures in the ensemble calculated using NMR restraints from **A)** mutant T458D (PDB ID: 2n4v; BMRB ID: 25682) and **B)** mutant T458Y (PDB ID: 2n4w; BMRB ID: 25683) are shown. A contrasting colour is used to display mutated residues. In **B)**, two different structures included in the structure ensemble display possible orientations of the Y458 side chain.

to measure the stability of biologically functional proteins and to comparatively analyse potential effects introduced by point mutations in the overall stability of the protein. The melting temperature (T_m) is a parameter widely used to describe thermal stability which is defined as the temperature at which half of the protein is unfolded.

There are several techniques to probe thermal denaturation of proteins, including classical methods exploiting fluorescence emission or direct measurements of heat using calorimetry.

Intrinsic fluorescence spectroscopy was the method of choice since WW domains are rich in aromatic, fluorescent residues and it also circumvents the need of an external fluorescent probe that relies on the exposure of large hydrophobic surfaces due to protein unfolding. The emission λ_{max} shift phenomena caused by the sensitivity of the tryptophan ring to local environment (Vivian and Callis 2001) was used to follow protein melting caused by temperature. Buffer exchange to an acetate buffer was performed on samples purified in a Tris buffer to avoid temperature-driven pH changes.

Fluorescence emission spectra of the single-change mutant proteins and the wild-type (used as a control) were acquired across a broad temperature span which included the expected range of temperature for the melting transition. For obvious reasons, fluorescence intensity of proteins containing a mutation to an aromatic residue was higher, especially that of L455W. Correlating the emission λ_{max} value with temperature results in a sigmoidal curve. The coordinates of the sigmoidal model representing 50% unfolding correspond to the T_m (see Fig. 3.9 in Materials and Methods for a schematic representation).

Analysis of the intrinsic fluorescence spectroscopy reveals a high data-point dispersion in each experimental run. Moreover, despite reaching experimental temperatures close to 100°C, the resulting melting curves do not reach saturation. It is the case of L455D, with several data points falling far from the adjusted curve and a short saturation plateau **(Fig 4.23A)**. Most remarkably, the T_m value presented by the wild-type protein differs by 15K from the T_m obtained by far-UV circular dichroism (Nguyen et al. 2003). These differences could indicate that fluorescence data is unreliable in this specific case. There are several elements supporting this interpretation: a) the effect of temperatures on the quantum yield of fluorophores, since higher temperatures cause a decrease in quantum yield and therefore the determination of the precise λ_{max} becomes increasingly complex in later experimental points. Moreover, spatial proximity of fluorophores (as is the case in a small, aromatic-rich domain like FBP28 WW2) can derive in internal fluorescence quenching (Hofmann 2010). Also, technical limitations regarding temperature control may have affected the measurements.

T_m measurements were also acquired using differential scanning calorimetry (DSC). DSC can measure the heat capacity (C_p) of a sample through a range of temperatures. Since an unfolding event is a transition involving heat exchange, the heat capacity maximum ($C_{p,\,max}$) can be correlated to that temperature where most unfolding events occur; that is, define the T_m. Even though the small size of the WW domain implies that C_p changes caused by unfolding transitions are expected

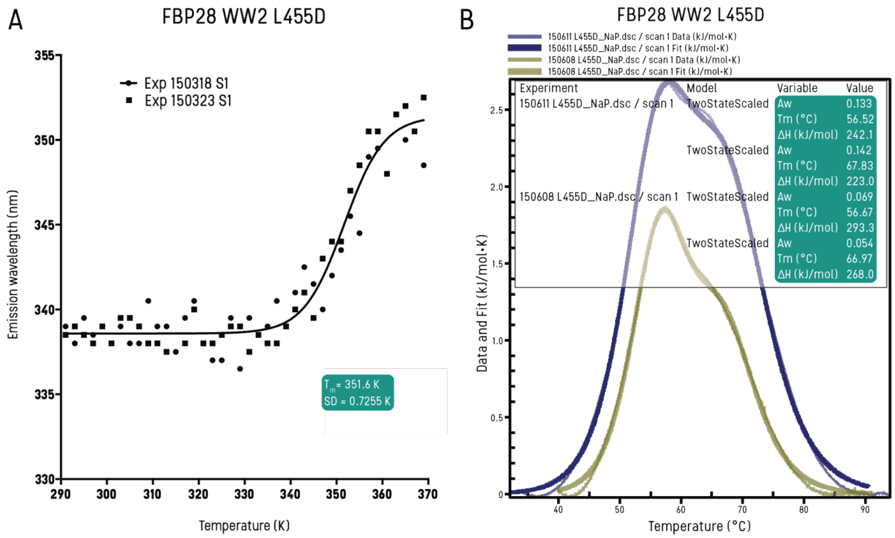


Figure 4.23. Thermal stability analysis of FBP28 L455D mutant domain

A) Intrinsic fluorescence spectroscopy data represented as λ_{max} against temperature. Data points from duplicate experiments are represented. Data is globally fitted to a sigmoidal curve. **B)** Differential scanning calorimetry data and fitting curves for duplicate experiments. Data acquired from the rest of FBP28 WW2 mutant domain is presented in Appendix section **(Fig. S3-S5)**.

Table 4.6. Melting temperatures of the FBP28 WW2 wild-type and single-change mutants

T_m as measured using:	Intrinsic Fluorescence Spectroscopy	Differential Scanning Calorimetry	Published data[7]
FBP28 WW2 WT	352K	339K	337K
FBP28 WW2 L455D	352K	330K‡	N. A.
FBP28 WW2 L455E	348K	334K	N. A.
FBP28 WW2 L455W	340K	332K	N. A.
FBP28 WW2 E456Y	334K	-	N. A.
FBP28 WW2 T458D	334K	329K	N. A.
FBP28 WW2 T458Y	351K*	326K	N. A.

*High error, curve did not reach saturation

‡Main peak value applies

to be small, technological advances in calorimetry equipment permitted to obtain reliable data in a NanoDSC calorimeter. Fresh buffer was used as the reference cell blank and a temperature gradient from 20ºC to 95ºC was applied. The DSC data fitted to a Gaussian model and the $C_{p,\,max}$ correlated to the T_m.

The T_m value obtained by the wild-type domain is within a difference of 2ºC respect to the value reported by Kelly and co-workers (Nguyen et al. 2003). Based on this observation and the lower intra- and inter-experiment variability, the values obtained by DSC have been considered to be more representative than the fluorescence data. All mutant proteins present a lower T_m than the wild-type. Interestingly, L455D presents a composed melting curve **(Fig. 4.23B)**, suggesting that two unfolding events occur, each one presenting a different T_m. On the other hand, mutations in position 455 oriented similarly to the wild-type (L455E and L455W) show a moderately higher T_m than the rest of the mutants (see **Fig. S3-S5** for the rest of melting data).

[7] Data extracted from: Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW. 2003. **Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation.** *Proc Natl Acad Sci U S A* **100**: 3948–3953.

The T_m values obtained for all FBP28 WW2 mutant domains by both methodologies are shown in **Table 4.6**.

4.2.4. Alternative Folding Trajectories are Described for FBP28 WW2 L453D and L453W Mutant Domains[8]

Folding trajectories of the single-change mutant proteins were simulated by canonical Langevin molecular dynamics using the UNRES force field. To deal with the multidimensionality of the system, dihedral angle principal component analysis (PCA) was applied to the mean-square fluctuations (MSF). A set of principal components (PC) modes is considered to significantly represent the MSF if they describe ≤40% of the fluctuations. Free-energy landscapes (FEL) were plotted along the modes of the two principal components (PCs) contributing the most to the MSF. Each local minimum in a FEL is correlated to a folding state.

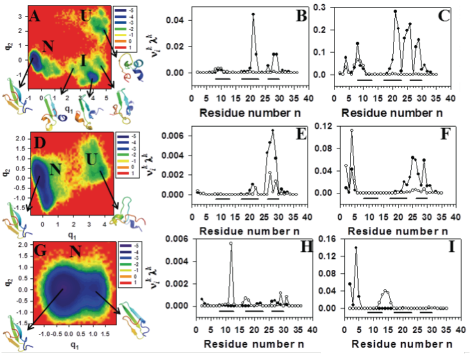


Figure 4.24. Free-energy landscapes and Mean-square fluctuations of FBP28 WW2 L455D mutant domain[9]

Free-energy landscapes (FELs) and mean-square fluctuations (MSF) representative for the three-state (A-C), the two-state (D-F) and downhill (G-I) trajectories specific for L26D. In FELs, I is intermediate, N is native and U is unfolded. MSF plots B, E and H correspond to MSF along θ while plots C, F and I correspond to MSF along γ. In MSF, black points represent the contribution to MSF by PC1 while white points represent the contribution by PC2.

FELs display that the three-state scenario (that is, folding through the formation of an intermediate) is dominant for all systems (**Fig 4.24**). Interestingly,

[8] This part of the work was carried out by collaborators at the Baker Laboratory of Chemistry and Chemical Biology led by H. A. Scheraga and it is briefly described for contextualisation.

[9] This figure was gently remitted by G. G. Maisuradze and H. A. Scheraga.

mutants L455D and L455W also present minor two-state and downhill folding trajectories.

The existence of alternative folding pathways and the dominance of biphasic kinetics is in agreement with previous simulation studies on the wild-type domain (Xu et al. 2011). MSFs of dihedral angle PCs indicate which residues fluctuate the most in each folding step. For all cases but for downhill folding, residues in β-turn2 are among the most fluctuating. Even though different temperatures were evaluated, its influence in the frequency of different folding trajectories remains unclear.

By analysing the distances between Cα through the simulation time, contact and stabilisation times were stablished. This information correlates a biphasic folding with hydrophobic collapse (Dinner et al. 1999) of β-hairpin1 while indicating that in the two-state and downhill folding scenarios, β-hairpin1 and β-hairpin2 were formed simultaneously by the Matheson-Scheraga mechanism (Matheson and Scheraga 1978).

Chapter 5

Discussion

The common scope of this thesis is the application of structural biology and biophysics to obtain high-resolution, atomic details on the interaction and the folding particularities of nuclear proteins that modulate DNA transcription, either repressing or activating it. The proper regulation and functioning of these elements are essential to ensure organism development, homeostasis and survival. Provided that the biological mechanisms under study in this work are not directly related, the discussion sections exposed hereafter have interpreted the result obtained in each of the project independently.

5.1. The LP Motif as a Potential Binding Determinant in Smad Partners

The Smad transcription factors present an unusually low affinity for DNA. This particularity has been suggested to be key in the capability of Smad signalling to be involved in a myriad of cellular activities and gene transcription programs. This is, in turn, the reason why the pathway is usually affected in proliferative and autoimmune diseases. For its correct activity, a series of proteins involved in gene transcription regulation partner with the active Smad complex in the nucleus, conferring response specificity and an overall higher affinity for DNA. Provided the great variety of cellular processes in which Smad proteins participate, it is not surprising that Smad-binding proteins belong to different groups according to their functionality. These include general and lineage-defining transcription factors, co-repressors, co-activators, chromatin remodelling proteins, gene silencing machinery (DNA methyltransferases) and pathway cross-talk signalling elements. Smad-interacting proteins have been reported to bind either to the MH2 domain or the linker region between domains, with the former being the most commonly-reported binding site.

Despite the importance of Smad signalling and the relatively high number of binding proteins as seen by biochemical assays such as co-precipitation, there is a surprising lack of structural and biophysical data related to these interactions. In the case of the TGF-β-activated Smad2 and Smad3, only the structure of the former with the interacting fragment from membrane anchor SARA has been solved. As a consequence, any new details revealed on how the interactions of Smad2 (and in extension, any other Smad) with its nuclear partners occur at the atomic label and the physico-chemical details governing the interactions will likely have an impact on the basic and applied research on the topic.

To provide a wide perspective on the subject, three Smad-interacting proteins with no connections in the literature beside its reported capability to bind to Smad2 MH2 have been studied. These are the transcription factor FoxH1, the transcriptional coactivator NCOA6 and the E3 ubiquitin-ligase/ chromatin reader TRIM33. This variety of ligands was looked for with the aim to obtain data on the presence of binding hotspots on the surface of the MH2 domain that are complemented by several high-affinity sites that confer specificity.

Unfortunately, the experimental difficulties in the effort to obtain soluble and stable Smad2 MH2 protein translated in a delay in the project that obliged for a re-orientation towards more modest objectives. To achieve this, the cloning and testing of a series of constructs and purification conditions and methods were required. In a first phase of the project, the use of a Smad2 MH2 construct using the practical totality of the inter-domain linker permitted to assess the binding capabilities of the different fragments derived from Smad-binding proteins. A linker-length-dependent stability of the Smad2 MH2 protein was demonstrated as first suspected by the good behaviour of the Smad2 186-467(EEME) construct in comparison with the constructs presenting a shorter N-terminal.

Solid-phase peptide synthesis has been successfully used to obtain short protein fragments of about 20 to 40 residues that had been described or hypothesised to be the main contributors to the interaction of nuclear proteins FoxH1, NCOA6 and TRIM33 with the Smad2 MH2 domain. Since more detailed information regarding was available on FoxH1 and NCOA6 specific sequences, these were studied first.

Surprisingly, the FoxH1 FM peptide is not able to bind to the Smad2 MH2 domain as corroborated by EMSA experiments. This results contradicts previous evidence (Randall et al. 2004). The main difference between the studies is the use of *in-cell* methods in contrast to the biochemical approach used in this thesis. Thus, an explanation for this discrepancy could be a potential requirement for post-transcriptional modifications (PTMs) in either the FoxH1 FM sequence or in Smad2 MH2. The former is more likely, provided that the Smad2 MH2 protein used in EMSA positively interacts with other peptides. New studies specifically addressing the potential of the FoxH1 FM sequence to bind the Smad2 MH2 could include the synthesis of modified FoxH1 FM peptides simulating PTMs using modified or non-natural amino acids.

The positive results in gel shift assays for peptides FoxH1 SIM and NCOA6 LXXLL-2 was a motivation to look for similar sequences in the TRIM33 middle region. By comparative sequence analysis, two patterns within the aforementioned peptides and a segment of TRIM33 middle were identified. These sequences conform the LP motif described in this thesis and proposed to drive the interaction with Smad2 MH2. This motif consists on a leucine-rich part and a double proline

(PP) part separated by a <10-residue gap, found in FoxH1 SIM and NCOA6 LXXLL-2 as a single element and in TRIM33 middle as a tandem sharing the leucine-rich part. Despite the striking sequence similarity, EMSA results indicate that the hypothesis is flawed, and that the presence of an LP motif defines by its own cannot define the capability of a sequence to interact with R-Smad MH2. Yet, the discovery of new Smad-binding sequences will provide for additional data that may lead to a more detailed and specific definition in the future.

On the other hand, binding affinities as determined by ITC are inconclusive regarding the requirement of the full LP motif for optimal MH2 binding. The affinities of three peptides for the Smad2 MH2 domain were analyse, FoxH1 E-LP, FoxH1 D-SIM and NCOA6 LXXLL-2. The data obtained shows that the three interactions present a dissociation constant in the same magnitude range. Instead, had the presence of the full leucine-rich part been fundamental, a higher affinity would be expected when using FoxH1 E-LP.

Given the size of a trimeric Smad2 MH2 - peptide complex, macromolecular crystallography (MX) was the most viable method available to obtain the tri-dimensional structures of the different complexes. Since protein crystallisation is heavily hindered by flexible and unstructured parts and provided the relationship observed between domain stability and linker length in the recombinant MH2 domain, the optimisation of the N-terminal boundary of the Smad2 MH2 was required to improve the possibilities of a successful crystallisation. Two new constructs, the Smad2 231-467 and Smad2 238-467, present a promising balance between both factors. Alternatively, an *in-situ* proteolysis approach was applied, derived from previous studies on the MH2 domain (Chacko et al. 2001). Another problem that required to be tackled was the low water solubility presented by FoxH1 SIM-containing peptides, interfering with the protein-peptide interaction and the crystallisation event by the formation of thick aggregate layers. New, solubility-optimised peptides were obtained by sequence extension to include a higher number of charged residues.

