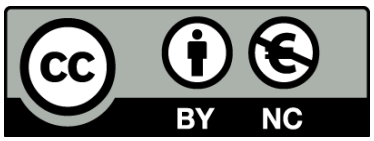
UNIVERSITAT DE BARCELONA

Structural studies of recombinant TGIF1 and FBP28 WW domains using NMR and peptide ligation strategies

David Suñol Moreno



UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

Structural studies of recombinant TGIF1 and FBP28 WW domains using NMR and peptide ligation strategies

David Suñol Moreno, 2016

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT EN BIOMEDICINA

Structural studies of recombinant TGIF1 and FBP28 WW domains using NMR and peptide ligation strategies

Memòria presentada per David Suñol Moreno per optar al títol de doctor per la universitat de Barcelona

Dra. Maria J. Macias Hernández
Directora de Tesi
Prof. d'investigació ICREA

David Suñol Moreno
Doctorand

Dr. Pedro F. Marrero González
Tutor de Tesi
Prof. Facultat de Farmàcia

Tesi realitzada a l'Institut de Recerca Biomèdica de Barcelona
Parc Científic de Barcelona
Barcelona, 2016

UNIVERSITAT DE BARCELONA

INSTITUTE FOR RESEARCH IN BIOMEDICINE
IRB BARCELONA

Declarations

1. This work has been carried out under the supervision of Maria J. Macias, Ph.D. (ICREA research professor and PI of the Structural Characterization of Macromolecular Assemblies group at the IRB Barcelona) at the department of Structural and Computational Biology of the Institute for Research in Biomedicine (IRB Barcelona).

2. No portion of the work referred to in this thesis has been submitted in support of an application for another degree qualification of this or any other University.

3. The work described in this thesis has been published in two peer-reviewed papers.

4. Financial support was obtained from the "La Caixa"/IRB Barcelona International PhD Programme Fellowship. This work was supported in part by the Ministerio de Economía y Competitividad, Gobierno de España (SAF2011-25119 and BFU2014-53787-P).

Acknowledgements

First of all, I would like to acknowledge my thesis director, Dr. Maria J Macias, for giving me the opportunity to carry out this thesis in her group. Your support and guidance have been essential to make me a discerning scientist and a wiser person, while your words have endless encouraged to achieve the best of myself. I would also want to thank the financial support I received from "la Caixa" as "La Caixa" /IRB Barcelona International PhD Programme fellow.

I highly thank Dr. Phillipp Selenko for accepting me in his lab in Berlin and enlightening me with his ideas and advices, that expanded my conception of what is science nowadays. Likewise, this gratitude is also extensible to François and all the other lab members, Akis, Marchel and Marleen. And impossible to forget are Camille, Floriant, Isaline and Jeff for those great times in German land. Special thanks to Dr. Sophie Zinn-Justin, not only for the clone she provided us, but for her kindness and suggestions.

I am grateful for the advice and suggestions from Dr. Phillipp Selenko, Dr. Modesto Orozco, Dr. Antoni Riera and Dr. Conchita Civera, who have been the members on my thesis advisory committee.

Essential and interesting were all my lab colleagues, with whom I play in the park: Macho team, first as always. M001 Toni, Благодарам што ми го пренесе твоето знаење мој господару, мој пријателе. M010 Tiago, foda-se! És do caralho! M100 Albert, merci per ser un far que ha guiat el meu camí al lab. Constanze, Danke dass du mein seil gehalten hast als ich gefallen bin und geschrien habe: BANANA!. Jordi: la SMAD2 mai hagués estat el mateix sense tu....soooort en el teu camí! ;). Ewelina, Dzięki Bogu Jest Piątek! Dzięki za mądre porady. DO PRZODU!. Regina & Marco, Ihr kamt, ihr saht, ihr siegtet! Eric, gràcies per ensenyar-me a no ser una L mai més! Pau, gràcies per descobrir-me secrets dels ordinadors, dels colors i del "sis-cents". Jimmy, you're a great mate. Mads, I could learn many things in many ways. Àngela, gràcies per fer que cada dia comenci de nou la primavera. Marta, filla de Davy, gràcies per fer veurem el món d'un altre prisma, sota la llum d'Hèlios i Selene, on tot comença i res acaba. And finally, the lab Soul, Lidia, gràcies pel teu entusiasme, per la teva paciència, i pel teu somriure infinit.
And this acknowledgement is extended also to all the summer students of the lab: Júlia, the original great Minion, Lluc, thanks for the peptides!, Alessia, Lluís and all the others, and Sandra, the young padawan that will surpass the master.

Very funny and amazing were the SC members and people from others labs, from whom I learnt a variety of human personalities. Specially Laura, Rosa and Júlia for those cafes, talks, parties and 'oh yeah, there is life outside IRB!'.

In the lands on the second floor, two great human beings, Patricia and Leyre, have made my life much easier and funnier.
Agrair també a aquells treballadors anònims que netegen el material i el lab, tenen cura dels instruments comuns, que ens porten regals cada dia, i en definitiva fan que tot això funcioni.

I would also want to thank the support of all my people outside the scientific-geek world:

M'agradaria agrair el suport de tots aquells amics que han estat sempre allà: Nina, Yoio, gràcies per poder comptar sempre amb vosaltres, per qualsevol cosa. Pau, Àngel, Albert, Víctor, potser no ens veiem tant com ens agradaria, però tots sou uns grans amics! Carol, hi ha amics que són per a tota la vida i tu ets un d'ells. Gràcies per estar sempre allà. Ferran & Jessica, que haría yo sin vosotros y sin las escapadas norteñas! Gràcies per acollir-me al vostre món. Agnieszka, Obraz jest niesamowity, ale Ty jeszcze bardziej! Noelia, tu, que viniste de las estrellas, con luz y calor me abrazaste, siempre resplandecerás, brillante, sin que nada ensombrezca jamás tu camino celestial.
També gràcies Andrea, per fer la portada de tesis més xula i verda que es té noticia!
Gràcies també als increïbles monitors de l'esplai, i als grans i petits, per convidar-me a passar tants bons moments amb vosaltres dintre i fora de l'esplai, i per mai oblidar que no som més que nens grans! GdG al ∞!
Finally, voldria agrair el suport inestimable de la meva família cada dia més gran! :) A tots els tiets, tietes, cosins i primo per aquelles estones agradables entre experiment i experiment. Especialment, voldria referir-me a l'Àlex i al seu germà baby XY, per ser les alegries més grans que he tingut en aquests 4 anys.
I sobretot als meus pares, Angelines i Josep, gràcies per estar allà, per ésser pesats, per suportar-me tots aquests anys, por ser especiales, por no rendiros nunca y en definitiva por ser como sois y seréis siempre.

Simplement gràcies, simply thank you,

Atwenter Elandhar - David Suñol Moreno

"Science is playing like a kid but with wisdom"

“Many of life’s failures are people who did not realise how close they were to success when they gave up.”

– Thomas Alva Edison

Table of Contents

1 Introduction

1.1 TGF-β pathway

Transforming growth factor-β (TGFβ) is one of the main signalling pathways that regulates a plethora of functions in metazoans, including many fundamental cellular processes such as cell differentiation, proliferation, tissue homeostasis or regeneration, among many others [1]. The control of a multicellular environment in animals rise as the main objective of TGFβ as it has been observed in all the sequenced metazoans, even in *Trichoplax adhaerens*, the most basal animal sequenced [2].

One of the best-studied networks activated by the TGF-β family involves the small mothers against decapentaplegic (SMADs) transcription factors as the central mediators of the pathway [3]. Briefly, external TGFβ-related cytokines trigger the activation of their receptors upon binding. Once TGFβ receptor I is phosphorylated, R-SMAD proteins get in turn also phosphorylated, favouring the formation of a heterotrimeric complex between two R-SMAD and one Co-SMAD. This complex accumulates in the nucleus and participates in transcriptional regulation in association with other proteins like transcription factors, co-activators and co-repressors and also with DNA. Finally, dephosphorylation and ubiquitination of the activated SMADs ends the signal (Figure 1.1).

The TGFβ-related cytokines are divided in two subfamilies; first: the TGF-β / Activin / Nodal and second: bone morphogenetic protein (BMP) / growth and differentiation factors (GDF) / Muellerian inhibiting substance (MIS) subfamily. The two groups bind to different combinations of receptors I and II, which subsequently activate different SMAD proteins [3]. Curiously, these are the only cell surface receptors that use

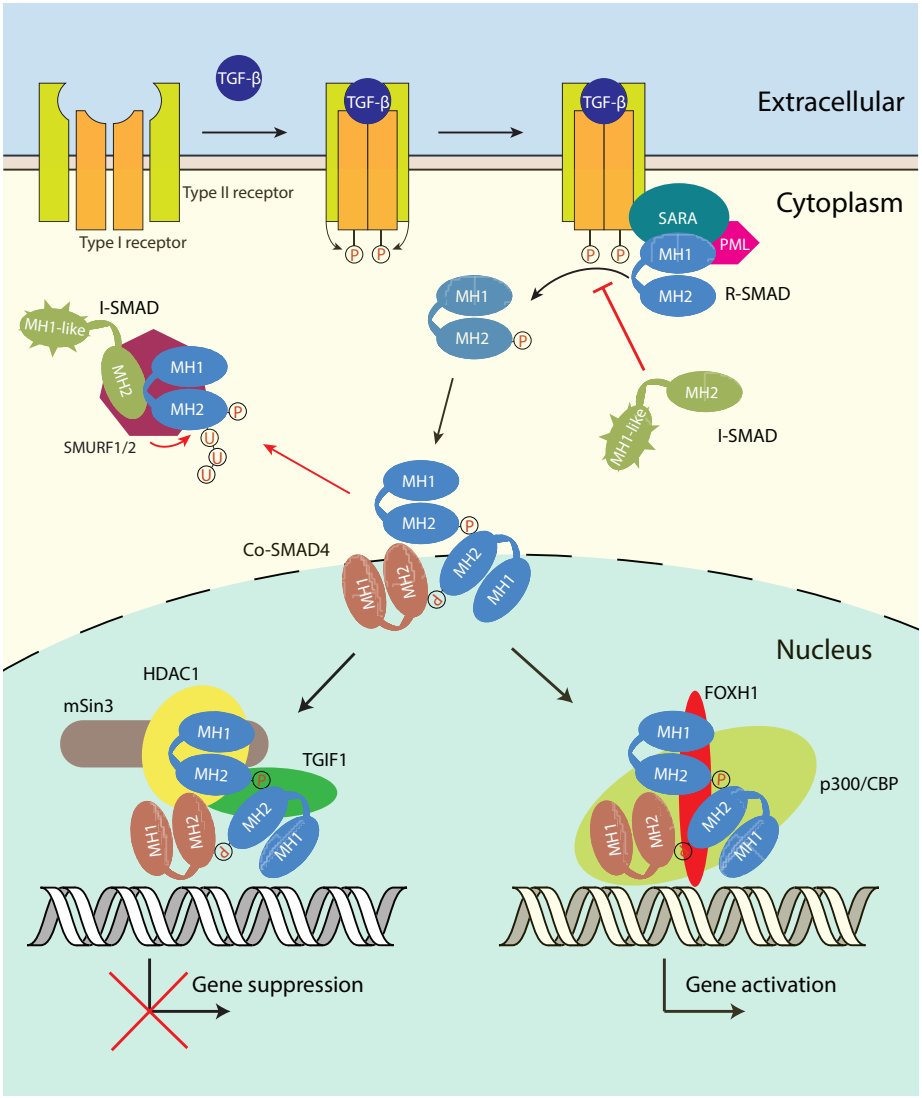


Figure 1.1: General scheme of the TGFβ signalling pathway.

serine/threonine kinases, instead of the commonly used tyrosine kinases. Generally, TGF-β/Activin/Nodal cytokines trigger the response of SMAD2/3 while BMP/GDF/MIS activate SMAD1/5/8. These five proteins are commonly named receptor-regulated SMAD (R-SMAD) because they are the ones that undergo phosphorylation by the receptor after the activation of the pathway. Once activated, the R-SMAD pro-

teins are translocated to the nucleus, where they form a heterotrimer with SMAD4, also called Co-SMAD, mediator of all R-SMADs. The heterotrimer is likely to be the functional unit of the pathway. It interacts with many partners and cofactors to, at the end, bind to the DNA and regulate the expression of hundreds of genes [4]. Finally, the inhibitory SMADs (I-SMADs : SMAD6 and SMAD7), act as a negative regulators of the pathway.

The apparent simplicity of the pathway contrasts with the broad variety of responses that it generates. Indeed, TGFβ pathway can cause contrary outcomes depending on the cellular context. For instance, DNA-binding protein inhibitor ID-1 is suppressed by TGFβ pathway in mammary epithelial cells [4] but it is induced in metastatic breast cancer cells [5]. Actually, only few proteins are expressed through all the cell types, such as SMAD7 [6]. The broad variety of responses is tightly regulated by several factors, which can be classified in three main groups. The first regulation level describes how and which cytokine signal triggers which receptor. In humans, there are seven type I and five type II TGFβ receptors. A tetramer of two type I and two type II receptors is needed to bind the cytokine. The availability of cytokine and type I and II receptors will influence on the final response. The second level includes all the cofactors (activators or suppressors) and others proteins that could regulate the SMADs-DNA interaction. Finally, the cell epigenetic status is basic for the final output.

Moreover, SMAD proteins act as an integrative hub from other cellular signalling pathways. For example, the interaction with FOXO factors links TGFβ with AKT pathway [7]. In a similar way, activated LEF1 (lymphoid enhancer-binding factor 1) and TCF7L2 transcription factor are the connection to the WNT pathway [8]. And the cross-talk with mitogen-activated protein kinases (MAPKs) is centred in the phosphorylations of the linker region [9].

1.2 SMAD proteins

1.2.1 SMAD structure

The roughly 500 amino acids of the SMAD proteins are structured in two globular domains connected through an unstructured linker (Figures 1.2, 1.3, 1.4). The N-terminal domain, MAD homology 1 (MH1) domain, is conserved in all SMAD proteins with the exception of SMAD6 and SMAD7. A zinc cation (Zn^{2+}) stabilises the DNA binding compact structure encompassing four α helices, six short β-strands and five loops. A β-hairpin between the strands B2 and B3 is responsible for the interaction with the DNA [10]. All R-SMADs and SMAD4 bind to a specific sequence of DNA called SMAD binding element (SBE), containing the sequence 5'-AGAC-3'. SMAD2, which has in its most common splicing variant a 30-residue insertion next to the DNA binding site, is believed to have limited DNA binding capacity [3]. In addition to SBE, SMAD1 and SMAD5 can also recognise GC-rich sequences [11]. On the other hand, neither SMAD6 nor SMAD7 interact efficiently with DNA [3].

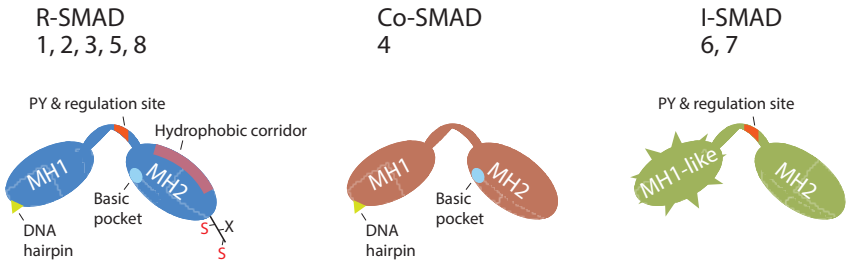


Figure 1.2: General scheme of SMAD proteins.

The C-terminal domain, MH2 domain, is conserved among the SMAD proteins. It is formed by a central β sandwich surrounded by three α-helices and one β-strand in one side and one loop-helix region in the other side [12]. This domain is responsible for most of the protein-protein interactions. The MH2 domain of the R-SMADs is characterised by a conserved C-terminal SXS motif. When phosphorylated by the type I receptor, these amino acids increase their affinity for a basic pocket located also in the MH2. The interaction between the phos-

phorylated C-terminal and the MH2 of another SMAD induces the formation of homo- or heterotrimers. The heterotrimer consisting of two R-SMADs and one SMAD4 is considered as the functional unit of the pathway. The MH2 in the R-SMAD also contains a hydrophobic corridor that mediates in many protein interactions such as nucleoporins, cytoplasmatic retention proteins or other transcription factors [3].

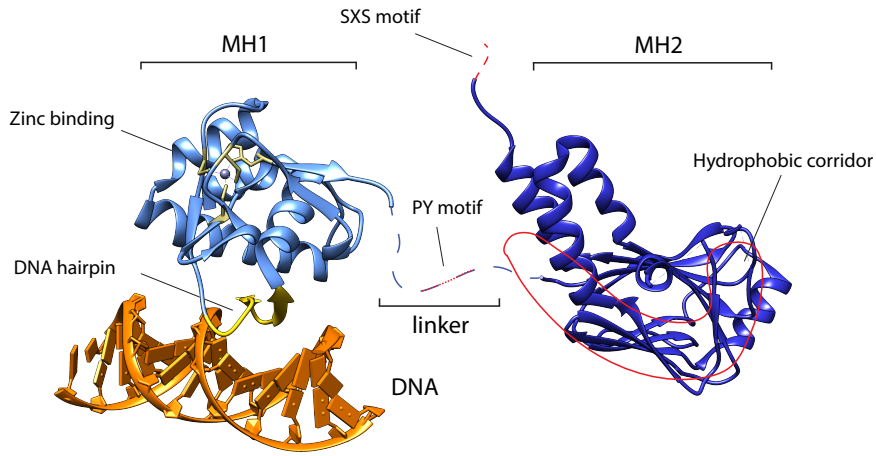


Figure 1.3: SMAD3 representation as an example of a R-SMAD. MH1 is labelled in light blue, MH2 in dark blue and DNA in orange. DNA hairpin is highlighted in yellow as well as Zinc binding region. In red are shown the regions corresponding to the PY motif, the hydrophobic corridor and the SXS motif, respectively. The structures were taken from PDB id: 1OZJ for MH1 and DNA and id: 1MJS for MH2.

The sequences connecting the MH1 and MH2 domains differ between SMADs but are conserved for a given type of SMADs. In particular, the PY motifs are present in the linkers of both R-SMADs and SMAD7. Also, the pattern of phosphorylated residues responsible for the regulation of the function and fate of R-SMADs is preserved among different SMADs [13, 14].

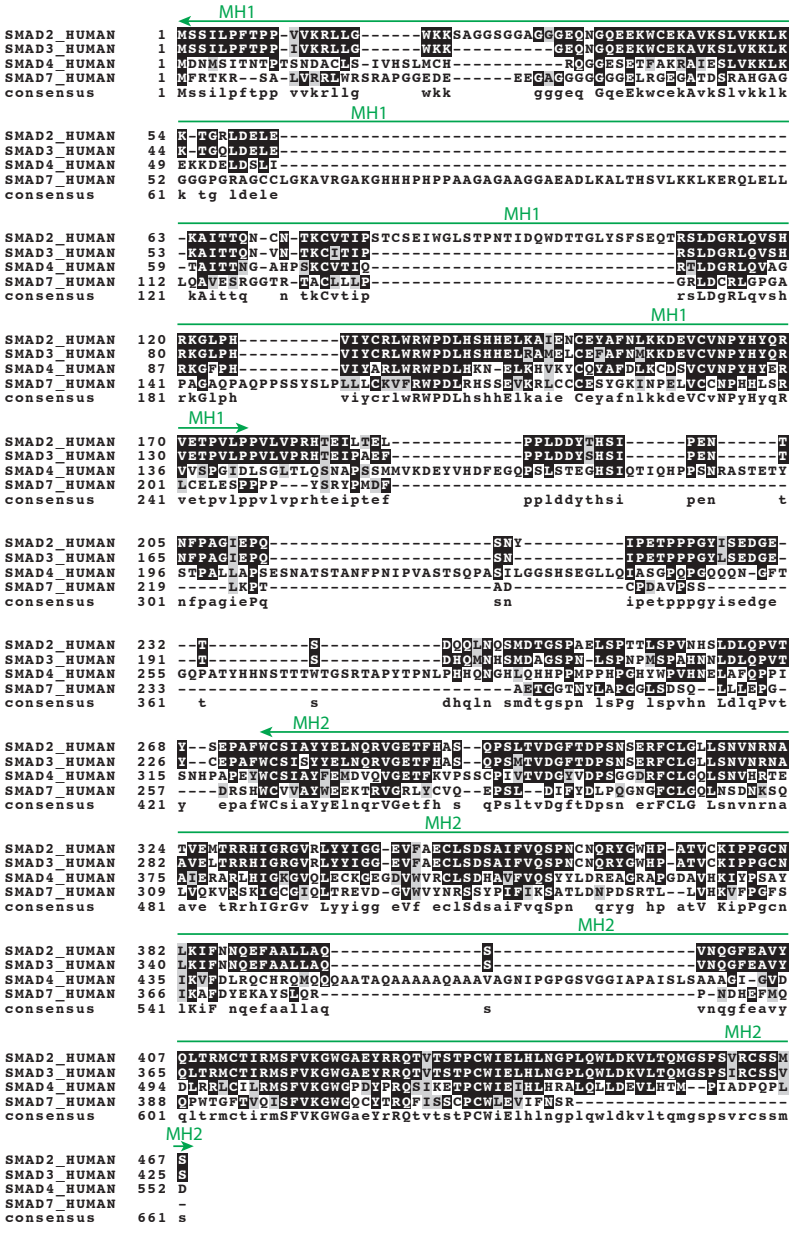


Figure 1.4: Sequence alignment between SMAD2, SMAD3, SMAD4 and SMAD7. MH1 and MH2 domains are indicated in green. Sequence alignment was done using M-Coffee software [15].

1.2.2 SMAD interacting proteins

Many SMAD interacting proteins have been characterised in the last decade [1]. In this section some examples will be introduced with special emphasis on the mode of binding to the SMAD proteins.

Smad anchor for receptor activation (SARA) is a protein that serves as a mediator between the TGFβ type I receptor and SMAD2/3 [16]. After the phosphorylation of the SMAD, the affinity of SARA to SMAD2/3 decreases significantly, allowing the release of the activated SMAD2/3. SARA interacts with SMAD2/3 through its SMAD binding domain (SBD), which is accommodated on the hydrophobic corridor of SMAD2/3 adopting an extended conformation [12]. The large area of SMAD2/3 covered by the SBD inhibits the efficient formation of SMAD2/3 trimers and thus SARA only binds to monomeric, unactivated SMAD2/3 [12].

Inside the nucleus, the activated R-SMADs form protein complexes with cofactors and DNA promoter sites. Among these cofactors, one of the best studied is the forkhead member FOXH1, a transcriptional co-activator that recruits the SMAD2/3-SMAD4 complex to the activin response element (ARE) [17]. Specifically, FOXH1 can bind to SMAD2/3 through its SMAD interacting motif (SIM), a proline-rich motif that binds to the α-helix 2 of SMAD2 MH2 [18]. This motif is not exclusive to FOXH1 related proteins but it is also found in other SMAD2-binding proteins such as the homeodomain Mixer [19]. In contrast to SARA, the interaction between the FOXH1 and SMAD2 does not interrupt the binding between two activated SMAD2 proteins, allowing the interaction between FOXH1 and the trimer [20]. In addition, FOXH1 also has a Fast/FoxH1 motif (FM) that can interact with a hydrophobic pocket of SMAD/SMAD interface [21]. This novel interacting motif is specific only to phosphorylated SMAD2, therefore, in contrast to SIM motif, FM do not bind to SMAD3. Moreover, this motif is only present in FOXH1 related proteins, which differentiates the interaction between FOXH1 and Mixer versus SMAD2 [21].

Another SMAD2/3 binding partners are the transcriptional co-activators p300 and cyclic AMP response element-binding protein (CBP). These

two similar proteins, both with histone acetyltransferase (HAT) activity, are recruited to chromatin by SMAD2/3 complexes. There, acetylation of the H3 and H4 histone tails promotes a rearrangement of the nucleosomes, enabling a better access for the transcriptional machinery [22]. Moreover, p300/CBP also acetylates SMAD proteins at their MH1 to enhance its affinity to DNA [23].

In addition to the interaction with co-transcriptional activators, SMAD proteins can also interact with transcriptional suppressors. For instance, SKI and Ski-like proteins bind to the MH2 of SMAD2 and SMAD4 interfering with the formation of the SMAD trimeric complexes [24]. In parallel to these mechanisms, SMADs can also recruit the complex formed by the nuclear receptor corepressor 1(N-CoR), mSin3A and Histone deacetylase 1 (HDAC1), which suppresses the gene transcription in an opposite way to p300/CBP [25]. Another examples include 5'-TG-3' interacting factor 1 (TGIF1) that competes with p300/CBP for SMADs interaction and ecotropic virus integration site 1 protein homolog (EVI1) that inhibits SMAD3 binding to DNA [20].

1.3 TGIF1

5'-TG-3' interacting factor 1 (TGIF1) was discovered in 1995 by Bertolino *et al.* as a novel protein that binds to the retinoid X receptor responsive element (RXRE), promoter of the rat retinol-binding protein II (CRBPII) [26] (Figure 1.5). Homology analysis identified TGIF1 as a member of the atypical homeodomain (HD) family of proteins. Further sequence comparison of TGIF1 and other HD members revealed a three amino acid insertion between helices 1 and 2. These three residues are shown to form a common feature among some of the proteins of the growing atypical homeodomain group. Thus, the authors proposed three amino acid loop extension (TALE) homeodomain class as a name to this new group of proteins. Nowadays, this arrangement is still valid as TGIF1 is classified as a member of the TGIF family, under the TALE superclass [27]. Bertolino *et al.* also determined, by EMSA experiments, the 5'-TGTCA-3' motif as the DNA sequence where the homeodomain TGIF1

preferentially binds. Besides, they found TGIF1 mRNA in most tissues, especially in placenta, prostate, testis and ovary. In contrast, there was no detection of mRNA in brain neither in muscle tissues [26].

Few years later, in 1999, Wotton *et al.* discovered the corepressor role of TGIF1 in TGFβ pathway [28]. The first evidence was the identification of TGIF1 as an interacting factor of SMAD2 by a two-hybrid library. Further immunochemistry experiments confirmed the binding between TGIF1 and SMAD2 and SMAD3, weaker with SMAD1 and none with SMAD4. In all positive cases, the binding was enhanced by the activation of the TGFβ pathway. Moreover, the authors pointed out the nuclear localisation of TGIF1 and, for first time, it was found that TGIF1 inhibits the induced TGFβ pathway [28]. Indeed, TGIF1 expression represses the transcription of plasminogen activator inhibitor-1 (PAI-1), a protein widely used to monitor TGFβ signalling. Moreover, when TGIF1 expression is reduced, the levels of PAI-1 increases again.

In other studies, TGIF1 function was also related with holoprosencephaly (HPE) [29]. HPE is a developmental defect that provokes brain malformation among many other consequences [30]. The causes are both environmental and genetic, including mutations in Sonic hedgehog protein (SHH) and SIX3 proteins. Further genetic analysis of patients carrying this disease led to discover that heterozygous mutations in the *Tgif* gene (localised in chromosome 18p11.3 [31]) were also causing HPE [29]. This result was supported by the fact that TGIF1 has a relevant role during embryogenesis, specially for the formation of the central nervous system [32]. Further research found other HPE associated mutations [33] and linked the disease to TGIF1 regulation upstream of SHH [34].

The implication of TGIF1 with the myelopoiesis (subprocess of hematopoiesis) was confirmed with the revelation of 1655 direct TGIF1 targets in HL60 cell line [35]. 35% of those genes were related with cell function, 18% with hematopoiesis, 16% with organ development among others aspects of cell function. Not surprisingly, proteins that are related to TGFβ pathway are among them, as TGFβ receptor 2 (TGFBR2), FOXH1 or TGIF2 [35].

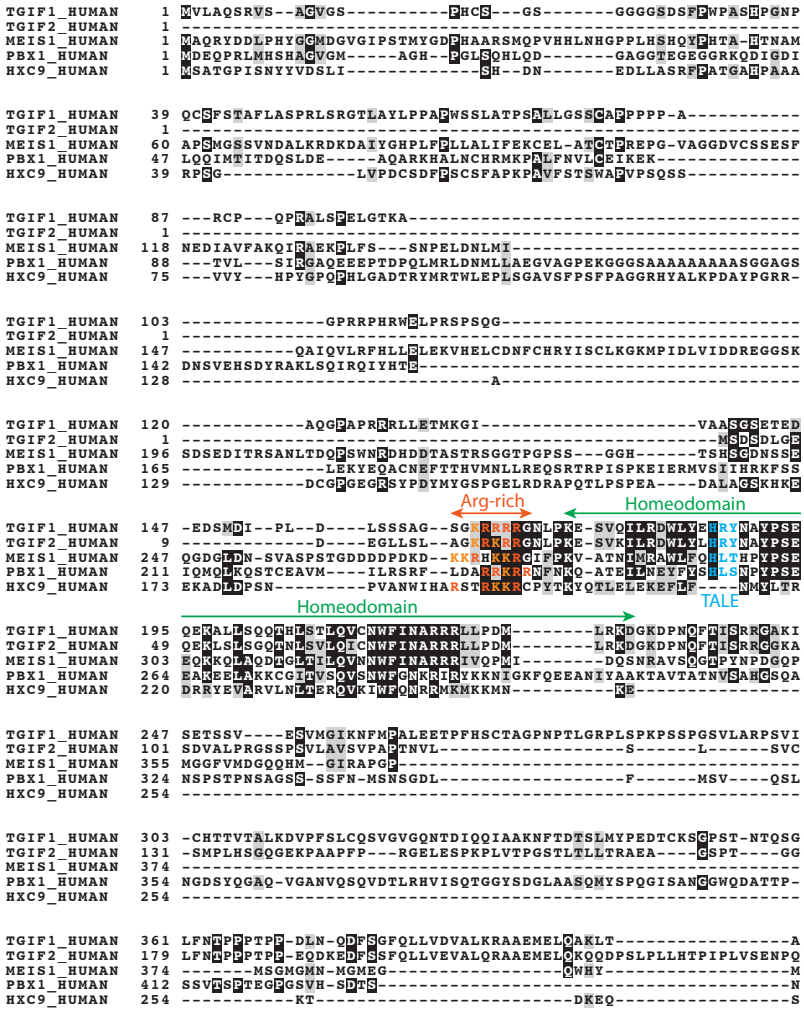


Figure 1.5: Sequence alignment between TGIF1 and TGIF2 (43.1% identity with TGIF1), MEIS1 (20.3%), PBX1 (18.6%) and HOXC9 (18.3%) proteins. The first four belong to Homeodomain-TALE super class proteins while HOXC9 is a HOX-nonTALE Homeodomain protein. In green is indicated the homeodomain, in red is highlighted the arginine-rich region before the homeodomain and in blue is shown the TALE motif for those TALE proteins. Sequence alignment was done using M-Coffee software [15].

TGIF1 is a short-lived protein with a half-life of less than 30 min [36].

Nevertheless, phosphorylation of threonines 364 and 368 by epidermal growth factor (EGF) (via ERK kinases) increases the half-life up to 1 h [36]. Interestingly, if these two threonines are mutated to valines, the half-life is also incremented. These results led to the conclusion that both threonines 364 and 368 are critical for the degradation of the protein, and their phosphorylation or mutation prevents the process [36]. Indeed, in 2010, F-box/WD repeat-containing protein 7 (FBXW7), a substrate-recognition factor of SCF ubiquitin ligase, was shown to interact with TGIF1 [37]. Nevertheless, FBXW7 only recognises phosphorylated substrates, meaning the phosphorylated long-lived TGIF1 was the only one to be recognised for degradation. This paradox could be explained by the identification of a second way of TGIF1 degradation by PHD and RING finger domain-containing protein 1 (PHRF1) ubiquitin ligase [38], not related to FBXW7.

In 2000, a second member of TGIF family, named TGIF2 (43.1% identity with TGIF1, Figure 1.5), was discovered [39]. TGFI2 is shorter than TGIF1 (237 amino acids long), missing the first 130 residues plus the ones between 297 and 347 of TGIF1. Nevertheless, the homeodomain is conserved with high similarity. Further research revealed that TGIF2 also represses transcription and binds to HDAC1, but because of missing amino acids, TGIF2 was unable to bind C-terminal-binding protein 1 (CtBP) [40]. Overall, it was suggested that TGIF2 has many overlapped functions with TGIF1.

1.3.1 TGIF1 structure

The NMR structure of the homeodomain (173-230) was deposited in the PDB by the Northeast Structural Genomics Consortium (Target HR4411B; PDB id: 2LK2). The structure maintains the typical HD shape formed by three α-helices. There is no structure of TGIF1 in complex with DNA, but the α-helix 3 is considered as the potential DNA binding site by comparison with the other HD structures. The histine-arginine-tyrosine residues, defined as the TALE element, are localised between α-helix 1 and α-helix 2, at positions 186-188 (Figure 1.6 A).

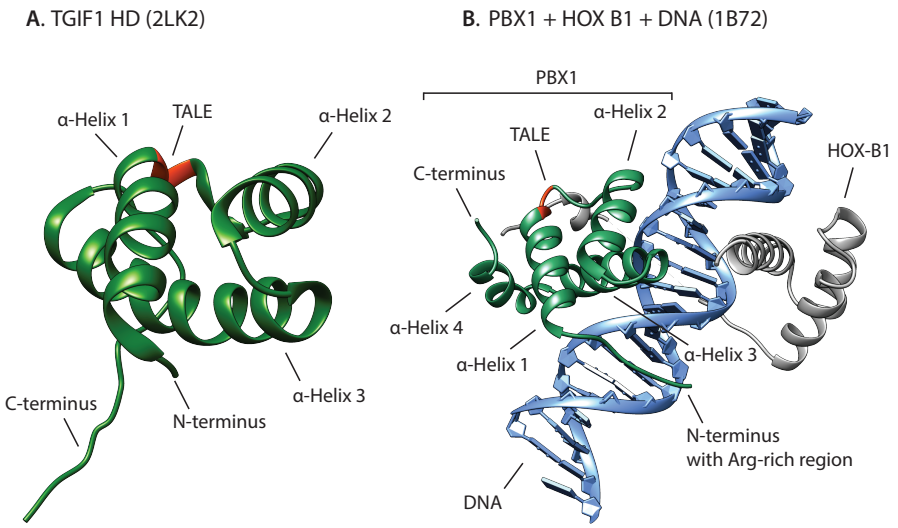


Figure 1.6: A) Structure of TGIF1 Homeodomain (173-230) by NMR (PDB id: 2LK2). **B)** Crystal structure of the complex PBX1 - HOX B1 - DNA (PDB id: 1B72).

The crystal structure of PBX1, another TALE superclass protein (18.6% identity with TGIF1, Figure 1.5), in a complex with HOX B1 and DNA (PDB id:1B72), reveals that the interaction between the two proteins is mediated through the hydrophobic C-terminal tail of HOX B1 that binds to the pocket formed between the α-helix 3 and the three residues of the TALE in the PBX1 [41] (Figure 1.6 B). This interaction suggests that the three amino acids of the TALE segment might be relevant for a protein-protein interactions in other proteins that belong to the TALE superclass. Moreover, the structure of the PBX1 reveals a fourth α-helix, just after the third one in the sequence (Figure 1.6 B). However, this fourth α-helix is not observed in another TALE protein (MEIS1, 20.3% identity with TGIF1, Figure 1.5, PDB id: 5BNG [42]) and may be specific to the PBC family. Also, in the crystal structure, PBX1 binds DNA not only through the third α-helix (as commonly for HD proteins) but also through the N-terminal region of the HD, rich in arginines (Figures 1.5 and 1.6 B).

In contrast, the structure of the rest of TGIF1 is poorly understood. In

order to have more information about it, we have performed protein structure prediction using the MetaDisorderMD2 programme [43]. The result suggests that the protein is mainly disordered, with the exception of the HD and a 15 residues segment near to the C-terminus (Appendix Figure 7.1)

1.3.2 Role of TGIF1 in the TGFβ signalling pathway

TGIF1 is a transcription factor that suppresses the transcription of genes under the regulation of the TGFβ signalling [28]. The interaction with SMAD2/3, enhanced by the activation of TGFβ, suggests a regulation at nuclear level. In addition, it was demonstrated that TGIF1 bridges between SMAD2 and HDAC1, a known co-suppressor. Based on this information, a model where activated SMAD2/3 complex could bind either to p300/CBP, thus promoting gene expression, or to TGIF1 bridging HDAC1, suppressing the gene transcription, was proposed [28] (Figure 1.1).

Further experiments revealed that TGIF1 not only represses different promoters but also inhibits the gene expression even if the promoters are localised far away from the beginning of the transcription site [44]. Moreover, TGIF1 is capable to repress without direct binding to DNA. With these results, it has been suggested that TGIF1 can act as a general suppressor factor instead of being specific to one cell response. In this context, three different regions were identified to be able to repress transcription independently [44] (Figure 1.8 A). One region, located in 237-321, seems to be HDAC1 dependent in order to suppress the transcription. However, the two other suppressing regions (130-171 and 337-401) may not need the presence of HDAC1 to inhibit the gene expression.

Interestingly, a new study demonstrated that the region 130-171 of TGFI1 binds to CtBP1 through the conserved motif PLDLS located in that region [45]. The complex TGIF1-CtBP1 can repress independently from HDAC1 but it was found that HDAC1 can also interact with TGIF1 even when CtBP1 is already bound to TGIF1. Furthermore, it was also found that the TGIF1 region (337-401) also recruits its own co-repressor,

mSin3 [46]. mSin3 has been shown to be associated with HDAC1s forming a suppressor complex. The global picture suggests that TGIF1 serves as a bridge between SMAD2 and the general co-repressor complex HDAC1/mSin3 [46].

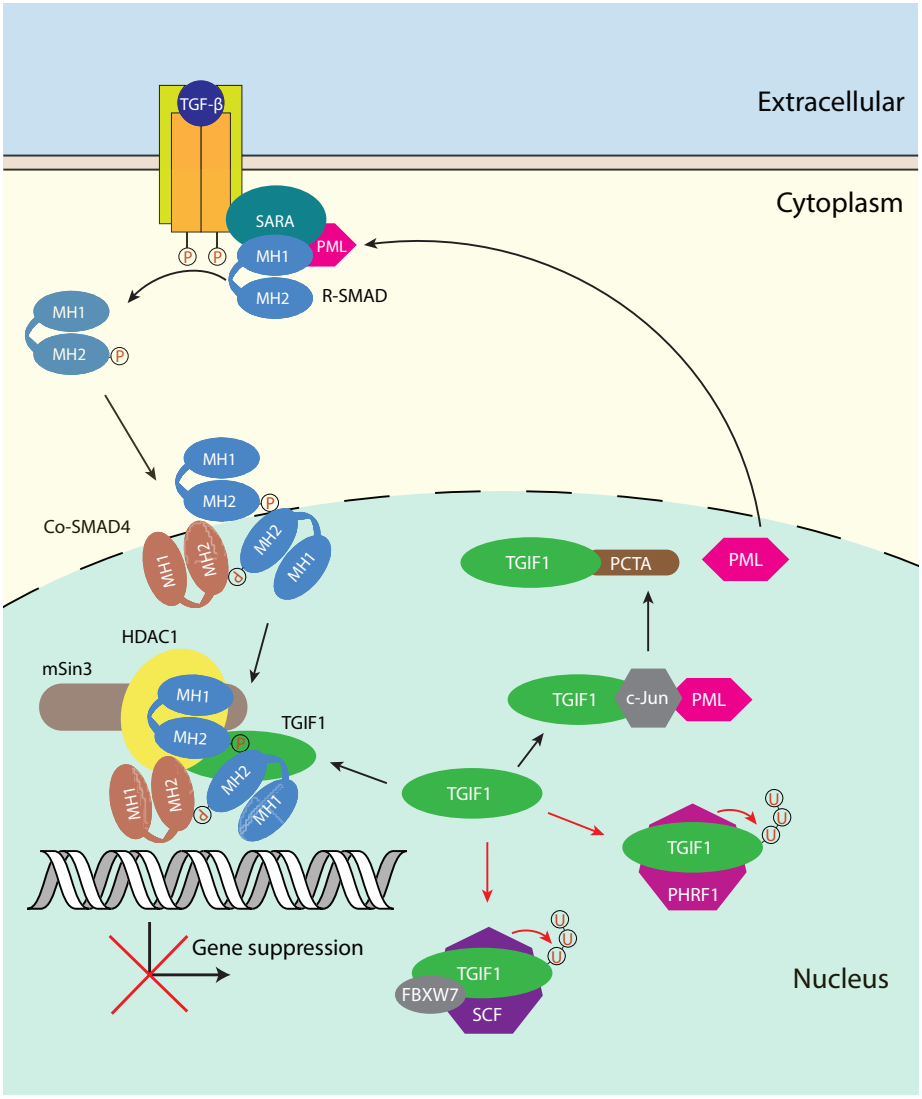


Figure 1.7: Scheme of TGIF1 roles in the TGFβ signalling pathway.

Overall, one repressor protein has been identified to every repressor do-

main of TGIF1. It is tempting to say that the mechanism explaining how TGIF1 represses the TGFβ pathway has been already solved. However, in another study, the authors tested the inhibition of the TGFβ pathway by TGIF1 wild type (WT) and TGIF1 with the mutations that were found to cause HPE. Surprisingly, most of the mutants behaved as the wild type, except for the mutation that, located in the homeodomain, impair the binding with DNA [29]. Even the mutant that impairs the binding with CtBP1 (substitution of serine to cysteine on the position 159) had a similar transcription inhibition. From these results, it seems that the transcription suppression is performed by different regions of TGIF1, being their functions overlapped and that TGIF1 plays more roles in the cell that the ones already demonstrated.

Indeed, a new TGFβ inhibiting pathway was discovered, which involves promyelocytic leukaemia (PML), c-Jun (previously identified as TGIF1 partner [47]) and TGIF1 [48]. PML is a protein required for the SMAD2 association with SARA and thereafter activation by type I receptor [49]. But for doing so, PML needs to be located in the cytoplasm. However, the complex of c-Jun/TGIF1 can bind and sequester PML in the nucleus and thus avoid SMAD2 activation.

The identification of a PML competitor for TGIF association (PCTA; also known as IRF2BP1) added more complexity to the pathway [50]. PCTA has the ability to compete with cPML for TGIF1. Then, when PCTA is present, PML will be released and the SMAD2 activation will proceed. Globally, this new scheme details a novel role of TGIF1 in the suppression of the TGFβ pathway, inhibiting, not only the gene transcription as it has been studied before [28], but also the activation of the R-SMADs at the beginning of the signalling cascade (Figure 1.7). These studies [44, 48, 50] highlight TGIF1 as one of the main suppressors of the TGFβ signalling pathway.

1.3.3 Roles of TGIF1 in another pathways

The roles of TGIF1 in human cells are not restricted to the TGFβ pathway, but TGIF1 is also involved in the regulation of another pathways. For instance, TGIF1 also binds to Axin-1 or Axin-2, sequestering them

in the nucleus [51]. Axin-1 or Axin-2 are the limiting components for the degradation complex of β-catenin, the final transductor of the Wnt signalling pathway [52]. The function of TGIF1 is to fine tune the levels of β-catenin in the cell for a correct signal response upon Wnt binding [53]. TGIF1 has also been associated with p38 MAPK in the context of proinflammatory phenotype of endothelial cells after irradiation or after TNF-α exposure [54]. Moreover, the binding with c-Jun also facilitates a cross-talk between both pathways [47]. Other functions of TGIF1 include the regulation mechanism of the mouse axial patterning [55] and the inhibition of the retinoid signalling via interaction with the retinoid response elements [56].

Collectively, these discoveries make TGIF1 a key actor in many pathways. Although it has not been explored in detail, TGIF1 could also act as a regulator of the cross talk between different pathways as a transversal transcriptional suppressor.

1.3.4 TGIF1 - SMADs interaction

The discovery of the SMAD2/TGIF1 interaction was performed through a two-hybrid library with a LexA/SMAD2 (100-467) protein. This means that the region of SMAD2 bound to TGIF1 included a fragment of MH1, the linker and the MH2 [28]. However, SMAD2 and SMAD3 are very similar in sequence (83.94% of identity, Figure 1.5) and the antibodies raised against one protein can recognise also the other. Therefore, the results performed with SMAD2 can be extrapolated also to SMAD3.
On the TGIF1 side, the region between the residues 267 and 321, known as SMAD interacting domain (SID), has been proposed to interact with SMAD2 [28]. In Figure 1.8 B is displayed where is this region localised along TGIF1 sequence, together with the other TGIF1 interacting proteins.

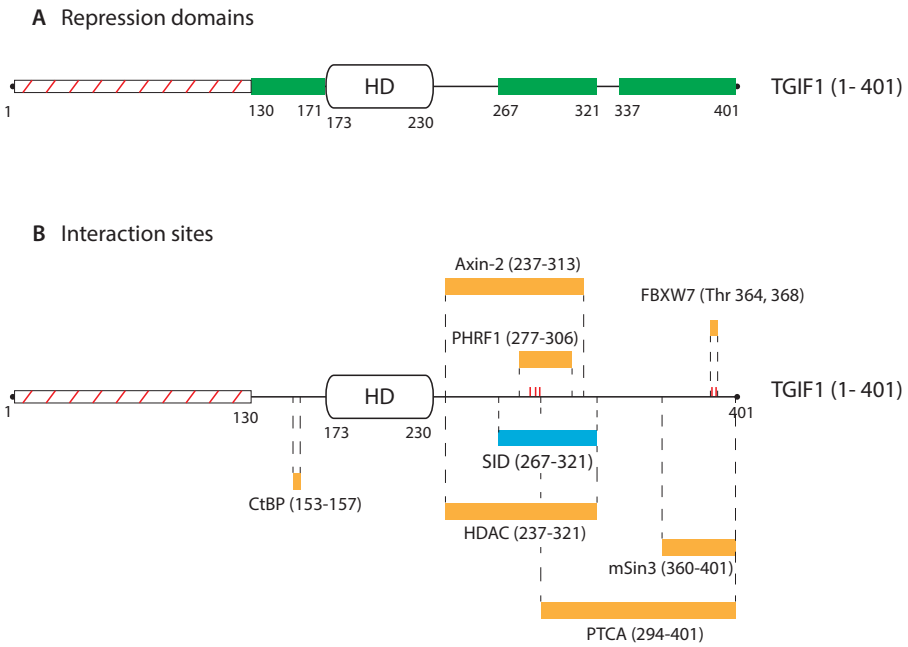


Figure 1.8: Scheme of TGIF1 repression domains (**A**) [44] and interactions sites with partner proteins (**B**). In green are shown the repressor domains, in orange are drawn the regions where the proteins that interact with TGIF1 bind and in blue is highlighted the interacting region with SMAD2. The red lines on TGIF1 sequence represent the phosphorylated residues found in different studies [36, 57]. The first 130 residues are not considered for TGIF1 study.

1.4 Peptide ligation

For more than 100 years chemists have been trying to synthesise the full proteome [58]. The full chemical synthesis of proteins will permit not only the introduction of post-translational modifications (PTM) (including phosphorylation, glycosylation, acetylation, etc. [59]), but also the incorporation of non natural amino acid and specific labels. Overall, it will allow a complete protein manipulation at atomic level opening the door to new strategies regarding protein modification, protein labelling and structure determination [60, 61, 62].

Although it is still a long term objective, many steps in that direction have been already done. In 1963, the development of the Solid-Phase Peptide Synthesis (SPPS) by Merrifield [63] revolutionised the way to synthesise peptides allowing a fast and efficient synthesis on solid supports. Many improvements have been incorporated to the original method such as the development of the Fmoc chemistry, the adoption of enhanced resins and reagents or the use of peptide synthesisers assisted with microwave technology, among others. However, incomplete reactions, accumulations of by products and aggregation of growing peptides are still some of the barriers that limit the synthesis of long peptides chains [64]. Recently, the use of microwave heating synthesisers have been proved to solve some of the aggregations issues, enabling to reach more than 100 residues long polypeptides [65, 66]. However, since the average protein length is about 360 amino acids for Eukarya (270 for Bacteria) [67], it still exists a gap between the main peptide synthesis technique and the protein world. To overcome this limitation, many new techniques have been developed that enable the ligation of two or several peptide fragments to get a longer one.

Generally speaking, there are three basic strategies to ligate two peptides. One involves the use of thiol auxiliary (via cysteine or external thiol group). Another strategy is based on the direct aminolysis between the two peptides, without the requirement of a thiol group. Finally, the last strategy comprise a diverse array of methods that, being orthogonal, and thus chemoselective, to the most of the peptide and biological chemistry, can generate an amide bond between two peptides.

1.4.1 Thiol assisted strategies

In 1989, Kemp and coworkers presented the last part of a new ligation method in which two peptides are bound through a peptide bond [68]. This method established the basis for the thiol based methodology as it was the first demonstration of a chemoselective ligation of unprotected peptide fragments. In this reaction, the 4-hydroxy-6-mercaptodibenzofuran moiety at C-terminus of the first peptide is used to react with the sulfenyl derivate at N-terminus of the second peptide (Figure 1.9 A). The reaction, favourable due to the good leaving group generated, links both peptides. A subsequent acyl transfer reaction creates a native bond between both peptides. Finally, a reduction with phosphine reduction yields the desired ligated product. In addition, the reaction proceed without racemization of the coupled amino acids.

In 1953, Wieland and coworkers demonstrated for the first time a dipeptide ligation between a phenyl thioester valine and a cysteine residue under aqueous conditions [69]. Based on this achievement, Kent and coworkers developed in 1994 the most successful peptide ligation, the Native Chemical Ligation (NCL) [70], between a C-terminal thioester peptide and an N-terminal cysteine peptide (Figure 1.9 B). It starts with the attack of the side-chain cysteine to the thioester carbonyl in a reversible transthioesterification step. A subsequent spontaneous intramolecular $S \rightarrow N$ acyl transfer through a five-membered ring forms the amide bond between both peptides. Interestingly, the reaction proceeds in aqueous medium with all the amino acids unprotected, even other cysteines; an indication of the high selectivity of the reaction. Guanidinium chloride or any equivalent chaotrope is usually added to prevent aggregation during the reaction.

The resultant product, however, ends with a cysteine residue in the ligation junction. As cysteine is one of the less common amino acid in the proteome (between 1-2 %, [71, 72]), it raises as the major limitation. Moreover, back in 1994 there were no methods to prepare peptide thioesters using Fmoc/*t*Bu chemistry. Thus, it was mandatory to use Boc/Bzl chemistry, which prevents the synthesis of phosphopeptides and glycopeptides due its strong acid conditions. Both conditions plus the long reaction time of the ligation reaction (48-72 h) limited the prac-

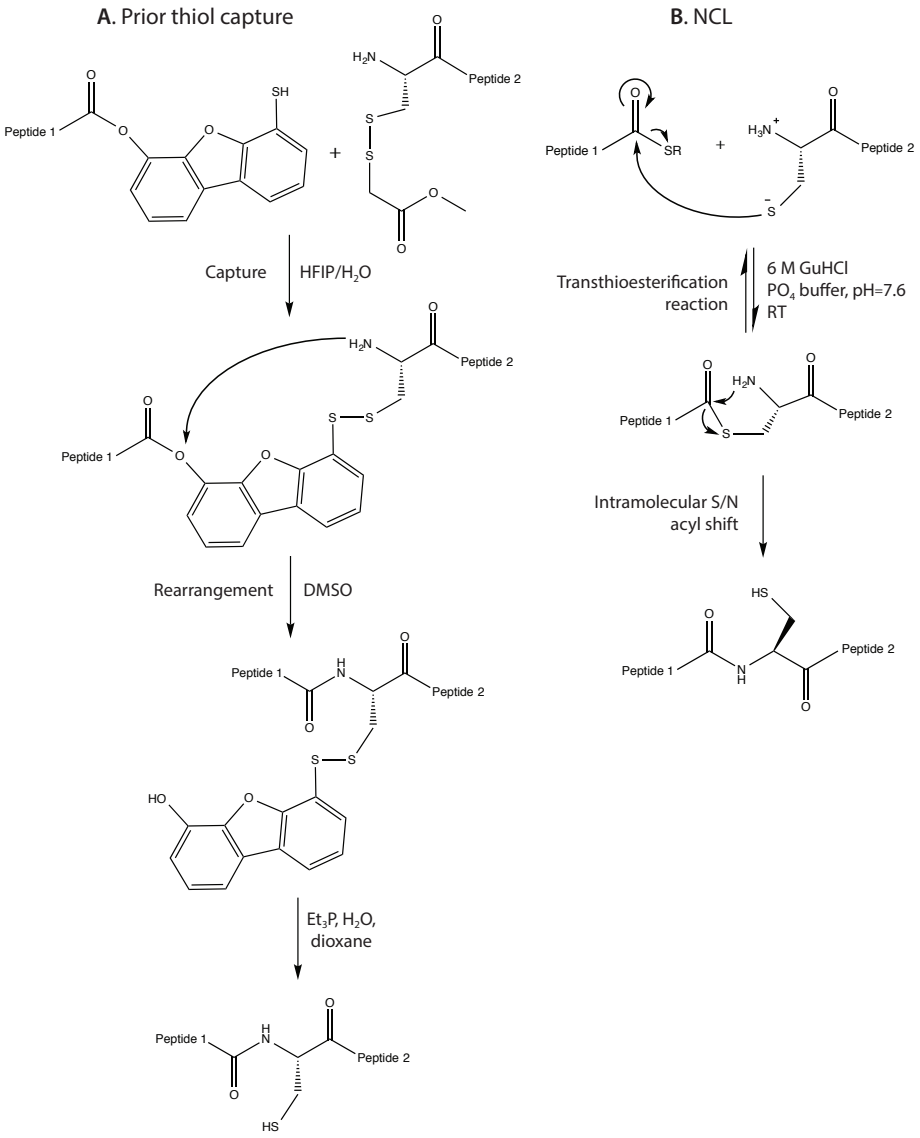


Figure 1.9: Scheme of prior thiol capture from Kemp *et al.* (**A**) and Native Chemical Ligation (NCL) from Kent and coworkers (**B**) peptide ligations strategies.

tical applicability of this new ligation method.

Nowadays, most of the above-mentioned are greatly overcome. The in-

corporation of a thiol catalyst, first thiophenol [73] and later 4-(carboxymethyl)thiophenol (MPAA) [74] speeds up the reaction up to 2 h. More advances include the development of the linker chemistry that allowed the synthesis of thioesters with Fmoc/*t*Bu strategy at high yields [75, 76]. Finally, post-ligation desulfuration processes of the thiol group have increased the versatility of the NCL. First, in 2001, Yan and Dawson explored the desulfuration of the cysteine to alanine with different combinations of metal catalysts [77]. Although the use of metals is inconvenient (peptide absorption on the metal surface, reduction of methionine and thiazolidine-protected cysteines), this work opened the door to the expansion of native residues that can be used at the ligation site. Interestingly, in 2007, Wan and Danishefsky presented a new radical-promoted method for the desulfuration of the cysteine to alanine [78]. This new method requires only tris(2-carboxyethyl)phosphine (TCEP), radical initiator 2,20-azobis[2-(2-imidazolin-2-yl)propane]dihydrochloride (VA-044) and *t*BuSH in water medium to achieve the selective desulfuration of non-protected cysteines, even in the presence of methionines, thioesters or glycopeptides. Nowadays, not only alanine, but also valine, lysine, leucine, threonine or proline can be generated at the ligation junction in metal-free post-ligation reactions. Overall, these innovations make the native chemical ligation method much more attractive.

Since the publication of the Native Chemical Ligation, the peptide ligation field has expanded with new approaches that complement the NCL [79]. The auxiliary-based methods relay on the use of a removable moiety to perform the ligation. This auxiliary group mimics the function of the cysteine and the process proceeds in a similar way as the prior thiol capture from Kemp and coworkers [68]. In general, the auxiliary motif contains one thiol group, which carries out the transthioesterification step with the thioester at C-terminal of the first peptide. Similarly to NCL, an $S \rightarrow N$ acyl shift generates an amide bond between both peptides. Finally, the auxiliary group is removed. This auxiliary group can be linked to the N-terminal [80] or to the side-chain of some amino acids, as sugar-assisted glycopeptide ligation (SAL) [81, 82]. Curiously, Wong and coworkers found inspiration in this reaction to develop the cysteine-free direct aminolysis [83], when they discovered that the re-

action could even proceed in the absence of a thiol auxiliary. However, the multistep auxiliaries preparation and the moiety removal drop the overall yield of the ligation reaction. Moreover, the ligation rate of the reactions is always lower than the original NCL, increasing side reactions such as hydrolysis and epimerization.

With the objective to synthesise longer and more complex proteins, sequentially ligations have been studied. The use of 1,3-thiazolidine-4-carboxo (Thz) group to protect the N-terminal cysteine opened the door to sequential C-to-N ligations. Thz, orthogonal to the main peptide synthesis conditions (including NCL), can be mild removed under methozyamine at pH 4.0, generating an N-terminal cysteine residue ready to react [84]. On the other hand, kinetically controlled ligation (KCL), developed also by Kent and coworkers [85], enables a selective NCL in the presence of a second thioester. It takes advantage of the different ligation speed depending whether the thioester is aryl (faster) or alkyl (slower), controlling at the same time the catalysers added (such as thiophenol or MPAA).

The scope of the NCL was further expanded with the development of the expressed protein ligation (EPL) strategy [86]. EPL takes advantage of the intein proteins to generate a C-terminal thioester that would proceed with an standard NCL (Figure 1.10). The inteins are the force-driving proteins in the enzymatic process known as "protein splicing". Analogue to RNA splicing, the intein protein, surrounded by a C and N-terminal exteins, cleaves itself at the same time that joins both extremes of the exteins. The resulting joined extein does not take part in the process. Remarkably, a mutation of the C-terminal aspartic acid to alanine in the intein protein prevents the final cleave and stops the process in the equilibrium between the thioester and the amide form [87]. Thus, a protein of interest can be expressed in a heterologous system together with the mutated intein. After protein purification, the addition of a thiol agent (as 2-mercapoethanesulfonate (MESNA)) generates the recombinant thioester (now without the intein) ready to react with an N-terminal cysteine of a synthetic peptide or another recombinant protein.
Thanks of the EPL, nowadays it is possible to ligate two recombinant

proteins (or a mixture recombinant-synthetic) via NCL, broadening the ligation options to a new high level [88]. However, as common desulfuration techniques cannot be used (all cysteines are unprotected), there is still the limitation to have a cysteine at the ligation junction. Nevertheless, the use of a photocleavable auxiliary have been already reported [89].

