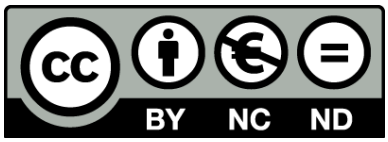
UNIVERSITAT DE BARCELONA

Into the structure of human full-length Smad proteins and the impact of cancer mutations

Tiago Lopes Gomes



UNIVERSITAT DE BARCELONA
Facultat de Farmàcia
i Ciències de l'Alimentació

UNIVERSITAT DE BARCELONA - FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

DOCTORAL THESIS

Into the structure of human full-length Smad proteins and the impact of cancer mutations

Author:
Tiago Lopes Gomes

Supervisor:
María J. Macías Hernández

Tutor:
Pedro Marrero González

Institute for Research in Biomedicine (IRB Barcelona) - Structural characterization of macromolecular assemblies group
Programa de doctorat en biomedicina

Barcelona, September 30, 2019

Declaration

I, Tiago Lopes Gomes, declare that this thesis titled, "Into the structure of human full-length Smad proteins and the impact of cancer mutations" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a PhD degree at the University of Barcelona.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where the thesis is based on work done by myself jointly with others, it is clearly stated.
- The author Tiago Lopes Gomes was a recipient of a FPI Severo Ochoa PhD Fellowship. This work was supported by a BFU2014-53787-P grant and by the BBVA foundation. Maria J. Macias is an ICREA programme investigator.
- This dissertation was typeseted using the LaTeX typesetting system based on the thesis template available in `http://www.latextemplates.com/`. The font used was Palatino 10pt. Front cover design by Mateu Marcet (`http://www.matema.org/`).

List of publications

Eric Aragón, Nina Goerner, Qiaoran Xi, **Tiago Gomes**, Sheng Gao, Joan Massagué, and Maria J. Macias. *Structural Basis for the Versatile Interactions of Smad7 with Regulator WW Domains in TGF-β Pathways.*
Structure, 20(10):1726–1736, October 2012

Albert Escobedo, **Tiago Gomes**, Eric Aragón, Pau Martín-Malpartida, Lidia Ruiz, and Maria J. Macias. *Structural basis of the activation and degradation mechanisms of the E3 ubiquitin ligase Nedd4L.*
Structure, 22(10):1446–1457, October 2014

Pau Martín-Malpartida, Marta Batet, Zuzanna Kaczmarska, Regina Freier, **Tiago Gomes**, Aragón, Yilong Zou, Qiong Wang, Qiaoran Xi, Lidia Ruiz, Àngela Vea, José A. Márquez, Joan Massagué, and Maria J. Macias. *Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors.*
Nature communications, 8(1):2070, December 2017

Lidia Ruiz*, Zuzanna Kaczmarska*, **Tiago Gomes***, Eric Aragón, Pau Martín - Malpartida, Carles Torner, Natalia de Marín Garrido, Regina Freier, José Márquez, and Maria J. Macias. *Dimer/monomer propensities of MH1 domains help define how Smad proteins select DNA motifs.* (in revision in **PNAS**). *Equal contribution

Eric Aragón*, Qiong Wang*, Yilong Zou*, Sophie M. Morgani, Lidia Ruiz, Zuzanna Kaczmarska, Jie Su, Carles Torner, Lin Tian, Jing Hu, Weiping Shu, Saloni Agrawal, **Tiago Gomes**, José A. Márquez, Anna-Katerina Hadjantonakis, Maria J. Macias[#] and Joan Massagué[#]. *Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-β signaling.* (accepted in **Genes & Development**). *Equal contribuition. [#]Corresponding authors.

In preparation:

Tiago Gomes, Pau Martín - Malpartida, Lidia Ruiz, Eric Aragón, Ángela Vea, Tiago N. Cordeiro*, and Maria J. Macias*. *Into the structure of human full-length Smad proteins.* *Corresponding authors

Tiago Gomes*, Ángela Vea*, Zuzanna Kaczmarska, Eric Aragón, Lidia Ruiz, Pau Martín - Malpartida, Roma Domenech, and Maria J. Macias. *Smad proteins cancer mutants affect protein stability, by disrupting salt bridge networks, while maintaining DNA binding functionality.* *Equal contribution

Acknowledgements

First of all i would like to thank Maria Macias, my supervisor, for letting me carry out my PhD studies in her laboratory. Thank you Maria for your patience, guidance, giving me the freedom to think out of the box and for our fruitful discussions about science. Thank you for your support Maria.

Secondly I would like to thank Tiago Cordeiro for his availably in collaborating with this work. Without his expertise part of my PhD would be much more difficult to tackle. Thank you Tiago for our friendship for so many years, almost 20!!!, and for our many hours of discussions. Science can be fun, honest, exciting and rigorous. You're an inspiration. Thank you.

Thank you Àngela for your collaboration and friendship and for doing a very good job in all things related to biochemistry. Part of this work, without you, with be much more difficult.

I would like to thank all of my lab mates that shared the bench with me throughout the years. Thank you Pau for your friendship. Our shared love for weird things in general and some very high quality comedic tirades, made my days easier. Also your help with programming and setting up NMR experiments was fundamental, thanqqqqq youuuuu (arms crossed, bowing down).

Thank you Lidia for your friendship, for introducing me to world of protein production and help with day to day activities in the lab. Without you i would not have learned so much for sure. Thank you!

My travel companions: David, one of a kind, thank you for your friendship xuclin. Toni, my master jedi, i appreciate you my friend, thank you for being such a nice guy. Jordi thank you for our shared love for uncomfortable humor. Eric thanks for being around the lab and for being such a relaxed guy.

Alberinho thanks for helping me out and for our discussions about science and everything else, without your friendship this work would be harder for sure!

Carles and Jorge, thanks for helping me out whenever I needed and for everything involving having a laugh.

Marco you're a good man, good luck!

Toni, Jimmy, Regina, Ewelina, Mads, Marco, our post-docs, many thanks guys for our fruitful discussions.

Thank you Roma Domenech for your help in setting up molecular dynamics simulations. Thank you Roman for proofreading the thesis and for your insightful comments.

On a more informal note i also would like to thank some, outside-of-the-lab, friends that made my life easier while undertaking my PhD research. Claudia és a maior, gosto muito de ti! Sr Peras uma beijoca nessa pança. Marco Bellini é que a sabe toda. Meu Pirralhinha uma beijoca nessa cara que sei que gostas. Luis e Marta gosto de voçês! Freaks deste mundo uni-vos! Elies y Gaby que guapos sois! Lucho me tenés repodrido! Aram, one of a kind, thank you man for your unconditional friendship. Freaks of this world unite! Dani, nuestras comidas me hacen

bien. Gracias por tu amistad amic! Mateu, gracias por la portada y tu amistad. Algún dia tendremos nuestro dúo ostia! Leonor, Giovanni beijinhos, baci! Albert, David, Edgar, Roman, Toni us estimo molt! Gracias por la vostra amistat amics. Lastly i would like to thank my parents, António and Lurdes. Without their love and unconditional support this work would not have been possible. To them I dedicate this work.

Contents

List of Figures

List of Tables

List of Abbreviations

3C	Human rhinovirus 3C protease
Å	Angstrom
χ^2	Pearson's chi-squared goodness of fit test
CcpNmr	Collaborative computational project for NMR
CCS	Collisional cross section
Co-Smad	Common partner Smad
COM	Centre-of-mass
Dmax	Maximum distance
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift assay
EOM	Ensemble optimisation method
FID	Free induction decay
FL	Full-length
FM	Flexible-Meccano
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)
HSQC	Heteronuclear single-quantum correlation spectroscopy
IM-MS	Ion mobility mass spectrometry
IPTG	Isopropyl β-D-1-thiogalactopyranoside
I-Smad	Inhibitory Smad
kDa	Kilodaltons
KDE	Kernel density estimate
LB	Luria broth
MBP	Maltose binding protein
MES	2-(N-morpholino)ethanesulfonic acid
µM	Micromolar
mM	Milimolar
ms	Miliseconds
M	Molar
nm	Nanometers
NMR	Nuclear magnetic ressonance
NOE	Nuclear overhauser effect
NOESY	Nuclear overhauser effect spectroscopy
PCR	Polymerase chain reaction

PDB	Protein data bank
PMSF	Phenylmethanesulfonyl fluoride
PY	Proline-Tyrosine motif
Rg	Radius of gyration
RMSD	Root mean squared deviation
RMSF	Root mean squared deviation
R-Smad	Receptor regulated Smad
SAD	Smad activation domain
SARA	Smad anchor for recetor activation
SAXS	Small angle x-ray scattering
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SMAD	Small mother against decapentaplegic
SOC	Super optimal broth with catabolite repression
SUMO	Small ubiquitin-like modifier
TCEP	Tris(2-carboxyethyl)phosphine
TEV	Tobacco etch virus
TGFβ	Transforming growth beta
TGFβR	Transforming growth beta receptor
TRIS	2-Amino-2-hydroxymethyl-propane-1,3-diol
TROSY	Transverse relaxation optimized spectroscopy
WT	Wild type

Para os meus pais...

Chapter 1

Introduction

1.1 The Transforming growth factor beta (TGFβ) signalling pathway

Signal transduction pathways are very important mechanisms that cells use to perform a myriad of biological processes that when aberrant, can cause serious diseases. Understanding the molecular basis of these pathways in detail, can thus help the interpretation of these signalling cascades in health and disease. The TGFβ (Transforming Growth Factor beta) signalling pathway is one of the best-studied ones and highly conserved in metazoans. Studying some of the critical complexes in this pathway, specifically those involving Smad (Sma and Mad related) proteins and their interactions, with protein partners and DNA, is thus an active area of research [1–3].

1.1.1 Principal effectors - The Smad protein family

Smad proteins are the principal effectors of the Smad-dependent TGFβ family of pathways and are extremely well conserved in all metazoans. They were firstly described in the late nineties as the principal driving force in TGFβ signalling. Smad proteins are classified in three functional classes: Receptor-regulated Smad (R- Smad, 1/5 2/3 and 8), the Co-mediator Smad (Co-Smad or Smad4) and the inhibitors I-Smad (Smad6/7). Smad2 and Smad3 mainly interact with TGFβ receptors, whereas Smad1, Smad5, and Smad8 are mostly activated by BMP receptors. R-Smads and Smad4 consist of two Mad Homology domains, MH1 and MH2, connected by a linker region of variable length poorly conserved. The MH1 domains of R-Smads and Smad4 bind to DNA, whereas the MH2 domain and the linker function as scaffolds for receptors, regulator proteins, and transcription cofactors, that interact and determine the outcome of the signal [3, 4] (see figure 1.1). The linker region acts as a protein scaffold, acting as a substrate, for ubiquitin ligases and is also a target for post-translational modifications, namely phosphorylation by CDK8/9 and GSK3 [5]. Ubiquitination and acetylation usually target the MH1 domain as well as nuclear export and localization signals [4, 6] (see figure 1.1B).

The MH1 domain is a zinc binding protein, with a zinc finger motif strictly conserved in Smad proteins, with three cysteines an one histidine residues (C3H1), that is important for fold stability [7]. The MH1 alpha-helix, beta-sheet mixed fold is unique among protein structures. It also possesses an anti-parallel β-hairpin that binds DNA [4] (see figure 1.1A).

The MH2 domain is also highly conserved between all Smads, with a fold consisting of eleven beta-stands organised in two sheets adopting the Greek key topology and a four helix bundle [8] (see figure 1.1A). As a protein-protein interaction hub it has various protein binding interfaces; an hydrophobic patch where SARA [1] binds to co-localize Smads to the membrane [9, 10] and where some transcription factors (e.g. FOXH1, SKI [2]) also bind. It also possesses a positively charged surface, the L3 loop, opposite to C-terminus phosphorylation sites, that drive protein oligomerization, transport into the nucleus and TGFβ signalling propagation (see 1.1.2) [10].

Compared to R-Smads and Co-Smads, the I-Smads have low sequence similarity at the MH1 domain, they contain a divergent MH2 domain and a linker region with a PY motif. I-Smads are expressed in response to TGF- ß or BMPs to provide negative feedback in the pathway [11–13] and in response to other pathways, such as STAT, to oppose TGFβ signaling [14]. Recently the molecular basis for the interaction of Smad7 with R-Smads, to promote TGFβ inhibition, has been proposed [15].

1.1.2 Pathway overview

The TGFβ family of cytokines (including BMP, nodal, activin, myostatin and TGFβ itself) regulates many processes during the life of metazoans. Encoded in the human genome there are twelve TGFβ receptors and thirty two receptors in total [1, 4]. This network of signals, which is highly conserved during evolution, plays important roles in embryo development, in differentiated tissue homeostasis and also in immune responses [4, 16]. After more than fifty thousand published papers, and forty years of research on the TGFβ family of cytokines, key questions such as TGFβ signalling is so much context-dependent remain open. Indeed, the TGFβ molecule can trigger, with similar ability, a given function and its opposite. For instance, it can inhibit cell proliferation and promote cell growth, or enhance cell pluripotency and differentiation. This context-dependence signalling of TGFβ is caused, at least in part, by a family of DNA binding proteins named Smad, that can act as mediators in the transmission of the signal, created by the TGFβ hormone, from the membrane receptor to the nucleus [16]. Smads have the ability to interact with many other proteins such as transcription factors, transcription coactivators or corepressors, in addition to DNA (see chapter 1.1.1). Furthermore,

[1] Smad anchor for receptor activation protein, also known as, zinc finger FYVE-type containing 9 or ZFYVE9 protein

[2] SNON; also called SKI-like proto-oncogene or SKIL. FOXH1 is also called Forkhead box protein H1.

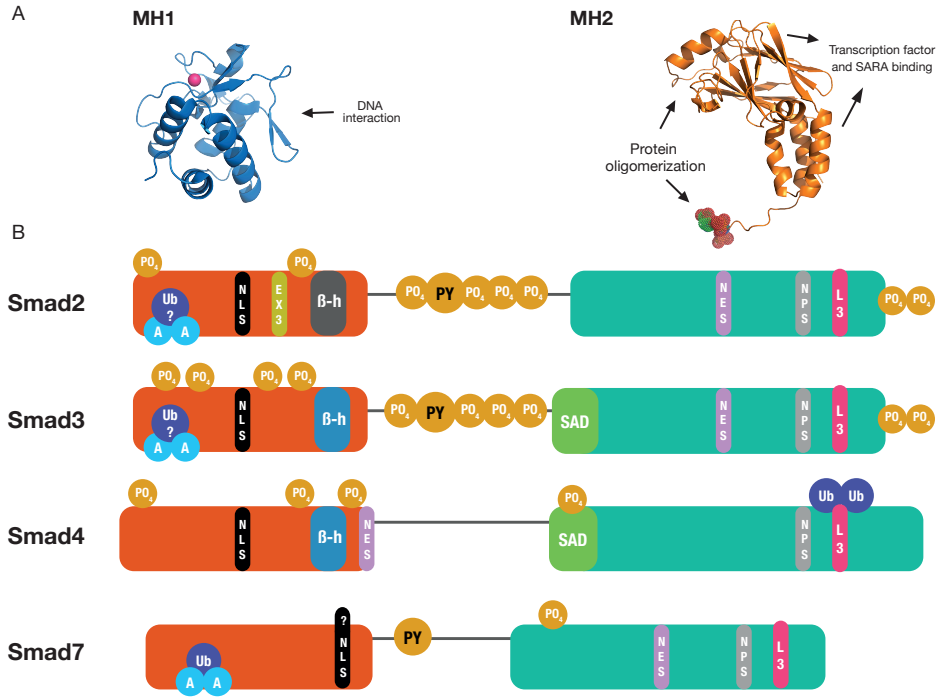


FIGURE 1.1: The protein Smad family. **A**: Representative PDBs of the MH1 (PDB code: 3QSV) and MH2 (PDB code: 1KHX) domains with the most characteristic protein and DNA interaction sites represented. The zinc atom is in pink and the phosphate atoms in dot representation. **B**: In red is represented the MH1 domain, in green the MH2 domain, connected by the inter-domain linker in grey. SAD is the Smad activation domain. All other annotations are post-translational modifications; A is acetylation, NLS is the nuclear localisation signal, Ub is ubiquitination, NES is the nuclear export signal, L3 is the L3 loop (essential for Smad activation), PY is the PY motif and PO_4 are phosphorylation sites. β-h is the beta hairpin involved in DNA binding.

their functions can be finely tuned in multiple cellular contexts [4]. Thus, Smad proteins constitute a general system present in all metazoan cells, which can easily be expanded to form versatile functional complexes (see chapter 1.1.1).

Resulting from an interaction with the TGFβ hormone, at the membrane, the TGFβ receptor type II phosphorylates the receptor type I, through its kinase domain [17] (see figure 1.2). The receptor type I is phosphorylated, at its GS domain [3], with the subsequent phosphorylation of an R-Smad (see figure 1.2, represented by Smad2) at the C-terminus in a Ser–X–Ser motif [4]. Once the signalling cascade is activated downstream, the phosphorylated R-Smads form homomeric and heterotrimeric complexes with Smad4 that translocate to the nucleus and interact with DNA and other proteins, to activate or repress transcription. Disentangling this complex architecture has been prone to some debate but it is thought that it exists predominantly in dimer and trimer oligomeric forms, combining R-smads and Smad4 [17–21]

There is a constant shuttling of Smads from the cytoplasm to the nucleus and vice-versa, with the exact duration of the signal being dependent on a plethora of effects, not all quite well understood [17]. Among them the concentration dependence of Smads in the nucleus and cytoplasm, their interaction with transcription activators (e.g FOXH1 [22]) and corepressors (TGIF1 [23]), pathological mutations [14, 24] and DNA interactions (e.g. Smad binding elements (CAGAC) and GC rich motifs in regulatory promoter sequences [25])

For signal termination Smads can be ubiquitinated at the MH1 domain acting as substrates for E3 ubiquitin ligases (e.g. NEDD4, SMURF2) at the HECT domain [26] (see figure 1.2). These ligases target the PY motif, of Smads via their WW domains for subsequent proteasomal degradation.

1.1.3 TGFβ in disease

Since the discovery of the TGFβ family around the early eighties and Smad proteins in the mid nineties, the involvement of TGFβ in pathological processes was almost immediately apparent. Looking for keywords in the literature relating TGFβ, Smads and cancer one can see an almost direct correlation since their discovery (see figure 1.3).

Smad4 is the fourth most mutated protein in pancreatic cancer and colorectal cancer, with a alteration frequency of 25 to 30% in gastrointestinal cancers [27, 28]. Mutations, in Smads, usually occur at the MH2 domain disrupting oligomerization and also affecting protein stability. In the MH1 domain mutations usually affect zinc coordination and are located at the core of the protein affecting protein stability.