A few conditions yielding microcrystals were detected, but only a small number resulted in mid-sized crystals that could be diffracted. Condition optimisation was not capable to reproduce crystal formation. Unfortunately, most of the crystals obtained were salt crystals and a promising crystal could not be diffracted due to an operational error at the beamline. Since the optimised constructs provide more promising expectations and solubility-improved peptides have already been synthesised, this branch of the project would have benefited from extended research and it will likely be continued by the Macias laboratory in the future.

Even if the whole complex cannot be studied by NMR, the peptides were studied individually. An all α-helix structure was discovered in the FoxH1 D-SIM peptide, with the PP segment defining an accessible loop. Interestingly, the tendency to form any secondary structure was lost in the larger FoxH1 E-LP peptide, a behaviour described for some IDPs.

The presence of secondary structure in an MH2-interacting sequence prompted the search for secondary structure propensity in TRIM33 middle, preferentially α-helixes. Two regions fulfilling the requirements were found and peptides containing the sequences were synthesised. Despite the negative binding results obtained by EMSA using the TRIM33 PH-2, these are not conclusive in the case of peptide TRIM33 PH-1 due to experimental difficulties related to the isoelectric point (pI) presented by the peptide. Future research on the Smad2 MH2 - TRIM33 interaction will entail the use of alternative methods to determine the binding for the TRIM33 PH-1 to the Smad2 MH2. Other possibilities may be addressed by using the full TRIM33 middle region.

The results presented in this thesis determine the sequence regions in transcription factor FoxH1 and co-activator NCOA6 that effectively interact with the Smad2 MH2 domain and points to likely interacting sequences in TRIM33. It has identified, as well, a tendency for an α-helix secondary structure. This thesis has set all the elements required for complex crystallisation with the aim to obtain structural data regarding interaction surfaces on the Smad2 MH2 domain. If specificity-defining sites can be determined, it opens a door for drug design targeting Smad2 specifically, even the blocking of specific interactions. Such specificity is certainly of interest in oncologic treatments. Moreover, it will contribute to the general knowledge on the activities and specific events governing TGF-β signalling.

5.2. Mutation of Residues in the Third β-strand of the FBP28 WW2 Domain Results in Alternative Folding Trajectories

Sequence and solvation environment are the main determinants of the folding mechanism followed by any individual protein. In some cases, for a polypeptide chain to obtain its native form, discrete folding steps, known as folding intermediates, appear towards the final folding stages. Often, transient folds play a role in the definition of subdomain architecture and in-cell folding efficiency rates. Lifetime is one key characteristic of folding intermediates. Interestingly, long intermediate lifetimes have been linked to health problems related to the misfolding of complex proteins, accumulation and aggregation. Well-known misfolding-related conditions include Alzheimer's and prion diseases, which present amyloid accumulations.

The main secondary structure typology of amyloid plaques is anti-parallel β-sheet. The second WW domain of Formin binding protein 28 (FBP28) has been long used as a β-sheet folding model. Moreover, its folding kinetics and its propensity to form amyloid-like fibrils are the reason for an ongoing controversy regarding its folding mechanism. To explain both circumstances, the existence of an uncharacterised intermediate fold was hypothesised, later supported by several molecular dynamics simulation studies.

In a first stage, the collaboration between the Macias Laboratory and the Baker Laboratory of Chemistry and Chemical Biology at Cornell University established to describe the folding mechanisms of FBP28 WW2 addressed the potential role of strand-crossing hydrophobic contacts that define WW domain structure and domain boundaries in the arising of the folding intermediate. The engineering of these positions plus the structural characterisation of the mutant domains and molecular dynamics simulations based on the UNRES force field could not reveal any effect regarding the formation of the intermediate.

An alternative hypothesis to explain the folding mechanisms and the role of the proposed intermediate state in FBP28 WW2 is studied in this thesis. This hypothesis states that the intermediate species appears due to the formation of contacts defining the β-hairpin2 at a slower rate than those defining β-hairpin1 and, specifically, that local hydrophobic contacts in β-turn2 heavily determine turn registry.

To address this hypothesis, seven FBP28 WW2 mutant domains designed to affect the registry of β-turn2 and β-strand3 have been designed and generated by single-point mutagenesis and recombinant expression. Six of these mutants have been successfully obtained, and their structures have been solved by NMR. Unfortunately, despite the use of denaturing purification methods, mutant E456W could not be obtained.

In all cases, the NMR structures have revealed that the domains present the typical triple-stranded WW domain folding with particularities exclusive of each mutant domain. The most interesting mutants from a structural point of view are mutants L455D, L455W and E456Y. In the first, the orientation of the D455 side chain is oriented differently than in the wild-type or any other mutant domain, with the carboxylic acid group in D455 and the hydroxyl in the Y450 phenol ring brought together, indicating the probable existence of a water-mediated bond. Mutant domain L455W, on the other hand, presents the peculiarity of a solvent-exposed W455 side-chain, probably due to steric collisions with Y450 not allowing for hydrophobic shielding with the domain's backbone, yet not heavily affecting the overall domain stability. This is not the case of E456Y, where the inclusion of the bulky Y456 successfully manages to force a rearrangement of surrounding side chains in its side of the domain and even modifies the structure of β-turn2 to fit in. Essential for the stability of this mutant domain is the π-π stacking of Y456 and W436. The rest of structures are more similar to the wild type domain, with some being practically identical to it but for the mutated residue.

Additionally, temperature-induced protein denaturing experiments have been performed and the melting temperatures (T_m) values for each FBP28 WW2 mutant domain, obtained. The determination of the melting data of each mutant domain was considered of interest not only for its later inclusion as a variable in bioinformatics simulations but also due to a previous observation describing an apparent two-state folding behaviour, with no presence of intermediates, of the wild-type domain when studied near its melting temperature. Intrinsic fluorescence spectroscopy and differential scanning calorimetry were the methods selected. Both are generally equivalent methods, yet very differing results have been obtained. Method robustness, equipment quality and T_m value comparison with published FBP28 WW2 wild-type data were considered to favour the Tm values obtained by DSC. All mutant domains were less stable than the wild-type. Surprisingly, those mutants presenting less thermal stability are those affecting residue T458. There is no obvious explanation for this circumstance, at least from the structural data obtained by NMR.

The experimental data presented in this thesis has been applied to the UNRES force-field in the computational work performed by the collaborators led by Dr. Harold A. Scheraga and Dr. Gia G. Maisuradze. Molecular dynamics simulations describe that all FBP28 WW2 mutant domains follow a dominant, common three-state folding trajectory. Nevertheless, two of the six mutant domains, L455D and L455W, present additional folding scenarios independent of an intermediate fold. Despite their minor proportion respect the three-state folding trajectories obtained in the simulations, the observation of two-state and downhill folding trajectories for these mutant domains indicates that it is in fact slow β-turn2 formation, most likely due to hydrophobic contacts between Y450 and L455, that favour the formation of the folding intermediate implicated in fibril formation. The increase in hydrophobic surface in mutant L455W may favour a faster formation of the correct interaction with Y450, while in mutant L455D the role of Van der Waals forces may be substituted by the hydrogen bond that co-ordinates D455 and Y450, accelerating β-turn2 formation.

Later studies on the matter elaborate on these observations by applying all-atom simulation force fields, instead (Kachlishvili et al. 2017). The relevance of the Y450-L455 contacts are supported and a role of residue T454 on the formation of surface to reduce solvent exposure of hydrophobic contacts is added to the equation. On the other hand, the downhill folding trajectories of L455D mutant seen using UNRES disappear in all-atom simulations, although alternative single-kinetics folding trajectories are still observed.

Chapter 6

Conclusions

The results of the research projects performed on protein complex formation and protein folding and dynamics as described in this thesis are resumed in the following conclusions:

Project 1. Characterisation of Protein–Protein Interfaces on the Smad2 MH2 Domain

1. Several Smad2 MH2 domain constructs were designed at different stages of the research. It was discovered that there is a balance regarding linker length and domain solubility. Smad2 MH2 231-467 and Smad2 MH2 238-467 were found to be the most balanced.

2. A series of peptides containing putative Smad-interacting sequences from FoxH1, NCOA6 and TRIM33 were obtained by solid-phase peptide synthesis. Premises for peptide design evolved as new results were obtained.

3. Subtype-specific binding between Smad2 MH2 186-467(EEME) and peptides FoxH1 SIM, FoxH1 LP (and their more soluble variants) and NCOA6 LXXLL-2 can be detected by electrophoretic mobility shift assays

(EMSA). Despite the several approaches, the Smad-interacting fragment in TRIM33 could not be identified.

4. The constant of dissociation (K_d) defining the different Smad2 MH2 - ligand complex kinetics was defined by isothermal titration calorimetry (ITC). All interactions occur in the low-micromolar range, with small differences.

5. The data collected is insufficient to prove existence of a leucine-rich - poli-proline motif (LP motif) as proposed in this thesis.

6. The unrefined tri-dimensional structure of the FoxH1 D-SIM peptide has been described. Interestingly, its all-helix structure is lost in the larger FoxH1 E-LP. Such behaviour has been described in some IDPs.

7. The tri-dimensional structures of the Smad2 MH2 - ligand complexes could not be obtained despite several crystallisation screenings have been attempted. Time was a limiting factor to obtain and successfully diffract crystals with the optimised Smad2 MH2 constructs.

Project 2. Discerning Residues in the FBP28 WW2 Domain Defining Folding Kinetics.

1. From the designed seven single-point mutant FBP28 WW2 domains, six were successfully obtained in quality and concentration that are suitable for structural studies.

2. The tri-dimensional structures of six FPB28 WW2 mutant domains was solved by NMR. In some cases, the addition of DMSO was required to favour one conformation to solve the structure. All domains present the typical WW fold. Particularities appear around the mutated domains, with the most singular being L455D, likely coordinating a water molecule with Y450, L455W, with a solvent-exposed, unrestricted indole group and L456Y, that

manages to fit in the crowded back side of the domain by aromatic stacking with W437.

3. Thermal stability was stablished for six mutant domains and the wild-type using intrinsic fluorescence spectroscopy and differential scanning calorimetry (DSC). The T_m values reveal that all single-point FBP28 WW2 mutants are less thermally stable than the wild-type, with those presenting a structure differing the most from the wild-type being significantly less stable.

4. The application of the experimental data to the UNRES force filed resulted in the identification of alternative folding trajectories for mutants L455D and L455W and the attribution of the intermediate formation to slow formation of the β-hairpin2. Work performed by collaborators.

Abbreviations, Symbols and Units

AcN	Acetonitrile
AMH	Anti-Müllerian hormone
ATP	Adenosine triphosphate
BMP	Bone morphogenic protein
BMRB	Biological magnetic resonance bank
C-terminal	Carboxyl terminal
CD	Circular dichroism
CDK	Cyclin-dependent kinase
ChIP-Seq	Chromatin Immunoprecipitation - Sequencing
DCM	Dichloro methane
DIC	N,N'-Diisopropylcarbodiimide
DLS	Dynamic light scattering
DMF	Dimethyl formamide
DMSO	Dimethyl sulfoxide
DSC	Differential Scanning Calorimetry
EM	Electron microscope
EMSA	Electrophoretic mobility shift assay
EMT	Epithelial-mesenchymal transition
ER	Endoplasmic reticulum
ESC	Embryo stem cells
FEL	Free-energy landscape
FH	Forkhead (domain)
FID	Free induction decay
FITC	Fluorescein isothiocyanate
FM	FoxH1 motif
FT	Fourier transform
GDF	Growth and differentiation factor
H	Enthalpy
$\hbar$	Plank constant over 2π
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HECT	Homologous to the E6-AP Carboxyl Terminal

HMQC	Heteronuclear multiple quantum coherence
HoBT	1-Hydroxybenzotriazole
HPLC	High-performance liquid chromatography
HRV	Human Rhinovirus
HSC	Haematopoietic stem cells
HSQC	Heteronuclear single quantum coherence
I	Spin quantum number
I-Smad	Inhibitory Smad
IDP	Intrinsically disordered protein
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
K_d	Dissociation constant
LB	Luria broth
LC-MS	Liquid chromatography – mass spectrometry
LXR	Liver X receptor
m	Magnetic quantum number
M	Molar
M	Net magnetisation
MALDI-TOF	Matrix-assisted laser desorption/ionization - Time of flight
MD	Molecular dynamics
MET	Mesenchymal-epithelial transition
MH	Mad-homology (domain)
MX	Macromolecular crystallography
N-terminal	Amino terminal
NES	Nuclear export sequence
NLS	Nuclear localisation sequence
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NR	Nuclear receptor
OD	Optical density
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDB	Protein data bank
PEG	Polyethylene glycol
PIP_3	Phosphatidylinositol triphosphate
PS	Polystyrene
PTM	Post-transcriptional modification
Q	Heat
R-Smad	Receptor-activated Smad
RP-HPLC	Reverse-phase HPLC

s	Spin
SARA	Smad anchor for receptor activation
SAXS	Small angle x-ray scattering
SBD	Smad-binding domain
SBE	Smad-binding element
SDS	Sodium dodecyl sulfate
SEC	Size-exclusion chromatography
SIM	Smad-interacting motif
SNP	Single-nucleotide polymorphism
SOC	Super-optimal with repression over catabolite
SPPS	Solid-phase peptide synthesis
TB	Terrific broth
TEV	Tobacco etch virus
TF	Transcription factor
TFA	Trifluoroacetic acid
TGF-β	Transforming growth factor - β
TIS	Triisopropylsilane
T_m	Melting temperature
TOCSY	Total correlation spectroscopy
UPLC	Ultra-performance liquid chromatography
UV	Ultraviolet (light)
WT	Wild-type
γ	Gyromagnetic ratio
δ	Chemical shift
λ	Wavelength
μ	Nuclear magnetic moment
ω	Larmor frequency

References

Agricola E, Randall RA, Gaarenstroom T, Dupont S, Hill CS. 2011. Recruitment of TIF1γ to chromatin via its PHD finger-bromodomain activates its ubiquitin ligase and transcriptional repressor activities. Mol Cell 43: 85–96.

Alarcón C, Zaromytidou A-I, Xi Q, Gao S, Yu J, Fujisawa S, Barlas A, Miller AN, Manova-Todorova K, Macias MJ, et al. 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. Cell 139: 757–769.

Antonson P, Jakobsson T, Almlöf T, Guldevall K, Steffensen KR, Gustafsson J-A. 2008. RAP250 is a coactivator in the transforming growth factor beta signaling pathway that interacts with Smad2 and Smad3. J Biol Chem 283: 8995–9001.

Aragón E, Goerner N, Zaromytidou A-I, Xi Q, Escobedo A, Massagué J, Macias MJ. 2011. A Smad action turnover switch operated by WW domain readers of a phosphoserine code. Genes Dev 25: 1275–1288.

Attisano L, Silvestri C, Izzi L, Labbé E. 2001. The transcriptional role of Smads and FAST (FoxH1) in TGFbeta and activin signalling. Mol Cell Endocrinol 180: 3–11.

BabuRajendran N, Palasingam P, Narasimhan K, Sun W, Prabhakar S, Jauch R, Kolatkar PR. 2010. Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors. Nucleic Acids Res 38: 3477–3488.

Bai X, Kim J, Yang Z, Jurynec MJ, Akie TE, Lee J, LeBlanc J, Sessa A, Jiang H, DiBiase A, et al. 2010. TIF1gamma controls erythroid cell fate by regulating transcription elongation. Cell 142: 133–143.

Balch WE, Morimoto RI, Dillin A, Kelly JW. 2008. Adapting proteostasis for disease intervention. Science 319: 916–919.

Baldwin RL, Rose GD. 2013. Molten globules, entropy-driven conformational change and protein folding. Curr Opin Struct Biol 23: 4–10.