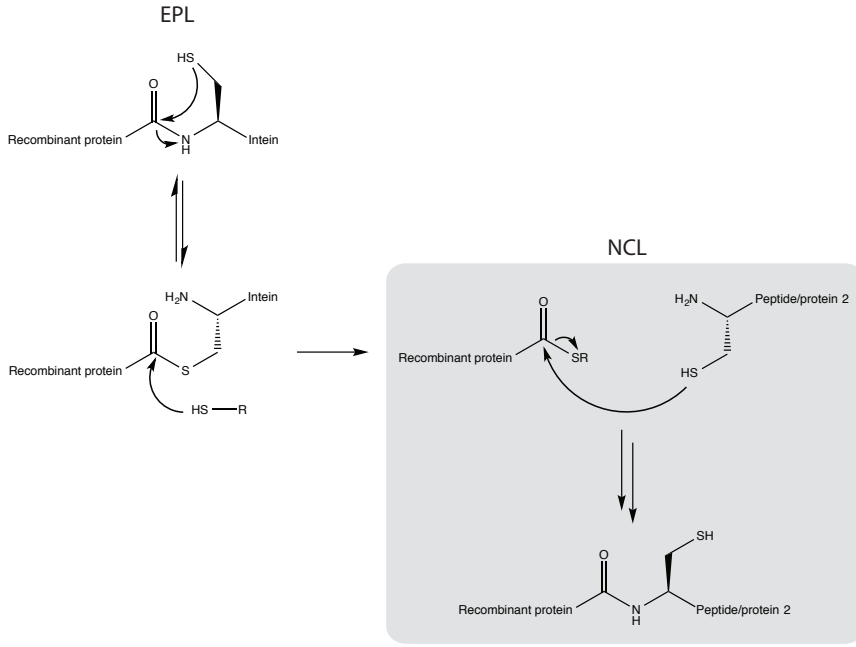


Figure 1.10: Scheme of Expressed Protein Ligation (EPL) from Muir *et al.*.

1.4.2 Direct aminolysis strategies

Since the first successful direct coupling between two peptides described by Kemp *et al.* in the 70s [90, 91], many direct aminolysis strategies have been developed. The first attempts were performed with protected peptides dissolved in organic solvents due their poor solubility in water. However, the reactions suffered from epimerization of the C-terminal α-carbons.

One step forward was the method designed by Blake [92], based on a thiocarboxyl segment condensation strategy in the presence of silver (Figure 1.11 A). The cations of silver selectively activate the carbonyl group at the C-terminal of the thiocarboxylic ending peptide. The activated group reacts then with the N-terminal amino group of the second peptide, binding both peptides through a peptide bond. No protection groups are needed for side-chain carboxyl but they are for the side-chain amino groups. However, the reaction causes the epimerization of the activated thiocarboxyl residue, limiting this protocol to glycine or proline at the C-terminal position.

An improvement of this method was carried out by Aimoto some years later [93] (Figure 1.11 B). The basic modification was the exchange of the thiocarboxyl by a S-alkyl thioester. Because of that, the activation of the carbonyl group is milder and the nucleophilicity is reduced. This reduction allows the presence of various protecting groups, including the acetamidomethyl (Acm) – Cys protecting group, previously discarted due to the side reactions caused by the activated thiocarboxyl group. In addition, he found that the addition of 1-hydroxybenzotriazole (HOBt) or 3,4-dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazine (HOOBt) and *N*,*N*-diisopropylethylamine (DIPEA) to the mixture reaction rises to more active esters, ending in a efficient segment condensation. In his paper, Aimoto highlights the importance of the peptide thioester synthesis, as the lower yield obtained stands to be one of the majors limitations of the method. Other limitations continued to be the racemization at the C-terminal position and the need of the protecting groups for side-chain amines and cysteines.

In 2007, Danishefsky and coworkers continued the development of the ligation method in their research of a reliable strategy for glycopeptides ligations [94] (Figure 1.11 C). In their study, they started from a glycopeptide ending at C-terminus with a phenolic ester bearing a protected *ortho*-thio moiety. Apart from Aimoto's AgCl assisted ligation protocol, they also tried TCEP as a substitute for the Ag cations, obtaining a similar result. Moreover, the use of TCEP was specially interesting for orthogonal reactions. They proved that TCEP only activates acyl donors even in the presence of alkyl thioesters, allowing two

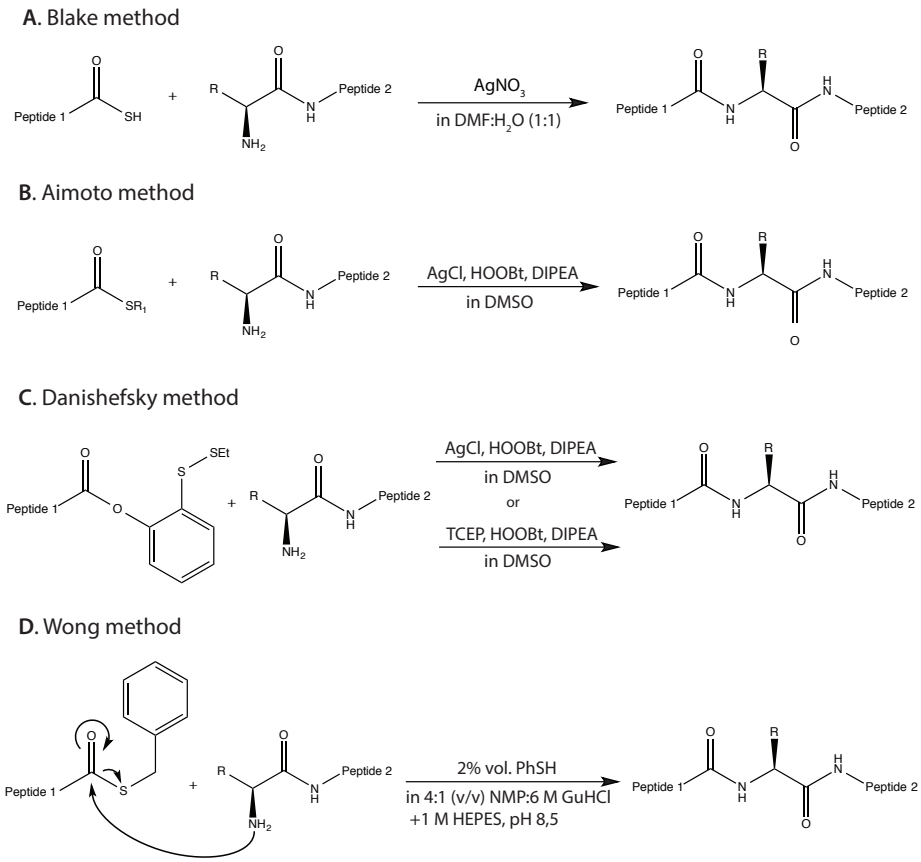


Figure 1.11: Schemes for a direct aminolysis peptide ligation reaction.

step ligation. On the other hand, AgCl activates both at the same time. Nevertheless, the previous limitations were still present. Side-chains from lysine and cysteine residues were protected by 1-(4,4- dimethyl-2,6-dioxocyclohex-1-ylidene)-3-methylbutyl (ivDde) and Acm, respectively. Moreover, although they made some efforts, they could not find any satisfactory condition avoiding the racemization of the C-terminal amino acid. Hence, all the reactions were performed with a glycine or a proline at C-terminus.

One year later, Wong and coworkers published the method of a non-epimerizable cysteine free direct aminolysis [83] (Figure 1.11 D). Based

on the previous works already explained, they focused their efforts on determining the most suitable buffer for a direct aminolysis between a C-terminal thioester and a free N-terminal peptide without causing epimerization at the C-terminal residue neither hydrolysis of the thioester. After several trials, they ended with the ligation buffer 4:1 v/v NMP:6 M GuHCl, 1 M HEPES, pH 8,5, where NMP:*N*-Methyl-2-pyrrolidone, GuHCl: guanidinium chloride and HEPES: 2-[4-(2-hydroxyethyl)-piperazin-1-yl]ethanesulfonic acid. Interestingly, the addition of the peptides prepared in trifluoroacetate salts drops the pH to 7.3-7.6, which actually is the pH that showed less hydrolysis. After the dilution of the peptides in the buffer, 2% thiophenol was added to the mixture in order to generate a better leaving group at the thioester. The reaction was set at 37°C with gentle mixing every 12 h.

The ligation reactions performed showed an absence of epimerization at the C-terminal amino acid. Moreover, the hydrolysis observed was minimal. In order to study the effect of different combinations at the ligation junction, they synthesised several thioesters and free N-terminal peptides only changing the last or first amino acid, respectively. Some of the combinations, specially those that contain glycine at N-terminal of the free peptide achieve a quantitative yield. However, those ligations with a tyrosine at N-terminal of the free peptide ended with a particularly low yield (range 12-39%). The authors also figured out that the N-terminal amino acid of the free peptide seems to be more influential to the reaction yield.

The reaction time varied from 48 to 96 h depending on the amino acids at the ligation junction. However, some additional experiments with cysteine-containing sequences showed that the native chemical ligation reaction is faster. The same cysteine-containing peptide ligations allow the authors to prove the chemoselectivity of the reaction when unprotected cysteine was present in the ligated peptides, as they obtained similar yields with equivalent sequences without cysteines. However, they could not find a way to avoid undesired side-reactions with unprotected lysines. Thus, when lysines are present in the sequence, they must be protected, e.g. by ivDde. Finally, the authors successfully synthesised a native trimeric 60-mer MUC1 glycoprotein taking advantage of the described ligation method.

Overall, this work described a great advance to the final goal of a cysteine-

free direct aminolysis reaction. It shows for the first time a chemoselective cysteine-free direct aminolysis reaction without epimerization at the C-terminal residue. In addition, it has already been used in the synthesis of microcin B17 [95] and polydiscamides B, C and D [96]. Nevertheless, the long reaction times (up to 96 h) and the low yield exhibit by tyrosine remain the main challenges to be addressed. Moreover, other combinations at the ligation junction, especially those with relevant difficulty due to sterically hindrance reasons (e.g. valine or leucine) [97], need to be studied.

1.4.3 Other peptide ligation strategies

Besides the thiol-assisted and thiol-free peptide ligation reactions, alternative strategies to achieve a native chemical bond between two peptides have been developed [98].

Bode and coworkers presented in 2006 a new strategy to ligate two peptides by the condensation of α-keto carboxylic acids and N-alkylhydroxyalamines (KAHA) [99]. Theoretically applicable to any residue par at the ligation junction, this reaction proceeds in mild conditions and requires no additional reagents or catalyst generating only carbon dioxide and water as by-products (Figure 1.12 A). Moreover, the reaction is "absolute" chemoselective and thus can be performed with the fully unprotected peptides. Furthermore, KAHA is completely orthogonal to NCL. The main limitation of the otherwise perfect peptide ligation method is the generation, precisely, of the α-keto carboxylic acid and N-alkylhydroxy-alamines peptides. The first have been nicely solved by the synthesis of the peptide on sulfur ylide linker [100]. After a fast treatment with oxone, the peptide ends with α-keto carboxylic acid. However, the obtention of the N-alkylhydroxyalamines peptides has been more problematic [100]. Nevertheless, in 2015, Pusterla and Bode managed, using an oxazetidine amino acid strategy, to synthesise S100A4 (metastasin), a remarkably difficult molecule to be synthesised by NCL due to its sequence and properties [101].

Another approach is based on the Staudinger-Meyer reaction between azides and triaryl phosphanes to form iminophosphoranes [102]. This

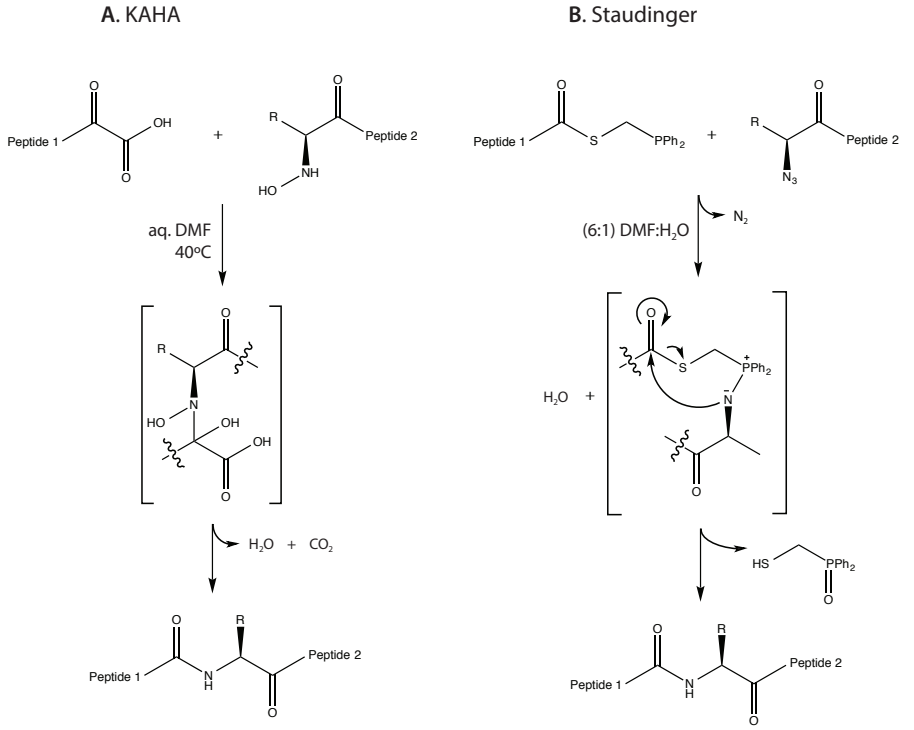


Figure 1.12: Scheme of KAHA from Bode and coworkers (**A**) and traceless Staudinger from Bertozzi and Raines and coworkers (**B**) peptide ligations.

reaction is performed in mild conditions, almost quantitatively, generating only molecular nitrogen as by-product. More than 80 years later, in 2000, Bertozzi and Raines and their coworkers – independently – started a new era of Staudinger reaction [103, 104]. In the non-traceless ligation, the phosphane ligand designed by Bertozzi, acts as an electrophilic trap containing an ester moiety. After reacting with the azide, it proceeds with an intramolecular cyclization generating an amide bond. Just 3 months later, a modification of the phosphine ligand led to traceless Staudingder ligation where the phosphane oxide moiety is cleaved during the hydrolysis (Figure 1.12 B) [105]. In parallel, Raines and coworkers first applied this concept in the peptide ligation [104]. Since then, this reaction has been applied elsewhere in the chemical biology field because of their orthogonally from *in vivo* reactions, including biotin or fluorescence labelling or DNA conjugation, among many

other examples [106]. Peptide ligation application, however, has been restricted to glycine amino acid at the ligation junction as it has not been found a phosphinothiol that proceed with the reaction at efficient rates with other nonglycyl amino acids [107].

Nowadays, click chemistry has became a common term in chemical biology field. The term, coined by Sharpless [108], refers to those reactions that, achieving high yields, are chemoselective to the biological reactions and thus could be used *in vivo* environment without side reactions. The already presented Staudinger reaction is considered one "click" reaction. Nevertheless, the more famous "click" reaction is the Huisgen 1,3-dipolar cycloaddition of azides and acetylenes to give 1,2,3-triazoles. This reaction, catalysed by Cu(I), ligates two peptides through a non native bond. Although it is not a native amide bond, the structure adopted is very similar to the amide moiety. Plus, triazoles are more stable against proteolytic degradation than native bonds [109]. Furthermore, its chemoselectivity and easy manipulation makes this reaction widely employed.

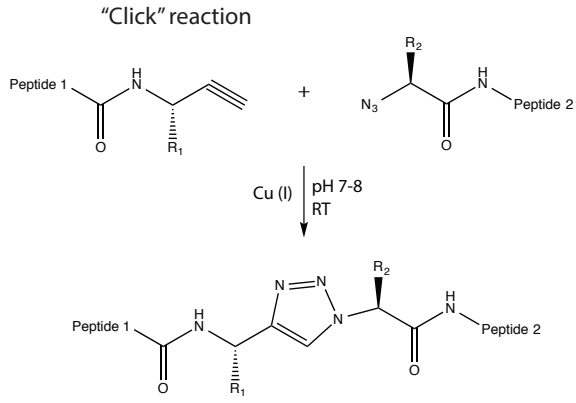


Figure 1.13: Scheme of "click" reaction, Cu(I)-catalyzed Huisgen 1,3-dipolar cycloaddition.

Finally, a completely different strategy for ligate two peptides is being developed. This novel approach is based on the use of enzymes to catalyse an efficient peptide ligation. A recent study presented peptiligase, a

modified mutant of peptide amide hydrolyse subtilisin BPN', that could ligate in high yields a carboxamidomethyl ester (cam-ester) C-terminal peptide with a free N-terminal peptide in water conditions [110].

2 Aims and objectives

2.1 Characterisation of the interaction between TGIF1 and SMAD proteins

The aim of our project was to elucidate the mechanism of interaction between TGIF1 and SMAD proteins from an structural point of view. We first decided to focus on the interaction between TGIF1 and SMAD2 as it is the most studied interaction. In this sense, we focussed our attention on the TGIF1 region (267-321). This region, named SMAD interacting domain (SID), has been suggested to be essential for SMAD2 binding [28]. Interestingly, mass spectrometry high-throughput analysis have revealed that three serines (286, 291, 294) localised in the SID are phosphorylated in human embryonic stem cells [57, 111]. In addition, two mutations (Thr280 and Ser291), also located in the middle of the interacting region, have been identified to cause HPE in humans [29]. Moreover, this region also is essential for the interaction between TGIF1 and HDAC1 [44], Axin-2 [51] and PHRF1 [38], plus it has been identified as one of the repressor domains of TGIF1 [44] (Figure 1.8).

Globally, the region between the residues 267-321 of TGIF1 seems to be a general region for protein-protein interaction. In addition, the phosphorylated serines detected suggest a regulatory mechanism that could govern the different interactions.

Our interest about TGIF1 structure also included the characterisation of the homeodomain region. Recent findings involving a similar protein, HOXC9 (18.3% identity with TGIF1, Figure 1.5) suggest that TGIF1 could also interact with SMAD4 through a preserved region rich in arginines at the N-terminus of the homeodomain. We then planned to characterise how the complex between SMAD2, SMAD4 and TGIF1

might be defined.

We also considered if there are certain intramolecular interactions in the TGIF1 protein that could contribute to define its global structure and to regulate its binding properties.

2.2 Study about the cysteine-free direct aminolysis ligation reaction

Up to date, there are numerous published ligation strategies enabling synthesis of long polypetides, in some cases even proteins [79]. In this study we focused our efforts on the direct aminolysis ligation strategy between a C-terminal thioester peptide and a free N-terminus peptide. This strategy has the advantage to ligate two peptides with any combination of amino acids at the ligation junction. Our project is based on the previous study done by Wong and coworkers [83], where they reported a direct aminolysis reaction with no detectable epimerization on a aqueous buffer at room temperature. Although several combinations at the ligation junction were successfully achieved, in the cases where sterically hindered amino acids, such as aromatic or β-branched, were present at the ligation junction, additional optimisation were required. Moreover, the main limitation of this method is the slow ligation rate in some reactions (up to 96 h), especially in those where sterically hindered amino acids are present.
In this context, the aim of this project was to improve the rate and the conversion of ligation reactions, at the same time that we check the applicability of the reaction using β-branched amino acid. With that objective in mind, we thought that converting the thioester in a better leaving group should help us to achieve our aim. HOBt is a reagent widely used in the SPPS for its ability to generate active esters because the low pKa of HOBt favours the formation of a good leaving group. Therefore, we decided that the addition of HOBt to the ligation buffer would be the ideal choice.
Our idea here was to define a method to prepare phosphorylated fragments of TGIF1 for the binding assays with SMAD2.

2.3 Determination of six mutant FBP28-WW2 structures

As a side project and to learn the basics of NMR assignment, we were also interested in the structural determination of several WW domain mutants, designed to clarify the folding and unfolding properties of this domain. We also wanted to understand the key contacts that maintain the domain folded, using several constructs containing deletions at the domain termini.

2.4 Thesis objectives

Therefore, the objectives of this thesis were:

1. Investigate the interactions between SMAD2 and TGIF1 using mainly NMR and other biophysical techniques.
2. Analyse the effect of serine phosphorylation (at positions 286, 291, 294) in the regulation of TGIF1 and SMAD2 interactions.
3. Determine whether TGIF1 and SMAD4-MH1 also interact with one another.
4. Characterise potential interactions between the two fragments of TGIF1, to detect the presence of open and close intramolecular conformations.
5. Improve the methods of direct aminolysis for the ligation of peptides.
6. Determine the structure of six selected mutants of FBP28-WW2 domain.

3 Materials and Methods

3.1 Chemistry

In this section, the methods for peptide synthesis and purification are explained.
Following the standard protocols, peptides were synthesised using Solid-Phase Peptide Synthesis (SPPS) [63] technique. The peptides were afterwards purified through Reverse Phase-High Performance Liquid Chromatography (RP-HPLC), if required. The characterisation of the products was performed by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI-MS) and Liquid Chromatography Mass Spectrometry (LC-MS) techniques.

3.1.1 Solid-Phase Peptide Synthesis

Basic principles

Initially developed by Merrifield in the 60s, Solid-Phase Peptide Synthesis (SPPS) has nowadays become the main strategy to synthesise peptides without the use of living beings. The chemical synthesis of the peptide is achieved through a repetitive loop consisting on: (1) amino acid coupling, (2) washing and (3) deprotection steps carried on an insoluble support of porous resin beads where the growing chain of amino acids attaches. The anchorage of the peptide to the solid support allows the use of several equivalents of coupling reagents because after the reaction the peptide stays and the excess of reagents is removed by simple washing and filtration.

When the first α-amino protected amino acid is added, it binds to the linker of the resin via its activated carboxyl group. After one washing step, the temporary α-amino protecting group of the amino acid is re-

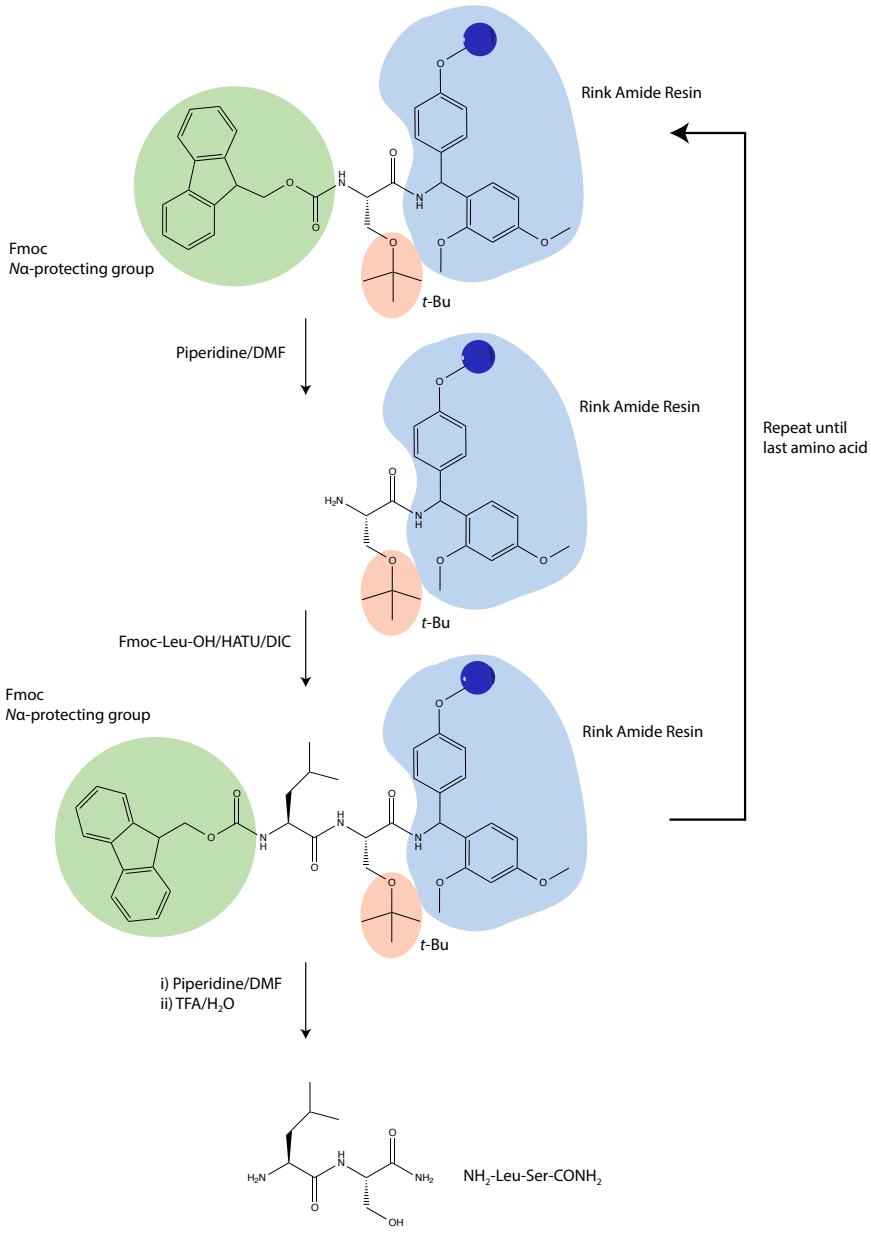


Figure 3.1: Scheme of Fmoc/*t*Bu SPPS.

moved. By adding and activating the next α-amino protected amino acid a peptide bond is formed between the carboxyl of the new amino acid and the free *N*-terminus from the amino acid attached on the resin (Figure 3.1). The repetition of this process proceed in a $N \leftarrow C$ way until the last amino acid of the sequence is incorporated. Finally, the peptide is cleaved from the resin together with the side-chain's protecting groups of the amino acids, yielding the peptide of interest.

SPPS allows the full chemical synthesis of a peptide in a fast and an efficient fashion. The versatility of the technique simplify the incorporation of phosphorylated or non-natural residues, in contrast with heterologous systems. However, despite all advantages that SPPS possess, it faces some limitations as well. For instance, incomplete coupling reactions generates byproducts difficult to purify and aggregation of the growing peptides during the synthesis drops the efficiency of the couplings. Overall, it is commonly accepted that standard SPPS has a limit of 40-60 amino acids long [64].

In order to avoid side reactions, SPPS uses two kinds of protecting groups. While the α-amino group of the incoming amino acid has a temporary protection – removed in each cycle – , the side-chain protecting groups and the link between the peptide and the resin are permanent and resistant to the chemistry used to cleave the temporary protection group and therefore are only cleaved in the final step. Following this idea, two main strategies have been developed: Boc/Bzl and Fmoc/*t*Bu (Table 3.1).

Boc/Bzl relies on the different behaviour of the protectors groups against a distinct acid condition. The *tert*-butoxycarbonyl (Boc), used for temporary protections, is cleaved in the presence of trifluoroacetic acid (TFA). In contrast, benzyl (Bzl) based protecting groups are stable under TFA conditions and they require the more strong acidic environment provided by HF to be cleaved from side-chain moieties and resin linkage. Although this method has high efficiency [112], the hazardous nature of HF and the special equipment needed to handle it led to develop an alternative.

This full orthogonal alternative, Fmoc/*t*Bu system, was developed by Carpino in 1972 [113]. It takes advantage of the fact that 9-fluorenylmethoxycarbonyl (Fmoc) – protecting α-amino group – needs base conditions (piperidine) to be cleaved while *tert*-butyl – protecting side-chain – and the linkage between the C-terminal residue to the resin matrix require acid (TFA). As this strategy needs milder cleavage conditions than Boc/Bzl, most labs have adopted it. All the peptides in this thesis have been synthesised using Fmoc/*t*Bu system and further in this thesis the explanation will only focus on this strategy.

Table 3.1: Main differences between Boc and Fmoc strategies.

	Boc/Bzl	Fmoc/*t*Bu
α-amino protecting group	Boc	Fmoc
Cleavage α-amino protecting group	TFA	Piperidine
Side-chain protecting group	Benzyl	*tert*-butyl
Cleavage side-chain protecting group	HF	TFA
Cleavage from the resin linker	HF	TFA

Resins and Linkers

The resin is the solid support where the peptide attaches during the synthesis. It exhibits (1) chemical and mechanical stability; (2) high swelling in several solvents, allowing optimal access for reagents; (3) uniform and functionalised beads that permits the binding of the linker and (4) good loading [114]. Polymeric resins fulfill these requirements. Merrifield 1% divinylbenzene cross-linked polystyrene (PS) resin was the first resin widely adopted in the field. Inexpensive to produce, it allows an optimal diffusion of the most used solvents in peptide synthesis. This resin has been used extensively although it was found that the incorporation of some amino acids decreases when the peptide length increases. To solve this problem, polyethylene glycol (PEG) was introduced in many combinations (TentaGel®, ArgoGel®). Nowadays, total PEG resin such as ChemMatrix® (CM) are used, achieving superior performance, especially when long peptide sequences need to be synthe-

sised [115].

The linker is the chemical moiety that connects the resin with the C-terminal of the first amino acid of the sequence, keeping the carboxyl group protected. The wide chemistry of the several linkers available explains the difference in the C-terminal ending of the peptide after the final cleavage. For example, Wang resin or 2-chlorotrityl chloride resin linker provides a free peptide ending in acid group; Rink amide resin linker yields peptides with amide function at their C-terminal part while 4-sulfamylbutyryl resin linker gives – through a "safety-catch" linker – a peptide thioester.

Reagents and reactions

N,*N*-Dimethylformamide (DMF) and dichloromethane (DCM) are the main solvents used in SPPS. They solubilise all the reagents involved in the reaction such as Fmoc amino acids, coupling reagents, piperidine, etc..

Activation of the carboxy terminal of the Fmoc-protected amino acid is required for an efficient binding with the growing peptide. Carbodiimide based reagents, such as *N*,*N*'-diisopropylcarbodiimide (DIC), react, as electrophiles, with the carboxyl acid group to obtain an ester bond, that is also a good leaving group. However, carbodiimides are so reactive that can cause side reactions such as racemerization through the formation of a symmetrical anhydride. To avoid them, an auxiliary nucleophile as 1-hydroxybenzotriazole (HOBt) is added, generating active, but milder, OBt esters. It suppress the racemerization, and at the same time improves the rate of the reaction. An alternative of it is ethyl 2-cyano-2-(hydroxyimino)acetate (trade name Oxyma Pure), which has the advantage of being not explosive when is fully dehydrated. Recently, in situ coupling reagents were developed to reduce racemization and increase the rate of the reaction. Families of compounds as acylphosphonium (benzotriazol-1-yloxytris(dimethylamino)phosphonium hexafluorophosphate, BOP; benzotriazol-1-yloxy tris(pyrrolidino)phosphonium hexafluorophosphate, PyBOP®) and acyluronium/aminium salts (*N*-[(dimethylamino)-1*H*-1,2,3-triazolo[4,5-*b*]pyridin-1-ylmethylene]-*N*-

methylmethanaminium hexafluorophosphate *N*-oxide, HATU; *N*-[(1*H*-benzotriazol-1-yl)(dimethylammo)methylene]-*N*-methylmethanaminium hexafluorophosphate *N*-oxide, HBTU) have become very popular in the labs around the world. They convert efficiently the Fmoc-protected amino acid into an active OBt ester in the presence of a base with no nucleophile component, as *N*,*N*-diisopropylethylamine (DIPEA) (Figure 3.2).

The standard procedure includes adding several equivalents of the Fmoc-amino acid and its activators in order to achieve the completeness of the reaction. The coupling reaction time depends on the reagent used, the amino acid that is going to be coupled and the length and aggregation of the peptide. As a general rule, the combination HATU/DIPEA takes about 30 min to complete the reaction, whereas DIC/HOBT takes 90 min.

In the deprotection step, the Fmoc group is eliminated via base induced β-elimination. The most common agent is piperidine that is diluted at 20% in DMF.

The final cleavage of the peptide from its solid support is achieved by applying acidic conditions. Usually TFA at different concentrations (varying from 82 to 97%) is used, which also removes all side-chain protecting groups. During the removal of the side-chain protecting groups there is a formation of highly reactive cationic species. Therefore, different type of so-called "scavengers" at the concentrations from 3% to 18%, are added together with the TFA in order to prevent side-reactions between the cleaved side-chain protecting groups and the peptide sequence. The scavengers used depends on the protecting groups it needs to react with. Combinations of water, triisopropylsilane (TIPS), 1,2-ethanedithiol (EDT) and other thiols or phenol are commonly used. An special case is the cleavage from the 4-sulfamylbutyryl resin linker, that uses the "safety-catch" technique, reported by Backes and Ellman [75], as a modification of previous Kenner's sulfoamide "safety-catch" [116]. This linker is stable under base conditions making it compatible with Fmoc/*t*Bu strategy. The activation proceeds with the reaction of iodoacetonitrile with DIPEA. In a second reaction, benzylmercaptan is

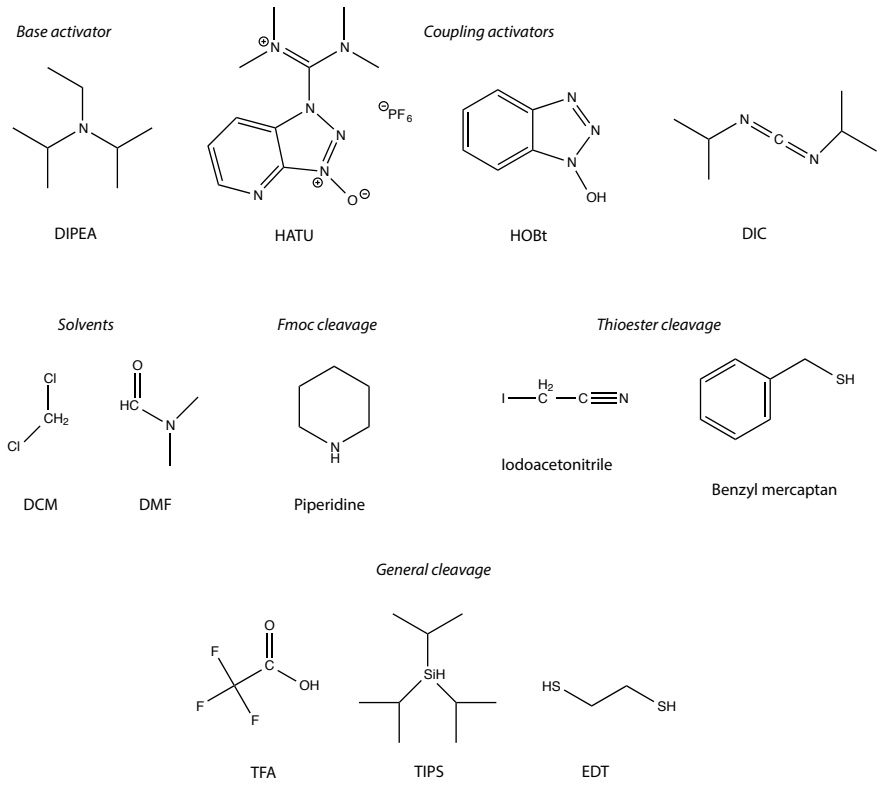


Figure 3.2: Chemical structure of the main reagents used in Fmoc/*t*Bu SPPS.

added, cleaving the peptide from the resin in a thioester form (Figure 3.3).

Quality test

Resin load determination is a test that measures the quantity of Fmoc release after a treatment with piperidine and thus, gives an exact value of the coupling efficiency. If it is done after the first coupling it provides the real quantity of sites (equivalents) able to generate a peptide.
However, for the rest of the couplings, the completeness of the reaction is checked by others methods. Ninhydrin test, conceived by Kaiser [117], examines the presence of free primary amines. If still some free amines are present in the analysed beads, the otherwise light yellow

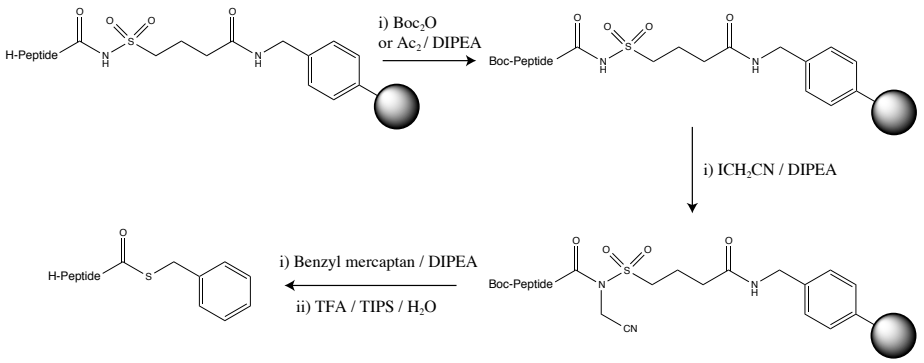


Figure 3.3: Scheme of cleavage of a peptide thioester from a sulfamylbutyryl resin. Prior the reaction with iodoacetonitrile, the N-terminus is protected with Boc_2O (temporary protection) or Ac_2 (permanent protection).

solution turns dark blue. However, the solution may adopt another colours if the last amino acid is serine, asparagine or aspartic acid. Moreover, this detection method has problems detecting secondary amines, what happens when the last amino acid is proline. In this case, the use of the chloranil test is recommended [118].

3.1.2 Peptide Purification

Reverse-Phase High Performance/Pressure Liquid Chromatography (RP-HPLC) is the most common method to purify the crude peptide from its impurities. The peptide is first dissolved in an aqueous solution with a minimal percentage of acetonitrile (MeCN) and then injected in a HPLC with a preparative RP-column. An acetonitrile gradient, typically from 5 to 60%, is applied while the elution is monitored at 220 nm (peptide bond). TFA is often added to the solvents as ion pair reagent to improve resolution and selectivity. The eluted fractions are collected and analysed by MALDI-MS.

3.1.3 Experimental procedures

As previously described in Biopolymers (Peptide Science) under the title ”Addition of HOBt improves the conversion of thioester-amine chemical ligation” [119].

General peptide synthesis

All manual synthesis were done in a polypropylene syringes with fitted porous filters, allowing a removal of the liquid by a vacuum pump. At the end of each day, resins were dried and stored at 4°C ON, swelling them again with DMF for 30 min at the beginning of the day. The amino acids were incorporated by using 5 eq of the corresponding Fmoc-amino acid derivatives activated with 4.9 eq of DIC in the presence of 4.9 eq of HOBt in DMF for 90 min. All the couplings were done with mechanical stirring at room temperature, unless otherwise stated. After each coupling the resin was washed 6 times with DMF (1 min) and 2 times more with DCM (1 min) before the Kaiser or chloranil test was performed. The efficiency of the couplings was verified with the Kaiser - or chloranil if the coupled amino acid was proline. If the Kaiser test was positive, recoupling was performed with 5 eq of Fmoc-amino acid, 4.9 eq of HATU and 10 eq of DIPEA in DMF for 50 min. When the use of the test was not enough, a mini cleavage was performed. It consist in a full cleavage of a very little quantity of resin and afterwards an MS analysis. In all cases, the Fmoc was release adding 40% piperidine in DMF for 5 min, and then 20% piperidine in DMF for 5 min. Both reactions were done at room temperature.

Synthesis of the peptides thioesters

Peptide thioesters were synthesised using the Fmoc/*t*Bu strategy and 4-sulfamylbutyryl aminomethyl (AM) resin (substitution: 0.73 mmol/g), at 100 μmol scale. For the coupling of the first residue, Fmoc-amino acid (4 eq), DIC (4 eq) and 1-methylimidazole (4 eq) were dissolved in DCM (12 μL/μmol). The solution was added to 1 eq of the resin placed in a glass tube. The mixture, sealed with a septum, was gently agitated for the next 18 h at 25°C. Once the reaction is over, the resin was transferred to the polypropylene syringes and then washed 5 times with DMF and 5 times with DCM before being dried. Coupling efficiency was estimated through resin load determination by UV analysis of Fmoc-release. The remaining amino acids were coupled as previously described.
After completion of the sequences, all peptides were either acetylated (with 5% acetic anhydride, 8.5% DIPEA and 86.5% DMF (in vol.), 30 min stirring at room temperature) or Boc-protected (with 20 eq of Boc_2 and 5

eq DIPEA in DMF, 2h stirring at room temperature) at the *N*-terminus prior to final peptide cleavage from the resin. The procedure for the linker activation was the following: iodoacetonitrile (67 eq) and DIPEA (13 eq) in DMF (72 μL/μmol) were added to the resin and gently agitated for the next 18 h. The resin was then washed six times with DMF, six times with DCM and then dried. In the next step, the peptide was cleaved from the activated resin using benzyl mercaptan (50 eq) and DIPEA (13 eq) in DMF (72 μL/μmol) in an 18 h reaction with gentle agitation. The peptide solutions were filtered and collected in a round bottom flask. The resin was washed twice with DMF (4 mL each time) and the combined filtrates were concentrated under vacuum in order to completely remove DMF. The crude product was dissolved in 6 mL of a TFA/TIPS/water mixture (95:2.5:2.5 by vol.) and stirred at room temperature for 3 h, yielding the peptide thioester without the side-chain protecting groups. The peptide thioesters were then precipitated in cold diethyl ether and centrifuged (3,000 g). The pellet was washed twice with cold ether, dried, and stored at -20°C. The crude thioesters were analysed by LC-MS and, depending on their purity, some of them were additionally purified by preparative RP-HPLC prior to their use in the ligation reactions.

Synthesis of free *N*-terminus amide peptides

Peptides containing a free *N*-terminus were synthesised manually using a Rink amide AM resin (substitution: 0.68 mmol/g) at 100 μmol scale with Fmoc/*t*Bu chemistry. The coupling was done as previously described.

The resin-bound peptides were cleaved and deprotected with TFA containing a scavenger mixture of water, thioanisol and EDT (90:5:2.5:2.5 by vol) at RT for 2 h. Afterwards, the peptides were then precipitated in cold diethyl ether and centrifuged (3,000 g). The pellet was washed twice with cold ether, dried and stored at -20°C. Finally, the crude peptides were purified by preparative RP-HPLC. Pure fractions were collected, lyophilised, and stored at -20°C. They were then analysed by MALDI-MS and LC-MS prior to their use in ligation reactions.

Peptide purification

Crude peptides, peptide thioesters, and ligation mixtures were purified using an Aqua C_{18}-column (internal diameter 4.6 mm, length 150 mm, particle size 5 μm, pore size 12.5 nm, Phenomenx) with a linear gradient from 10 to 24% aqueous acetonitrile (0.1% TFA) over 1.2 min, followed by a gradient of 24 to 57% for the next 42 min with a flow rate of 1 mL/min using an ÄKTA purifier 10 HPLC System (GE Healthcare, Uppsala Sweden). Fractions were analysed by MALDI-MS and those containing the products were collected, lyophilised, and stored at -20°C. Ligation mixtures were analysed on the microbore Aqua C_{18}-column using a linear gradient from 4.75 to 57% or from 9.5 to 57% aqueous acetonitrile (0.1% TFA) for 25 min at a flow rate of 1 mL/ min.

Peptide ligation

The free N-terminus peptides (1.5 eq or 0.5 eq) were dissolved in 120 μL of ligation buffer (*N*-Methyl-2-pyrrolidone (NMP) : 6M Guanidinium chloride (GuHCl) + 1 M 2-[4-(2-hydroxyethyl)piperazin-1-yl]eth-anesulfonic acid (HEPES) at 4:1 (v/v). A buffer containing 6 M GuHCl and 1 M HEPES was prepared (50 mL) and adjusted to pH 8.5 using 25% NaOH solution and degassed. For each ligation trial 100 μL of the buffer was mixed with 400 μL of NMP). The resulting solution was used to dissolve the thioester peptide (between 0.3 and 1.5 μmol). Depending of the peptide thioester amount, the final volume of the ligation buffer was different (the volume of 80 μL and the described procedure is for peptide thioester amount of 1 μmol). However, the free NH_2 peptide was always 1.5 eq higher (Table 4.7, combination I, II, V, VI, VII, VIII and X) or 0.5 eq lower (Table 4.7, combinations III, IV and IX) from the peptide thioester equivalents). For each reaction, the solution was separated into two equal parts, and 15 μL of HOBt, dissolved in the same ligation buffer (2 eq based on the amount of peptide thioester), were added to the ligation series with HOBt, while 15 μL of the ligation buffer was added to the other part, for comparison, as a control. For the ligations, 2 eq of with 1-hydroxy-7-azabenzotriazole (HOAt), 4 eq of 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) or 4 eq of DIPEA were added to the ligation mixture, respectively. Finally, thiophenol (2% by volume, 1.5 μL) was added to the reactions in the presence or absence of activa-

tors/bases, and the resulting mixture was incubated at 37°C with gentle agitation until the reaction was completed. At various time points the reaction was followed by LC-MS, and in some cases by MALDI-MS and RP-HPLC, depending on the peptide sequences and amino acids at the ligation junction (Table 4.7). At each time point, 8 μL aliquots of the ligation mixture were taken and quenched by the addition of 0.1% TFA in water (32 μL) or tris(2-carboxyethyl)phosphine (TCEP) solution (32 μL, of a 10 mg/mL solution) when the products contained cysteine residues. The quenched mixtures were diluted up to 1.5 mL with % MeCN/0.1 % FA in water and stored at -20°C.

Reagents

Acetic acid (ACS reagent >99.7%), NaOH (Bioxtra, >98%), 1-methylimidazole (>99%), benzyl mercaptan (>99%), HOBt, DBU, and thiophenol (97% GC) were purchased from Sigma-Aldrich Chemie GmbH (Taufkirchen, Germany), TFA (HPLC grade, >99,5%) from Alfa Aesar (Ward Hill, Massachusetts, USA). Formic acid (FA, >98%), DIC (>98%) and NMP from Fluka (Sigma-Aldrich Chemie GmbH). 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, >99.5%), guanidinium chloride (GuHCl, >99%) from Melford Laboratories Ltd. (Ipswich, United Kingdom). Acetonitrile (for HPLC PLUS gradient-ACS-Reag. Ph. Eur-Reag. USP) and DMF (for peptide synthesis) were purchased from Carlo Erba Reagents (Milano, Italy), while DIPEA (for synthesis) and DCM (99.9%) from Merck KGaA (Darmstadt, Germany). HOAt was purchased from Medalchemy (Alicante, Spain). Standard Fmoc-amino acid derivatives, supplied by Iris Biotech GmbH (Marktredwitz, Germany), containing tert-butyl protecting groups at serine, threonine and tyrosine, the 2,2,4,6,7-pentamethyl-dihydrobenzofuran-5-sulfonyl (Pbf) protecting group for arginine, the Boc protecting group for tryptophan, and 1-(4,4-dimethyl-2,6-dioxocyclohex-1-ylidene)isovaleryl (ivDde) protecting group for the side-chain amine of lysine. Rink amide AM resin (200-400 mesh) and 4-Sulfamylbutyryl AM resin were purchased from Novabiochem (Merck, Spain). Water was purified in-house with a Milli-Q Advantage A10 Ultra Pure Water Purification System (Millipore, Merck, Spain).

3.2 Biology

In this section the molecular biology methods are introduced. As most structural biology techniques require large amounts of highly pure and concentrated protein, express the protein in some of the different heterologous systems available and afterwards purify it is the common way to obtain it. Nowadays, the most used system is the heterologous expression in bacteria, and specifically in *Escherichia coli* (*E.coli*). The choice of this organism provides the user plenty of advantages:

- It allows an easy insertion of the DNA sequence of the protein of interest (compared to different expression systems, e.g. in insects).
- It permits a high expression of the protein at low cost (especially important with ^{15}N or ^{13}C labelled samples that are necessary for NMR experiments).
- It grows fast in a high density cultures.
- *E.coli* metabolism and genetics are well known.

Several *E.coli* strains, each of them with a plethora of genetic modifications, are available. Every modification bestows unique characteristics to *E.coli*, such as protease deficiency (avoiding degradation of the newly expressed protein of interest), T1 Phage Resistant (which gives *E.coli* resistance against a virulent phage) or DE3 (containing the T7 RNA Polymerase for induction with the T7 promoter), among many others.
In this thesis, three main strains of *E.coli* were used. DH5-α, with a very high efficiency in transformation, was used for cloning while the strains BL-21(DE3) and Rosetta were used for protein expression. Rosetta differentiates from BL-21(DE3) since it has an extra plasmid that contains rare codons tRNA. If the expressed sequence has codons that are not normally used by *E.coli*, Rosetta may give better expression than BL-21. Otherwise, it is more recommended the use of BL-21 as it grows better because it lacks this second plasmid.

3.2.1 Cloning

The *E.coli* strain expresses the protein that is encoded in the vector we have inserted in the bacterial cell. The vectors are circular DNA plasmids of about 5-7 KDa long. They contain, at least, all the genetic information related to: the construct to be expressed and its promoter region, the antibiotic resistance (that keeps the plasmid in the bacteria cells through positive selection) and its promoter region, the origin of replication (which allows the replication of the plasmid) and the multiple cloning site (MCS) (for an easy insertion of the construct of interest). In order to get the required vector, cloning method may be used. This method consist of the amplification of the desired insert (usually obtained from another vector or from a cDNA library) and then the ligation (or recombination) of it with the chosen vector. After the ligation, DH5-α cells are transformed with the ligated vector. Some colonies are then selected to be sequenced. Only the ones that gives the correct sequence are stored at -20ºC to be used for further experiments.

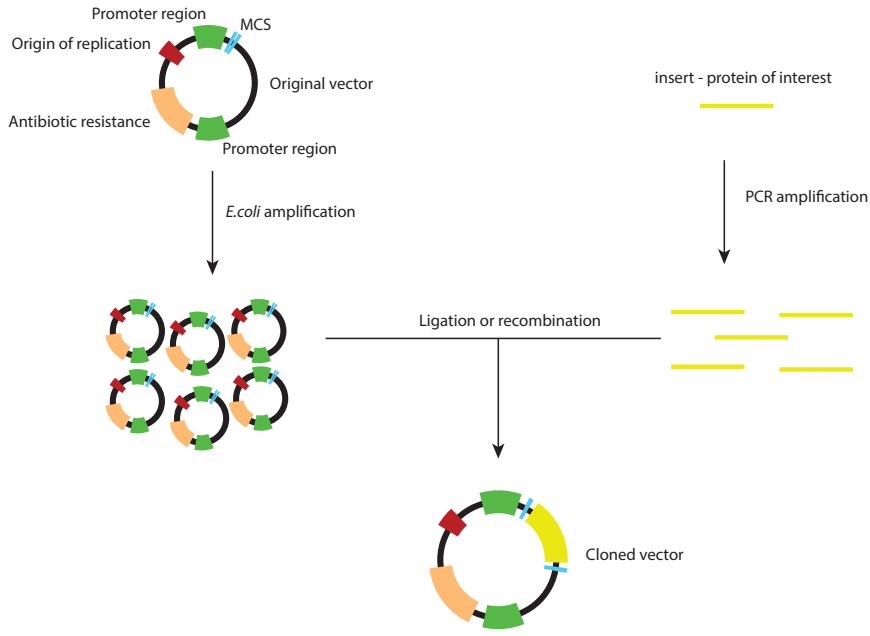


Figure 3.4: General scheme of the cloning method.

If only mutations of a certain vector are required, then site-directed

mutagenesis method is preferred. It requires a whole vector PCR amplification with the primers containing the modified sequence. In this PCR, proofreading Pfu polymerase is used to avoid any undesirable mutation. Afterwards, the PCR solution is transformed in DH5-α cells and the resulting colonies are verified like the previous cloning method.
Transformation is the process where the plasmid is inserted in the bacteria cell. Two main methods are widely used. Heat-shock method is based on a quick (45 s) thermal shock that creates cell wall pores from where the plasmid DNA enters into the cell. For a better efficiency, chemically competent cells are used. In contrast, electroporation method takes advantage of the fact that the cells can be shocked with an electrical field of 10-20 kV/cm creating cell wall holes that permit DNA plasmid to enter into it. This method requires electroporation-competent cells for the best performance. In both cases, the bacteria can repair quickly the weakness of the wall and, after 2 min on ice, super optimal broth with catabolite repression (SOC) medium (Table 3.2) is added. The cells keep growing at 37ºC for 1 h and finally they are plated in agar plates (Table 3.2) with the suitable antibiotic.

3.2.2 Protein expression and purification

As milligrams of protein are required for structural experiments, the growth, induction and expression of the protein is performed in 2 L erlenmeyer with 1 L of culture medium. Briefly, a colony from a plate is taken and submerged in the appropriate medium and antibiotic. The erlenmeyer is then leaved at 37ºC with 220 revolutions per minute (RPM) to optimise the growth. When the culture reaches OD_{600} of 0.6-0.8, isopropyl β-D-1-thiogalactopyranoside (IPTG, an allolactose mimic that triggers the expression of the proteins under lac operator; most used promoter of protein expression) is added to induce the expression of the protein of interest. The common medium used is Luria Broth (LB) (Table 3.2), but there are other formulations such as Terrific Broth (TB). However, for some NMR experiments, such as HSQC or 3D experiments (see section 3.3.1), the protein sample have to be labelled with ^{15}N or double labelled with ^{15}N and ^{13}C. In these cases, after an overnight (ON) preculture, 1 L culture is grown in minimal medium containing ^{15}N or ^{15}N and ^{13}C sources (Table 3.2). After the growth, the culture is cen-

trifuged and the pellet is immediately purified or stored at -20°C until purification.

To facilitate the protein purification, the construct is inserted with a peptide/protein tag (such as His-tag or glutathion S-transferase (GST)-tag) joined to the protein through a cleavage site. Thus, the protein is expressed together with the peptide/protein tag. This method allows the separation of the tag-protein of interest with immobilised metal ion affinity (IMAC) chromatography. In this kind of chromatography, the high affinity between the tagged protein and the resin implies an specific binding between them. After washing the resin from the rest of the components with the washing buffer, the purified protein elutes in the elution buffer. Overall, this technique allows a fast purification under mild conditions. However, the protein of interest is still ligated with a non-desire tag. To overcome this, a cleaving protein (such as Tobacco etch virus protease (TEV) or 3C protease) is added to the eluted protein.

Size exclusion chromatography (SEC) is a widespread technique in protein purification. In this kind of chromatography, the protein is purified according to its size thanks to the different porous size of the column beads. Thus, the biggest particle do not enter to most of the porous, making a short itinerary. On the other hand, the smaller molecules are able to enter in the space between the beads, making a longer itinerary and thus eluting after the bigger particles.
Ion exchange chromatography is another type of chromatography that permits the separation of the different components by their affinity to the ion exchanger. Cation exchange chromatography is used when cations need to be separated while in anion exchange are the anions who bind to the column to be separated. Thus, the relation between the pI of the protein and the pI of the other non-desired substances determines the chosen column, cation or anion. After the loading of the sample with low concentration of salts in the column, the gradient with increasing concentrations of salt elutes the desired protein.

Overall, these techniques allow an easy separation of the protein from its tag previously cleaved as well as from other proteins not previously separated.

Table 3.2: Common solutions composition.

LB (1 L; autoclaved)	
Tryptone	10 g
Yeast extract	5 g
NaCl	10 g
Suitable antibiotic:	
Ampicillin	100 μg/mL
Kanamycin	50 μg/mL

SOC (0.5 L; autoclaved)	
Tryptone	10 g
Yeast extract	2.5 g
NaCl (5 M)	1 ml
KCl (1 M)	1.25 ml
$MgCl_2$ (1 M)	5 ml
$MgSO_4$ (1 M)	5 ml
Glucose (1 M)	10 ml

LB Agar (0.5L; autoclaved)	
LB Broth	25 g
Bacteriological Agar	15 g
Suitable antibiotic:	
Ampicillin	100 μg/mL
Kanamycin	50 μg/mL

M9 medium 1 L (10x)	
Na_2HPO_4	60 g
KH_2PO_4	30 g
NaCl	5 g
$N^{15}H_4Cl$	5 g

Minimal medium (1 L)	
M9 medium (10x)	100 ml
Trace elements (100x)	10 ml
20%(w/v) Glucose	20 ml
/ or $^{13}C_6$-glucose	/ or 2 g
$MgSO_4$ (1 M)	1 ml
$CaCl_2$ (1 M)	0.3 ml
Biotin (1 mg/ml)	1 ml
Thiamin (1 mg/ml)	1 ml
Suitable antibiotic:	
Ampicillin	100 μg/mL
Kanamycin	50 μg/mL

Trace elements 1 L (100x)	
EDTA	5 g
$FeCl_3$ x 6 H_2O	0.83 g
$ZnCl_2$	84 mg
$CuCl_2$ x 2 H_2O	13 mg
$CoCl_2$ x 6 H_2O	10 mg
H_3BO_3	10 mg
$MnCl_2$ x 6 H_2O	1.6 mg

3.2.3 Experimental procedures

TGIF1-SMADs project

Cloning

Since the discovery of TGIF1 [26], up to 12 different isoforms have been identified [120]. The most abundant isoform (number 4, according to the NCBI nomenclature) was the first characterized by Bertolino *et al.* and the most used afterwards by the scientific community [28, 48]. This isoform lacks the first 116 amino acids from the canonical one and the amino acids corresponding to 116-134 are not conserved. Nevertheless, since residue number 135 to 401 both canonical and isoform 4 sequences are identical. Because of the very low expression of the canonical one respect to the isoform 4, the first 130 residues are not considered to be important for the role of TGIF1 and are not treated in this thesis. However, for consistency with the recommended nomenclature regarding protein isoforms, in this thesis we are going to use Uniprot canonical sequence as a reference (Q15583), although most of the studies used the isoform 4 as a reference.