Acting as, a context-dependent, tumour suppressor, TGFβ can inhibit epithelial growth, endothelial, hematopoietic and immune cell proliferation as well as

[3] The GS domain is a glycine- serine rich flexible motif

[4] Ser is serine and X are methionine or valine residues

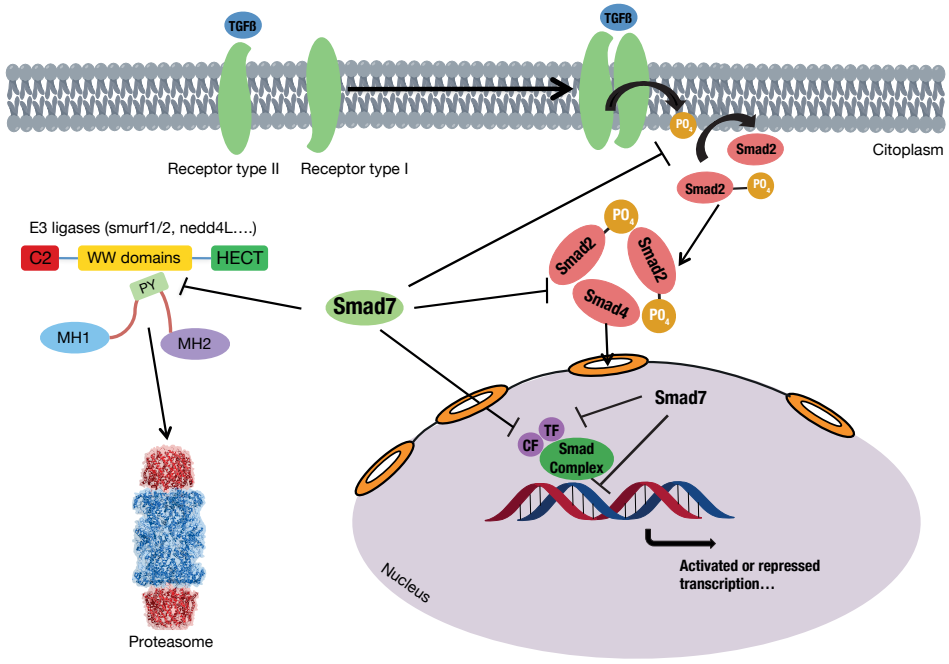


FIGURE 1.2: The TGFβ canonical pathway. PY represents the PY motif, C2, HECT and WW domains are protein domains of E3 ubiquitin ligases. CF and TF are corepressor and coactivator transcription factors, respectively [6].

extracellular matrix regulation. When unregulated it can stimulate EMT [5] and promote tumour invasion, metastasis and tissue fibrosis. TGFβ works on a delicate balance between normal tissue homeostasis and tumour proliferation, being a difficult target for drug therapy. Of the drugs that went to clinical trials it was observed disruption of normal cardiac development, aortic aneurisms and other fibrotic events. The drugs currently being evaluated usually block TGFβ activity by disrupting the phosphorylation of R-Smads by the serine/threonine kinase domain of the receptor type two [29]. Tumours, usually, use TGFβ to escape immune surveillance promoting EMT and tumour invasion. Recently, combining immunotherapy with a TGFβ inhibitor has showed promising results in modulating tumour proliferation in an animal model of colon cancer [30]. Also upregulation of TGFβ, by an upregulated deubiquinating enzyme, promotes glioblastoma

[6] The C2 domain is a membrane anchoring domain. The HECT (Homologous to the E6-AP Carboxyl Terminus) domain is involved in the ubiquitination cascade. WW domains, interacting with PY motifs in substrates are, generally, arranged in two to four units [32].

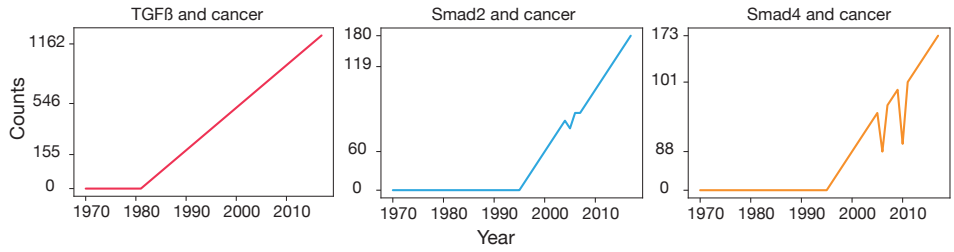


FIGURE 1.3: Number of occurrences (counts), of the keywords TGFβ and cancer, Smad2 and cancer, Smad4 and cancer, from articles in PUBMED, for each year.

progression [31]. Downregulating this pathway decreases tumour proliferation.

The abnormal behavior of TGFβ, in cancer and fibrosis, contrasts with its normal functioning in non-pathological processes, this duality renders TGFβ a subject of increased interest for the research community. Its context-dependent behaviour and the crosstalk with other pathways and the immune system needs to be better understood for a subsequent success in designing effective drug therapies.

1.2 Biophysical Methods

1.2.1 Small angle X-ray scattering in structural biology

Small angle X-ray scattering (SAXS), and its applications to structural biology, has become one of the most powerful techniques for atomic level description of biomolecules in solution, especially for flexible and multi-domain proteins. Despite its low resolution, size independence is one of the great advantages in using this technique. Its resurgence in recent years allowed researchers unprecedented views into the architecture of big and flexible molecular machines, a task almost impossible to tackle using other techniques [33, 34].

In its essence, SAXS records the diffracted waves of atoms, in solution, and correlates the angles of diffraction with the distance between those atoms. Starting from the early seventies, as seen in figure 1.4, the applications of SAXS to biomolecules have seen a resurgence, built upon the discovery of X-rays, in the late nineteenth century and the first applications off small angle scattering in the late thirties [35].

Biomolecules are scatterers, they scatter incident radiation and if its wavelength λ is on the order of atomic distances its internal structure can be infered, by analyzing the diffraction patterns. The most relevant data for scattering studies

[5]Epithelial to mesenchymal transition is a phenomenon by which cells lose their polarity and adhesion, gaining migration and invasion capabilities, becoming mesenchymal stem cells.

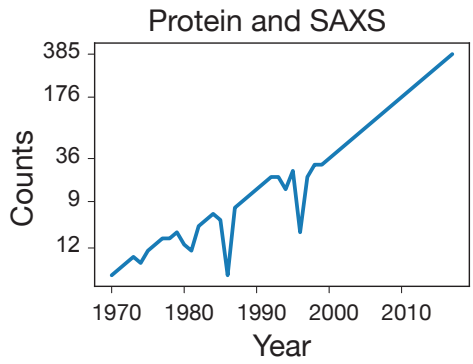


FIGURE 1.4: Number of occurrences (counts), of the keywords SAXS and protein from articles in PUBMED, for each year.

is inelastic scattering, without energy transfer from the incident wave to the scattered one. Starting from two single atoms, the phase difference of the scattered waves is $\delta = \Delta 2\pi/\lambda$, as seen in equation 1.1, where λ is the wavelength of the incident beam [7] and $\Delta = (\vec{k_1} - \vec{k_0}) \cdot r$ is the path difference between the two waves [35], as depicted in figure 1.5. The wave, A of one atom relative to another with scattering amplitude f is given by equation 1.1, where i is an imaginary number, r is the distance between the two atoms and q is the momentum vector; q is the difference between the vectors $\vec{k_o}$ and $\vec{k_1}$ that are the incident and scattered waves respectively; each one with magnitude $2\pi/\lambda$ (see figure 1.5).

$$A = f \exp(i\frac{\Delta}{\lambda}2\pi) = f \exp(i\frac{2\pi}{\lambda}(\vec{k_1} - \vec{k_0}) \cdot r) = f \exp(iq \cdot r) \tag{1.1}$$

The final term in equation 1.1 retains the momentum vector q, that can be calculated by simple geometric calculations with the aid of the diagram represented in figure 1.5A and depicted in equation 1.2. The momentum vector depends only on the scattering angle θ and the incident beam wavelength and is a fundamental quantity in SAXS [35, 36].

$$q = \frac{4\pi \sin(\theta)}{\lambda} \tag{1.2}$$

Looking only at this apparently simple equations we can start to appreciate the interplay between scattered angles θ and the distance between scatterers r; meaning the relationship between reciprocal space and real space. This duality is a fundamental concept in diffraction theory and is interconverted by Fourier transformations.

[7] The energy of the incident beam in modern synchrotrons is on the order of 12 keV, allowing sampling of distances around 1 Å.

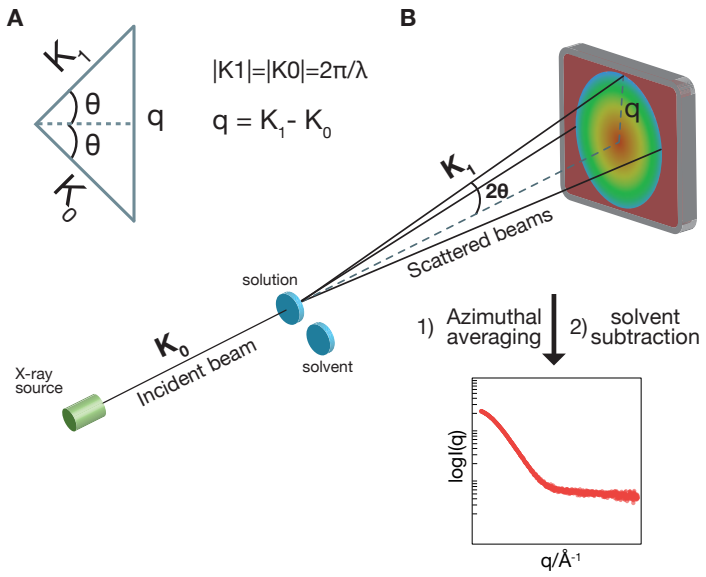


FIGURE 1.5: Typical setup for a SAXS experiment. **A**: Schematic representation of the scattering experiment; K_0 and K_1 are the incident and scattered vectors respectively, q is the momentum vector, θ is the scattered angle, $|K_0|$ and $|K_1|$ are the magnitude of vectors K_0 and K_1 respectively. **B**: SAXS measurement of a solution and its corresponding solvent. After azimuthal averaging, for both solution and solvent measurements, the curves are subtracted to obtain the final SAXS profile, where the intensity of the scattered waves I(q) is plotted against the momentum vector q.

Building upon the previous conceptual example of a simple two scatterer experiment, for practical use one should focus in analyzing much greater assemblies of atoms. This would be simply the sum of equation 1.1 over all atoms j as in equation 1.3,

$$F(q) = \sum_j f_j \exp(iq \cdot r_j) \tag{1.3}$$

and the intensity of the diffracted waves are the square of their amplitude $F(q)$ for N scatterers as in equation 1.4,

$$I_N(q) = \sum_{j=1}^{N} |F_n(q)|^2 \tag{1.4}$$

Due to the isotropic rotational motion of the biomolecules in solution, scattering intensity becomes only a function of the magnitude of the momentum transfer q, causing the intensity $I(q)$ to be symmetric around the direction of the beam as depicted in figure 1.5B. The calculated amplitudes $F(q)$ have to be rotational averaged, for all atoms, represented by equation 1.5. When measuring the intensities the phase information (see equation 1.1) is lost; the denominated phase problem in x-ray scattering.

$$I_N(q) = N\langle |F_n(q)|^2 \rangle \tag{1.5}$$

In 1915 Peter Debye calculated the rotational average, over all particle orientations, of the intensities that gave rise to the equation 1.7 [37] [8],

$$I_N(q) = \sum_k \sum_j m_j m_k \frac{\sin(qr_{jk})}{qr_{jk}} \tag{1.7}$$

where r_{ij} is the distance between two atoms, q_{ij} the momentum vector, m_j and m_k are terms that include the scattering contrast amplitudes of atoms in solution subtracted by the solvent with a scattering density. Equation 1.7 is integrated across the volume of all scatterers. This equation is another fundamental result in scattering theory where the relationship between intensities, momentum vectors, atomic distances and particle orientations is fully established. All major subsequent analyses of SAXS measurements will be based on this equation.

Building upon the major breakthrough of Debye, André Guinier published in 1939 the relationship between the radius of gyration Rg, an overall estimate of the particle size, and the scattering at small angles [38], the denominated Guinier's law. The principal assumption, as in the first term of equation 1.9, relies on the approximation of $sin(qr/qr)$ to a Taylor series of powers applied to qr [9].

$$\begin{aligned} I(q) &= \sum_{jk} \frac{m_j m_k}{qr_{jk}} (qr_{jk} - \frac{q^3 r_{jk}^3}{3!} + \cdots) \\ &= (\sum_j m_j)^2 (1 - \frac{1}{3} R_g^2 q^2 + \cdots) \end{aligned} \tag{1.9}$$

[8] The derivation of the Debye equation uses the approximation where the scattered intensity is spherically averaged.

$$\langle \exp(iq \cdot r) \rangle = \frac{\sin(qr)}{qr} \tag{1.6}$$

[9]

$$\frac{sin(qr)}{qr} \approx 1 - \frac{qr^2}{3!} + \frac{qr^4}{4!} \cdots \tag{1.8}$$

Where the radius of gyration Rg is given by,

$$Rg^2 = \frac{\sum m_j r_j^2}{\sum m_j} \tag{1.10}$$

The equality in equation 1.10 is used to derive the second term in equation 1.9. Finally having demonstrated the relationship between the radius of gyration and the scattering intensity and with another simplification, we reach the final expression for Guinier's law, expressed in equation 1.11 [35, 36].

$$\begin{aligned} I(q) &= I_0 \exp(-1/3 R_g^2 q^2) \\ \ln I(q) &= \ln I_0 - 1/3 R_g^2 q^2 \end{aligned} \tag{1.11}$$

Equation 1.11 establishes a linear relationship between the intensity and the momentum vector and it is valid, for globular proteins, if $qRg < 1.3$, the Guinier's law regimen. For retrieval of Rg it is sufficient to produce a linear fit of $I(q)$ vs q^2, with Rg being the slope and I_0 the intensity at $q = 0$. $I(0)$ depends on the square of the number of electrons and is proportional to the molecular weight and the square of the concentration, in molar, of the scattered species [34]. From this apparently simple relationship, as given in figure 1.6C, one can estimate the overall size and mass of the particle. For biomolecules it is possible to establish the oligomeric state and even to have a rough estimation of the particle flexibility. Another important derivation of this law is establishing if the species is aggregated or prone to aggregation in a concentration dependent manner. Deviation from this linear relationship, to higher intensities, at low q angles, indicate particle aggregation.

Another important derivation of a typical SAXS curve is the distance distribution function, $P(r)$. As represented in figure 1.6A, it estimates the distance distribution of all distances present in the biomolecule and is analogous to the Patterson map in x-ray crystallography [33, 35]. For a perfect sphere one would expect that the distribution exhibits a gaussian shape, with the maximum of the peak corresponding to the center-of-mass. For more evolved shapes, e.g in figure 1.6A for a dumbbell shape in red, this visual interpretation could be impaired. Nonetheless it is trivial to observe that, for the dumbbell case, one observes two peaks; one corresponding to the distances inside the spheres and the other to the inter-sphere distances. If one establishes a parallelism with a protein structure, with each sphere representing a domain connected by a flexible linker, the inter-domain distance could be easily estimated from this very simple but yet tremendously powerful analysis. Calculating the $P(r)$ also determines the maximum distance, $Dmax$, of the particle by retrieving the maximum value where the $P(r)$ equals zero.

The $P(r)$ is derived from the intensity, $I(q)$, by an inverse Fourier transformation to obtain the real space representation as seen in equation 1.13. This transformation changes the angle dependence, of the reciprocal space, to a distance

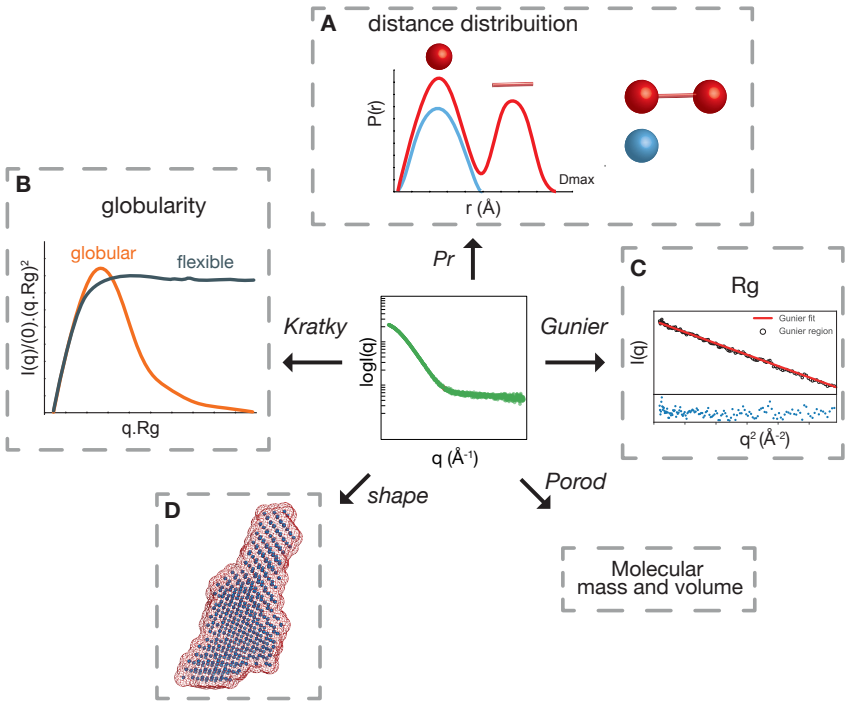


FIGURE 1.6: SAXS data analysis. **A**: Distance distribution, P(r), of a SAXS curve where r is the the distance in Angstroms and Dmax is the maximum distance observed. In blue is represented the P(r) for a sphere and in red for a dumbbell (two spheres connected by a linker). **B**: Kratky plot. Rg is the radius of gyration and I(0) is the intensity extrapolated for q=0. **C**: Gunier plot for retrieval of the radius of gyration Rg. The red line is a linear fit for $0 < q.Rg < 1.3$, the denominated Gunier region. The residuals for the fit are below in blue. **D**: Envelope shape for a SAXS curve in red. In blue are represented the dummy atoms used for calculation of the low resolution envelope.

dependence r, to the real space. The integral over all q is truncated between $q = 0$ and $q = Dmax$ to give the final $P(r)$ distribution see in 1.13 and figure 1.6A. To obtain this distribution is not trivial, especially for bad quality experimental data. The $Dmax$ is varied and the $P(r)$ distribution is back calculated from equation 1.12 and fitted back to the experimental data, until a reasonable convergence is obtained [35, 39].

$$I(q) = 4\pi \int_0^{Dmax} P(r) \cdot \frac{\sin(qr)}{qr} dr \tag{1.12}$$

$$P(r) = \frac{2r^2}{\pi} \int_0^{\infty} q^2 I(q) \frac{\sin(qr)}{qr} dq \tag{1.13}$$

The Rg from the $P(r)$ distribution, as seen in equation 1.14, is obtained by integrating with r^2 over r and is determined using all experimental data and not just the small angles, as in the Guinier approximation [10]. Discrepancies between the Rg and $I(0)$ determined from Guinier's law or the $P(r)$ distribution could indicate protein aggregation and poor buffer matching.