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K. 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6: 1–10.

Bax A, Clore GM, Gronenborn AM. 1990. 1H–1H correlation via isotropic mixing of 13C magnetization, a new three-dimensional approach for assigning 1H and 13C spectra of 13C-enriched proteins. Journal of Magnetic Resonance (1969) 88: 425–431.

Bax A, Davis DG. 1985. MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. Journal of Magnetic Resonance (1969) 65: 355–360.

Bodenhausen G, Ruben DJ. 1980. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69: 185–189.

Bourgeois B, Gilquin B, Tellier-Lebègue C, Östlund C, Wu W, Pérez J, Hage P El, Lallemand F, Worman HJ, Zinn-Justin S. 2013. Inhibition of TGF-β signaling at the nuclear envelope: characterization of interactions between MAN1, Smad2 and Smad3, and PPM1A. Sci Signal 6: ra49.

Brennan RG. 1993. The winged-helix DNA-binding motif: another helix-turn-helix takeoff. Cell 74: 773–776.

Broome BM, Hecht MH. 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. J Mol Biol 296: 961–968.

Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54: 905–921.

Buchanan LE, Dunkelberger EB, Tran HQ, Cheng P-N, Chiu C-C, Cao P, Raleigh DP, de Pablo JJ, Nowick JS, Zanni MT. 2013. Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient β-sheet. Proc Natl Acad Sci U S A 110: 19285–19290.

Caira F, Antonson P, Pelto-Huikko M, Treuter E, Gustafsson JA. 2000. Cloning and characterization of RAP250, a novel nuclear receptor coactivator. J Biol Chem 275: 5308–5317.

Carpino LA, El-Faham A. 1994. Effect of Tertiary Bases on O-Benzotriazolyluronium Salt-Induced Peptide Segment Coupling. J Org Chem 59: 695–698.

Carpino LA, Han GY. 1972. 9-Fluorenylmethoxycarbonyl amino-protecting group. J Org Chem 37: 3404–3409.

Chacko BM, Qin B, Correia JJ, Lam SS, de Caestecker MP, Lin K. 2001. The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization. Nat Struct Biol 8: 248–253.

Chacko BM, Qin BY, Tiwari A, Shi G, Lam S, Hayward LJ, De Caestecker M, Lin K. 2004. Structural basis of heteromeric smad protein assembly in TGF-beta signaling. Mol Cell 15: 813–823.

Chan WC, White PD. 2000. Fmoc solid phase peptide synthesis: A practical approach. Oxford University Press, New York.

Chen G, Nomura M, Morinaga H, Matsubara E, Okabe T, Goto K, Yanase T, Zheng H, Lu J, Nawata H. 2005. Modulation of androgen receptor transactivation by FoxH1. A newly identified androgen receptor corepressor. J Biol Chem 280: 36355–36363.

Chen W, Dijke P Ten. 2016. Immunoregulation by members of the TGFβ superfamily. Nat Rev Immunol 16: 723–740.

Chen X, Weisberg E, Fridmacher V, Watanabe M, Naco G, Whitman M. 1997. Smad4 and FAST-1 in the assembly of activin-responsive factor. Nature 389: 85–89.

Cheung MS, García AE, Onuchic JN. 2002. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. Proc Natl Acad Sci U S A 99: 685–690.

Cleary JP, Walsh DM, Hofmeister JJ, Shankar GM, Kuskowski MA, Selkoe DJ, Ashe KH. 2005. Natural oligomers of the amyloid-beta protein specifically disrupt cognitive function. Nat Neurosci 8: 79–84.

Davis CM, Dyer RB. 2014. WW domain folding complexity revealed by infrared spectroscopy. Biochemistry 53: 5476–5484.

de Caestecker M. 2004. The transforming growth factor-beta superfamily of receptors. Cytokine Growth Factor Rev 15: 1–11.

Derynck R, Zhang YE. 2003. Smad-dependent and Smad-independent pathways in TGF-beta family signalling. Nature 425: 577–584.

Dinner AR, Lazaridis T, Karplus M. 1999. Understanding beta-hairpin formation. Proc Natl Acad Sci U S A 96: 9068–9073.

Dong A, Xu X, Edwards AM, Midwest Center for Structural Genomics, Structural Genomics Consortium, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, et al. 2007. In situ proteolysis for protein crystallization and structure determination. Nat Methods 4: 1019–1021.

Dryland A, Sheppard RC. 1986. Peptide synthesis. Part 8. A system for solid-phase synthesis under low pressure continuous flow conditions. Journal of the Chemical Society, Perkin Transactions 1 125.

Dupont S, Zacchigna L, Cordenonsi M, Soligo S, Adorno M, Rugge M, Piccolo S. 2005. Germ-layer specification and control of cell growth by Ectodermin, a Smad4 ubiquitin ligase. Cell 121: 87–99.

Eanes ED, Glenner GG. 1968. X-ray diffraction studies on amyloid filaments. J Histochem Cytochem 16: 673–677.

Eisenberg D, Jucker M. 2012. The amyloid state of proteins in human diseases. Cell 148: 1188–1203.

Ericsson UB, Hallberg BM, Detitta GT, Dekker N, Nordlund P. 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal Biochem 357: 289–298.

Fändrich M, Forge V, Buder K, Kittler M, Dobson CM, Diekmann S. 2003. Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. Proc Natl Acad Sci U S A 100: 15463–15468.

Feige MJ, Groscurth S, Marcinowski M, Yew ZT, Truffault V, Paci E, Kessler H, Buchner J. 2008. The structure of a folding intermediate provides insight into differences in immunoglobulin amyloidogenicity. Proc Natl Acad Sci U S A 105: 13373–13378.

Ferguson N, Berriman J, Petrovich M, Sharpe TD, Finch JT, Fersht AR. 2003. Rapid amyloid fiber formation from the fast-folding WW domain FBP28. Proc Natl Acad Sci U S A 100: 9814–9819.

Ferguson N, Johnson CM, Macias M, Oschkinat H, Fersht A. 2001. Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. Proc Natl Acad Sci U S A 98: 13002–13007.

Ferreira JC, Marcondes MF, Icimoto MY, Cardoso THS, Tofanello A, Pessoto FS, Miranda EGA, Prieto T, Nascimento OR, Oliveira V, et al. 2015. Intermediate Tyrosyl Radical and Amyloid Structure in Peroxide-Activated Cytoglobin. PLoS ONE 10: e0136554.

Fersht AR. 2008. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. Nat Rev Mol Cell Biol 9: 650–654.

Fesik SW, Zuiderweg ER. 1988. Heteronuclear three-dimensional nmr spectroscopy. A strategy for the simplification of homonuclear two-dimensional NMR spectra. Journal of Magnetic Resonance (1969) 78: 588–593.

Fischer R, Mader O, Jung G, Brock R. 2003. Extending the applicability of carboxyfluorescein in solid-phase synthesis. Bioconjug Chem 14: 653–660.

Freire E. 1995. Thermal denaturation methods in the study of protein folding. Meth Enzymol 259: 144–168.

Froimchuk E, Jang Y, Ge K. 2017. Histone H3 lysine 4 methyltransferase KMT2D. Gene 627: 337–342.

Fuentealba LC, Eivers E, Ikeda A, Hurtado C, Kuroda H, Pera EM, De Robertis EM. 2007. Integrating patterning signals: Wnt/GSK3 regulates the duration of the BMP/Smad1 signal. Cell 131: 980–993.

Gao S, Alarcón C, Sapkota G, Rahman S, Chen P-Y, Goerner N, Macias MJ, Erdjument-Bromage H, Tempst P, Massagué J. 2009. Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. Mol Cell 36: 457–468.

Germain S, Howell M, Esslemont GM, Hill CS. 2000. Homeodomain and winged-helix transcription factors recruit activated Smads to distinct promoter elements via a common Smad interaction motif. Genes Dev 14: 435–451.

Goldstrohm AC, Albrecht TR, Suñé C, Bedford MT, Garcia-Blanco MA. 2001. The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. Mol Cell Biol 21: 7617–7628.

Gorrec F, Palmer CM, Lebon G, Warne T. 2011. Pi sampling: a methodical and flexible approach to initial macromolecular crystallization screening. Acta Crystallogr D Biol Crystallogr 67: 463–470.

Grzesiek S, Bax A. 1992a. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. Journal of Magnetic Resonance (1969) 99: 201–207.

Grzesiek S, Bax A. 1992b. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. J Am Chem Soc 114: 6291–6293.

Hartl FU, Bracher A, Hayer-Hartl M. 2011. Molecular chaperones in protein folding and proteostasis. Nature 475: 324–332.

He W, Dorn DC, Erdjument-Bromage H, Tempst P, Moore MAS, Massagué J. 2006. Hematopoiesis controlled by distinct TIF1gamma and Smad4 branches of the TGFbeta pathway. Cell 125: 929–941.

Hendrickson W. 1985. Protein-DNA interactions studied by the gel electrophoresis-DNA binding assay. BioTechniques. https://uic.pure.elsevier.com/en/publications/protein-dna-interactions-studied-by-the-gel-electrophoresis-dna-b.

Hermanson O, Glass CK, Rosenfeld MG. 2002. Nuclear receptor coregulators: multiple modes of modification. Trends Endocrinol Metab 13: 55–60.

Hesling C, Lopez J, Fattet L, Gonzalo P, Treilleux I, Blanchard D, Losson R, Goffin V, Pigat N, Puisieux A, et al. 2013. Tif1γ is essential for the terminal differentiation of mammary alveolar epithelial cells and for lactation through SMAD4 inhibition. Development 140: 167–175.

Higgins SJ, Hames BD. 1999. Protein expression: A practical approach. Oxford University Press, Oxford.

Hill CS. 2016. Transcriptional control by the smads. Cold Spring Harb Perspect Biol 8.

Hofmann A. 2010. Spectroscopic techniques: I Spectrophotometric techniques. In Principles and Techniques of Biochemistry and Molecular Biology (eds. K. Wilson and J. Walker), pp. 477–521, Cambridge University Press.

Holmes C, Boche D, Wilkinson D, Yadegarfar G, Hopkins V, Bayer A, Jones RW, Bullock R, Love S, Neal JW, et al. 2008. Long-term effects of Abeta42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial. The Lancet 372: 216–223.

Horwich A. 2002. Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions. J Clin Invest 110: 1221–1232.

Ikushima H, Miyazono K. 2010. TGFbeta signalling: a complex web in cancer progression. Nat Rev Cancer 10: 415–424.

Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CSH, Yu J, Hersi K, Raaijmakers J, Gish G, Mbamalu G, et al. 2005. WW domains provide a platform for the assembly of multiprotein networks. Mol Cell Biol 25: 7092–7106.

Ingólfsson HI, Lopez CA, Uusitalo JJ, de Jong DH, Gopal SM, Periole X, Marrink SJ. 2014. The power of coarse graining in biomolecular simulations. Wiley Interdiscip Rev Comput Mol Sci 4: 225–248.

Jäger M, Nguyen H, Crane JC, Kelly JW, Gruebele M. 2001. The folding mechanism of a beta-sheet: the WW domain. J Mol Biol 311: 373–393.

Jahn TR, Parker MJ, Homans SW, Radford SE. 2006. Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. Nat Struct Mol Biol 13: 195–201.

Jancarik J, Kim SH. 1991. Sparse matrix sampling: a screening method for crystallization of proteins. J Appl Crystallogr 24: 409–411.

Jeener J, Meier BH, Bachmann P, Ernst RR. 1979. Investigation of exchange processes by two-dimensional NMR spectroscopy. J Chem Phys 71: 4546.

Johns H. 2013. Fiscal Year 2014 Appropriations for Alzheimer's-related Activities at the U.S. Department of Health and Human Services. Alzheimer's Association, Washington, DC.

Ju B-G, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG. 2006. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. Science 312: 1798–1802.

Kachlishvili K, Dave K, Gruebele M, Scheraga HA, Maisuradze GG. 2017. Eliminating a protein folding intermediate by tuning a local hydrophobic contact. *J Phys Chem B* **121**: 3276–3284.

Kaiser E, Colescott RL, Bossinger CD, Cook PI. 1970. Color test for detection of free terminal amino groups in the solid-phase synthesis of peptides. Anal Biochem 34: 595–598.

Karanicolas J, Brooks CL. 2003. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? Proc Natl Acad Sci U S A 100: 3954–3959.

Kato Y, Ito M, Kawai K, Nagata K, Tanokura M. 2002. Determinants of ligand specificity in groups I and IV WW domains as studied by surface plasmon resonance and model building. J Biol Chem 277: 10173–10177.

Kato Y, Nagata K, Takahashi M, Lian L, Herrero JJ, Sudol M, Tanokura M. 2004. Common mechanism of ligand recognition by group II/III WW domains: redefining their functional classification. J Biol Chem 279: 31833–31841.

Kavsak P, Rasmussen RK, Causing CG, Bonni S, Zhu H, Thomsen GH, Wrana JL. 2000. Smad7 binds to Smurf2 to form an E3 ubiquitin ligase that targets the TGF beta receptor for degradation. Mol Cell 6: 1365–1375.

Kendrew JC. 1962. Nobel Lecture: Myoglobin and the Structure of Proteins. In Nobel Lectures, Chemistry 1942-1962, Elsevier Publishing Company, Amsterdam.

Kent SB. 1988. Chemical synthesis of peptides and proteins. Annu Rev Biochem 57: 957–989.

Kim S-W, Park K, Kwak E, Choi E, Lee S, Ham J, Kang H, Kim JM, Hwang SY, Kong Y-Y, et al. 2003. Activating signal cointegrator 2 required for liver lipid metabolism mediated by liver X receptors in mice. Mol Cell Biol 23: 3583–3592.

Ko L, Cardona GR, Chin WW. 2000. Thyroid hormone receptor-binding protein, an LXXLL motif-containing protein, functions as a general coactivator. Proc Natl Acad Sci U S A 97: 6212–6217.

Kofron M, Puck H, Standley H, Wylie C, Old R, Whitman M, Heasman J. 2004. New roles for FoxH1 in patterning the early embryo. Development 131: 5065–5078.

Kondé E, Bourgeois B, Tellier-Lebegue C, Wu W, Pérez J, Caputo S, Attanda W, Gasparini S, Charbonnier J-B, Gilquin B, et al. 2010. Structural analysis of the Smad2-MAN1 interaction that regulates transforming growth factor-β signaling at the inner nuclear membrane. Biochemistry 49: 8020–8032.

Kusanagi K, Inoue H, Ishidou Y, Mishima HK, Kawabata M, Miyazono K. 2000. Characterization of a bone morphogenetic protein-responsive Smad-binding element. Mol Biol Cell 11: 555–565.

Labbé E, Silvestri C, Hoodless PA, Wrana JL, Attisano L. 1998. Smad2 and Smad3 positively and negatively regulate TGF beta-dependent transcription through the forkhead DNA-binding protein FAST2. Mol Cell 2: 109–120.

Ladbury JE, Chowdhry BZ. 1996. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. Chem Biol 3: 791–801.

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8: 477–486.

Lee J, Culyba EK, Powers ET, Kelly JW. 2011. Amyloid-β forms fibrils by nucleated conformational conversion of oligomers. Nat Chem Biol 7: 602–609.

Lee S-H, Kim D-H, J. Han J, Cha E-J, Lim J-E, Cho Y-J, Lee C, Han K-H. 2012. Understanding Pre-Structured Motifs (PreSMos) in Intrinsically Unfolded Proteins. Curr Protein Pept Sci 13: 34–54.

Lee SK, Jung SY, Kim YS, Na SY, Lee YC, Lee JW. 2001. Two distinct nuclear receptor-interaction domains and CREB-binding protein-dependent transactivation function of activating signal cointegrator-2. Mol Endocrinol 15: 241–254.

Leichsenring M, Maes J, Mössner R, Driever W, Onichtchouk D. 2013. Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science 341: 1005–1009.