- TGIF1 sequence was obtained from an pCMV5 Flag-TGIF1 (Isoform 2) clone (Addgene number: 14047) from Dr. Massagué lab [28]. The sequences corresponding to amino acids 256-347 & 256-339 (all residue numbering are referenced to isoform 1) were amplified by PCR and recombined in pETM11 (vector engineered by G. Stier, EMBL, Heidelberg, Germany) using $RecA_f$ recombinase (New England Biolabs, Massachusetts, USA). The vector includes kanamycin resistance, T7-lactose protein promoter, N-terminal His_6 tag and TEV cleaving site between the His_6-tag and the protein.
- TGIF1 (150-248) was amplified by PCR and recombined in pOPINJ (vector engineered by Oxford Protein Production Facility, UK) using $RecA_f$ recombinase (New England Biolabs, Massachusetts, USA). The vector includes ampicillin resistance, T7-lactose protein promoter, N-terminal His_6-GST tag and 3C protease cleaving site between the His_6-GST tag and the protein.
- SMAD2-EEE (186-467) (Uniprot entry: Q15796) clone was obtained as a gift from Dr. Zinn-Justin (CEA, Gif-sur-Yvette, France). It was

cloned in pETM-10 (vector engineered by G. Stier, EMBL, Heidelberg, Germany). The vector includes kanamycin resistance, T7-lactose protein promoter, N-terminal His_6 tag and no cleaving site between the His_6-tag and the protein. The sequence contains three mutations corresponding to the last three serines (464, 465, 467) that were substituted by three glutamic acids.
- SMAD4-MH1 (10-140) (Uniprot entry: Q13485) and SMAD2-MH1 (10-174) (Uniprot entry: Q15796) are cloned in pETM-11 (vector engineered by G. Stier, EMBL, Heidelberg, Germany). The vector includes kanamycin resistance, T7-lactose protein promoter, N-terminal His_6 tag and TEV cleaving site between the His_6-tag and the protein.
All clones were confirmed by DNA sequencing (Macrogen, Amsterdam, The Netherlands).

- DNA. Two different DNA were used in this project (Table 3.3). Having the same sequence, both are full palindromic and the double stand (ds) molecule end with both blunt extreme. Each of them contains two repetitions of the TGIF1 canonical binding DNA sequence, TGTCA (in red) ([26]). One of them was labelled at the N-terminus with Cy5 [121], what turns the DNA fluorescent and thus able to be used in EMSA experiments. Single stand purified DNA (Biomers, Germany) were dissolved in a minimum quantity of buffer (20 mM HEPES, 150 mM NaCl, 1 mM sodium azide, at pH 7.5) and then annealed by heating it at 90°C for 3 min and allowed to cold down slowly to room temperature.

Table 3.3: DNA sequences used in this thesis.

Name	Sequence	ds Mass (mol/g)
DNA1	ATTGACAGCTGTCAAT	9,758
DNA1-Cy5	ATTGACAGCTGTCAAT	11,030

Protein Expression and Purification

- TGIF1 (256-347) & TGIF1 (256-339) were expressed in BL-21 *E.Coli* cells (Novagen, Darmstadt, Germany) using LB medium or minimal medium for isotopical labelling (Table 3.2). The bacteria were induced

at OD_{600} 0.8 with 1 mM IPTG ON at 20ºC. After resuspension with guanidinium chloride (GuHCl) (Table 3.5), the bacteria were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. Then they were centrifuged (30,000 g, 30 min, 4ºC) and the supernatant was collected. The first step of purification involved Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer (Table 3.5). The proteins were eluted with 500 mM imidazole (elution buffer) (Table 3.5), and afterwards TEV protease was added to the protein solution for the cleaving step. After two hours incubating at room temperature (RT), the solutions were dialysed ON at 4ºC against 50 mM HEPES, 150 mM NaCl, 1 mM TCEP at pH 8.0 in order to decrease the concentration of imidazole. The cleaved TGIF1 (256-347) was passed again through Ni^{2+} affinity resin to remove the His-tag. Finally, the flow through that contains the protein was purified on a Hiload 16/60 Superdex 30 prepgrade column (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 20 mM HEPES, 100 mM NaCl, 1 mM TCEP at pH 6.8. The elution was collected in 2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 3 KDa (Merck Millipore, Ireland). The mass was checked by MALDI-TOF.

- TGIF1 (150-248) was expressed in Rosetta *E.Coli* cells (Novagen, Darmstadt, Germany) using LB medium or minimal medium for isotopical labelling (Table 3.2). The bacteria were induced at OD_{600} 0.8 with 0.5 mM IPTG, ON at 37ºC. After resuspension with lysis buffer (Table 3.5) supplemented with lysozyme, DNAse I and phenylmethanesulfonyl fluoride (PMSF), the bacteria were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. Then they were centrifuged (30,000 g, 30 min, 4ºC) and the supernatant was collected. The first step of purification involved Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer (Table 3.5). In the washing step, a solution of 1M NaCl was added in order to weak the interactions of the homeodomain with the DNA and remove the last one from the resin. The protein was eluted with 500 mM imidazole (elution buffer) (Table 3.5), and afterwards 3C protease was added to the protein solution for the ON cleaving step at 4ºC. Next day, the cleaved TGIF1 (150-248) was purified on a Hiload 16/60 Superdex 75 prepgrade column

(GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 20 mM HEPES, 150 mM NaCl, 1 mM TCEP at pH 6.4. The elution was collected in 2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 3 KDa (Merck Millipore, Ireland).

- SMAD2-EEE (186-467) was expressed in BL-21 *E.Coli* cells (Novagen, Darmstadt, Germany) using LB medium (Table 3.2). The bacteria were induced at OD_{600} 0.8 with 1 mM IPTG ON at 20ºC. After resuspension with lysis buffer (Table 3.5) supplemented with lysozyme, DNAse I and PMSF, the bacteria were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. Then they were centrifuged (30,000 g, 30 min, 4ºC) and the supernatant was collected. The first step of purification involved Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer. The protein was eluted with 500 mM imidazole (elution buffer) (Table 3.5). Without a cleaving step, the protein was purified on a Hiload 16/60 Superdex 30 prepgrade column (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 20 mM HEPES, 150 mM NaCl, 1 mM TCEP at pH 8.0. The elution was collected in 2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 10 KDa (Merck Millipore, Ireland).

- SMAD4-MH1 (10-140) was expressed in BL-21 *E.Coli* cells (Novagen, Darmstadt, Germany) using LB medium (Table 3.2). The bacteria were induced at OD_{600} 0.8 with 0.5 mM IPTG ON at 20ºC. After resuspension with lysis buffer (Table 3.5) supplemented with lysozyme and DNAse I, the bacteria were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. Then they were centrifuged (30,000 g, 30 min, 4ºC) and the supernatant was collected. The first step of purification involved Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer. The protein was eluted with 500 mM imidazole (elution buffer) (Table 3.5), and afterwards TEV protease and TCEP (2 mM final concentration) was added to the protein solution for the ON cleaving step at 4ºC. Next day, the cleaved SMAD4-MH1 (1-140) was purified on a Cation Exchange Hi Trap SP HP 5 ml (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE

Healthcare, Uppsala, Sweden). The gradient was done with a buffer A: 20 mM HEPES, 1 mM TCEP at pH 8.0; and buffer B: 20 mM HEPES, 1 M NaCl, 1 mM TCEP at pH 8.0. The elution was collected in 0.2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 10 KDa (Merck Millipore, Ireland). The purification was done together with more members in the lab.

- SMAD2-MH1 (10-174) was expressed in BL-21 *E.Coli* cells (Novagen, Darmstadt, Germany) using LB medium (Table 3.2). The bacteria were induced at OD_{600} 0.9 with 0.5 mM IPTG ON at 20ºC. After resuspension with lysis buffer (Table 3.5), the bacteria were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. Then they were centrifuged (30,000 g, 30 min, 4ºC) and the pH of the supernatant was adjusted to 9.2 adding 1 M $NaHCO_3$. The purification continues through a Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), equilibrated with 1 M $NaHCO_3$. After a washing with the elution buffer without imidazole, the protein was eluted with 300 mM imidazole (elution buffer) (Table 3.5). Before the cleaving step, the protein was purified on a Hiload 16/60 Superdex 75 prepgrade column (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 50 mM 2-Amino-2-(hydroxymethyl)propane-1,3-diol (TRIS), 100 mM NaCl, 2 mM dithiothreitol (DTT), 2% glycerol at pH 7.0. The elution was collected in 2 ml fractions and those corresponding with the protein were joined and TEV protease was added for the ON cleaving step at 4ºC. Next morning, more TEV was added and the sample was left at RT for 3 h. After the cleaving, a second size exclusion chromatography was set in a Hiload 16/60 Superdex 75 prepgrade column (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 20 mM TRIS, 80 mM NaCl, 2 mM DTT at pH 7.0. The elution was collected in 2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 10 KDa (Merck Millipore, Ireland). The purification was done by others members in the lab.

All proteins were monitored at every step by Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE). The concentrations were measured in a NanoDrop-1000 (Thermo Fisher Scientific,

Waltham, USA).

The theoretical values of molecular weight, isoelectric point and extinction coefficient at 280 nm of each protein were calculated using the ProtParam tool, hosted by ExPASy portal (http://web.expasy.org/protparam/).

Table 3.4: Overview of the proteins used in this project.

Protein	MW (kDa)	Aas	pI	Samples
TGIF1 (256-339)	9.0	86	7.95	non labelled, ^{15}N
TGIF1 (256-347)	9.9	94	6.05	non labelled, ^{15}N, ^{15}N-^{13}C
TGIF1 (150-248)	11.7	101	10.08	non labelled, ^{15}N, ^{15}N-^{13}C
S2-EEE (186-467)	33.2	297	5.31	non labelled
S2-MH1 (10-174)	19.0	169	8.81	non labelled
S4-MH1 (10-140)	15.2	135	9.07	non labelled

Table 3.5: TGIF1-SMADs solutions composition.

GuHCl buffer TGIF1 (256-347)	
GuHCl	6 M
Imidazole	30 mM
	pH 8.0

Elution buffer TGIF1 (256-347)	
TRIS	20 mM
NaCl	150 mM
Imidazole	500 mM
	pH 8.0

Lysis buffer TGIF1 (150-248)	
TRIS	40 mM
NaCl	300 mM
Imidazole	30 mM
Tween 20	0.2%
	pH 7.4

Elution buffer TGIF1 (150-248)	
TRIS	40 mM
NaCl	300 mM
TCEP	1 mM
Imidazole	500 mM
	pH 7.4

Lysis buffer S2-EEE (186-467)	
TRIS	20 mM
NaCl	150 mM
Imidazole	30 mM
	pH 8.0

Elution buffer S2-EEE (186-467)	
TRIS	20 mM
NaCl	150 mM
Imidazole	500 mM
	pH 8.0

Lysis buffer S4-MH1 (10-140)	
TRIS	40 mM
NaCl	300 mM
Imidazole	40 mM
Tween 20	0.2%
	pH 8.0

Elution buffer S4-MH1 (10-140)	
TRIS	40 mM
NaCl	150 mM
Imidazole	300 mM
	pH 8.0

Lysis buffer S2-MH1 (10-174)	
$NaCO_3$	786 mM
NaCl	500 mM
DMSO	10 %
Trace Elements	1x
TCEP	4 mM

Elution buffer S2-MH1 (10-174)	
HEPES	50 mM
NaCl	500 mM
Glycerol	20 %
Imidazole	300 mM
TCEP	4 mM
	pH 7.0

WW domains project. As previously described in PNAS under the title "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements" [122].

Cloning

The FBP28 sequence were cloned from 9-11 days mice embryo cDNA. The WW2 sequence from mice and humans is identical. All variants were cloned into a pGAT2 vector (vector engineered by M. Hyvönen, University of Cambridge, UK). The vector includes ampicillin resistance, T7-lactose protein promoter, N-terminal His_6-GST tag and Thr protease cleaving site between the His_6-GST tag and the protein. Point mutations of the variants were introduced by site-mutagenesis PCR using Pfu DNA-polymerase (Agilent Technologies, USA) and the appropriate primers containing the mutations. After the digestion with DPN1, the vectors were transformed in DH5-α strain (Invitrogen, Carlsbad, CA, USA) with the heat-shock method. Some colonies were selected, their plasmids were amplified by miniPrep (Quiagen, Hilden, Germany) and afterwards verified by DNA sequencing.

Protein Expression and Purification

All proteins were expressed in *E.coli* BL21(DE3) (Novagen, Germany) using LB medium. WT and Y19L ΔNY11R, ΔNΔCY11R, ΔNΔCY11RL26A mutants were induced at OD_{600} 0.6 with 0.3-0.5 mM IPTG, for 3 h at 37°C. Y11R and W30F were otherwise induced ON at 37°C. After resuspension with lysis buffer A (Table 3.6) supplemented with lysozyme, DNAse I and PMSF, they were lysed using EmulsiFlex-C3 (Avestin, Mannheim, Germany) cell disruptor. They were then centrifuged (30,000 g, 30 min, 4°C) and purified through Ni^{2+} affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer. The proteins were eluted with 50 mM ethylenediaminetetraacetic acid (EDTA) (elution buffer, Table 3.6) and then thrombin was added to the solutions and incubated ON in order to remove the tag. The cleaved proteins were purified from the His_6-GST-tag with a Hiload 16/60 Superdex 30 prepgrade column (GE Healthcare, Uppsala, Sweden) in an Äkta® Purifier10/FPLC system (GE Healthcare, Uppsala, Sweden) equilibrated with 20 mM sodium phosphate, 100 mM NaCl, 1 mM Azide at pH 6.5.

The elution was collected in 2 ml fractions and those corresponding with the protein were concentrated in a Amicon® Ultra-15 centrifugal filter 3KDa (Merck Millipore, Ireland). The proteins were monitored at every step by Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) and the masses of each mutant were checked by MALDI-TOF. The concentration were measured in a NanoDrop-1000 (Thermo Fisher Scientific, Waltham, USA). The theoretical values of molecular weight, isoelectric point and extinction coefficient at 280 nm of each protein were calculated using the ProtParam tool, hosted by ExPASy portal
(http://web.expasy.org/protparam/).

Table 3.6: WW domain project solutions composition.

Buffer A		Elution buffer	
TRIS	20 mM	TRIS	20 mM
NaCl	150 mM	NaCl	150 mM
Imidazole	10 mM	Imidazole	10 mM
	pH 8.0	EDTA	50 mM
			pH 8.0

3.3 Techniques

3.3.1 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance spectroscopy, or simply NMR, is a spectroscopy technique that provides information at atomic level of molecules and solids. It is based on the ability of certain nuclei atom to absorb and emit electromagnetic radiation when they are under a magnetic field. In the structural biology field, it is considered as one of the main techniques, along with X-ray crystallography and cryo-electron microscopy.
NMR was born in 1937 when Isidor Rabi measured for first time the nuclear magnetic moment of an atom nuclei [123]. A couple years later, in 1946, Felix Bloch and Edward Purcell, independently, applied the measurement also to liquids and solids, respectively, opening the door for the study of small molecules [124, 125]. But it was in 1964 when Richard Ernst realised that the replacement of the slow and long radio pulses by a short but intense ones while processing them by Fourier transform increases dramatically the sensitivity of the NMR. Later on, based on the ideas of Jean Jenner, Ernst proved the possibility of realising 2D NMR spectra [126]. Finally, in 1985, Kurt Wüthrich solved the first structure of a small protein (BUSI) based on the measurement of distances between atoms by 2D NOESY experiments [127]. Nowadays NMR is a full established versatile technique capable to determine the full structure of a protein up to roughly 100 KDa [128], to follow a phosphorylation reaction in-cell [129], and to provide structural information about large multi domain complexes in solution [130].

Basic principles

NMR is based on the manipulation of the spin of the atom nuclei. It appears that some atomic nuclei own a nuclear spin angular momentum. This momentum is an intrinsic property of the atoms nuclei as it is the charge or the mass. The possession or not of the nuclear spin angular momentum is specified by the spin quantum number, *I*. *I* can be an integer (0, 1 ...) or half-integer (1/2, 3/2 ...). Only when I is not equal to 0 ($I \neq 0$), the nucleus of the atom has spin. For instance, the most abundant H isotope (99.9885%), ^{1}H, has an $I = 1/2$, making it appropriate for

NMR spectroscopy. However, the quantum number only states which nuclei can be manipulated, but not the intensity of the manipulation. As every nuclei has a positive charge, the spin angular momentum generates a magnetic momentum. The relation between the spin angular momentum and the magnetic momentum is given by the gyromagnetic ratio (γ), which explains the intensity of the magnetic momentum generated and thus, the intensity of the manipulation:

$$\vec{\mu} = \gamma \vec{I} \tag{3.1}$$

where $\vec{\mu}$ is the nuclear magnetic moment and $\vec{I}$ the nuclear spin moment.
The ^{1}H nuclei has a γ of 267.513 10^{6} rad s^{-1} T^{-1}. This high value of γ in conjunction with the fact that it is the most abundant nuclei in biologic samples, makes ^{1}H the most suitable atom for biological NMR. The other abundant nuclei in living being, ^{12}C (98.93% isotope abundance), with a I = 0, thus no spin angular momentum; and ^{14}N (99.632%) with a I = 1, integer spin (that introduces high quadrupole interferences making inviable the measurement) are not capable to be used in NMR. This problem is solved enriching artificially (labelling) the samples with ^{13}C and ^{15}N, respectively, both with an adequate I of -1/2. However, their γ is not so high as ^{1}H, and so less signal is acquired if the experiment is recorded with ^{13}C or ^{15}N atoms nuclei. For this reason, most of the NMR experiments are actually acquired in ^{1}H.

When a nuclei is under the influence of an external magnetic field (B_0), the spin adopts different conformations, each of them with different energy in a process known as Zeeman effect (Figure 3.5). The number of the conformations is regulated by the quantic number *m*, that adopt *m* = *I*, *I*-1, ... -*I*, with a total of 2*I* + 1 conformations. The energy of each configuration is explained by:

$$E_m = -\gamma \vec{I} B_0 = -m\hbar\gamma B_0 \tag{3.2}$$

If we apply this formula for I = 1/2; then *m* adopt the values +1/2 (α) and -1/2 (β).

$$E_\alpha = -\frac{1}{2}\hbar\gamma B_0 \qquad E_\beta = +\frac{1}{2}\hbar\gamma B_0 \tag{3.3}$$

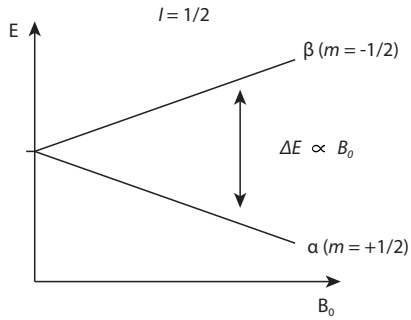


Figure 3.5: Zeeman effect in a atom nuclei with $I = 1/2$. The external magnetic field B_0 splits the otherwise one conformation into two energetic levels.

And so the energy difference between both levels is,

$$\begin{aligned} \Delta E_{\alpha\to\beta} &= +\frac{1}{2}\hbar\gamma B_0 - (-\frac{1}{2}\hbar\gamma B_0) \\ &= \hbar\gamma B_0 \end{aligned} \tag{3.4}$$

Applying Plank's law $E = h\nu$, the frequency of the transition between the two spin states, also known as the Larmor frequency, is,

$$\begin{aligned} w_0 &= -\gamma B_0 \quad \text{in rad s}^{-1} \\ v_0 &= -\frac{\gamma B_0}{2\pi} \quad \text{in Hz} \end{aligned} \tag{3.5}$$

Considering a commonly used 14.0921 T external magnetic field, the difference of energy between alpha and beta for ^{1}H turn out to be 3.97555 10^{-25} J, or as Larmor frequency, 599,985,479.179 Hz, or 600 MHz (located in the Ultra high frequency, UHT, radio spectrum band).

In thermal equilibria, the energy levels are occupied following the Boltzmann distribution.

$$\frac{N_\beta}{N_\alpha} = e^{-\frac{\Delta E}{k_B T}} = e^{-\frac{\hbar\gamma B_0}{k_B T}} \tag{3.6}$$

For ^{1}H, in the same external magnetic field and at room temperature (T = 298 K), N_β is 0.99990 times N_α, meaning that the lower state is only

0.01% more occupied than the higher energetic level. This small difference in population is the reason why NMR has low sensitivity. Apart from decreasing the T and increasing the B_0, scientists have commonly employ two techniques to overcome the low signal. First, use high concentrated samples; the more atoms are, the strongest the signal. And second, acquire the same spectrum many times. The idea is to accumulate several spectra and process them altogether. The signal from the sample is added linearly, so after *N* spectra the signal is *N* times stronger. However, the noise, as randomly, only increases by a factor of $\sqrt{N}$. Considering both factors, the signal-to-noise ratio (SNR) increases the by a factor of $\sqrt{N}$, meaning that doubling the number of scans (spectra acquired) the signal increases 1.41 times. As doubling the number of scans also means doubling the time required for acquiring it, there is a practical limit of using this technique, that depends on every experiment.

So far, one can imagine that if all atom nuclei are equal, there will be only one signal per atom nuclei and thus it will be impossible to distinguish between the atoms with different connectivity. However, it happens that the electron density that surrounds the nuclei is different from nucleus to nucleus. The different electron density shields the nuclei from the external magnetic field in a different way, resulting that every nucleus provides a slightly different frequency depending of the environment. This phenomena links the 3D structure of the molecule with the signals that are acquired in the NMR.
Having different chemical environments, the nuclei have slightly different Larmor frequencies. As the transition between the two states depends on the strength of the external magnetic field, the frequency differences between the same nucleus are different for each spectrometer. The concept of chemical shift (δ), defined as,

$$\delta(ppm) = \frac{v - v_{ref}}{v_{ref}} 10^6 \tag{3.7}$$

overcome this limitation. Due to this, the same spin have the same chemical shift regardless the magnet used to measure it. The huge difference between the absolute frequency range (in 14.0921 T; 600 MHz for ^{1}H) and the relative frequency range between the most and the less

shielded nuclei (in 14.0921 T, 10 kHz for ^{1}H) makes part per million (ppm) the units of reference for the chemical shift.

As well as the nuclei spins are affected by the external magnetic field, they are also altered by the other nuclei that surround them. In other words, one nucleus spin has a different Larmor frequency depending if the spin, to which it is connected, is aligned or not with the external magnetic field. This interaction between the spins is called J-coupling (also known as scalar coupling). Although the J-coupling have a low affection to the total frequency (it is in the rage of 1 to 100 Hz), it causes the splitting of the signal, which lead to the broadening of the peak. Furthermore, the J-coupling allows the transfer of magnetisation between nuclei, including heteronuclear nuclei, being the basis of the multidimensional experiments.

Vector model

The vector model is a simple framework that although it do not explain a full NMR experiment (only quantum mechanics does it) it helps to understand the basics of it. The vector model will be used to explain briefly how a spectrum is actually acquired.
The vector model is based on the simplification that the atom nuclei is like a positive charged particle that spins on its axis. When a non-zero radius charged nucleus spins, generates a tiny magnetic field, $\vec{\mu}$. Under the influence of an external magnetic field (B_0), such as the one generated by an NMR spectrometer, the $\vec{\mu}$ aligns with it along the same z-axis. However, this is an oversimplification. Actually, as the spinning tiny magnet has also angular momentum, the nuclei actually precess around the z-axis at precisely the Larmor frequency before explained. Plus, quantum mechanics shows us that the angle of precession (θ) for the ^{1}H nuclei can only be 54.7°. As mentioned before, when $I = 1/2$, the energy is split in two levels. In the vector model this is represented with the lower levels corresponding to the precession around +z while the higher is precessing around -z (Figure 3.6).
Moreover, the thermal motion – much greater than the interaction between the tiny magnet and the strong external one – randomise the direction of each nuclei. Overall the net magnetisation is equal as if

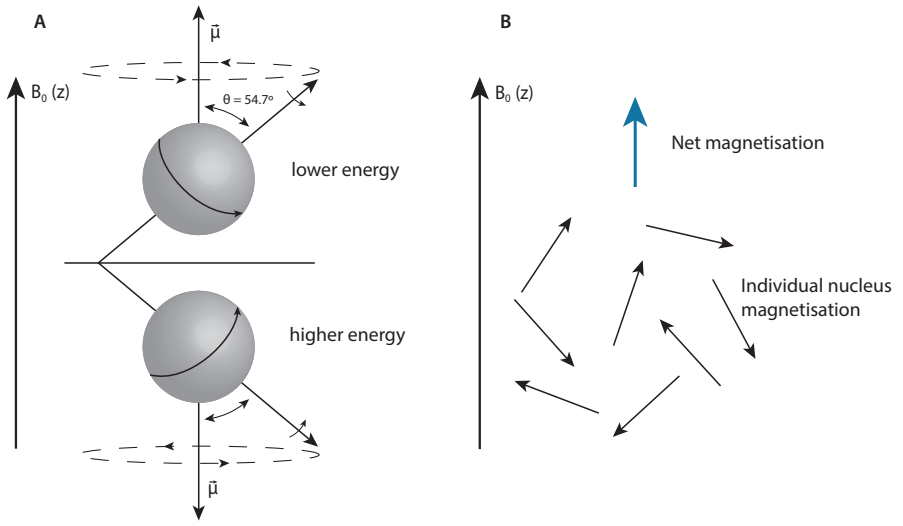


Figure 3.6: A) Under the influence of an external magnetic field (B_0), a charged nucleus not only spins around its axis, but precess around the direction of B_0. **B**) The sum of the individuals magnetisations results in an overall magnetisation aligned with the external field.

only 1 spin over 10,000 is perfectly aligned with the external field while the other become totally random. For this reason, the magnetisation of each nuclei is summed and only the total magnetisation is considered.

NMR can only record the displace of the total magnetisation, being impossible to detect each one of the tiny magnets. At equilibrium, the net magnetisation is stable on the +z axis, so no record can be done. To do so, the magnetisation has to move to the xy plane and record the precession movement when the magnetisation return to the equilibrium state at +z.

A simple NMR experiment starts with a strong but short (around 10 μs for ^{1}H) radio pulse (radio frequency, RF) applied at +x. This pulse, in the Larmor frequency, shifts the net magnetisation towards the -y axis, on the xy plane. In most of the NMR experiments the magnetisation is transferred to another atoms and finished usually on ^{1}H nuclei (because as explained before, this nuclei gives the highest signal). Finally, the magnetisation returns to the +z axis in a process known as relax-

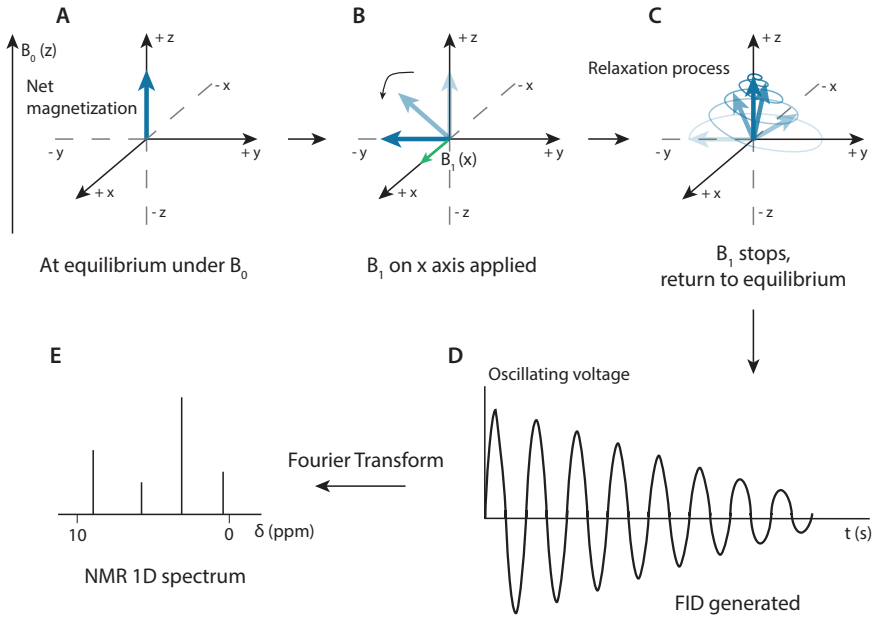


Figure 3.7: Scheme of a very simple NMR experiment. **A**) Under the B_0 field, the net magnetisation at equilibrium points toward +z axis. **B**) A short RF pulse on B_1 (x) shifts the magnetisation towards -y axis. **C**) Once the pulse is over, relaxation process start, recovering the equilibrium. **D**) While the magnetisation is oscillating, recovering the equilibrium, the FID is recorded. **E**) The FT is applied to the FID, obtaining an standard NMR 1D spectrum.

ation. In this moment, the probe records the precession around +z as a free induction decay (FID) signal. Applying the Fourier Transform (FT) the oscillating voltage versus time function recorded is transformed in intensity versus frequency plot, the common axis of an NMR spectrum (Figure 3.7).

The relaxation process is one of the most important parameters in NMR as it has a direct influence on the signal recorded. It can be divided in the sum of two different processes. The spin-lattice relaxation, also called longitudinal, refers to the recovery of the magnetisation on the +z axis and is characterised by the time constant T_1. On the other hand, the spin-spin relaxation, also called transverse, explains the recovery of

the randomness on the xy plane and is defined by the time constant T_2.

Experiments used

In this thesis, all the NMR samples were dissolved in aqueous buffers. The 2 ^{1}H of water also are affected by the NMR and the pulses. As the quantity of molecules is several times higher than any protein we dissolved in it, many techniques (as watergate or presaturation) have been developed to suppress the otherwise gigantic water peak at ∼4.7. All the experiments recorded in this thesis have proteins as sample to be analysed. From this point, all the explanation will be focused in proteins.

1D experiment

^{1}H one dimension (1D) experiment is the most simple experiment and the first executed in any structural biology project (Figure 3.8). This experiment requires only an enough concentrated sample containing ^{1}H. It is a fast experiment (less than a minute) that shows all the peaks related to ^{1}H of the sample. The range from -2 to +12 ppm allows the observation of all kinds of shielded ^{1}H of a protein. It provides information about (1) the presence of protein in the tube (2) the fold of the protein (3) the good suppression of the water (required if we want to perform more complex experiments) (4) the presence of any impurity that is affecting the spectrum.
However, the utility of this experiment is limited to check how is the protein globally because it contains all the peaks from ^{1}H in only 1D, making impossible to distinguish and assign them.

2D experiments

The two dimension (2D) experiments represent the correlation between two different nuclei spins, indicating how two different spins are connected between themselves. These experiments are based on the transfer of magnetisation between different nuclei through J-coupling or another spin interaction. The 2D experiments are displayed in a two dimension plot (every axis correspond to the δ of each nuclei) with the in-

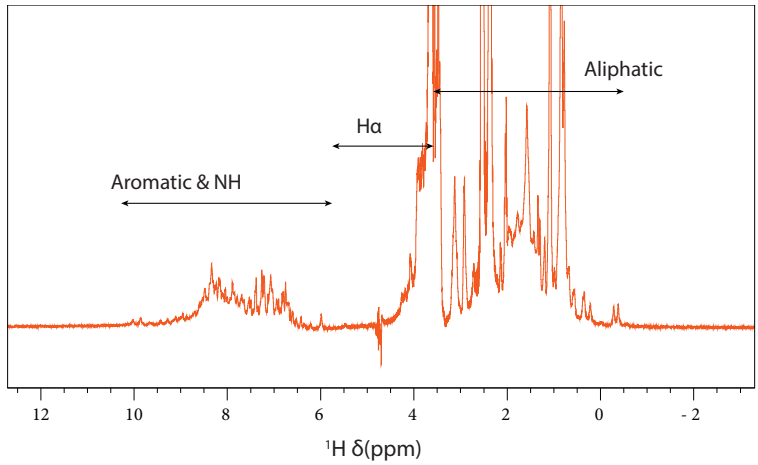


Figure 3.8: 1D ^{1}H spectrum of a folded protein. The highest peaks are cut to allow better visualisation of the smaller ones. The two peaks located below 0 ppm indicate the folding of the protein. In the region from 0 to 4 the aliphatic H are located. The negative peaks at ∼4.7 refers to the suppressed water. Finally after 6 and until 10, the aromatic H and the H bound to N are localised.

tensity given by isopleths. There are two main kinds of 2D experiments used in this thesis: homonuclears (TOCSY/NOESY) and heteronuclears (HSQC).

The homonuclear experiments are those that transfer the magnetisation between the same type of nuclei, in most cases ^{1}H. In all the homonuclear experiments done in this thesis both nuclei are ^{1}H, thus the further explanation is based only on ^{1}H -^{1}H experiments. In these 2D spectra, both axis corresponds to ^{1}H. The spectra show two kinds of peaks, the ones in the diagonal, like in the 1D spectra; and the cross-peaks, that are symmetrical left/right of the diagonal peak and represent the link between two protons with different δ. The nature of the link distinguishes which kind of experiment is. Total correlation spectroscopy (TOCSY) provides information about the ^{1}H that are connected by a chain of couplings while Nuclear Overhauser effect spectroscopy (NOESY) show the ^{1}H that are close in space with an intensity that drops $1/r^{6}$ with the distance.

TOtal Correlation SpectroscopY (TOCSY) is an evolution of the previous correlation spectroscopy (COSY) experiment. As mentioned previously, a cross-peak indicates an unbroken chain of couplings between two protons. For example, in a threonine the HA proton is coupled to HB, and the HB is coupled to the three HG. So, in TOCSY, there is a cross-peak not only between HA and HB or HB and HG but also between HA and HG. TOCSY is then very useful in order to see the connection of all the protons in the same spin system. As every amino acid is an isolated spin system (the peptide bond breaks the coupling chain), this spectra is very useful in the assignment.

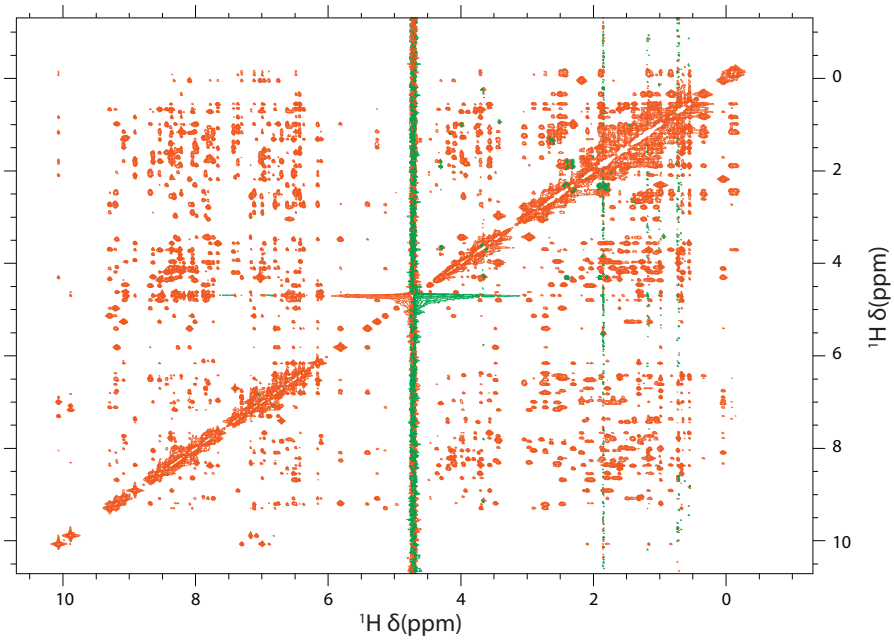


Figure 3.9: NOESY spectrum. The positive peaks are displayed in red and the negative in green. It can be seen as 1D spectrum in the diagonal that is expanded with the cross-peaks symmetrically located at both sides. The broad vertical line in the middle of the spectra is the water peak.

Nuclear Overhauser Effect SpectroscopY (NOESY). However, only with a TOCSY is very difficult to assign completely a protein as many amino acids are repeated in the sequence and some of them have similar

patterns of δ. For this reason, the NOESY spectra result very helpful (Figure 3.9). NOESY cross-peaks represents the link between those ^{1}H that are close in the space (up to 5 Å) and so, not only the cross-peaks from the same spin system are seen but also some cross-peaks from ^{1}H from different spin systems. Specially interesting are the cross-peaks from the $^{1}H^{N}$. As the $^{1}H^{N}$ is very near in the space from the ^{1}H of the -1 residue, the two different system couplings can be connected. Thus, all the protein can be sequentially connected, identified and assigned with the only exception of proline that do not have $^{1}H^{N}$.

A second crucial advantage of the space-linked cross-peaks that NOESY provides is those connections close in space but far in sequence. If we have discarded aggregation, the only reason that two atoms far in sequence are together in space is because the protein is folded in a such way that these two atoms are in proximity. In addition, as the NOE effect depends $1/r^{6}$ on the distance, it also provides information about how close in the space are both atoms. With the information of these cross-peaks, software such as Crystallography & NMR System (CNS) [131] or Ambiguous Restraints for Iterative Assignment (ARIA) [132] can calculate the structures with the sequence of the protein that fulfill the NOESY restrictions.

The heteronuclear experiments are those that transfer the magnetisation between different kind of nuclei. The most used nuclei are ^{1}H, ^{13}C and ^{15}N. This implies that the sample must be labelled in ^{13}C or ^{15}N, being this the main drawback of the experiment.

The most prominent 2D heteronuclear experiment is the **Heteronuclear Single-Quantum Correlation/Coherence spectroscopy (HSQC)** [133]. It records the correlation between the chemical shift of ^{1}H and ^{13}C (more used for organic molecules) or ^{1}H and ^{15}N (more used for biologic molecules) through the J coupling interaction between the nuclei. Focusing on the ^{1}H-^{15}N HSQC experiment, it shows a peak for every ^{15}N-^{1}H bound in the protein. Every amino acid have one ^{15}N-^{1}H bound with the exception of the proline (that lacks the $^{1}H^{N}$). Also some side chains contains ^{15}N-^{1}H and give a peak too. Overall, the spectrum, with the x-axis for ^{1}H and the y axis for ^{15}N, allows a general vision of almost all amino acids in a fast and easy way (Figure 3.10). Such vision allows a quick analysis of the position of the peaks. If they are too close (from

7.5 to 8.5 $^{1}H\delta$(ppm)) may indicate that the protein is unfolded. The other way around, a good dispersion (from 6 to 10 $^{1}H\delta$(ppm)) means proper folding. Moreover, if some disruption is affecting the amino acid such a phosphorylation or any binding of a ligand, the correspondent affected peak will shift from the reference position. However, this information is much more relevant if the assignment of every ^{15}N-^{1}H peak is known. Unfortunately, with only an HSQC is impossible to known to which amino acid corresponds to each peak, as 3D experiments are needed for the assignment. **Heteronuclear Multiple-Quantum Correlation** (HMQC) is another 2D experiment that, similarly to the HSQC, also correlates the chemical shift, via the J coupling, between the ^{1}H and the ^{13}C or ^{15}N, giving very similar kind of spectra and information. It is based on a different pulse sequence compared to HSQC, that overall give worse resolution but higher signal-to-noise ratio.

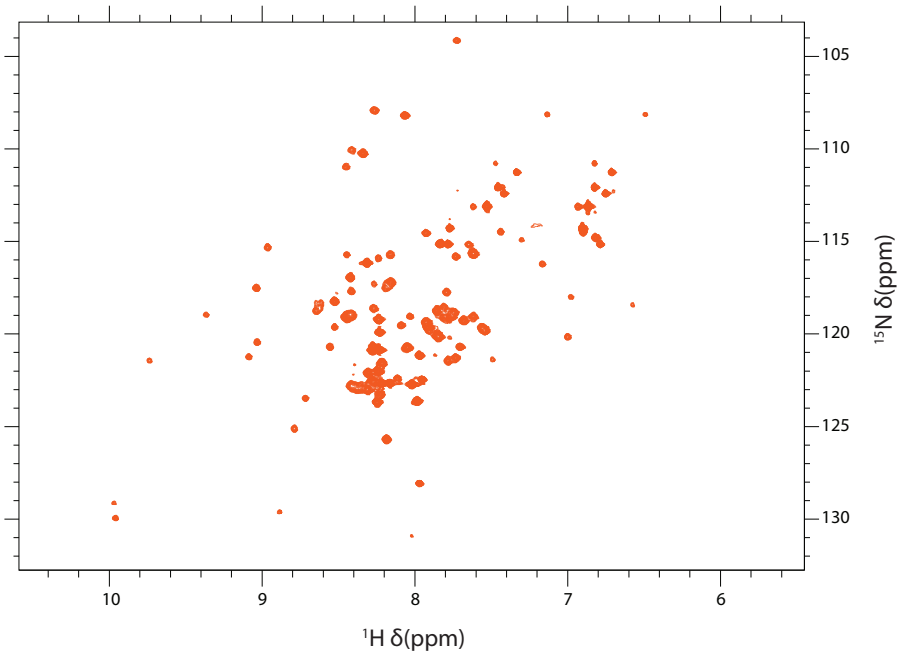


Figure 3.10: Figure of a typical ^{1}H-^{15}N HSQC spectrum, where each peak represents one ^{15}N-^{1}H correlation. The good dispersion observed indicates that the protein is well folded.

3D experiments

The three dimension (3D) experiments provide information about the link between three different spin nuclei. There are many types of 3D experiments, but in this thesis only the ones needed for assignment were used.

CBCA(co)NH. It provides information about the link between the $^{1}H^{N}$ and ^{15}N of the i amino acid and the $^{13}C\alpha$ and $^{13}C\beta$ of the previous i-1 amino acid. In the practice is like a ^{1}H-^{15}N HSQC with a new ^{13}C on the third dimension. In this experiment, the magnetisation from the $^{1}H\alpha$ and $^{1}H\beta$ is transferred to the corresponding ^{13}C, where they evolve together. Because of this, both ^{13}C appear in the same dimension. Afterwards the magnetisation goes through the ^{13}C' and evolves in the ^{15}N of the i amino acid to finally end in the $^{1}H^{N}$ for detection (Figure 3.11) [134].

CBCANH. It is similar to the CBCA(co)NH. In this experiment the magnetisation goes over the ^{13}C' allowing the ^{15}N to receive the magnetisation not only from the previous amino acid (i-1) but also for the same amino acid (i). This crucial change allows the link between the i and the i-1 amino acid (Figure 3.11) [135].

hN(coca)NH. The hN(coca)NH experiment connect the chemical shift of an $^{15}NH_{i}$ with the following $^{15}NH_{i+1}$, allowing a "backbone NH walk". This results very useful in the assignment process as provide information about which ^{15}NH peak is after which other peak. The magnetisation is transferred from the $^{15}NH_{i+1}$ to the $^{15}NH_{i}$ (evolving in different dimensions) through $^{13}C\alpha$ and ^{13}C'. Finally the magnetisation is recorded in $^{1}H^{N}$ (Figure 3.11) [136].

Overall, the protein is assigned following the amino acid sequence, in a similar way as NOESY/TOCSY works. While the CBCANH is main spectrum used for the assignment, the CBCA(co)NH and hN(coca)NH spectra help to distinguish the different spin systems.

Ligand binding principles

NMR is a tool that allows monitoring a binding reaction when a ligand is added to a given sample. As explained before, ^{1}H-^{15}N HSQC is the most common experiment used due to the broad spectrum of peaks that

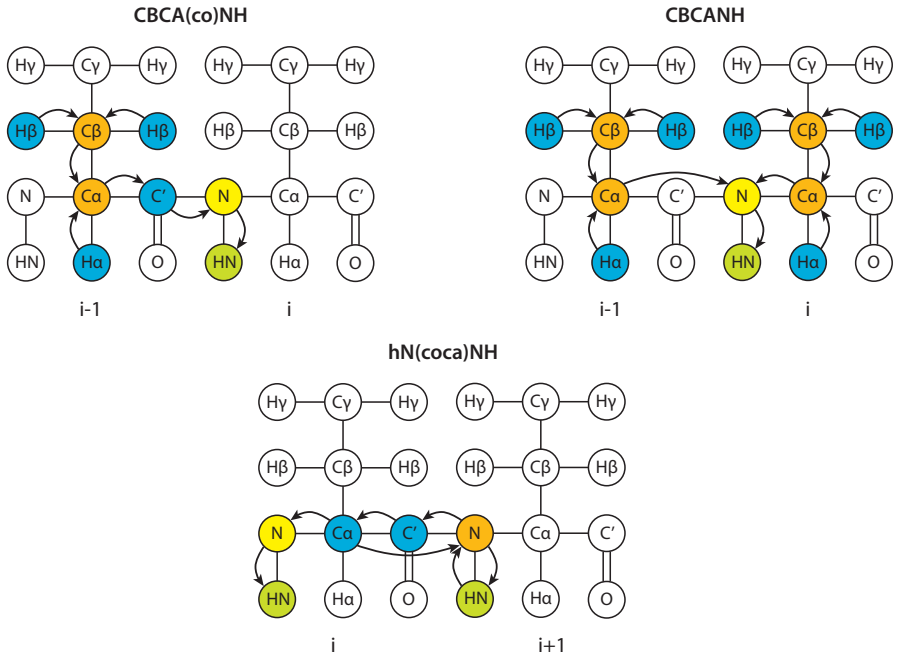


Figure 3.11: Scheme of CBCA(co)NH, CBCANH and hN(coca)NH 3D experiments. The graph show the way of the magnetisation. In blue are represented those nuclei where the magnetisation goes through while in orange, yellow and green are those nuclei where the magnetisation evolve and thus are a dimension of the spectra.

it provides (one for each amino acid excepting proline) in a fast manner (20-40 min each experiment). As the protein is labelled in ^{15}N, the ligand can be unlabelled and so, do not appear in the spectrum. If the ligand interacts with some region of the protein, the nuclei atoms that are in that region will be affected by changes in the electronic environment and so their δ will shift generating a chemical shift perturbation (CSP). The usual way of work includes a titration of the ligand versus the protein in a way that several spectra are acquired while equivalents of ligand are added.

The way the peak is shifted depends on how fast is the exchange of the protein-ligand interaction in comparison with the difference between the ν of both states ($\Delta\nu$). When the interaction exchange is much faster than the difference in Hz between the bound and the un-

bound state, there are several binding and unbinding processes during the same NMR experiment. This situation lead to a mixture of both situations in one single peak that shifts from the unbound δ to the bound δ without losing intensity (Figure 3.12 B).

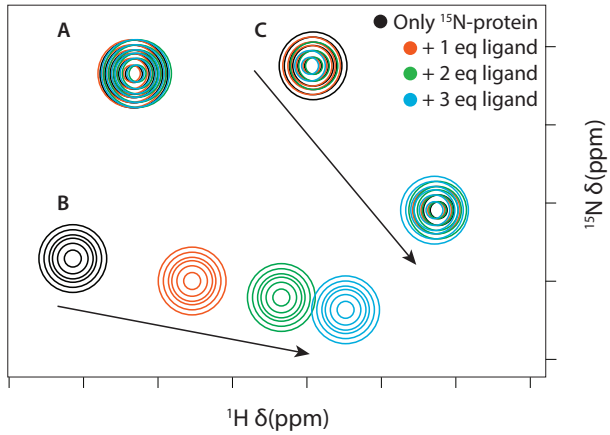


Figure 3.12: ^{1}H-^{15}N HSQC titration. **A**) The peak does not shift upon the addition of the ligand. **B**) Example of fast exchange. The unbound peak moves towards the final bound peak. **C**) Slow exchange. The unbound peak progressively disappears while the bound peak appears.

On the other hand, if the exchange rate is much slower than the difference of δ, then some molecules will be bound with the ligand and some no, leading to the appearance of two smaller peaks, from the bound and unbound state respectively. As the proportion of ligand increases, the peak related to the unbound state drops while the peak from the bound state increases at the same proportion (Figure 3.12 C).
In the intermediate state (when exchange rate is similar than $\Delta\delta$) something in between is observed. The mixture peak shifts but with lower intensity in the middle region, plus both peak at the bound and unbound state are also seen but also at lower intensity as the same quantity of spins have to be shared by all peaks.

Experimental procedures

TGIF1-SMADs project

- TGIF1 (256-339) & TGIF1 (256-347)

The experiments regarding TGIF1 (256-347) 2D HSQC, and 3D CBCANH-BEST, HNCANH-BEST assignment experiments were done at 200 μM in 20 mM HEPES, 100 mM NaCl, 2 mM DTT, 1.5 mM EDTA, 100 μM PMSF at pH 6.4. 5% of D_2O was added to a final volume of 350 μl in a 5 mm shigemi tube. The temperature was set at 277 K.

Generally, the phosphorylation reaction with p38α kinase was done adding ATP 2 mM (final concentration), $MgCl_2$ 5 mM (final concentration) and 10 μl of p38α at 0.1 μg/μl (SignalChem, Richmond, Canada) for every 100 μl of protein solution. The reaction was incubated at 25°C ON without stirring.

The further phosphorylation with Casein Kinase I (CK1) was done adding ATP 2 mM (final concentration), 1x protein kinase buffer (provided by the manufacturer) and 8 μl of CK1 at 1,000,000 U/ml (New England Biolabs, Massachusetts, USA) for every 100 μl of protein solution. The reaction was incubated at 25°C ON without stirring.

The experiments regarding pTGIF1 (256-347) with p38α, 2D HSQC, and 3D CBCANH-BEST, HNCANH-BEST assignment experiments were done at 200 μM in 20 mM HEPES, 100 mM NaCl, 2 mM DTT, 1.5 mM EDTA, 100 μM PMSF at pH 6.4. 5% of D_2O was added to a final volume of 320 μl in a 5 mm shigemi tube. The temperature was set at 277 K.

2D SOFAST-HSQC and SOFAST-HMQC experiments before and after phosphorylations with p38α and CK1 were done with 50 μM TGIF1 (256-347) in 20 mM HEPES, 100 mM NaCl, 2 mM DTT (or 1 mM TCEP) at pH 6.8. 5% of D_2O was added to a final volume of 150 μl in a 3 mm tube. The temperature was set at 277 K.

Real-time NMR experiment following p38α phosphorylations were done with 80 μM TGIF1 (256-339) in 20 mM HEPES, 100 mM NaCl, 2 mM DTT at pH 6.8. For the phosphorylation were added $MgCl_2$ 5 mM (final concentration), ATP 1 mM (final concentration), 8 μl p38α at 0.1 μg/μl (SignalChem, Richmond, Canada). 5% of D_2O was added to a final volume of 150 μl in a 3 mm tube. 21 non-stop SOFAST-HMQC experiments (34 min 42s each) were recorded at 298 K. The intensities of every ^{1}H-^{15}N unambiguously assigned peak were fitted to mono-exponential

equation ($y = y_0 * (1 - exp(-K * x))$ in GraphPad Prism software.
In the NMR ^{1}H-^{15}N HSQC titrations between ^{15}N TGIF1 (256-347) (unphosphorylated (75 μM) and with phosphorylated S286 and S291 (45 μM)) and SMAD2-EEE (186-467), the common buffer used was 20 mM HEPES, 100 mM NaCl, 1 mM TCEP at pH 6.8. The temperature was set at 277 K.
TGIF1 (256-347) HSQC titrations with SMAD2-MH1 (10-174) at around 200 μM in 20 mM PO_4, 80 mM NaCl, 1 mM TCEP at pH 6.4. 10% of D_2O was added to a final volume of 154 μl in a 3 mm tube. The temperature was set at 277 K.
TGIF1 (256-347) HSQC titrations with TGIF1 (150-248) at around 200 μM in 20 mM PO_4, 80 mM NaCl, 1 mM TCEP at pH 6.4. 10% of D_2O was added to a final volume of 148 μl in a 3 mm tube. The temperature was set at 277 K.
- **TGIF1 (150-248)**
The experiments regarding TGIF1 (150-248) 2D HSQC experiments were done at 100 μM in 20 mM PO_4, 150 mM NaCl, 1 mM TCEP at pH 4.9. 10% of D_2O was added to a final volume of 270 μl in a 5 mm shigemi tube. The temperature was set at 298 K.
The 3D CBCANH-BEST, CBCA(CO)NH-BEST assignment experiments were done at 166 μM in 20 mM PO_4, 80 mM NaCl, 1 mM TCEP at pH 5.25. 10% of D_2O was added to a final volume of 160 μl in a 3 mm tube. The temperature was set at 298 K.
TGIF1 (150-248) HSQC titrations with DNA were done at 231 μM in 20 mM PO_4, 80 mM NaCl, 1 mM TCEP at pH 5.25. 10% of D_2O was added to a final volume of 160 μl in a 3 mm tube. The temperature was set at 298 K.
TGIF1 (150-248) HSQC titrations with SMAD4-MH1 (10-140) at 65 μM in 20 mM HEPES, 150 mM NaCl, 1 mM TCEP at pH 6.4. 10% of D_2O was added to a final volume of 140 μl in a 3 mm tube. The temperature was set at 298 K.
TGIF1 (150-248) HSQC titrations with SMAD2-MH1 (10-174) at 200 μM in 20 mM PO_4, 80 mM NaCl, 1 mM TCEP at pH 5.25. 10% of D_2O was added to a final volume of 154 μl in a 3 mm tube. The temperature was set at 298 K.

The experiments regarding the TGIF1 (256-339) & TGIF1 (256-347) 2D

HSQC, SOFAST-HMQC; 3D CBCANH-BEST, HNCANH-BEST assignment experiments and Real-Time NMR following kinase phosphorylation, were recorded with a 600MHz or 750 MHz Bruker Avance spectrometer both equipped with a cryogenically cooled triple resonance $^{1}H^{13}C/^{15}N$ TCI probe located in the Department of NMR-supported Structural Biology, Leibniz Institute of Molecular Pharmacology (FMP Berlin), Robert-Rössle Strasse 10, 13125 Berlin, Germany. All other experiments, including all the titrations, were recorded in house Bruker Avance III 600-MHz spectrometer equipped with a cryoprobe (CPQCI ^{1}H-^{31}P/-$^{13}C/^{15}N$/D Z-GRD).

The spectra were processed with TopSpin v3.5 Bruker Software. Real-time NMR experiments were evaluated with Sparky [137] and the assignment and the titrations were analysed with CcpNMR Analysis [138]. For every titration, the change in the displacement as well as in the intensity of every chemical shift was analysed. In all cases the last titration point was compared with the first point with no ligand. The movement in the chemical shift was evaluated with the equation:

$$d = \sqrt{\frac{1}{2}[\delta_H^2 + (0.14 \cdot \delta_N^2)]} \tag{3.8}$$

where d is the averaged Euclidean distance moved, as is stated in [139]. Only the peaks above 2 standard deviation σ and with a d higher than 0.02 ppm are considered as shifted. The intensity ratio was calculated as the height of the peak at bound state divided by the height at free state. A uniform scaling number averaged from the ratio of 4 different non-interacting residues was applied to normalise the values. This value compensate the decrease of intensity due to dilution or precipitation of the ^{15}N sample. Again, only those values above or below 2 standard deviation σ were considered. Those values with a normalised intensity more than 2.5 times were not considered in order to calculate the average and the standard deviation.
Chemical shift index (CSI) was calculated using the random coil values provided by the Biological Magnetic Resonance Data Bank (BMRB). For those phosphorylated serines, the random coil values for Cα and Cβ were calculated from [140].

WW mutants project

As previously described in PNAS under the title "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements" [122].
1D-^{1}H, 2D-NOESY and 2D-TOCSY were acquired at 285 K unless otherwise stated. The mixing time for 2D-NOESY was 120 ms. The spectra were acquired on a Bruker Avance III 600-MHz spectrometer equipped with a z-pulse field gradient unit and a triple (^{1}H, ^{13}C, ^{15}N) resonance probe head and further processed with TopSpin v3.5 Bruker Software. NOEs were manually assigned using CARA software and integrated using the batch integration method of XEASY package [141]. Crystallography & NMR system (CNS) vs 1.1 was used for structure calculation [131] using the unambiguously assigned intra and inter-molecular NOEs. 200 structures were calculated with the calculation protocol using 100,000 cooling steps. The protocol includes a water refinement software [132] with a modified protocols developed in house, which include all experimental restraints during refinement. The 20 lowest energy structures for each mutant were selected and their quality was checked with iCing [142] and PROCHECK-NMR [143] software. Finally, the structures represented in this thesis were depicted using UCSF Chimera [144].

3.3.2 Mass Spectrometry

Basic principles

Mass spectrometry (MS) is a analytical technique that determines the mass of a particle (isolated or as mixtures). A mass spectrometer separates previously ionised particles by its mass to charge ratio to finally record them in a ion detector.
There are many kinds of MS techniques, differing basically in the way how they perform these three steps. Firstly, it can ionize the particle by Matrix-Assisted Laser Desorption/ionization (MALDI) or electrospray ionization method (ESI), among many others. Later, MS sorts the resulting ions based on their mass/charge (m/z) ratio. Time-of-flight (TOF), quadrupole, ion traps or Fourier transform (FT) analysers are some of the most used ways to sort them. In the last step a MS detects the ions

thanks of an electron multiplier, micro-channel plate detectors (MCP) or other methods. An important aspect of MS is that it requires few sample to work despite the sample is not recovered. In this thesis only MALDI-TOF and ESI-quadropole devices were used, and so from now the explanation will focus on these kind of MS spectrometers.

The idea behind MALDI is to shoot the sample with a laser to desorb them from a suitable matrix, at the same time that they are ionizated by being protonated or deprotonated. It is considered a soft technique as it can ionize, without major fragmentation, many kinds of biomolecules such as proteins, DNA or sugars. After the sample have been ionized, the difference of time of flight (TOF) of the electrical accelerated particles sorts the particles through their m/z ratio. MALDI-TOF configuration is very common as they both works in a batch mode and they allow the measurement of a wide mass range.

ESI is also considered a soft technique. It generates the ions when a high voltage is applied to a flowing liquid creating an aerosol. At the time that the little drops loses their liquid, the particle is being charged until no liquid remains and the particle stays ionized. Quadropole analyser is made by four parallel cylindrical rods that creates an electrical field so that only particles with the right m/z ratio passes through it to the detector. ESI-quadropole can work in a continuous mode and thus it can be coupled to an HPLC as a final detector to check the mass of the eluted fractions.

Experimental procedures

Peptide ligation project. As previously described in Biopolymers (Peptide Science) under the title ”Addition of HOBt improves the conversion of thioester-amine chemical ligation” [119].

MALDI-MS. Mass spectra were acquired on a 4700 Proteomic analyser or on a 4800 Plus MALDI TOF/TOF Analyser (AB Sciex, Framingham, USA) calibrated with Calmix (Calmix 4700 Proteomics Analyser Calibrating Mixture). The mass spectra were recorded in positive reflector TOF mode in the m/z range 500–2000 or 1500–4500 at a fixed laser in-

tensity of 4800 using alpha-cyano-4-hydroxycinnamic acid (ACH) as a matrix. Spectra were analysed by Data Explorer software (Version 4.6, Applied Biosystems GmbH).