The volume, V, of the particle can also be determined by using equations 1.15 and 1.16, where Q is the Porod invariant.

$$V = 2\pi^2 \frac{I(0)}{Q} \tag{1.15}$$

where,

$$Q = \int_0^{\infty} q^2 I(q) dq \tag{1.16}$$

For globular particles the calculated volume is around two times the molecular mass. The Porod invariant is calculated by integrating the area under the curve of the Kratky plot as represented in figure 1.6B. This calculation is valid for globular particles due to the convergence to zero, at higher q angles; for flexible particles the calculation of the Porod invariant is impaired by the ill-defined area under the curve (see flexible case in figure 1.6B). As a consequence of the latter, the Kratky plot can also be used to empirically estimate the biomolecule flexibility by plotting $q \cdot Rg$ vs $I(q)/I(0) \cdot (q \cdot Rg)^2$. The Rg normalization makes the system size independent allowing shape and flexibility to be assessed [40].

A typical biomolecular SAXS experiment usually follows the following steps:

1. Before measurement, assess the biomolecule's aggregation propensities and optimize experimental procedures to enforce monodispersity.

[10] The radius of gyration determined from the P(r) distribution is the following:

$$Rg^2 = \frac{\frac{1}{2}\int_0^{Dmax} P(r) r^2 dr}{\int_0^{Dmax} P(r) dr} \tag{1.14}$$

2. Measure the intensities for a standard with known molecular weight Mw. This standard can be water or a biomolecule. Record the $I(0)$, from the standard, for molecular mass determination [11].

3. Acquire solution and solvent intensities and subtract solvent intensities to obtain the final SAXS profile (see 1.5B) [12].

4. Determine Rg, $I(0)$, Mw and V and determine oligomerization state, flexibility and aggregation propensities for the biomolecule.

5. Calculate the ab-initio envelope for globular particles as seen in figure 1.6D [43].

6. Fit prior theoretical models or envelopes to the SAXS profiles, and retrieve three dimensional information. [13]

SAXS in the last two decades has matured into a much used and user-friendly technique. Nowadays the majority of experiments are performed in synchrotrons with automatic data acquisition and analysis [44]. All the steps referred before are routinely performed even without expert assistance. Current software development has also streamlined the data acquisition process [39]. Future developments will include further integrations with other techniques (e.g. Cryo-EM, NMR, mass spectrometry) and further hardware development.

1.2.2 Nuclear magnetic resonance spectroscopy in structural biology

Nuclear magnetic resonance spectroscopy (NMR) is a relativity recent technique in structural biology, with the first protein structure being published in the mid eighties [45], more than fifty years after the first measurement of the nuclear magnetic moment in 1938 [46]. Being an inherently quantum phenomenon its full description can only be attained by a full quantum mechanics description. Nonetheless for applications to structural biology and its usefulness in gathering biological information, a classical vector geometry approach will suffice, for the most part, without sacrificing rigour. In this classical approach atoms are treated as magnets inside an external magnetic field, the NMR spectrometer.

The nuclear magnetic dipole moment $\vec{\mu}$, referred earlier, comes from the spin angular momentum , $\vec{S}$, of the nucleus as presented in equation 1.17. The spin quantum number, $\vec{I}$, is discrete and for common nuclei in biomolecular NMR (e.g ^{1}H, ^{13}C and ^{15}N), it has two allowed states, $I = 1/2$ and $I = -1/2$ possessing

[11] Besides calculating the molecular mass by Porod analysis it can also be calculated by comparison with a standard or by the volume of correlation V_c [41]. All methods can be compared to estimate molecular weight convergence [42].

[12] Usually for the momentum vector interval: $0 < q < 0.5$.

[13] How to obtain these models will be further discussed in chapter 1.2.3.

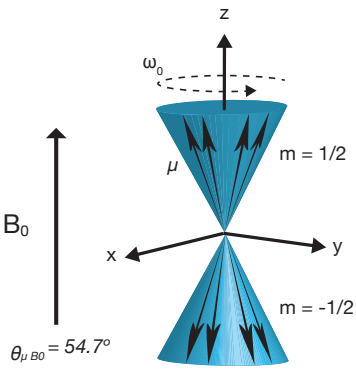


FIGURE 1.7: Magnetic dipole orientation and precession, for spin $1/2$, in the presence of an external magnetic field B_0, parallel to the zz axis. The precession angle between the magnetic dipole moment and B_0 is 54.7°. Larmor frequency is ω_0, μ is the magnetic dipole moment and m is the magnetic quantum number.

high and low energies, respectively. Using a simple analogy these states can be compared to the magnetic poles of a bar magnet spinning about the zz axis.

$$\vec{S} = \hbar\sqrt{I(I+1)} \tag{1.17}$$

The proportionally constant between the nuclear magnetic moment and the spin angular momentum is the nuclear gyromagnetic ratio, γ, as depicted in equation 1.18, and it is an intrinsic property of each nucleus.

$$\vec{\mu} = \gamma\vec{I} \tag{1.18}$$

When placed in an external magnetic field the nucleus will precess around the magnetic field, $\vec{B_0}$, due to the torque generated by the interaction with the nuclear moment $\vec{\mu}$. The splitting of states, as seen in figure 1.7, is called Zeeman splitting and for the allowed values of $\vec{I}$ the energy is given by equation 1.20. The precession frequency, for each state, is denominated Larmor frequency, ω_0, and is also a characteristic of each nucleus. The Larmor frequency is the product of the gyromagnetic ratio, γ, with the magnetic field $\vec{B_0}$ (see equation 1.19)

$$\frac{d\mu}{dt} = \gamma\vec{B_0} = \omega_0 \tag{1.19}$$

$$\Delta E = \hbar\gamma\vec{B_0} \tag{1.20}$$

The Larmor frequency refers to the resonante frequency of each nucleus and

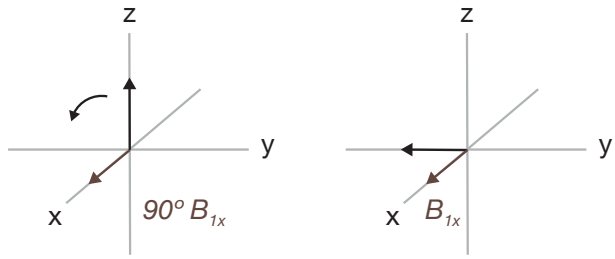


FIGURE 1.8: Radio frequency pulse along the xx axis in the rotating frame, where $90°B_{1x}$ is a 90° pulse. The black vector is the net magnetization. The transformation follows the right hand rule.

states that the absorption frequency for each transition, as seen in figure 1.7, depends on the strength of the applied field and the gyromagnetic ratio of the nuclei. This is a fundamental result in NMR and it can be derived that for an NMR signal to be detected, with a high signal-to-noise ratio, it is advantageous to have high-field spectrometers and high γ nuclei (e.g ^{1}H) [14].

Due to fast speed of precession it is advantageous to define a change of coordinates at the precession axis, the rotating frame transformation. This new coordinates, x and y will be centerd at the laboratory frame and rotating with the nuclear spin precession about the zz axis. For an NMR signal to be recorded the net magnetization has to shift, from its zz axis linearity, to the $x-y$ plane. The latter is achieved by applying a radio frequency pulse along an axis (e.g. the xx axis as depicted in figure 1.8), to shift the magnetization, splitting the energy states as referred earlier. This creates a new magnetic local field, $\vec{B}_1$, with the duration of the pulse determined by the angle of the rotation. From simple trigonometric relations, as seen in equation 1.21, the net magnetization would end on the $-yy$ axis, while applying the 90° pulse, with $\alpha = \pi/2$ as also seen in figure 1.8.

$$I_z \overset{\alpha I_x}{\rightarrow} I_x \cos(\alpha) - I_y \sin(\alpha) \tag{1.21}$$

After pulse termination the net magnetization returns to equilibrium, precessing about the zz axis. Recording this magnetization decay produces the free induction decay (FID). More evolved pulse sequences are needed to obtain relevant structural information for biomolecules, usually consisting of various pulse trains, for which the simplest case was stated.

Having a basic description of how NMR signals arise, three major nuclei properties can me measured and are the most important in biomolecular NMR spectroscopy.

[14]The population differences between the two Zeeman levels is only on the order of 1 to 10^5, for a 11.7 T magnetic field. NMR spectroscopy is a relatively insensitive technique [47].

J-coupling

Magnetization can be transferred through bonds connecting atoms. Each bond in a protein has a characteristic resonant frequency; applying a pulse at this frequency transfers magnetization between atoms, and a NMR signal can be detected. This type of magnetization transfer, involving spin-spin coupling, is called scalar or J-coupling. The scalar coupling is the isotropic part of the J-coupling, a 3×3 tensor. The magnitude of the coupling is measured by the scalar coupling constant, ${}^{n}J$ in which n designates the number of covalent bonds separating the two nuclei. The most useful J-coupling is the three bond,${}^{3}J$, because it is related with protein dihedral angles by the Karplus equation 1.22,

$$ {}^{n}J(\phi) = A\cos^2\phi + B\cos\,\phi + C \tag{1.22}$$

The A, B, and C parameters are empirically-derived, whose values depend on the atoms and substituents involved, and are calculated by studying conformationally restricted small molecules, *ab initio* calculations or protein structures derived from x-ray crystallography. The dihedral angles that were calculated from ${}^{3}J$-couplings, measured in proteins, can be compared with other dihedral angles, for restraints in structural calculations, analyzing a dataset of known protein structures. A range of NMR experiments are available for measuring protein ${}^{3}J$-couplings (e.g. HNHA for ${}^{3}J_{H^N H_\alpha}$).

The Karplus equation 1.22 describes the correlation between ${}^{n}J$-coupling constants and dihedral torsion angles formed by one, two or three bonds. As a reference, in equation 1.23 are presented the ${}^{3}J$-coupling constants for several atom pairs with the corresponding graphical representation given in figure 1.9.

$$\begin{aligned}
{}^{3}J_{H^N H^\alpha} &= 6.51\cos^2(\phi-60) - 1.76\cos(\phi-60) + 1.60\\
{}^{3}J_{COH^\alpha} &= 3.72\cos^2(\phi+120) - 2.18\cos(\phi-120) + 1.28\\
{}^{3}J_{H^N CO} &= 4.29\cos^2(\phi\pm180) - 1.01\cos(\phi\pm180)\\
{}^{3}J_{H^N C^\beta} &= 3.06\cos^2(\phi-60) - 0.74\cos(\phi-60) + 0.13
\end{aligned} \tag{1.23}$$

By calculating ${}^{3}J$-couplings one could obtain the protein backbone torsion angles, as this are the J-couplings that can give the most valuable structural information in proteins. As it depends on the torsion angle ϕ, it tends to be large, 8-12 Hz, in beta-sheet structures, and small, between 3 and 5 Hz, in alfa-helices. For small molecules it can be measured directly by the splitting of the resonances of interest, but for larger biomolecules, which possess a much larger number of resonances, this can be problematic due to the poor signal-to-noise ratio without isotopic labelling (e.g. ^{13}C and ^{15}N) and signal overlap. This obstacle has been overcome with the recently developed isotopic labelling techniques started in the late eighties, and subsequently the development of segmented labelling strategies, aimed at large biomolecules [48, 49].

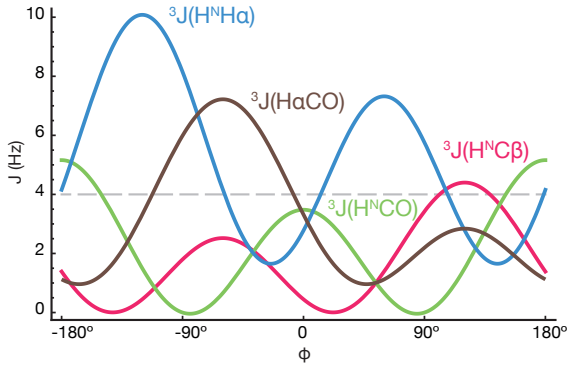


FIGURE 1.9: Representation of the parametrized curves of some backbone ^{3}J-couplings, showing the dependence with the ϕ dihedral angle. The grey dashed line is a J-couping of 4 Hz [47, 48].

By looking at the graphic depicted in figure 1.9, one can see that for one ^3J-coupling (e.g 4Hz) there are four possible angles ($\phi \equiv -180°, -60°, 20°, 110°$), for the $^3J_{H^N H_\alpha}$ coupling, so to retrieve an unequivocal solution one must look at the $^3J_{H^N C^\beta}$ coupling that has a maximum around 2.5 Hz, that yields a $\phi \equiv$ -60°. This points out the important aspect of measuring distinct ^3J-couplings, between different atoms, to give accurate dihedral angles as valuable distance restraints. This variability can also reflect dynamic aspects of the considered bond (in the latter case the HN-Cα bond).

The other dihedral angle ψ is not so amenable to this kind of treatment as its ^3J-couplings are of a small magnitude and prone to errors in the subsequent calculations. The side-chain angles χ_1 and χ_2 also have ^3J-couplings, but are much more difficult to calculate and there aren't many experiments that can measure them and some Karplus curves were not parametrized.

Chemical shift

Besides Zeeman splitting, nuclei experience other interactions that are dependent on the chemical environment around each nucleus. The local magnetic field slightly differs from the applied field B_0, due to the shielding effect of the electronic cloud surrounding each nucleus. This produces a local magnetic field B, by the precessing electrons, given by equation 1.24,

$$B = (1 - \sigma) \cdot B_0 \tag{1.24}$$

where σ is the shielding factor, an anisotropic quantity describing the electron density around the nucleus, described in three dimensional space by the tensor in equation 1.25 [48],

$$\sigma = \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{bmatrix} = \frac{1}{3}[\sigma_{xx} + \sigma_{yy} + \sigma_{zz}] \tag{1.25}$$

where σ_{xx}, σ_{yy} and σ_{zz} are the magnetic fields along the xx, yy and zz axis respectively. As the molecules are freely tumbling in solution the tensor can be averaged to the right side term in equation 1.25.

The observed frequency, due to the local magnetic field B, is called the chemical shift, δ, and is characteristic of each nuclei and chemical environment. To remove the dependency of the applied field, B_0, the chemical shift is measured against a reference compound that do not have any protons, being converted to a dimensionless scale [15], given by equation 1.26,

$$\delta = \frac{\nu - \nu_0}{\nu_0} \cdot 10^6 \tag{1.26}$$

where ν is local frequency an ν_0 is the frequency of the reference compound. The chemical shift is a very important quantity in the structural biology of biomolecules. Being dependent on the local environment it gives us information about the local molecular structure. Secondary structure elements, namely in proteins, have characteristic chemical shift patterns that can be used as restraints in structural determination and for defining structural ensembles of IDPs and multi-domain proteins. Parsing databases of protein dihedral angles with known chemical shifts and comparing them with experimental ones, is the most used method for using chemical shifts in structural bimolecular calculations [50]. Chemical shifts can also be used to map protein-protein interactions [51] and post-translational modifications [52].

The Nuclear Overhauser Effect

The interaction between two magnetic dipoles is called dipolar coupling and it depends on the inverse third power of the inter-nuclear distance (see equations 1.27 and 1.28). In an isotropic solution these couplings average to zero because of rotational diffusion, but if it is considered the effect of dipolar coupling on nuclear spin relaxation the outcome is quite different producing a measurable quantity, the NOE. Atomic nuclei can relax through many mechanisms, one of the most important is dipolar relaxation. This can be seen as an effect of a local magnetic field, generated by a moving nuclei, into a nearby nuclei that produces changes in the overall net magnetization, that returns to equilibrium after this effect ceases. As a pictorial view, the moving nuclei can be seen as a local radio-frequency pulse that can relax other nuclei or it can also be relaxed by the same nuclei, in other words, a spin S relaxes spin I and also spin I relaxes spin S. This effect is called cross-relaxation σ_{IS}, and is stated in equation 1.27.

[15] Normally given in ppm (parts per million).

$$\sigma_{IS} = \frac{1}{10} K^2 \tau_c \left(\frac{6}{1 + (\omega_I + \omega_S)^2 \tau_c^2} - \frac{1}{1 + (\omega_I - \omega_S)^2 \tau_c^2} \right) \tag{1.27}$$

and,

$$K = (\mu_0 / 4\pi) \gamma_1 \gamma_2 / r_{IS}^3 \tag{1.28}$$

The quantity r_{IS} is the internuclear distance, $\mu_0/4\pi$ is a scaling factor for converting into appropriate units, γ_I and γ_S are the gyromagnetic ratios of the I and S nuclei, ω_I and ω_S are the Larmor frequencies and τ_c^2 is the correlation time [16]. So the cross-relaxation depends on the nuclei involved, the correlation time and the distance.

If we consider that the protein tumbles isotropically and that I=S=^1H, then the Larmor frequencies for the I and S spins are identical and the correlation time would be in the order of 10 ns (for a 600 MHz spectrometer). As a consequence the first term in brackets in equation 1.27 would become negligible and the equation 1.27 could be simplified to equation 1.29.

$$\sigma_{IS} = \gamma_1 \gamma_2 r_{IS}^{-6} \tau_c \tag{1.29}$$

The previously mentioned action of a local magnetic field on another nuclei is nothing more than the transfer of spin polarization, i.e. the degree of alignment of the nuclear spin with the applied magnetic field B_0. The Nuclear Overhauser Effect (NOE) is the transference of spin-polarization (i.e. magnetization) by a cross-relaxation mechanism.

This transference of spin-polarization is distance depended, being weaker at longer distances and stronger at shorter ones and it is also transferred by space, as opposed to J-coupling that uses a through-bond correlation.