Leopold PE, Montal M, Onuchic JN. 1992. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci U S A 89: 8721–8725.

Liu F, Du D, Fuller AA, Davoren JE, Wipf P, Kelly JW, Gruebele M. 2008. An experimental survey of the transition between two-state and downhill protein folding scenarios. Proc Natl Acad Sci U S A 105: 2369–2374.

Liwo A, Khalili M, Scheraga HA. 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci U S A 102: 2362–2367.

Luo K. 2004. Ski and SnoN: negative regulators of TGF-beta signaling. Curr Opin Genet Dev 14: 65–70.

Macias MJ, Gervais V, Civera C, Oschkinat H. 2000. Structural analysis of WW domains and design of a WW prototype. Nat Struct Biol 7: 375–379.

Macias MJ, 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: 646–649.

Macias MJ, Martin-Malpartida P, Massagué J. 2015. Structural determinants of Smad function in TGF-β signaling. Trends Biochem Sci 40: 296–308.

Macias MJ, Wiesner S, Sudol M. 2002. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett 513: 30–37.

Maisuradze GG, Liwo A, Senet P, Scheraga HA. 2013. Local vs global motions in protein folding. J Chem Theory Comput 9: 2907–2921.

Malik S, Roeder RG. 2005. Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends Biochem Sci 30: 256–263.

Marion D, Driscoll PC, Kay LE, Wingfield PT, Bax A, Gronenborn AM, Clore GM. 1989a. Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-nitrogen-15 Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1.beta. Biochemistry 28: 6150–6156.

Marion D, Kay LE, Sparks SW, Torchia DA, Bax A. 1989b. Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins. J Am Chem Soc 111: 1515–1517.

Martin-Malpartida P, Batet M, Kaczmarska Z, Freier R, Gomes T, Aragon E, Zou Y, Wang Q, Xi Q, Ruiz L, et al. 2017. Structural basis for genome-wide recognition of 5-bp GC motifs by Smad transcription factors.

Massagué J, Seoane J, Wotton D. 2005. Smad transcription factors. Genes Dev 19: 2783–2810.

Massagué J. 2008. TGFbeta in Cancer. Cell 134: 215–230.

Matheson RR, Scheraga HA. 1978. A method for predicting nucleation sites for protein folding based on hydrophobic contacts. Macromolecules 11: 819–829.

Matsuura I, Denissova NG, Wang G, He D, Long J, Liu F. 2004. Cyclin-dependent kinases regulate the antiproliferative function of Smads. Nature 430: 226–231.

Mayor U, Guydosh NR, Johnson CM, Grossmann JG, Sato S, Jas GS, Freund SMV, Alonso DOV, Daggett V, Fersht AR. 2003. The complete folding pathway of a protein from nanoseconds to microseconds. Nature 421: 863–867.

Merrifield RB. 1965. Solid-Phase Peptide Syntheses. Endeavour 24: 3–7.

Morikawa M, Koinuma D, Miyazono K, Heldin CH. 2013. Genome-wide mechanisms of Smad binding. Oncogene 32: 1609–1615.

Morikawa M, Koinuma D, Tsutsumi S, Vasilaki E, Kanki Y, Heldin C-H, Aburatani H, Miyazono K. 2011. ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif. Nucleic Acids Res 39: 8712–8727.

Mu Y, Nordenskiöld L, Tam JP. 2006. Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. Biophys J 90: 3983–3992.

Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. 2011. Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell 147: 565–576.

Mullis KB. 1997. Nobel Lecture: The Polymerase Chain Reaction. In Nobel Lectures, Chemistry 1991-1995, World Scientific Publishing Co., Singapore.

Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, Grothe R, Eisenberg D. 2005. Structure of the cross-beta spine of amyloid-like fibrils. Nature 435: 773–778.

Neudecker P, Robustelli P, Cavalli A, Walsh P, Lundström P, Zarrine-Afsar A, Sharpe S, Vendruscolo M, Kay LE. 2012. Structure of an intermediate state in protein folding and aggregation. Science 336: 362–366.

Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW. 2003. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. Proc Natl Acad Sci U S A 100: 3948–3953.

Nilsson J, Ståhl S, Lundeberg J, Uhlén M, Nygren PA. 1997. Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. Protein Expr Purif 11: 1–16.

O'Neill MJ. 1964. The Analysis of a Temperature-Controlled Scanning Calorimeter. Anal Chem 36: 1238–1245.

Olofsson A, Ippel HJ, Baranov V, Hörstedt P, Wijmenga S, Lundgren E. 2001. Capture of a dimeric intermediate during transthyretin amyloid formation. J Biol Chem 276: 39592–39599.

Olzscha H, Schermann SM, Woerner AC, Pinkert S, Hecht MH, Tartaglia GG, Vendruscolo M, Hayer-Hartl M, Hartl FU, Vabulas RM. 2011. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. Cell 144: 67–78.

Park S-H, Raines RT. 2004. Fluorescence gel retardation assay to detect protein-protein interactions. Methods Mol Biol 261: 155–160.

Park YW, Wilusz J, Katze MG. 1999. Regulation of eukaryotic protein synthesis: selective influenza viral mRNA translation is mediated by the cellular RNA-binding protein GRSF-1. Proc Natl Acad Sci U S A 96: 6694–6699.

Pepys MB. 2006. Amyloidosis. Annu Rev Med 57: 223–241.

Peroutka Iii RJ, Orcutt SJ, Strickler JE, Butt TR. 2011. SUMO fusion technology for enhanced protein expression and purification in prokaryotes and eukaryotes. Methods Mol Biol 705: 15–30.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.

Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, et al. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29: 845–854.

Quéré R, Saint-Paul L, Carmignac V, Martin RZ, Chrétien M-L, Largeot A, Hammann A, Pais de Barros J-P, Bastie J-N, Delva L. 2014. Tif1γ regulates the TGF-β1 receptor and promotes physiological aging of hematopoietic stem cells. Proc Natl Acad Sci U S A 111: 10592–10597.

Ramirez-Espain X, Ruiz L, Martin-Malpartida P, Oschkinat H, Macias MJ. 2007. Structural characterization of a new binding motif and a novel binding mode in group 2 WW domains. J Mol Biol 373: 1255–1268.

Randall RA, Germain S, Inman GJ, Bates PA, Hill CS. 2002. Different Smad2 partners bind a common hydrophobic pocket in Smad2 via a defined proline-rich motif. EMBO J 21: 145–156.

Randall RA, Howell M, Page CS, Daly A, Bates PA, Hill CS. 2004. Recognition of phosphorylated-Smad2-containing complexes by a novel Smad interaction motif. Mol Cell Biol 24: 1106–1121.

Ranganathan R, Lu KP, Hunter T, Noel JP. 1997. Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. Cell 89: 875–886.

Rhodes G. 2006. Crystallography made crystal clear. 3rd ed. Elsevier/Academic Press, Amsterdam.

Roberts CWM, Orkin SH. 2004. The SWI/SNF complex--chromatin and cancer. Nat Rev Cancer 4: 133–142.

Salomon-Ferrer R, Case DA, Walker RC. 2013. An overview of the Amber biomolecular simulation package. Wiley Interdisciplinary Reviews: Computational Molecular Science 3: 198–210.

Sarantakis D, Teichman J, Lien EL, Fenichel RL. 1976. A novel cyclic undecapeptide, Wy-40,770, with prolonged growth hormone release inhibiting activity. Biochem Biophys Res Commun 73: 336–342.

Schiene C, Fischer G. 2000. Enzymes that catalyse the restructuring of proteins. Curr Opin Struct Biol 10: 40–45.

Shi W, Chang C, Nie S, Xie S, Wan M, Cao X. 2007. Endofin acts as a Smad anchor for receptor activation in BMP signaling. J Cell Sci 120: 1216–1224.

Shi Y, Hata A, Lo RS, Massagué J, Pavletich NP. 1997. A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature 388: 87–93.

Shi Y, Massagué J. 2003. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. Cell 113: 685–700.

Shi Y, Wang YF, Jayaraman L, Yang H, Massagué J, Pavletich NP. 1998. Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. Cell 94: 585–594.

Simonsson M, Kanduri M, Grönroos E, Heldin C-H, Ericsson J. 2006. The DNA binding activities of Smad2 and Smad3 are regulated by coactivator-mediated acetylation. J Biol Chem 281: 39870–39880.

Sipe JD, Benson MD, Buxbaum JN, Ikeda S-I, Merlini G, Saraiva MJM, Westermark P. 2016. Amyloid fibril proteins and amyloidosis: chemical identification and clinical classification International Society of Amyloidosis 2016 Nomenclature Guidelines. Amyloid 23: 209–213.

Slagle CE, Aoki T, Burdine RD. 2011. Nodal-dependent mesendoderm specification requires the combinatorial activities of FoxH1 and Eomesodermin. PLoS Genet 7: e1002072.

Sole NA, Barany G. 1992. Optimization of solid-phase synthesis of [Ala8]-dynorphin A. J Org Chem 57: 5399–5403.

Steffensen KR, Gustafsson J-Å. 2006. Liver X receptors: new drug targets to treat Type 2 diabetes? Future Lipidol 1: 181–189.

Sudol M, Sliwa K, Russo T. 2001. Functions of WW domains in the nucleus. FEBS Lett 490: 190–195.

Tsukazaki T, Chiang TA, Davison AF, Attisano L, Wrana JL. 1998. SARA, a FYVE Domain Protein that Recruits Smad2 to the TGFβ Receptor. Cell 95: 779–791.

Vanier GS. 2013. Microwave-assisted solid-phase peptide synthesis based on the Fmoc protecting group strategy (CEM). Methods Mol Biol 1047: 235–249.

Vivian JT, Callis PR. 2001. Mechanisms of tryptophan fluorescence shifts in proteins. Biophys J 80: 2093–2109.

Vojkovsky T. 1995. Detection of secondary amines on solid phase. Pept Res 8: 236–237.

Watanabe M, Whitman M. 1999. FAST-1 is a key maternal effector of mesoderm inducers in the early Xenopus embryo. Development 126: 5621–5634.

Weisberg E, Winnier GE, Chen X, Farnsworth CL, Hogan BL, Whitman M. 1998. A mouse homologue of FAST-1 transduces TGF beta superfamily signals and is expressed during early embryogenesis. Mech Dev 79: 17–27.

Wernimont A, Edwards A. 2009. In situ proteolysis to generate crystals for structure determination: an update. PLoS ONE 4: e5094.

Wolynes PG, Onuchic JN, Thirumalai D. 1995. Navigating the folding routes. Science 267: 1619–1620.

Wotton D, Lo RS, Lee S, Massagué J. 1999. A Smad transcriptional corepressor. Cell 97: 29–39.

Wrana JL, Attisano L, Wieser R, Ventura F, Massagué J. 1994. Mechanism of activation of the TGF-beta receptor. Nature 370: 341–347.

Wu G, Chen YG, Ozdamar B, Gyuricza CA, Chong PA, Wrana JL, Massagué J, Shi Y. 2000. Structural basis of Smad2 recognition by the Smad anchor for receptor activation. Science 287: 92–97.

Wu JW, Hu M, Chai J, Seoane J, Huse M, Li C, Rigotti DJ, Kyin S, Muir TW, Fairman R, et al. 2001b. Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling. Mol Cell 8: 1277–1289.

Wu JW, Krawitz AR, Chai J, Li W, Zhang F, Luo K, Shi Y. 2002. Structural mechanism of Smad4 recognition by the nuclear oncoprotein Ski: insights on Ski-mediated repression of TGF-beta signaling. Cell 111: 357–367.

Xi Q, Wang Z, Zaromytidou A-I, Zhang XH-F, Chow-Tsang L-F, Liu JX, Kim H, Barlas A, Manova-Todorova K, Kaartinen V, et al. 2011. A poised chromatin platform for TGF-β access to master regulators. Cell 147: 1511–1524.

Xu J, Huang L, Shakhnovich EI. 2011. The ensemble folding kinetics of the FBP28 WW domain revealed by an all-atom Monte Carlo simulation in a knowledge-based potential. Proteins 79: 1704–1714.

Xu J, Lamouille S, Derynck R. 2009. TGF-beta-induced epithelial to mesenchymal transition. Cell Res 19: 156–172.

Xu L, Kang Y, Çöl S, Massagué J. 2002. Smad2 nucleocytoplasmic shuttling by nucleoporins can/nup214 and nup153 feeds tgfβ signaling complexes in the cytoplasm and nucleus. Molecular cell 10: 271–282.

Yamada S, Bouley Ford ND, Keller GE, Ford WC, Gray HB, Winkler JR. 2013. Snapshots of a protein folding intermediate. Proc Natl Acad Sci U S A 110: 1606–1610.

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. J Biol Chem 291: 382–392.

Yang M, Lei M, Huo S. 2003. Why is Leu55-->Pro55 transthyretin variant the most amyloidogenic: insights from molecular dynamics simulations of transthyretin monomers. Protein Sci 12: 1222–1231.

Yeo CY, Chen X, Whitman M. 1999. The role of FAST-1 and Smads in transcriptional regulation by activin during early Xenopus embryogenesis. J Biol Chem 274: 26584–26590.

Yoon J, Lee KJ, Oh G-S, Kim GH, Kim S-W. 2017. Regulation of Nampt expression by transcriptional coactivator NCOA6 in pancreatic β-cells. Biochem Biophys Res Commun 487: 600–606.

Young RA. 2011. Control of the embryonic stem cell state. Cell 144: 940–954.

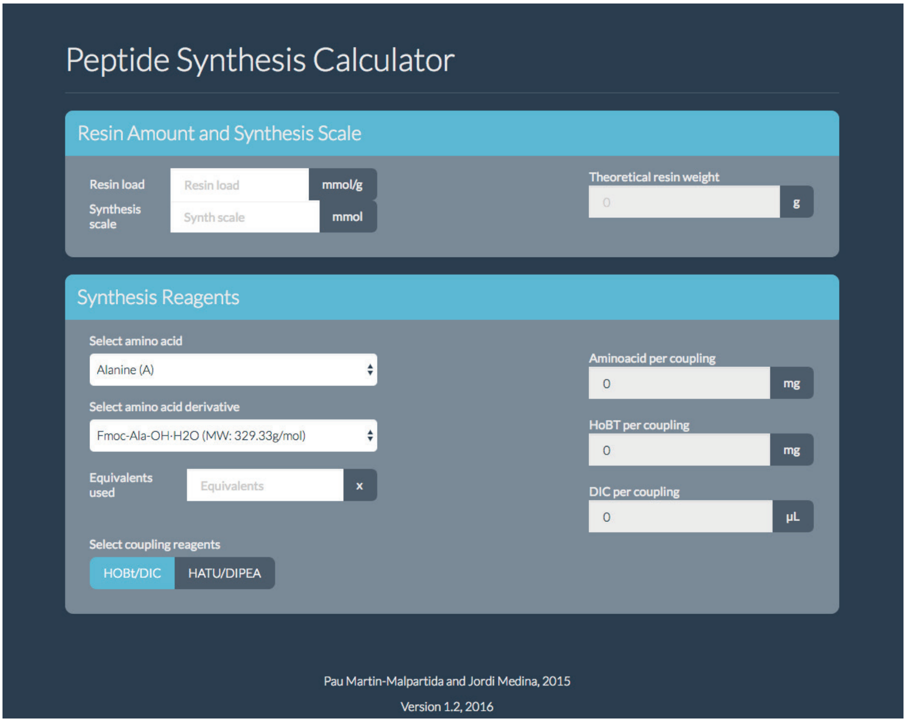
Zhou R, Maisuradze GG, Suñol D, Todorovski T, Macias MJ, Xiao Y, Scheraga HA, Czaplewski C, Liwo A. 2014. Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements. Proc Natl Acad Sci U S A 111: 18243–18248.

Zhou R, Maisuradze GG, Suñol D, Todorovski T, Macias MJ, Xiao Y, Scheraga HA, Czaplewski C, Liwo A. 2014. Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements. Proc Natl Acad Sci U S A 111: 18243–18248.