LC-MS. The ligation mixtures (30 μL, at final concentration of approx. 60 μmol/L for each time point) were injected into the HPLC-MS system (Waters, model Alliance 2796 with a quaternary pump and UV/Vis dual absorbance detector Waters 2487 connected with ESI-MS model Micromass ZQ). The separation was achieved on a Sunfire C18-column (internal diameter 2.1 mm, particle size 3.5 μm, length 100 mm) using a linear gradient from 10% or 20% to 100% aqueous acetonitrile (0.1% FA) in 8 min at a flow rate of 0.3 mL/min. The mass spectra were acquired for a mass range from m/z 500 to 2000 in positive ion mode using five different cone voltages ranging from 5 to 70 V. The TIC spectra used for peak integration corresponds to the cone voltage of 30 V and were analysed by Masslynx 4.0 software (Waters, Milford, USA).

WW mutants project. As previously described in PNAS under the title "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements" [122]. And **TGIF1-SMADs project**.

MALDI-MS. Mass spectra were acquired on a 4700 Proteomic analyser (AB Sciex) calibrated with Calmix (Calmix 4700 Proteomics Analyser Calibrating mixture). The spectra were recorded in positive linear mode or at positive reflector, when possible. 0.4 μl of sample were mixed with the same quantity of a sinapinic acid (SA) matrix solution. Spectra were analysed by Data Explorer software (Version 4.6, Applied Biosystems GmbH).

3.3.3 MicroScale Thermophoresis

Basic principles

MicroScale Thermophoresis (MST) is a technique used for quantify molecular interactions. It is based on the measurement of the displacement of a molecule caused by a gradient of temperature, an effect called "thermophoresis". This effect depends on the hydration shell, charge or size

of the molecule. If the molecule is bound to a ligand, the bound complex will have a different hydration shell, charge and size in comparison with the unbound one. Therefore both molecules will be displaced in a different rate when there is a gradient of temperature. This technique allows the determination of molecular interactions in their native state in a wide range of concentrations (nm to mM) [145].
The experiment works as following: the fluorescent sample is introduced in a narrow capillary by capillarity. Then the device start measuring the basal fluorescence by standard methods. Afterwards, an IR-laser is switched on and creates the temperature gradient. Because of the thermophoresis effect, the molecule moves away from the heat. As previous described, the molecule moves away in a different rate depending if it is bound to a ligand or not. This process is repeated with increasing equivalents of ligand. After the experiment, the normalised fluorescence (F_{norm}) is calculated from the fluorescence values with and without the IR laser. Finally, the (F_{norm}) is plotted versus the increasing concentrations of ligand and the K_D is derived.

Experimental procedures

As pTGIF1 (256-347) is not fluorescent, NT-647 fluorescence dye (NanoTemper Technologies, München, Germany) was bound to it. This dye only binds to cysteines (there are 3 in pTGIF1 (256-347) fragment) and only one cysteine per molecule is labelled. However, the cysteine where the dye binds varies from molecule to molecule. Following manufacturer instructions, 100 μl at 20 μM of pTGIF1 (256-347) reacted with 100 μl at 60 μM of dye in 20 mM $NaPO_4$, 80 mM NaCl at pH 6.4 buffer. After 35 min of reaction at room temperature in the dark, the free dye was separated from the resin with a PD-10 (General Electric, UK). The MST experiment was carried out mixing 5 μl of pTGIF1 (256-347) at 20 nM and 5 μl of SMAD2 at increasing concentrations doubling from 5.6 nM to 186 μM in 20 mM $NaPO_4$, 80 mM NaCl, 1 mM TCEP, 0.05% Tween at pH 6.4. 16 capillaries were analysed. Excitation power was set at 40% while MST was set at 20%.

3.3.4 Electrophoretic Mobility Shift Assay

Basic principles

Electrophoretic Mobility Shift Assay (EMSA) is a technique that detect the interaction between two molecules in a easy and fast way. It consists on a gel electrophoresis, similar to the SDS-PAGE, where the molecules (commonly DNA or RNA) are moved due to their size and charge. When a protein is added to the solution two things can happen: (1) the DNA do not bind to the protein and so, the DNA band is equal as the one without protein (2) the DNA actually binds to the protein. In this case, as the complex DNA-protein is bigger and has a different charge as the original DNA, the band shifts from its original position to a new one upper in the gel. In this experiment, only the DNA is labelled and thus the protein remains invisible. There are different labelling methods, radioactivity, fluorescence dye or biotin. In this thesis, only the fluorescence labelling was used because its security and simplicity. Overall, EMSA is a very sensible technique that requires little amount of DNA and protein allowing competition experiments at very reduced concentrations.

Experimental procedures

TGIF1 Homeodomain (150-248) mother solution was 114 μM in 20 mM $NaPO_4$, 150 mM NaCl, 1 mM TCEP, at pH 6.4 while the DNA was annealed in 20 mM HEPES, 150 mM NaCl, 1 mM sodium azide, at pH 7.5. All protein dilutions were done in 20 mM HEPES, 150 mM NaCl, 1 mM sodium azide, at pH 7.5. The protein concentration gradients were done in a serial dilution fashion. In each sample, 10 μl of protein were mixed with 3 μl of DNA and 10 μl of Orange dye. After the mixing, the samples were incubated for 30 min. Then, they were loaded in a 12% polyacrylamide gels under non-denaturating conditions. The buffer for the electrophoresis was 25 mM TRIS, 200 mM Glycine, pH 8.2. The gel was run 1h 15 min at 80V at room temperature, until the orange dye reach the bottom. Finally, the gels were revealed at the wavelength of 678/694 nm (excitation/emission) using a Typhoon imager (GE Healthcare, Uppsala Sweden). The EMSA assays were carried out together with Dr. Guca.

4 Results

4.1 Deciphering the binding of TGIF1 (256-347) to SMAD2

4.1.1 TGIF1 (256-347) is unstructured

The first step in the project was to design a fragment of TGIF1 that includes the region described to interact with SMAD2 (SID), between the residues 267 and 321 of TGIF1 [28]. We decided to perform our experiments with the TGIF1 fragment (256-347), which not only includes the SMAD2 binding region but also the contact region of HDAC1 [44], Axin-2 [51] and PHRF1 [38] (Figures 1.8 B and 4.1).

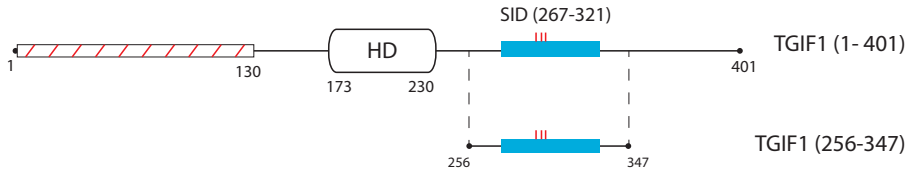


Figure 4.1: Scheme of TGIF1 (256-347) fragment in relation with the full-length TGIF1 protein. The SID region is shown in blue while the serine phosphorylations found in the region are indicated with a red stick.

In our first experiment, we prepared a ^{15}N labeled sample and recorded an ^{1}H-^{15}N HSQC of it (Figure 4.2). The spectrum shows a narrow distribution of the peaks, indicating that this region of TGIF1 lacks a stable tertiary structure. We proceed then to prepare a double labelled sample (^{15}N-^{13}C) in order to assign the amide resonances. With the information obtained combining the pair of CBCA(CO)NH and HNCANH experiments we could assign up to 78 of the 83 (94%) non-proline amino acids of the fragment (94 total amino acids).

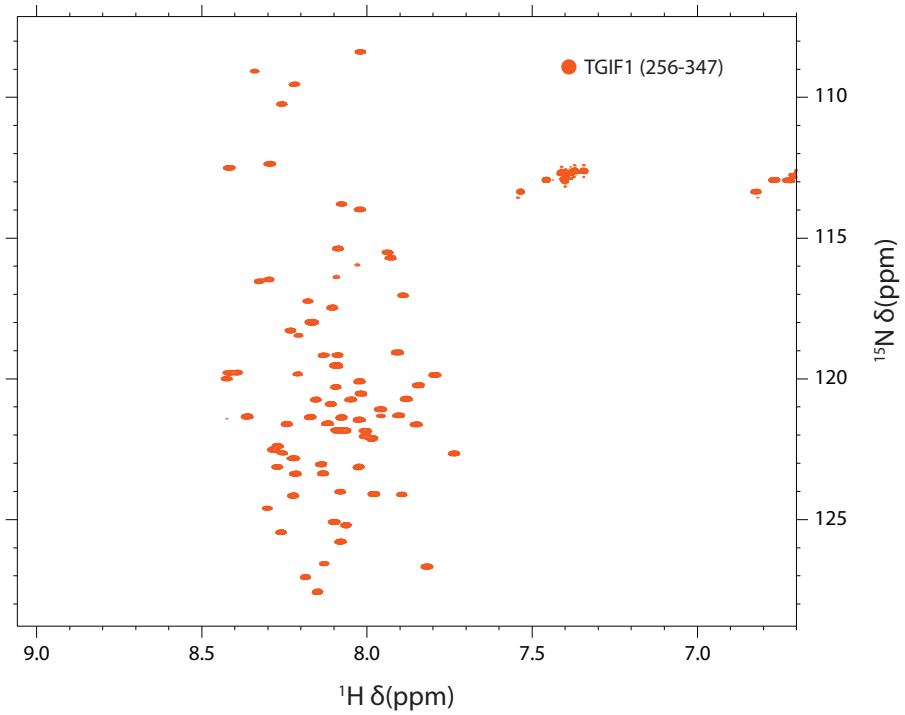


Figure 4.2: ^{1}H-^{15}N HSQC of TGIF1 (256-347). The poor dispersion of the peaks indicates a lack of stable tertiary structure. 91.6% of the expected peaks can be observed (76 of a total of 83 peaks).

The assignments of the carbon resonances also allowed us to compare the Cα and Cβ chemical shift (δ) of each amino acid with its random coil value (Chemical shift index (CSI)) (Figure 4.3). The lack of a tendency towards a positive (α-helix) or negative (β-sheet) difference versus the random coil values indicates the absence of differentiated α-helix or β-sheet structures, in agreement with the previous data obtained.

4.1.2 TGIF1 (256-347) does not interact with SMAD2-EEE (186-467)

In our next experiment we investigated the interaction between the TGIF1 (256-347) and SMAD2 protein using NMR. As explained in the introduction, SMAD2 binds to TGIF1 through part of MH1 and the full

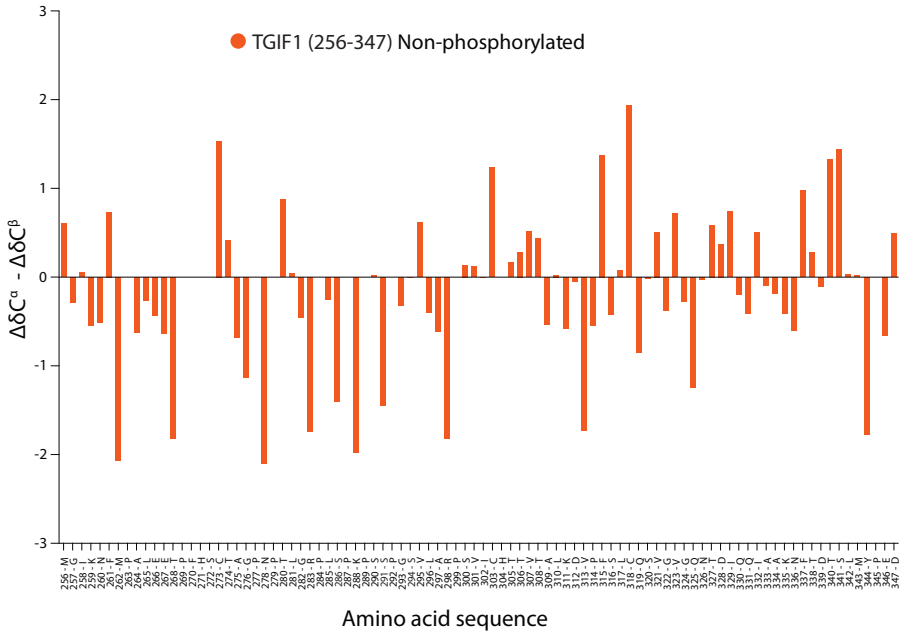


Figure 4.3: Chemical shift representation of $\Delta C\alpha$-$\Delta C\beta$ of TGIF1 (256-347). Tendency to negative values indicates the presence of β-sheet while a positive tendency reflects α-helix.

linker and MH2 domain [28]. Plus, as most of the proteins that interact with SMAD2 bind to the linker or MH2 domain of SMAD2 [20], we decided to start our titrating experiments with the linker-MH2 fragment of SMAD2 (amino acids 186-467). In addition, as previous experiments demonstrated [28], *in vivo* interaction is enhanced when TGF-β is present, or in other words, when SMAD2 is activated, hence, phosphorylated in the last three serines (464, 465, 467). In order to facilitate its obtention, glutamic acids can be used to mimics the effects of the phosphorylated serines [146]. Therefore, in our experiments we used a clone of SMAD2 (186-467) with the three serines 464, 465, 467 substituted by glutamic acid (from now, stated like SMAD2-EEE (186-467)).

The HSQC titration between the ^{15}N TGIF1 (256-347) with up to 7 equivalents of the ligand SMAD2-EEE (186-467) showed no variation in the δ neither in the intensity of the peaks, indicating that there is no inter-

action between them under the experimental conditions used (Figure 4.4 and Appendix Figure 7.3).

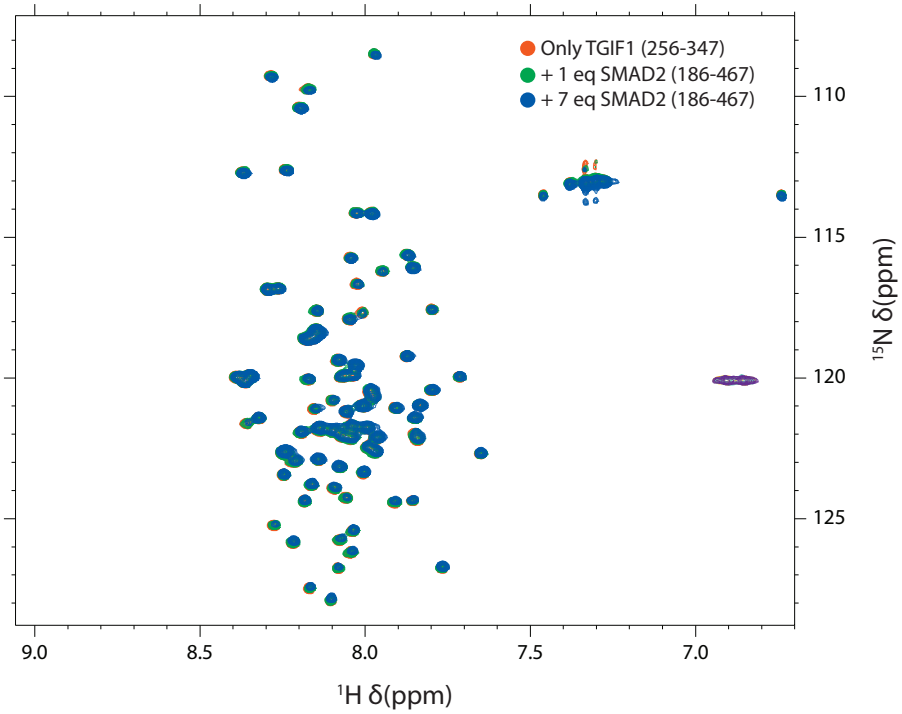


Figure 4.4: Superimposition of several ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (256-347) with increasing equivalents of SMAD2-EEE (186-467). Red colour refers to TGIF1 (256-347) alone; green to TGIF1 (256-347) + 1 equivalent of SMAD2-EEE (186-467) and blue to TGIF1 (256-347) + 7 equivalents of SMAD2-EEE (186-467).

4.1.3 p38α phosphorylates Ser286 and Ser291 of TGIF1 (256-347)

The negative result led us to think that the serines 286, 291 and 294, which have been found phosphorylated *in vivo* ([57], [111]) and are located in the middle of the TGIF1 region that interacts with SMAD2, may have a role in the regulation of the interaction. In particular, we

wanted to know if the presence of the phosphate groups may have some influence in the binding (Figure 4.5).

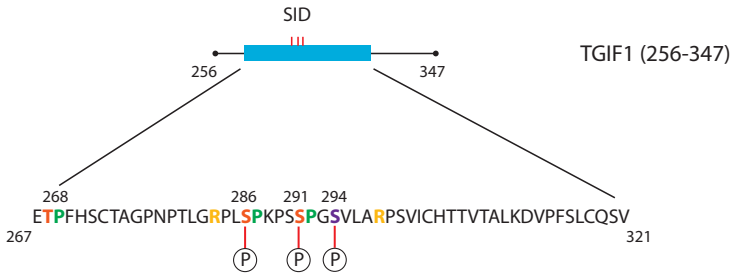


Figure 4.5: Detail of the TGIF1 sequence that is proposed to interact with SMAD2. In red there are the serines/threonines that are theoretically phosphorylated by p38α as they are before proline (in green). There are no more serines or threonines before proline in the rest of the construct sequence not shown. The phosphorylations that were found by HTP-MS methods are indicated by a P. Ser294, not theoretically phosphorylated by p38α but found *in vivo* with the phosphate group is labelled in purple. The arginines present in the fragment are highlighted in yellow.

In order to get the phosphorylated TGIF1 fragment, we first planned to apply the peptide synthesis strategy. As the fragment is too long (94 amino acid) to synthesise it by a single SPPS, we designed a ligation strategy with the improved cysteine-free direct aminolysis we have developed in our lab (Section 4.3). However, as explained in the subsection 4.4.4, the strategy for this particular peptide was unsuccessful.

Then we considered the possibility to phosphorylate the TGIF1 (256-347) fragment *in vitro* with specific kinases. In this sense, p38α kinase is known to phosphorylate serines or threonines before prolines [147]. In the TGIF1 (256-347) fragment there are three serine/threonines that could be phosphorylated by p38α as they are situated before a proline (Thr268, Ser286, Ser291; Figure 4.5). Two of them, Ser286 and Ser291, are two of the three serines that have been found phosphorylated *in vivo* [57]. Thus, we investigated the *in vitro* phosphorylation reaction by p38α kinase on a sample of TGIF1 (256-347) by NMR (Figure 4.6 A, B).

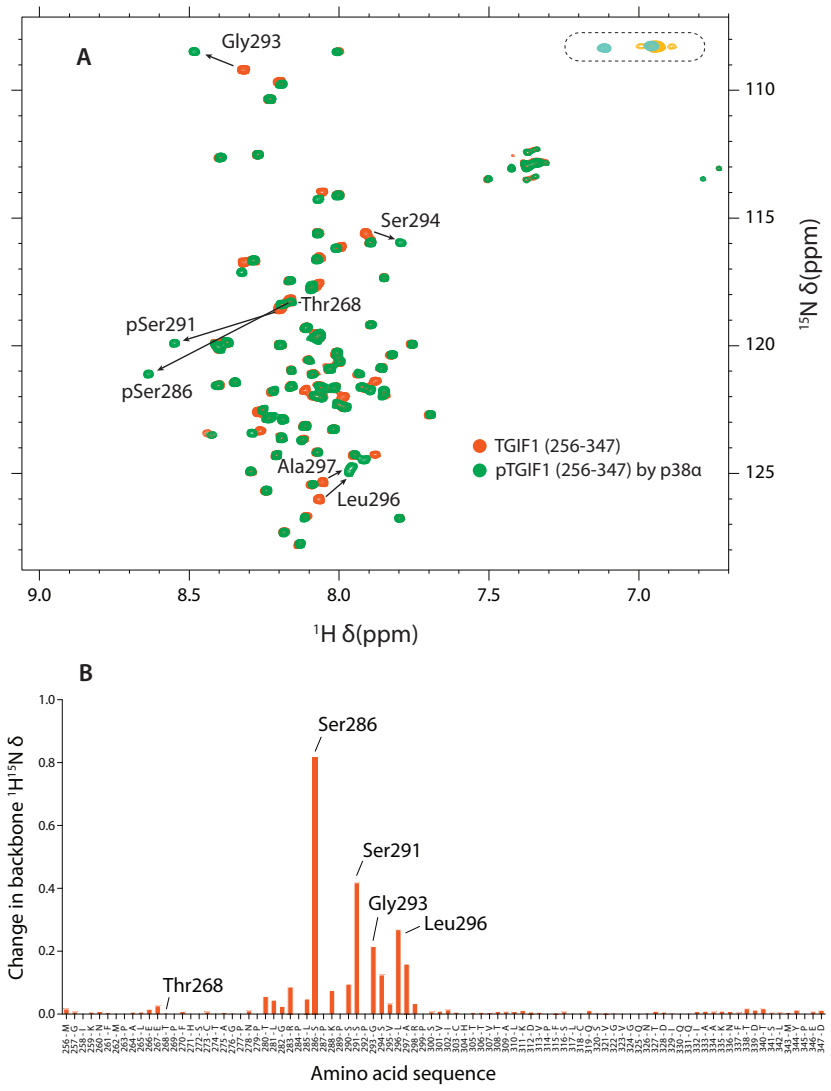


Figure 4.6: A) Superimposition of ^{1}H-^{15}N-SOFAST-HMQC spectra of TGIF1 (256-347) before (red) and after (green) the phosphorylation with p38α. The most relevant changes in δ and the unchanged position of threonine 268 are labelled. Surrounded by a dashed line are the side-chain peaks of the arginines: In yellow before phosphorylation and in light green after phosphorylation. **B**) Shift of every backbone ^{1}H-^{15}N δ caused by p38α phosphorylation. Prolines and non-assigned amino acids have a value of 0.

Interestingly, we found a great shift of the ^{1}H-^{15}Nδ relative to the Ser286 and Ser291 (but not to the Thr268) and smaller changes relative to the amino acids surrounding them. As it has been reported previously ([129], [140]), a large shift of a serine to higher chemical shift (meaning less shielded spin nuclei) indicates its phosphorylation. We therefore conclude that the addition of the p38α causes the specific phosphorylation of Ser286 and Ser291. Curiously, Ser286 and Thr268 have a very similar chemical shift and they appear superimposed in the ^{1}H-^{15}N HMQC spectrum before the phosphorylation. Therefore, in order to confirm the assignment, we also assigned the backbone ^{1}H-^{15}N peaks of TGIF1 (256-347) after p38α phosphorylation by performing 3D experiments (Appendix Figure 7.2).

Another shift can be observed in Figure 4.6 A. Located at low δ of both dimensions, a couple of negative peaks shift their positions due to the phosphorylation event (indicated by dashed line in Figure 4.6 A). These peaks likely belong to the ^{1}H-^{15}N side-chains of the two arginines present in the sequence of TGIF1 (256-347) (their peaks actually resonate at around ^{15}N 87 ppm but as the spectrum frequency is too narrow, the spectrum is folded, and this feature is reflected by the negative sign of the resonances). These shifts may indicate the presence of salt bridges between the phosphate group and the side-chain of the arginines [148]. This interpretation is supported by the fact that both arginines are situated just before and after the phosphorylated serines, specifically at -3 of Ser286 and at +7 of Ser291 (indicated in yellow in the Figure 4.5).

To investigate if the serines are getting phosphorylated at the same time or if the reaction is sequential, we performed a Real-Time NMR experiment. In this assay, we followed the appearance of the new peaks forming (and the disappearance of the old peaks) in the course of the phosphorylation reaction. We determined that the Ser286 was phosphorylated first (with half of population phosphorylated after 50.0 min) while the Ser291 took 191.8 min to be half phosphorylated (Figure 4.7).

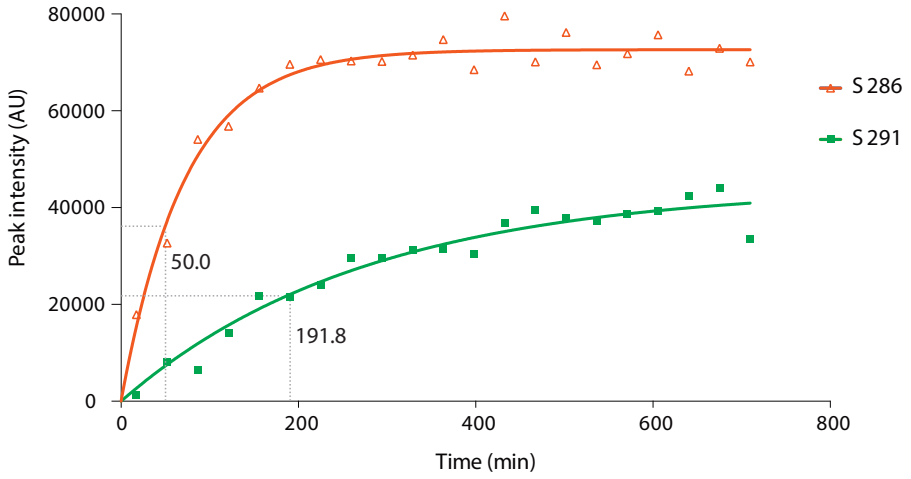


Figure 4.7: Kinetic of the appearance of the phosphorylated serines 286 and 291 peaks. The intensity versus time peaks are fitted to a mono-exponential equation.

pTGIF1 (256-347) by p38α does not change its secondary structure after phosphorylation.

As shown in Figure 4.6 A, the peak distribution in the ^{1}H-^{15}N HMQC spectra of pTGIF1 (256-347) by p38α did not change significantly, although some peaks have shifted due to the phosphorylation. Moreover, the CSI values of the phosphorylated TGIF1 (256-347) by p38α expose a similar distribution (Figure 4.8). Overall, we can conclude that TGIF1 (256-347) does not undergo any big structural rearrangement after the phosphorylation and it remains essentially unstructured.

4.1.4 pTGIF1 (256-347) by p38α spectrum does not change significantly after the addition of SMAD2-EEE (186-467)

To investigate the binding of phosphorylated TGIF1 (256-347) by p38α and SMAD2-EEE (186-467), we performed a NMR titration experiment on a ^{15}N-pTGIF1 (256-347) by p38α sample (100% phosphorylated at

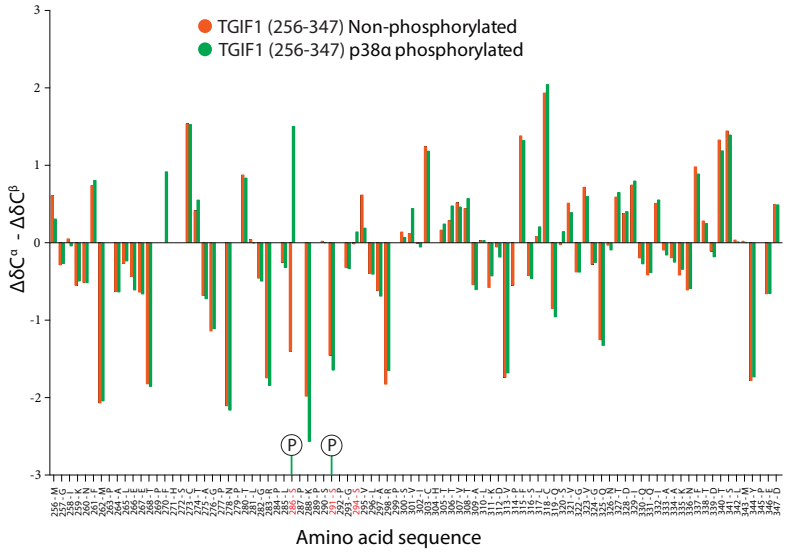


Figure 4.8: Superimposition of chemical shift index (CSI) values ($\Delta C\alpha$-$\Delta C\beta$) of TGIF1 (256-347) before and after being phosphorylated by p38α. Tendency to negative values indicate the presence of β-sheet while a positive tendency reflects α-helix. The circled P represents the phosphorylated serines. Labelled in red are all three serines found phosphorylated *in vivo*.

S286; 50% at S291) titrated by SMAD2-EEE (186-467). The results of the titration, showed in Figure 4.9 and Appendix Figure 7.4, revealed three chemical shift perturbations (CSPs): pSer286 and pSer291 are shifted an average distance of 0.021 ppm and 0.022 ppm, respectively, just above the cut-off and the third shift of 0.056 ppm, corresponds to one peak located at the bottom left region. This peak, and the one next to it, did not appear in the assignment spectra and therefore we do not know to which amino acid refers to. We speculate that they refer to some of the histidines that we could not observe (and assign) before. These two new peaks also change their intensity (1.70 and 0.78 times, comparing the final and the initial points of the titrations) as new peak 1 moves and overlaps the second one during the addition of SMAD2-EEE (186-467). Finally, Arg283 also increases its intensity (1.47 times) during the titration. All the changes are summarised in table 4.1. Overall, the small shifts of the δ and the intensity could indicate a weak interaction between pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467).

Table 4.1: List of the assigned peaks that changed in the titration between pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467).

Peak	Intensity variation (times)	CSP, *d* (ppm)
New peak 1	1.70	0.056
New peak 2	0.78	-
Arg 283	1.47	-
pSer 286	-	0.021
pSer 291	-	0.022

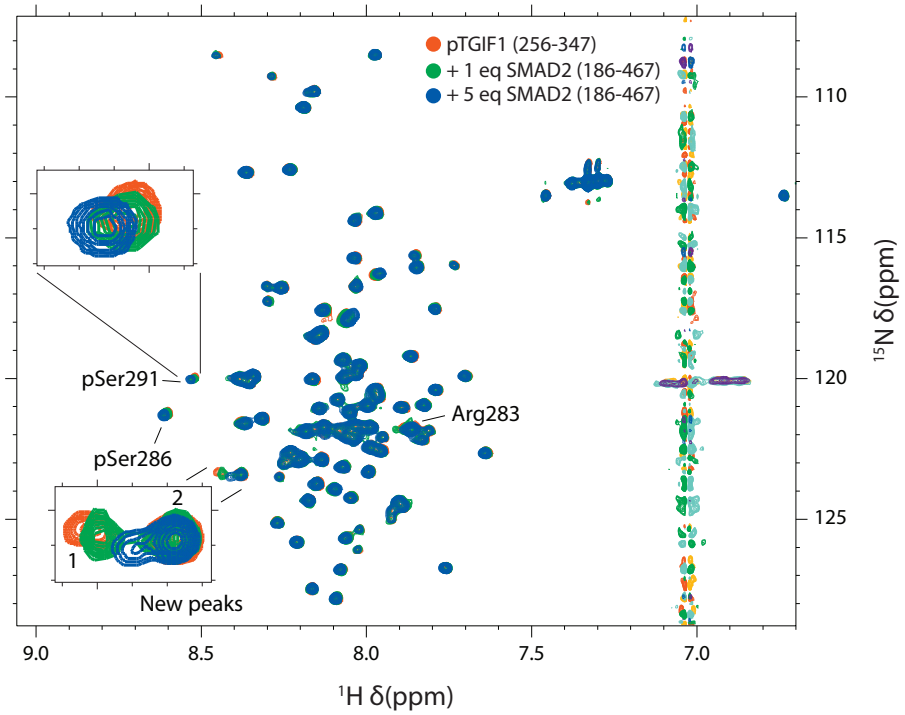


Figure 4.9: Superimposition of several ^{1}H-^{15}N-SOFAST-HSQC of pTGIF1 (256-347) by p38α with increasing equivalents of SMAD2-EEE (186-467). Red colour refers to pTGIF1 (256-347) alone; green to pTGIF1 (256-347) + 1 equivalent of SMAD2-EEE (186-467) and blue to pTGIF1 (256-347) + 5 equivalents of SMAD2-EEE (186-467). The verticals peaks located at ^{1}H 7.0 are related to an excess of PMSF added.

4.1.5 A direct interaction between pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467) was not detected by MST

In order to confirm the interaction between pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467) we proceed to analyse the interaction by MST technique. The flat curve of the normalised fluorescence indicates an absence of interaction between the two proteins. However, the last point, at very high concentration of SMAD2-EEE (186-467) may suggest some interaction at mmolar range (Figure 4.10).

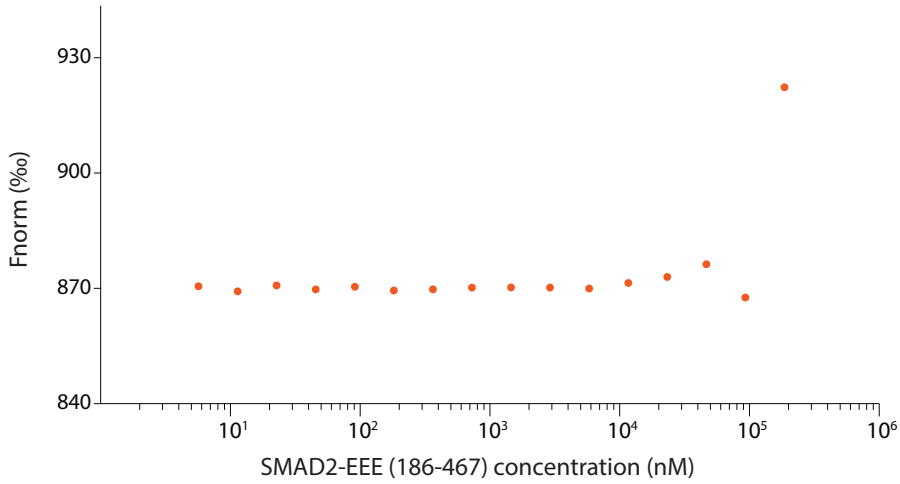


Figure 4.10: MicroScale Thermophoresis (MST) result on the interaction between pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467). The flat curve of the normalised fluorescence (Fnorm) versus increasing concentration of SMAD2-EEE (186-467) point out an absence of interaction.

4.1.6 pTGIF1 (256-347) by p38α is further phosphorylated by CK1

As mentioned earlier, High-throughput (HTP)-MS experiments found three phosphorylations in the TGIF1 (256-347) fragment [57], [111]. We have reproduced *in vitro* two of them (Ser286 and Ser291) using p38α

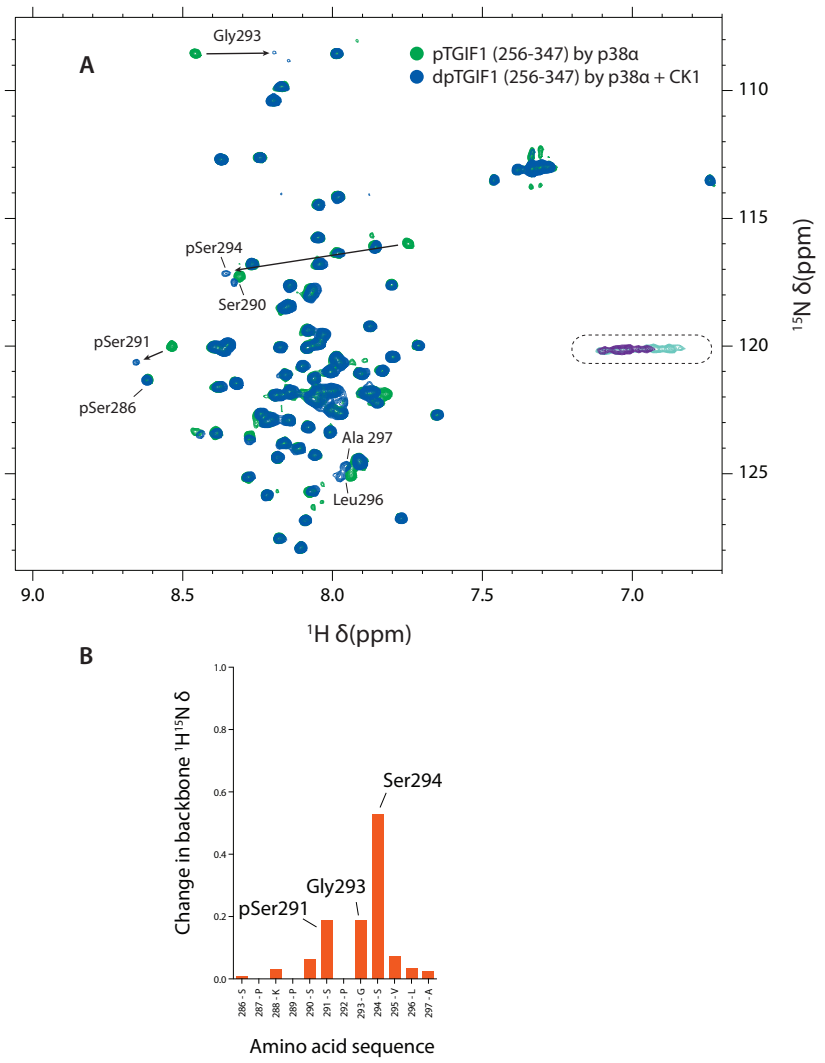


Figure 4.11: A) Superimposition of ^{1}H-^{15}N-SOFAST-HSQC spectra before (green) and after (blue) the phosphorylation with CK1 on a pTGIF1 (256-347) by p38α. The most relevant changes of δ are labelled. The most relevant changes in δ are labelled. Surrounded by a dashed line are the side-chain peaks of the arginines. In light green are labelled the negative peaks before phosphorylation while in purple are labelled the negative peaks after phosphorylation. **B**) Shift of the backbone^{1}H-^{15}N δ from the surrounding Ser294 amino acids. Prolines have a value of 0.

kinase, but still we lacked the phosphorylation over Ser294. To explore whether Ser294 could also be phosphorylated, we employed Casein Kinase I (CK1). CK1 is a kinase that has pSXXS/T as it most effective recognition motif [149], where the target serine or threonine is located at +3 of an already phosphorylated serine. Since Ser294 has the phosphorylated Ser291 at -3 (286-pSPKPS**pS**PG**S**-294) we decided to use CK1 to phosphorylate Ser294. Thus, we added CK1 kinase to a previously double-phosphorylated ^{15}N-pTGIF1 (256-347) by p38α. The figure 4.11 A shows a superimposition of a ^{1}H-^{15}N-SOFAST-HSQC spectrum before and after the phosphorylation by CK1. The shift from the Ser294 and the amino acids that surrounded it confirm the selective phosphorylation of Ser294 (Figure 4.11 B). Besides, we can also observe a shift from the side-chain of the arginines, indicating that some contacts have changed due to the presence of the phosphate group, in a similar way as explained before.

4.1.7 TGIF1 (256-347) spectrum does not change significantly after the addition of SMAD2-MH1 (10-174)

The absence of a strong direct contact between TGIF1 (256-347) or its double phosphorylated version and SMAD2-EEE (186-467) led us to test whether the MH1 domain of SMAD2 was participating in the interaction. Accordingly, we titrated the SMAD2-MH1 (10-174) fragment on ^{15}N-TGIF1 (256-347) using NMR (Figure 4.12 and Appendix Figure 7.5). The addition of 5 equivalents of SMAD2-MH1 (10-174) decrease the intensity of three peaks: Cys273 (to 0.66 times), Cys303 (0.60) and Leu317 (0.75) while all others keeps constant their intensity and their position (Table 4.2). This result indicates a weak interaction of both molecules, similarly to pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467).

Table 4.2: List of the assigned peaks that changed in the titration between TGIF1 (150-248) and SMAD4-MH1 (10-140).

Peak	Intensity variation (times)
Cys 273	0.66
Cys 303	0.60
Leu 313	0.75

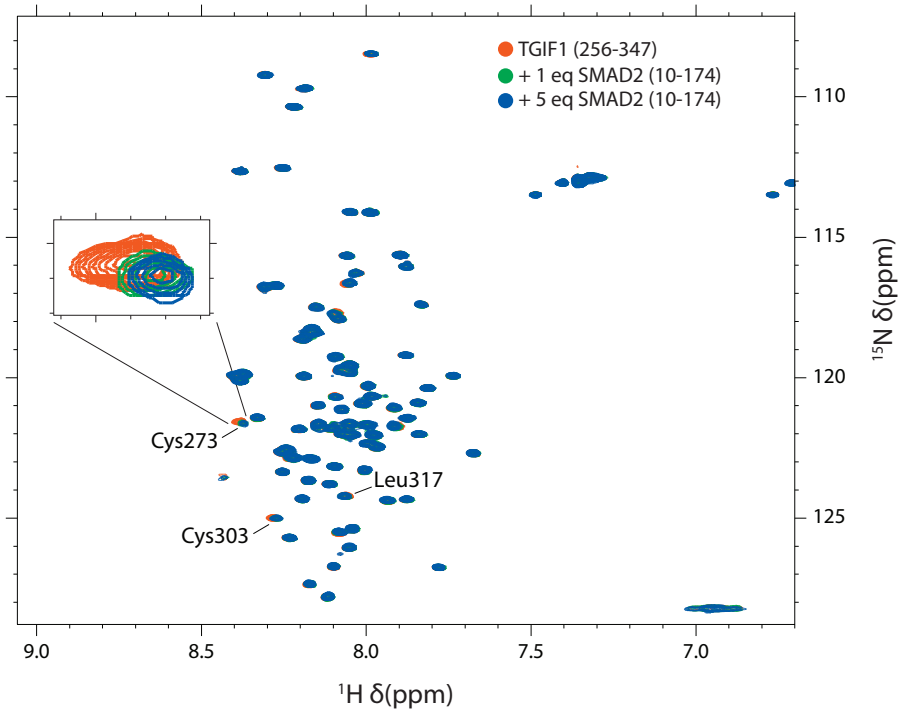


Figure 4.12: Superposition of several ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (256-347) with increasing equivalents of SMAD2-MH1 (10-174). Red colour refers to TGIF1 (256-347) alone; green to TGIF1 (256-347) + 1 equivalent of SMAD2-MH1 (10-174) and blue to TGIF1 (256-347) + 4 equivalents of SMAD2-(10-174). The peaks that are affected by the addition of SMAD2-MH1 (10-174) are labelled.

4.2 Investigating the interaction between TGIF1 homeodomain (150-248) and SMAD proteins

4.2.1 Characterisation of the TGIF1 homeodomain (150-248)

In our project we were also interested in the role of the TGIF1 homeodomain in the interactions with the SMAD proteins. As explained before, recent findings showed an interaction between the N-terminal region of the HOXC9 HD and SMAD4-MH1 [150]. Using NMR and EMSA techniques, we wanted to explore if this interaction also occurs with TGIF1 and whether this contact may be essential for the TGIF1 and SMAD2 interaction. In this sense, we designed a construct of TGIF1 that contains the homeodomain and the N-terminal region, TGIF1 (150-248), Figure 4.13.

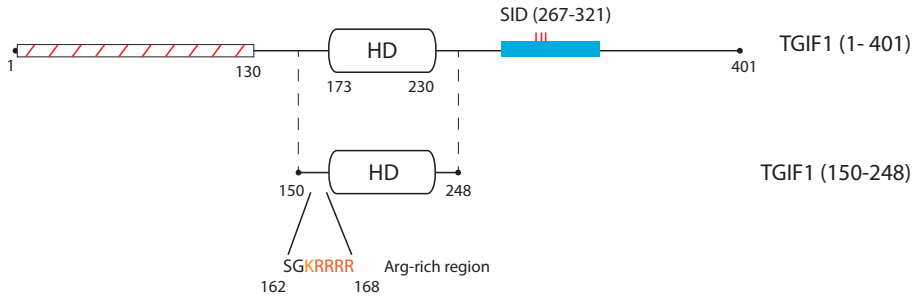


Figure 4.13: Construct of TGIF1 (150-248) with the detail of the homeodomain and the N-terminal arginine-rich region.

First of all, we compared the ^{1}H-^{15}N-HSQC spectrum of TGIF1 (150-248) with the resonances of the already published NMR structure of TGIF1 homeodomain (171-248) (PDB id: 2LK2, BMRB Entry 17971). As we can see in Figure 4.14, most of the peaks superimpose with the 2LK2 structure. The difference in the chemical shifts can be attributed to different buffer conditions of both samples. Moreover, we can also distinguish new peaks that are not present in the resonances corresponding to the sample deposited in the BMRB. In particular, three new peaks appear

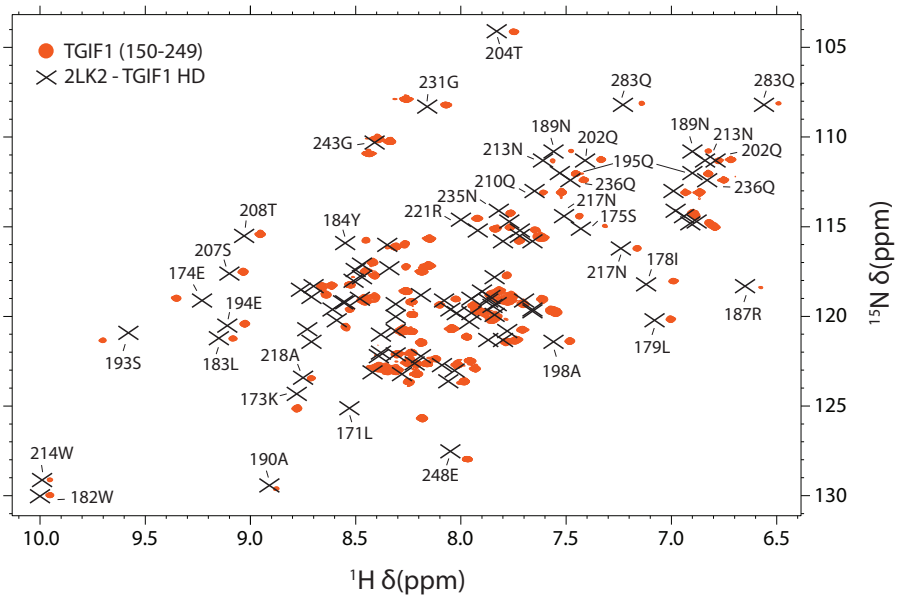


Figure 4.14: Superimposition of ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (150-248) (in red) with the peaks taken from the already published TGIF1 HD structure (171-248) (black crosses; PDB id: 2LK2). The assignment of those isolated peaks is shown.

in the glycine region ($<$ 113 ppm of δ ^{15}N), in agreement with the three new glycines that our construct has in the N-terminal region, missing in the 2LK2 structure. Finally, the wide distribution of the peaks (6.5 to 10.0 in ^{1}H δ dimension) indicates the presence of a well-folded structure.

In order to further confirm the activity of the TGIF1 (150-248) fragment we titrated it with a 16-mer DNA that contains two copies of the TGIF1 binding DNA canonical sequence [26] (Table 3.3). The final spectrum with 3 eq of DNA reveals a strong interaction between the protein and the DNA as most of the TGIF1 resonances shift and vary their intensity of the peaks (Figure 4.15).

Although we have not assigned the construct, we can deduce several residues – mainly from the isolated peaks – by comparing our spectrum with the already published 2LK2. Still, we miss the assignment of the N-terminal region that our construct has and it is not present in the 2LK2

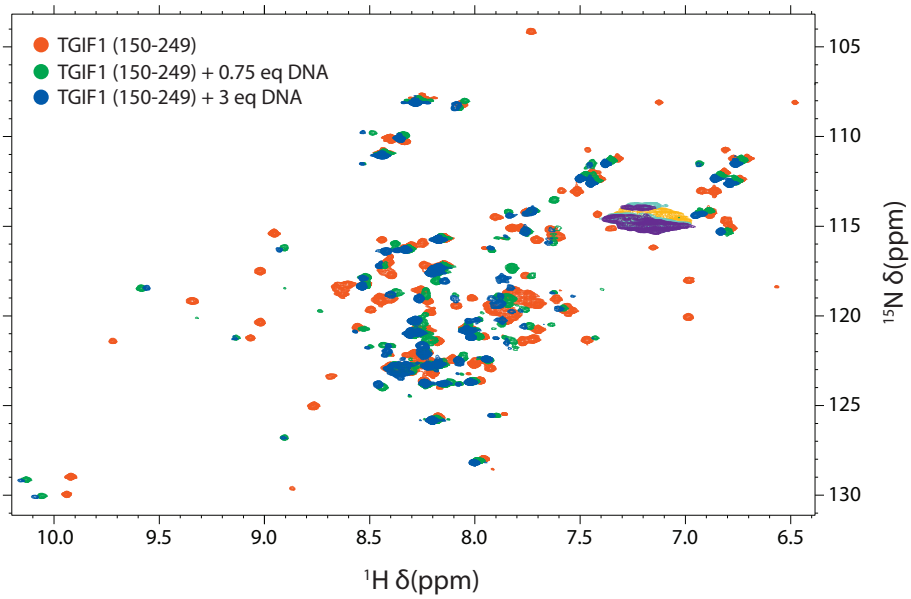


Figure 4.15: Superimposition of ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (150-248) with increasing equivalents of DNA1 (ATTGACAGCTGTCAAT). Red colour refers to TGIF1 (150-248) alone; green to TGIF1 (150-248) + 0.75 equivalents of DNA and blue to TGIF1 (150-248) + 3 equivalents of DNA.

construct. Thus, in the next sections only the change of the deduced peaks are described.

4.2.2 TGIF1 homeodomain (150-248) interacts with SMAD2/4-MH1

Our next step was to explore by NMR titration experiment if TGIF1 (150-248) also interacts with SMAD4-MH1 as HOXC9 does [150]. Thus, we titrated SMAD4-MH1 (10-140) domain on ^{15}N-TGIF1 (150-248) (Figure 4.18 and Appendix Figure 7.6). After the addition of 1.6 equivalents of SMAD4-MH1 several changes have been observed. Two peaks belonging to Thr208 and the side-chain of Gln210 changed their positions, 0.121 ppm and 0.062 ppm, respectively. Interestingly, the peak intensity of some resonances was also significantly affected by the addition of SMAD4-MH1. The peak corresponding to the Hϵ-Nϵ of the Trp214

Table 4.3: List of the assigned peaks that changed during the NMR titration between TGIF1 (150-248) and SMAD4-MH1 (10-140).

Peak	Intensity variation (times)	CSP, *d* (ppm)
Hϵ-Nϵ Trp 214	3.32	-
Glu 174	2.68	-
Hϵ22 Gln 210	1.86	-
Lys 173	1.75	0.046
Ile 178	1.75	-
Ser 175	1.70	-
Hδ22 Asn 217	1.61	-
Thr 208	-	0.121

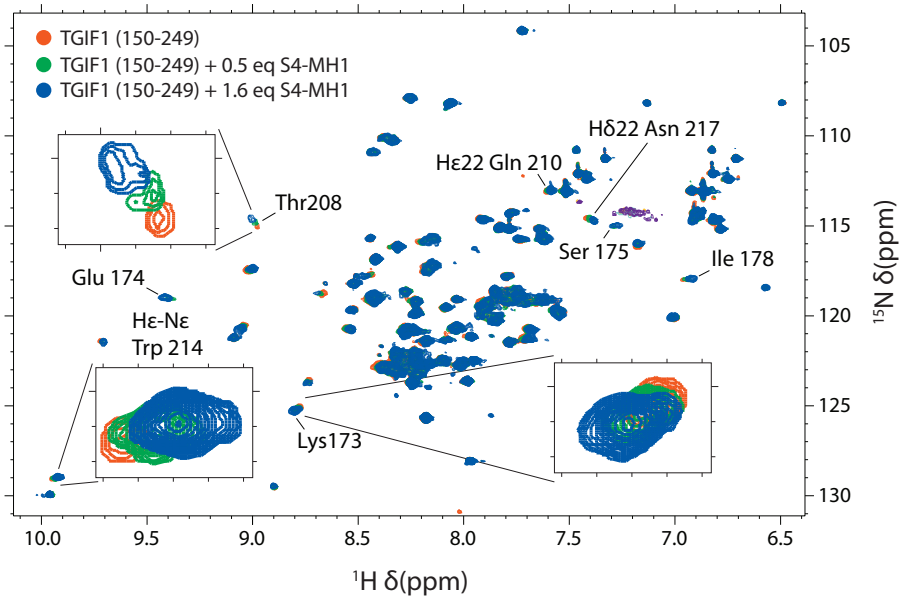


Figure 4.16: Superimposition of ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (150-248) with increasing equivalents of SMAD4-MH1 (10-140). Red colour refers to TGIF1 (150-248) alone; green to TGIF1 (150-248) + 0.5 equivalents of SMAD4-MH1 (10-140) and blue to TGIF1 (150-248) + 1.6 equivalents of SMAD4-MH1 (10-140). The peaks that are affected by the addition of SMAD4-MH1 (10-140) are labelled.

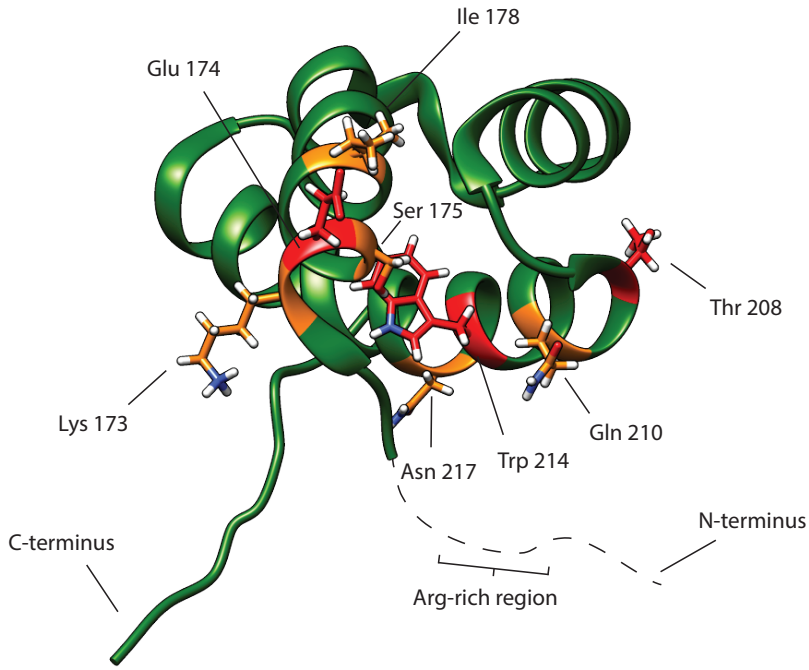


Figure 4.17: Structure of TGIF1 HD with the residues affected upon the presence of SMAD4-MH1 (10-140) highlighted. The red labelled amino acids are those with higher variation (214Trp, 174Glu , 208Thr) while in orange are labelled all the others modified residues. The residues are shown on the previously solved TGIF1-HD structure (PDB: 2LK2). With a dashed line is represented the extra region at N-terminus that the TGIF1 construct (150-248) has in comparison with the 2LK2.

increased its intensity by 3.32 times while the Glu174 increased 2.68 times. The table 4.3 summarises the others peaks affected more than the threshold values. All residues that varied during the titrations are localised in the same N-terminus region of the protein, which led to the conclusion that the interaction between TGIF1 homeodomain (150-248) and SMAD4-MH1 (10-140) affects to an specific region of the homeodomain (Figure 4.17).

The evidence of the interaction between the TGIF1 (150-248) and SMAD4-MH1 (10-140) made us to consider whether this interaction could also

Table 4.4: List of the assigned peaks that changed in the NMR titration between TGIF1 (150-248) and SMAD2-MH1 (10-174).

Peak	Intensity variation (times)
Hδ22 Asn 217	2.04
Ser 175	1.90
Hδ22 Asn 213	1.80
Asn 217	1.78
Hδ21 Asn 213	1.72
Hϵ22 Gln 210	1.72
Glu 174	1.72

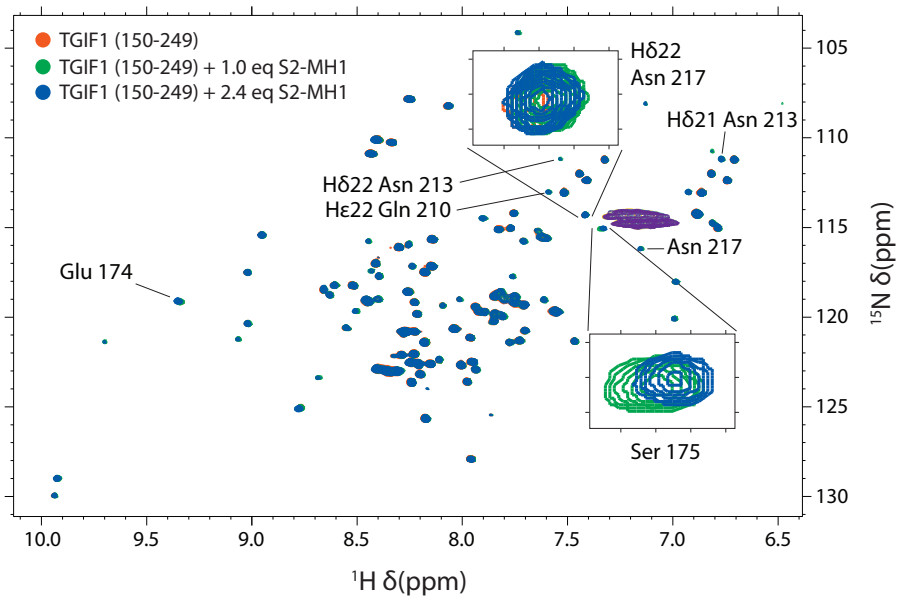


Figure 4.18: Superposition of several ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (150-248) with increasing equivalents of SMAD2-MH1(10-174). Red colour refers to TGIF1 (150-248) alone; green to TGIF1 (150-248) + 1.0 equivalent of SMAD2-MH1(10-174) and blue to TGIF1 (150-248) + 2.4 equivalents of SMAD2-MH1(10-174). The peaks that are affected by the addition of SMAD2-MH1(10-174) are labelled.

occur between the TGIF1 (150-248) and SMAD2-MH1 (10-174). Subsequently, we titrated SMAD2-MH1 (10-174) on ^{15}N-TGIF1 (150-248) (Figure 4.18 and Appendix Figure 7.7). The analysis of the ^{1}H-^{15}N HSQC revealed that no peak was shifted during the titration. However, up to seven peaks varied their intensity (see Table 4.4) upon the addition of SMAD2-MH1 (10-174). Interestingly, some of these peaks (Glu174, Ser175, Hϵ22 Gln210, Hδ22 Asn217) are the same as the ones that varied in the TGIF1 (150-248) - SMAD4-MH1 (10-140) titration (Figure 4.19). This evidence led to the conclusion that both proteins interact similarly, affecting at the same N-terminus region of TGIF1 (150-248). Moreover, the overall intensity shifts are higher in the SMAD4-MH1 (10-140) titration, which maybe is related to a higher affinity of the SMAD4-MH1 (10-140) in comparison with the SMAD2-MH1 (10-174).