If one is interested in determining protein structures using NMR spectroscopy, the NOE information is one of the most important sources of distance restraints. If an atomic nuclei has a NOE with another nuclei it is within a certain distance, because the rate of relaxation that produces the NOE varies with the inverse sixth power of the internuclear distance (see equation 1.29), so its intensity decays very rapidly with increasing distance and is only observed, in general, for protons separated up to 5-6 Å.

One of the bottlenecks in determining NOE's as restraints in biomolecular NMR spectroscopy is optimizing the mixing time (i.e. time that the magnetization is transfered between nuclei). For shorter times there is no NOE buildup, if the time is too large other phenomenon's [17], other than cross-relaxation could be responsible for the magnetization transfer, invalidating the inverse sixth power of the inter-nuclear distance proportionality.

[16] The correlation time is the time that the IS vector takes to move by one radian.

[17] One example could be the transfer of magnetization through the molecule by a diffusive process (spin-diffusion).

The NOE is an extremely valuable tool for transferring through-space magnetization, in biomolecular NMR, and can be measured by Nuclear Overhauser Effect Spectroscopy (NOESY) [47–49].

NMR experiments in structural biology

One of the most useful NMR experiments is the 2D HSQC (Heteronuclear Single Quantum Coherence) that correlates the hydrogen (H) and nitrogen (N) dimensions. Each frequency pair represents a protein residue [18] and it can be used to map, at an atomic level, residues involved in protein interactions and also to assess protein folding and flexibility propensities. A change in the chemical environment would shift the NH pair frequency. It starts by applying a 90° pulse in the proton channel, shifting the two populations, [19] of the proton magnetization, from the zz axis (H_z) to the $-yy$ axis ($-H_y$) [20] as depicted in figures 1.8 and 1.10. The 1/4J term reflects the evolution time for the $^1J_{NH}$ coupling, given schematically in figure 1.10 by the grey arrows in $-H_y$. After the first pulse two simultaneous 180° pulses, in both nitrogen and hydrogen channels, are applied to refocus magnetization evolution due to ^{1}H chemical shift and $^1J_{NH}$ evolution, transferring magnetization to ^{15}N. The net magnetization is thus $-H_xN_z$, and after two simultaneous 90° pulses it is shifted to $-N_yH_z$. The previous segment is called an INEPT [21] pulse sequence and it is one of the most important building blocks in biomolecular NMR spectroscopy. The net magnetization is thus set to evolve, during the t_1 delay, due to ^{15}N and $^1J_{NH}$ evolution. The 180° pulse between the two INEPT sequences is to preclude the chemical shift evolution of ^{1}H. After the 180° pulse, the magnetization is at $-N_yH_z$ and a reverse INEPT sequence transfers back the magnetization to ^{1}H to the transversal plane, at H_x. The system is thus ready for acquisition in the ^{1}H channel with a decoupling sequence being applied to convert the $^1J_{NH}$ doublets to singlets [53, 54].

The HSQC experiment explores the high $^1J_{NH}$ coupling of around 90 Hz as depicted in figure 1.11A. As a pictoric analogy we can say that the hydrogen signal amplitude changes according to the frequency of the nitrogen atom by amplitude modulation, according to the time t_1. The signals are subsequently 2D Fourier transformed to give the final HSQC spectrum.

Building upon the 2D HSQC experiment, higher dimensional experiments are performed to attribute the NMR signals to the biomolecule sequence. Instead of the $^1J_{NH}$ coupling multiple J-couplings are explored to assign frequencies to the remaining atoms in a protein backbone as in figure 1.11A.

[18] Also residues possessing nitrogenated side-chains (e.g asparagine and glutamine) appear on the spectrum. Prolines and the N-terminal residue usually don't appear in the spectrum.

[19] Recall that the populations are the ones described in figure 1.7.

[20] The notation follows the product operator formalism described in more advance NMR texts [47, 48, 53].

[21] INEPT stands for Insensitive nuclei enhanced by polarization transfer and essentially takes advantage of the higher gyromagnetic ratio of ^{1}H to transfer magnetization to ^{15}N.

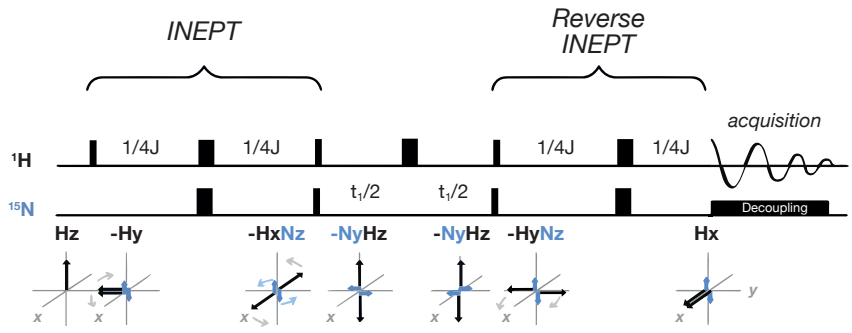


FIGURE 1.10: Pulse sequence for a protein HSQC experiment. J is the J-coupling, t_1 is the delay and H and N are hydrogen and nitrogen respectively. Narrower and wider black bars are 90° and 180° pulses, respectively. This figure was based on a schema from [53].

Two of the most common used higher dimensional experiments are the 3D CBCANH and CBCA(CO)NH experiments. They transfer the magnetization via the strong ${}^1J_{H^\beta C_\beta}$ and ${}^1J_{H^\alpha C_\alpha}$ couplings (see figure 1.11A). The magnetization is then transferred to the C_α and C_β atoms and then to the C_α as shown in figure 1.11A. For the CBCANH experiment the previous step is allowed to evolve for the same (i) and previous residue ($i-1$), exploring the weak ${}^2J_{C_\alpha N^H}$ coupling, and then transfer back from the N to the H^N for acquisition. For the CBCA(CO)NH experiment the magnetization only evolves for the previous residue, exploring the ${}^1J_{CON^H}$ coupling, followed by the same acquisition mode as in the CBCANH case. Extrapolating from the HSQC we have thus a 3D experiment represented in a cube where the $x-y$ plane is the N-H frequency pair and the third dimension corresponds to the carbon chemical shifts as represented in figure 1.11B. Extracting planes perpendicular to the ^{15}N axis allows the connection of adjacent $i-1$ and i residues. Joining this information with the residue specific dispersion patterns of carbon chemical shifts, we can assign in a sequence specific manner the residues in a HSQC experiment, paving the way for subsequent NMR studies [55]. This higher dimensional experiments can be combined with NOE pulse sequences, and other dimensions, to achieve an even higher dimensional space, aimed at reducing the inherent chemical shift degeneracy.

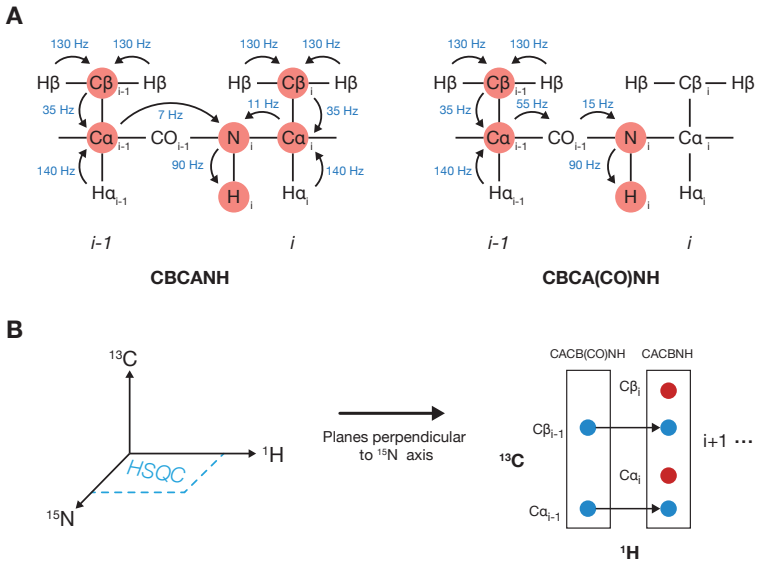


FIGURE 1.11: Protein J-couplings and 3D experiments. **A**: In orange are the protein backbone J-couplings explored in the CBCANH and CBCA(CO)NH 3D experiments. The *i-1* is the previous and *i* the same residue, respectively. **B**: Schematic representation of a backbone NMR 3D experiment.

1.2.3 Ensemble determination, for flexible proteins, derived from experimental data

Proteins are not rocks. Since solving the first protein structure, myoglobin in 1958, the structural biology field has been gradually employing, in its lexicon, flexibility to describe protein structures. Proteins are intrinsically flexible systems, the magnitude of which is directly connected to its function whereas by side-chain rotations, loop flexibility and even whole domain rearrangements. More recently intrinsically disordered proteins (IDP), proteins with no defined structure, were shown to be highly abundant in the human proteome among proteins involved in disease [56]. It is also estimated that about 44% of the human proteome contains intrinsically disordered segments of more than thirty residues [57]. It is thus urgent to develop methods that can describe this conformational variability.

Departing from the more classical approaches, using x-ray crystallography, determining ensembles for flexible proteins usually use a plethora of solution experimental techniques, often in combination. Nuclear magnetic resonance spectroscopy usually provides information on secondary structure while SAXS and Foster resonance energy transfer (FRET) usually determine particle overall size

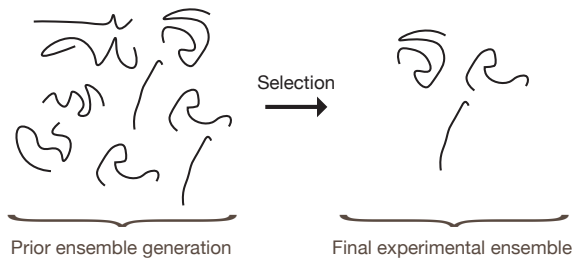


FIGURE 1.12: Ensemble selection from flexible biomolecules. General strategy for experimental ensemble determination. Selection represents any experimental technique and/or algorithm.

and shape [58].

As seen in figure 1.12, ensemble determination usually starts with a pool of random conformations, sampling the highest conformational space possible, after which a selection method is applied, to recover a subset that agrees best with the experimental data. Random conformations are generated by simple algorithms that usually rely on pre-generated databases of flexible loops, derived from the protein data bank (PDB) [59, 60] and combined using Monte-Carlo optimization techniques [61, 62]. Due to the broad conformational space accessible to flexible proteins, classical molecular dynamics approaches [22] are not commonly used to generate conformations, because of the high computational cost and low conformational variability. Another aspect impairing their use is the poor definition of protein-water interactions, usually leading to an artificially protein compaction [63].

After generating the conformational space it is needed to calculate the energy of the system and the most probable configurations. Conceptually this is performed by evaluating the energy according to equation 1.30,

$$E_{total} = E_{physical} + \omega E_{data} \tag{1.30}$$

where E_{total} is the total energy, $E_{physical}$ is the physical energy given by a force field imploying stereo-chemical restraints and E_{data} is the deviation from the models with respect to the experimental data, weighted by a factor ω. Using SAXS as an example, the intensities I_k would be calculated for each configuration k each with a weight v_k, [23] as in equation 4.8. The fitting to the experimental data would be evaluated using the chi squared distribution χ^2 [65]. In equation 1.32 the I_{exp} is the experimental intensity, I_{calc} the calculated intensity, σ the experimental errors and c a scaling factor.

[22] For a more in-depth analysis, on the foundations of molecular dynamics, please refer to chapter 1.2.4.

[23] Using, for example CRYSOL [43] or FOXS [64].

$$I(q) = \sum_k v_k I_k(q) \tag{1.31}$$

$$\chi^2 = \frac{1}{N_p - 1} \sum_i \left[\frac{I(q_i)_{exp} - cI(q_i)_{calc}}{\sigma(q_i)^2} \right]^2 \tag{1.32}$$

Methods that try to determine protein ensemble distributions usually minimize the χ^2 while trying to find the v_k distribution of weights for all accepted configurations; this is the selection step in figure 1.12. Other restraints, besides SAXS, could be derived from NMR, FRET, x-ray, cryo-EM or other experimental techniques for which back-calculation algorithms exist, for comparison with experimental data. Selecting the conformations usually employs genetic algorithms where the initial conformations are divided in chromosomes, and an evolution based on genetic traits is evaluated according to the experimental data [66, 67]. Another increasingly popular method is bayesian analysis, where the final ensemble is described as a posterior probability distribution, according to Bayes' law in equation 1.33 [68, 69].

$$P(E|D) = \frac{P(D|E)P(E)}{P(D)} \tag{1.33}$$

As defined in equation 1.33, $P(E|D)$ is the probability of the ensemble given the data D. The prior distribution is $P(E)$, meaning, the initial probability of the ensemble before the data [24]. The likelihood is $P(D|A)/P(D)$ and could be defined as the support D gives to E [69, 70]. Here the theoretical and experimental errors are defined explicitly, as probability distributions, exerting an advantage over genetic methods that define them empirically.

Hybrid methods for structural determination are an active topic of research, given the recent emergence of IDP's and flexible proteins as targets for disease. Steady improvements need to be effected; namely improving force field performance and increasing computational power that can also benefit "traditional" protein structure determination [63].

1.2.4 Molecular dynamics simulations

Understanding protein function, at an atomic level, is key for understanding biological phenomena. The importance of molecular dynamics and its contribution in helping to describe how proteins work, was recognised by the Nobel prize award in chemistry in 2013.

The atomic level knowledge described previously, to a full extent, would require a complete description of the electronic and atomic structure of a particle. From first principles this system would be described by the time-dependent

[24] Usually the prior models are scored according to a force field as depicted in chapter 1.2.4.

Schrödinger equation [71] (see equation 1.34), that is the fundamental equation in quantum mechanics.

$$i\hbar\frac{d}{dt}\Psi(r,t) = H\Psi(r,t) \tag{1.34}$$

For a solvated biomolecular system it is computationally intractable, nowadays, to fully describe it, using the wave function $\Psi(r,t)$, describing the system probabilistically as a function of time t and position r, where the Hamiltonian operator H acts on the total potential energy (see equation 1.34). To overcome these bottlenecks several approximations have to be undertaken to make the system more tractable and less computationally expensive. The three most important ones relate to approximating atoms movement and energy evaluation. The first one is the Born-Oppenheimer approximation, where the electronic movement is discarded due to its much higher speed, when compared to the nuclei. The second refers to treating the atom nuclei trajectories according to Newton's laws. Finally, the third approximation acts on simplifying the potential energy $V(r)$ to a semi-empirical one, usually encompassing experimental parameters derived from small molecules or from ab initio quantum chemical calculations and containing the most common interaction types found in nature; the so-called force field [35, 72] (see equation 1.35 [25]).

$$V(r) = \overbrace{\sum_{\text{bonds}} k_b(l-l_0)^2 + \sum_{\text{angles}} k_a(\theta-\theta_0)^2 + \sum_{\text{torsions}}\sum_{n}\frac{1}{2}V_n[1+\cos(n\omega-\gamma)]}^{\text{bonded terms}} + \underbrace{\sum_{\text{nonbonded}}\sum_{i,j} f_{ij}\left\{\underbrace{\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{\text{Lennard-Jones}} + \underbrace{\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}}_{\text{Coulomb}}\right\}}_{\text{nonbonded terms}} \tag{1.35}$$

In a force field the potential energy is approximated to a sum of bonded and nonbonded terms with simple and already known physics.

For bonded terms; the bonds term represents every covalently linked atom. Two covalent bonds, shared between three atoms correspond to the angles potential term. These two terms obey to Hooke's law [26], where k_b and k_a are the force constants. Two angles sharing a common bond, between four atoms, also known as a dihedral are described by the torsion term as seen in figure 1.13 and equation 1.35.

[25] The force field in equation 1.35 is based in the AMBER force field [73], one of the most used. CHARMM [74] and OPLS [75] are also other commonly used force fields.

[26] Here the bond behaves as a harmonic oscillator between the two atoms.

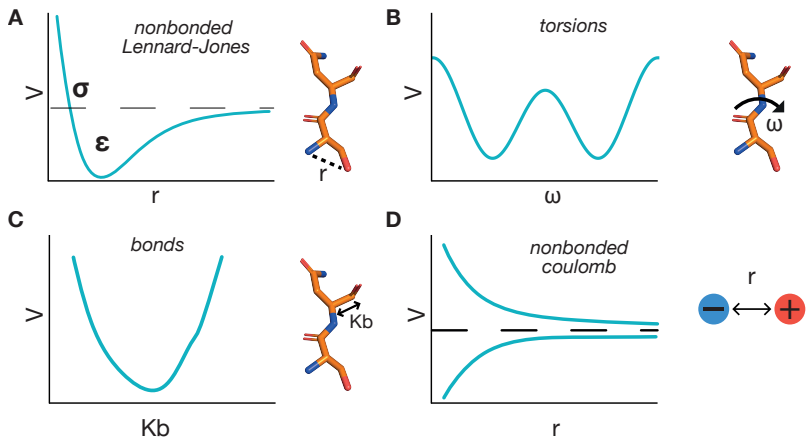


FIGURE 1.13: General terms represented in a force field. **A**: Lennard-Jones potential. r_0 is the Van der Walls radii, r is the inter-atomic distance and ϵ is the depth of the potential well. **B**: Torsion potential. ω is a dihedral angle. **C**: Bonds potential. K_b is the force constant according to Hooke's law. **D**: Coulomb's law for interactions of atomic point charges where r is the inter-charge distance. V in all panels is the potential energy.

For nonbonded terms, usually including all atom pairs separated by three bonds or more; the first member describes Van der Walls forces by the Lennard-Jones potential where the first term corresponds to the repulsive part describing Pauli repulsion and the second term is the attractive part, describing dispersion forces. The Van der Walls radii, for two interacting atom pairs is σ_{ij}, r_{ij} is the inter-atom distance and ϵ_{ij} is the potential well representing the function minimum. For the Coulomb term, $q_i q_j$ are a pair of point charges and r_{ij} is the inter-charge distance, as seen in figure 1.13. The bonds and angles potentials oscillate around an energy minimum and the torsion potential is a periodic function, oscillating between higher and lower energy, usually dependent on atomic steric clashes. The nonbonded terms are more complex functions and have high potential energy, for shorter distances and lower values for larger distances (see figure 1.13C and 1.13D). To a large extent force fields have remained remarkably almost unchanged since the first protein simulation in 1977 [76].

Having a description of the system's energy, one needs to describe its motion.