Zhou S, Zawel L, Lengauer C, Kinzler KW, Vogelstein B. 1998. Characterization of Human FAST-1, a TGFβ and Activin Signal Transducer. Molecular cell 2: 121–127. André J-M. 2014. The nobel prize in chemistry 2013. Chemistry International 36.

Zhu Y, Qi C, Cao WQ, Yeldandi AV, Rao MS, Reddy JK. 2001. Cloning and characterization of PIMT, a protein with a methyltransferase domain, which interacts with and enhances nuclear receptor coactivator PRIP function. Proc Natl Acad Sci U S A 98: 10380–10385.

Supplementary Data and Appended Documents



Peptide Synthesis Calculator

Resin Amount and Synthesis Scale

Resin load: Resin load mmol/g

Synthesis scale: Synth scale mmol

Theoretical resin weight: 0 g

Synthesis Reagents

Select amino acid: Alanine (A)

Select amino acid derivative: Fmoc-Ala-OH·H2O (MW: 329.33g/mol)

Equivalents used: Equivalents x

Select coupling reagents: HOBt/DIC | HATU/DIPEA

Aminoacid per coupling: 0 mg

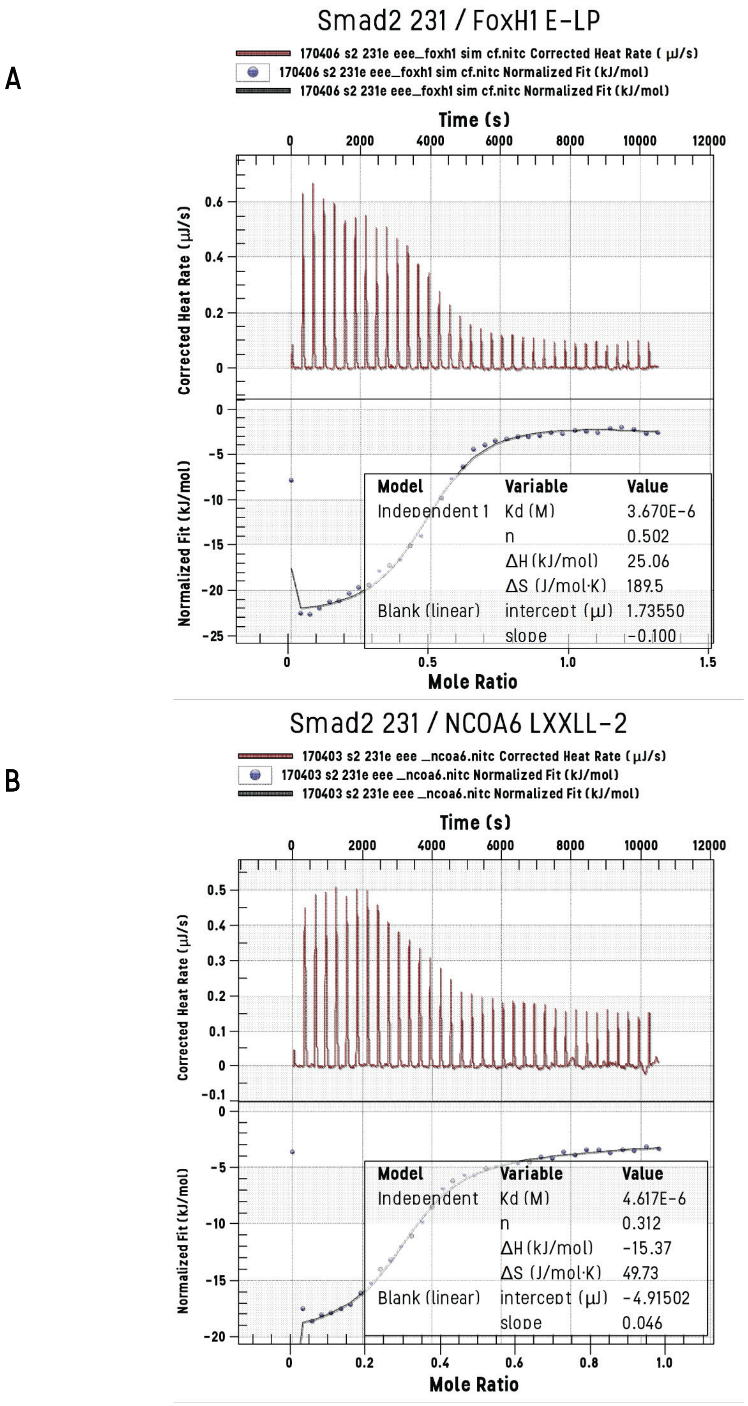
HoBT per coupling: 0 mg

DIC per coupling: 0 µL

Pau Martin-Malpartida and Jordi Medina, 2015

Version 1.2, 2016

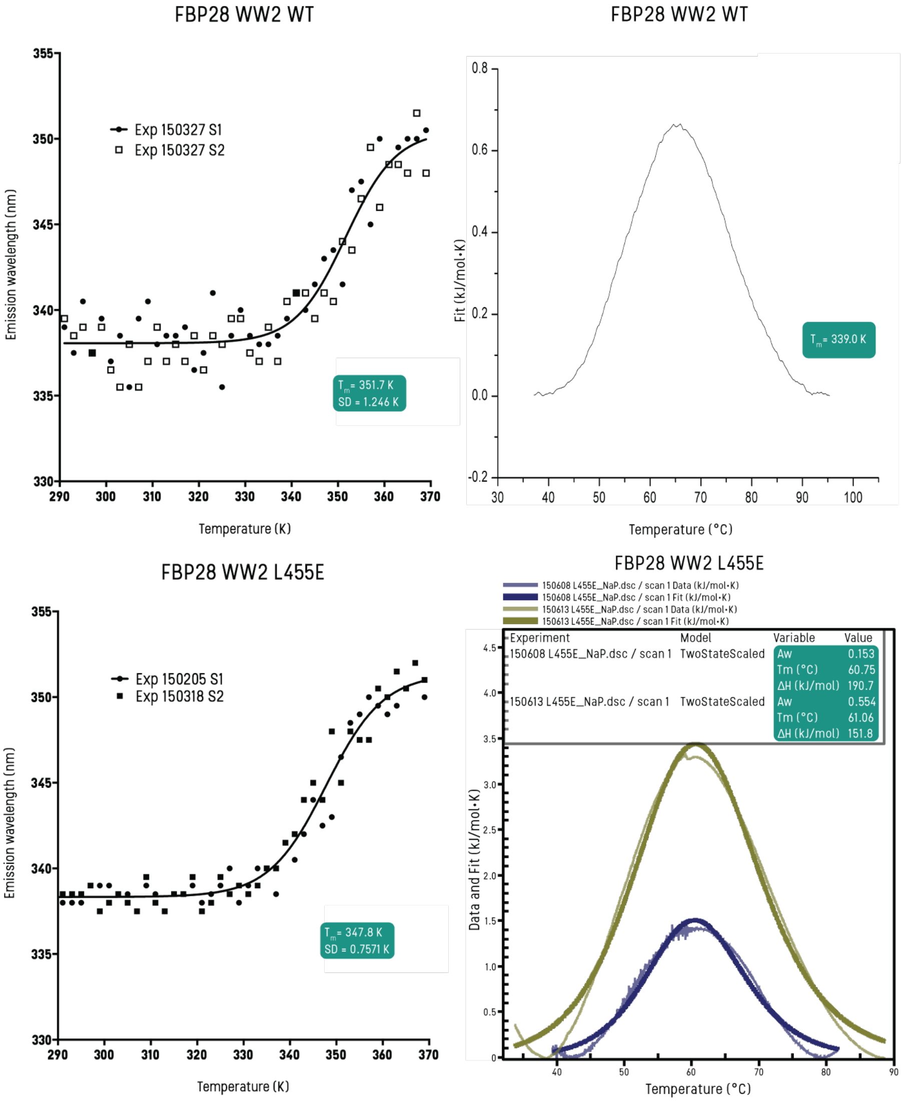
Figure S1. User interface of the web application for peptide synthesis assistance



Model	Variable	Value
Independent 1	Kd (M)	3.670E-6
	n	0.502
	ΔH (kJ/mol)	25.06
	ΔS (J/mol·K)	189.5
Blank (linear)	intercept (μJ)	1.73550
	slope	-0.100

Model	Variable	Value
Independent	Kd (M)	4.617E-6
	n	0.312
	ΔH (kJ/mol)	-15.37
	ΔS (J/mol·K)	49.73
Blank (linear)	intercept (μJ)	-4.91502
	slope	0.046

Figure S2. ITC data of FoxH1 E-LP and NCOA6 LXXLL-2 – Smad2 MH2 interactions



Experiment	Model	Variable	Value
150608 L455E_NaP.dsc / scan 1	TwoStateScaled	Aw	0.153
		Tm (°C)	60.75
		ΔH (kJ/mol)	190.7
150613 L455E_NaP.dsc / scan 1	TwoStateScaled	Aw	0.554
		Tm (°C)	61.06
		ΔH (kJ/mol)	151.8

Figure S3. Melting temperatures of the FBP28 WW2 wild-type and L455E mutant

On the left, intrinsic fluorescence curves of the maxim emission wavelength. On the right, heat exchange DSC curves.

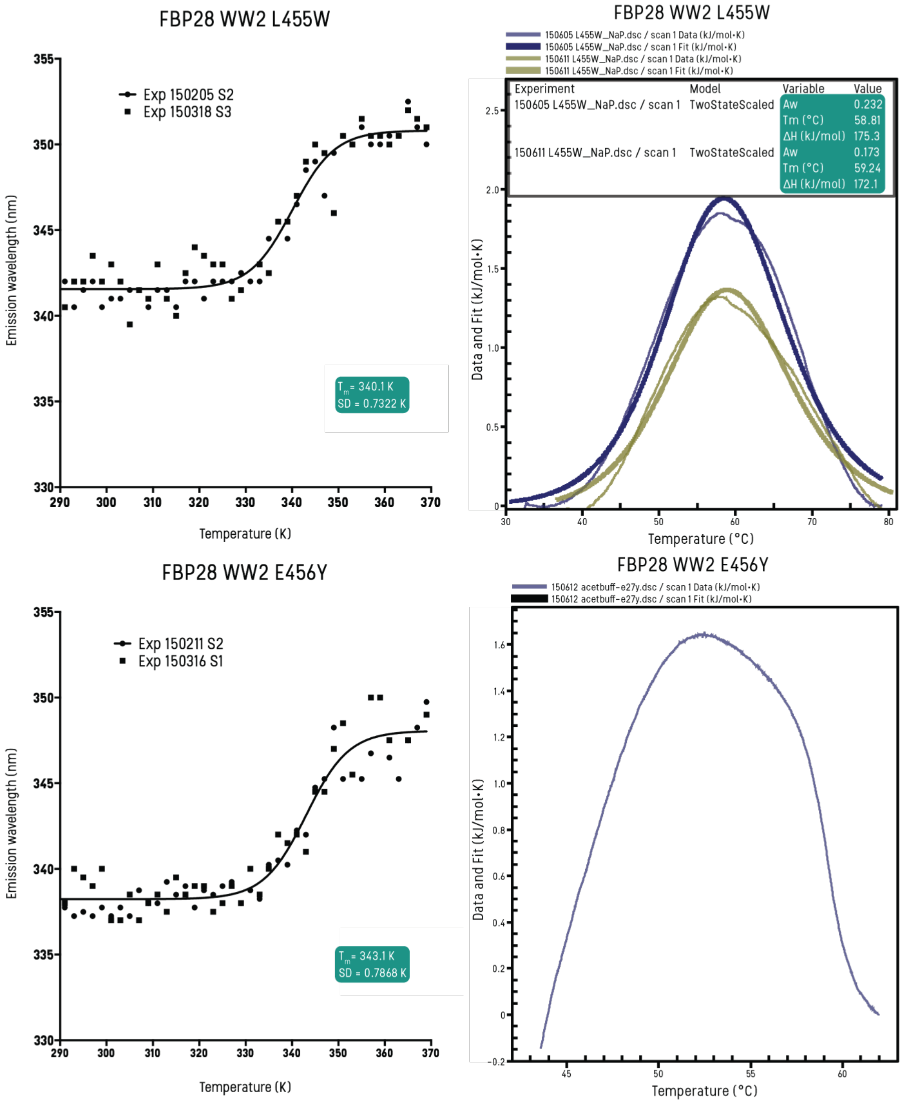
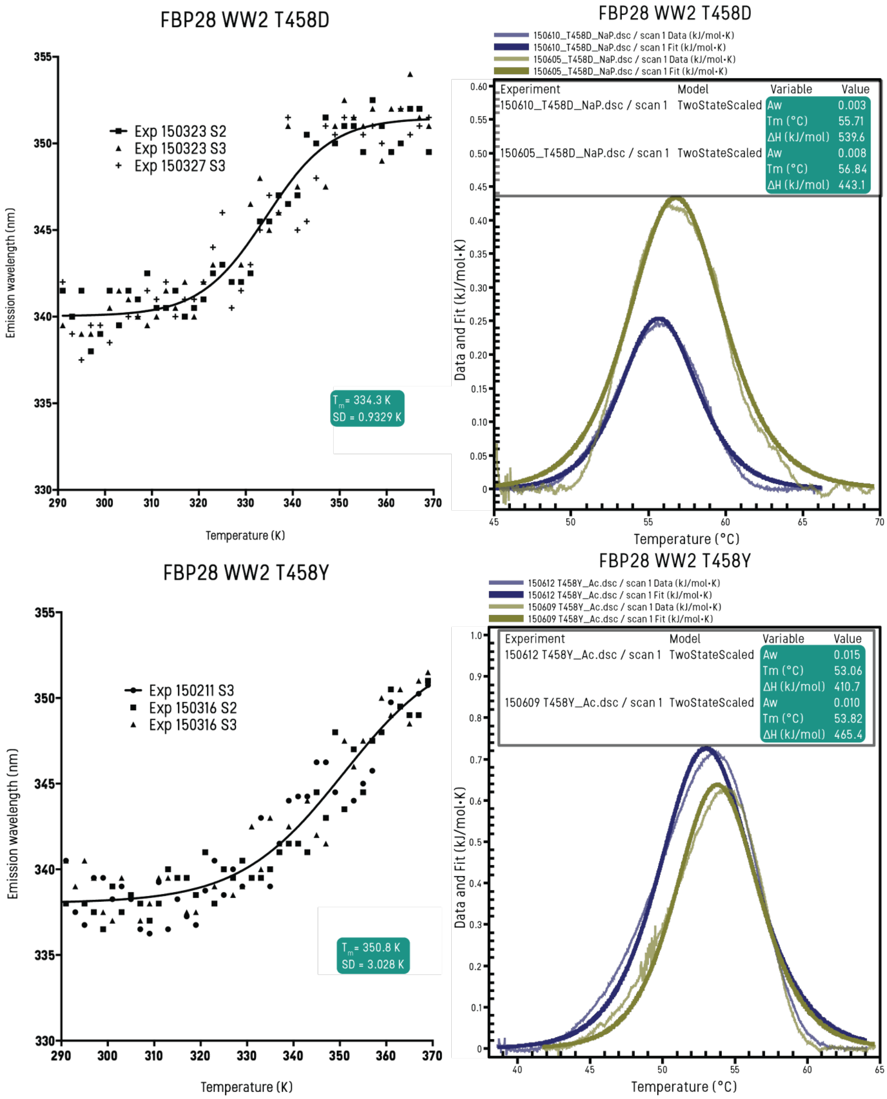


Figure S4. Melting temperatures of the FBP28 WW2 L455W and E456Y mutants

On the left, intrinsic fluorescence curves of the maxim emission wavelength. On the right, heat exchange DSC curves.