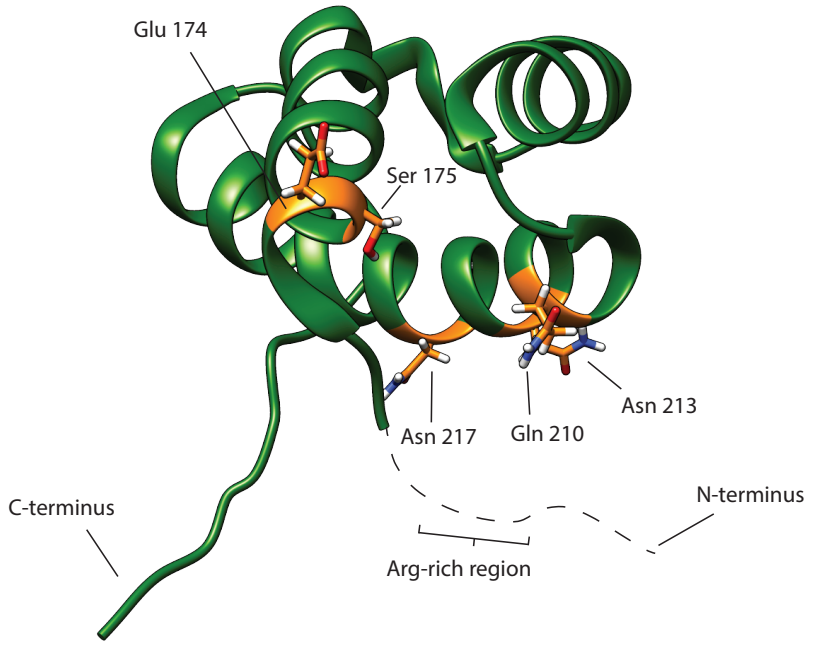


Figure 4.19: Structure of TGIF1 HD with the residues affected upon the presence of SMAD2-MH1 (10-174) highlighted in orange. The residues are shown on the previously solved TGIF1-HD structure (PDB: 2LK2). With a dashed line is represented the extra region at N-terminus that the TGIF1 construct (150-248) has in comparison with the 2LK2.

After the observation of the interaction of TGIF1 homeodomain (150-248) with SMAD4-MH1 (10-140) and SMAD2-MH1 (10-174), we wanted to analyse these interactions with a complementary technique. Taking advantage of the binding of TGIF1 HD with its canonical DNA, we performed electrophoretic mobility shift assays (EMSA) in order to confirm the interactions between proteins already seen. In the first experiment (Figure 4.20 A) we increased the concentration of TGIF1 homeodomain (150-248) while keeping constant the Cy5-DNA concentration at 3.75 nM. The shift produced by the complex between the DNA and the protein starts to be visible at 31.3 nM of protein. Increasing the protein concentration, the complex is formed by two molecules of protein by one of DNA, in agreement with the two protein binding sites present in the palindromic DNA. This assay was fundamental to determine the best protein concentration of the protein for the following experiments. For these experiments we choose a TGIF1 homeodomain (150-248) concentration between 54 and 212 nM because in this range we can see all four bands (ssDNA, dsDNA, complex 1:1 and complex 2:1) in equilibrium with each other. Therefore, any slight change in the binding between TGIF1 HD and the DNA will affect greatly in the abundance of all four bands.

In our next assay, keeping constant the DNA and the TGIF1 homeodomain (150-248), we added increasing concentrations of SMAD4-MH1 (10-140) (Figure 4.20 B). At both constant concentrations of TGIF1 (150-248) (54 and 212 nM), the bands corresponding to the 2:1 complex almost disappeared after the addition of two equivalents of SMAD4-MH1 (10-140) over TGIF1 (150-248). As SMAD4-MH1 (10-140) alone with the DNA did not interact, we can conclude that the presence of SMAD4-MH1 (10-140) interferes with the formation of the complex between the TGIF1 homeodomain and the DNA by interacting with the TGIF1 (150-248).

We obtained equivalent results when we tested the effect of the addition of SMAD2 (10-174) to TGIF1 HD (150-248) (Figure 4.20 C). In this case the binding between TGIF1 (150-248) and DNA is already disrupted with one equivalent of SMAD2 (10-174). These results added evidence about the interaction between SMAD2/4-MH1 with TGIF1 (150-248) al-

ready observed by NMR.

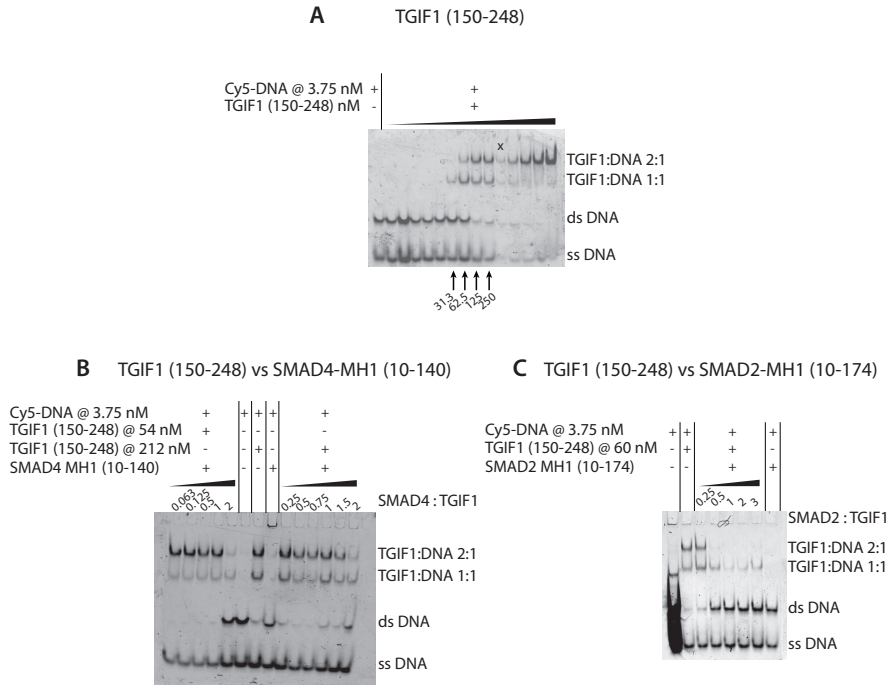


Figure 4.20: EMSA assays carried out with TGIF1 (150-248) and its canonical DNA labelled with Cy5 (Cy5-ATTGACAGCTGTCAAT) (at constant concentration of 3.75 nM in all assays). **A**) Simple assay with increasing concentrations of TGIF1 (150-248) (from 0.97 nM to 4 μM, doubling at each line) with while keeping constant the DNA concentration. A cross on a line indicates an error on the sample loaded. **B**) Competition assay with increasing concentrations of SMAD4-MH1 (10-140) while keeping constant the concentrations of TGIF1 (150-248) and DNA. Two different competitions assays were done at 54 and 212 nM concentrations of TGIF1 (150-248). The concentration of SMAD4-MH1 (10-140) in the control line with only the DNA was 500 nM. **C**) Competition assay with increasing concentrations of SMAD2-MH1 (10-174) while keeping constant the concentrations of TGIF1 (150-248) and DNA. The concentration of SMAD2-MH1 (10-174) in the control line with only the DNA was 180 nM, equivalent to 3 times 60 nM. The DNA alone control was taken from a sample of a previous day, and presents degradation, smirring along the line. ss and ds DNA stands for single and double stranded DNA, respectively.

4.3 TGIF1 (256-347) binds to TGIF1 homeodomain (150-248)

In order to decipher a possible intramolecular interactions in TGIF1 protein, we performed NMR titration experiments between the two fragments of the protein. The titration, represented in Figure 4.22 and Appendix Figure 7.8, revealed the δ displacement of three peaks (Table 4.5) and the decrease of intensity of two more peaks, up to 0.5 times in the case of Cys303. Altogether, these results pointed out an interaction between the two fragments of TGIF1. Engrossingly, Cys273, Cys303 and Leu317 were also disturbed when SMAD2-MH1 (10-174) was added to ^{15}N-TGIF1 (256-347), indicating a similar mode of interaction between these two proteins and TGIF1 (256-347).
Additionally, we also tested by EMSA whether TGIF1 (256-347) could also disturb the interaction between DNA and TGIF1 homeodomain (150-248). Indeed, after the addition of 4 equivalents of TGIF1 (256-347), the band corresponding to the 1:1 complex disappeared, similarly as with the addition of SMAD2/4-MH1, previously reported.

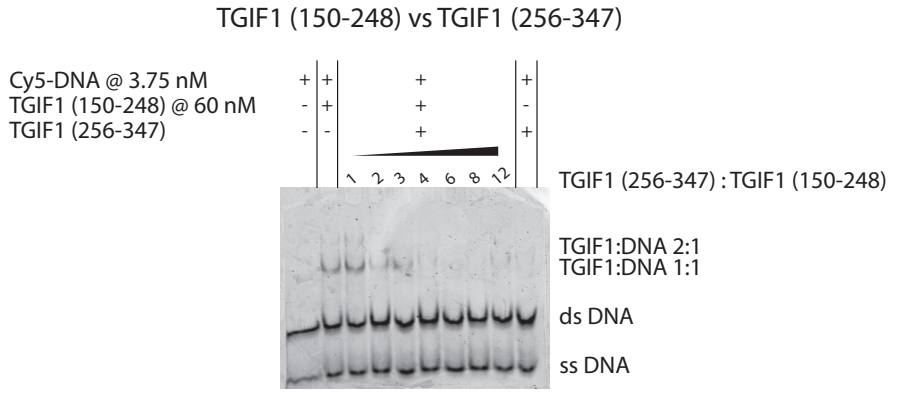


Figure 4.21: EMSA competition assay with increasing concentrations of TGIF1 (256-347) while keeping constant the concentrations of TGIF1 (150-248) and DNA. The concentration of TGIF1 (256-347) in the control line with only the DNA was 720 nM, equivalent to 12 times 60 nM. ss and ds DNA stands for single and double stranded DNA, respectively.

Table 4.5: List of the assigned peaks that changed in the titration between TGIF1 (256-347) and TGIF1 (150-248).

Peak	Intensity variation (times)	CSP, *d* (ppm)
Cys 303	0.50	-
Leu 317	0.58	-
*Ser 272 or Thr 306	-	0.026
Cys 273	-	0.025
Thr 274	-	0.023

*The assignment was confusing for this peak.

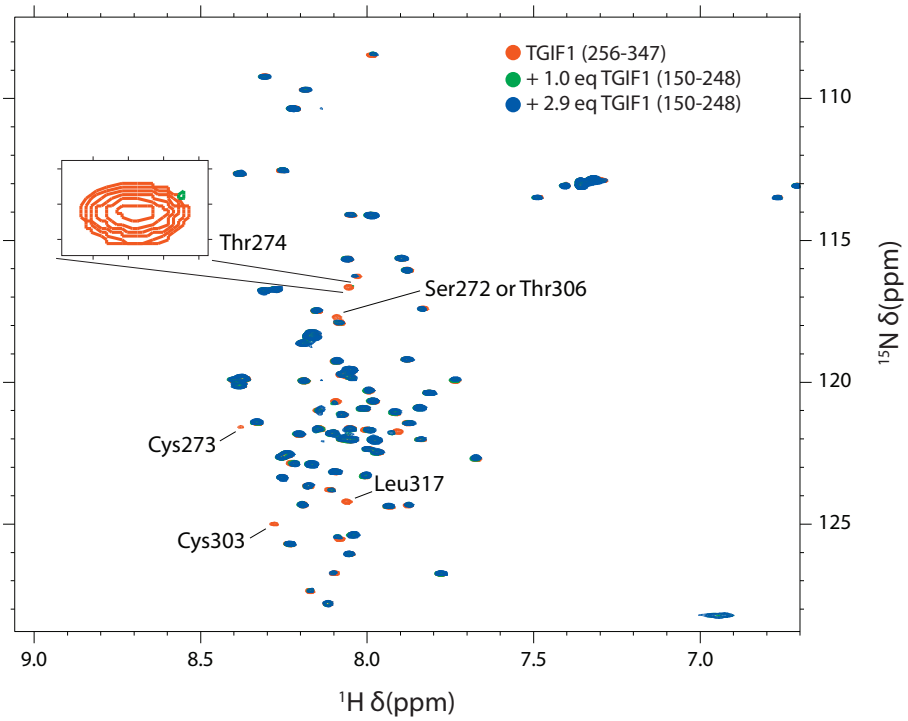


Figure 4.22: Superposition of several ^{1}H-^{15}N-SOFAST-HSQC of TGIF1 (256-347) with increasing equivalents of TGIF1 (150-248). Red colour refers to TGIF1 (256-347) alone; green to TGIF1 (256-347) + 1.0 equivalent of TGIF1 (150-248) and blue to TGIF1 (256-347) + 2.9 equivalents of TGIF1 (150-248). The peaks that are affected by the addition of TGIF1 (150-248) are labelled.

4.4 Study about the cysteine-free direct aminolysis ligation reaction

This project was done together with Dr. Todorovski and the result is published in Biopolymers (Peptide Science) under the title "Addition of HOBt improves the conversion of thioester-amine chemical ligation" [119] (Appendix).
The first step towards our aim was to synthesise the peptides that will be used in the peptide ligation trials. As mentioned in the introduction, the direct aminolysis cysteine-free ligation reaction requires that the N-terminal peptide ends in an thioester on its C-terminus. On the other hand, the C-terminal peptide do not have this requisite and thus can end in amide or acid moiety at its C-terminus (Figure 4.23).

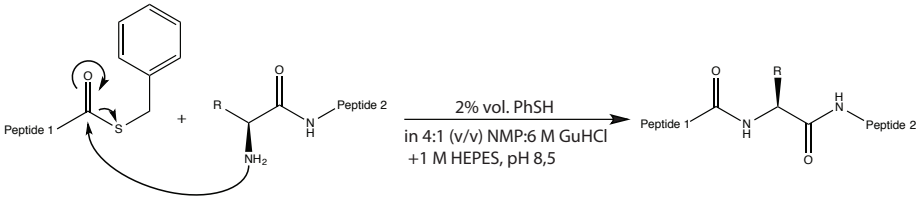


Figure 4.23: Direct aminolysis cysteine-free peptide ligation scheme.

In total, we synthesised 13 peptides (Tables 4.6). Five of them (1-5) were synthesised in sulfamylbutyryl resin in order to get the peptides ending with a benzyl mercaptan group at their C-terminus. The rest of the peptides (6-13) were synthesised in a common Rink-Amide resin. Because of the more complex procedure of cleaving the peptide from the sulfamylbutyryl resin, the yields of the thioesters peptides are considerably lower than those peptides ending with an amide group.

The peptide sequences were selected in order to test different amino acids combinations at the ligation junction while keeping all the other reaction parameters constant. Thus, peptides 1-3 only differ in the last amino acid and sequences 6-9, 10-12 in the first one. Seven of these peptides (1-3, 6-9) are 10 amino acid long as we also wanted to check the efficacy of the ligation method in longer sequences. Moreover, we

Table 4.6: List of the peptides synthesised. SB stands for Sulfamylbutyryl and RA for Rink-Amide AM.

	Peptide sequence	m/z theo.	m/z exp.	Yield* (%)	Resin used
1	Ac-GASATVSPL**G**-SC_7H_7	1007.46	1007.50	30	SB
2	Ac-GASATVSPL**V**-SC_7H_7	1049.51	1049.58	63	SB
3	Ac-GASATVSPL**S**-SC_7H_7	1037.49	1037.55	64	SB
4	Ac-LYRA**G**-SC_7H_7	727.87	727.53	20	SB
5	Ac-PSpSPGS**V**-SC_7H_7	858.87	858.25	48	SB
6	NH_2-**G**GPSPLGFLG-$CONH_2$	900.49	900.57	96	RA
7	NH_2-**V**GPSPLGFLG-$CONH_2$	942.54	942.61	92	RA
8	NH_2-**L**GPSPLGFLG-$CONH_2$	956.56	956.62	75	RA
9	NH_2-**C**GPSPLGFLG-$CONH_2$	946.48	946.46	50	RA
10	NH_2-**G**SPGYS-$CONH_2$	566.25	566.47	89	RA
11	NH_2-**A**SPGYS-$CONH_2$	580.26	580.48	59	RA
12	NH_2-**Y**SPGYS-$CONH_2$	672.29	672.55	86	RA
13	NH_2-**L**ARPSVI-$CONH_2$	754.49	754.44	59	RA

*The yield was calculated before the HPLC purification, except for 10 to 13, which were calculated after HPLC purification.

synthesised standard peptide sequences (4 and 10-12) [83, 76]. Finally, we also synthesised a phosphorylated thioester (5) that corresponds to the TGIF1 protein sequence (289-295; serine 291 phosphorylated) and the C-terminal amide (13) related C-terminal TGIF1 sequence (296-302). Notice that this sequence of TGIF1 is placed in the central interaction region with SMAD2 and the phosphorylated 291 serine is one of those found phosphorylated by MS experiments (See 4.1.3). All the thioester sequences are C-terminus acetylated to avoid intramolecular reaction.

4.4.1 HOBt increases the conversion, but not the rate, of peptide ligation reactions

In order to check if the addition of HOBt may improve the conversion or the rate of the ligation reactions, we performed the following series of experiments: for every combination of thioester and amide peptide, we did two parallel experiments. Starting from the same reaction solution, we divided the volume in two tubes. In one tube we added two equivalents of HOBt (based on the amount of peptide thioester) while in the second we added the same volume of ligation buffer. At each time point (different depending of the amino acid combination), we collected a little fraction of each tube, stopped the reaction and analysed it by LC-MS. We then integrated the area of the peak corresponding to the expected product. The area was plotted against time, obtaining a kinetic representation of the reaction (Figures 4.24 and 4.25).

We started with the Gly-Gly combination (in all cases, the first amino acid is the thioester and the second is the C-terminal peptide) (Figure 4.24, A; Table 4.7, I). Based on the kinetic plot, we can see that in both conditions (with and without HOBt) the reaction ended with a similar amount of product after 36 h, although in the first hours the reaction without HOBt generated more product. However, considering standard Gly-Gly sequences (C, III), the presence of HOBt increases the amount and the rate of the product formed. We continued the experiments with Ser-Gly combination (B, II). In this case, the total amount of product clearly increases when we added HOBt in the reaction solution. We observe similar result in the combination Gly-Ala (D, IV) with standard sequences although in this peptide combination the reaction without HOBt generated more product at the first hours of reaction. In contrast, the presence of a cysteine in the C-terminal peptide (E, V) enhances the speed of the ligation reaction (at 8 h the reaction was already finished) but the addition of HOBt, in this reaction, decreases the conversion.

In another set of experiments, we analysed the presence of HOBt when β-branched or aromatic amino acids are present at the ligation junction. These amino acid are known to have a lower conversion and a slower reaction rate most probably due to sterical hinderance [97]. Re-

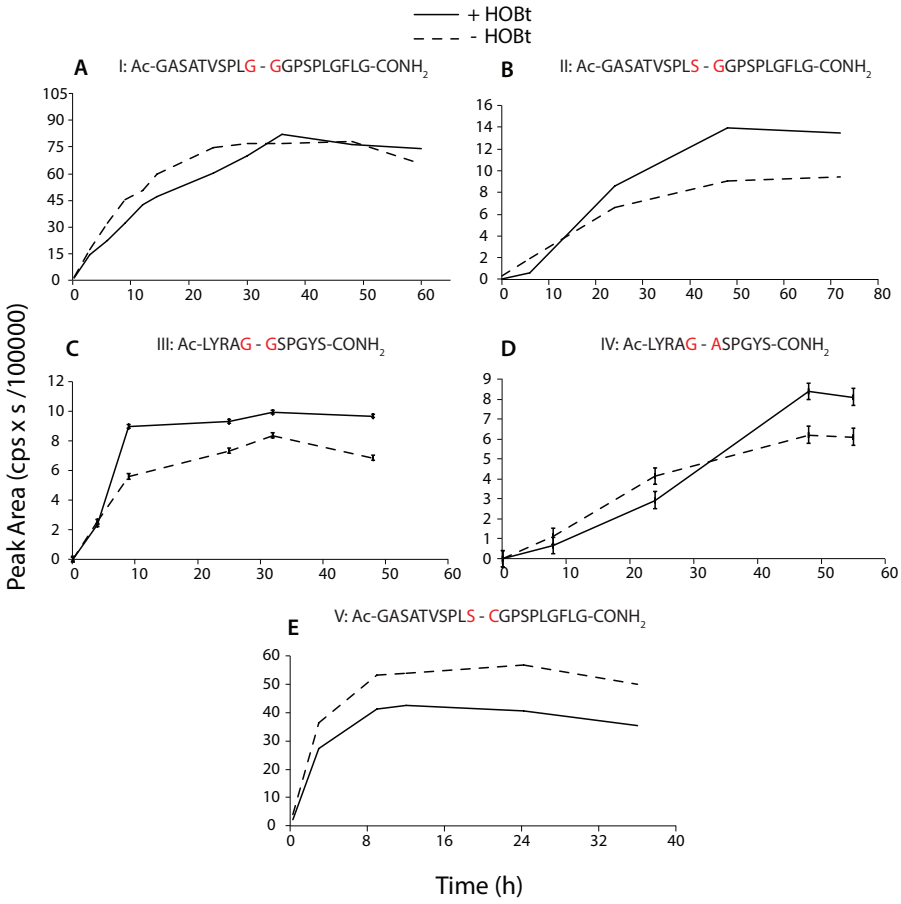


Figure 4.24: Kinetics of the ligated product formation in ligation reactions with HOBt (solid line) or without (dashed line). **A**) Ligation I; between peptides 1 and 6. **B**) Ligation II; between peptides 3 and 6. **C**) Ligation III; between peptides 4 and 10. **D**) Ligation IV; between peptides 4 and 11. **E**) Ligation V; between peptides 3 and 9. The error bars were calculated from duplicate injections in the LC-MS. In the Appendix Table 7.1 are listed the product ligation cations of every reaction.

markably, when we reacted Ser-Val (Figure 4.25 A; Table 4.7 VI), the presence of HOBt increases dramatically the conversion, but not the rate of the reaction. We got the same outcome when the combinations were Val-Val (B, VII), Val-Leu (C, VIII) or Gly-Tyr (D, IX).

Table 4.7: Summary of the peptide ligation results in different amino acid combinations.

	Ligation Combinations	Yield (Y) / Conversion (C) (%) without HOBt	with HOBt
I	Ac-GASATVSPL**G**-SC_7H_7 + NH_2-**G**GPSPLGFLG-$CONH_2$	n.d	n.d
II	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**G**GPSPLGFLG-$CONH_2$	n.d	n.d
III	Ac-LYRA**G**-SC_7H_7 + NH_2-**G**SPGYS-$CONH_2$	83 (Y)	89 (Y)
IV	Ac-LYRA**G**-SC_7H_7 + NH_2-**A**SPGYS-$CONH_2$	n.d	n.d
V	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**C**GPSPLGFLG-$CONH_2$	n.d	n.d
VI	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**V**GPSPLGFLG-$CONH_2$	10 (Y)	36 (Y)
VII	Ac-GASATVSPL**V**-SC_7H_7 + NH_2-**V**GPSPLGFLG-$CONH_2$	5 (Y)	34 (Y)
VIII	Ac-GASATVSPL**V**-SC_7H_7 + NH_2-**L**GPSPLGFLG-$CONH_2$	17 (C)	37 (C)
IX	Ac-LYRA**G**-SC_7H_7 + NH_2-**Y**SPGYS-$CONH_2$	33 (Y)	57 (Y)
X	Ac-PSpSPGS**V**-SC_7H_7 + NH_2-**L**ARPSVI-$CONH_2$	19 (C)	32 (C)

To confirm our results further, we performed the ligations reactions in a large scale, which it allowed us to purify the products and quantify the reaction yields (Figure 4.26 and Table 4.7, combinations III, VI, VII and VIII). For the rest of the reactions, we calculated, when possible, the conversion based of the integrated ions ratio (in the TIC spectra from the LC-MS) between the ligated product at the last point of reaction

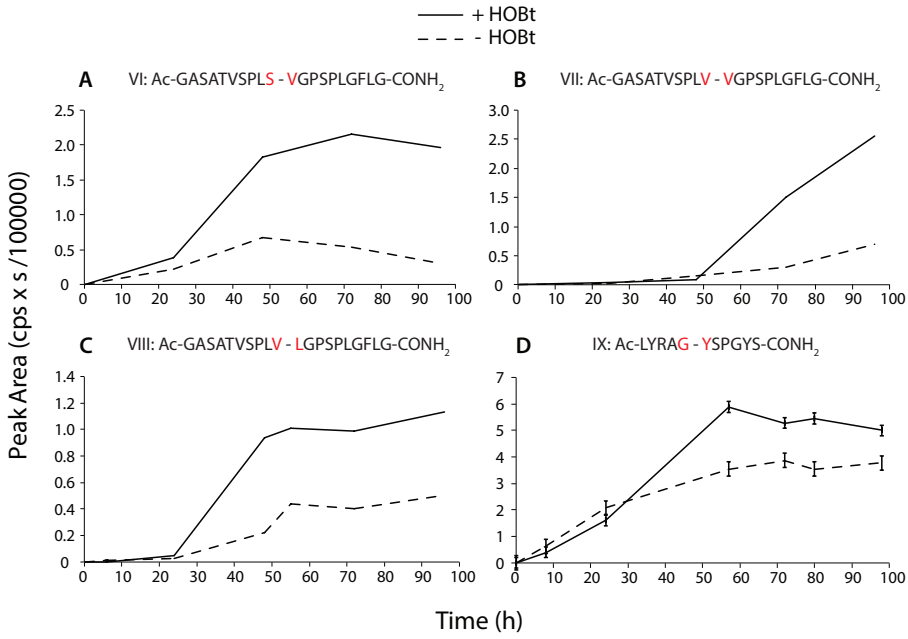


Figure 4.25: Kinetics of the ligated product formation in ligation reactions with HOBt (solid line) or without (dashed line). **A**) Ligation VI; between peptides 3 and 7. **B**) Ligation VII; between peptides 2 and 7. **C**) Ligation VIII; between peptides 2 and 8. **D**) Ligation IX; between peptides 4 and 12. The error bars were calculated from duplicate injections in the LC-MS. In the Appendix Table 7.1 are listed the product ligation cations of every reaction.

versus peptide thioester at the beginning of the reaction. In all cases, the yield/conversion obtained was higher when HOBt was present in the ligation, increasing in most cases more than 20 points, although in any case the final yield/conversion reaches more than 60%. Confirming our previous results, the effect of HOBt is limited in the combination Gly-Gly (III) but it is more intense when hindered amino acids such as valine or tyrosine (VI, VII, VIII and IX) are present in the ligation junction.

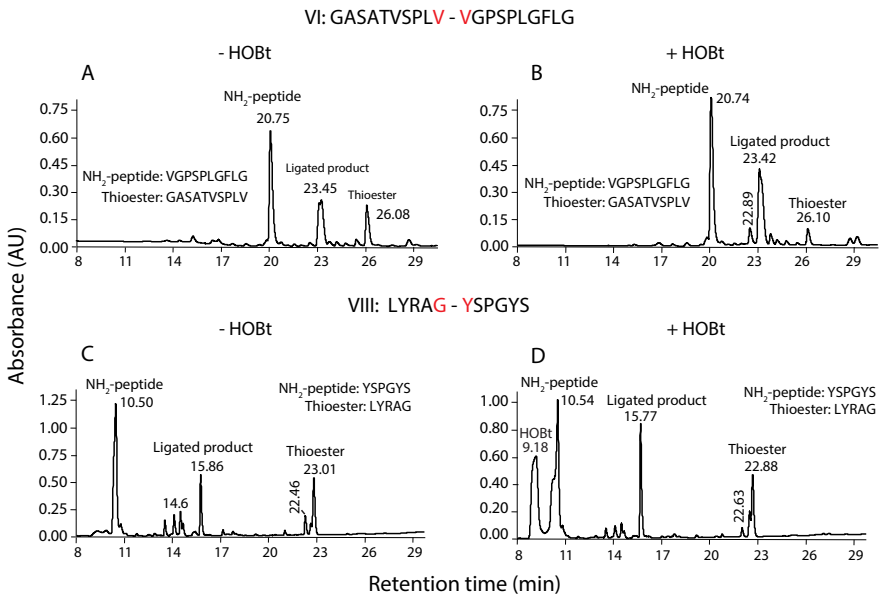


Figure 4.26: Example of an RP-HPLC of two peptide ligation reactions. In both cases the last point of the reaction was taken. **A-B**) Ligation VI (Table 4.7) without (A) and with HOBt (B); at 96 h of reaction. **C-D**) Ligation VIII (Table 4.7) without HOBt (C) and with HOBt (D) at 98 h of reaction.

4.4.2 Improving the aminolysis ligation reaction

Looking for further improvements in the conversion and rate of the peptide ligation reactions, we tested several variations of the ligation conditions (Figure 4.27). An increase of the number of equivalents of HOBt, up to 10, does not enhance the quantity of product formed, but decreases it (Figure 4.27 A). This effect could be explained by the drop of the pH caused by the acidic nature of HOBt.
We then focused our attention on the thiophenol. It is known that in the cysteine based ligations reactions the replacement of thiophenol by 4-(carboxymethyl)thiophenol (MPAA) can augment the rate and the conversion of the reaction [74]. However, in our system, the mentioned substitution causes a notorious drop in the conversion (Figure 4.27 B). One possible explanation is that the rate of the transthioesterification step was very low in this case (Appendix Figure 7.9) and therefore the

observed conversion rate was worse than in the case of thiophenol.

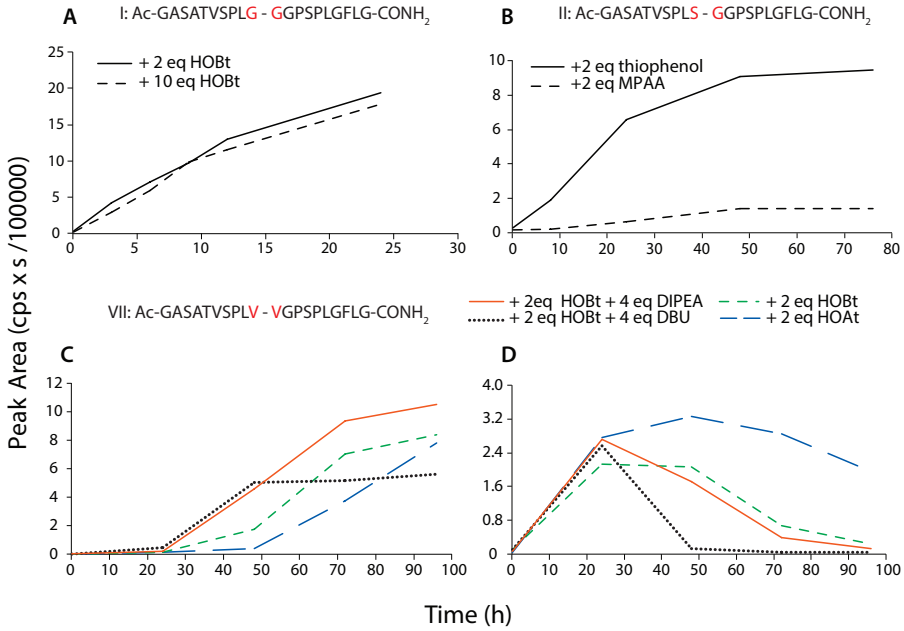


Figure 4.27: Kinetics of the ligated product formation in ligation reactions (**A**-**C**). **A)** Ligation I. Comparison between 2 equivalents of HOBt (solid line) and 10 equivalents HOBt (dashed line). **B)** Ligation II. Comparison between 2 equivalents of thiophenol (solid line) and 2 equivalents of MPAA (dashed line). **C)** Ligation VII. Comparison between HOBt + DIPEA (red solid line); HOBt + DBU (black dotted line); HOBt (green short dashed line) and HOAt (blue long dashed line). **D)** Same as C but following the kinetics of the phenyl thioester. In the Appendix Table 7.1 are listed the product ligation cations of every reaction.

In peptide synthesis, the addition of a base activator can improve the rate of a reaction [151]. We therefore tested whether the presence of *N*,*N*-diisopropylethylamine (DIPEA) or 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) could increase the rate of the reaction (Figure 4.27 C). Our results show that although both bases improve the conversion up to certain point (50 h of the reaction), for a reaction times longer than 50 h, only the presence of DIPEA had notable influence on the conversion

rate. The different behaviour of both bases can be explained by the different presence of the intermediate phenyl thioester, formed after the transthioesterification step (Figure 4.27 D). The plot indicates that in the case of the DIPEA the formed phenyl thioester lasts longer in the reaction mixture while when DBU is present, this intermediate disappears after 50 h, at the same point where the conversion of the product stops. This observation highlights the importance of the thiophenol intermediate in the reaction because, without the presence of it, the ligation reaction stops.

Finally, we tested the potential role of 1-hydroxy-7-azabenzotriazole (HOAt) as an alternative of HOBt. Indeed, HOAt has been demonstrated superior than HOBt in coupling reactions [152] because combines in one molecule the effect of HOBt itself and a tertiary amine, enhancing altogether its catalytic effect. In addition, its lower pKa (3.28 versus 4.60 of HOBt) and thus, potentially better leaving group, makes HOAt a good candidate to improve the effect of HOBt in the ligation reaction. However, our results indicate that the final conversion is similar using either HOAt or HOBt (Figure 4.27 C, D). Interestingly, the product appears much later in the presence of HOAt than when we use HOBt and thus can be hypothesised that with more reaction time the conversion will be higher with HOAt, although more reaction time would also increase side reactions. This phenomena could be attributed to the fact that HOAt requires more time to react with the thioester intermediate keeping the thiophenolic ester more time in the solution, as shown in Figure 4.27 D.

Overall, only the addition of 4 equivalents of DIPEA improves slightly the previous results obtained using HOBt only while the use of the others reagents turn out to be a worse option.

On the other hand, we were wondering whether the ligation reaction would be feasible with a free N-terminus thioester, that is, being chemoselective to the N-terminus of the other peptide. The logic behind this idea is to use the ligated product as an input to a second ligation reaction using a new thioester containing peptide, allowing the synthesis of longer and more complex peptides in a stepwise manner. To test this

hypothesis, we synthesised three more thioester peptides (Table 4.8) with the same sequence as their homologs 1, 2, 3 (to 14, 15, 16, respectively) but with a free N-terminus. In this set of experiments we evaluated the presence of the ligation product respect to the intramolecular cyclisation product. In the combinations we tried, shown in Figure 4.28, the formation of the cyclisation product overwhelms the obtention of the ligation product. Therefore, we dismissed this option.

Table 4.8: List of the peptides synthesised. SB stands for Sulfamylbutyryl.

	Peptide sequence	m/z theo.	m/z exp.	Yield* (%)	Resin used
14	NH_2-GASATVSPL**G**-SC_7H_7	965.54	965.58	15	SB
15	NH_2-GASATVSPL**V**-SC_7H_7	1007.53	1007.55	48	SB
16	NH_2-GASATVSPL**S**-SC_7H_7	995.49	995.46	44	SB

*The yield was calculated before the HPLC purification.

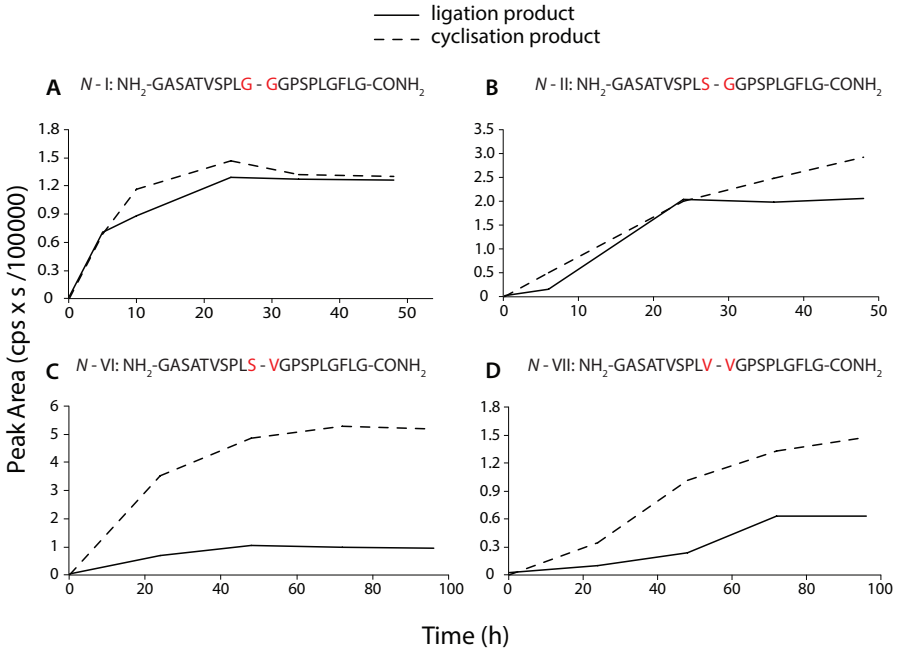


Figure 4.28: Kinetics of the ligated product formation (solid line) or the cyclisation product (dashed line) in different ligation reactions. The *N* before the reaction number indicates that for this reaction a free N-terminal thioester was reacted instead an acetylated one. **A**) Ligation *N* - I; between peptides 14 and 6. **B**) Ligation *N* - II; between peptides 16 and 6. **C**) Ligation *N* - III; between peptides 16 and 10. **D**) Ligation *N* - IV; between peptides 15 and 11.

4.4.3 Proposal of a mechanism of reaction

The results we have obtained led us to propose a plausible reaction mechanism for the direct aminolysis peptide ligation in the presence of HOBt (Figure 4.29).

In our proposed mechanism, the reaction starts with a transthioesterification step where the thiophenol substitutes the benzyl mercaptan at the C-terminus of the peptide thioester. The presence of the phenyl thioester intermediate in all our reactions confirms its existence (Figure 4.30). The lower pKa of thiophenol compared to benzyl mercaptan (6.6 versus 9.67, respectively) make this thioester a better leaving group,

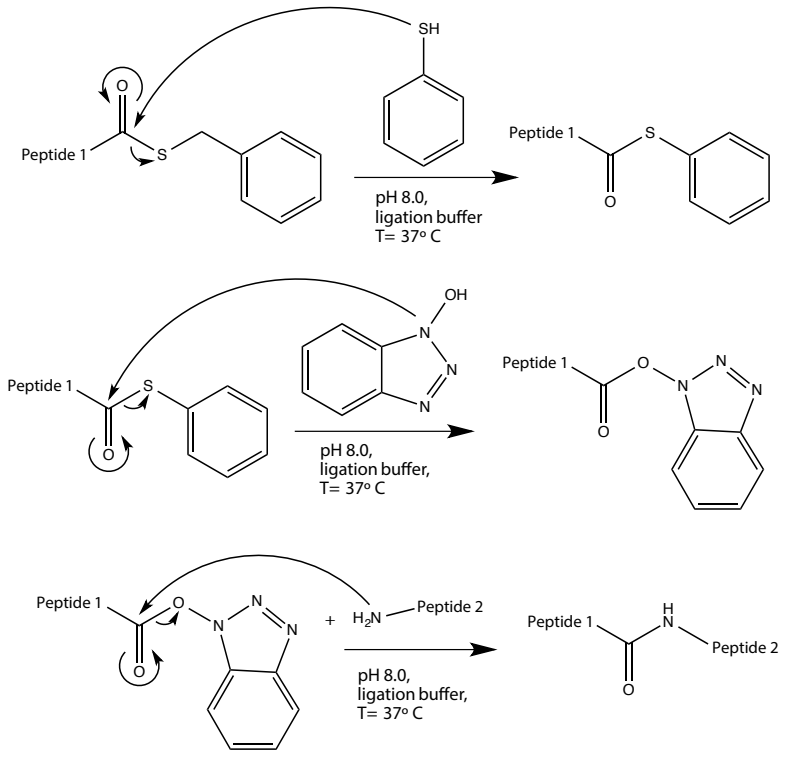


Figure 4.29: Scheme of the proposed mechanism for the ligation reaction.

conditioning the following steps. In fact, this step is crucial for the reaction, demonstrated by the absence of product when thiophenol is missing (data not shown) due to the poor leaving group that the alkanethiols generate. However, when thiophenol is substituted by MPAA (with a similar pKa of 6.6) (Figure 4.27 B, Appendix Figure 7.9), the product formed drops dramatically. The low formation of the intermediate thioester by MPAA reveals to be the key reason of the low conversion obtained. In this aspect, the ligation buffer used (more suitable to thiophenol exchange) could be the explanation for this fact. HOBt, however, seems that does not have a big influence on this step, as the phenyl thioester intermediate formation is similar regardless HOBt presence.

The fact that we could detect the phenyl thioester intermediate in the solution demonstrates the slower rate of the second part of the reac-

tion. This part is composed by two separated steps. Firstly, HOBt reacts with the phenyl thioester intermediate, generating an HOBt ester that is more prone to be reactive with the N-terminus of the second peptide, due to the better leaving group of HOBt ester over phenyl ester (lower pKa of the HOBt, 4.6 versus 6.6 of thiophenol). Finally, the nucleophilic attack performed by the amine at the N-terminus ends the reaction generating the native peptide bond between the two peptides.

Overall, we believe that although most of the reaction may follow this mechanism, we cannot discard another alternative mechanism that could be happening at the same time as, for instance, the reaction mechanism without HOBt.

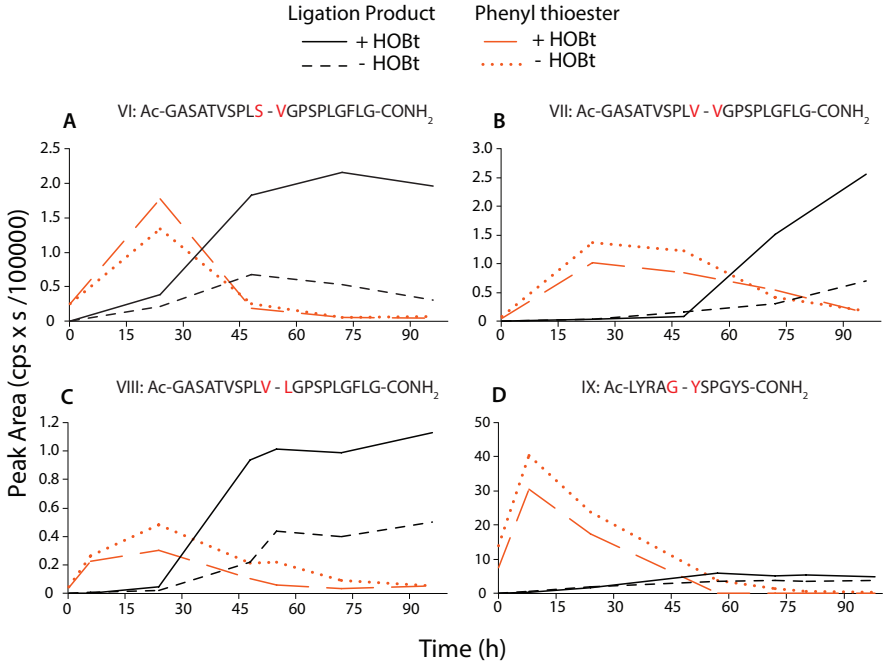


Figure 4.30: Same kinetic plot as Figure 4.25 but with the thiophenolic ester formation represented too in red, with HOBt (solid line) or without HOBt (dashed line). In the Appendix Table 7.1 are listed the product ligation cations of every reaction.

4.4.4 Ligation with TGIF1 phosphorylated peptides

One of the aims of the peptide ligation project was to find an easy method to synthesise long fragments of phosphorylated peptides selected from the TGIF1 protein sequence. To test the feasibility of this hypothesis, we synthesised two peptides: a serine phosphorylated thioester and an N-terminus free peptide (Table 4.6, 5 and 13). Once ligated, these peptide fragments would generate a sequence corespondent to 289-302 of TGIF1 that we plan to use in the analysis of the interactions with SMAD2.
In Figure 4.31 there is the kinetic plot for the reaction. Again, the presence of HOBt in the reaction increases the conversion of the reaction. However, the final product does not reach one third of the initial thioester quantity (Table 4.7, X).

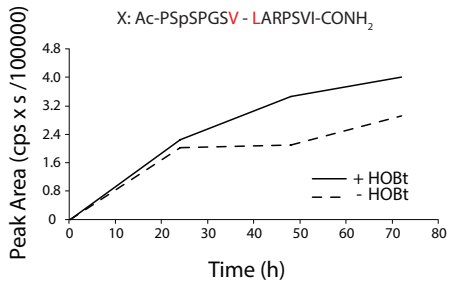


Figure 4.31: Kinetics of the ligated product formation of the X combination, with HOBt (solid line) or without (dashed line).

Lastly, the fact that we did not observe a clear interaction between TGIF1 and SMAD2 made the synthesis of the peptide less attractive as we have thought to use the peptides in order to investigate the importance of the various phosphorylation sites on the binding with SMAD2. However, we finally discarded the peptide synthesis approach because the yield obtained was too low, not only regarding the peptide ligation itself, but also with the peptide thioester synthesis. In conclusion, in our system the protein expression plus kinase phosphorylation strategy appears to be easier to generate and more powerful than the peptide synthesis approach.

4.5 Determination of six mutant structures of FBP28-WW2

4.5.1 Introduction

Computing simulations have been developed during the last 60 years as a useful tool to study molecular processes, otherwise impossible to analyse by experimental techniques. Specifically, all-atoms force fields have recently facilitated the study of the folding of small proteins [153]. However, the validation of these methods remains uncompleted because the different timescale between the computational (from hundreds of nanoseconds to microseconds) and the experimental studies (from several microseconds to seconds). Thanks to their accessible quantitative observables, the study of folding and unfolding kinetics have been used as a model to connect microscopic – theoretical – and macroscopic – experimental – world. In order to achieve the millisecond scale in the molecular dynamics simulations, it is required the use of a coarse grained force field. Thus, several coarse-grained methods have been developed. One of them, designed by the group of Dr. Harold Scheraga (Cornell University), makes use of a physics-based united-residue (UNRES) force field, being able to simulate the folding of small proteins in the millisecond timescale [154]. The objective of this project was to validate the use of the UNRES force field for the study of protein folding mechanism using the FBP28-WW2 domain as a model.

The WW domain of the Formin Binding Protein 28 (FBP28, also known as Transcription elongation regulator 1 (Tcerg1), Uniprot: O14776, Figure 4.32) has been used as a model for the validation of the molecular dynamics methods thanks to the availability of its structure, its fast-kinetics folding, biological relevance and small size (is one of the smallest monomeric triple-stranded antiparallel β-sheet protein known [155]). Thus, many efforts have been made in order to study its folding via experimental and theoretical methods. Still, there is some controversy about whether FBP28-WW2 follows a two or three-state mechanism during its folding [156].

In order to study the folding mechanism of the FBP28-WW2, UNRES

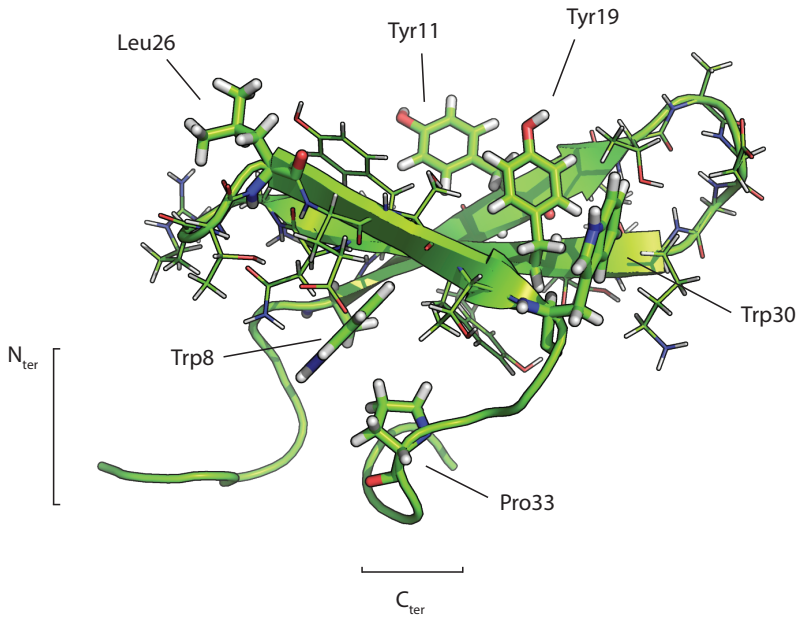


Figure 4.32: Wild-Type FBP28-WW2 domain structure. The structure was solved by Macias *et al.* in 2000 [155]. PDB id: 1E0L. The main amino acids are represented with their side-chain. Among them, the most relevant are labelled.

force field simulations introducing specific point mutations were performed to detect the presence or absence of some potential intermediates. To do so, canonical Langevin simulations with the UNRES force fields were performed for the WW2 wild-type domain and six mutants: Y11R, Y19L, W30F, ΔNY11R, ΔNΔCY11R, ΔNΔCY11R/L26A (Figure 4.33). For each protein, 512 independent folding trajectories were calculated, starting from an extended structure. Afterwards, each trajectory was classified as native (N), intermediate (I) or unfolded (U) based on the root-mean-square deviation (RMSD) from the respective reference structure. To obtain the rate constant, the fractions of the N, I and U states were averaged over the 512 trajectories at each time interval and were fitted to exponential kinetics equations. Lastly, the results were comparable with the experimental data previously determined [156].

Figure 4.33: Sequence of all six mutant compared with the wild-type (WT) sequence. In red are distinguished the mutated amino acids. For simplicity, all the amino acid number are referenced to the WW2 domain of 37 amino acids total, where the first glycine (1) corresponds to the amino acid number 428 of the full length protein and the last lysine (37) to the number 464, respectively.

4.5.2 Results

My participation in this project was (together with Dr. Todorovski and Dr. Macias) to express, purify and determine the structure of the six mutants of the FBP28 WW2 protein: Y11R, Y19L, W30F, ΔNY11R, ΔNΔCY11R and ΔNΔCY11R/L26A (Figure 4.33). The election of the six mutants was based on the previous experimental results done by Nguyen *et al.* [156]. As described in the materials and methods section, after the purification of each mutant we carried out 2D-NOESY and TOCSY NMR experiments in order to determine the structure of every mutant. The structures of each mutant are represented in Figures 4.34, 4.35, 4.36, 4.37, 4.38 and 4.39. In the table 4.9 are summarised the statistics of every structure we determined.

Overall, in all mutants the main three-β-strand structure is maintained thanks to the interaction between the proline 33 and the tryptophan 8, which act as a "staple" keeping the folding of the domain. Nevertheless, near the mutations there are minor changes affecting the position of some side-chains.

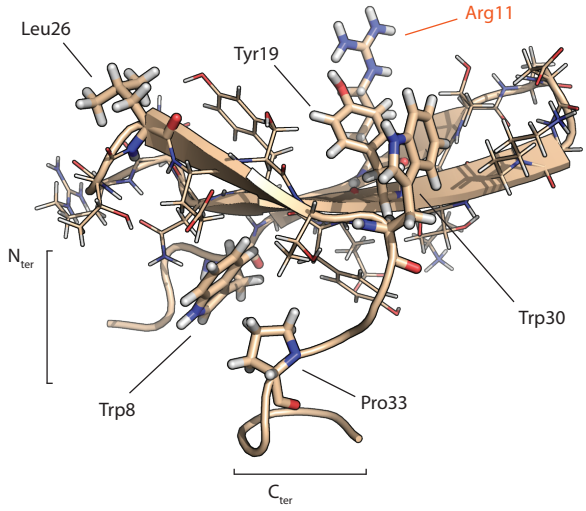


Figure 4.34: Experimental NMR structure of the mutant Y11R (Y438R). The main amino acids are represented with their side-chain. Among them, the most relevant are labelled. In red is distinguished the mutated residue. PDB id: 2MW9.

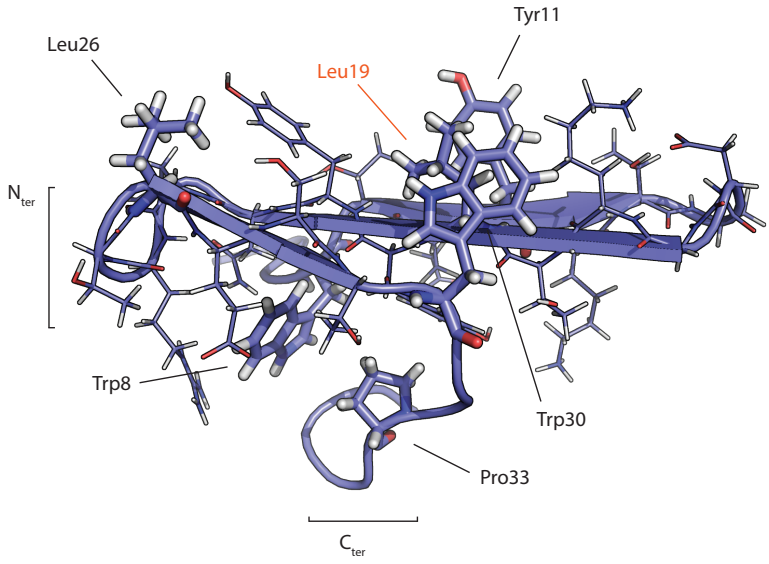


Figure 4.35: Same as Figure 4.34 but with the mutant Y19L (Y446L). PDB id: 2MWA.

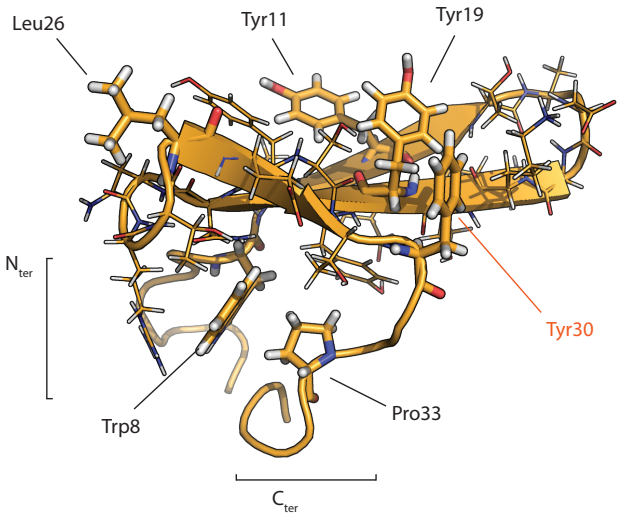


Figure 4.36: Same as Figure 4.34 but with the mutant W30F (W457F). PDB id: 2MWB.

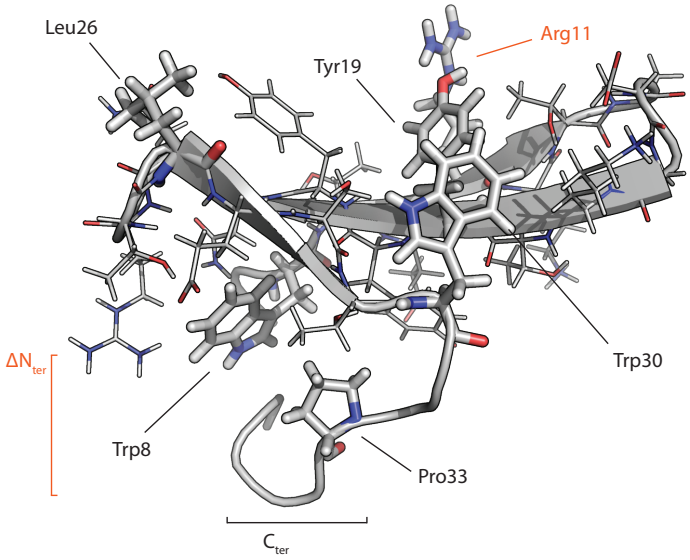


Figure 4.37: Same as Figure 4.34 but with the mutant ΔNY11R (Y438R). PDB id: 2MWF.

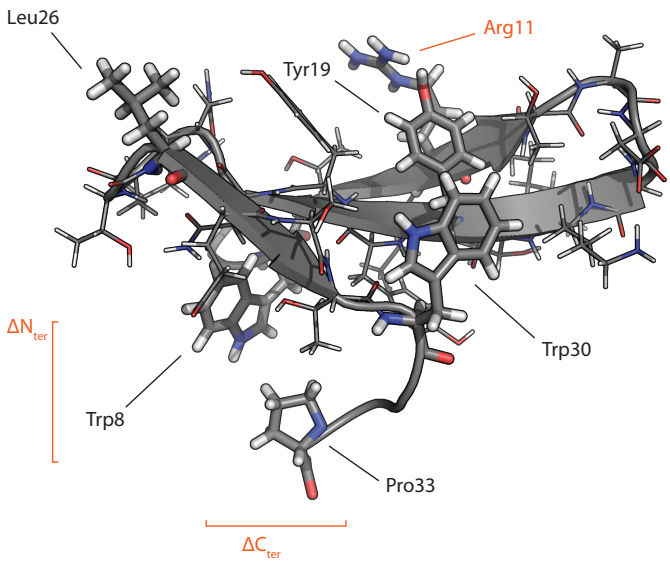


Figure 4.38: Same as Figure 4.34 but with the mutant ΔNΔCY11R (Y438R). PDB id: 2MWD.

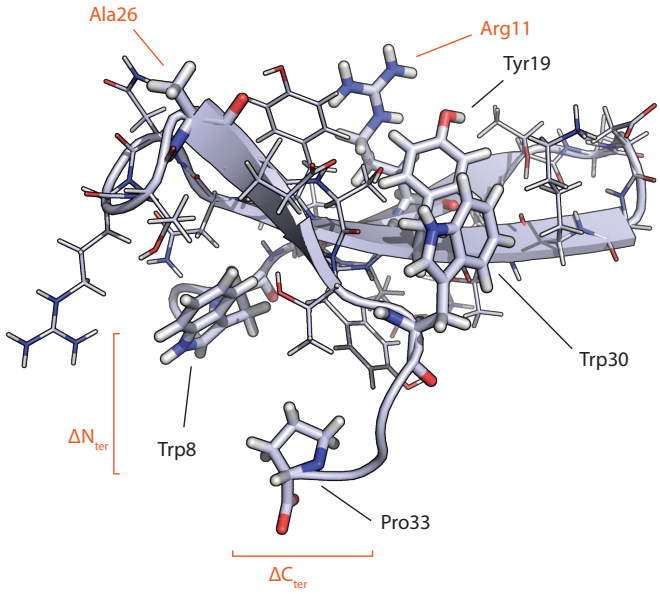


Figure 4.39: Same as Figure 4.34 but with the mutant ΔNΔCY11RL26A (Y438R/L453A). PDB id: 2MWE.