Knowing the force $F(x)$ (derivative of the potential energy $V(x)$) acting in atom at position x, at time t, it is possible to know the positions, velocity v and acceleration a acting on that atom, allowing the system to advance to the next step. Integration of the equation 1.37, derived from 1.36, yields a time-dependent trajectory. This cycle is repeated, over all atoms of the system, until convergence

of some physical metric or according to user preference, to retrieve the final trajectory.

$$\frac{dv}{dt} = F(x) = -\frac{dV(x)}{dx} = m \cdot a \tag{1.36}$$

$$\begin{aligned} v_i(t + \frac{\Delta t}{2}) &= v_i(t - \frac{\Delta t}{2}) + \frac{F_i(t)}{m_i}\Delta t \\ x_i(t + \Delta t) &= x_i(t) + v_i(t + \frac{\Delta t}{2})\Delta t \end{aligned} \tag{1.37}$$

The Leapfrog algorithm, as depicted in equation 1.37, for integration of Newton's equations of movement (see 1.36) is one of the most used in biomolecular simulations. The system is updated with a timestep Δt [27] to propagate the velocities v, positions x for each atom i with mass m. Having a description of the system's energy and its motion, one can start a molecular dynamics simulation protocol.

A protocol for performing a biomolecular dynamics simulation typically follows the following steps:

- **Starting conformation:** Usually from X-ray crystallography, NMR spectroscopy or homology modelling. To this conformation initial velocities and positions are assigned and also all subsequent iterations.
- **Preparation:** In this step partial charges are assigned to the system, the system is solvated and minimized, to relieve local clashes between atoms.

  Biomolecules, in vivo, are solvated systems with water as the solvent. For a realistic simulation, besides taking into account the protein interactions, protein-water interactions have to be taken into account. Water stabilizes protein folding transition states, mediates hydrogen bond formation, among many other interactions. It does so by lowering the energy barrier, when compared to an in vacuo system. There are many models for water representation and ideally the best model will balance accurate description of the water molecule against the computational tractability of the calculations. Explicit water models range from a simple point charge, the SPC model, to the more computationally expensive TIP5P model [77] that includes a correct representation of the water dimer and the inclusion of the electron lone pairs. A good compromise is the TIP3P model, a three site model where interactions of the three atoms are calculated, excluding the electron lone pair [78]. The latter is one of the most commonly uses in biomolecular simulation.

[27] The timestep should be very small, and below the protein bond stretching timescale; for proteins simulations this value is around 2fs.

Implicit water models are less computationally expensive and less rigourous. Here the solvent is represented by a continuum with a distance dependent dielectric constant, with no individual water molecules. Once a solvent box is defined, periodic boundary conditions [28] are established and the system is ready for equilibration.

- **Equilibration:** An equilibration molecular dynamics generally allows the pressure and/or the temperature to equilibrate to the production simulation value. Normally the starting conformation is positionally restrained and the solvent is allowed to equilibrate around the biomolecule.

- **Production run:** The relevant molecular dynamics simulation where all subsequent analysis will be performed.

Molecular dynamics of biomolecules have come a long way since the first protein simulation [29], in vacuo, in the late seventies [76] with the simulation running for a total time of 9.2 ps [30]. Roughly a decade later a simulation of the same protein, the first of a solvated system in water, ran for 210 ps [31], about an order of magnitude increase from the previous simulation [79]. In 2010, roughly another decade afterwords, the millisecond barrier was crossed with a 1 ms simulation, of the same protein, within a solvated system [80]. This represented an increase of one hundred million in trajectory length. Since then, with the increase in computational power, simulations can reach timescales of hundreds of nanoseconds, for small proteins, in desktop computers with graphical processing units (GPU). For supercomputers whole virus can now be simulated and complete protein folding simulations, in the microsecond to millisecond timescale, are now accessible [81].

[28] For the water molecules to be spatially contained inside the defined solvent box, adjacent boxes are created and water molecules are allowed to transfer between these boxes, maintaining the overall energy constant. All calculations are only made inside the define solvent box and not the adjacent ones [35].

[29] The simulated protein was the bovine pancreatic trypsin inhibitor (BPTI).

[30] Only bond vibrations and maybe some early methyl rotations could be explored.

[31] Methyl group rotations and some side-chain rotamers could be accessible in this timescale.

Chapter 2

Aims and objectives

2.1 State of the art

The work presented in this thesis builds upon previous work involving Smad proteins both at the structural and biological/biochemical levels. Namely, the structures of the MH1 (except Smad2 and the I-Smads) and MH2 domains have been solved. Also several hypothesis were put forward for the oligomeric equilibria of Smad2 and Smad4 and the consequent Smad2/Smad4 complex formation. The previous research was built mainly on individual domains and no structural information exists on the human full-length proteins. Being large multi-domain oligomeric proteins connected by flexible long linkers, the standard protocols used in the biomolecular structural field would not suffice in obtaining a complete picture of the conformational space. We plan to bridge this gap by combining multiple techniques, with method development for data analysis, for attaining the clearest picture possible.

2.2 Objectives

The main objectives of the presented work are the following:

- Establish expression and purification protocols for human full-length proteins and derived protein constructs.
- Acquire structural and biochemical data to establish Smad proteins conformational landscape.
- Develop and/or adapt data analysis methods to achieve the previous item.
- Establish an hypothesis for TGFβ activation, based on the data acquired for the full-length proteins.
- Study the effect, at an atomic-level, that cancer mutations have in Smad protein domains.

Chapter 3

Materials and methods

3.1 Protein cloning and production

3.1.1 Cloning and site-directed mutagenesis

For protein site-directed mutagenesis the parent plasmid was used for each mutation reaction. Each 25 µl reaction contained 100 ng of DNA template, 62.5 ng of sense mutagenic primer, 62.5 ng of antisense mutagenic primer, 2.5 µl of dNTP mix (stock at 2 mM), and 2.5 µl of 10X manufacturer's reaction buffer to which 1 µl of Pfu DNA polymerase (Agilent Technologies, CA, USA) was added. The mutation PCR (Polymerase chain reaction) was performed in a BIO-RAD T100 Thermal Cycler according to the following temperature ramps : T1, 95 °C for 60 s, (1 cycle) followed by T1, 95 °C for 60 s; T2, 50 °C for 1 min (15 cycles); and T3, 68 °C for 12 min. To digest the parental template strain, 1 µl of DpnI restriction enzyme (10 units/µl) was incubated with the reaction for 2 h at 37 °C. After this 5 µl of the digested reaction mix containing the mutated plasmid was transformed into Dh5α cells in SOC media by the heat shock method. The transformation mix was plated in LB agar (buffer in table B.1) and incubated at 37 °C overnight. Clones were isolated, and each variant was sequenced by Sanger-Sequencing, to confirm the presence of the specific mutation.

For protein cloning first the parental plasmid was linearised with the restriction enzymes depicted in table 3.1 for each vector family.

TABLE 3.1: Restriction enzymes used for vector linearisation.

Vector family	Enzymes
pCoofy	BamHI | HindIII
pTEM	NCOI | HindIII
pOPIN	KpnI | HindIII

Subsequently each insert, previously amplified by PCR, was cloned by ligation independent cloning (LIC) using primers that partially overlapped, around 15bp, with both the insert and the parental plasmid. The enzyme that catalyzed the

reaction was a recombinase (New England Biolabs inc.) and the recombination mixture contained 0.25 µl of the enzyme, 0.5 µl of the enzyme buffer, 100 ng of the linearized vector and around 10x the amplified insert, for a total volume of 5 µl. The previous mixture was incubated at 37 °C for 15-30 min without mixing and transformed into Dh5α cells in SOC media and plated in LB agar. Clones were isolated, and each variant was sequenced by Sanger-Sequencing, to confirm the successful insert cloning.

3.1.2 Protein production

For unlabelled protein production all protein constructs were expressed in the E. coli BL21 (DE3) strain. Cells were cultured at 37 °C in Luria-Bertani (LB) medium until an OD600 of 0.6-0.8. After the cultured cells were induced with IPTG (Isopropyl β-D-1-thiogalactopyranoside), at a final concentration of 0.5 mM, and expresse overnight at 20 °C, bacterial cultures were centrifuged at 4000g for 20 min and resuspended in lysis buffer. Cells were lysed using an EmulsiFlex-C5 (Avestin) at 20000 psi and centrifuged at 35000g for 45 min to discard insoluble material. The lysis buffers were protein dependent and are presented in table 3.2. The soluble supernatants were purified by nickel-affinity chromatography (HiTrap Chelating HP column, GE Healthcare Life Science) using a NGC Quest 10 Plus Chromatography System (BIO-RAD). For S4FL a StrepTrap column (GE Healthcare Life Science) was employed. The used gradient, for all constructs, spanned from 0% to 100% elution buffer B in 15 column volumes. The buffers, for streptavidin and nickel-affinity chromatography, were also protein dependent and are presented in table 3.2.

Eluted proteins were digested at 4 °C with the vector specific protease (see table 3.2), and further purified by ion exchange chromatography using a HiTrap SP HP or Q (GE Healthcare) columns and a gradient running from 0% to 100% buffer B ion exchange buffers, as seen in table B.3. As a final purification step size-exclusion chromatography was performed and the column was equilibrated in gel filtration buffer as seen in table B.3.

For the purification of the inter-domain linkers, S2L and S4L, connecting the folded domains a different protocol was used. The proteins were expressed as described above but the lysis and elution were performed in denaturing conditions. After expression and lysis, using the same protocol describe above, the proteins were allowed to bind to a benchtop nickel column and were washed in refolding buffer, as seen in table 3.2. The proteins were refolded, while column bound, using a stepwise approach; four subsequent washes were employed increasing in each wash the ratio of refolding/lysis buffers from zero to four. A final elution step with S2EB buffer was performed after which the proteins were cleaved and further purified by gel filtration chromatography as described above. Protein purity was accessed by SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) gels and mass spectrometry. Protein concentration was performed

TABLE 3.2: Protein constructs properties and buffers. Constructs are the protein constructs, MW are the molecular weights in kilodaltons, aa are the total number of residues, PI is the isoelectric point calculated using the PROTPARAM webserver [82], vector and protease are the ones used for expression and lysis and elution are the lysis and elution buffers, respectively, described in table B.3. RB is the refolding buffer and LB is the lysis buffer in table B.3.

Constructs	MW	aa	PI	Vector	Protease	Lysis	Elution
Smad2							
S2FL	53.32	475	6.4	pETM10	none	LB	S2EB
S2460*	52.48	467	6.3	pETM10	none	LB	S2EB
S2L	12.81	115	4.3	pETM11	TEV	RB	S2EB
S2LMH2	33.08	297	5.6	pETM10	none	LB	S2EB
S2MH1	19.01	169	9.7	pETM11	TEV	LB	S2EB
S2MH1-E3	15.66	139	10.1	pETM11	TEV	LB	S2EB
S2FLEEE	53.44	475	6.2	pETM10	none	LB	S2EB
Smad4							
S4FL	66.44	609	6.6	pCOOFY34	3C	LB	S4EB
S4LMH2	46.87	430	6.2	pETM11	TEV	LB	S4EB
S2SADMH2	34.03	307	6.5	pETM11	TEV	LB	S4EB
S4MH2	29.29	266	6.2	pETM11	TEV	LB	S4EB
S4L	17.22	162	5.7	pETM11	TEV	RB	S4EB
S4MH1	18.15	158	9.5	pTEM11	TEV	LB	S4EB
Smad7							
S7FL	48.57	445	9.5	pOPINF	3C	LB	S2EB
S7FLdel20-85	42.72	380	9.7	pOPINF	3C	LB	S2EB

using Amicon® Ultra Centrifugal Filters with appropriate molecular weight cut-offs.

For isotopically labelled protein expression, the proteins were firstly expressed in LB media until an OD600 of 0.6-0.8. Before induction the cultures were centrifuged at 2000g for 10 min and the resulting pellets were resuspended in minimal media (for buffer recipe see table B.2). After incubating for one hour at 20 °C the cultures were induced with 0.5 mM IPTG and left expressing overnight. All subsequent purification steps were identical to the unlabelled proteins protocol.

3.2 Electrophoretic mobility shift assays

For a total volume of 20 µl, 7.5 nM of 5′-end Cy5-labeled duplex DNA was incubated with increasing amounts of proteins (from 31nM to 2 µM) in binding buffer

(20 mM TRIS pH 7.2, 80 mM NaCl). Binding reactions were carried out for 30 minutes at room temperature. Electrophoresis was performed in non-denaturing 12.0% (19:1) polyacrylamide gels. The gels ran for 1 hour and 20 minutes in 1X TG buffer (25 mM TRIS, pH 8.4, 192mM Glycine) at 90 V at 20 °C. The gels were exposed in a Typhoon imager (GE Healthcare) using a wavelength of 678/694 nm (excitation/emission maximum) for the Cy®5 fluorophore.

3.3 Small angle X-ray scattering

SAXS data was acquired at Beamline 29 (BM29) at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). Protein samples were centrifuged for 10 minutes at 10000g prior to data acquisition. Experiments on BM29 were collected, on 45 µl samples, with the following settings: 12.5 keV, 100% transmission, low viscosity and 0s wait time. Data was recorded on a Pilatus 1M detector, at 10 °C. Ten frames were collected for 1s each, for each sample. Solvent scattering data was collected for each sample to account for buffer contribution. Image conversion to the 1D profile, scaling, buffer subtraction and radiation damage assessment was done by the in-house software pipeline available at the BM29 beamline. Further processing was done by the ATSAS software suite [39] and the Scatter package (`http://www.bioisis.net/scatter`).

3.4 Nuclear magnetic resonance spectroscopy

NMR experiments were acquired in a Bruker AVIII 600MHz spectrometer using a 5mm TXI cryoprobe. All samples were recorded in 40 mM TRIS, 150 mM NaCl, 10% D_2O, pH 6.6. Parameters for acquisition of all experiments are presented in table 3.3. The HSQC experiments were processed using TOPSPIN v3.5 (Bruker), all other experiments were processed using NMRPIPE [83] and analyzed with the CcpNmr Analysis [84] software suite. Acquisition modes were the planes and the Non-Uniform Sampling [1] (NUS) methods for the HSQC and all other experiments respectively. All experiments were recorded using BEST-TROSY [2] pulse sequences.

The NMR backbone spectral assignment strategy followed established protocols by using carbon frequency strips employing CBCANH and CBCA(CO)NH experiments as described in 1.2.2. Due to the highly flexible nature of the linker of Smad4 (S4L) additional, and more recently developed, experiments were also recorded. The HN(COCA)NH and HN(CA)NH set, that sequentially connect backbone amides, aids in reducing spectral overlap analysis due to the higher dispersion of the ^{15}N atom when compared to the ^{1}H. Also HN(CA)CO and HNCO

[1] NUS methods allow for increased resolution and decreased acquisition time, by employing sampling schedules that only record the FID partially [85].

[2] The BEST principle is based in an increased ^{1}H steady-state polarization and TROSY refers to the transverse relaxation-optimized spectroscopy, that selects favourable relaxation properties [86].

TABLE 3.3: Spectra parameters for NMR experiments where NS is the number of scans, T is the temperature, SW is the spectral width and TD is the size of the FID.

Constructs	Experiment	NS	T(°C)	Nuclei	SW(ppm)	TD
S2L	HSQC	8	5	^{1}H	12.01	4096
				^{15}N	25.00	256
S4FL	HSQC	8	5	^{1}H	12.01	4096
				^{15}N	25.00	256
S4MH1	HSQC	8	25	^{1}H	7.00	2048
				^{15}N	34.00	256
S4L	HSQC	8	5	^{1}H	12.01	4096
				^{15}N	25.00	256
	CBCA(CO)NH	32	5	^{1}H	12.00	2048
				^{15}N	36.00	90
				^{13}C	70.00	60
	CBCANH	32	5	^{1}H	12.00	2048
				^{15}N	36.00	90
				^{13}C	70.00	60
	HN(COCA)NH	64	5	^{1}H	12.00	2048
				^{15}N	36.00	80
				^{15}N	36.00	80
	HN(CA)NH	64	5	^{1}H	12.00	2048
				^{15}N	36.00	80
				^{15}N	36.00	80
	HNCO	16	5	^{1}H	12.00	2048
				^{15}N	36.00	60
				^{13}C	20.00	96
	HN(CA)CO	32	5	^{1}H	12.00	2048
				^{15}N	36.00	60
				^{13}C	20.00	96

experiments, that connect backbone carbonyls, were recorded to further reduce spectral overlap degeneracy analysis.

NMR relaxation experiments, ^{15}N T1, T2, and heteronuclear NOE (HetNOE) were acquired at 5ºC. The inversion recovery delays used for T1 relaxation experiment were 20, 110, 160, 270, 430, 540, 700, 860, 1080, 1400, 1720 and 2000 ms. The delays used for the T2 experiment were 0, 20, 40, 60, 80, 120, 160, 200, 280 and 400 ms. The size of the fid for all experiments was (^{1}H)1024 × (^{15}N)256 points and

the interscan delay was set to 3s. Relaxation rates were retrieved by fitting peak intensities to an exponential function implemented in CcpNnmr analysis [84].

Peak movement d in NMR titrations, using HSQC experiments described above, was quantified using equation 3.1,

$$d = \sqrt{\frac{1}{2}\left[\delta_H^2 + (0.15 \cdot \delta_N^2)\right]} \tag{3.1}$$

were δ_H^2 and δ_N^2 are the ^{1}H and ^{15}N chemical shift differences, respectively.

3.5 Protein ensemble generation

3.5.1 Structural modelling of Smad2 and Smad4 linkers

Random coil ensemble models of S2L and S4L containing 10000 conformations each were generated employing Flexible-Meccano (FM) [59], followed by side-chains modelling with SCCOMP [87], and energy-minimization in explicitly solvent using GROMACS 5.1.1 [88]. We used the force field AMBER99sb-ILDN [89] with no ions added, and TIP3P [78] for the water model. We used CRYSOL [43] to compute the theoretical SAXS profiles from conformational ensembles of S2L and S4L. All theoretical curves were obtained with 101 points and a maximum scattering vector of 0.5 Å^{-1} using 25 harmonics. Using the ensemble optimization method (EOM) [66, 67] we selected from the S2L and S4L structural pools, the linker structures whose theoretical SAXS profiles collectively fit their experimental SAXS profiles.

3.5.2 Missing fragment reconstruction for the Smad2 and Smad4 MH1 and MH2 domains

Ensembles of terminal disordered fragments were built using FM and attached to the X-ray templates using in-house scripts. For each built segment side-chains were added using SCCOMP and then pre-processed with Rosetta3.5 $fixbb$ module to alleviate steric clashes. Internal disordered loops were built using the Rosetta software framework using the $rosettaCM$ application [90, 91] outputting 5000 structures with the final average structure being reported and used for subsequent calculations, for the full-length and oligomeric constructs. Previously solved x-ray structures for Smad4 MH1 (PDB:3QSV) and MH2 (PDB:1DD1) domains and Smad2 MH1-E3 (PDB:6H3R) and MH2 (PDB:1KHX) domains, were used as templates. For the S2MH1E3 construct the α-helix between residues 90-98, at the E3 exon, was built with the Modeller package [92], by enforcing secondary structure elements determined by analysing NMR data.