Experiment	Model	Variable	Value
150610_T458D_NaP.dsc / scan 1	TwoStateScaled	Aw	0.003
		Tm (°C)	55.71
		ΔH (kJ/mol)	539.6
150605_T458D_NaP.dsc / scan 1	TwoStateScaled	Aw	0.008
		Tm (°C)	56.84
		ΔH (kJ/mol)	443.1

Experiment	Model	Variable	Value
150612 T458Y_Ac.dsc / scan 1	TwoStateScaled	Aw	0.015
		Tm (°C)	53.06
		ΔH (kJ/mol)	410.7
150609 T458Y_Ac.dsc / scan 1	TwoStateScaled	Aw	0.010
		Tm (°C)	53.82
		ΔH (kJ/mol)	465.4

Figure S5. Melting temperatures of the FBP28 WW2 T458D and T458Y mutants

On the left, intrinsic fluorescence curves of the maxim emission wavelength. On the right, heat exchange DSC curves.

Preventing fibril formation of a protein by selective mutation

Gia G. Maisuradze[a,1], Jordi Medina[b], Khatuna Kachlishvili[a], Pawel Krupa[a,c], Magdalena A. Mozolewska[a,c], Pau Martin-Malpartida[b], Luka Maisuradze[a], Maria J. Macias[b,d,1], and Harold A. Scheraga[a,1]

[a]Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301; [b]Structural Characterization of Macromolecular Assemblies, Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, 08028, Spain; [c]Laboratory of Molecular Modeling, Faculty of Chemistry, University of Gdańsk, 80-308 Gdańsk, Poland; and [d]Catalan Institution for Research and Advanced Studies, Barcelona, 08010, Spain

Contributed by Harold A. Scheraga, September 15, 2015 (sent for review July 13, 2015; reviewed by Ulrich H. E. Hansmann and Gregory A. Petsko)

The origins of formation of an intermediate state involved in amyloid formation and ways to prevent it are illustrated with the example of the Formin binding protein 28 (FBP28) WW domain, which folds with biphasic kinetics. Molecular dynamics of protein folding trajectories are used to examine local and global motions and the time dependence of formation of contacts between C^αs and C^βs of selected pairs of residues. Focus is placed on the WT FBP28 WW domain and its six mutants (L26D, L26E, L26W, E27Y, T29D, and T29Y), which have structures that are determined by high-resolution NMR spectroscopy. The origins of formation of an intermediate state are elucidated, viz. as formation of hairpin 1 by a hydrophobic collapse mechanism causing significant delay of formation of both hairpins, especially hairpin 2, which facilitates the emergence of an intermediate state. It seems that three-state folding is a major folding scenario for all six mutants and WT. Additionally, two-state and downhill folding scenarios were identified in ~15% of the folding trajectories for L26D and L26W, in which both hairpins are formed by the Matheson–Scheraga mechanism much faster than in three-state folding. These results indicate that formation of hairpins connecting two antiparallel β-strands determines overall folding. The correlations between the local and global motions identified for all folding trajectories lead to the identification of the residues making the main contributions in the formation of the intermediate state. The presented findings may provide an understanding of protein folding intermediates in general and lead to a procedure for their prevention.

fibril formation | selective mutation | FBP28 WW domain | millisecond-timescale MD simulations | high-resolution NMR spectroscopy

An intermediate state in protein folding is involved in amyloid fibril formation, which is responsible for a number of neurodegenerative diseases (1–7). Therefore, prevention of the aggregation of folding intermediates is one of the most important problems to surmount. Hence, it is necessary to determine the mechanism by which an intermediate state is formed. For example, one of the members of the WW domain family (8, 9), the triple β-stranded WW domain from the Formin binding protein 28 (FBP28; Protein Data Bank ID code 1E0L) (10) (Fig. 1*N*), has been shown to fold with biphasic kinetics exhibiting intermediates during folding (3, 5, 6, 11–16). We address this problem here with the design of new FBP28 WW domain mutants and by examining their structural properties and folding kinetics.

Because of the small size, fast folding kinetics, and biological importance, the formation of intermolecular β-sheets is thought to be a crucial event in the initiation and propagation of amyloid diseases, such as Alzheimer's disease, and spongiform encephalopathy, FBP28, and other WW domain proteins (e.g., Pin1 and FiP35) have been the subjects of extensive experimental (4, 11, 17–23) and theoretical (3, 5, 6, 12–16, 24–27) studies. However, a folding mechanism of the FBP28 was debatable for a long time because of its complexity. There are not only discrepancies between experimental and theoretical results but also, different experiments that reveal different folding scenarios.

In particular, Nguyen et al. (11) studied the folding kinetics of the WT FBP28 and its full-size and truncated mutants by temperature denaturation and laser temperature–jump relaxation experiments. Nguyen et al. (11) found that the folding of the WT FBP28 involves intermediates (three-state folding) below the melting temperature and that the strand-crossing hydrophobic cluster of Tyr11, Tyr19, and Trp30 residues, which were mutated, is not a likely origin of the three-state scenario; also, truncation at the C terminus and an increase of temperature can modulate the two- and three-state folding behavior. The conclusion regarding three-state folding was challenged by Ferguson et al. (4), who observed single-exponential folding kinetics for the FBP28 by using fluorescence measurements and concluded that the biphasic kinetics observed by Nguyen et al. (11) might be related to aggregation and rapidly forming ribbon-like fibrils at physiological temperature and pH, with morphology typical of amyloid fibrils.

Our recent theoretical studies (12–16) of the same systems (11) showed that (*i*) folding of all of these systems involves intermediates; (*ii*) the strand-crossing hydrophobic cluster of residues 11, 19, and 30 is not associated with biphasic kinetics; and (*iii*) neither an increase of temperature nor truncation can alter the folding scenario. Moreover, discrepancies between experimental and theoretical results for some of these mutants caused by experimental limitations were clarified (16).

It also was found (3, 5) that the WT FBP28 folds with biphasic kinetics attributed to independence in the slow formation of turn 2

Significance

Protein folding intermediates are associated with formation of amyloid fibrils, which are responsible for a number of neurodegenerative disorders, such as Alzheimer's, Parkinson's, Huntington's, and Creutzfeldt-Jakob's diseases. Alzheimer's disease is one of the largest problems nowadays, with a significant rise in the number and cost (more than $1 trillion a year by 2050 will be spent in the United States alone unless action is taken to shift current trends; www.alz.org). Therefore, elucidation of the origins of formation of intermediates and finding ways to prevent them are very important. The presented findings may lead to an understanding of the structural mechanisms by which intermediates initiate fibril aggregation.

Author contributions: G.G.M. designed research; G.G.M., J.M., K.K., P.K., M.A.M., P.M.-M., and M.J.M. performed research; G.G.M., K.K., P.K., M.A.M., L.M., M.J.M., and H.A.S. analyzed data; and G.G.M., M.J.M., and H.A.S. wrote the paper.

Reviewers: U.H.E.H., University of Oklahoma; and G.A.P., Weill Cornell Medical College.

The authors declare no conflict of interest.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.rcsb.org [PDB ID codes 2n4r (L26D), 2n4s (L26E), 2n4t (L26W), 2n4u (E27Y), 2n4v (T29D), and 2n4w (T29Y)], and the NMR assignments have been deposited in the BioMagResBank, www.bmrb.wisc.edu [BMRB ID codes 25678 (L26D), 25679 (L26E), 25680 (L26W), 25681 (E27Y), 25682 (T29D), and 25683 (T29Y)].

[1]To whom correspondence may be addressed. Email: gm56@cornell.edu, maria.macias@irbbarcelona.org, or has5@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518298112/-/DCSupplemental.

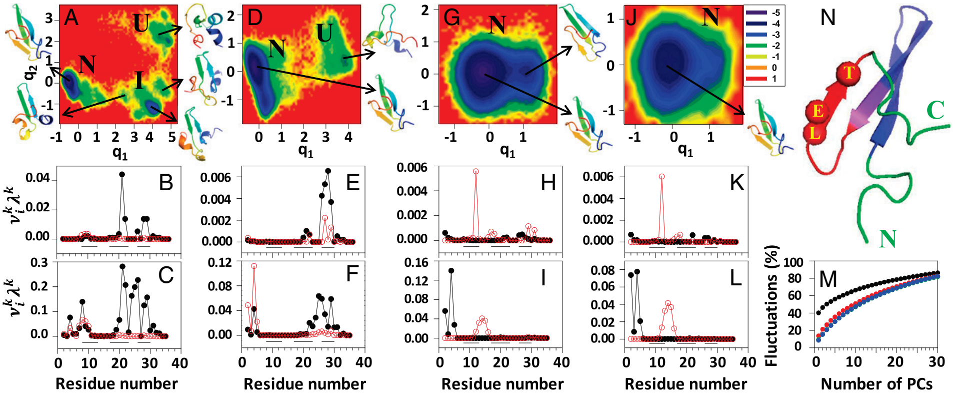


Fig. 1. FELs (kilocalories per mole) along the first two PCs with representative structures at the minima, and contributions of the principal modes (defined in *SI Materials and Methods*) [$\nu_i^k\lambda^k$; black lines with black circles (principal mode 1) and red lines with white circles (principal mode 2)] to the MSFs along the θ- and γ-angles for the (*A*–*C*) three-state, (*D*–*F*) two-state, and (*G*–*I*) downhill folding trajectories of L26D and (*J*–*L*) the downhill folding trajectory of L26W. The black lines on the bottoms of *B*, *C*, *E*, *F*, *H*, *I*, *K*, and *L* correspond to the β-strand regions. I, intermediate; N, native; U, unfolded. *M* represents percentages of the total fluctuations captured by the PCs for three-state (black line), two-state (red line), and downhill (blue line) trajectories of L26D and the downhill folding trajectory (green line; indistinguishable from the blue line) of L26W. *N* represents the experimental structure of FBP28, in which the mutated residues are represented by spheres, and hairpins 1 and 2 are represented by blue and red, respectively (the purple region corresponds to the overlap of these hairpins). C, C terminus; E, glutamic acid; L, leucine; N, N terminus; T, threonine.

contacts with respect to the remainder of the protein and identified a key surface-exposed hydrophobic contact (Tyr21 with Leu26) for enforcing the correct registry of the residues of turn 2. To show the importance of the surface-exposed hydrophobic contact (Tyr21 with Leu26) and the involvement of turn 2 in a slow formation phase, the L26A mutant was studied (3). The fast phase (formation of hairpin 1) was not affected by this mutation, whereas the slow phase became even slower, which also was confirmed experimentally (11). These results suggested that the replacement of leucine by alanine actually stabilizes the misregistered turn 2 conformations relative to the WT; hence, it was concluded (3) that the surface-exposed hydrophobic contact (Tyr21 with Leu26) might be responsible for tying down turn 2 with a correctly formed hairpin. It should be noted that this surface-exposed hydrophobic contact is not present in other members of the WW domain family, which fold with monophasic kinetics.

Later theoretical studies (6, 12–16) of the WT FBP28 confirmed the results of ref. 3, showing biphasic folding kinetics with a stable intermediate state. Therefore, to prevent the formation of the intermediate state, it is logical to make mutations in the region of turn 2 and the third β-strand to speed up the formation of hairpin 2 as implemented here. However, based on the results of mutant L26A (3), it is not an easy task to ensure the elimination of intermediates and therefore, requires a detailed understanding of folding/misfolding mechanisms, folding/misfolding pathways, and effect of temperature on folding mechanism, etc., to identify proper sites for mutations.

Based on previous studies (3, 15), Leu26 is one of the main residues in which mutation might speed up the correct registry of turn 2. Moreover, the FBP28 is the only WW domain among 200 WW domain sequences that contains leucine at this position (3). Usually, this position is almost always occupied by a charged residue or glycine; therefore, following the natural tendency of the WW domain family, two mutants were designed, replacing leucine 26 with negatively charged polar amino acids: aspartic acid and glutamic acid (L26D and L26E, respectively). Also, replacement of leucine by alanine (the smallest nonpolar aliphatic amino acid) was found to slow down the process (3, 11); hence, for replacement of leucine 26, we also selected a very nonpolar and larger aromatic amino acid, tryptophan (L26W). It should be noted that leucine at position 26 is not a reflection of negative design by evolution but rather, is a result of pressure to maximize specificity through use of polar residues (3). Based on earlier results on the binding affinity of the WW domain, it was proposed (3) that requirements for ligand specificity have led to a local sequence with a strong propensity for a misregistered turn.

The next mutant was made by substituting a negatively charged polar amino acid, glutamic acid 27, with a nonpolar aromatic amino acid, tyrosine (E27Y). Finally, two more mutants were designed by replacing a neutral polar amino acid, threonine 29, with a negatively charged polar amino acid, aspartic acid (T29D) and a nonpolar aromatic amino acid, tyrosine (T29Y). Both Glu27 and Thr29 are critically placed residues contributing the most to the mean-square fluctuations (MSFs) (15), and mutation of these residues by disfavored amino acids might destabilize the misregistered turn 2 and β-strand 3 and speed up the correct registry.

To characterize the effects of these mutations, the six recombinant proteins carrying a single-point mutation were expressed, and their structures were studied by high-resolution NMR spectroscopy (*SI Materials and Methods*). All mutants adopt the triple-stranded antiparallel β-sheet characteristic of the WW fold, with slight variations caused by each specific mutation (Fig. S1). The experimental and theoretical melting temperatures (T_m values) for each mutant were determined with differential scanning calorimetry and multiplexed replica exchange molecular dynamics (MD) simulations, respectively (Table S1). We also ran simulations consisting of 120 (for WT and L26D) and 96 (for L26E, L26W, E27Y, T29D, and T29Y) canonical MD trajectories generated with the coarse-grained united residue (UNRES) force field (*SI Materials and Methods*) (28–30) at five and four different temperatures, respectively (24 MD trajectories, with ~1.4 μs formal time and effectively ~1.4 ms of each at each temperature), which were below, very close to, and above (for some mutants) the melting temperatures. The folding dynamics of each system were analyzed in terms of principal component analysis (PCA) (*SI Materials and Methods*) (12, 15, 31) describing the global motions of the protein, local motions of each residue [free-energy profiles (FEPs) along the amino acid sequence], and distances between the C^{α}s and C^{β}s of selected pairs of residues forming hairpins 1 and 2 over time.

Results

We first determined the foldability of the mutant domains in silico. It appeared that all of these mutants can fold, and the percentages of folding trajectories increase with temperature at most times (Fig. S2). The mutants of Leu26 appeared as better foldable systems (the foldability of L26D and L26E was even higher than that of the WT) than the mutants of Thr29 and especially, Glu27, for which the number of nonfolding trajectories exceeded the number of folding trajectories. We interpret this as the replacements of Glu27 and Thr29 by Tyr with a large aromatic side chain may force a reorganization of surrounding side chains and make folding more difficult (details are in *SI Materials and Methods*).

All systems were examined to determine the number of pathways through which they can fold. By calculating the rmsd of the first and second hairpins for all mutants and WT with respect to the native state, we determined that all systems can fold mainly through two different folding pathways: (*i*) hairpin 1 forms first, and then, the rest of the protein folds (dominant pathway with 51–100% occurrence); (*ii*) hairpin 2 forms first, and then, the rest of the protein folds (minor pathway with 0–49% occurrence). However, two mutants, L26D and L26W, and WT can, in addition, fold through a third folding pathway, in which both hairpins form simultaneously. The emergence of the third folding pathway at temperatures lower than the melting points for L26D (13%) and L26W (20%) and the melting temperatures for WT (4%) and L26W (5%) is a first indication that these systems may fold without intermediates; however, detailed analyses of these trajectories are required to validate this observation. Also, the frequency of occurrence of any type of folding pathway does not depend on temperature (details are in Table S2). Our findings regarding two folding pathways and the dominance of the first folding pathway for WT are in agreement with the recent work by Xu et al. (32).