Table 4.9: Structural statistics of the six FBP28-WW2 mutants.

	Y11R (Y438R) PDB: 2MW9	Y19L (Y446L) PDB: 2MWA	W30F (W457F) PDB: 2MWB
Restraints used for the calculation (120 structures ensemble)			
Sequential ($\|i-j\|=1$)	124	133	115
Medium range ($1<\|i-j\|\leq 4$)	48	41	53
Long range ($\|i-j\|>4$)	165	193	147
Dihedrals	45	57	71
Hydrogen Bonds	10	10	10
RMSD (Å) from experimental			
NOE (x10^{-3})	3.2 ± 0.8	3.5 ± 0.8	2.2 ± 0.4
Bonds (x10^{-3}) (Å)	2.2 ± 0.5	4.2 ± 0.5	3.3 ± 0.5
Angles (°)	1.6 ± 0.02	0.6 ± 0.02	3.6 ± 0.05
Coordinate Precision (Å)			
Backbone	0.50	0.40	0.35
CNS potential energy (kcal mol^{-1})			
Total energy	-1321 ± 54	-1219 ± 27	-1409 ± 39
Electrostatic	-1428 ± 56	-1434 ± 40	-1516 ± 43
Van der Waals	-138.2 ± 8	-86.73 ± 13	-139.6 ± 10
Bonds	9.63 ± 0.6	10.4 ± 1	7.8 ± 0.6
Angles	38.5 ± 3	57.8 ± 5.5	33.8 ± 4
Structural quality (% residues) 20 best structures			
In most favoured region of Ramachandran plot	89.1	88	93.6
In additionally allowed region	10.2	12	6.4

	ΔNY11R (Y438R) PDB: 2MWF	ΔNΔCY11R (Y438R) PDB: 2MWD	ΔNΔCY11RL26A (Y438R/L453A) PDB: 2MWE
Restraints used for the calculation (120 structures ensemble)			
Sequential ($\|i-j\|=1$)	102	90	93
Medium range ($1<\|i-j\|\leq 4$)	38	43	54
Long range ($\|i-j\|>4$)	170	178	146
Dihedrals	70	54	48
Hydrogen Bonds	10	10	10
RMSD (Å) from experimental			
NOE (x10^{-3})	4.3 ± 0.3	2.8 ± 0.3	3.1 ± 0.1
Bonds (x10^{-3}) (Å)	4.6 ± 0.1	3.2 ± 0.1	4.1 ± 0.1
Angles (°)	0.6 ± 0.02	0.6 ± 0.02	0.5 ± 0.02
Coordinate Precision (Å)			
Backbone	0.36	0.45	0.40
CNS potential energy (kcal mol^{-1})			
Total energy	-1168 ± 22	-1038 ± 22	-1047 ± 22
Electrostatic	-1365 ± 22	-1075 ± 40	-1157 ± 25
Van der Waals	-89.6 ± 6	-86.73 ± 13	-96.8 ± 8
Bonds	11.7 ± 0.8	10.6 ± 1	8.1 ± 0.6
Angles	55.5 ± 4	38.5 ± 6	32.3 ± 3
Structural quality (% residues) 20 best structures			
In most favoured region of Ramachandran plot	90	89	86
In additionally allowed region	10	11	14

5 Discussion

5.1 Characterisation of the interaction between TGIF1 and SMAD proteins

Description of the interaction between TGIF1 (256-347) and SMAD2

The first objective of the present work was the characterisation at the structural level of the interaction between TGIF1 and SMAD2, previously described biochemically [28]. For each protein, the region that was described to interact was expressed and purified. The ^{1}H-^{15}N HSQC of TGIF1 (256-347) fragment revealed the absence of tertiary structure, in agreement with the predictions generated by MetaDisorderMD2 disorder predictor (Appendix Figure 7.1). Nevertheless, we could assign up to 94% of the backbone peaks observed in the HSQC.

In order to observe the interaction between both fragments, HSQC titration experiments were performed. First, SMAD2-EEE (186-467) was titrated over ^{15}N labelled TGIF1 (256-347), resulting in no chemical shift differences. We also titrate the TGIF1 (256-347) previously doubly phosphorylated by p38α, as we wanted to investigate if the phosphorylations of Ser286 and Ser291 have influence on the binding to SMAD2. The addition of five equivalents of SMAD2-EEE (186-467) only causes small changes in chemical shifts and intensities. However, the fact that three of those modified peaks were related to the phosphorylation event (both phosphorylated serines and Arg283) support our hypothesis that the phosphorylation could modulate the interaction between the pair of proteins in a more biological context.

In order to validate the interaction a MST experiment was carried out. The results confirmed the absence of a strong interaction between both

proteins in the experimental conditions used, suggesting that *in vitro* the interaction of both protein fragments is weak.

A final titration was performed between TGIF1 (256-347) and SMAD2. In this experiment we used a fragment of SMAD2 corresponding to the MH1 (SMAD2-MH1 (10-174)) in order to discard an interaction through the MH1 domain. With only three amino acids that were modified upon the titration (Cys273, Cys303, Leu313), the result suggested that the interaction, if happens, is also weak.

Several explanations might account for these observations. (1) The interaction perhaps requires a third (or more) partner/s to facilitate the TGIF1 (256-347) and SMAD2-EEE (186-467) interaction. This hypothesis is supported by the western-blot experiments performed by Wotton *et al.* where they demonstrated that the addition of SMAD4 or FOXH1 enhances the binding between SMAD2 and TGIF1 [28]. (2) Another explanation lies on the results – somehow antagonistic with the former – where TGIF1 does not bind to phosphorylated SMAD2 [48]. That is, TGIF1 and SMAD2 interaction is enhanced by the activation of the TGFβ, but, by some unknown mechanism, it is the not activated SMAD2 which binds to TGIF1 and not the phosphorylated one. In our experiments we used a SMAD2 with three glutamic acid residues at the C-terminus, which mimics the phosphorylated serines. If TGIF1 only binds to unphosphorylated SMAD2, that could be a reasonable explanation why we do not observe a thigh interaction between both fragments. (3) It could be possible that TGIF1 only binds to the full-length SMAD2. As we did not test this interaction, we cannot discard this possibility. (4) Finally, it may be possible that the TGIF1 fragment requires another post-translational modification (PTM), such as phosphorylation on the serine 294 (also located in the SID domain), to interact with SMAD2.

TGIF1(256-347) is phosphorylated *in vitro* by p38α and CK1

Post-translational modifications (PTM) are one of the main ways to regulate protein function and binding to other molecules [59]. Phosphorylation on serines, threonines and tyrosines are one of the most studied

PTM. In this work, we report how three serines of TGIF1(256-347) are phosphorylated *in vitro* by p38α and CK1 kinases. These results mimic the phosphorylations detected by mass spectrometry high-throughput analysis in human embryonic stem cells [57].

We described that p38α phosphorylates *in vitro* Ser286 and Ser291 of TGIF1 (256-347) fragment in a stepwise manner using Real-Time NMR experiments. The subsequent addition of CK1 phosphorylates Ser294. This is the first time that TGIF1 is suggested to be a substrate for both kinases. The phosphorylations do not modify greatly the structure of the TGIF1 (256-347) fragment. However, we have reported that the chemical shifts attributed to the side-chains of the arginines shift upon phosphorylation. The localisation of these arginines – surrounding the phosphorylated serines – make possible the assumption that these shifts were due to the salt bridges formed between the guanidinium group of the arginines and the phosphate group of the phosphorylated serines [148].
We are aware that we only investigated the phosphorylation in this specific fragment of TGIF1 *in vitro*. This evidence does not demonstrate that these serines would be also phosphorylated by p38α and CK1 when the full-length TGIF1 is the substrate or that others kinases might participate in the process *in vivo*.

Description of the interaction between TGIF1 (150-248) and SMAD2/4-MH1 proteins

We were also interested in the role of the homeodomain region in DNA and SMADs interactions. In previous studies, it was found that the preserved region rich in arginines localised at N-terminal of the HOXC9 HD interacts with SMAD4-MH1 [150]. In contrast, this region has been found to bind DNA in another homeodomain-TALE protein, PBX1 [41]. Based on this information, we generated a new TGIF1 homeodomain construct (150-248) that contains not only the HD but also the arginine-rich N-terminal region.

The TGIF1 (150-248) HSQC spectra evidences the presence of tertiary structure, in agreement with the already determined TGIF1 HD struc-

ture (PDB id: 2LK2). Unfortunately, we were incapable to assign the TGIF1 (150-248) protein. Instead, we have used the information available in the BMRB to identify the ligand binding. Nevertheless, peaks that were not isolated and the peaks corresponding to the N-terminal region (absent in the 2LK2 construct) we could not extract any information, even for those modified because of the titrations.

The first titration with the canonical DNA confirm the tight binding between them, as most of the peaks were greatly modified upon the binding.

In the next experiment, we wanted to confirm whether SMAD4-MH1 could bind to the TGIF1 (150-248) fragment. After the addition of 1.6 equivalents of SMAD4-MH1 over the ^{15}N labelled TGIF1 fragment, several peaks changed, specially regarding their intensity. The localisation of the modified peaks in the TGIF1 HD 2LK2 structure shows that all of them are situated in the same region, towards the N-terminal region of the construct, where the arginine-rich is located. However, these peaks are different from the ones that are shifted in the HOXC9 interaction with SMAD4-MH1 [150], where only the peaks at the N-terminal of the Arg-rich region are moved upon SMAD4-MH1 titration. Because we have not assigned our TGIF1 (150-248) construct, we cannot assure that the peaks related to the N-terminal Arg-rich region are also modified by SMAD4-MH1 addition.

Moreover, the change in the intensity was positive, not negative. Usually, when some specific residues bind to another molecule, their tumbling decreases. Therefore, their related peak gets broader, decreasing its intensity [157]. In contrast, when there is no contact, the peak keeps the same intensity. However, if the peak increases its intensity, as in our case, that means that before the addition of SMAD4-MH1 those residues were bound. And the presence of SMAD4 enhances the freedom of these residues, with more tumbling, narrower peak, and thus higher intensity [157].

We hypothesise two possible reasons for those results. (1) TGIF1-HD might be involved in contacts with other TGIF1-HD molecules in solu-

tion, characteristic of a monomer-dimer equilibrium. Hence, the presence of SMAD4-MH1 (1-140) disturbs or impedes the dimer formation. (2) Before the addition of SMAD4-MH1 domain, the TGIF1-HD might be in a closed conformation, with the N-terminal tail firmly bound to the domain. However, in the presence of the SMAD4-MH1, the open conformation is favoured, allowing a free tumbling of those residues. Furthermore, some of these residues are located in the third α-helix and may be important for the interaction with DNA. Unfortunately, we cannot confirm the interaction of SMAD4 with the N-terminal tail because the lack of the full assingment. We also tried some docking simulations between TGIF1 and SMAD4-MH1 without any result.

The titration of the SMAD2-MH1 over the same labelled protein revealed analogous behaviour as with SMAD4-MH1, with almost the same modified peaks. However, less peaks were affected and the intensity variation was not so high. We conclude that we were observing the same phenomena but in a lesser intensity.

Further EMSA experiments provide more evidences of the binding. In particular, the addition of SMAD4-MH1 or SMAD2-MH1 interferes with the binding between TGIF1 (150-248) and its canonical DNA. This result highlights the importance of the binding as it interferes with one of the main TGIF1 functions. Curiously, in the EMSA experiments, the SMAD2-MH1 domain seems to interfere more than SMAD4-MH1, in contrast with the NMR results.

Globally, these results complement the previous and controversial findings [28, 48]. On one hand, Wotton *et al.* demonstrated that the presence of SMAD4 enhances the binding between SMAD2 and TGIF1, discarding the interaction between SMAD4 and TGIF1 [28]. On the other hand, Seo *et al.* proved that the addition of TGIF1 prevents the formation of the SMAD2/SMAD4 complex [48]. To that previous knowledge, we confirmed that SMAD4-MH1 and SMAD2-MH1 interact with the TGIF1 HD region interfering with its DNA binding. In addition, we cannot discard an interaction between the SID region of TGIF1 and SMAD2, maybe effective only when TGIF1 is triple phosphorylated at serine 286, 291 and 294.

TGIF1 (256-347) and TGIF1 (150-248) interact with each other.

The NMR titration experiment confirmed the interaction between the two TGIF1 fragments (150-248) and (256-347). This binding event argues for the closed conformation of the full length protein. It could be that, in physiological conditions, the two regions interact with each other and the phosphorylation upon TGIF1 (256-347) would disrupt the interaction. Moreover, the EMSA results indicate that the addition of TGIF1 (256-347) interferes with the normal binding between TGIF1 HD and its canonical DNA, in a similar way as SMAD4-MH1 and SMAD2-MH1.

General perspectives

This PhD project was aimed to structurally characterise the binding of TGIF1 protein with its partners, SMAD2, SMAD4 and also DNA. The results obtained within this thesis might be used as preliminary information for further structure-based rational drug design against several diseases such as cancer or HPE. By knowing the exact interaction surface of TGIF1 with the binding partners we could find interesting and promising hot spots that can serve as targets of chemical compounds that would specifically inhibit these interactions. The first step was then to describe those interactions biophysically and structurally.
To conclude, this project is an ongoing project in our lab. We have planned more experiments in order to confirm our results. Such experiments include TGIF1-HD bound to DNA X-Ray structure, K_D determination by isothermal titration calorimetry (ITC) and NMR titrations with the protein combinations that we could not perform.

5.2 Study about the cysteine-free direct aminolysis ligation reaction

Nowadays there are many published strategies to ligate peptides in order to synthesise longer polypeptides, even proteins in some cases. However, there is no yet standard protocol as each strategy has its own advantages and limitations, depending of the protein of interest. Native

chemical ligation (NCL) is currently the most popular and established technique. Moreover, recent advances in desulfuration techniques have opened the door to replace the required cysteine in the ligation junction for up to 12 different amino acids [158]. These discoveries solved, at least partially, the main limitation of the NCL. Therefore, the NCL is currently feasible, at least theoretically, with almost any peptide sequence.

However, other methods are under development with the goal to achieve a peptide ligation that would be chemoselective, free of epimerisation, independent of the amino acid at the ligation junction and with high reaction rate. One of these methods is focus on the direct aminolysis between the two peptides, without the requirement of any external thiol at the ligation junction. Based on the ligation method developed by Wong and coworkers [83], in our work we have demonstrated that the addition of HOBt to the ligation buffer increases the conversion, but not the rate, of every peptide ligation tried, regardless which amino acids are at the ligation junction (with the exception of cysteine). Especially interesting was the increasing of the conversion when valine, leucine or tyrosine where at the ligation junction. In those reactions, the addition of HOBt was crucial to obtain a satisfactory yield (between 34 and 57%).
The use of HOBt has also been reported in others peptide ligation studies. Thus, Danishefsky and coworkers used HOBt as a catalyst for a direct aminolysis peptide ligation between a thioacid and a free N-terminus peptide [159]. Based on their experiments, the authors proposed an oxidation mechanism, where the presence of an oxidant in the reaction (such air versus argon) makes a difference in the yields reported. In our work, we hypothesise that HOBt catalyses the reaction between the thiophenolic intermediate and the final product. The better HOBt leaving group versus the thiophenolic leaving group (based on their different pKa) would explain this mechanism of reaction.

Regarding the special case of cysteine, we hypothesise that another reaction pathway, presumably NCL, was also taken place. As the NCL is faster than the direct aminolysis reaction, the presence of HOBt induce the direct aminolysis pathway, sequestering reagents for the NCL and

thus, slowing the reaction.

In addition of HOBt, we also tried the effect of the presence of HOAt (an HOBt analogue) and DIPEA or DBU (two bases) in the ligation reaction. Only the addition of 4 equivalents of DIPEA demonstrated better results in comparison with HOBt alone, while the other reagents were either worse or equal to the HOBt option. In another set of experiments, we investigated the same reaction but using free N-terminus thioesters instead of protected with acetyl group. However, in all these trials the competitive intramolecular cyclisation reaction was dominating over the aminolysis and thus we discarded this option.

Moreover, we demonstrated that the method is also applicable to phosphorylated peptides. However, for the phosphorylated TGIF1 peptide, the use of the peptide synthesis followed by peptide ligation method ends in a final low yield. Therefore, the heterologous expression with *in vitro* phosphorylation was the preferred option in our case.

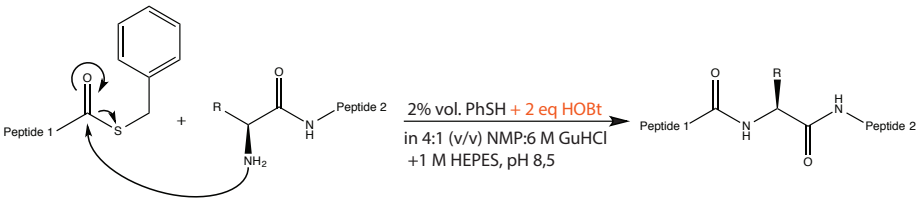


Figure 5.1: Scheme of the Wong direct aminolysis peptide ligation with the improvement we presented in this thesis (in red).

To sum up, the improvement presented in this thesis is a small step towards having a perfectly valid direct aminolysis peptide ligation reaction (Figure 5.1). Moreover, the easy separation of the ligation product by HPLC is another good argument for the applicability of the technique. However, much more effort has to be done in order to establish this method as an appropriate alternative to the NCL. Some of the issues that need to be solved in the future include the improvement of the ligation rate and conversion and the expansion of the chemoselectivity also for non protected lysines. Here, we want to point out that

the addition of DIPEA to the ligation buffer is a promising start point for further developments.

Finally, all the results of this section are published in Biopolymers (Peptide Science) under the title "Addition of HOBt improves the conversion of thioester-amine chemical ligation" [119] (Appendix).

5.3 Determination of six mutant structures of FBP28-WW2

The determination of the six mutants structures of the FBP28-WW2 domain was key to validate the simulations performed by the UNRES force field. In particular, the solved structures indicate that the point mutations on the WW2 domain, including deletions at the N and C-terminus, do not alter the overall structure. Interestingly, the shortest structures, ΔNΔCY11R and ΔNΔCY11RL26A, with only 28 residues, are the smallest WW domain structures determined so far.

The UNRES simulations analysis revealed that the WT and four mutants exhibit double-exponential kinetics (three-state folding) while the Y19L and W30F mutants show single-exponential kinetics (two-state folding). Most of the results are in agreement with the experimental kinetics studies done by Nguyen *et al.* [156] while the discrepancies observed can be explained in terms of the free-energy barrier heights.

Overall, this study concluded that UNRES force field is a valid computational tool as it can simulate accurately the kinetic constants obtained experimentally.
This work was published in PNAS under the title "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements" [122] (Appendix).

6 Conclusions

1.- Interaction between SMAD2 and TGIF1 by NMR.

We have prepared several fragments of TGIF1 and analysed their interactions with SMAD2 *in vitro*. No direct interaction between TGIF1 (256-347) and with SMAD2-EEE (186-467) was detected in our experimental conditions. The phosphorylation on Ser286 and Ser291 favours a weak interaction between both proteins. On the other hand, we also observed weak interaction between TGIF1 (256-347) and the MH1 of SMAD2.

2.- p38α and CK1 phosphorylate TGIF1 (256-347) *in vitro*.

For the first time it is reported a stepwise phosphorylation of Ser286 and Ser291 by p38α and the subsequent phosphorylation of Ser291 by CK1. The phosphorylations do not disturb the previous TGIF1(256-347) global structure but modify the side-chains of nearby arginines.

3.- SMAD2/4-MH1 interactions with TGIF1 (150-248).

The addition of SMAD4-MH1 or SMAD2-MH1 domains disrupts the binding of TGIF1 homeodomain with the DNA. Moreover, the NMR titrations suggest that the presence of either SMAD4-MH1 or SMAD2-MH1 domains affects the conformational freedom of certain residues located in the N-terminal region of the homeodomain.

4.- TGIF1 (150-248)-TGIF1 (256-347) interaction might suggest the presence of open and closed conformations of full-length TGIF1.

NMR titrations confirmed the interaction and EMSA experiments showed

that TGIF1 (256-347) disrupts the binding of the homeodomain with DNA.

5.- The addition of HOBt improves the cysteine-free direct aminolysis ligation strategy.

The addition of HOBt improves the reaction conversion, but not the rate, of the ligation reaction, especially when sterically hindered amino acid (such as valine or leucine) are present at the ligation junction. The reaction is also compatible with phosphorylated peptides but intramolecular cyclisation side-reactions appear when the peptide thioester is not protected at the N-terminus.

6.- Six selected mutants structures of FBP28-WW2 were determined *de novo* by NMR.

The six mutants maintain the main characteristics of WW domain structure, even for the mutations introducing deletions at the N and C terminus. The structures were the experimental confirmation of the effect of this set of mutations in the structure. The existence of the WW fold in all these mutants was key to provide the grounds for the simulated folding curves generated using the UNRES force field.

7 Appendix

Characterisation of the interaction between TGIF1 and SMAD proteins

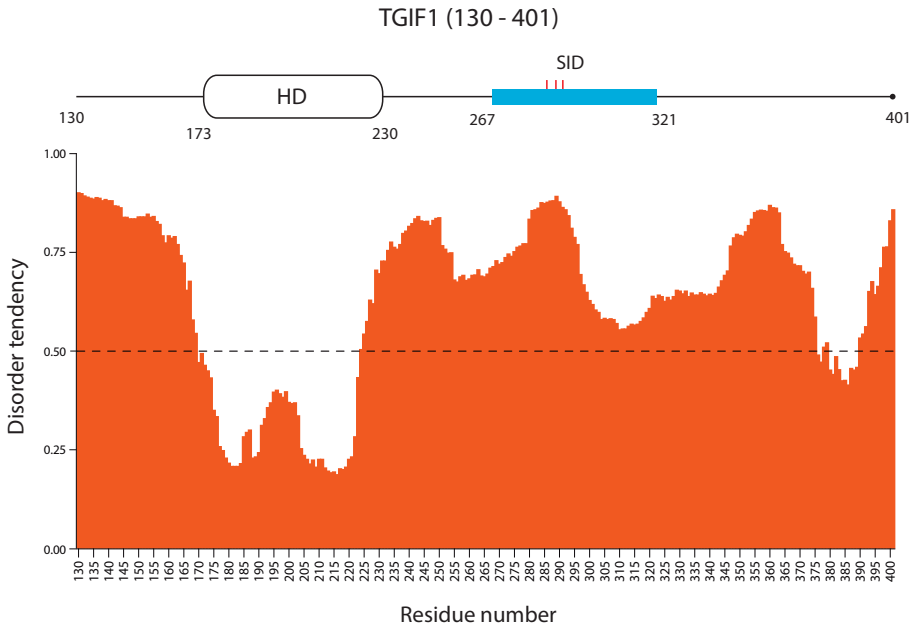


Figure 7.1: MetaDisorderMD2 prediction plot [43]. All residues whose disorder probability is over 0.5 are considered as disordered. The sequence starts at the TGIF1 amino acid number 130.

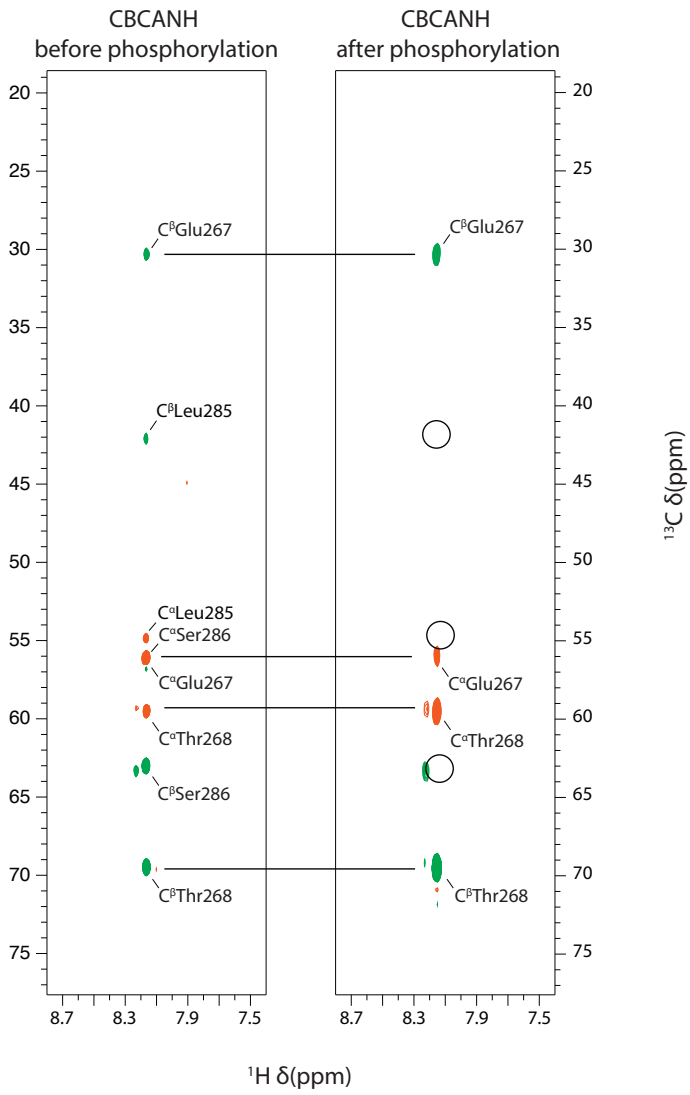


Figure 7.2: δ^{15}N 117.96 ppm strips from the CBCANH experiments on TGIF1 (256-347) before and after phosphorylation by p38α. The strips correspond to the superimposed Thr268 and Ser286 peaks. The horizontal lines connect equal peaks while the circle indicates the place where the peaks related to Ser286 should have been. Cα peaks are red; Cβ are green.

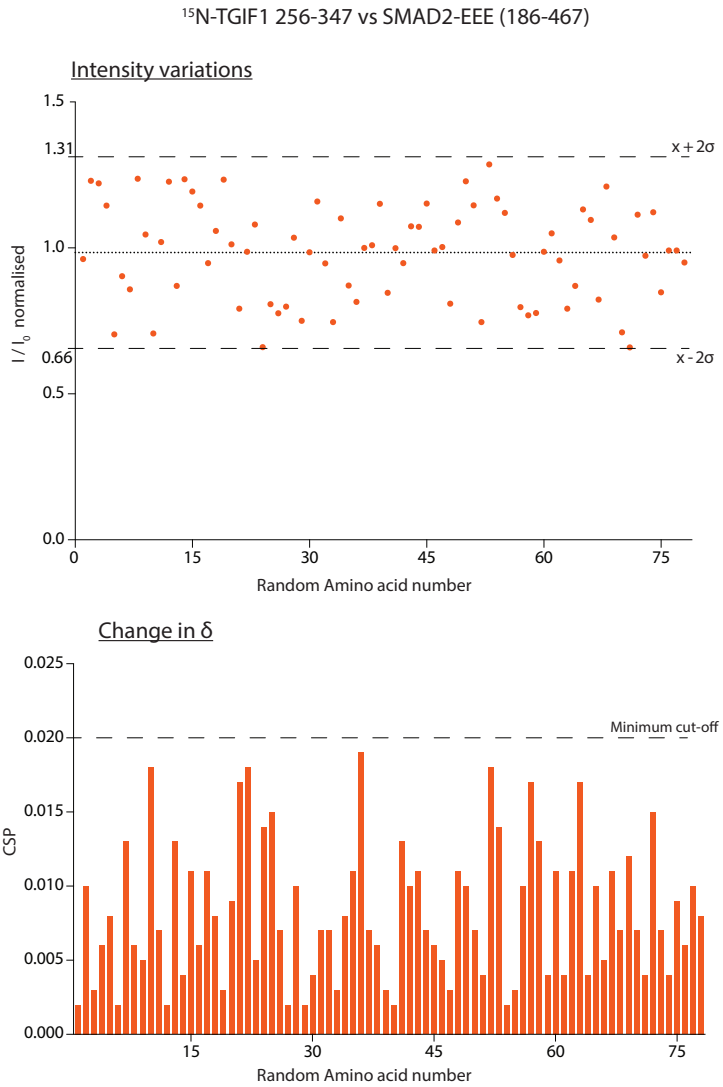


Figure 7.3: Intensity and CSP graph for the titration between ^{15}N-TGIF1 (256-347) and SMAD2-EEE (186-467). The amino acids are displaced randomly, but in the same order between both graphs. In green are labelled those peaks above the cut-off.

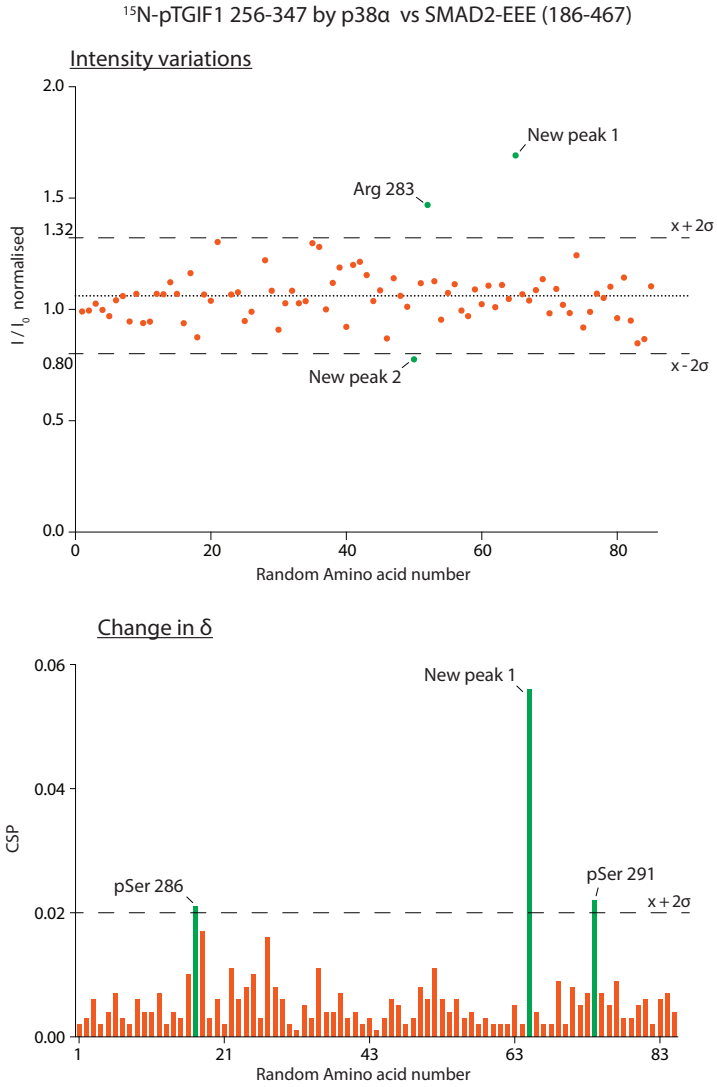


Figure 7.4: Intensity and CSP graph for the titration between ^{15}N-pTGIF1 (256-347) by p38α and SMAD2-EEE (186-467). The amino acids are displaced randomly, but in the same order between both graphs. In green are labelled those peaks above the cut-off.

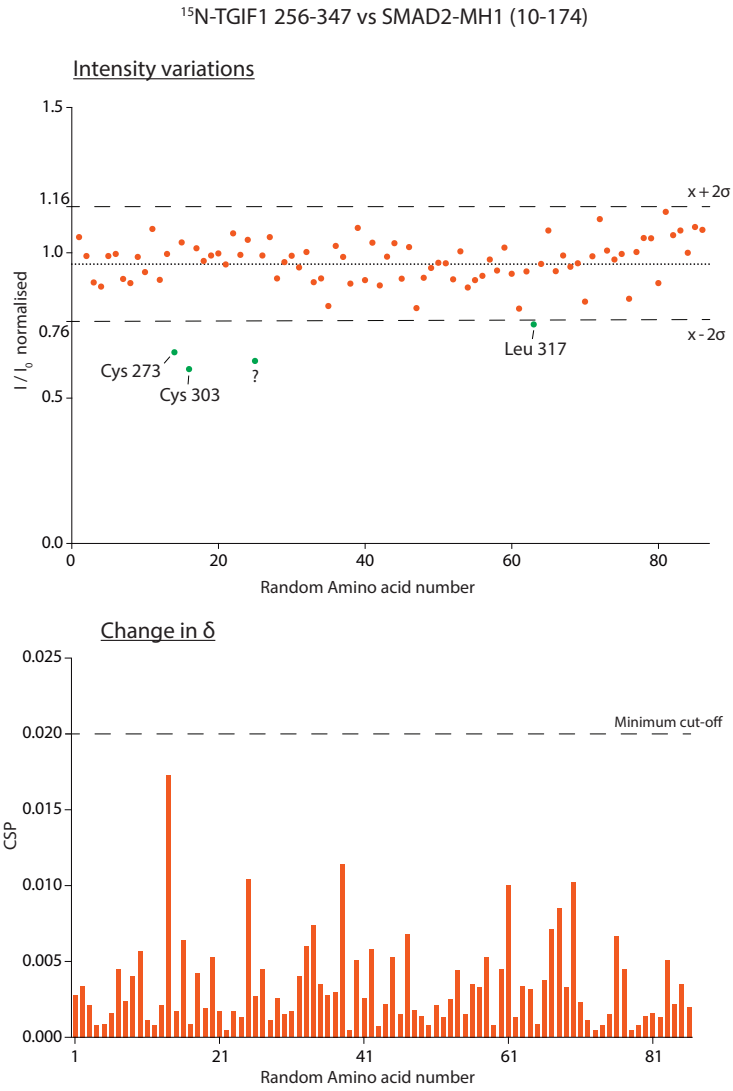


Figure 7.5: Intensity and CSP graph for the titration between ^{15}N-TGIF1 (256-347) and SMAD2-MH1 (10-174). The amino acids are displaced randomly, but in the same order between both graphs. In green are labelled those peaks above the cut-off.

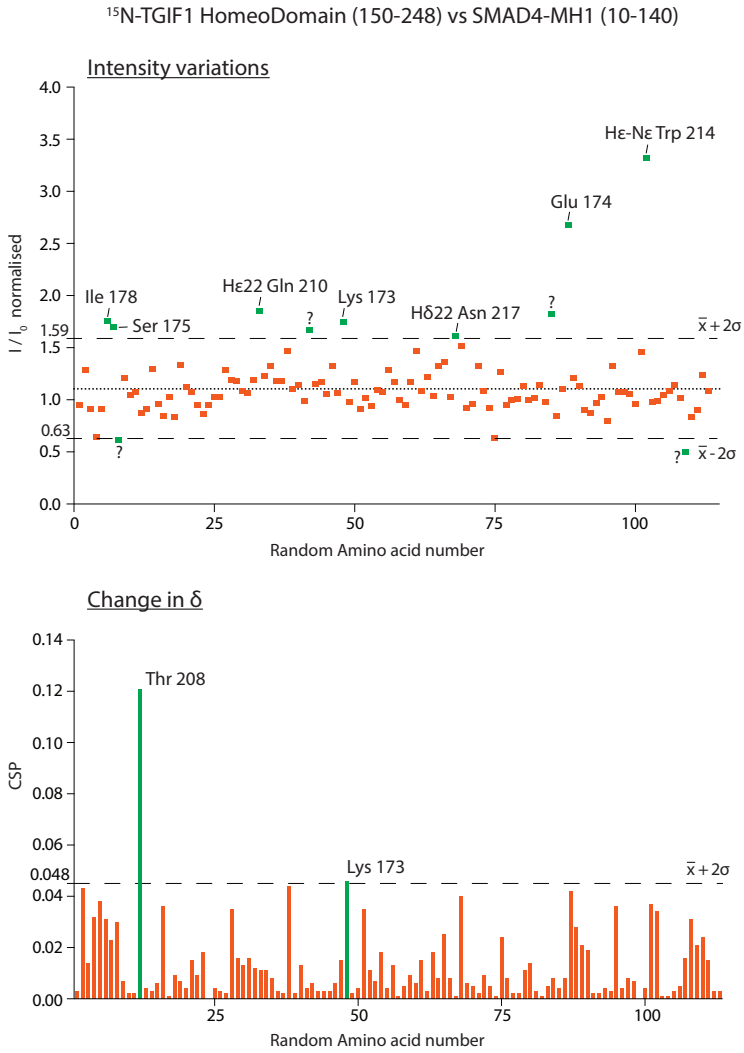


Figure 7.6: Intensity and CSP graph for the titration between ^{15}N-TGIF1 homeodomain (150-248) and SMAD4-MH1 (10-140). The amino acids are displaced randomly, but in the same order between both graphs. The interrogant mark indicates that those peaks could not be assigned. In green are labelled those peaks above the cut-off.

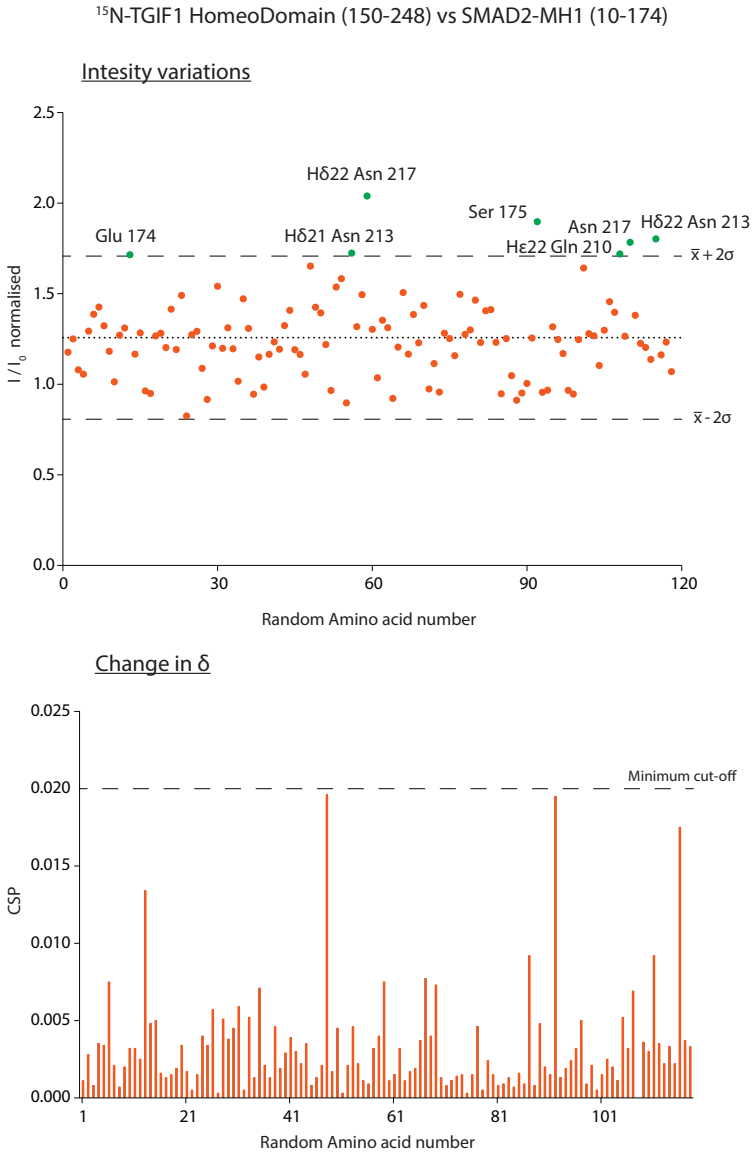


Figure 7.7: Intensity and CSP graph for the titration between ^{15}N-TGIF1 homeodomain (150-248) and SMAD2-MH1 (10-174). The amino acids are displaced randomly, but in the same order between both graphs. In green are labelled those peaks above the cut-off.

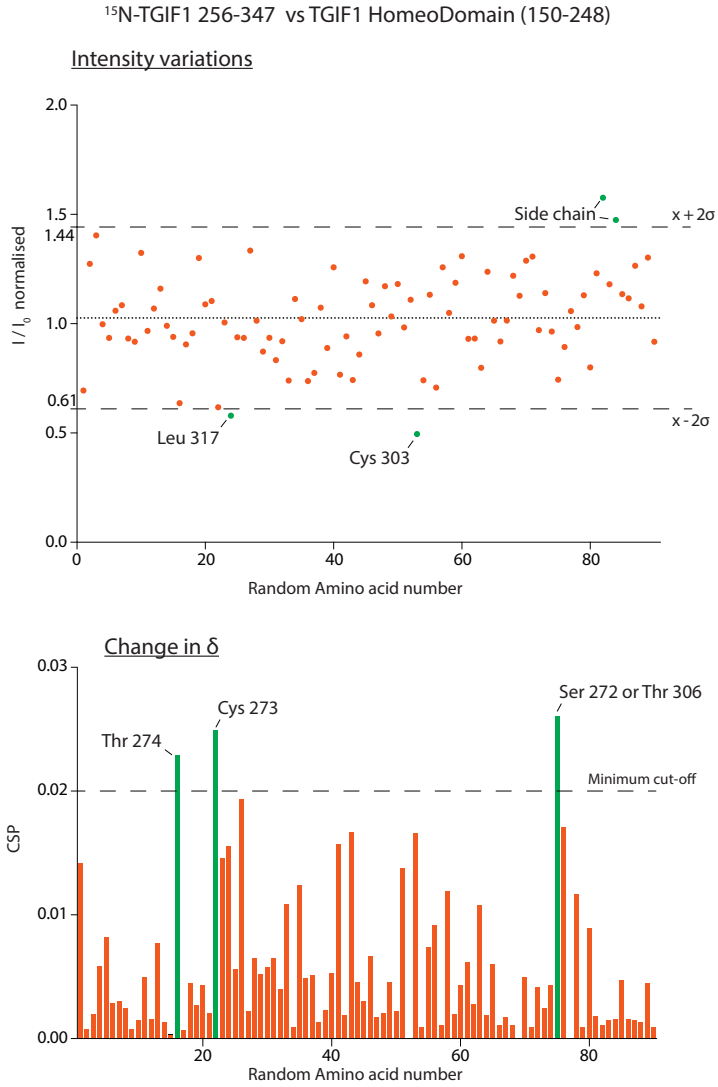


Figure 7.8: Intensity and CSP graph for the titration between ^{15}N-TGIF1 (256-347) and TGIF1 homeodomain (150-248). The amino acids are displaced randomly, but in the same order between both graphs. In green are labelled those peaks above the cut-off.

Study about the cysteine-free direct aminolysis ligation reaction

Table 7.1: List of the cations followed for each amino acid combination.

	Ligation Combinations	Experimental m/z of the product	Experimental m/z of exchanged thiophenol thioster
I	Ac-GASATVSPL**G**-SC_7H_7 + NH_2-**G**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1783.4 $[M + 2H]^{2+}$ = 892.2	$[M+H]^+$= 993.5
II	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**G**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1813.2 $[M + 2H]^{2+}$ = 907.1	$[M+H]^+$= 1023.5
III	Ac-LYRA**G**-SC_7H_7 + NH_2-**G**SPGYS-$CONH_2$	$[M + H]^+$ = 1168.5	$[M + H]^+$ = 713.5
IV	Ac-LYRA**G**-SC_7H_7 + NH_2-**A**SPGYS-$CONH_2$	$[M + H]^+$ = 1182.5	$[M+H]^+$= 713.5
V	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**C**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1859.0 $[M + 2H]^{2+}$ = 930.0	*
VI	Ac-GASATVSPL**S**-SC_7H_7 + NH_2-**V**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1855.4 $[M + 2H]^{2+}$ = 928.2	$[M + H]^+$= 1023.5
VII	Ac-GASATVSPL**V**-SC_7H_7 + NH_2-**V**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1867.2 $[M + 2H]^{2+}$ = 934.1	$[M + H]^+$= 1035.5
VIII	Ac-GASATVSPL**V**-SC_7H_7 + NH_2-**L**GPSPLGFLG-$CONH_2$	$[M + H]^+$ = 1881.4 $[M + 2H]^{2+}$ = 941.2	$[M + H]^+$= 1035.5
IX	Ac-LYRA**G**-SC_7H_7 + NH_2-**Y**SPGYS-$CONH_2$	$[M + H]^+$ = 1274.6	$[M + H]^+$= 713.5
X	Ac-PSpSPGS**V**-SC_7H_7 + NH_2-**L**ARPSVI-$CONH_2$	$[M + H]^+$ = 1488.6 $[M + 2H]^{2+}$ = 744.8	$[M + H]^+$= 844.8

*As the N-terminal peptide starts with cysteine, the intermediate through thiophenol thioester was not observed.

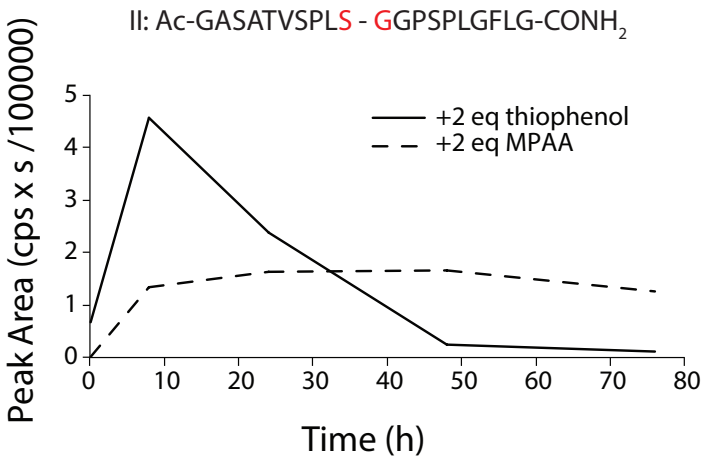


Figure 7.9: Kinetics of the thiophenolic peptide ester in the reaction II. Comparison between 2 equivalents of thiophenol (solid line) and 2 equivalents of MPAA (dashed line).

Addition of HOBt Improves the Conversion of Thioester-Amine Chemical Ligation

Toni Todorovski,[1] David Suñol,[1] Antoni Riera,[1,2] Maria J. Macias[1,3]

[1]*Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10 08028 Barcelona, Spain*

[2]*Departament de Química Orgànica, University of Barcelona, Martí i Franquès, 1, Barcelona 08028, Spain*

[3]*Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluis Companys, Barcelona 23 08010, Spain*

Received 6 May 2015; revised 10 September 2015; accepted 18 September 2015
Published online 22 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/bip.22745

ABSTRACT:

The syntheses of large peptides and of those containing non-natural amino acids can be facilitated by the application of convergent approaches, dissecting the native sequence into segments connected through a ligation reaction. We describe an improvement of the ligation protocol used to prepare peptides and proteins without cysteine residues at the ligation junction. We have found that the addition of HOBt to the ligation, improves the conversion of the ligation reaction without affecting the epimerization rate or chemoselectivity, and it can be efficiently used with peptides containing phosphorylated amino acids.

Keywords: *ligation of peptide fragments through direct aminolysis; cysteine-free native chemical ligation; native chemical ligation with β-branched residues at ligation junction; formation of highly reactive HOBt esters*

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Maria J. Macias, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain;
e-mail: maria.macias@irbbarcelona.org

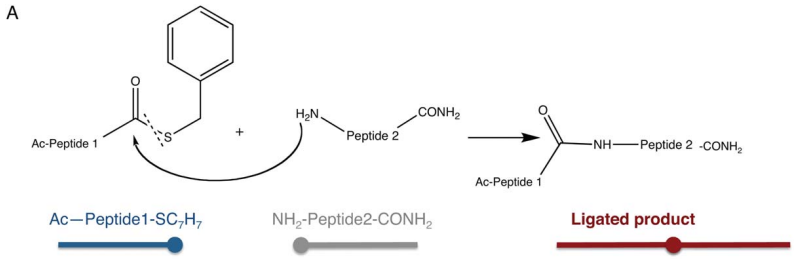
Contract grant sponsor: European Union's Seventh Framework Programme for research, technological development and demonstration
Contract grant number: IRBPostPro 246557
Contract grant sponsor: Spanish National Research Program
Contract grant number: SAF2011-25119, BFU2014-53787-P, CTQ2014-56361, 2009SGR00901
Contract grant sponsor: Consolider
Contract grant number: CSD2009-00080

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of any preprints from the past two calendar years by emailing the Biopolymers editorial office at biopolymers@wiley.com.

INTRODUCTION

Solid-phase peptide synthesis (SPPS), developed in the 60s by B. Merrifield,[1] is nowadays the standard methodology for chemical preparation of peptides. Although peptides up to 50 residues can be obtained using this methodology [2,3] the linear synthesis of large or complex sequences frequently yield mixtures containing undesired products and is usually not practical. Improvement of these syntheses often require the application of a convergent approach, dissecting the native sequence in segments that are connected using a ligation reaction (Figure 1, Panel A).[4–6] These protocols also facilitate the introduction of modified amino acids in proteins (including phosphorylated or acetylated residues or non-natural amino acids) as well as to the addition of specific labels for antibody recognition, protein detection and cell imaging.[7]

The direct coupling of two peptides was reported for the first time by Kemp *et al.* in 1974.[8,9] This method allows a ligation reaction to occur independently of the amino acids pairs at the ligation junction. The reaction couples a peptide containing an activated carboxylic derivative at its C-terminus to a second peptide carrying an unprotected NH_2-group at its N-terminus. The success of the method depends on the aqueous solubility of the designed segments and suffers from the epimerization of the product at the ligation site. Improvements to the direct aminolysis method were described later by Blake



B

m/z exp	Ac—Peptide1-SC_7H_7	Peptide2-$CONH_2$	*m/z* exp
1037.5	GASATVSPLS•	•ASPGYS	580.3
1007.5	GASATVSPLG•	•CGPSPLGFLG	946.5
727.8	LYRAG•	•GGPSPLGFLG	900.5
1049.5	GASATVSPLV•	•VGPSPLGFLG	942.5
1063.5	GASATVSPLL•	•GSPGYS	566.3
858.8	PSpSPGSV•	•YSPGYS	672.3
		•LGPSPLGFLG	956.6
		•LARPSVI	754.8

C

Ligated products	HOBt +	HOBt -	
LYRAG•GSPGYS	89%	83%	Yield
GASATVSPLS•VGPSPLGFLG	36%	10%	Yield
GASATVSPLV•VGPSPLGFLG	34%	5%	Yield
LYRAG•YSPGYS	57%	33%	Yield
GASATVSPLV•LGPSPLGFLG	36.7%	16.7%	Conv.
PSpSPGSV•LARPSVI	31.5%	18.8%	Conv.

FIGURE 1 Schematic representation of the ligation reaction through direct aminolysis (Panel A) Peptides designed for the ligation reactions used in this study. The m/z experimental values (which are same as theoretically calculated ones) obtained for each peptide is shown next to the corresponding sequences prior the ligation (Panel B). Ligated peptides sequences (the ligation site is represented with a circle) are listed, with yields and conversion percentages obtained for ligation reactions carried in the presence or in the absence of HOBt (Panel C).

et al. and Aimoto *et al.*[10–12] However, in these protocols the reaction conditions cause epimerization of the C-terminal thioester residue, and therefore only non-epimerizable amino acids (glycine or proline) can be introduced at this position.

A breakthrough in the ligation strategy was the development of the native chemical ligation (NCL) approach.[4] In this reaction, carried out in aqueous buffers, the peptide containing a C-terminal thioester is attacked by the thiol group of a cysteine residue located at the N-terminus of the second peptide. The thioester intermediate forms a native amide bond after an intramolecular rearrangement. This ligation reaction can also be used with recombinant proteins to generate the segment carrying the thioester group at the C-terminus *via* intein-mediated cleavage.[13] A drawback in this process is that proteins synthesized by NCL are generated mainly from two segments and are often limited by the presence of cysteines in the sequence. If cysteines are not present in the native sequence, the ligated peptide would require a desulfurization step which is not always straightforward.[14,15]

Novel approaches have recently been described by Wong and coworkers, that partially overcome the above mentioned drawback.[16,17] Their works report the chemical ligation of peptides and proteins without cysteine residues at the ligation junction with high chemoselectivity and low levels of

epimerization. The main limitation of these strategies is the ligation rate, normally requiring long reaction times, ranging from a few hours to several days.

We considered that adding 1-hydroxybenzotriazole (HOBt), a common reagent used in SPPS, to the protocol described by Wong and coworkers,[16] could improve the ligation yield and the conversion of the ligation reaction without affecting the epimerization rate or chemoselectivity. To test this hypothesis, we prepared various peptides with different amino acids at the ligation junction including a WW domain sequence (of 35 aa) containing at non-natural residue. Some of the peptides included phosphorylated amino acids, which are often a challenge due to the steric hindrances that they introduce during couplings, as well as their tendency to undergo alpha, beta-elimination under strong reaction conditions. Our results indicate that, the presence of HOBt increases the yield and conversion of the ligation, thus reducing the possibility of side-products and degradation. These experimental conditions are especially suited for the preparation of intrinsically unfolded protein segments that are particularly prone to degradation.

MATERIALS AND METHODS

General Procedure for the Synthesis of Peptide Thioesters. Six peptide thioesters (Figure 1, Panel B) were synthesized using the SPPS approach and 4-sulfamylbutyryl resin (Supporting Information). The coupling of the first residue was performed as recommended by the manufacturer.[18] In brief: Fmoc-AA-OH (4 equiv.), DIC (4 equiv.) and 1-methylimidazole (4 equiv.) were dissolved in DCM (12 μL/μmol). The solution was added to 1 equiv. of the resin and the mixture was gently agitated for the next 18 h at 25°C. The resin was then washed 5 times with DMF and 5 times with DCM before being dried. Coupling efficiency was estimated through resin load determination by UV analysis of Fmoc-release. The remaining amino acids were incorporated manually using 5 equiv. of the corresponding amino acid derivatives activated with 4.9 equiv. of DIC in the presence of 4.9 equiv. of HOBt in DMF at 100 μmol scale using Fmoc/tBu chemistry. The efficiency of the couplings was verified with the Kaiser test.[19]

After completion of the sequences, all peptides were either acetylated or Boc-protected at the *N*-terminus prior to final peptide cleavage from the resin. The procedure for the peptide activation was the following: iodoacetonitrile (67 equiv.) and DIPEA (13 equiv.) in DMF (72 μL/μmol) were added to the resin and gently agitated for the next 18 h. The resin was then washed six times with DMF, six times with DCM and then dried. In the next step, the peptide was cleaved from the activated resin using benzyl mercaptan (50 equiv.) and DIPEA (13 equiv.) in DMF (72 μL/μmol) in an 18 h reaction with gentle agitation. The peptide solutions were filtered and collected in a round-bottom flask. The resin was washed twice with DMF (4 mL each time) and the combined filtrates were concentrated under vacuum in order to completely remove DMF. The crude product was dissolved in 6 mL of a TFA/triisopropylsilane/water mixture (95:2.5:2.5 by vol.) and stirred at room temperature for 3 h, yielding the peptide thioester without the side-chain protecting groups. The peptide thioesters were then precipitated in cold diethyl ether and centrifuged (3000*g*). The pellet was washed twice with cold ether, dried, and stored at −20°C.

The crude thioesters were analyzed by LC-MS and, depending on their purity, some of them were additionally purified by preparative RP-HPLC prior to their use in the ligation reactions.

General Procedure for Free N-Terminus Peptide Synthesis. Peptides containing a free *N*-terminus (Figure 1, Panel B) were synthesized manually using a Rink amide AM resin and Fmoc/tBu chemistry with 5 equiv. of the amino acids derivatives activated by 4.9 equiv. of DIC in the presence of 4.9 equiv. of HOBt in DMF at the 100-μmol scale. The efficiency of the manual coupling was verified by the Kaiser test.[19]

The resin-bound peptides were cleaved and deprotected with TFA containing a scavenger mixture of water, thioanisol, and ethanedithiol (90:5:2.5:2.5 by vol) at RT for 2 h. They were then precipitated in cold diethyl ether and centrifuged (3000 *g*). The pellet was washed twice with cold ether, dried and stored at −20°C.

Finally, the crude peptides were purified by preparative RP-HPLC. Pure fractions were collected, lyophilized, and stored at −20°C. They were then analyzed by MALDI-MS and LC-MS prior to their use in ligation reactions.

General Procedure for Peptide Ligation Reactions. The free *N*-terminus peptides (1.5 equiv. or 0.5 equiv.) were dissolved in 120 μL of ligation buffer *(Ligation buffer = NMP:6M GnHCL + 1M HEPES = 4:1 (v/v). A buffer containing 6 M GuHCl and 1M HEPES was prepared (50 mL) and adjusted to pH 8.5 using 25% NaOH solution and degassed. For each ligation trial 100 μL of the buffer was mixed with 400 μL of NMP)*. The resulting solution was used to dissolve the thioester peptide (between 0.3 and 1.5 μmol) (*Depending of the peptide thioester amount the final volume of the ligation buffer was different (the volume of 80 μL and the described procedure is for peptide thioester amount of 1 μmol). However, the free NH_2 peptide was always 1.5 equiv. higher (*Table I, *entry 1, 2, 5, 6 and 7) or 0.5 equiv. lower (*Table I, *entry 3, 4 and 8) from the peptide thioester equivalents)*. For each reaction, the solution was separated into two equal parts, and 15 μL of HOBt dissolved in the same ligation buffer (2 equiv. based on the amount of peptide thioester) were added to the ligation series with HOBt, while 15 μL of the ligation buffer was added to the other part, for comparison, as a control. For the ligations with HOAt and DIPEA, 2 equiv of HOAt, or, 4 equiv of DBU and DIPEA were added to the ligation mixture, respectively.

Finally, thiophenol (2% by volume, 1.5 μL) was added to the reactions in the presence or absence of activators, and the resulting mixture was incubated at 37°C with gentle agitation until the reaction was completed. At various time points the reaction was followed by LC-MS, and in some cases by MALDI-MS and RP-HPLC, depending on the peptide sequences and amino acids at the ligation junction (Table I, Supporting Information Table I). At each time point, 8 μL aliquots of the ligation mixture were taken and quenched by the addition of 0.1% TFA in water (32 μL) or *tris*(2-carboxyethyl)phosphine (TCEP) solution (32 μL, of a 10 mg/mL solution) when the products contained cysteine residues. The quenched mixtures were diluted up to 1.5 mL with % MeCN/0.1% FA in water (Supporting Information Table I) and stored at −20°C.