3.5.3 Smad2 and Smad4 full-length ensemble generation

For generating full-length conformations and building upon the results from sections 3.5.1 and 3.5.2, we started by adapting the FM pipeline for use with multi-domain proteins. The linker positions were randomised and 10000 conformations were generated and selected using the EOM method as reported in section 3.5.1. The ensemble optimization method was slightly modified for reporting more robust statistics following previous published results [56, 62]. For Smad4, starting from the 10000 conformation random pool, 200 sub-ensembles of 50 members each were generated and 750 genetic operations were performed, until convergence of the χ^2 metric, fitted to the experimental data (see section 1.2.3). For each cycle a new sub-ensemble was outputted and subjected to further analyses.

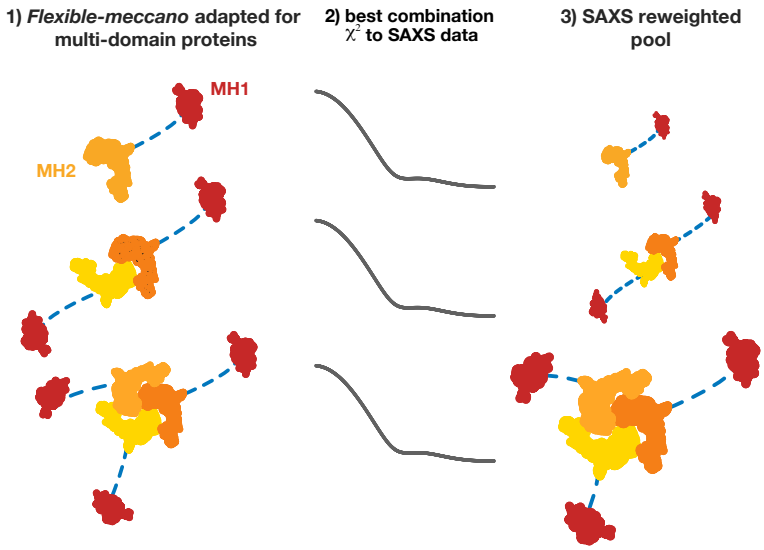


FIGURE 3.1: Smad2 ensemble generation strategy. *Left panel*: Monomer, dimer and trimer schematic representations for the random pool ensemble. The linker of Smad2 is represented in dashed lines together with the MH1 and MH2 domains (colouring scheme same as in figure 4.1). *Middle panel*: Theoretical SAXS profiles for the monomer, dimer and trimer configurations, referred above. *Right panel*: SAXS derived ensemble reweighted from the random pool.

For Smad2 the same strategy was adopted taking into account the oligomeric nature of this protein. The procedure reported above was repeated for each state (monomer, dimer and trimer) with 6000 conformations being generated for each oligomeric state, totalling 18000 conformations. Subsequently and after the EOM

procedure the SAXS reweighted population distributions, for each state, were further analyzed (see figure 3.1).

3.6 Ion mobility mass spectrometry

Ion mobility equipment and a mass spectrometry (IM-MS) is an analytical technique that combines an ion-mobility and a mass spectrometer. While the first separates ions according to their mobility in the gas phase, the second separates them according to their mass-to-charge ratio. In the biomolecular application field, the time that the ionized biomolecules take to move across a drift tube is recorded; the so called drift time [93, 94]. The drift time can be converted to collision cross sections (CCS) if standards with known mass and CCS are used. The rotationally averaged CCS, a property with area units, gives low resolution structural information about the biomolecules in the gas phase and is given by equation 3.2,

$$\Omega = \frac{3e}{16N}\sqrt{\frac{2\pi}{\mu k_B T K}} \tag{3.2}$$

where Ω is the colisional cross section, N is the drift gas number density, μ is the reduced mass of the ion and drifts gas, k_B is the Boltzmann constant, T is the drift gas temperature and K is the mobility of the ion [3].

Ion mobility mass spectrometry experiments were performed using a Synapt G1-HDMS mass spectrometer (Waters, Manchester, UK). Samples were delivered in 150 mM ammonium acetate buffer. They were injected directly and infused by an automated chip-based nanoelectrospray using a Triversa Nanomate system (Advion BioSciences, Ithaca, NY, USA) as the interface. The ionization was performed in positive mode using a spray voltage and a gas pressure of 1.70 kV and 0.5 psi, respectively. Cone voltage, extraction cone and source temperature were set to 40 V, 2 V and 20 °C, respectively. Trap and transfer collision energies were set to 10 V and 10 V, respectively. The pressure in the Trap and Transfer T-Wave regions were 5.84·10-2 mbar of Ar and the pressure in the IMS T-Wave was 0.460 mbar of N2. Trap gas and IMS gas flows were 8 and 24 mL/sec, respectively. The travelling wave used in the IMS T-Wave for mobility separation was operated at a velocity of 300 m/sec. The wave amplitude was fixed to 10 V. The bias voltage for entering in the T-wave cell was 15 V. The instrument was calibrated over the m/z range 500-8000 Da using a solution of cesium iodide. MassLynx version 4.1 SCN 704 and Drift scope version 2.4 softwares were used for data processing. Ion mobility data analysis was performed with Driftscope software vs. 2.5 integrated in MassLynx software.

[3] The mobility K is given by $\nu = KE$, where ν is the ion velocity and E is the electric field.

3.7 Differential scanning fluorimetry

Differential scanning fluorimetry (DSF), also denominated thermofluor, is a technique that allows for protein stability determination by calculating its melting temperature (Tm). In general terms a protein is incubated with a fluorescent dye that when in contact with the hydrophobic core of the protein, emits radiation [95–97]. The hydrophobic core is exposed usually by a temperature gradient until the protein is unfolded as depicted in figure 3.2.

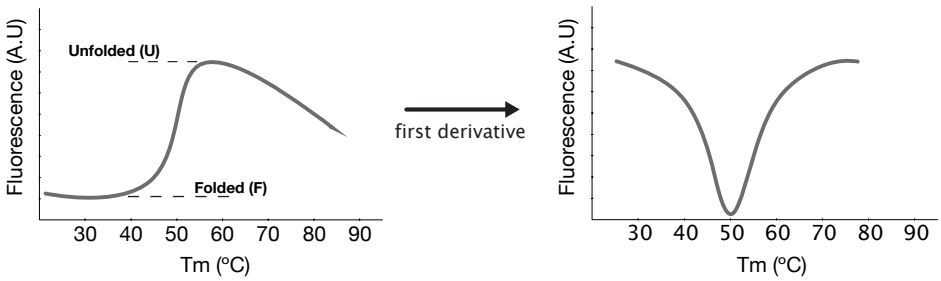


FIGURE 3.2: Differential scanning fluorimetry profile. Fluorescence is represented in arbitrary units (A.U.) and U and F are the unfolded and folded fractions respectively. The right panel is the first derivative with the Tm being the minimum of the function.

The Tm can be extracted by calculating the first derivative, of the unfolding profile, as seen in the right panel of figure 3.2, or by doing a non-linear fitting to a Boltzmann sigmoidal curve, to extract the Tm, as in equation 3.3,

$$Y = F + \frac{U - F}{1 + \exp(\frac{Tm - x}{a})} \tag{3.3}$$

where Y is the fluorescence emission in arbitrary units, F and U are the minimum and maximum fluorescence respectively, Tm is the melting temperature, x is the temperature and a is the slope of the curve.

Experiments were performed in a StepOnePlus Real-time PCR System (Applied Biosystems). The assay was performed on 96-well plates (MicroAmp Fast 96-Well Reaction Plate, Applied Biosystems), in a total volume of 25 µl for each reaction. The melting curves were acquired in triplicate, and the average melting temperature was reported. For each measurement lysozyme (positive control) and proteins were used at 0.5 mg ml. All samples were exchanged into the gel filtration buffer. SYPRO orange dye (sigma) was used at 10X starting from a 5000X solution. The plate was sealed with optical quality sealing tape (Platemax) and centrifuged at 100g for 30s. The samples were equilibrated for 60s, at 25 °C, followed by a linear gradient from 25 °C to 95 °C in 1 °C increments with the SYPRO

orange fluorescence being recorded throughout the gradient. Melting temperatures were extracted by using the first derivative method applied to the sigmoidal melting curve. For the protein plus EDTA (Ethylenediaminetetraacetic Acid) melting temperature determination, EDTA concentration was varied from 32 to 0 mM in half dilution increments.

3.8 Molecular dynamics

The smad4 MH1 structure was retrieved from the protein data bank with the code 3QSV. Mutant models were generated with PYMOL. Molecular dynamics simulations were performed with the GROMACS package [88] using the ZAFF force field for metalloproteins [98]. The system was solvated in a dodecahedron box with TIP3P water, using the ambertools package from AMBER15 [73]. AMBER topologies were converted to GROMACS ones using the ACPYPE package [99]. The system was minimized for a maximum of 50000 steps or until the force constant was less than 1000 kJ/mol/nm, using the steepest descent algorithm implemented in GROMACS. The cutoff distance used for the non-bonded interactions, using the particle mesh Ewald method, was 10 Å. Prior to the final production simulation, the system was equilibrated using the NPT ensemble for 500 ps, followed by 50 ps in the NVT ensemble. Finally, for the production run, the system was simulated for 500 ns at 300K and 100 ns for the 450K simulations, using a 2 fs integration step for both cases. For the DNA-protein simulations a 100 ns total simulation time was used for each mutant. Temperature coupling was done with the Nose–Hoover algorithm at 300K. Pressure coupling was done with the Parrinello–Rahman algorithm at 1 bar. Simulation analysis was performed with the MDTraj [100] and MDAnalysis [101] python libraries. For each mutant, the reported analyses metrics were reported as an average of three simulations, with an accumulated time of 1.5 µS. For every 10 ps a frame was saved, totalling 50000 structures for further analysis.

Regarding analysis metrics, the residue mean squared deviation (RMSF) was calculated as given in equation 3.4,

$$RMSF(i) = \sqrt{\langle(\mathbf{x}_i - \langle\mathbf{x}_i\rangle)^2\rangle} \tag{3.4}$$

where x is the the atomic position for atom i and $\langle {x_i}^2 \rangle$ the average of all atomic positions for each atom i.

The root mean squared deviation (RMSD) for all atom pairs was calculated as given by equation 3.5,

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i(t) - \mathbf{x}_i^{\text{ref}}\right)^2} \tag{3.5}$$

where N is the number of atoms, x^{ref} and x are the coordinates at $t = 0$ and $t = t$ respectively.

The fraction of native contacts [102] were calculated according to equation 3.6, as implemented in MDanalysis [101].

$$Q(r, r_0) = \frac{1}{1 + \exp(\beta(r - \lambda r_0))} \tag{3.6}$$

where r and r_0 are the contact distances at time t and $t = 0$ respectively, β is the softness of the switching function and λ is the reference distance tolerance.

Salt bridges were defined as the fraction of native contacts between acidic and basic residues with a cutoff distance of 6 Å.

3.9 *In silico* stability calculations

In silico stability calculations were performed with the rosetta software, one of the state-of-the-art modelling software's, for calculating protein stability effects upon mutation [103]. We used the $ddg\_monomer$ application inside the rosetta modelling suite, with the author's previously described parameters [104]. Stability effects are ascertained by calculation of an approximated Gibbs free energy upon mutation ($\Delta\Delta$G), in the so-called rosetta energy units (REU). Values higher than 1.5 Kcal/mol and lower than -1.5 Kcal/mol are considered destabilizing and stabilizing mutations, respectively. Values between the previous values are considered to be neutral. The calculations were repeated ten times. Error bars represent the 95% confidence intervals for the mean.

3.10 Smad2 and Smad4 conservation entropy and disorder propensity calculations

Protein conservation entropy was computed using Smad2 (Uniprot:Q15796) and Smad4 (Uniprot:Q13485) pre-clustered alignments as computed by the GREMLIN webserver [105], using default parameters. A final set of 99 and 226 sequences were used for Smad4 and Smad2 respectively. The conservation entropy profile was extracted from the alignments using Skylign [106]. The protein disorder propensity was calculated with MetaDisorder [107]; a state-of-the-art algorithm employing twelve protein disorder predictors, with the final result being a consensus protein disorder propensity prediction.

Chapter 4

Results and discussion

4.0.1 The inter-domain linker of Smad2 and Smad4 is an intrinsically disordered protein

Smads are multi-domain proteins connected by linkers of variable length. Looking at the conservation entropy from the multiple sequence alignments shown in figure 4.1A, we can observe, as expected, that the MH1 and MH2 domains are conserved and with disorder propensity profiles, captured for the human full-length sequence, characteristic of globular proteins [108, 109]. The latter is stressed by the average disorder propensity below the 0.5 threshold. The linker segment is less conserved and has a higher disorder propensity suggesting a highly flexible protein, both for Smad2 (S2L) and Smad4 (S4L), approaching the maximum threshold value. This predicted inter-domain linker flexibility was confirmed by a qualitative analysis of the HSQC spectra depicted in figure 4.1B. The spectra displays signals characteristic of intrinsically disordered proteins (IDP) with a narrow dispersion for the ^{1}H frequencies [110–112]. The SAXS-derived Kratky plot (see figure 4.1C) also showed a profile typical for IDPs; it monotonically increases with a short plateau between 2 and 4 s.Rg [40]. The asymmetric pair distance distribution, Pr (see figure 4.1C) also indicated that both linkers are extended particles with a radius of gyration (Rg) and a maximum distance (Dmax) of 29.9±0.4 Å and 111.0±2 Å, respectively, for S2L. For S4L, the Rg and Dmax were 36.8±0.1 Å and 128.5±2 Å, respectively.

The Rg and Dmax for both linkers are substantially bigger than would be expected for a globular protein of the same sequence length (see figure 4.2A) and their Rg values are identical to proteins behaving as random coils, with a slightly bigger Rg ($Rg_{S2L}{}^{Rc}$ = 27 Å, $Rg_{S4L}{}^{Rc}$ = 33 Å), for the same sequence length [113]. Also the Uversky plot, (see figure 4.2B), a quantitative sequence based metric for assessing protein globularity, situates the two linkers at the disordered half of the plot with the S2L possessing a higher absolute charge, while the hydrophobicity is identical for both linkers.

To fully quantitatively characterize the conformational ensemble we applied the EOM method to the SAXS data (see section 3), assuming the intrinsically disordered behaviour, for the linkers, reported above. As can be seen in figure 4.1D

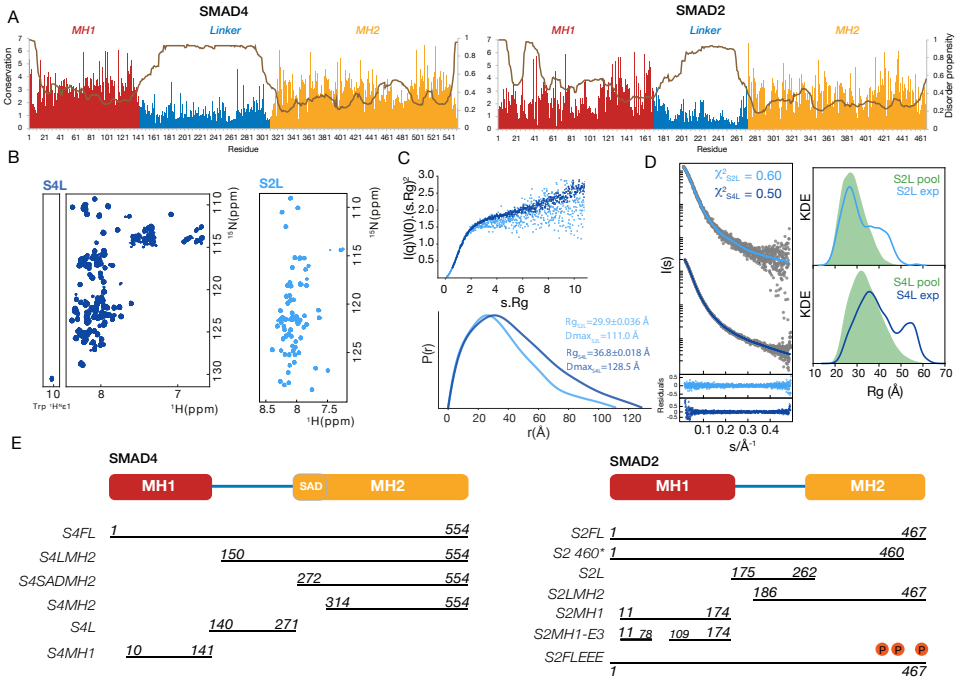


FIGURE 4.1: **A**: Conservation and disorder propensity for Smad2 and Smad4 proteins. In red, blue and yellow are depicted the MH1, linker and MH2 domains conservation entropy, respectively. In brown is represented the protein disorder propensity, where zero is fully ordered and one fully disordered. **B**: S2L (light blue) and S4L (dark blue) ^{15}N HSQC spectra revealing the narrow ^{1}H chemical shift dispersion characteristic of an IDP. **C** Kratky plots (upper panel) and distance distributions (lower panel) derived from the SAXS experimental profiles for S2L and S4L. **D**: In the left panel the experimental SAXS profiles in grey and the solid line simulated curve from the EOM, are displayed. In the right panel the kernel density estimate (KDE) for the random pool, in green, and the experimental Rg for S2L and S4L, are represented. In the bottom panel are the point by point residuals. **E**: Depicted is the numbering spanning all protein constructs used in this work. MH1 and MH2 are the Smad protein domains, L is the inter-domain linker, E3 is the Exon3 of Smad2, SAD is the Smad activation domain, and EEE are the three Glutamic acids acting as phosphomimetics.