PCA. Free-energy landscapes (FELs) provide an understanding as to how proteins fold and function (33–35). It is impossible to present an FEL as a function of all degrees of freedom of a protein. Consequently, we have to rely on the coordinates along which the intrinsic folding pathways can be viewed. The folding dynamics of the WT and all mutants are investigated here by constructing FELs along the principal components (PCs) obtained from PCA, which typically capture most of the total displacement from the average protein structure with the first few PCs during a simulation (12–15, 36).

As was expected, a dominant folding scenario for all studied systems is a three-state folding (i.e., WT and all mutants fold through an intermediate state). An illustrative FEL along the first two PCs, $\mu(q_1,q_2) = -k_BT\ln P(q_1,q_2)$, of one of the mutants, L26D at 305 K, is plotted in Fig. 1*A*, in which three states (unfolded, intermediate, and native) can be identified. The representative structures of the states indicate that the L26D mutant in this particular trajectory folds through the first, most dominant pathway. Similar FELs but with different representative structures of intermediate states, in which the second hairpin is formed, are characteristic for the MD trajectories, in which the systems fold through the second type pathway (not shown).

The analyses of the MD trajectories, in which the WT and mutants fold through the third type of pathway (both hairpins form simultaneously), show that only two mutants, L26D and L26W, exhibit two-state and downhill folding scenarios. In particular, L26D at 315 K can undergo both a two-state (Fig. 1*D*) and downhill folding (Fig. 1*G*), and L26W at 310 K can fold through the downhill folding scenario (Fig. 1*J*). The representative structure of the unfolded state in Fig. 1*D* (two-state folding) is not a typical unfolded structure. It is a mixture of representative structures of intermediate and unfolded states with partially formed hairpin 1. In the FEL, in which L26D undergoes downhill folding (Fig. 1*G*), two minima can be identified, with native and native-like representative structures; however, both minima are located in the native basin, and the barrier between them is <0.3 kcal/mol. Therefore, we consider it as downhill folding. The rest of the trajectories of WT and L26W, in which both hairpins formed simultaneously, did not exhibit two-state or downhill folding, because the rmsd (unlike PCA) was unable to capture a subtle behavior of one of the hairpins inducing an intermediate state (13).

Apart from the FELs, we have calculated the contributions of the two main principal modes [solid lines with filled (principal mode 1) and empty (principal mode 2) circles in Fig. 1] to the MSFs along the θ- (Fig. 1 *B*, *E*, *H*, and *K*) and γ-angles (Fig. 1 *C*, *F*, *I*, and *L*) and the percentages of the total fluctuations captured by the PCs (Fig. 1*M*) for the three-state, two-state, and downhill folding trajectories.

The main contributions to the fluctuations in a three-state folding trajectory come from all three β-strands and the second turn (Fig. 1 *B* and *C*). These results are in agreement with our earlier results on WT FBP28 (15). The main contributions to the fluctuations in a two-state folding trajectory come from the N terminus, the third β-strand, and the second turn (Fig. 1 *E* and *F*). There is no contribution from the first β-strand and the first turn, and there are minimal contributions from the middle β-strand, which indicate that the largest part of hairpin 1 forms very fast. These results explain why the representative structure of the unfolded state in a two-state folding trajectory (Fig. 1*D*) differs from a typical unfolded structure. Contributions to the fluctuations in downhill folding trajectories of L26D (Fig. 1 *H* and *I*) and L26W (Fig. 1 *K* and *L*) are almost identical (i.e., they come from the N terminus, the first β-strand, and the first turn). There is no contribution from the second turn, and there are some minor contributions from the third β-strand, the main "players" in the emergence of the intermediate state, which explains why L26D and L26W fold through the downhill folding scenario (i.e., without an intermediate state).

The percentage of total fluctuations captured by the first PC in the three-state folding trajectory (Fig. 1*M*, black line) is ∼40% [the same results were obtained for WT FBP28 (13)], whereas the first PCs in two-state and downhill folding trajectories capture only ∼14% and 9% (Fig. 1*M*, red, blue, and green lines) of total fluctuations, respectively. We have shown previously that the FEL constructed along PCs can describe the folding dynamics correctly if these PCs can capture at least 40% of the total fluctuations (13). Therefore, here, we examined the two-state and downhill folding trajectories in 7D and 8D PC spaces, respectively; however, we could not find any new major basins. Hence, 2D FELs are sufficient in these folding trajectories. A large difference in the percentages of the captured fluctuations between three-state folding trajectories and two-state and downhill folding trajectories can be explained by the fact that the largest contribution to the fluctuations in two-state and downhill folding trajectories comes from the very flexible part of the protein, the N terminus; also, PCA has proven to be an effective tool for the analysis of protein folding trajectories involving concerted motions of many residues, which can be captured by a few PCs with the largest eigenvalues (15). Interestingly, the distribution of the percentages of the total fluctuations captured by the PCs obtained for the B domain of staphylococcal protein A (13), a three-helical bundle, which folds through the two-state or downhill folding scenario, is similar to those for two-state and downhill folding trajectories of L26D and L26W.

Because the L26D and L26W mutants are the only ones exhibiting (with a small percentage) a folding scenario other than three state, we ran an additional 500 MD trajectories for each mutant to eliminate the possibility that two-state or downhill folding was an accidental folding scenario. Indeed, after examining 500 MD trajectories of both mutants, we found that, in ∼15% of all folding trajectories, these mutants fold through either two-state or downhill folding. The rest of this paper will consider only these two mutants.

Before scrutinizing the folding mechanisms of L26D and L26W, we examined the structures of all mutants determined by high-resolution NMR spectroscopy (Fig. S1 and Table S3) to find out the structural basis for the aforementioned theoretical findings.

The NMR data corresponding to L26D reveal the presence of a well-folded domain. However, no NOEs were identified from the D26 side chain to the surrounding residues. The calculated structures revealed that the D26 side chain is consistently oriented toward the Y21 hydroxyl (Fig. S1*A*). These results suggest the presence of a water-mediated hydrogen bond that stabilizes that specific orientation, which may allow some "flexibility" during the correct registry of turn 2. In other words, it may either speed up (two-state or downhill folding) or slow down (three-state folding) the correct registry of turn 2 in contrast to the WT, in which surface-exposed hydrophobic contact enforces the slow correct registry of turn 2.

In L26E and L26W, the E26 and W26 substitutions presented contacts more similar to the WT, with the orientation of their side chains in the calculated structures resembling that of L26 (Fig. S1 *B* and *C*). In L26W, we observed contacts between the indole of W26 and the aromatic ring of Y21, but the calculated structures result in several orientations of the W26 ring that are compatible with the experimental restraints and do not affect the turn structure (Fig. S1*C*); hence, their correlation with different folding scenarios is not straightforward. Structural properties of the rest of the mutants are provided in *SI Materials and Methods*.

Mechanisms of Hairpin Folding. To elucidate the origins of a significant time difference between formation of the first and second hairpins, which is the cause of the induction of the intermediate state, we focus on the folding mechanisms of each hairpin. In particular, we examined the behavior of the distances between the C^{α}s of selected residues, pertaining to the first and second and the second and third β-strands (Fig. S3), and also, the behavior of the distances between the C^{α}s and C^{β}s of nonpolar residues, pertaining to (*i*) the solvent-exposed hydrophobic cluster (Tyr11, Tyr19, and Trp30), (*ii*) the delocalized hydrophobic core (Trp8, Tyr20, and Pro33), (*iii*) the surface-exposed hydrophobic contact (Tyr21 and Leu26), and (*iv*) the contact (Ala14 and Gly16) over time. It should be noted that the surface-exposed hydrophobic contact is formed by a different pair (Tyr21 and Trp26) in L26W and does not exist in L26D. We also calculated the time when the distances between the C^{α}s of each selected pair of residues reach (or are very close to) the experimental distance for the first time; we will designate it below as "the first contact time" (black circles connected by black lines in Fig. 2) and the time when the distance between each selected pair of residues gets stabilized (i.e., does not undergo significant changes after that) (red circles connected by red lines in Fig. 2). The same first contact and stabilization time was calculated for the C^{β}s of nonpolar residues (white circles connected by dashed lines in Fig. 2).

By comparing the first contact and stabilization times of the three-state (Fig. 2 *A* and *B*), two-state (Fig. 2 *C* and *D*), and downhill (Fig. 2 *E* and *F*) folding trajectories, we had the following findings. (*i*) A first contact between C^{α}s and C^{β}s of most of the selected pairs of residues occurs within a short time ($\leq 10^{-3}$ μs) for all trajectories, whereas a stabilization time in the three-state folding trajectory is greater by about one and two orders of magnitude than that in the two-state and downhill trajectories, respectively, which finally causes the emergence of an intermediate state. (*ii*) There is a correlation between the first contact time and the location of pairs of residues pertaining to the first hairpin in both downhill folding trajectories (Fig. 2 *E* and *G*) and partially in the two-state folding trajectory (Fig. 2*C*). In other words, the first contact time depends on how far a pair of residues is located from the turn and increases with the distance between the turn and the location of pairs of residues. A similar correlation is observed between the first contact time and the location of pairs of residues, pertaining to the second hairpin, in the downhill folding trajectory of L26W (Fig. 2*G*). There is no such correlation in a three-state folding trajectory (Fig. 2*A*). (*iii*) The stabilization time for all selected pairs in the downhill folding trajectories (Fig. 2 *E* and *G*) and the two-state trajectory (Fig. 2*C*) either increases with the distance between the turn and the location of pairs of residues or stays constant. (*iv*) The first contact time between C^{α}s and C^{β}s and the location of pairs of nonpolar residues, pertaining to the second hairpin, are correlated in all trajectories (Fig. 2 *B*, *D*, *F*, and *H*), whereas the pairs of nonpolar residues, pertaining to the first hairpin, are correlated for both C^{α}s and C^{β}s the downhill folding trajectories (Fig. 2 *F* and *H*). In three- and two-state folding trajectories, the first contact time between C^{α}s and C^{β}s and location of pairs of nonpolar residues, pertaining to the first hairpin, are correlated only for the C^{β} distances. For some pairs of nonpolar residues, the first contact time between C^{α}s does not follow the one between C^{β}s. (*v*) The stabilization time for the pairs of nonpolar residues in the downhill folding trajectories and the two-state trajectory behaves the same way as described in *iii*.

The foregoing results indicate that the hairpins of L26D and L26W fold through two different mechanisms. In particular, both hairpins in both downhill folding trajectories fold through the mechanism proposed by Matheson and Scheraga (37), which is based on transient hydrophobic interactions and considers the nucleation process as an initial aspect of folding, converting an

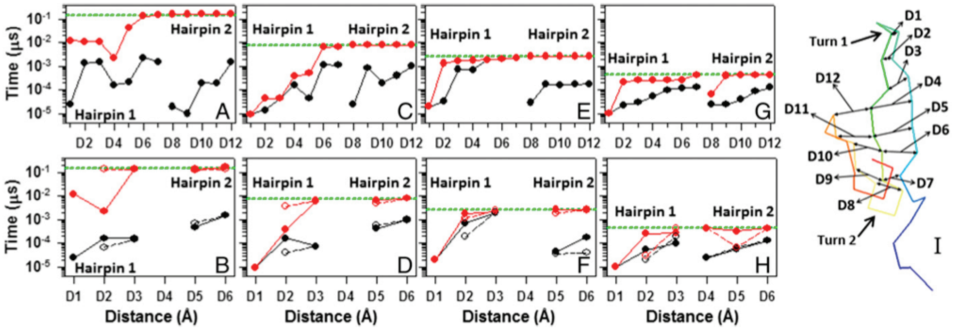


Fig. 2. The first contact time (black circles connected by black lines) and stabilization time (red circles connected by red lines) vs. the distances between C^{α}s of selected pairs of residues of hairpin 1 (D1 → Ala14 and Gly16, D2 → Thr13 and Lys17, D3 → Lys12 and Thr18, D4 → Tyr11 and Tyr19, D5 → Glu10 and Tyr20, D6 → Thr9 and Tyr21, and D7 → Trp8 and Asn22) and hairpin 2 (D8 → Asn23 and Asp26, D9 → Asn22 and Glu27, D10 → Tyr21 and Ser28, D11 → Tyr20 and Thr29, and D12 → Tyr19 and Trp30) and also, vs. the distances between C^{α}s (C^{β}s are represented by white circles connected by dashed lines) of only nonpolar residues of hairpin 1 (D1 → Ala14 and Gly16, D2 → Tyr11 and Tyr19, and D3 → Trp8 and Tyr20) and hairpin 2 (D5 → Tyr20 and Pro33 and D6 → Tyr19 and Trp30) for (*A* and *B*) three-state, (*C* and *D*) two-state, and (*E* and *F*) downhill folding trajectories of the L26D mutant and (*G* and *H*) the downhill folding trajectory of the L26W mutant (D8 → Asn23 and Trp26 in *G* and D4 → Tyr21 and Trp26 in *H*). Structure of L26D is illustrated in *I*. Horizontal green dashed lines indicate the folding time of L26D and L26W.

extended chain to a collapsed hydrophobic pocket. This method predicts the nucleation regions for protein folding by estimating the free energy of formation of the nucleation sites. This model was later referred to in the literature as a zipper model (38). The first hairpin in a three-state folding trajectory seems to fold through the hydrophobic collapse mechanism proposed by Dinner et al. (39), in which the hydrophobic collapse initiates hairpin formation (Fig. 2*A*). Two pairs of nonpolar residues, (Tyr11 and Tyr19) and (Trp8 and Tyr20), which make the first contact at the same time (based on the distance between C^{α}s and almost the same time based on C^{β}s), are the only driving force for formation of hairpin 1, because the pair (Ala14 and Gly16), responsible for the formation of the first turn, makes the first contact in the early stage of the trajectory but becomes deformed and does not stabilize for a long time (Fig. 2*B*). The folding of the first hairpin in the two-state folding trajectory is a "mixture" of these two mechanisms (i.e., it does not exhibit the "order") (Fig. 2*D*) during formation of nucleation sites proposed by Matheson and Scheraga (37) (i.e., the contact between Trp8 and Tyr20 forms faster than the contact between Tyr11 and Tyr19) and looks more like hydrophobic collapse; however, because of the pair Ala14 and Gly16, which makes a contact in the early stage of the trajectory and remains stabilized (i.e., the first turn is formed), hairpin 1 manages to restore the order (Fig. 2 *C* and *D*). Based on the first contact time between the nonpolar residues (Fig. 2 *B* and *D*), the second hairpin in both three-state and two-state folding trajectories folds through the model by Matheson and Scheraga (37); however, the order is distorted if we consider all of the selected pairs of hairpin 2 (Fig. 2 *A* and *C*). Based on these results, it is crucial when the pairs of nonpolar residues make the first contact and how fast the first turn is formed. If the pair of nonpolar residues, located farther from the turn, makes the first contact faster (or almost at the same time) than the pair of nonpolar residues located closer to the turn and the first turn does not form at the beginning of the trajectory, this order can cause a delay of stabilization of contacts between not only these residues but also, the residues of all selected pairs (especially the pairs of residues pertaining to the second hairpin), which finally induces the emergence of the intermediate state.

It should be noted that we also examined several three-state folding trajectories of the four other mutants examined above and found that the folding mechanisms of the hairpins in these trajectories do not differ from one of those (the three-state folder) presented in this section.

FEPs Along θ_i- and γ_i-Angles of Folding Trajectories. To explain the origins of three different (downhill, two-state, and three-state) folding scenarios, we studied the local motions of each residue along the sequence. In particular, we investigated the FEPs along the backbone virtual bond angle-θ and backbone virtual bond dihedral angle-γ of each residue (*SI Materials and Methods* and Fig. S4).

Fig. S5 illustrates the FEPs along the θ_i- (Fig. S5*A*) and γ_i-angles (Fig. S5*B*) computed from the above-discussed four MD trajectories. The black, blue, red, and green curves in Fig. S5 correspond to the FEPs of three-state, two-state, and two downhill folding trajectories, respectively. Small red and blue circles at the bottom of each panel in Fig. S5 are the NMR-derived structural data of L26D and L26W, respectively.