Table I Ligation Combinations Used in the Study With the Experimental Masses for the Product and Exchanged Thiophenol Thioester Used for Monitoring the Reaction Kinetics

	Ligation Combinations	Experimental *m/z* of the Product	Experimental *m/z* of Exchanged Thiophenol Thioester SC_7H_7» SC_6H_5	Yield (Y)/conv.(C) With HOBt addition (%)	Yield (Y)/conv.(C) Without HOBt Addition (%)
1	Ac-GASATVSPL***G***-SC_7H_7 + NH_2-***G***GPSPLGFLG-$CONH_2$	$[M+H]^+=1783.4$ $[M+2H]^{2+}=892.2$	$[M+H]^+=993.5$	n.d	n.d
2	Ac-GASATVSPL***S***-SC_7H_7 + NH_2-***G***GPSPLGFLG-$CONH_2$	$[M+H]^+=1813.2$ $[M+2H]^{2+}=907.1$	$[M+H]^+=1023.5$	n.d	n.d
3	Ac-LYRA***G***-SC_7H_7 + NH_2-***G***SPGYS-$CONH_2$	$[M+H]^+=1168.5$	$[M+H]^+=713.5$	89 (Y)	83 (Y)
4	Ac-LYRA***G***-SC_7H_7+ NH_2-***A***SPGYS-$CONH_2$	$[M+H]^+=1182.5$	$[M+H]^+=713.5$	n.d	n.d
5	Ac-GASATVSPL***S***-SC_7H_7+ NH_2-***V***GPSPLGFLG-$CONH_2$	$[M+H]^+=1855.4$ $[M+2H]^{2+}=928.2$	$[M+H]^+=1023.5$	36 (Y)	10 (Y)
6	Ac-GASATVSPL***V***-SC_7H_7+ NH_2-***V***GPSPLGFLG-$CONH_2$	$[M+H]^+=1867.2$ $[M+2H]^{2+}=934.1$	$[M+H]^+=1035.5$	34 (Y)	5.2 (Y)
7	Ac-GASATVSPL***V***-SC_7H_7+ NH_2-***L***GPSPLGFLG-$CONH_2$	$[M+H]^+=1881.4$ $[M+2H]^{2+}=941.2$	$[M+H]^+=1035.5$	36.7 (C)	16.7 (C)
8	Ac-LYRA***G***-SC_7H_7+ NH_2-***Y***SPGYS-$CONH_2$	$[M+H]^+=1274.6$	$[M+H]^+=713.5$	57 (Y)	33 (Y)
9	Ac-PSpSPGS***V***-SC_7H_7+ NH_2-***L***ARPSVI-$CONH_2$	$[M+H]^+=1488.6$ $[M+2H]^{2+}=744.8$	$[M+H]^+=844.8$	31.5 (C)	18.8 (C)
10	Ac-GASATVSPL***S***-SC_7H_7+ NH_2-***C***GPSPLGFLG-$CONH_2$	$[M+H]^+=1859$ $[M+2H]^{2+}=930$		n.d.	n.d

The ligation junctions are in bold and italic.

Synthesis of Modified Human Pin 1 Protein Using the Cysteine-Free Ligation Protocol With HOBt. Human Pin 1 protein was synthesized using the ligation protocol with HOBt, through ligation of segment 1 with segment 2 (Scheme 1, Supporting Information figures). Segment 1 was synthesized as thioester following the above described procedure and acetylated at the *N*-terminus. Segment 2, up to the Asn, was synthesized in a microwave-assisted peptide synthesizer (Liberty Blue, CEM Corporation). Starting from Phe, the rest of the sequence of segment 2 was coupled manually following the general procedure for peptide synthesis. The ligation was performed following the protocol described previously, using 3.2 μmol of segment 1 and 4.8 μmol of segment 2. The final volume of the ligation buffer was 700 μL due to the tendency of segment 2 to form a viscous gel (theoretically the final volume of the ligation buffer should be 240 μL). The reaction was followed by MALDI-MS and stopped after 78 h. The crude product was purified by RP-HPLC, and the fractions corresponding to the ligation product were collected, lyophilized, and stored at -20°C. In order to remove the Lys sidechain protecting group (ivdDe), the product was dissolved in DMF containing 3% of hydrazine. The solution was left at room temperature for 1h with a gentle mixing. Afterwards, DMF was removed under reduced pressure and the solid residue was dissolved in water solution containing 20% MeCN/0.1% FA. The solution was lyophilized and afterwards stored in the freezer at -20°C.

MALDI-MS. Mass spectra were acquired on a 4700 Proteomic analyzer or on a 4800 Plus MALDI TOF/TOF Analyzer (AB Sciex) calibrated with Calmix (Calmix 4700 Proteomics Analyzer Calibrating Mixture). The mass spectra were recorded in positive reflector TOF mode in the *m/z* range 500–2000 or 1500–4500 at a fixed laser intensity of 4800 using alpha-cyano-4-hydroxycinnamic acid as a matrix. Spectra were analyzed by Data Explorer software (Version 4.6, Applied Bisystems GmbH).

RP-HPLC. Crude peptides, peptide thioesters, and ligation mixtures were purified using an Aqua C_{18}-column (internal diameter 4.6 mm, length 150 mm, particle size 5 μm, pore size 12.5 nm, Phenomenx) with a linear gradient from 10% to 24% aqueous acetonitrile (0.1% TFA) over 1.2 min, followed by a gradient of 24% to 57% for the next 42 min with a flow rate of 1 mL/min using an ÄKTA purifier 10 HPLC System (GE Healthcare Life Sciences). Fractions were analyzed by MALDI-MS and those containing the products were collected, lyophilized, and stored at -20°C.

Ligation mixtures were analyzed on the microbore Aqua C_{18}-column using a linear gradient from 4.75% to 57% or from 9.5% to 57% aqueous acetonitrile (0.1% TFA) for 25 min at a flow rate of 1 mL/min.

LC-MS. The ligation mixtures (30 μL, at final concentration of approx. 60 μmol/L for each time point) were injected into the HPLC-MS system (Waters, model Alliance 2796 with a quaternary pump

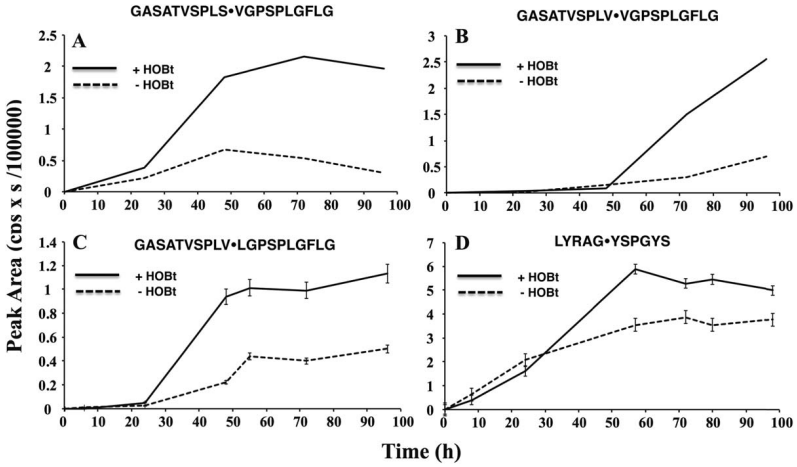


FIGURE 2 Kinetics of ligated product formation in ligation reactions with HOBt and without (solid or dashed lines respectively). The combinations shown are GASATVSPLS-VGPSPLGFLG (panel A), GASATVSPLVVGPSPLGFLG (panel B), GASATVSPLV-LGPSPLGFLG (panel C) and LYRAG-YSPGYS (panel D). The reaction was followed by LC-MS and through peak integration of the product ions at m/z 928.2 (doubly charged, panel A), m/z 934.1 (doubly charged, panel B), m/z 941.2 (doubly charged, panel C), m/z 1274.6 (singly charged, panel D).

and UV/Vis dual absorbance detector Waters 2487 connected with ESI-MS model Micromass ZQ). The separation was achieved on a Sunfire C_{18}-column (internal diameter 2.1 mm, particle size 3.5 μm, length 100 mm) using a linear gradient from 10% or 20% to 100% aqueous acetonitrile (0.1% FA) in 8 min at a flow rate of 0.3 mL/min. The mass spectra were acquired for a mass range from *m/z* 500 to 2000 in positive ion mode using five different cone voltages ranging from 5 to 70 V. The TIC spectra used for peak integration correspond to the cone voltage of 30V and were analyzed by Masslynx 4.0 software (Waters).

Kinetics of Cysteine-Free Ligation With and Without HOBt. As mentioned above, depending on the peptide sequence and amino acids at the ligation junction, the kinetics of the ligation reaction was followed using different time intervals (Supporting Information Table I). At each time point, aliquots of 8 μL of the ligation mixture were taken and quenched by the addition of 0.1% TFA in water (32 μL) or of a TCEP solution (32 μL, of a 10 mg/mL solution) if the products contained cysteine residues. The samples were diluted up to 1.5 mL with solution containing H_2O/MeCN/FA at different ratios by volume (Supporting Information, Table I, column 4) and analyzed by LC-MS, in some cases additionally by MALDI-MS or analytical RP-HPLC. The kinetics was followed through integrating the peak areas of the single or double charged ions of ligated product and exchanged thiophenol thioester (Table I). Yields were calculated after the RP-HPLC purification of the crude ligation mixture, while the given conversions where calculated from the integrated ions ratio of ligated product vs peptide thioester in the TIC spectra. In those cases, were the experiments were performed in triplicate the standard error was calculated (Reagents provided as supplementary material).

RESULTS AND DISCUSSION

To test if the addition of HOBt to the reaction protocol described by Wong and coworkers[16] can improve the rate and conversion of ligation we first selected a set of different native sequences, present in TGIF1 and FoxH1 proteins (Supporting Information Figure 1) and collected in Figure 1. The selected peptides contain β-branched (Table I, entries 5, 6 and 7) or aromatic residues (Table I, entry 8) at the ligation site, since these residues -and also phosphorylated amino acids- are often present in protein binding interfaces and are known to add complexity to the ligation reaction.[20,21] All sequences of the acetylated *C*-terminal peptide thioesters and free *N*-terminal amidated peptides used here are given in Figure 1. These peptides were subjected to the chemical ligation with and without HOBt as additive. We performed the ligation studies following the protocol described by Wong et al.,[16] in conjunction with a set of experiments where 2 equivalents of HOBt (1:2 molar ratio of peptide thioester:HOBt) were added to the ligation

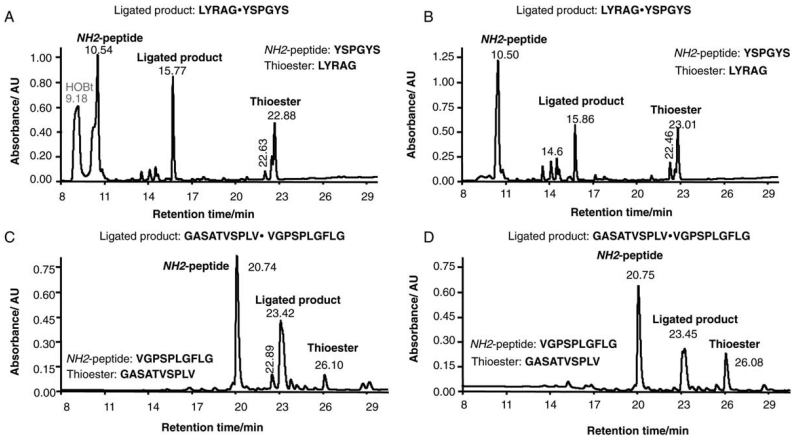


FIGURE 3 Reversed-phase HPLC chromatograms of the ligation reaction with HOBt (Panel A and C) and without HOBt (Panel B and D) using two peptides, LYRAG-YSPGYS (Panel A and B) and GASATVSPLVVGPSPLGFLG (Panel C and D).

mixture. The combinations are shown in Figure1, panel C and Table I.

In all cases, the final conversion, based on the peak integration of the product ions in the LC-MS experiments, and yield was higher when HOBt was present in the ligation mixture (Figures 1 panel C, Figure 2, Supporting Information Figure 2, Table I) independently of the peptide thioester/peptide molar ratio. Moreover, in the cases when aromatic or β-branched amino acids were present at the ligation junction, the influence of the added HOBt on the final conversion/yield was greater (Figure 1 panel C, Figure 2 bold lines) than when the peptide had less sterically hindered amino acids (Figure 1 panel C, Supporting Information Figure 2 bold lines). Interestingly, the HOBt addition specifically increased the final conversion/yield, while having very little or no influence on the rate of ligation (except in the case of GY ligation junction, Figure 2 panel D). The addition of 10 equiv. of HOBt did not improve the final conversion or the reaction rate further (Supporting Information Figure 3), but led to a reduction in the conversion rate (Supporting Information Figure 3 dashed lines). We attribute this observation to a decrease of the pH of the ligation mixture caused by the acidic nature of HOBt.[22] Furthermore, using the above-described conditions the product can be obtained with high purity since it elutes at a different retention time than the starting peptide reagents (thioester and the free N-terminus peptide), as displayed in Figure 3.

In another set of experiments performed as controls, we ligated the same set of NH_2-peptide sequences as before, in the

Table II NH_2-Free *C*-Terminal Peptide Thioesters and *N*-Terminal Peptide-Free Amines Used in the Ligation Studies

Peptide Thioesters	*m/z* theor	*m/z* exper	Peptides	*m/z* theor	*m/z* Exper	Lig. Com.
1. NH_2-GASATVSPLG-SC_7H_7	965.54	965.58	A. NH_2-GGPSPLGFLG-$CONH_2$	900.49	900.47	1-A 2-A 2-B 3-B 4-C
2. NH_2-GASATVSPLS-SC_7H_7	995.49	995.46	B. NH_2-VGPSPLGFLG-$CONH_2$	942.54	942.55	
3. NH_2-GASATVSPLV-SC_7H_7	1007.53	1007.55	C. NH_2-CGPSPLGFLG-$CONH_2$	946.48	946.46	
4. NH_2-GASATVSPLL-SC_7H_7	1021.54	1021.57				

Lig. Com.: ligation combinations.

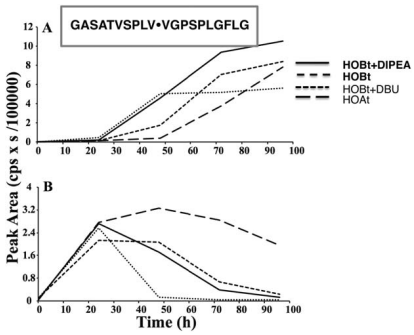


FIGURE 4 Kinetics of ligated product formation (panel A) and transthioesterification step (panel B) in the ligation reaction with HOBt+DIPEA (solid line), HOBt (dashed lines), HOBt+DBU (dotted lines) and HOAt (long-dashed lines). The ligation was performed using ligation combinations GASATVSPLV-VGPSPLGFLG. The reaction was followed by LC-MS and through peak integration of the product ion at m/z 934.1 (doubly charged, panel A), and through peak integration of the ion at m/z 1035.5 (singly charged, panel B).

presence of HOBt but using unprotected thioesters (at the N-terminus) (Table II). In all cases the formation of the expected ligation product was observed (Supporting Information Figure 4 solid lines); however, in the presence of cyclic side-products (Supporting Information Figure 4 dashed lines). According to the observed *m/z* values, the cyclization products are formed due to the intramolecular cyclization of the peptide thioesters, a competitive side reaction favored by the unprotected *N*-termini of the peptide thioester.

Finally, ligation reactions with or without *N*-terminal protected thioesters and peptides containing a N-term Cysteine were much faster than in all the other ligation reactions tested, (Supporting Information Figure 5).

The addition of HOBt suppresses the 5(4H)-oxazolone formation, preventing plausible racemization, and thus, improving the aminolysis rate through the formation of more effective acyl donors esters.[23] For this reason, we hypothesized that the presence of HOBt in the ligation mixture improves the ligation reaction yield and also contribute to increasing the conversion.

Effects of the Presence of DBU and DIPEA Activators

The addition of a base activator can improve the final conversion and speed up the reaction. To test this hypothesis, we performed ligation reactions with HOBt and DBU and DIPEA as activating bases, since they are strong bases but weak nucleophiles (Figure 4, panel A dotted line and solid line, correspondingly). To test the effect of the activators we chose Val-Val residues at the ligation site because HOBt-enhances the ligation reaction in cases where β-branched or aromatic amino acids were present at the junction (Figure 1 panel C, Figure 2 and Table I). In the case of DIPEA, we observed limited improvement in the final conversion (Figure 4, panel A solid line), while for DBU the final conversion was lower, despite the fact that up to 50 h the kinetics of the reaction was almost identical to that achieved using DIPEA (Figure 4, panel A dotted line). The observation is directly related to a transthioesterification step (which can be followed by a singly charged ion at *m/z* 1035.5), as after 50 h the number of ions corresponding to the thiophenol thioester exchange was close to zero in the case of DBU (Figure 4, panel B dotted line). Our results indicate that the presence of DBU do not contribute to increasing the ligation rate under the experimental conditions we have investigated, while in the case of DIPEA we could detect small improvements regarding the final conversion.

Potential Alternatives to HOBt

Regarding the choice of an alternative to HOBt, we used HOAt, a molecule similar to HOBt but with a lower pKa (pKa 3.28 and 4.60 respectively)[24,25] and therefore potentially a better leaving group. However, we observed that the final conversion and also the reaction rate were similar to those obtained with HOBt in the experimental conditions tested (Figure 4, panel A). This observation could be attributed to the fact that HOAt requires double the induction time relative to all other cases (Figure 4, panel A long-dashed line). This extended time can be explained by the fact that the transthioesterification is also slower for HOAt (Figure 4, panel B long-dashed line). These observations imply that if reactions were allowed to proceed for longer (more than 96 h), we would achieve a higher final conversion; however, this would also cause an increase in side products, thereby reducing the yield of the product of interest and adding complexity to the purification step. However, when the ligation reactions are more favorable, the use of HOAt could be beneficial with regards to the final conversion, as increases in the reaction time will not have a strong influence on the amount of side products.

Synthesis of a WW Domain Containing a Non-Natural Amino Acid

Finally, as an application of the ligation protocol here described we synthesized a modified WW domain using the peptidyl-prolyl isomerase I (Pin I, 34 AA, Supporting Information Scheme 1) sequence as the template. Pin1 is a modular protein that contains a WW domain, responsible for target recognition,

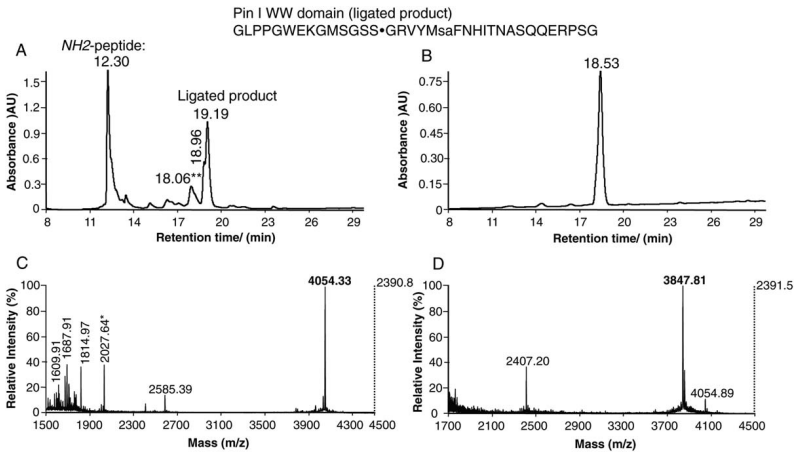


FIGURE 5 Reversed-phase chromatograms (Panel A and B) and MALDI-MS (Panel C and D) of modified human Pin I protein. After purification of the crude Pin I (Panel A) the pure sample (Panel B) was analyzed by MALDI-MS with ivdDe group present at the Lys residue (Panel C) and after removal of the ivdDe group (Panel D). The masses corresponding to the Pin 1 protein are labeled in bold, while the doubly charge ion of the Pin 1 protein is with asterisk. The peak in Panel A, labeled with double asterisk corresponds to the Pin I protein with a 16 Da higher mass.

and the catalytic domain involved in the cis-trans isomerization of phosphor-Ser/Thr-Pro bonds.[26,27] The folding pathway of the Pin1 WW domain has been the subject of many studies,[28] which have shown that the domain can be unfolded and refolded with high efficiency. In addition, many structures of this WW domain in complex with target peptides are described in the literature.[29] Since folding and binding studies often require the production of many WW proteins carrying mutations, we explored the possibility of using peptide synthesis and refolding as an efficient strategy to prepare these samples. If some of these mutations are designed to include non-natural amino acids, a ligation strategy might be the best choice to optimize the segment's design and purification. As an example, we prepared a WW sequence with a non-natural residue 2,4,6-trimethylphenyl Ala (Msa)[30–32] at position 19 and several point mutations: Lys 1 was replaced with Gly, whereas Arg 9 and Arg 12 were replaced by Ala. The replacement of both arginines prevented possible intramolecular cyclization with the guanidinium group of the Arg side-chain, while the Lys at position 8 was protected with ivdDe. The ligation junction was chosen at a Ser-Gly (position 14-15) due to the expected faster ligation rate when compared with the other options (Supporting Information Figure 2, panel B). The crude product was purified by RP-HPLC (Figure 5, panel A) and the modified Pin I was obtained with high purity (Figure 5, panel B and C). After successful removal of the ivdDe side chain group (Figure 5, panel D) the pure product was obtained at final yield of 15%. Despite the low yield obtained in this example, the synthesis of the modified Pin1 is a proof of principle of the applicability of the method, and we consider that the yield could be significantly improved with further optimizations.

Addition of HOBt to Peptides Containing Phosphorylated Residues

We also tested the HOBt-ligation protocol using peptide sequences that correspond to TGIF1 protein containing phosphorylated Ser (Table I entry 9) in the sequence of the acetylated *C*-terminal peptide thioester. The expression of proteins containing phosphorylated residues is complicated and expensive. In addition, the synthesis of phosphorylated polypeptides using microwave (MW)-assisted SPPS is inefficient due to the alpha,beta-elimination of the phosphate group under MW conditions catalyzed by piperidine.[33] Therefore, the cysteine-free ligation with HOBt could also be useful in the synthesis of phosphorylated polypeptides. The addition of HOBt significantly improved the final conversion (Figure 1 panel C, Figure 6, Table I entry 9 and Supporting Information). Furthermore,

the phosphorylated residue did not participate in any side reaction with the added HOBt, and the products that correspond to truncated peptide sequences or sequences with dephosphorylated Ser residue were not detected.

Mechanism of Reaction

Our experimental results suggest a possible general mechanism for the ligation reactions (Supporting Information Scheme 2). The initial conversion of the benzyl to the phenyl thioester is followed by the subsequent substitution by the more reactive HOBt ester. The lower pKa of HOBt (pKa 4.6) compared to that of thiophenol (pKa 6.6) would make the former a better leaving group.

Our results indicate that the limiting reaction step is the transthioesterification, as the ligation reaction does not occur in the absence of thiophenol. It is known that the thiophenol is able to undergo rapid and almost complete thiol-thioester exchange[34] and is an excellent leaving group. Therefore, in the absence of thiophenol, the initial benzyl thioesters won't be able to undergo aminolysis or to form active HOBt esters, since alkanethiols are poor leaving groups.[34] When thiophenol is replaced by 4-(carboxymethyl)thiophenol (MPAA), which has a pKa of 6.6, similar to that of thiophenol, the rate of transthioesterification is lower (Supporting Information Figure 7, panel A dashed line). We believe that the ligation buffer can affect this step since the presence of MPAA, which shows better solubility in aqueous buffers, induces a higher and faster transthioesterification than thiophenol.[34] Moreover, in our experiments, the concentration of the MPAA thioester formed after the transthioesterification remained constant during the experiment. Consequently, the ligation conversion is six times lower, as shown by peak integration in the LC-MS experiments (Supporting Information Figure 7, panel B solid and dashed line, respectively). In addition, when HOBt, was added to the ligation mixture in the absence of thiophenol, no ligation product was detected (data not shown). Again, as we mentioned before, we believe that this is due to the fact that alkanethiols are poor leaving groups and therefore the preformed benzyl thioesters are not able to react giving the active HOBt esters. In summary, the addition of HOBt facilitates the reaction of the exchanged phenyl thioester with the free *N*-terminus peptide, probably through formation of highly reactive HOBt-ester, However, it does not have a significant effect on the extent of transthioesterification. These observations are shown in Supporting Information Figure 8 (panel A-D), where the final conversion of the transthioesterification step (Supporting Information Figure 8, dotted and long-dashed lines) differs from the final conversion of the product (Supporting Information Figure 8, dashed and solid lines). In addition, the induction time of the reaction is related to the transthioesterification step, i.e., when the transthioesterification reaches the maximum and begins to drop, the formation of the product starts immediately.

CONCLUSIONS

We have found that the addition of HOBt to the ligation of *C*-terminal peptide thioesters with various free *N*-terminus peptides increases the ligation conversion, especially when β-branched or aromatic amino acids are present at the ligation junction. This increase is probably due to the *in-situ* formation of a highly reactive HOBt-ester. Furthermore, all reactions proceeded without epimerization and the method is chemoselective. Only when Lys residues are present in the sequence it is expected that the residue's side chain react with the *in-situ* formed HOBt-ester in a similar way as with a thioester previously reported by Wong et al.[16] Despite the long reaction times, the method shows potential for application in the synthesis of peptide sequences that cannot be generated by other ligation techniques, especially in the presence of β-branched or aromatic amino acids at the junction site. The improvement of the method is also applicable to the synthesis of phosphorylated peptides that, despite recent developments in protein chemistry, are not readily obtainable due to their tendency to suffer alpha,beta-elimination of the phosphate group.

T.T. is a recipient of a FP7 Marie Curie Actions (COFUND Grant program), and D.S. is a recipient of "la Caixa"-IRB Barcelona International PhD Fellowship. The authors thank IRB Barcelona for support, Dr. M. Royo for discussions regarding some peptide synthesis, Dr. M. Vilaseca and of Dr. I. Fernandez for MS analysis. They also thank Alvaro Rol Rua for the preparation of the Msa amino acid. M.J.M. is an ICREA Programme Investigator.

REFERENCES

1. Merrifield, R. B. J Am Chem Soc 1963, 85, 2149–2154.
2. Alewood, P.; Alewood, D.; Miranda, L.; Love, S.; Meutemans, W.; Wilson, D. Methods Enzym 1997, 289, 14–29.
3. Chandrudu, S.; Simerska, P.; Toth, I. Molecules 2013, 18, 4373–4388.
4. Dawson, P.; Muir, T.; Clark-Lewis, I.; Kent, S. Science 1994, 266, 776–779.
5. Dirksen, A.; Meijer, E. W.; Adriaens, W.; Hackeng, T. M. Chem Commun (Camb) 2006, 15, 1667–1669.
6. Dawson, P. E.; Kent, S. B. H. Annu Rev Biochem 2000, 69, 923–960.
7. Wombacher, R.; Cornish, V. W. J Biophotonics 2011, 4, 391–402.
8. Kemp, D. S.; Bernstein, Z. W.; McNeil, G. N. J Org Chem 1974, 69, 2831–2835.

9. Kemp, D. S.; Choong, S. L. H.; Pekaar, J. J Org Chem 1974, 39, 3841–3847.
10. Blake, J. Int J Pept Protein Res 1981, 17, 273–274.
11. Aimoto, S.; Mizoguchi, N.; Hojo, H.; Yoshimura, S. Bull Chem Soc Jpn 1989, 62, 524–531.
12. Aimoto, S. Biopolymers 1999, 51, 247–265.
13. Vila-Perelló, M.; Muir, T. W. Cell 2010, 143, 191–200.
14. Pentelute, B. L.; Kent, S. B. H. Org Lett 2007, 9, 687–690.
15. Crich, D.; Banerjee, A. J Am Chem Soc 2007, 129, 10064–10065.
16. Payne, R. J.; Ficht, S.; Greenberg, W. A.; Wong, C. H. Angew Chem Int Ed 2008, 47, 4411–4415.
17. Wang, P.; Danishefsky, S. J. J Am Chem Soc 2010, 132, 17045–17051.
18. Novabiochem, Novabiochem Innovations 4/99; 1999.
19. Kaiser, E.; Colescott, R. L.; Bossinger, C. D.; Cook, P. I. Anal Biochem 1970, 34, 595–598.
20. Hackeng, T. M.; Griffin, J. H.; Dawson, P. E. Proc Natl Acad Sci USA 1999, 96, 10068–10073.
21. Coltart, D. M. Tetrahedron 2000, 56, 3449–3491.
22. Saha, A.; Nadimpally, C.; Paul, A.; Kalita, S.; Mandal, B. Protein Pept Lett 2014, 21, 188–193.
23. König, W.; Geiger, R. Chem Ber 1970, 103, 788–798.
24. Novabiochem. Novabiochem Innovations: 2/09; 2009.
25. Carpino, L. A. J Am Chem Soc 1993, 115, 4397–4398.
26. Ping Lu, K.; Hanes, S. D.; Hunter, T. Nature 1996, 380, 544–547.
27. Ranganathan, R.; Lu, K. P.; Hunter, T.; Noel, J. P. Cell 1997, 89, 875–886.
28. Luo, Z.; Ding, J.; Zhou, Y. Biophys J 2007, 93, 2152–2161.
29. Aragón, E.; Goerner, N.; Zaromytidou, A. I.; Xi, Q.; Escobedo, A.; Massagué, J.; Macias, M. J. Genes Dev 2011, 25, 1275–1288.
30. Medina, E.; Moyano, A.; Pericas, M. A.; Riera, A. Helv Chim Acta 2000, 83, 972–988.
31. Martín-Gago, P.; Gomez-Caminals, M.; Ramón, R.; Verdaguer, X.; Martin-Malpartida, P.; Aragón, E.; Fernández-Carneado, J.; Ponsati, B.; López-Ruiz, P.; Cortes, M. A.; Colás, B.; MacIas, M. J.; Riera, A. Angew Chem Int Ed 2012, 51, 1820–1825.
32. Martín-Gago, P.; Aragón, E.; Gomez-Caminals, M.; Fernández-Carneado, J.; Ramón, R.; Martin-Malpartida, P.; Verdaguer, X.; López-Ruiz, P.; Colás, B.; Cortes, M. A.; Ponsati, B.; Macias, M. J.; Riera, A. Molecules 2013, 18, 14564–14584.
33. Attard, T. J.; O'Brien-Simpson, N. M.; Reynolds, E. C. Int J Pept Res Ther 2009, 15, 69–79.
34. Johnson, E. C. B.; Kent, S. B. H. J Am Chem Soc 2006, 128, 6640–6646.

Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements

Rui Zhou[a,b,c,1], Gia G. Maisuradze[a,1,2], David Suñol[d], Toni Todorovski[d], Maria J. Macias[d,e], Yi Xiao[b], Harold A. Scheraga[a,2], Cezary Czaplewski[c], and Adam Liwo[c,1,2]

[a]Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301; [b]Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, China; [c]Laboratory of Molecular Modeling, Faculty of Chemistry, University of Gdańsk, 80-308 Gdańsk, Poland; [d]Structural Characterization of Macromolecular Assemblies, Structural and Computational Biology Programme, Institute for Research in Biomedicine, 08028 Barcelona, Spain; and [e]Catalan Institution for Research and Advanced Studies, 08010 Barcelona, Spain

Contributed by Harold A. Scheraga, November 5, 2014 (sent for review September 18, 2014)

To demonstrate the utility of the coarse-grained united-residue (UNRES) force field to compare experimental and computed kinetic data for folding proteins, we have performed long-time millisecond-timescale canonical Langevin molecular dynamics simulations of the triple β-strand from the Formin binding protein 28 WW domain and six nonnatural variants, using UNRES. The results have been compared with available experimental data in both a qualitative and a quantitative manner. Complexities of the folding pathways, which cannot be determined experimentally, were revealed. The folding mechanisms obtained from the simulated folding kinetics are in agreement with experimental results, with a few discrepancies for which we have accounted. The origins of single- and double-exponential kinetics and their correlations with two- and three-state folding scenarios are shown to be related to the relative barrier heights between the various states. The rate constants obtained from time profiles of the fractions of the native, intermediate, and unfolded structures, and the kinetic equations fitted to them, correlate with the experimental values; however, they are about three orders of magnitude larger than the experimental ones for most of the systems. These differences are in agreement with the timescale extension derived by scaling down the friction of water and averaging out the fast degrees of freedom when passing from all-atom to a coarse-grained representation. Our results indicate that the UNRES force field can provide accurate predictions of folding kinetics of these WW domains, often used as models for the study of the mechanisms of proein folding.

FBP28 WW domain | nonnatural variants | folding rates | free-energy landscapes | millisecond-timescale canonical MD simulations

Recent advances in computer simulation techniques have facilitated the direct study of the folding process of small fast-folding proteins, using all-atom force fields (1). However, it is important to validate the simulation methodologies, and the only way to accomplish this is a quantitative comparison with experimental data with proper statistics. The validation of all-atom simulation methodologies is still a major problem because of the differences between the experimental timescale (from multiple microseconds to seconds) and the theoretical one (from hundreds of nanoseconds to microseconds). To overcome this problem, many approximate coarse-grained methods have been developed during the past decade (2–5). One of them makes use of a physics-based united-residue (UNRES) force field developed in our group over the past years (6–14) (*SI Appendix*, Fig. S1 and *SI Materials and Methods*).

The folding and unfolding rates are among the most accessible quantitative observables for two- and multistate folding proteins; therefore, a study of protein folding kinetics can bridge microscopic motions and the world of experimental measurements. In analyzing protein folding kinetics, the differential rate equations and their integrated forms become more complex as the number of intermediate forms between the completely unfolded form and the native form increases. Therefore, to determine the mechanisms and the microscopic rate constants, it is necessary to vary them to obtain a computed folding trajectory that matches the one that is simulated by molecular dynamics.

To cover a sufficiently large timescale and obtain a stable folding trajectory theoretically, it is necessary to use a coarse-grained, rather than an all-atom, force field. For this purpose, use is made of the UNRES force field to compute folding trajectories by canonical Langevin dynamics simulations. Then, any intermediate states are identified, and the dependence of the fractions of unfolded, intermediate, and native states of the protein (averaged over all trajectories) are determined as a function of time, and the kinetic equations are fitted to these data to compare the calculated rate constants with those determined experimentally (15).

This general approach is illustrated here, as an example, with the triple-β–stranded WW domain from the Formin binding protein 28 (FBP28) (PDB ID 1E0L) (16) and its full-size and truncated mutants (15) (*SI Appendix*, Fig. S2). The FBP28 WW domain is a member of the WW-domain family (17), with its kinetics examined by possible two-state and three-state models. The FBP28 WW domain is a good model with which to study

Significance

In spite of recent advances made in computer simulation techniques, one of the main challenges in the protein-folding field is to bridge microscopic motions and experimental measurements. This paper demonstrates that the physics-based, coarse-grained united-residue (UNRES) force field, which has the ability to simulate folding of small- and midsize proteins in the millisecond timescale, can predict the folding kinetics correctly and bridge theoretical and experimental worlds. The results suggest that the use of the UNRES force field will open a new door to the understanding of protein motions at much longer timescales and help explain the differences between theoretical results and experimental observations.

Author contributions: G.G.M., H.A.S., and A.L. designed research; R.Z., G.G.M., D.S., T.T., M.J.M., C.C., and A.L. performed research; R.Z., G.G.M., M.J.M., Y.X., H.A.S., and A.L. analyzed data; and R.Z., G.G.M., M.J.M., H.A.S., and A.L. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank (PDB), www.pdb.org [PDB ID codes 2mw9 (Y11R), 2mwa (Y19L), 2mwb (W30F), 2mwf (ΔNY11R), 2mwd (ΔNΔCY11R), and 2mwe (ΔNΔCY11R/L26A)], and the BioMagResBank (BMRB), www.bmrb.wisc.edu [BMRB ID codes 25309 (Y11R), 25310 (Y19L), 25311 (W30F), 25315 (ΔNY11R), 25313 (ΔNΔCY11R), and 25314 (ΔNΔCY11R/L26A)].

[1]R.Z., G.G.M., and A.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: gm56@cornell.edu, has5@cornell.edu, or adam@chem.univ.gda.pl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1420914111/-/DCSupplemental.

β-sheet formation. It should be noted that the WW domains have been the subject of extensive theoretical (1, 14, 18–28) and experimental (15–17, 29–34) studies because of their small size, biological importance (35), and interesting fast-folding kinetics.

As indicated here, a controversy still exists about whether the FBP28 WW domain folding proceeds by a two- or three-state mechanism. Using temperature denaturation and laser temperature-jump relaxation experiments, Nguyen et al. (15) concluded that the FBP28 WW domain can be tuned between two-state and three-state kinetics by temperature changes, selected point mutation, and truncation. These experiments (15) indicated that, below the transition midpoint, the wild-type (WT) FBP28 WW domain deviates from single-exponential kinetics, implying at least a three-state folding scenario, with two folding rate constants of 0.030 μs^{-1} and < 0.0011 μs^{-1}, respectively. Above the folding-transition temperature, single-exponential kinetics were observed with a folding-rate constant of 0.071 μs^{-1} (15). On the other hand, Ferguson et al. (31) argued that the kinetics proceed by a two-state mechanism and that the biphasic kinetics observed by Nguyen et al. (15) might be related to misfolding and aggregation, rapidly forming ribbon-like fibrils at physiological temperature and pH, with morphology typical of amyloid fibrils. Recently, by using infrared and fluorescence spectroscopy in studying the FBP28 WW domain and its tryptophan mutants (W8Y and W30Y), Davis and Dyer (36) found that the folding mechanism for the FBP28 WW domain is similar to that proposed by others (15, 29, 32), but their W8Y and W30Y mutants (36) provide evidence of an intermediate dry molten globule state.

Many computational studies were also performed on the FBP28 WW domain to gain atomic details and femtosecond temporal resolution. By carrying out all-atom molecular dynamics simulations with a Gō-like model on the FBP28 WW domain, Karanicolas and Brooks (18) demonstrated that biphasic kinetics originated from slow formation of the C-terminal β-hairpin in the FBP28 WW domain. Xu et al. (25) applied an all-atom Monte Carlo simulation to study the folding kinetics and revealed that two major folding pathways exist that differ in the order and mechanism of hairpin formation. Mu et al. (20) carried out replica exchange molecular dynamics (REMD) with explicit water at the all-atom level to explore the details of the kinetics. The results of Mu et al. showed that the formation of the second turn in the transition-state structure was responsible for the stable intermediate state that would lead to aggregation and misfolded structures as proposed by Ferguson et al. (31). By studying the effects of macromolecular crowding and confinement on the transition state structures in comparison with bulk for the 1PIN WW domain, Cheung and Thirumalai (37) found that (*i*) the folding rates of this protein in the presence of crowding or in confined spaces typically, but not always, increase because of entropic destabilization of the denatured states and, (*ii*) depending on which phenomenon is dominant, the entropic stabilization or enthalpic interactions, the transition state structures can be similar to or different from those in bulk.

One way to settle the discrepancies between the theoretical and experimental studies is to compare the same parameters obtained from theoretical simulations and experimental data. Therefore, we have generated a large number of trajectories by canonical Langevin dynamics simulations with the UNRES force field to analyze the folding kinetics of the FBP28 WW domain and its six mutants (Y11R, Y19L, W30F, ΔNY11R, ΔNΔCY11R, and ΔNΔCY11R/L26A, where ΔN and ΔC denote the deletion of the five N-terminal and the four C-terminal residues, respectively). These mutants, as well as the WT FBP28 WW domain, are associated with a strand-crossing hydrophobic cluster Tyr11/Tyr19/Trp30, involved in either hairpin of each protein (15). We ran 512 trajectories for each system (about 5–8 μs formal time and effectively 5–8 ms of each trajectory, per the UNRES timescale) of canonical Langevin dynamics simulations, starting from the fully extended structure. Then, we identified the native, intermediate, and unfolded states and determined the dependence of the fractions of these three states of each protein (averaged over all trajectories) as a function of time and fitted kinetic equations to these time-dependent fractions. The calculated rate constants were compared with those determined experimentally (15). Moreover, the MD trajectories were analyzed in terms of free-energy landscapes (FELs) along the C^{α}-rmsd from the native structure and radius of gyration (R_g). We have also characterized the structures of all variants, using the data obtained from high-resolution NMR (*SI Appendix*, Fig. S2). The variants investigated adopt a similar fold to that of the wild-type protein, with small differences in and around the sites where mutations and deletions have been introduced.

Results

Kinetic Studies for Wild-Type FBP28 Domain and Its Mutants. The simulated results for the three full-size and three truncated mutants, as well as the WT FBP28 WW domain (*SI Appendix*, Fig. S2), are presented here. All parameters obtained from the fitting of the simulation data by Eq. **1** (single-exponential kinetics) and Eqs. **2** and **3** (double-exponential kinetics) are summarized in Table 1. Plots of the fractions of native structures vs. time, $[N](t)$ (light blue ragged lines), and the fractions of intermediate structures vs. time, $[I](t)$ (light blue ragged lines), along with fitting curves [single-exponential kinetics (dotted black line) and double-exponential kinetics (solid black line)] for

Table 1. Fitting results of native states by single- (Eq. 1) and double-exponential (Eqs. 2 and 3) kinetics

	Fitting results											
	Two-state model			Three-state model							Experimental data[§]	
Name	C_0	λ_0*, $\times 10^{-3}$ ns^{-1}	$\chi^{2\dagger}$, $\times 10^{-2}$	C_1	m_1	C_2	m_2	λ_1 (sim)[‡], ns^{-1}	λ_2 (sim)[‡], ns^{-1}	$\chi^{2\dagger}$, $\times 10^{-2}$	λ_1 (exp)[¶], μs^{-1}	λ_2 (exp)[¶], μs^{-1}
Wild type	0.67	0.43	2.4	0.72	0.11	0.15	0.57	0.085	0.00032	1.4	0.030(1)[#]	<0.0011
Y11R	0.80	3.0	1.8	0.80	0.12	0.10	1.97	0.040	0.0026	1.3	0.025(4)	<0.0014
Y19L	0.018	6.2	0.23	0.019	0.79	0.57	0.35	0.24	0.00041	0.65	0.035(2)	<0.0021
W30F	0.16	1.3	0.84	0.31	0.26	0.68	0.33	0.36	0.00012	1.8	0.054(2)	—
ΔNY11R	0.37	2.8	2.2	0.37	0.05	0.24	0.81	0.76	0.0025	2.0	0.026(3)	<0.0016
ΔNΔCY11R	0.33	3.2	1.8	0.33	0.17	0.37	0.51	0.061	0.0025	1.7	0.050(2)	—
ΔNΔCY11R/L26A	0.46	2.2	2.0	0.47	0.14	0.29	0.55	0.039	0.0019	1.7	0.044(1)	<0.0020

*λ_0 of Eq. **1**.
†Sum of squares of *SI Appendix*, Eq. **S24**, divided by the number of degrees of freedom.
‡λ_1 (sim) and λ_2 (sim) of Eqs. **2** and **3**.
§Ref. 15.
¶λ_1 (exp) and λ_2 (exp) are the same as k_1 and k_2 in ref. 15.
#SDs are in parentheses.

the WT FBP28 WW domain (Fig. 1 *A* and *B*) and its two mutants, Y11R (Fig. 1 *C* and *D*) and ΔNΔCY11R/L26A (Fig. 1 *E* and *F*), are shown (similar plots for the rest of the mutants are illustrated in *SI Appendix*, Fig. S3). The fitting results, illustrated in Fig. 1 and *SI Appendix*, Fig. S3, and χ^2 values in Table 1 indicate that all systems, except for the Y19L and W30F mutants (*SI Appendix*, Fig. S3 *A*–*D*), exhibit double-exponential kinetics, i.e., are three-state folders. The single-exponential kinetics of W30F are in agreement with experimental results (15), whereas the single- and double-exponential kinetics of Y19L (*SI Appendix*, Fig. S3 *A* and *B*) and ΔNΔCY11R (*SI Appendix*, Fig. S3 *G* and *H*) mutants, respectively, differ from experimental results, in which Y19L and ΔNΔCY11R mutants exhibit double- and single-exponential kinetics, respectively (15).

For the three-state folding systems, both macroscopic rate constants λ_1 and λ_2, in Table 1, have meaningful (positive) values, and the intermediate state is present in remarkable quantity even at the end of the simulations for all molecules, sometimes in an amount comparable to or greater than that of the native state, as illustrated for ΔNΔCY11R/L26A, ΔNY11R, and ΔNΔCY11R in Fig. 1*F* and *SI Appendix*, Fig. S3 *F* and *H*, respectively. Table 1 also includes the experimental rate constants, determined by Nguyen et al. (15). The rate constants determined by fitting to simulation results are about three orders of magnitude greater compared with experimental values (except for ΔNY11R), which is consistent with the fact that the timescale of UNRES is extended by about three orders of magnitude because of averaging out the fast motions of the secondary degrees of freedom (19) and scaling down water friction in our Langevin dynamics simulations by a factor of 1,000.

The values of the fast-phase rate constants (λ_1) determined by simulations correlate well with their experimental counterparts except for the truncated ΔNY11R, ΔNΔCY11R, and ΔNΔCY11R/L26A mutants, which are clear outliers (Fig. 2). Unfortunately, the experimental values of the slow-phase rate constants (λ_2) are not accurate enough to make the similar meaningful comparison with their simulated counterparts.

The preexponential parameter m_1 in Eq. **2** determines the percentage of pathways following the fast-phase (single-exponential) folding kinetics; consequently, the parameter $(1 - m_1)$ indicates the percentage of pathways following the slow-phase (double-exponential) folding kinetics. Indeed, the results for the m_1 parameter, shown in Table 1, are in agreement with the fitting results. For example, the m_1 parameter for WT and Y19L is 0.11 and 0.79, respectively; this indicates that 11% and 79% of the pathways of WT and Y19L, respectively, follow the fast-phase route and 89% and 21% of the pathways of WT and Y19L, respectively, follow the slow-phase route. A disagreement between the fitting results and the m_1 parameter occurs only for the W30F mutant. A plausible explanation of this discrepancy is given below in *Discussion*.

Free-Energy Landscapes of Wild-Type FBP28 WW Domain and Its Mutants. To obtain more insights into the folding kinetics of these proteins and explain the causes of the discrepancies between the experimental and computational results shown above (Figs. 1 and 2 and *SI Appendix*, Fig. S3), we analyzed the distribution of the conformational states in terms of free-energy landscapes along the C^α-rmsd from the native structure and the radius of gyration (R_g) as order parameters in three time intervals: initial, for which the fraction of the native structures is below 20% of the maximum (equilibrium) fraction; intermediate, for which the fraction of the native structures is from 20% to 50% of the maximum fraction of the native structures; and final, for which the fraction of the native structures exceeds 50% of the maximum value. The FELs along C^α-rmsd and R_g [$\mu(\mathrm{rmsd}, R_g) = -k_B T \ln P(\mathrm{rmsd}, R_g)$, where P, T, and k_B are the probability distribution function (pdf), the absolute temperature, and the Boltzmann constant, respectively] for the WT FBP28 WW domain and its Y11R and ΔNΔCY11R/L26A mutants are shown in Figs. 3–5, respectively. The FELs for the rest of the mutants are illustrated in *SI Appendix*, Figs. S4–S7.

As was expected, and can be deduced easily from the fraction curves of Fig. 1 and *SI Appendix*, Fig. S3, the FELs of WT and all full-size and truncated mutants exhibit three-state folding kinetics; however, there are some discrepancies in the folding pathways and in the depths of the states. For example, mutants Y11R and Y19L fold along two different folding pathways (Fig. 4 and *SI Appendix*, Fig. S4), whereas WT and the rest of the mutants (W30F, ΔNY11R, ΔNΔCY11R, and ΔNΔCY11R/L26A) follow a single folding pathway (Figs. 3 and 5 and *SI Appendix*, Figs. S5–S7). Both Y11R and Y19L fold either along the pathway, in which hairpin 1 forms first in the intermediate state and then the protein jumps to the native state, or with a different order of formation of hairpins; i.e., hairpin 2 forms first in the intermediate state before the protein reaches the native state. These findings are in agreement with our previous study (27). The replacement of a nonpolar aromatic amino acid (Tyr) by a charged extremely hydrophilic amino acid (Arg) for the Y11R mutant and by a very nonpolar branched aliphatic amino acid (Leu) for the Y19L mutant, both in hairpin 1, may destabilize this hairpin and induce the second folding pathway, as was pointed out in our earlier work (27) (see the spherical representation of the mutated residue in the representative structures of the intermediate state in Fig. 4 and *SI Appendix*, Fig. S4). The WT and the rest of the mutants fold along a pathway, in which hairpin 1 forms first in the intermediate state before the protein reaches the native state (Figs. 3 and 5 and *SI Appendix*, Figs. S5–S7). These findings are in agreement with previous experimental (15, 30, 32, 33, 36) and theoretical (18, 24, 27, 28) studies. However, by performing unfolding MD simulations of the FBP28 WW domain, Petrovich et al. (32) found that, in 30% of the trajectories, the nonnative helical structure is formed, instead of the first β-strand, allowing the second hairpin to form first. We also observed the nonnative helical structure formed in

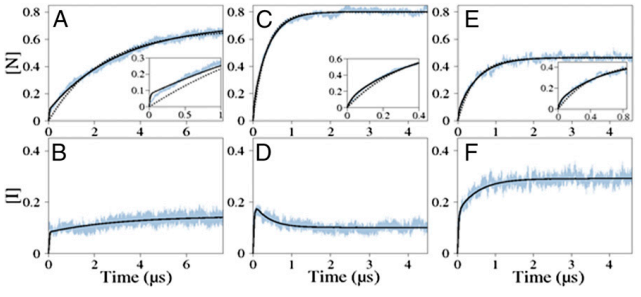


Fig. 1. Fractions of native (*A*, *C*, and *E*) and intermediate (*B*, *D*, and *F*) structures as functions of time for the wild-type FBP 28 WW domain (*A* and *B*) and its Y11R (*C* and *D*) and ΔNΔCY11R/L26A (*E* and *F*) mutants. Light blue ragged lines present the simulation data; dotted lines correspond to the fits by single-exponential kinetics (Eq. **1**, fractions of only native structures); solid lines correspond to the fits by double-exponential kinetics (Eqs. **2** and **3**, fractions of the native and intermediate structures, respectively). *Insets* represent the enlarged values of fractions of native structures for the first several hundred nanoseconds of time.

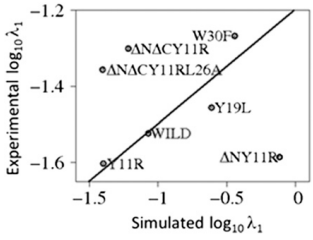


Fig. 2. A correlation plot of the experimental (15) and simulated fast cumulative constants (λ_1 of Table 1) with a least-squares fitting line.

the N-terminal region in the unfolded state for WT; however, it did not help the second hairpin to form first. Moreover, Petrovich et al. (32) explained the formation of the first hairpin in the remaining 70% of the trajectories; in particular, they found that the first loop is present in the early stage as a kink in the backbone of the protein to allow long-range interactions to occur, followed by side-chain interactions and hydrophobic collapse of residues to enable the first and second β-strands to form the first hairpin. This finding is in agreement with our recent study (28).

The folding scenarios illustrated in the FELs for Y19L and W30F mutants (*SI Appendix*, Figs. S4 and S5) contradict our findings obtained by fitting the fractions of native and intermediate structures (*SI Appendix*, Fig. S3 *A–D*). In particular, the FELs of Y19L and W30F mutants clearly show a three-state folding scenario, whereas the fitting results in *SI Appendix*, Fig. S3 indicate single-exponential kinetics, which normally implies a two-state folding scenario.

In the end, it should be noted that, based on the results shown in *SI Appendix*, Figs. S3 *A* and *B* and S4, the replacement of Tyr-19 by Leu had the most destabilizing influence among all mutations, which indicates the importance of the Tyr-19 residue for the stability of the protein. These results are in agreement with the findings of an earlier experimental study (16), in which the structural roles of the conserved residues were studied by using site-directed mutagenesis of the FBP28 WW domain.

Discussion

To account for the discrepancies between the results obtained from the kinetic studies and FELs, we drew the free-energy diagrams (*SI Appendix*, Fig. S8) based on the distributions of the populations of the native, intermediate, and unfolded structures (P_N, P_I, P_U) for all of the systems studied.

The main differences, illustrated in these free-energy diagrams, between the mutants exhibiting single- and double-exponential kinetics are (*i*) in the depth of the intermediate state and (*ii*) in the barrier heights of the unfolded/intermediate and intermediate/native states (*SI Appendix*, Fig. S8). In particular, the intermediate state is the deepest and the barrier height between the unfolded and intermediate states is much lower than the one between the intermediate and native states [ratios (R) between the barrier heights for all systems are in *SI Appendix*, Fig. S9] for the mutants exhibiting single-exponential kinetics (*SI Appendix*, Fig. S8 *D* and *E*); whereas the systems exhibiting double-exponential kinetics have the deepest native state, and the barrier height between the unfolded and intermediate states is higher than (or comparable to) that between the intermediate and native states (*SI Appendix*, Figs. S8 *A–C* and *F* and S9). This explains why single-exponential kinetics emerge during three-state folding. The point is that the barrier height between the intermediate and native states is so high compared with that between the unfolded and intermediate states [three (for Y19L) or four (for W30F) times higher than that between the unfolded and intermediate states] that the timescale separation occurs and only the slowest step is observed by experiment due to experimental limitations; hence, single-exponential kinetics arise. Among the studied mutants, ΔNΔCY11R is the only exception, which exhibits double-exponential kinetics from the fitting, but the features (depth of the intermediate state, barrier heights) illustrated in the free-energy diagram (*SI Appendix*, Fig. S8*G*) are characteristic of single-exponential kinetics, which actually were observed in the experiment (15) for this mutant. The reason for the double-exponential kinetics for the ΔNΔCY11R mutant is that the difference between the barrier heights of the unfolded/intermediate (0.53 k_BT) and intermediate/native (0.60 k_BT) states is not large enough to lead to single-exponential kinetics. However, the difference between the values of χ^2 (0.018 and 0.017, respectively) (Table 1) of the single- and double-exponential fits for the ΔNΔCY11R mutant is the smallest among all mutants, which indicates that there is a tiny threshold between single- and double-exponential kinetics for this mutant.

In a perceptive analysis of the timescales for protein folding kinetics, Thirumalai (38) showed that, at the atomic level, the free-energy barrier height scales as $N^{1/2}$, where N is the number of residues; i.e., the free-energy barrier heights for the FBP28 WW domain and its full-size ($n = 37$) and truncated ($n = 32$ and $n = 28$) mutants should vary between 5.3 k_BT and 6 k_BT. However, because of averaging out the fast motions of the secondary degrees of freedom, at the coarse-grained level, the free-energy barriers, illustrated in *SI Appendix*, Fig. S8, are lower than those at the atomic level.

Thus, we have illustrated that single-exponential kinetics do not arise only in two-state folding. They may emerge during

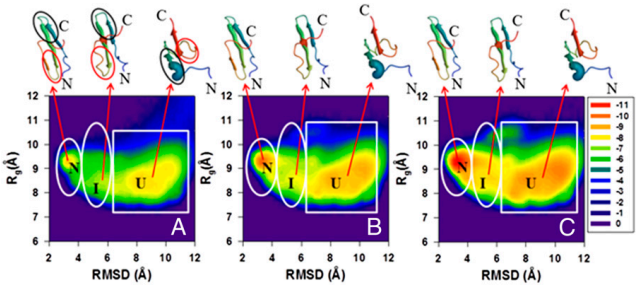


Fig. 3. Variation of the distribution of conformational states in terms of FELs (in kcal/mol) along the C^{α}-rmsd and R_{gyr} order parameters for the wild-type FBP28 WW domain. The data have been collected from different sections of all 512 trajectory sets for the molecule (shown in *A–C*, respectively). The FEL corresponding to the initial parts of the trajectories (with the average fraction of the native structures up to 20% of the maximum fraction) is shown in *A*; the FEL from the middle parts of the trajectories (the fraction of the native structures between 20% and 50% of the maximum fraction) is shown in *B*; and the FEL from the final parts of the trajectories (the fraction of the native structures exceeds 50% of the maximum fraction) is shown in *C*, respectively. The letters "U," "I," and "N" correspond to unfolded, intermediate, and native states, respectively. The representative structures of unfolded, intermediate, and native states are plotted on top of each state. Hairpin 1 and hairpin 2 are circled by black and red lines, correspondingly, in *A*.

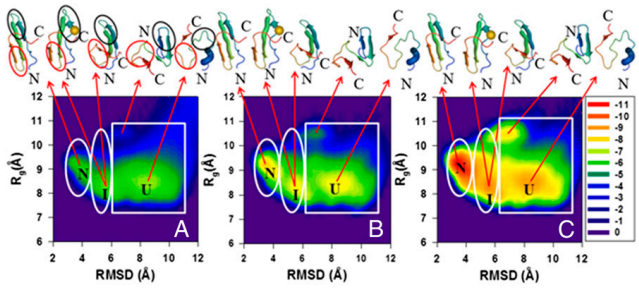


Fig. 4. (*A–C*) Same as in Fig. 3 but for the Y11R full-size mutant.

three-state (or multistate) folding when one of the free-energy barriers is much higher than the other; consequently, a separation of timescales occurs and single-exponential kinetics arise.

To explain the discrepancy between the fitting results (*SI Appendix*, Fig. S3 *C* and *D*) and the m_1 parameter for the W30F mutant, we have calculated the changes in the populations of the intermediate state over the above-defined time intervals (shown in Fig. 3 legend). It turns out that, unlike in the other mutants (in which P_I increases gradually over time), P_I increases rapidly (~28%) in the initial part of the trajectories, in W30F, then decreases slightly (~25%) in the middle part of the trajectories, and then increases again (~38%). This has been reflected in a bit of unusual behavior of the fitted curve of the fractions of the intermediate structures (light blue ragged line in *SI Appendix*, Fig. S3*D*); consequently, the m_1 parameter obtained from this fitting differs from the expected one (>0.5).