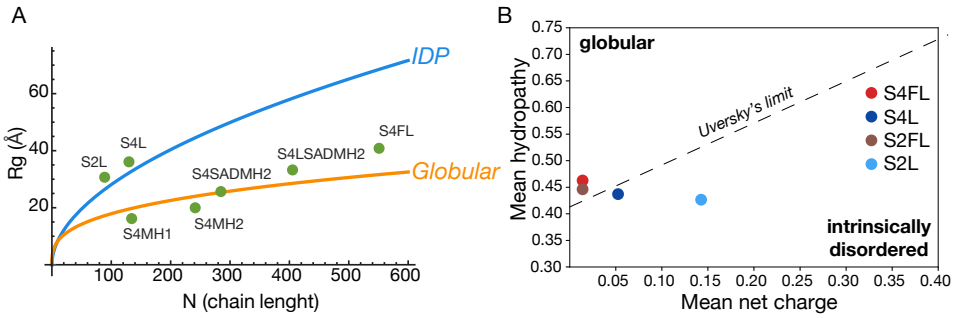


FIGURE 4.2: **A**: Radius of gyration, Rg, versus residue length, N, for Smads constructs. In blue is the Flory relationship for intrinsically disordered proteins and in yellow the Rg versus N for globular proteins, retrieved from the PDB [114]. **B**: Uversky plot for Smad2 and Smad4 linkers and full-length constructs calculated using the localCIDER python library [115].

left panel, the resulting experimental ensemble is extended with two major populations; one similar to the random coil ensemble and a more extended minor population, for both linkers. The EOM derived ensembles fully explain the experimental SAXS curves with an χ^2 of 0.6 and 0.5 for S2L and S4L, respectively. The fit to the random coil ensemble, without filtering with SAXS data also yielded good statistics (χ^2=1.03 for S4L and χ^2=0.63 for S2L) reinforcing the flexible nature of these linkers (see figure A.3C). Taken together these results indicate that Smad linkers behave as intrinsically disordered proteins with slightly more extended conformations, when compared to the random coil distributions.

To gather further evidence for this expanded behaviour we performed a quantitative sequence analysis, mainly regarding charge partitioning and hydrophobicity analysis. Both sequences have identical hydrophobicity, $f_{disorder}$ the fraction of disorder promoting residues and similar charged amino acid mixing, K. The major differences relate to the higher content of negatively charged residues, f-, of S2L and a subsequently higher fraction of charged residues, FCR, and net charge per residue, NCPR, as seen in table 4.1. Another striking difference is the value of Ω for S4L, more than four times that of S2L.

The Ω parameter has been related to protein expansion with lower values correlating with more expanded proteins; a value of 0 reports that proline and charged residues are mixed, while a value of 1 that they are segregated [115, 117]. For S2L the high proline content (16%) together with a low Ω would explain its expanded conformations. For S4L with a lower proline content (10%) and a higher Ω, one would expect more collapsed conformations when compared with the random coil conformations, this is not the case with similar relative extended profiles being observed as in the S2L case (see figure 4.1D, left panel). The reason for

TABLE 4.1: Sequence parameters for Smad2 and Smad4 linkers calculated using CIDER [115]. N is the number of residues, f- and f+ the fraction of positively and negatively charged residues, respectively. FCR is the total fraction of charged residues and NCPR the net charge per residue. K is the extent of charged amino acid mixing and Ω the patterning of charged plus proline residues. The parameter σ is the charge asymmetry along the sequence and $f_{disorder}$ the fraction of disorder promoting residues following the classification of Campen et al. [116]. Hydropathy is the Kyte-Doolittle hydropathy scale from 0 (least hydrophobic) to 9 (most hydrophobic).

	S2L	**S4L**
N	91	134
f-	0.154	0.067
f+	0.011	0.015
FCR	0.165	0.082
NCPR	-0.143	-0.052
K	0.177	0.20
Ω	0.089	0.416
$\mathbf{f_{disorder}}$	0.692	0.739
Hydropathy	3.88	3.97

the latter observation is not immediately apparent from sequence analysis alone. Also both linkers, regarding their amino acid sequence fractional charge, behave as globular proteins (see figure A.2); so the expanded conformations could be a result of charge partitioning and the high content of proline residues [108]. Recently it has been proposed that a higher serine content, in protein inter-domain linkers, could increase disorder propensities [118]. S4L and S2L have 22 and 9 serines, respectively. The higher serine content of S4L could compensate for its lower Proline content, in maintaining similar extended conformational distributions, when comparing to S2L.

4.0.2 Smad4 is a monomeric and flexible protein in an "open-closed" equilibrium of conformations

Building upon the described IDP behaviour of Smads linkers we recombinantly expressed ^{15}N labeled Smad4 full-length (S4FL) protein (see section 3.1) and superimposed it with the HSQC of its linker (S4L), as seen in figure 4.3A. About 60-70% of the resonances were superimposable, between the two constructs, reporting that in the full-length context the linker maintained a similar IDP-like behaviour. This method has been previously used to infer similar extrapolations [119].

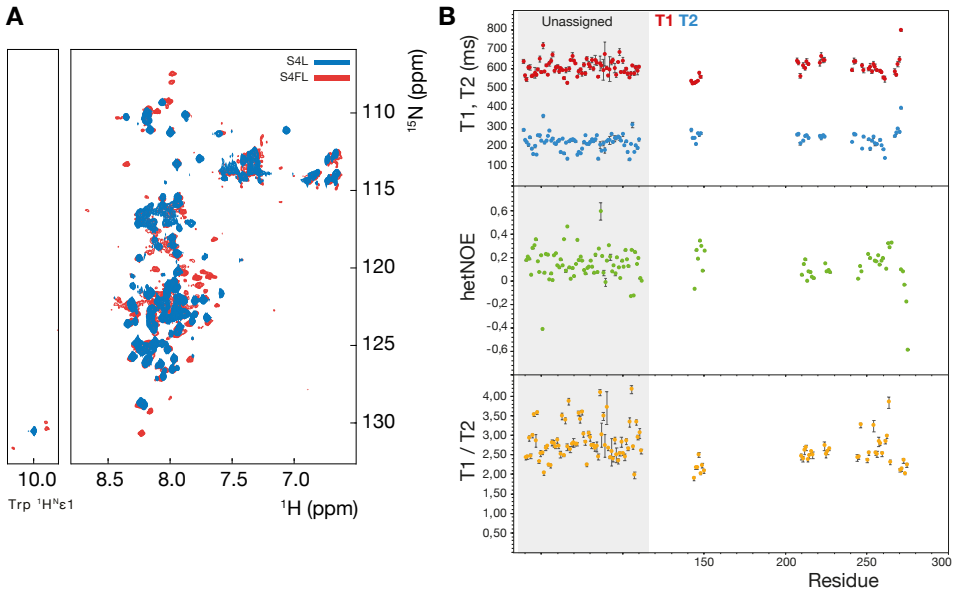


FIGURE 4.3: **A**: NMR HSQCs of Smad4 linker (blue) overlaid with the full-length (red) protein, showing also a typical low dispersity pattern for the backbone amides resonances, for the Smad4 full-length. **B**: T1, T2 and hetNOE relaxation experiments data for the Smad4 linker construct. Unassigned residues that could not be assigned are indicated at the grey banner and represented as gaps to the right.

Relaxation measurements were also undertaken for S4L. The protein is currently only partially assigned, due to the low dispersion and sequence variability of this linker. The assigned areas are located at the N- and C-terminal of the protein and we can observe typical values reported for IDPs (see figure 4.3B). The hetNOE values are mostly below the 0.5 threshold with the C-term reporting negative values; values above this threshold indicate a tendency for decreased flexibility with values below reporting decreased flexibility [111, 120]. The hetNOE values essentially dictate that for motions on the ps-ns timescale the S4L is a flexible protein with no major secondary structure observed, regarding this metric, at least for the assigned stretches of the protein. For the unassigned part the values are also mostly below the 0.5 values. For the T1/T2 ratio [1] some value oscillation is observed for the unassigned part and higher values for the C-term that could indicate some partial structure formation. Regarding the later, conclusions are

[1]This ratio is usually used as a reporter of secondary structure propensity for IDPs, with higher values indicating a tendency for secondary structure formation.

merely of a speculative nature as further assignment completeness is needed for a full secondary structure mapping.

After determining that the Smad4 linker is inherently flexible, even in a full-length context, we proceeded to analyse the behaviour of its individual domains and refine them in solution, using small-angle x-ray scattering (SAXS) data. The MH1 domain (S4MH1) is monomeric (Rg=16Å±0.8) and the experimental SAXS profile fully explains the determined x-ray structure previously published [7, 25] (χ^2=0.76), as seen in figure 4.4A. Regarding the MH2 domain (S4MH2) the x-ray structure reported a trimer in the unit cell [121]. Our SAXS data indicated that this domain is a monomer (χ^2=1.08, Rg=22Å±0.4) and not a trimer (χ^2=21.33), in solution (see figure 4.4A), even at concentrations higher than 20 mg/ml (see section 3). The first determined x-ray-structures of S4MH2 reported its oligomeric state as being trimeric [121, 122], while other studies indicated that it behaved as a monomer [20]. Our results provide the first solution structural information for this domain class and indicate the monomer as the major oligomeric state.

To fully characterize domain boundaries we produced another construct that extended towards the N-terminal of S4MH2 (see figure 4.1E), the S4SADMH2, incorporating the Smad activation domain (SAD) [4, 123]. This structural motif (M294-P312) was not seen in the electron density of the S4SADMH2 crystal structure (PDB:1DD1) with only a small β-sheet stretch of seven residues (N285-P293) being fitted, packing against another β-sheet (A425-Y430). We rebuilt the missing loops (see section 3) and checked if this packing could be a crystal artefact and if the linker started at E321 instead of Q289. Analysing the SAXS profiles the best scenario that was compatible with the experimental data was the one with the linker starting at Q289 (χ^2=0.87, Rg=25.1Å±0.1) (see figure A.3B, right panel). The model with the fully extended SAD motif was not compatible with the SAXS data (χ^2=2.01) (see figure A.3B, left panel).

Having established Smad4 domain boundaries, the IDP behaviour of the inter-domain linker and its domains oligomeric state, we thus proceeded to characterize the full-length protein. We randomized linker positions (see section 3) and analyzed the EOM derived ensemble with the SAXS data. The selected ensemble agreed with the SAXS experimental data (χ2=0.5, Rg=47 Å±1.0) with a low χ^2 and a random dispersion of the residuals (see figure 4.4B and table 4.2). The selected pool spanned a myriad of conformations (see figure 4.4C, top panel) with three major populations being detected; one with a MH1-MH2 inter-domain distance of ≈50Å a second with ≈100Å and a third around 160Å, corresponding to a Rg of ≈35Å, ≈50Å and ≈70Å, respectively (see figure 4.4C). The ensemble is overall slightly more compact than the random coil, given by the shift to lower dimensions (MH1-MH2 distance and Rg), after filtering with the SAXS data. The pre-filtered random coil ensemble followed a reasonable approximation to a gaussian distribution, not biasing any particular conformation (see figure 4.4C bottom panel in pink and figure A.4) and not explaining the experimental SAXS data (χ^2=1.98) (see figure A.3B).

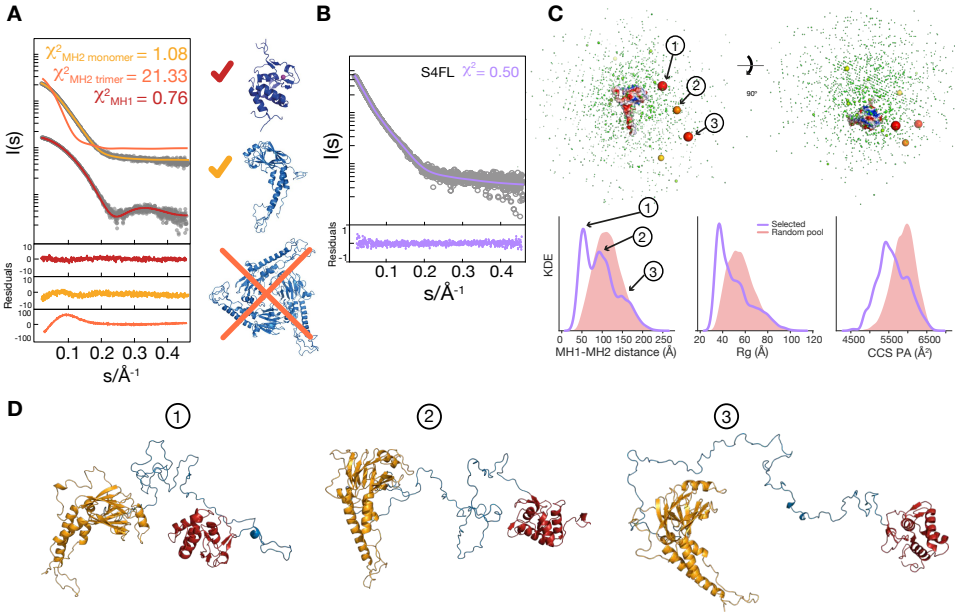


FIGURE 4.4: **A**: In solid lines are represented the SAXS simulated profiles from the MH1 (PDB:3QSV in red), MH2 monomer and MH2 trimer (PDB:1DD1 in orange), Smad4 domains. Underlaid are the simulated profiles, in grey, are the SAXS experimental intensities and in the bottom panel the experimental residuals. **B**: In pink is depicted the SAXS EOM simulated profile for the Smad4 full-length, overlaid with the experimental profile in grey and the respective residuals at the bottom panel. **C**: At the top panel is displayed the EOM Smad4 full-length ensemble. The MH2 domain is depicted in surface representation and the center-of-mass of the MH1 in sphere representation. The most representative conformations are numbered and the sphere radius is proportional to the probability occurrence, with the highest probability given in red and the lowest in green. In the middle panel are the kernel density estimates (KDE), in purple for the EOM ensemble and in pink for the random coil ensemble, corresponding to the MH1-MH2 domain inter center-of-mass distance, for the radius of gyration and for the collisional cross-section, from left to right respectively. The numbering refers to one stated in the top panel. **D**: Conformations from the EOM ensemble, referred in panel C, with the same colour code as figure 4.1.

At figure 4.4D a representative conformation for each of the three clusters reported above is presented and we can observe that the MH1 and MH2 domains

TABLE 4.2: SAXS parameters for Smad4 constructs. Reported values are for the average SAXS curve of the concentrations reported. Rg is the radius of gyration, Dmax is the maximum distance from the P(r) distribution and χ^2 is the Pearson's chi-squared goodness of fit test. Nomenclature is the same as in figure 4.1.

Protein	Construct	Concentration (mg/ml)	Rg (Å)	Dmax (Å)	χ^2
Smad4	MH1	1 - 2 - 4.2 - 7.4	16.0 ± 0.8	55.8	0.76
	MH2	0.5 - 1.2 - 3.2 - 12 - 20	22.0 ± 0.4	74.7	1.08
	SADMH2	1 - 2.2 - 5.5 - 17	25.1 ± 0.1	90.0	0.87
	L	0.5 - 1 - 2	36.8 ± 0.1	128.5	0.50
	LSADMH2	0.6 - 1 - 1.3	37.1 ± 0.2	146.5	0.86
	FL	1 - 3 - 6.6	47.0 ± 1.0	171.4	0.50

are non-interacting and are not in a predominantly auto-inhibited conformation, in our experimental conditions, as previously postulated [**hata_1997a**, 124, 125]. The asymmetry of the P(r) distribution (Dmax=171.4Å) (see figure 4.5) strongly suggests a moderate to high flexibility content of the full-length construct. The Kratky plot (see figure 4.5) also reports a plateau characteristic of flexible multi-domain proteins, at s.Rg ≈4 [113]. The linker, in the full-length context, is slightly more compact when compared to the S4L construct, approaching the random coil distribution (see figure A.5). The Rg for S4FL, as seen in see figure 4.2A), is located at the intrinsically disordered half and between the IDP and globular curves, for the same residue number; this results supports the flexible nature for the Smad4 full-length construct.

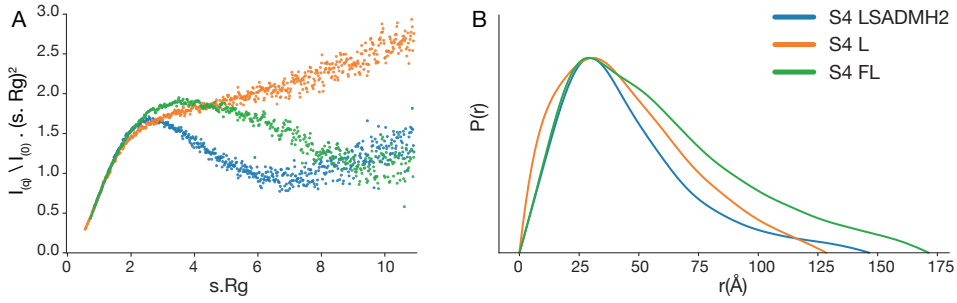


FIGURE 4.5: Kratky and pair distance distribution plots for Smad4 constructs. **A**: Kratky plot, where s is the momentum transfer, Rg the radius of gyration, I(q) is the intensity and I(0) the intensity at time zero. **B**: Pair distance distributions, where r is the distance and P(r) the pair distance distribution function.

To gain further insights into the nature of the compact conformations, while maintaining flexibility, we produced a deletion mutant, S4LMH2 (see figure 4.1E) including the linker and the MH2 domain, without the MH1 domain. The Rg for this construct was 37.1±0.2 Å, a value similar to the S4L Rg (see table 4.2). From the Rg values alone it is suggested that the linker in the S4LMH2 is more compact than for the S4L, this is bolstered by the shift to smaller Rg's when comparing the EOM ensemble (χ^2=0.86) to the random coil one (χ^2=2.01) (see figure 4.6). The latter observation suggests that the more compact conformations for S4FL are not triggered by a MH1-MH2 interaction, as previously stated.

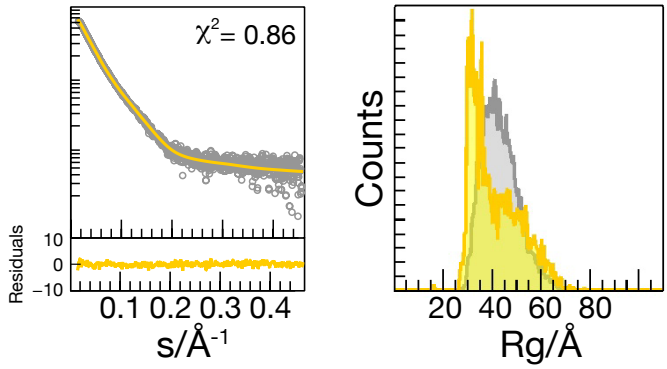


FIGURE 4.6: S4LMH2 construct SAXS data. *Left panel*: EOM derived SAXS profile, in yellow, fitted to the experimental profile, in grey. *Right panel*: Histogram of the EOM derived Rg, in yellow, fitted to the random coil distribution, in grey.