By comparing the FEPs along all of the θ- and γ-angles of four trajectories (Fig. S5), we found that the FEPs along most of the θ-angles for all four trajectories are similar to each other; however, there are several θ-angles (θ_i; i = 6, 7, 14, and 19–21) along which the FEPs are different (Fig. S5*A*). These differences are based on the formation of local minima, which are deepest for the FEP corresponding to the three-state trajectory and gradually become shallow (or disappear) for the FEPs corresponding to two-state and downhill folding trajectories. As was shown previously (15), the γ-angles are more sensitive and correlated to the global motions of the protein than the θ-angles; hence, differences between the FEPs along more γ-angles were found (Fig. S5*B*). These differences can more or less be observed along almost every γ-angle, except the γ-angles belonging to the C and N termini. They are more complicated than the FEPs along the θ-angles; however, we can identify how gradually the deepness of local minima changes (or vanishes) with the folding scenario. As in the FEPs along the θ-angles, the significant differences between the FEPs occur along the γ-angles consisting of residues pertaining to the solvent-exposed hydrophobic cluster (Tyr11, Tyr19, and Trp30), the delocalized hydrophobic core (Trp8, Tyr20, and Pro33), and the surface-exposed hydrophobic contact (Tyr21 and Leu26).

There is a clear correlation between the FEPs along the θ- and γ-angles and the contributions of the principal modes to the MSF (Fig. 1) (i.e., the θ- and γ-angles, along which the FEPs exhibit one or more local minima, have peaks in the graphs of the principal modes), which is understandable, because an existence of local minima on an FEP is a manifestation of jumps that the angles make, back and forth, between the local and global minima; hence, these angles contribute to the MSF. Contributions to the MSF increase with the deepness of local minima. Therefore, the principal modes of downhill folding trajectories have the least number of peaks. Because the FEPs and principal modes are correlated, we can conclude that the FEPs (Fig. S5) and FELs (Fig. 1) are correlated as well. In other words, the local minima on the FEPs are correlated to the local minima on the FELs. For the three-state trajectory, most of the FEPs along the γ-angles representing the β-strands and their edges exhibit three minima [one (deepest) corresponds to the native state, and the other two (shallow) correspond to the unfolded and intermediate states, respectively], whereas the FEPs along the θ-angles have mainly two minima [one (deepest) corresponds to the native state, and the second (shallow) corresponds to either the unfolded or intermediate state]. By comparing the FEPs along the γ-angles of three-state and two-state trajectories, we can easily identify which local minima of the FEPs along both angles correspond to the unfolded and the intermediate states. The results show that the main contributions in the formation of the intermediate state come from the θ_{21}-, γ_8-, γ_9-, γ_{21}-, γ_{25}-, and γ_{26}-angles pertaining to the surface-exposed hydrophobic contact (Tyr21 and Leu26) and part of the delocalized hydrophobic core (Trp8, Tyr20, and Pro33).

Discussion and Conclusions

One of the important problems in protein folding, the emergence of intermediates implicated in amyloid fibril formation, was addressed in this study in the example of the FBP28, which folds with biphasic kinetics. To understand the origins of biphasic folding kinetics of the FBP28, the structures of six new mutants (L26D, L26E, L26W, E27Y, T29D, and T29Y) have been determined by high-resolution NMR spectroscopy, and extensive MD simulations at different temperatures (below, very close to, and above the melting point) were performed with the coarse-grained UNRES force field. By analyzing the MD trajectories of these six mutants together with the WT in terms of the local motions of each residue and the distances between the C^{α}s and C^{β}s of selected pairs of residues over time and by PCA, we made the following findings.

i) All six mutants fold, maintaining the canonical WW structure as revealed by NMR. Their foldability increases, at most times, with increasing temperature. The mutations of Leu26 create better foldable systems (even better than WT in some cases) than the mutations of Thr29 and especially, Glu27.

ii) All six mutants and WT can fold through two different folding pathways with a different order of formation of the hairpins. Also, two mutants, L26D and L26W, can fold through a third folding pathway, in which both hairpins form simultaneously.

iii) Three-state folding is a major folding scenario of all six mutants and WT. However, two other folding scenarios, two-state and downhill folding, have been identified in ∼15% of folding MD trajectories for L26D and L26W.

iv) For formation of intermediates, it is crucial how each hairpin, especially hairpin 1, folds. If both hairpins are formed by the mechanism by Matheson and Scheraga (37), then the system may fold through a downhill (or two-state) folding scenario. If

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

hairpin 1 is formed by the hydrophobic collapse mechanism, then an intermediate state emerges, and the protein folds through a three-state folding scenario. Apart from the contacts between nonpolar residues, it is also important for a folding scenario when the first contacts occur between selected pairs of polar–polar and polar–nonpolar residues (some of them form hydrogen bonds). In a downhill folding trajectory, the first contacts between all selected pairs of residues occur in order, starting from the pairs closest to the turns. This order is distorted in two-state and three-state folding trajectories.

Previous experimental and theoretical studies (3, 11, 14) have shown that it is not easy to eliminate the intermediate state by mutation. However, the correlations between the local and global motions found here enabled us to identify the residues making the main contribution in the formation of the intermediate state. This approach can be applied to other proteins to identify the residues, mutations of which may help to eliminate intermediates.

Finally, the problems addressed in this study (i.e., elucidation of the origins of formation of intermediates and finding ways to prevent them) are very important for understanding folding/misfolding in general. The findings regarding the folding of the hairpins by different mechanisms, their role in the formation of intermediates, and the correlations between the local and global motions have general importance and can be applied to a broader class of proteins.

Materials and Methods

Canonical MD simulations were carried out with the UNRES force field parameterized (29) on the β-strand protein 1E0L and the α-helical protein 1ENH. The UNRES force field takes the solvent into account implicitly through the mean–force potential of interactions between united side chains (29). The Berendsen thermostat (40) was used to maintain constant temperature. The time step in MD simulations was $\delta t = 0.1$ mtu (molecular time unit) [1 mtu = 48.9 fs is the "natural" time unit of MD (41)], and the coupling parameter of the Berendsen thermostat was $\tau = 1$ mtu. In total, $\sim 3 \times 10^8$ MD steps were run for each trajectory, starting from the fully extended structure. Details of the protein purification and structural determination are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Dr. Adam Liwo for helpful discussions. This research was conducted by using the resources of (*i*) our 588-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University; (*ii*) the Informatics Center of the Metropolitan Academic Network in Gdansk; (*iii*) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center; and (*iv*) the Institute for Research in Biomedicine Barcelona. This work was supported by National Institutes of Health Grant GM-14312, National Science Foundation Grant MCB10-19767, Polish Ministry of Science and Education Grant 530-8370-D498-14, the Foundation for Polish Science Grant Mistrz 7./2013, and Spanish National Research Program, Ministry of Economy and Competitiveness Grants SAF2011-25119 and BFU2014-53787-P. M.J.M. is a Catalan Institution for Research and Advanced Studies Programme Investigator.

1. Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM (1998) Amyloid fibril formation by an SH3 domain. *Proc Natl Acad Sci USA* 95(8):4224–4228.
2. Ramirez-Alvarado M, Merkel JS, Regan L (2000) A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proc Natl Acad Sci USA* 97(16):8979–8984.
3. Karanicolas J, Brooks CL, 3rd (2003) The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design? *Proc Natl Acad Sci USA* 100(7):3954–3959.
4. Ferguson N, et al. (2003) Rapid amyloid fiber formation from the fast-folding WW domain FBP28. *Proc Natl Acad Sci USA* 100(17):9814–9819.
5. Karanicolas J, Brooks CL, 3rd (2004) Integrating folding kinetics and protein function: Biphasic kinetics and dual binding specificity in a WW domain. *Proc Natl Acad Sci USA* 101(10):3432–3437.
6. Mu Y, Nordenskiöld L, Tam JP (2006) Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. *Biophys J* 90(11):3983–3992.
7. Neudecker P, et al. (2012) Structure of an intermediate state in protein folding and aggregation. *Science* 336(6079):362–366.
8. Sudol M (1996) Structure and function of the WW domain. *Prog Biophys Mol Biol* 65(1-2):113–132.
9. Sudol M, Hunter T (2000) NeW wrinkles for an old domain. *Cell* 103(7):1001–1004.
10. Macias MJ, Gervais V, Civera C, Oschkinat H (2000) Structural analysis of WW domains and design of a WW prototype. *Nat Struct Biol* 7(5):375–379.
11. Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW (2003) Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proc Natl Acad Sci USA* 100(7):3948–3953.
12. Maisuradze GG, Liwo A, Scheraga HA (2009) Principal component analysis for protein folding dynamics. *J Mol Biol* 385(1):312–329.
13. Maisuradze GG, Liwo A, Scheraga HA (2010) Relation between free energy landscapes of proteins and dynamics. *J Chem Theory Comput* 6(2):583–595.
14. Maisuradze GG, Zhou R, Liwo A, Xiao Y, Scheraga HA (2012) Effects of mutation, truncation, and temperature on the folding kinetics of a WW domain. *J Mol Biol* 420(4-5):350–365.
15. Maisuradze GG, Liwo A, Senet P, Scheraga HA (2013) Local vs global motions in protein folding. *J Chem Theory Comput* 9(7):2907–2921.
16. Zhou R, et al. (2014) Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements. *Proc Natl Acad Sci USA* 111(51):18243–18248.
17. Jäger M, Nguyen H, Crane JC, Kelly JW, Gruebele M (2001) The folding mechanism of a β-sheet: The WW domain. *J Mol Biol* 311(2):373–393.
18. Ferguson N, Johnson CM, Macias M, Oschkinat H, Fersht A (2001) Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. *Proc Natl Acad Sci USA* 98(23):13002–13007.
19. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR (2006) Φ-analysis at the experimental limits: Mechanism of β-hairpin formation. *J Mol Biol* 360(4):865–881.
20. Ferguson N, et al. (2006) General structural motifs of amyloid protofilaments. *Proc Natl Acad Sci USA* 103(44):16248–16253.
21. Liu F, et al. (2008) An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc Natl Acad Sci USA* 105(7):2369–2374.
22. Jager M, et al. (2008) Understanding the mechanism of β-sheet folding from a chemical and biological perspective. *Peptide Sci* 90(6):751–758.
23. Davis CM, Dyer RB (2014) WW domain folding complexity revealed by infrared spectroscopy. *Biochemistry* 53(34):5476–5484.
24. Ferguson N, et al. (2001) Using flexible loop mimetics to extend phi-value analysis to secondary structure interactions. *Proc Natl Acad Sci USA* 98(23):13008–13013.
25. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106(45):19011–19016.
26. Piana S, et al. (2011) Computational design and experimental testing of the fastest-folding β-sheet protein. *J Mol Biol* 405(1):43–48.
27. Beccara SA, Škrbić T, Covino R, Faccioli P (2012) Dominant folding pathways of a WW domain. *Proc Natl Acad Sci USA* 109(7):2330–2335.
28. Liwo A, Czaplewski C, Pillardy J, Scheraga HA (2001) Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J Chem Phys* 115(5):2323–2347.
29. Liwo A, et al. (2008) Optimization of the physics-based united-residue force field (UNRES) for protein folding simulations. *Proceedings of the NIC Symposium*, eds Munster G, Wolf D, Kremer M (NIC Directors, Julich, Germany), pp 63–70.
30. Maisuradze GG, Senet P, Czaplewski C, Liwo A, Scheraga HA (2010) Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. *J Phys Chem A* 114(13):4471–4485.
31. Jolliffe IT (2002) *Principal Component Analysis* (Springer, New York).
32. Xu J, Huang L, Shakhnovich EI (2011) The ensemble folding kinetics of the FBP28 WW domain revealed by an all-atom Monte Carlo simulation in a knowledge-based potential. *Proteins* 79(6):1704–1714.
33. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598–1603.
34. Brooks CL, 3rd, Onuchic JN, Wales DJ (2001) Statistical thermodynamics. Taking a walk on a landscape. *Science* 293(5530):612–613.
35. Wales DJ (2003) *Energy Landscapes* (Cambridge Univ Press, Cambridge, United Kingdom).
36. Maisuradze GG, Liwo A, Scheraga HA (2009) How adequate are one- and two-dimensional free energy landscapes for protein folding dynamics? *Phys Rev Lett* 102(23):238102.
37. Matheson RR, Scheraga HA (1978) A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules* 11(4):819–829.
38. Dill KA, Fiebig KM, Chan HS (1993) Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA* 90(5):1942–1946.
39. Dinner AR, Lazaridis T, Karplus M (1999) Understanding beta-hairpin formation. *Proc Natl Acad Sci USA* 96(16):9068–9073.
40. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684–3690.
41. Khalili M, Liwo A, Rakowski F, Grochowski P, Scheraga HA (2005) Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J Phys Chem B* 109(28):13785–13797.

JORDI**MEDINA**

PhD in Biomedicine Candidate

EDUCATION

PhD in Biomedicine (Exp. Dec. 2017)
University of Barcelona

MSc in Biomedicine (2013)
University of Barcelona

BSc in Biotechnology (2012)
Autonomous University of Barcelona

BSc in Biochemistry (2012)
Autonomous University of Barcelona

LANGUAGES

English	●●●●●●○
French	●●●●○○○
Spanish	●●●●●●●
Catalan	●●●●●●●

IT SKILLS

Database search

Pubmed	●●●●●●○
Espacenet	●●●●○○○

Data analysis

Prism	●●●○○○○

PERSONAL INFO

Jordi Medina Vives

January 3rd, 1989

Spanish

Calderon de la Barca, 19 B-1
08032, Barcelona - SPAIN

PROFILE

I am a highly-motivated fourth-year PhD student with over 5-years' experience in structural biology and protein biophysics. During these years I have developed particularly strong analytical and communication skills as well as competencies to perform my duties independently. My near-term career objectives are to transfer my aptitudes and background to the promotion and evaluation of new inventions in the life sciences area.

RESEARCH EXPERIENCE

PhD Student. Institute for Research in Biomedicine (IRB Barcelona)
Sept 2013 - Dec 2017 (Expected)

PhD Thesis director: Maria J Macias

PhD thesis: Structure, dynamics and complex formation of eukaryotic transcriptional regulators.

MSc Student. Institute for Research in Biomedicine (IRB Barcelona)
Oct 2012 - July 2013

Master thesis: NMR study of the E3 ubiquitin-ligase WWP2 - Smad3 interaction.

Barcelona Science Park Intern. Molecular Biology Institute of Barcelona (IBMB-CSIC)
July 2011 - Sept 2011

Project: Effects of the suppression of Histone H1x by short-hairpin RNAs.

SELECTED COURSES AND CONGRESSES

Summer School on Intellectual Property. Universität Bonn.
31st July - 11th August, 2017. Bonn, Germany

EMBO Course on Structural Characterization of Macromolecular Assemblies
21st - 27th May, 2016. Grenoble, France

Poster presented: Finding Hotspots on the Surface of Smad2 MH2 Domain

IRB Barcelona PhD Students Symposium
12th-13th October 2015. Barcelona.

Poster presented: Preventing Fibril Formation of Protein FBP28 by Selective Mutation

Winter School on Quantum Theory of NMR. Universität Leipzig.
23rd-27th February 2015. Colditz, Germany

SELECTED FELLOWSHIPS AND AWARDS

IRB Barcelona International PhD Fellowship. 2013
Awarded by: IRB Barcelona

University Entrance Test (PAU) With Honors Distinction. 2007
Awarded by: Catalan Ministry for Innovation, Universities and Research

PUBLICATIONS

Maisuradze GG, Medina J, Kachlishvili K, Krupa P, Mozolewska MA, Martin-Malpartida P, et al. **Preventing fibril formation of a protein by selective mutation.** PNAS. 2015

✉ jordi.medina@protonmail.com ☎ +34 676 854 675 in linkedin.com/in/jordi-medinavives