It should be noted that, for a similar reason, i.e., a rapid increase of P_I in the initial part of the trajectories, the fast-phase rate constant of the ΔNY11R mutant increased by one order of magnitude.

In the end, the truncated mutants that are outliers in Fig. 2 might be caused by the unusual states with structures similar to the native states, but with the second hairpin shifted (Fig. 5 and *SI Appendix*, Figs. S6 and S7), which suggests that these states might derail folding to misfolded states (20, 39). One of the reasons for the "distortion" of structures of the ΔNΔCY11R and ΔNΔCY11R/L26A mutants could be the deletion of the C-terminal Leu36 residue, part of a delocalized hydrophobic core, Trp8/Tyr20/Pro33. As for the ΔNY11R mutant, based on earlier experimental studies (15, 16), the truncation of the N-terminal residues has no observable effect on the stability of the domain; hence, the reasons for the distortion of the structure must be different. An investigation of the reasons for this distortion is beyond the scope of the present study but it is worth pursuing in the future.

Conclusions

In this study, we carried out a quantitative analysis of the simulated kinetics of the folding of the FBP28 WW domain and its six mutants at the coarse-grained level. By analysis of the fractions of the native structures, we found double-exponential kinetics (three-state folding) for the WT and its mutants except for the Y19L and W30F mutants, which exhibited single-exponential kinetics implying two-state folding. The results from the FELs along the C^{α}-rmsd and R_g indicate that the WT and all its mutants are three-state folders. For most of the mutants, the obtained results are in agreement with experiment (15). The discrepancies between the results of our simulation and the experimental (15) kinetics studies and the FELs for some mutants, as well as the origins of single- and double-exponential kinetics and their correlations with two- and three-state folding, are explained in terms of the free-energy barrier heights. In particular, we have shown that single-exponential kinetics can emerge even in three-state (or multistate) folding when one of the free-energy barriers is much higher than the other. The calculated fast-phase rate constants correlate with the experimental values determined by Nguyen et al. (15) except for the truncated mutants. For most of the systems, the rate constants obtained by simulations are greater by about three orders of magnitude, which is caused by averaging out the fast degrees of freedom in our coarse-grained treatment (19) and by scaling down water friction by a factor of 1,000. It should be noted that scaling down the friction is a common practice in coarse-grained Langevin-dynamics simulations to accelerate the simulations (40, 41).

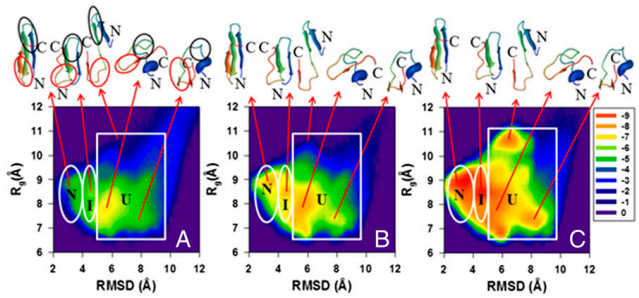


Fig. 5. (*A–C*) Same as in Fig. 3 but for the ΔNΔCY11R/L26A truncated mutant.

An important finding of this work is that, for such small proteins as the FBP28 WW domain and its variants, the intermediate and unfolded states are at equilibrium with the native state in water solutions (Fig. 1 and *SI Appendix*, Fig. S3). Therefore, estimating the content of the native state based on, e.g., the CD or fluorescence spectra of a protein solution must be considered with caution because the unfolded protein/intermediate state can be present even at temperatures well below the folding-transition temperature.

As in our previous studies (27, 28), consistent with previous findings (15, 18, 24, 30, 32, 33, 36), we found that the intermediate state consists predominantly of conformations with hairpin 1 well established and hairpin 2 only outlined. The Y11R and Y19L mutants are exceptions (Fig. 4 and *SI Appendix*, Fig. S4) because of the destabilization of hairpin 1 caused by the replacement of a hydrophobic tyrosine residue with a charged extremely hydrophilic (arginine) and a very nonpolar branched aliphatic (leucine) residue.

Materials and Methods

To determine the rate constants for protein folding kinetics, we first derived the rate equations for two- and three-state folding models. All steps of derivations are given in *SI Appendix, SI Materials and Methods*. Here, we present only the equations, in final form, for the mole fractions of the native state as a function of time, $[N](t)$, for a two-state folding model,

$$[N](t) = C_0[1 - \exp(-\lambda_0 t)], \quad [1]$$

and for the mole fractions of native and intermediate states as functions of time, $[N](t)$ and $[I](t)$, respectively, for a three-state folding model

$$[N](t) = C_1\{1 - m_1\exp(-\lambda_1 t) - (1 - m_1)\exp(-\lambda_2 t)\} \quad [2]$$

$$[I](t) = C_2\{1 - m_2\exp(-\lambda_1 t) - (1 - m_2)\exp(-\lambda_2 t)\}. \quad [3]$$

To obtain the rate constants, the fractions of the U, I, and N states averaged over 512 MD trajectories at each time interval were fitted by Eq. **1** for a two-state model and by Eqs. **2** and **3** for a three-state model. In other words, the mole fractions of the native state (two-state model) and of the native and intermediate states (three-state model) calculated from simulation data were fitted by Eq. **1** (with C_0 and λ_0 as determinable parameters) and Eqs. **2** and **3** (with C_1, C_2, m_1, m_2, λ_1, and λ_2 as determinable parameters) (details of all parameters, the simulation, and data analysis are in *SI Appendix*).

Details of the preparation of FBP28 WW mutants and of the calculated structures are given in *SI Appendix*.

ACKNOWLEDGMENTS. 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*) our 96-processor cluster at the Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology; and (*iv*) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center. This work was supported by grants from the National Institutes of Health (GM-14312) and the National Science Foundation (MCB10-19767), by Grant 530-8370-D498-14 from the Polish Ministry of Science and Education, by Grant 31300598 from the National Science Foundation of China, and by the Spanish National Research Program [Ministry of Economy and Competitiveness, SAF2011-25119 (to M.J.M.)]. D.S. has a "laCaixa"/Institute for Research in Biomedicine Barcelona International PhD Programme fellowship, and T.T. is a recipient of a European Union Co-funding of Regional, National, and International Programmes - Marie Curie Actions grant. M.J.M. is a Catalan Institution for Research and Advanced Studies Programme Investigator.

1. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517–520.
2. Skolnick J, et al. (2003) TOUCHSTONE: A unified approach to protein structure prediction. *Proteins* 53(Suppl 6):469–479.
3. Scheraga HA, et al. (2004) The protein folding problem: Global optimization of the force fields. *Front Biosci* 9:3296–3323.
4. Tozzini V (2005) Coarse-grained models for proteins. *Curr Opin Struct Biol* 15(2):144–150.
5. Thorpe IF, Zhou J, Voth GA (2008) Peptide folding using multiscale coarse-grained models. *J Phys Chem B* 112(41):13079–13090.
6. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1993) Prediction of protein conformation on the basis of a search for compact structures: Test on avian pancreatic polypeptide. *Protein Sci* 2(10):1715–1731.
7. Liwo A, et al. (1997) A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem* 18:849–873.
8. Liwo A, et al. (1997) A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J Comput Chem* 18:874–887.
9. 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* 115:2323–2347.
10. Liwo A, Oldziej S, Czaplewski C, Kozlowska U, Scheraga HA (2004) Parametrization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. *J Phys Chem B* 108:9421–9438.
11. Oldziej S, Liwo A, Czaplewski C, Pillardy J, Scheraga HA (2004) Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 2. Off-lattice tests of the method with single proteins. *J Phys Chem B* 108:16934–16949.
12. Oldziej S, et al. (2004) Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. *J Phys Chem B* 108:16950–16959.
13. Liwo A, et al. (2007) Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J Phys Chem B* 111(1):260–285.
14. 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.
15. 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.
16. 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.
17. Sudol M, Hunter T (2000) NeW wrinkles for an old domain. *Cell* 103(7):1001–1004.
18. 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.
19. 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 USA* 102(7):2362–2367.
20. Mu Y, Nordenskiöld L, Tam JP (2006) Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. *Biophys J* 90(11):3983–3992.
21. Maisuradze GG, Liwo A, Scheraga HA (2009) Principal component analysis for protein folding dynamics. *J Mol Biol* 385(1):312–329.
22. Maisuradze GG, Liwo A, Scheraga HA (2010) Relation between free energy landscapes of proteins and dynamics. *J Chem Theory Comput* 6(2):583–595.
23. Shaw DE, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330(6002):341–346.
24. Piana S, et al. (2011) Computational design and experimental testing of the fastest-folding β-sheet protein. *J Mol Biol* 405(1):43–48.
25. 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.
26. A Beccara S, Škrbić T, Covino R, Faccioli P (2012) Dominant folding pathways of a WW domain. *Proc Natl Acad Sci USA* 109(7):2330–2335.
27. 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.
28. Maisuradze GG, Liwo A, Senet P, Scheraga HA (2013) Local vs global motions in protein folding. *J Chem Theory Comput* 9(7):2907–2921.
29. 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(2):373–393.
30. 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.
31. 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.
32. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR (2006) Phi-analysis at the experimental limits: Mechanism of beta-hairpin formation. *J Mol Biol* 360(4):865–881.
33. Jäger M, et al. (2006) Structure-function-folding relationship in a WW domain. *Proc Natl Acad Sci USA* 103(28):10648–10653.
34. Sharpe T, Jonsson AL, Rutherford TJ, Daggett V, Fersht AR (2007) The role of the turn in beta-hairpin formation during WW domain folding. *Protein Sci* 16(10):2233–2239.
35. Serpell LC (2000) Alzheimer's amyloid fibrils: Structure and assembly. *Biochim Biophys Acta* 1502(1):16–30.
36. Davis CM, Dyer RB (2014) WW domain folding complexity revealed by infrared spectroscopy. *Biochemistry* 53(34):5476–5484.
37. Cheung MS, Thirumalai D (2007) Effects of crowding and confinement on the structures of the transition state ensemble in proteins. *J Phys Chem B* 111(28):8250–8257.
38. Thirumalai D (1995) From minimal models to real proteins: Time scales for protein folding kinetics. *J Phys I* 5:1457–1467.
39. Neudecker P, et al. (2012) Structure of an intermediate state in protein folding and aggregation. *Science* 336(6079):362–366.
40. Veitshans T, Klimov D, Thirumalai D (1997) Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold Des* 2(1):1–22.
41. Cieplak M, Hoang TX, Robbins MO (2002) Thermal folding and mechanical unfolding pathways of protein secondary structures. *Proteins* 49(1):104–113.

8 Abbreviations and Units

Å	Ångström
ACH	Alpha-cyano-4-hydroxycinnamic acid
Acm	Acetamidomethyl
APL	Acute promyelocytic leukemia
ARE	Activin response element
ARIA	Ambiguous Restraints for Iterative Assignment
AU	Arbitrary unit
BMP	Bone morphogenetic protein
BMRB	Biological Magnetic Resonance Data Bank
Boc	*tert*-Butoxycarbonyl
BOP	Benzotriazol-1-yloxytris(dimethylamino)phosphonium hexafluorophosphate
BSA	Bovine serum albumin
Bzl	Benzyl
°C	Celsius
cam-ester	Carboxamidomethyl ester
CBP	Cyclic AMP response element-binding protein
cDNA	Complementary DNA
CM	ChemMatrix®
CNS	Crystallography & NMR System
Co-SMAD	Common partner SMAD
cps	Counts per second
CRBPII	Retinol-binding protein II
CSI	Chemical shift index
CSP	Chemical shift perturbation
CtBP	C-terminal-binding protein 1
δ	Chemical shift
Da	Dalton (g/mol)
DBU	1,8-Diazabicyclo[5.4.0]undec-7-ene
DCM	Dichloromethane

DIC	*N*,*N*'-Diisopropylcarbodiimide
DIPEA	*N*,*N*-Diisopropylethylamine
DMF	*N*,*N*-Dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
ds DNA	double-stranded DNA
DTT	Dithiothreitol
E.coli	*Escherichia coli*
EDT	Ethanedithiol
EDTA	Ethylenediaminetetraacetic acid
EGF	Epidermal growth factor
EMSA	Electrophoretic mobility shift assays
EPL	Expressed Protein Ligation
eq	Equivalents
ESI	Electrospray ionization
EVI1	Ecotropic virus integration site 1 protein homolog
ExPASy	Expert protein analysis system
FBXW7	F- box/WD repeat-containing protein 7
FID	Free induction decay
FM	Fast/FoxH1 motif
Fmoc	9-Fluorenylmethoxycarbonyl
FPLC	Fast protein liquid chromatograph
FT	Fourier transform
GDF	Growth and differentiation factors
GST	Glutathion S-transferase
GuHCl	Guanidinium chloride
h	Hour
HAT	Histone acetyltransferase
HATU	*N*-[(dimethylamino)-1*H*-1,2,3-triazolo[4,5-*b*]pyridin-1-ylmethylene]-*N*-methylmethanaminium hexafluorophosphate *N*-oxide
HBTU	*N*-[(1*H*-benzotriazol-1-yl)(dimethylammo)methylene]-*N*-methylmethanaminium hexafluorophosphate *N*-oxide
HD	Homeodomain
HDAC1	Histone deacetylase 1

HEPES	2-[4-(2-Hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HFIP	1,1,1,3,3,3-Hexafluoro-2-propanol
HOAt	1-Hydroxy-7-azabenzotriazole
HOBT	1-Hydroxybenzotriazole
HOOBt	3,4-Dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazine
HPE	Holoprosencephaly
HPLC	High Performance/Pressure Liquid Chromatography
HSQC	Heteronuclear Single-Quantum Correlation/Coherence spectroscopy
IMAC	Immobilised metal ion affinity
IPTG	Isopropyl β-D-1-thiogalactopyranoside
I-SMAD	Inhibitory SMAD
ITC	Isothermal titration calorimetry
ivDde	1-(4,4-Dimethyl-2,6-dioxocyclohex-1-ylidene)-3-methylbutyl
K	Kelvin
KAHA	α-ketoacid–hydroxylamine
KCL	Kinetically controlled ligation
K_D	Dissociation constant
LB	Luria broth
LC-MS	Liquid Chromatography Mass Spectrometry
MALDI	Matrix-Assisted Laser Desorption/Ionization
MAPK	Mitogen-activated protein kinases
MCP	Micro-channel plate
MCS	Multiple cloning site
MeCN	Acetonitrile
MESNA	2-Mercapoethanesulfonate
MH	MAD homology
min	Minute
MIS	Muellerian inhibiting substance
MPAA	4-(Carboxymethyl)thiophenol
mRNA	messenger RNA
MS	Mass spectrometry
MST	MicroScale Thermophoresis
NBD	4-Chloro-7-nitrobenzofurazan
NCL	Native chemical ligation

N-CoR	Nuclear receptor corepressor 1
NLS	Nuclear localisation signal
NMP	*N*-Methyl-2-pyrrolidone
NMR	Nuclear magnetic resonance
NOE	Nuclear overhauser effect
NOESY	Nuclear overhauser effect spectroscopy
OD	Optical density (Absorbance)
ON	Overnight
Oxyma Pure	Ethyl 2-cyano-2-(hydroxyimino)acetate
PAI-1	Plasminogen activator inhibitor-1
Pbf	2,2,4,6,7-pentamethyldihydrobenzofuran-5-sulfonyl
PCR	Polymerase chain reaction
PDB	Protein data bank
PEG	Polyethylene glycol
PHRF1	PHD and RING finger domain-containing protein 1
PML	Promyelocytic leukaemia protein
PMSF	phenylmethanesulfonyl fluoride
ppm	Parts per million
PS	Polystyrene
PCTA	PML competitor for TGIF association
PTM	Post-translational modifications
PyBOP®	Benzotriazol-1-yloxy tris(pyrrolidino)phosphonium hexafluorophosphate
RF	Radio frequency
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RP	Reverse Phase
RPM	Revolutions per minute
R-SMAD	Receptor-regulated SMAD
RT	Room temperature
RXRE	Retinoid X receptor responsive element
SAL	Sugar-assisted ligation
SARA	Smad anchor for receptor activation
SBD	SMAD binding domain
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SID	SMAD interacting domain

SIM	SMAD interacting motif
SMAD	Small mothers against decapentaplegic
SNR	Signal-to-noise ratio
SOC	Super optimal broth with catabolite repression
SPPS	Solid-Phase Peptide Synthesis
ss DNA	single-stranded DNA
SSH	Sonic hedgehog protein
TALE	Three amino acid loop extension
TB	Terrific broth
*t*Bu	*tert*-butyl
TCEP	Tris(2-carboxyethyl)phosphine
TEV	Tobacco etch virus
TFA	Trifluoroacetic acid
TGFβ	Transforming growth factor-β
TGFBR	TGFβ receptor
TGIF1	5'-TG-3' interacting factor 1
Thz	1,3-Thiazolidine-4-carboxo
TIPS	Triisopropylsilane
TOCSY	Total correlation spectroscopY
TOF	Time of flight
TRIS	2-Amino-2-(hydroxymethyl)propane-1,3-diol
UNRES	United-residue
VA-044	2,20-azobis[2-(2-imidazolin-2-yl)propane]dihydrochloride
vol.	Volume
WT	Wild type

Bibliography

[1] J. Massagué, "TGFβ signalling in context.," *Nature reviews. Molecular cell biology*, vol. 13, pp. 616–30, oct 2012.

[2] L. Huminiecki, L. Goldovsky, S. Freilich, A. Moustakas, C. Ouzounis, and C.-H. Heldin, "Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom.," *BMC evolutionary biology*, vol. 9, p. 28, 2009.

[3] Y. Shi and J. Massagué, "Mechanisms of TGF-beta signaling from cell membrane to the nucleus.," *Cell*, vol. 113, pp. 685–700, jun 2003.

[4] Y. Kang, C. R. Chen, and J. Massagué, "A self-enabling TGFβ response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells," *Molecular Cell*, vol. 11, no. 4, pp. 915–926, 2003.

[5] D. Padua, X. H. F. Zhang, Q. Wang, C. Nadal, W. L. Gerald, R. R. Gomis, and J. Massagué, "TGFβ Primes Breast Tumors for Lung Metastasis Seeding through Angiopoietin-like 4," *Cell*, vol. 133, no. 1, pp. 66–77, 2008.

[6] P. Kavsak, R. K. Rasmussen, C. G. Causing, S. Bonni, H. Zhu, G. H. Thomsen, and J. L. Wrana, "Smad7 binds to Smurf2 to form an E3 ubiquitin ligase that targets the TGFβ receptor for degradation," *Molecular Cell*, vol. 6, no. 6, pp. 1365–1375, 2000.

[7] J. Seoane, H.-V. Le, L. Shen, S. A. Anderson, and J. Massagué, "Metabolic rate depression. The biochemistry of mammalian hibernation," *Cell*, vol. 117, pp. 211–223, 2004.

[8] N. Nakano, S. Itoh, Y. Watanabe, K. Maeyama, F. Itoh, and M. Kato, "Requirement of TCF7L2 for TGF-β-dependent transcriptional activation of the TMEPAI gene," *Journal of Biological Chemistry*, vol. 285, no. 49, pp. 38023–38033, 2010.

[9] M. Kretzschmar, J. Doody, and J. Massagué, "Opposing BMP and EGF signalling pathways converge on the TGF-B family mediator Smad1," *Nature*, vol. 389:, no. 6651, pp. 618–22., 1997.

[10] Y. Shi, Y. F. Wang, L. Jayaraman, H. Yang, J. Massagué, and N. P. Pavletich, "Crystal structure of a Smad MH1 domain bound to DNA: Insights on DNA binding in TGF-beta signaling," *Cell*, vol. 94, no. 5, pp. 585–594, 1998.

[11] O. Korchynskyi and P. Ten Dijke, "Identification and functional characterization of distinct critically important bone morphogenetic protein-specific response elements in the Id1 promoter," *Journal of Biological Chemistry*, vol. 277, no. 7, pp. 4883–4891, 2002.

[12] G. Wu, Y.-G. Chen, B. Ozdamar, C. A. Gyuricza, P. A. Chong, J. L. Wrana, J. Massagué, and Y. Shi, "Structural Basis of Smad2 Recognition by the Smad Anchor for Receptor Activation," *Science*, vol. 287, pp. 92–97, jan 2000.

[13] C. Alarcón, A. I. Zaromytidou, Q. Xi, S. Gao, J. Yu, S. Fujisawa, A. Barlas, A. N. Miller, K. Manova-Todorova, M. J. Macias, G. Sapkota, D. Pan, and J. Massagué, "Nuclear CDKs Drive Smad Transcriptional Activation and Turnover in BMP and TGF-beta Pathways," *Cell*, vol. 139, no. 4, pp. 757–769, 2009.

[14] E. Aragón, N. Goerner, A.-I. Zaromytidou, Q. Xi, A. Escobedo, J. Massagué, and M. J. Macias, "A Smad action turnover switch operated by WW domain readers of a phosphoserine code.," *Genes & development*, vol. 25, pp. 1275–88, jun 2011.

[15] I. M. Wallace, O. O'Sullivan, D. G. Higgins, and C. Notredame, "M-Coffee: Combining multiple sequence alignment methods with T-Coffee," *Nucleic Acids Research*, vol. 34, no. 6, pp. 1692–1699, 2006.

[16] T. Tsukazaki, T. A. Chiang, A. F. Davison, L. Attisano, and J. L. Wrana, "SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor," *Cell*, vol. 95, no. 6, pp. 779–791, 1998.

[17] S. Zhou, L. Zawel, C. Lengauer, K. W. Kinzler, and B. Vogelstein, "Characterization of human FAST-1, a TGF beta and activin signal transducer.," *Molecular cell*, vol. 2, no. 1, pp. 121–127, 1998.

[18] Y. G. Chen, a. Hata, R. S. Lo, D. Wotton, Y. Shi, N. Pavletich, and J. Massagué, "Determinants of specificity in TGF-beta signal transduction.," *Genes & development*, vol. 12, no. 14, pp. 2144–2152, 1998.

[19] S. Germain, M. Howell, G. M. Esslemont, and C. S. Hill, "Homeodomain and winged-helix transcription factors recruit activated Smads to distinct promoter elements via a common Smad interaction motif," *Genes and Development*, vol. 14, no. 4, pp. 435–451, 2000.

[20] J. Massagué, J. Seoane, and D. Wotton, "Smad transcription factors," *Genes & development*, vol. 19, pp. 2783–2810, 2005.

[21] R. A. Randall, M. Howell, C. S. Page, A. Daly, P. A. Bates, and C. S. Hill, "Recognition of phosphorylated-Smad2-containing complexes by a novel Smad interaction motif.," *Molecular and cellular biology*, vol. 24, no. 3, pp. 1106–21, 2004.

[22] S. Ross, E. Cheung, T. G. Petrakis, M. Howell, W. L. Kraus, and C. S. Hill, "Smads orchestrate specific histone modifications and chromatin remodeling to activate transcription.," *The EMBO journal*, vol. 25, no. 19, pp. 4490–502, 2006.

[23] M. Simonsson, M. Kanduri, E. Grönroos, C. H. Heldin, and J. Ericsson, "The DNA binding activities of Smad2 and Smad3 are regulated by coactivator-mediated acetylation," *Journal of Biological Chemistry*, vol. 281, no. 52, pp. 39870–39880, 2006.

[24] J. W. Wu, A. R. Krawitz, J. Chai, W. Li, F. Zhang, K. Luo, and Y. Shi, "Structural mechanism of Smad4 recognition by the nuclear oncoprotein Ski: Insights on Ski-mediated repression of TGF-beta signaling," *Cell*, vol. 111, no. 3, pp. 357–367, 2002.

[25] K. Luo, S. L. Stroschein, W. Wang, D. Chen, E. Martens, S. Zhou, and Q. Zhou, "The Ski oncoprotein interacts with the Smad proteins to repress TGFbeta signaling," *Genes and Development*, vol. 13, no. 17, pp. 2196–2206, 1999.

[26] E. Bertolino, B. Reimund, D. Wildt-Perinic, and R. G. Clerc, "A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif.," *The Journal of biological chemistry*, vol. 270, pp. 31178–88, dec 1995.

[27] T. R. Bürglin and M. Affolter, "Homeodomain proteins: an update," *Chromosoma*, no. 458, pp. 497–521, 2016.

[28] D. Wotton, R. S. Lo, S. Lee, and J. Massagué, "A Smad transcriptional corepressor.," *Cell*, vol. 97, pp. 29–39, apr 1999.

[29] K. W. Gripp, D. Wotton, M. C. Edwards, E. Roessler, L. Ades, P. Meinecke, a. Richieri-Costa, E. H. Zackai, J. Massagué, M. Muenke, and S. J. Elledge, "Mutations in TGIF cause holoprosencephaly and link NODAL signalling to human neural axis determination.," *Nature genetics*, vol. 25, pp. 205–8, jun 2000.

[30] J. E. Ming and M. Muenke, "Holoprosencephaly: from Homer to Hedgehog.," *Clin. Genet.*, vol. 53, no. 3, pp. 155–163, 1998.

[31] M. C. Edwards, N. Liegeois, J. Horecka, and S. Ronald, "Human CPR (Cell Cycle Progression Restoration) genes impart a far phenotype on yeast cells," 1997.

[32] E. Bertolino, S. Wildt, G. Richards, and R. G. Clerc, "Expression of a novel murine homeobox gene in the developing cerebellar external granular layer during its proliferation," *Developmental Dynamics*, vol. 205, no. 4, pp. 410–420, 1996.

[33] C. Aguilella, C. Dubourg, J. Attia-Sobol, J. Vigneron, M. Blayau, L. Pasquier, L. Lazaro, S. Odent, and V. David, "Molecular screening of the TGIF gene in holoprosencephaly: identification of two novel mutations.," *Human genetics*, vol. 112, pp. 131–4, feb 2003.

[34] K. Taniguchi, A. E. Anderson, A. E. Sutherland, and D. Wotton, "Loss of Tgif function causes holoprosencephaly by disrupting the SHH signaling pathway.," *PLoS genetics*, vol. 8, p. e1002524, jan 2012.

[35] R. Hamid and S. J. Brandt, "Transforming growth-interacting factor (TGIF) regulates proliferation and differentiation of human myeloid leukemia cells.," *Molecular oncology*, vol. 3, pp. 451–63, dec 2009.

[36] R. S. Lo, D. Wotton, and J. Massagué, "Epidermal growth factor signaling via Ras controls the Smad transcriptional co-repressor TGIF.," *The EMBO journal*, vol. 20, pp. 128–36, jan 2001.

[37] M. T. Bengoechea-Alonso and J. Ericsson, "Tumor suppressor Fbxw7 regulates TGFβ signaling by targeting TGIF1 for degradation.," *Oncogene*, vol. 29, pp. 5322–8, sep 2010.

[38] A. Ettahar, O. Ferrigno, M.-Z. Zhang, M. Ohnishi, N. Ferrand, C. Prunier, L. Levy, M.-F. Bourgeade, I. Bieche, D. G. Romero, F. Colland, and A. Atfi, "Identification of PHRF1 as a Tumor Suppressor that Promotes the TGF-β Cytostatic Program through Selective Release of TGIF-Driven PML Inactivation.," *Cell reports*, vol. 4, pp. 530–41, aug 2013.

[39] I. Imoto, a. Pimkhaokham, T. Watanabe, F. Saito-Ohara, E. Soeda, and J. Inazawa, "Amplification and overexpression of TGIF2, a novel homeobox gene of the TALE superclass, in ovarian cancer cell lines.," *Biochemical and biophysical research communications*, vol. 276, no. 1, pp. 264–70, 2000.

[40] T. A. Melhuish, C. M. Gallo, and D. Wotton, "TGIF2 Interacts with Histone Deacetylase I and Represses Transcription," *Journal of Biological Chemistry*, vol. 276, no. 34, pp. 32109–32114, 2001.

[41] D. E. Piper, a. H. Batchelor, C. P. Chang, M. L. Cleary, and C. Wolberger, "Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation.," *Cell*, vol. 96, no. 4, pp. 587–597, 1999.

[42] A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale, "DNA-dependent formation of transcription factor pairs alters their binding specificity," *Nature*, vol. 527, no. 7578, pp. 384–388, 2015.

[43] L. P. Kozlowski and J. M. Bujnicki, "MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins," *BMC Bioinformatics*, vol. 13, no. 1, p. 111, 2012.

[44] D. Wotton, R. S. Lo, L. a. Swaby, and J. Massagué, "Multiple modes of repression by the Smad transcriptional corepressor TGIF.," *The Journal of biological chemistry*, vol. 274, pp. 37105–10, dec 1999.

[45] T. a. Melhuish and D. Wotton, "The interaction of the carboxyl terminus-binding protein with the Smad corepressor TGIF is disrupted by a holoprosencephaly mutation in TGIF.," *The Journal of biological chemistry*, vol. 275, pp. 39762–6, dec 2000.

[46] D. Wotton, P. S. Knoepfler, C. D. Laherty, R. N. Eisenman, and J. Massagué, "The Smad transcriptional corepressor TGIF recruits mSin3.," *Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research*, vol. 12, pp. 457–63, sep 2001.

[47] M. Pessah, C. Prunier, J. Marais, N. Ferrand, a. Mazars, F. Lallemand, J. M. Gauthier, and a. Atfi, "c-Jun interacts with the corepressor TG-interacting factor (TGIF) to suppress Smad2 transcriptional activity.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 6198–203, may 2001.

[48] S. R. Seo, N. Ferrand, N. Faresse, C. Prunier, L. Abécassis, M. Pessah, M.-F. Bourgeade, and A. Atfi, "Nuclear retention of the tumor suppressor cPML by the homeodomain protein TGIF restricts TGF-beta signaling.," *Molecular cell*, vol. 23, pp. 547–59, aug 2006.

[49] H.-K. Lin, S. Bergmann, and P. P. Pandolfi, "Cytoplasmatic PML function in TGF-beta signalling," *Nature*, vol. 431, no. September, pp. 205–211, 2004.

[50] N. Faresse, F. Colland, N. Ferrand, C. Prunier, M.-F. Bourgeade, and A. Atfi, "Identification of PCTA, a TGIF antagonist that promotes PML function in TGF-beta signalling.," *The EMBO journal*, vol. 27, pp. 1804–15, jul 2008.

[51] M.-Z. Zhang, O. Ferrigno, Z. Wang, M. Ohnishi, C. Prunier, L. Levy, M. Razzaque, W. C. Horne, D. Romero, G. Tzivion, F. Colland, R. Baron, and A. Atfi, "TGIF Governs a Feed-Forward Network that Empowers Wnt Signaling to Drive Mammary Tumorigenesis," *Cancer Cell*, vol. 27, no. 4, pp. 547–560, 2015.

[52] R. Nusse, "Wnt signaling in disease and in development.," *Cell research*, vol. 15, no. 1, pp. 28–32, 2005.

[53] M. S. Razzaque and A. Atfi, "TGIF function in oncogenic Wnt signaling," *Biochimica et Biophysica Acta - Reviews on Cancer*, vol. 1865, no. 2, pp. 101–104, 2016.

[54] M. Hneino, K. Blirando, V. Buard, G. Tarlet, M. Benderitter, P. Hoodless, A. François, and F. Milliat, "The TG-interacting factor TGIF1 regulates stress-induced proinflammatory phenotype of endothelial cells," *Journal of Biological Chemistry*, vol. 287, no. 46, pp. 38913–38921, 2012.

[55] T. A. Melhuish, K. Taniguchi, and D. Wotton, "Tgif1 and Tgif2 Regulate Axial Patterning in Mouse," *Plos One*, vol. 11, no. 5, p. e0155837, 2016.

[56] L. Bartholin, S. E. Powers, T. A. Melhuish, S. Lasse, M. Weinstein, and D. Wotton, "TGIF Inhibits Retinoid Signaling," *Molecular and Cellular Biology*, vol. 26, no. 3, pp. 990–1001, 2006.

[57] L. M. Brill, W. Xiong, K.-B. Lee, S. B. Ficarro, A. Crain, Y. Xu, A. Terskikh, E. Y. Snyder, and S. Ding, "Phosphoproteomic analysis of human embryonic stem cells.," *Cell stem cell*, vol. 5, pp. 204–13, aug 2009.

[58] E. Fischer and E. Fourneau, "Ueber eigine Derivate des Glykocolls," *Ber. Dtsch. Chem. Ges*, vol. 34, no. August, pp. 2868–2877, 1901.

[59] C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto, "Protein post-translational modifications: The chemistry of proteome diversifications," *Angewandte Chemie - International Edition*, vol. 44, no. 45, pp. 7342–7372, 2005.

[60] R. M. Wilson, S. Dong, P. Wang, and S. J. Danishefsky, "The winding pathway to erythropoietin along the chemistry-biology frontier: A success at last," *Angewandte Chemie - International Edition*, vol. 52, no. 30, pp. 7646–7665, 2013.

[61] A. Dirksen, E. W. Meijer, W. Adriaens, and T. M. Hackeng, "Strategy for the synthesis of multivalent peptide-based nonsymmetric dendrimers by native chemical ligation.," *Chemical communications (Cambridge, England)*, no. 15, pp. 1667–1669, 2006.

[62] B. L. Nilsson, M. B. Soellner, and R. T. Raines, "Chemical synthesis of proteins.," *Annual review of biophysics and biomolecular structure*, vol. 34, pp. 91–118, jan 2005.

[63] R. B. Merrifield, "Solid Phase Peptide Synthesis. I. The Synthesis of a tetrapeptide," *Journal of the American Chemical Society*, vol. 85, no. 14, pp. 2149–2154, 1963.

[64] S. Chandrudu, P. Simerska, and I. Toth, "Chemical methods for peptide and protein production.," *Molecules*, vol. 18, pp. 4373–88, jan 2013.

[65] S. L. Pedersen, A. P. Tofteng, L. Malik, and K. J. Jensen, "Microwave heating in solid-phase peptide synthesis.," *Chemical Society reviews*, vol. 41, no. 5, pp. 1826–44, 2012.

[66] D. Singer, T. Zauner, M. Genz, R. Hoffmann, and T. Zuchner, "Synthesis of pathological and nonpathological human exon 1 huntingtin," *Journal of Peptide Science*, vol. 16, no. 7, pp. 358–363, 2010.

[67] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3390–3400, 2005.

[68] N. Fotouhi, N. G. Galakatos, and D. S. Kemp, "Peptide synthesis by prior thiol capture. 6. Rates of the disulfide-bond-forming capture reaction and demonstration of the overall strategy by synthesis of the C-terminal 29-peptide sequence of BPTI," *Journal of Organic Chemistry*, vol. 54, no. 12, pp. 2803–2817, 1989.

[69] T. Wieland, E. Bokelmann, L. Bauer, H. U. Lang, and H. Lau, "Über Peptidsynthesen. 8. Mitteilung. Bildung von S-haltigen Peptiden durch intramolekulare Wanderung von Aminoacylresten," *Annalen der Chemie*, vol. 583. Band, no. 1951, pp. 129–149, 1953.

[70] P. Dawson, T. Muir, I. Clark-Lewis, and S. Kent, "Synthesis of proteins by native chemical ligation," *Science*, vol. 266, no. 5186, pp. 776–9, 1994.

[71] P. McCaldon and P. Argos, "Oligopeptide Biases in Protein Sequences and their use in Predicting Protein Coding Regions in Nucleotide Sequences," *Proteins*, vol. 4, no. 2, pp. 99–122, 1988.

[72] I. Pe'er, C. E. Felder, O. Man, I. Silman, J. L. Sussman, and J. S. Beckmann, "Proteomic Signatures: Amino Acid and Oligopeptide Compositions Differentiate among Phyla," *Proteins: Structure, Function and Genetics*, vol. 54, no. 1, pp. 20–40, 2004.

[73] P. E. Dawson, M. J. Churchill, M. R. Ghadiri, and S. B. H. Kent, "Modulation of reactivity in native chemical ligation through the use of thiol additives," *Journal of the American Chemical Society*, vol. 119, no. 19, pp. 4325–4329, 1997.

[74] E. C. B. Johnson and S. B. H. Kent, "Insights into the mechanism and catalysis of the native chemical ligation reaction.," *Journal of the American Chemical Society*, vol. 128, pp. 6640–6, may 2006.

[75] B. J. Backes and J. A. Ellman, "An Alkanesulfonamide "Safety-Catch" Linker for Solid-Phase Synthesis," *The Journal of Organic Chemistry*, vol. 64, pp. 2322–2330, apr 1999.

[76] J. B. Blanco-Canosa and P. E. Dawson, "An efficient Fmoc-SPPS approach for the generation of thioester peptide precursors for

use in native chemical ligation," *Angewandte Chemie - International Edition*, vol. 47, no. 36, pp. 6851–6855, 2008.

[77] L. Z. Yan and P. E. Dawson, "Synthesis of Peptides and Proteins without Cysteine Residues by Native Chemical Ligation Combined with Desulfurization," *Journal of the American Chemical Society*, no. 123, pp. 526–533, 2001.

[78] Q. Wan and S. J. Danishefsky, "Free-radical-based, specific desulfurization of cysteine: A powerful advance in the synthesis of polypeptides and glycopolypeptides," *Angewandte Chemie - International Edition*, vol. 46, no. 48, pp. 9248–9252, 2007.

[79] L. Liu, ed., *Protein Ligation and Total Synthesis I*, vol. 362. Springer, 2016.

[80] L. E. Canne, S. J. Bark, and S. B. H. Kent, "Extending the applicability of native chemical ligation," *Journal of the American Chemical Society*, vol. 118, no. 25, pp. 5891–5896, 1996.

[81] A. Brik, Y. Y. Yang, S. Ficht, and C. H. Wong, "Sugar-assisted glycopeptide ligation," *Journal of the American Chemical Society*, vol. 128, no. 17, pp. 5626–5627, 2006.

[82] S. Ficht, R. J. Payne, A. Brik, and C. H. Wong, "Second-generation sugar-assisted ligation: A method for the synthesis of cysteine-containing glycopeptides," *Angewandte Chemie - International Edition*, vol. 46, no. 31, pp. 5975–5979, 2007.

[83] R. J. Payne, S. Ficht, W. A. Greenberg, and C.-H. Wong, "Cysteine-free peptide and glycopeptide ligation by direct aminolysis.," *Angewandte Chemie (International ed. in English)*, vol. 47, pp. 4411–5, jan 2008.

[84] D. Bang and S. B. H. Kent, "A one-pot total synthesis of crambin," *Angewandte Chemie - International Edition*, vol. 43, no. 19, pp. 2534–2538, 2004.

[85] D. Bang, B. L. Pentelute, and S. B. H. Kent, "Kinetically controlled ligation for the convergent chemical synthesis of proteins," *Angewandte Chemie - International Edition*, vol. 45, no. 24, pp. 3985–3988, 2006.

[86] T. W. Muir, D. Sondhi, and P. A. Cole, "Expressed protein ligation: a general method for protein engineering.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. June, pp. 6705–6710, 1998.

[87] M. Q. Xu and F. B. Perler, "The mechanism of protein splicing and its modulation by mutation," *Embo Journal*, vol. 15, no. 19, pp. 5146–5153, 1996.

[88] V. Muralidharan and T. W. Muir, "Protein ligation: an enabling technology for the biophysical analysis of proteins.," *Nature methods*, vol. 3, no. 6, pp. 429–438, 2006.

[89] C. Chatterjee, R. K. McGinty, J. P. Pellois, and T. W. Muir, "Auxiliary-mediated site-specific peptide ubiquitylation," *Angewandte Chemie - International Edition*, vol. 46, no. 16, pp. 2814–2818, 2007.

[90] D. Kemp, Z. W. Bernstein, and G. N. McNeil, "Nucleophilic reactivity of peptides towards 2-Acyloxy-N-ethylbenzamides. The utility of free peptides as nucleophiles in amide bond forming reactions," *Journal of Organic Chemistry*, vol. 69, no. 19, pp. 2831–2835, 1974.

[91] D. Kemp, S.-L. H. Choong, and J. Pekaar, "Rate constants for peptide p-Nitrophenyl ester coupling reactions in dimethylformamide. A model for steric interactions in the peptide bond forming transition state," *Journal of Organic Chemistry*, vol. 39, no. 26, pp. 3841–3847, 1974.

[92] J. Blake, "Peptide segment coupling in aqueous medium: silver ion activation of the thiolcarboxyl group," *International journal of peptide and protein research*, vol. 17, no. 2, pp. 273–4, 1981.

[93] S. Aimoto, "Polypeptide synthesis by the thioester method.," *Biopolymers*, vol. 51, pp. 247–65, jan 1999.

[94] G. Chen, Q. Wan, Z. Tan, C. Kan, Z. Hua, K. Ranganathan, and S. J. Danishefsky, "Development of efficient methods for accomplishing cysteine-free peptide and glycopeptide coupling," *Angewandte Chemie - International Edition*, vol. 46, no. 39, pp. 7383–7387, 2007.

[95] R. E. Thompson, K. a. Jolliffe, and R. J. Payne, "Total synthesis of microcin B17 via a fragment condensation approach.," *Organic letters*, vol. 13, pp. 680–3, feb 2011.

[96] G. Santhakumar and R. J. Payne, "Total Synthesis of Polydiscamides B, C, and D via a Convergent Native Chemical Ligation: Oxidation Strategy," pp. 8–11, 2014.

[97] T. M. Hackeng, J. H. Griffin, and P. E. Dawson, "Protein synthesis by native chemical ligation: expanded scope by using straightforward methodology.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 18, pp. 10068–10073, 1999.

[98] V. R. Pattabiraman and J. W. Bode, "Rethinking amide bond synthesis.," *Nature*, vol. 480, no. 7378, pp. 471–9, 2011.

[99] J. W. Bode, R. M. Fox, and K. D. Baucom, "Chemoselective amide ligations by decarboxylative condensations of N-alkylhydroxylamines and alpha-ketoacids," *Angewandte Chemie - International Edition*, vol. 45, no. 8, pp. 1248–1252, 2006.

[100] J. W. Bode, "Reinventing amide bond formation," *Top Organomet Chem*, vol. 44, pp. 13–34, 2013.

[101] I. Pusterla and J. W. Bode, "An oxazetidine amino acid for chemical protein synthesis by rapid, serine-forming ligations," *Nature Chemistry*, vol. 7, no. June, pp. 1–5, 2015.

[102] H. Staudinger and J. Meyer, "Über neue organische Phosphorverbindungen," *Helvetica Chimica Acta*, vol. 2, no. 1, pp. 612–618, 1919.

[103] E. Saxon and C. R. Bertozzi, "Cell Surface Engineering by a Modified Staudinger Reaction," *Science*, vol. 287, no. 5460, pp. 2007–2010, 2000.

[104] B. L. Nilsson, L. L. Kiessling, and R. T. Raines, "Staudinger ligation: a peptide from a thioester and azide.," *Organic letters*, vol. 2, no. 13, pp. 1939–1941, 2000.

[105] E. Saxon, J. I. Armstrong, and C. R. Bertozzi, "A "traceless" Staudinger ligation for the chemoselective synthesis of amide bonds.," *Organic letters*, vol. 2, no. 14, pp. 2141–2143, 2000.

[106] C. I. Schilling, N. Jung, M. Biskup, U. Schepers, and S. Bräse, "Bioconjugation via azide-Staudinger ligation: an overview.," *Chemical Society reviews*, vol. 40, no. 9, pp. 4840–4871, 2011.

[107] A. Tam and R. T. Raines, *Chapter 2 Protein Engineering with the Traceless Staudinger Ligation*, vol. 462. Elsevier Inc., 1 ed., 2009.

[108] H. C. Kolb, M. G. Finn, and K. B. Sharpless, "Click Chemistry: Diverse Chemical Function from a Few Good Reactions," *Angewandte Chemie - International Edition*, vol. 40, no. 11, pp. 2004–2021, 2001.

[109] A. A. H. Ahmad Fuaad, F. Azmi, M. Skwarczynski, and I. Toth, "Peptide conjugation via CuAAC 'click' chemistry," *Molecules*, vol. 18, no. 11, pp. 13148–13174, 2013.

[110] A. Toplak, T. Nuijens, P. J. Quaedflieg, B. Wu, and D. B. Janssen, "Peptiligase, enzyme for efficient chemoenzymatic peptide synthesis and cyclisation in water," *Advanced Synthesis & Catalysis*, vol. 358, pp. 2140–2147, 2016.

[111] T. Shiromizu, J. Adachi, S. Watanabe, T. Murakami, T. Kuga, S. Muraoka, and T. Tomonaga, "Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the phosphositeplus database as part of the chromosome-centric human proteome project," *Journal of Proteome Research*, vol. 12, no. 6, pp. 2414–2421, 2013.

[112] S. B. H. Kent, "Chemical Synthesis of Peptides and Proteins," *Annual Review of Biochemistry*, vol. 57, no. 1, pp. 957–989, 1988.

[113] L. A. Carpino and G. Y. Han, "The 9-Fluorenylmethoxycarbonyl Amino-Protecting Group," *Journal of Organic Chemistry*, vol. 37, no. 22, pp. 3404–3409, 1972.

[114] F. García-Martín and F. Albericio, "Solid supports for the synthesis of peptides," *Chimica Oggi*, vol. 26, no. 4, pp. 29–34, 2008.

[115] F. García-Martín, M. Quintanar-Audelo, Y. García-Ramos, L. Cruz, C. Gravel, R. Furic, S. Côté, J. Tulla-Puche, and F. Albericio, "ChemMatrix, a poly (ethylene glycol)-based support for the solid-phase synthesis of complex peptides," *Journal of combinatorial chemistry*, vol. 8, no. 2, pp. 213–220, 2006.

[116] G. Kenner, J. Mc Dermott, and R. Sheppard, "The safety catch principle in solid phase peptide synthesis," *Journal of the Chemical Society D: Chemical Communications*, no. 12, pp. 636–637, 1971.

[117] E. Kaiser, R. L. Colescott, C. D. Bossinger, and P. I. Cook, "Color test for detection of free terminal amino groups in the solid-phase synthesis of peptides," *Analytical biochemistry*, vol. 34, pp. 595–598, 1970.

[118] T. Christensen, "A qualitative test for monitoring coupling completeness in solid phase peptide synthesis using chloranil," *Acta Chemica Scandinavica*, vol. 33B, pp. 763–766, 1979.

[119] T. Todorovski, D. Suñol, A. Riera, and M. J. Macias, "Addition of HOBt improves the conversion of thioester-amine chemical ligation," *Biopolymers (Peptide Science)*, vol. 104, no. 6, pp. 693–702, 2015.

[120] R. Hamid, J. Patterson, and S. J. Brandt, "Genomic structure, alternative splicing and expression of TG-interacting factor, in human myeloid leukemia blasts and cell lines.," *Biochimica et biophysica acta*, vol. 1779, pp. 347–55, may 2008.

[121] R. B. Mujumdar, L. a. Ernst, S. R. Mujumdar, C. J. Lewis, and A. S. Waggoner, "Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters.," *Bioconjugate Chemistry*, vol. 4, no. 2, pp. 105–11, 1993.

[122] R. Zhou, G. G. Maisuradze, D. Suñol, T. Todorovski, M. J. Macias, Y. Xiao, H. A. Scheraga, C. Czaplewski, and A. Liwo, "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 18243–8, dec 2014.

[123] I. Rabi, J. Zacharias, S. Millman, and P. Kusch, "A New Method of Measuring Nuclear magnetic Moment," *Physical Review*, vol. 53, no. February, p. 318, 1938.

[124] F. Bloch, W. W. Hansen, and M. Packard, "The Nuclear Induction Experiment," *Phys. Rev.*, vol. 70, no. 474, 1946.

[125] E. Purcell, H. Torrey, and R. Pound, "Resonance Absorption by Nuclear Magnetic Moments in a Solid," *Physical Review*, vol. 69, no. 1-2, pp. 37–38, 1946.

[126] W. P. Aue, E. Bartholdi, and R. R. Ernst, "Two dimensional spectroscopy. Application to nuclear magnetic resonance," *The Journal of Chemical Physics*, vol. 64, no. 5, 1976.

[127] M. P. Williamson, T. F. Havel, and K. Wuthrich, "Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry.," *Journal of molecular biology*, vol. 182, pp. 295–315, mar 1985.

[128] V. Tugarinov, W.-Y. Choy, V. Y. Orekhov, and L. E. Kay, "Solution NMR-derived global fold of a monomeric 82-kDa enzyme," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 3, pp. 622–627, 2005.

[129] P. Selenko, D. P. Frueh, S. J. Elsaesser, W. Haas, S. P. Gygi, and G. Wagner, "In situ observation of protein phosphorylation by

high-resolution NMR spectroscopy.," *Nature structural & molecular biology*, vol. 15, pp. 321–9, mar 2008.

[130] C. Göbl, T. Madl, B. Simon, and M. Sattler, "NMR approaches for structural analysis of multidomain proteins and complexes in solution," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 80, pp. 26–63, 2014.

[131] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren, "Crystallography & NMR system: A new software suite for macromolecular structure determination.," *Acta crystallographica. Section D, Biological crystallography*, vol. 54, no. 5, pp. 905–921, 1998.

[132] M. Nilges, M. Clore, and A. M. Gronenborn, "Determination of three dimensional structures of proteins from interproton distance data by hybrid distance geometry calculations.," *FEBS Lett.*, vol. 229, no. 2, pp. 317–324, 1988.

[133] G. Bodenhausen and D. J. Ruben, "Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy," *Chemical Physics Letters*, vol. 69, no. 1, pp. 185–189, 1980.

[134] S. Grzesiek and A. Bax, "Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR," *Journal of the American Chemical Society*, vol. 114, no. 16, pp. 6291–6293, 1992.

[135] S. Grzesiek and A. Bax, "An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins," *Journal of Magnetic Resonance*, vol. 99, no. 1, pp. 201–207, 1992.

[136] Z. Y. J. Sun, D. P. Frueh, P. Selenko, J. C. Hoch, and G. Wagner, "Fast assignment of 15N-HSQC peaks using high-resolution 3D HNcocaNH experiments with non-uniform sampling," *Journal of Biomolecular NMR*, vol. 33, no. 1, pp. 43–50, 2005.

[137] T. D. Goddard and D. G. Kneller, "SPARKY 3," *University of California, San Francisco*, 2008.

[138] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue, "The CCPN data model for NMR spectroscopy: Development of a software pipeline," *Proteins: Structure, Function and Genetics*, vol. 59, no. 4, pp. 687–696, 2005.

[139] M. P. Williamson, "Using chemical shift perturbation to characterise ligand binding," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 73, pp. 1–16, 2013.

[140] E. a. Bienkiewicz and K. J. Lumb, "Random-coil chemical shifts of phosphorylated amino acids," *Journal of Biomolecular NMR*, vol. 15, no. 3, pp. 203–206, 1999.

[141] C. Bartels, T. H. Xia, M. Billeter, P. Guntert, and K. Wuthrich, "The program XEASY for computer-supported NMR spectral analysis of biological macromolecules.," *Journal of biomolecular NMR*, vol. 6, pp. 1–10, jul 1995.

[142] J. F. Doreleijers, W. F. Vranken, C. Schulte, J. L. Markley, E. L. Ulrich, G. Vriend, and G. W. Vuister, "NRG-CING: Integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB," *Nucleic Acids Research*, vol. 40, no. D1, pp. 519–524, 2012.

[143] R. A. Laskowski, J. A. Rullmannn, M. W. MacArthur, R. Kaptein, and J. M. Thornton, "AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.," *Journal of biomolecular NMR*, vol. 8, pp. 477–486, dec 1996.

[144] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera - A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[145] C. J. Wienken, P. Baaske, U. Rothbauer, D. Braun, and S. Duhr, "Protein-binding assays in biological liquids using microscale thermophoresis," *Nat Commun*, vol. 1, no. 7, p. 100, 2010.

[146] B. M. Chacko, B. Qin, J. J. Correia, S. S. Lam, M. P. de Caestecker, and K. Lin, "The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization.," *Nature structural biology*, vol. 8, no. 3, pp. 248–253, 2001.

[147] D. L. Sheridan, Y. Kong, S. A. Parker, K. N. Dalby, and B. E. Turk, "Substrate discrimination among mitogen-activated protein kinases through distinct docking sequence motifs," *Journal of Biological Chemistry*, vol. 283, no. 28, pp. 19511–19520, 2008.

[148] D. J. Mandell, I. Chorny, E. S. Groban, S. E. Wong, E. Levine, C. S. Rapp, and M. P. Jacobson, "Strengths of hydrogen bonds involving phosphorylated amino acid side chains," *Journal of the American Chemical Society*, vol. 129, no. 4, pp. 820–827, 2007.

[149] H. Flotow, P. R. Graves, A. Wang, C. J. Fiol, R. W. Roeske, and P. J. Roach, "Phosphate groups as substrate determinants for casein kinase I action," *Journal of Biological Chemistry*, vol. 265, no. 24, pp. 14264–14269, 1990.

[150] B. Zhou, L. Chen, X. Wu, J. Wang, Y. Yin, and G. Zhu, "MH1 domain of SMAD4 binds N-terminal residues of the homeodomain of Hoxc9," *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1784, no. 5, pp. 747–752, 2008.

[151] W. Chan and P. White, eds., *Fmoc solid phase peptide synthesis: a practical approach*. Oxford University Press, 2000.

[152] L. A. Carpino, "1-Hydroxy-7-azabenzotriazole. An efficient peptide coupling additive," *Journal of the American Chemical Society*, vol. 115, no. 13, pp. 4397–4398, 1993.

[153] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How Fast-Folding Proteins Fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[154] G. G. Maisuradze, P. Senet, C. Czaplewski, A. Liwo, and H. A. Scheraga, "Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field," *Journal of Physical Chemistry A*, vol. 114, no. 13, pp. 4471–4485, 2010.

[155] M. J. Macias, V. Gervais, C. Civera, and H. Oschkinat, "Structural analysis of WW domains and design of a WW prototype," *Nature structural & molecular biology*, vol. 7, no. 5, pp. 375–379, 2000.

[156] H. Nguyen, M. Jager, A. Moretto, M. Gruebele, and J. W. Kelly, "Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 3948–53, apr 2003.

[157] J. M. R. Baker, R. P. Hudson, V. Kanelis, W.-Y. Choy, P. H. Thibodeau, P. J. Thomas, and J. D. Forman-Kay, "CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices," *Nature Structural & Molecular Biology*, vol. 14, no. 8, pp. 738–745, 2007.

[158] L. Liu, ed., *Protein Ligation and Total Synthesis II*. No. February, 2016.

[159] P. Wang and S. J. Danishefsky, "Promising general solution to the problem of ligating peptides and glycopeptides.," *Journal of the American Chemical Society*, vol. 132, pp. 17045–51, dec 2010.

Curriculum Vitae

David Suñol Moreno
Born September 15th, 1988
in Barcelona, Spain
Nationality: Spain

Cell phone: (+34) 609.35.50.44
Email: d.sunol.m@gmail.com
Address: C/Manila 61 8è 2a
08034 Barcelona

Education

09/2012 – present	**Universitat de Barcelona**, Barcelona, Spain PhD Student in Biomedicine
09/2011 – 09/2012	**Universitat Pompeu Fabra**, Barcelona, Spain MSc in Biomedical Research
09/2009 – 09/2011	**Universitat de Barcelona**, Barcelona, Spain BSc in Biochemistry Graduate with honors
09/2006 – 06/2012	**Universitat de Barcelona**, Barcelona, Spain BSc in Chemistry

Work experience

09/2012 – present	**Institute for Research in Biomedicine (IRB Barcelona)**, Barcelona, Spain Pre-doctoral researcher Structural characterization of macromolecular assemblies (Dr Maria J. Macias) ”La Caixa” /IRB Barcelona International PhD Programme fellowship
03/2015 – 06/2015	**Forschunginstitut für Molekulare Pharmakologie (FMP-Berlin)**, Berlin, Germany Visiting Student In-Cell NMR (Dr Philipp Selenko)
01/2012 – 07/2012	**Institute for Research in Biomedicine (IRB Barcelona)**, Barcelona, Spain MSc Student

	Developmental Neurobiology and Regeneration (Dr Eduardo Soriano)
08/2011 – 08/2011	**BTI BIOTECHNOLOGY INSTITUTE I+D, S.L.** Vitoria-Gasteiz, Spain Visiting Student Laboratory of Regenerative Medicine (Dr Gorka Orive)
07/2010 – 08/2010	**Universitat de Barcelona** Departament de Biologia Cel·lular, Barcelona, Spain Undergraduate Student Developmental Neurobiology and Regeneration (Dr. Eduardo Soriano)

Languages

Catalan:	native
Spanish:	native
English:	fluent
German:	basic knowledge

Conferences & Courses attended

2015	IRB PhD Student Symposium. November 12-13 Barcelona, Spain. Poster presentation
2015	Course: Scientific Communication: Getting started writing & speaking.
2014	IRB Barcelona BioMed Seminars. Transporters and other molecular machines. November 17-20. Barcelona, Spain
2014	IRB PhD Retreat, November 14-15. Barcelona, Spain. Member of the Organizer team. Poster presentation
2014	Workshop: "Cell-Free protein synthesis", September 9-11 Swedish NMR center, Goteborg, Sweden
2014	RIMLS PhD Retreat 2014, May 8-9. Nijmegen, The Netherlands Poster presentation
2013	IRB PhD Student Symposium. November 14-15. Barcelona, Spain

2013 Course: "Future of Biophysics". International School of Biological Magnetic Resonance. June 9-19. Erice, Sicily, Italy

2012 8th Forum of Neuroscience, FENS (Federation of European Neuroscience Societies). July 14-18. Barcelona, Spain

Publications

R. Zhou, G. G. Maisuradze, D. Suñol, T. Todorovski, M. J. Macias, Y. Xiao, H. A. Scheraga, C. Czaplewski, and A. Liwo: **Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements.** Proc. Natl. Acad. Sci., vol. 111, no. 51, pp. 18243–8, Dec. 2014.

T. Todorovski, D. Suñol, A. Riera, and M. J. Macias: **Addition of HOBt improves the conversion of thioester-amine chemical ligation.** Biopolym. (Peptide Sci.), vol. 104, no. 6, pp. 693–702, 2015

C. Schelhorn, P. Martín-Malpartida, D. Suñol, and M. J. Macias: **Structural Analysis of the Pin1-CPEB1 interaction and its potential role in CPEB1 degradation**. Sci. Rep., vol. 5, no. April, p. 14990, 2015