To further increase the robustness of the above results we decided to perform multi-curve fitting of the SAXS data using multiple constructs simultaneously, allowing us to compare individual domains when in isolation and in a full-length context, by increasing artificially the available SAXS resolution. To accomplish the latter we devised a strategy by which, starting from a full-length conformation, the corresponding domain was deleted and fitted to the experimental SAXS profile of the remaining construct, as seen in figure 4.7. In detail, for example, if one would want to disentangle the individual contributions of the S4L and the S4LMH2 constructs in a S4FL context, the FM-generated conformations to which the individual SAXS profiles were to be fitted, are given by the full-length ones. If the χ^2 is reasonable one could assume that the behaviour of S4L and S4LMH2, when in isolation, is similar to a full-length context, at the available SAXS resolution. Fitting of multiple SAXS curves has also been recently reported applied to protein fibrillation processes [126].

Looking at figure 4.7, the multi-curve χ^2_{multi} statistics are 0.78, 0.95 and 0.76 for S4L, S4LMH2 and S4L, respectively. These values agree with a scenario of a

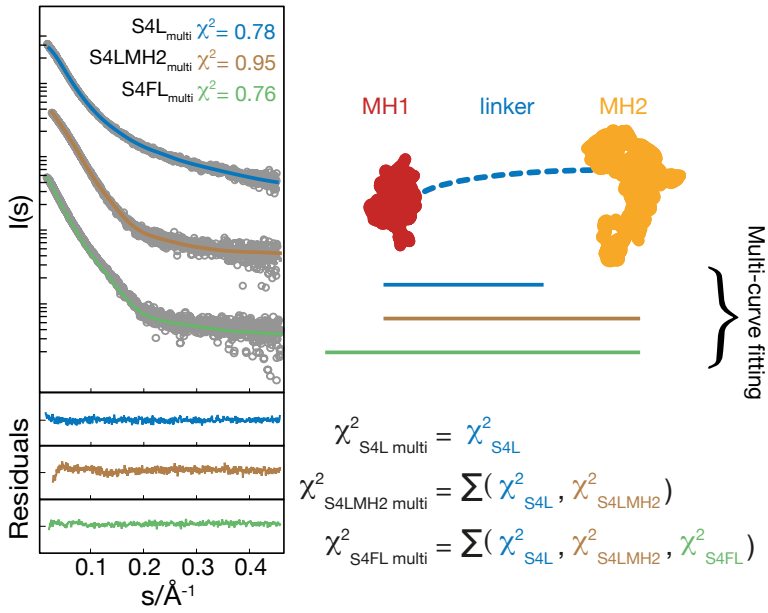


FIGURE 4.7: Smad4 SAXS multi-curve fitting using the S4FL, S4LMH2 and S4L constructs. Conformations used for the fitting, depicted in the left panel, were retrieved from full-length conformations by deleting the corresponding domains, exemplified at the right panel and fitted to the individual SAXS curves S4L, S4LMH2 and S4FL in blue, green and brown, respectively. Global fitting χ^2_{multi} is the best fit curve given by EOM using one, two or three individual curves for S4L, S4LMH2 and S4FL, respectively. The algorithm was adapted, to allow multi-curve fitting, from the original EOM algorithm [66, 67].

similar conformational landscape for the S4L and S4LMH2 when is isolation and when in a S4FL context, even if slightly more compacted in a full-length context as for the case of the S4L construct (see figures 4.1 and A.5). Multi-curve fitting reinforces the flexible nature of the Smad4 linker in its full-length counterpart adding robustness to the Smad4 full-length results described in figure 4.4.

4.0.3 Smad4 is not predominantly in an auto-inhibited conformation

To obtain atomic level detail into a putative MH1-MH2 interaction and to validate the results reported above, we performed a titration of 3 equivalents of the S4SADMH2 construct into the ^{15}N labelled MH1 domain (see figure 4.8A). The

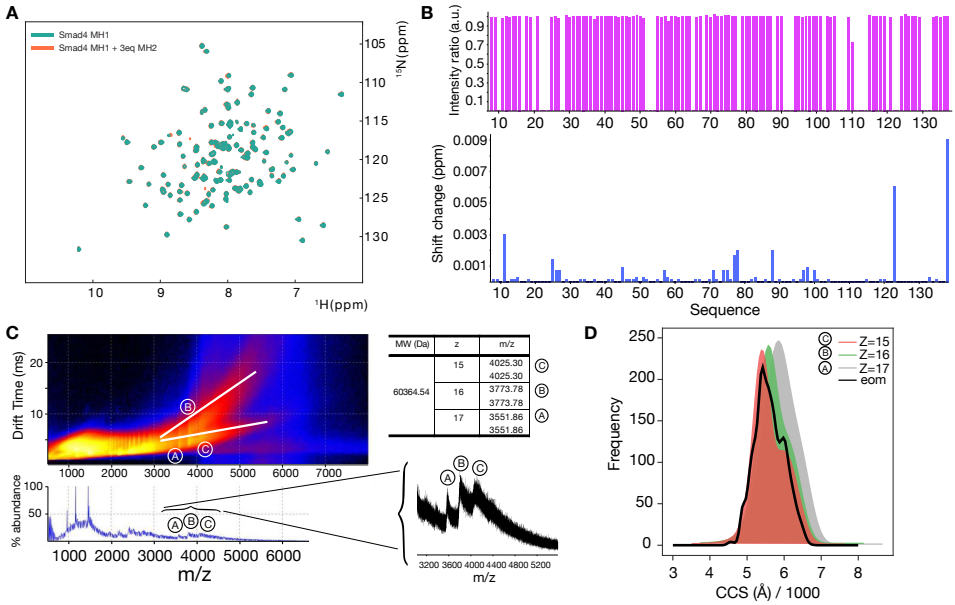


FIGURE 4.8: Smad4 MH1-MH2 titration and IM-MS of the full-length Smad4. **A**: NMR titration of 3 equivalents of the Smad4 MH2 domain, in orange, into the MH1 domain, in green. **B**: NMR titration peak intensities, in pink, and chemical shift perturbations in light purple. **C**: IM-MS of the Smad4 full-length protein. A, B and C are the mass-to-charge ratios, m/z, analyzed and z is the charge. Each white line indicates a protein conformation. **D**: Collisional cross section derived from drift time distributions depicted in panel C. The EOM curve is the theoretical collisional cross section (CCS) for the SAXS-derived ensemble, calculated with the IMPACT software [127] using the PA method.

chemical shift difference mapping experiment did not show any major differences in the chemical environment, around the MH1 domain, with no peak shifts (highest value was 0.009 ppm) and intensity changes (see figure 4.8A and 4.8B) being observed, indicating that the MH1 and MH2 domains do not form a stable complex.

We estimated that the MH1-MH2 interaction affinity, for 3 equivalents and using a one-site binding model [128], is higher than 20 mM (Kd>20mM) (see figure A.6). We then calculated the estimated inter-domain affinity (K'd) and an estimated effective concentration ($Ceff$ > 2.2 mM) (see scheme 1 and equations 4.1

to 4.2), where f_{closed} and f_{open} are the closed and open populations of S4FL, respectively, determined from the EOM SAXS-derived ensemble [2]. When we took the average curves of the open and closed conformations and calculated the optimal linear combination, we found that the retrieved values were ≈0.15 and ≈0.85 for the closed and open curves, respectively. These values are in good agreement with the ones, derived from the ensemble analysis, of 0.10 and 0.90 for the closed and open populations, respectively.

$$\mathrm{MH_1} + \mathrm{MH_2} \rightleftharpoons \mathrm{MH_1{-}MH_2} \quad \{1\}$$

$$K_d = \frac{[MH1][MH2]}{[MH1 - MH2]}$$

$$K'_d = f_{open}/f_{closed} \tag{4.1}$$

$$K'_d = K_d/Ceff \tag{4.2}$$

$$Ceff = K_d \cdot f_{closed}/f_{open} = 20 \cdot 0.10/0.90 = 2.2\, mM \tag{4.3}$$

The effective concentration is a metric that is seldom measured experimentally and poorly discussed in the context of multi-domain proteins. Applied to multi-domain proteins, it provides a quantitative description of the effect of disordered linkers by estimating the intra-molecular affinity if the inter-molecular affinity is known or estimated [130, 131]. When at a given concentration in an intra-molecular interaction the encounter rate between connected domains equals the rate of the same unconnected interaction, we determine the effective concentration. By comparing the intra- and inter-molecular affinity estimates, the presence of the linker increases by an almost an order of magnitude, from 20 to 2.2 mM , the apparent affinity between the MH1 and MH2 domains (see equation 4.3). In either case the low affinities indicate that the linker could act as merely a mechanical element to approximate both domains without a stable interaction. The latter could enhance the affinity with transcription factors and Smad responsive DNA motifs, as proposed for other systems [131, 132]. The linker influence can be quantitatively ascertained by the effective local concentration, $Ceff$. When the Kd is smaller than $Ceff$, closed conformations are favoured and if the Kd is larger than $Ceff$, open conformations are preferred [133]. For S4FL the relation $Kd > \approx 10 \cdot Ceff$, indicates that open conformations could be favoured, reinforcing the non-interacting MH1-MH2 domain hypothesis, postulated earlier.

[2] Open and closed conformations were determined empirically by considering all closed conformations the ones with at least one distance between two atoms, between the MH1 and MH2 domains, below the 4.5Å threshold. This value was considered the minimum interaction threshold for a putative atomic interaction [101, 129]. Open conformations where the ones with values higher than 4.5Å.

We validated the EOM SAXS-derived ensemble using ion mobility mass spectrometry (IM-MS), by comparing the theoretical collisional cross-sections (CCS) with the ones derived from the experimental drift time distributions (see figure 4.8C) [127, 134]. We could extract drift time distributions from three mass-to-charge ratios (m/z) 15 (C), 16 (B) and 17 (A) corresponding to values 4025.30, 3773.78 and 3551.86Å^2, respectively. For each m/z we observed two putative conformations, given by the two white lines in figure 4.8C. The mass determined from all m/z was ≈60380 Da that correlates well with the theoretical value of 60364 Da (see table in figure 4.8C), reinforcing once again a monomeric state for Smad4 full-length. The EOM derived CCS could be qualitatively compared with the IM-MS ones (see figure 4.8D); the distributions spanned similar CCS values and reported the two major populations observed from figure 4.8C, one at 5500Å^2 (see figure 4.8D), slightly more compact than the random coil ensemble (see figure 4.4C, bottom panel) and another at ≈6000 Å^2, with the CCS comparable to the random coil maximum CCS. The three m/z gave different relative contributions from the two conformations reported above; as we increase the m/z to 17 the more extended conformations are favoured. Despite this qualitative comparison solution and gas-phase experiments have to be compared with care. For IDPs it is known that IM-MS experiments can sometimes lead to more compact and/or extended conformations in a solvent depleted environment [135]. As seen in figure 4.8C there are a multitude of conformations to lower m/z ratios, possible indicating a plethora of more compact conformations that couldn't be analyzed due to conformational heterogeneity. The CCS distributions [3] reported in figure 4.8D had to be normalised to the experimental CCS due to a compaction of around 2000Å^2. The later value is highly unlikely to occur in solution. Nonetheless the relevance of this results is stressed by the similar relative population ratios for the two populations observed, that could be qualitatively compared to populations 1 and 2 in figure 4.4C.

We also determined the Tm (see section 3) for all globular or partial globular constructs of Smad4. The retrieved values were 63.4±1.1 ºC, 59.6±0.1 ºC, 57.8±0.3 ºC, 60.1±0.1ºC and 56.8 ºC for S4MH1, S4MH2, S4SADMH2, S4LMH2 and S4FL, respectively (see figure 4.9). The Tm of the linker S4L couldn't be determined due to the lack of globularity of this construct (see section 3). The rationale for determining these melting temperatures was to ascertain protein stability in the buffer conditions used for all experiments. All values reported stable conditions for all constructs with all values around 60ºC. Looking at the data, the S4MH1 is the most stable and the S4FL the least stable protein. Adding the SAD element to S4MH2 seems to slightly decrease protein stability and adding the linker to the

[3] Drift time values were obtained by fitting Gaussian curves to the distributions, with the maximum being the drift time for the particular m/z ratio. The CCS distributions in figure 4.8D were converted from the drift time distributions in figure 4.8C using the following calibrants with known CCS: β-Lactoglobulin (bovine milk), transthyretin (human plasma), avidin (egg white), serum albumin (bovine), concavalin A (Canavalia ensiformis), alcohol dehydrogenase (Saccharomyces cerevisiae) and pyruvate kinase (rabbit heart) in 200 mM ammonium acetate at 20 µM, using a standard curve. Final CCS for each m/z are reported as the maximum of the distribution.

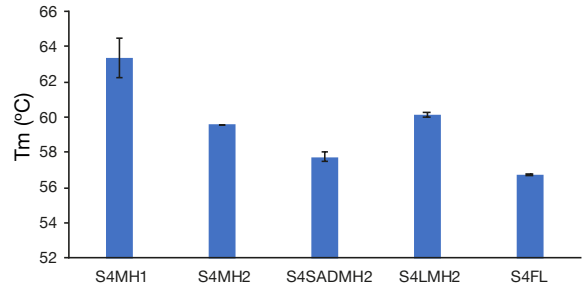


FIGURE 4.9: Melting temperatures for Smad4 constructs using differential scanning fluorimetry. The protein naming follows figure 4.1 scheme.

S4SADMH2 recovers the stability to a higher value than the S4MH2 construct. We can thus conclude that the buffer conditions used did not compromise protein stability.

4.0.4 Solution characterization of the Smad2 MH1 domain isoforms reveals their monomeric nature

Having established a methodological framework to describe the Smad4 conformational landscape, we applied the same reasoning to Smad2, a representative of the R-Smad family (see section 1). For Smad2 two splicing variants have been described at the MH1 domain, with and without the exon 3 (E3). [4] (see figure 4.1 and A.1). Both variants agree well with the experimental SAXS profiles ($\chi^2_{S2MH1E3}$=0.81 and $\chi^2_{S2MH1\text{-}E3}$=0.69) as seen in table 4.3 and figure 4.10A. The S2MH1E3 has bigger measured dimensions (Dmax=74Å, Rg=19.3±0.6Å) than the S2MH1-E3 (Dmax=66Å, Rg=17.3±0.3Å). This was expected due to the E3 exon presence. Both proteins are monomeric as reflected by the invariant Rg and Dmax at increasing concentrations.

The GS loop and the E3 exon were both treated as flexible regions and were allowed to span a large conformational space (see section 3). For the GS loop, present in both S2MH1E3 and -E3 constructs, its highly flexible nature was confirmed by calculating the per-residue S^2 order parameter, that reported the average value to be below 0.7; a quantitative threshold that correlates with protein flexibility [136] (see figure 4.10D). For the E3 exon, despite some unassigned residues, we detected some flexibility at its extremes and a short α-helix spanning residues 90 to 98. How the formation of this helix, that was previously unreported, affects protein function is currently not clear. For all other residues, excepting the ones in the GS loop and the E3 exon, all S^2 values agreed with flexibility profiles characteristic for secondary structure elements in globular proteins, above the 0.7 threshold.

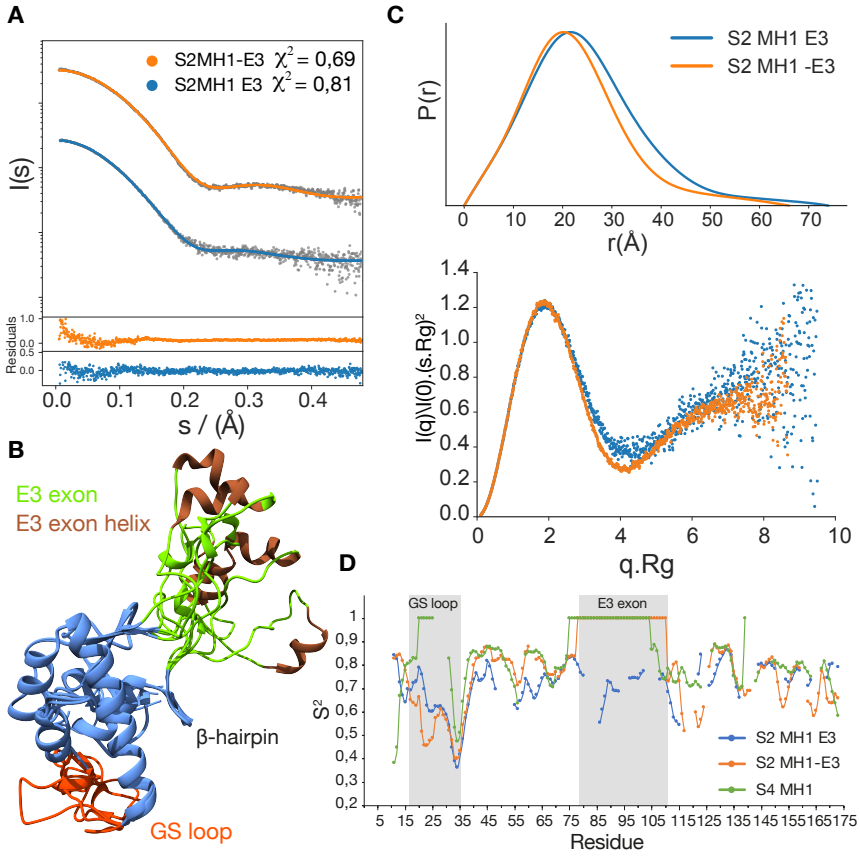


FIGURE 4.10: Smad2 MH1 SAXS and NMR data analysis. **A**: EOM-derived fits for S2MH1E3, in blue and S2MH1-E3, in orange, to the SAXS experimental profiles, in grey. The residuals are below with the same colour-code. **B**: Five representative structures of the EOM-derived ensemble for S2MH1E3. Colouring follows label naming. The helix in brown corresponds to residues 90-98. The β-hairpin is the DNA binding hairpin. **C**: Pair distance distribution function and Kratky plots for profiles described in **A**, following the same colour arrangement. **D**: Order parameter S^2 calculated using TALOS [50] for the S4MH1, S2MH1-E3 and S2MH1E3 proteins. Residue numbering follows S2MH1E3, the biggest protein, for easy comparison. A value of 1 for S^2 represents a gap in the sequence alignment for the corresponding protein and blank residues correspond to an unassigned residue.

TABLE 4.3: SAXS parameters for Smad2 MH1 an linker constructs. Reported values are for the average SAXS curve of the concentrations reported. Rg is the radius of gyration, Dmax is the maximum distance from the P(r) distribution and χ^2 is the Pearson's chi-squared goodness of fit test. Nomenclature for the constructs is the same as in figure 4.1.

Protein	Construct	Concentration (mg/ml)	Rg (Å)	Dmax (Å)	χ^2
Smad2	MH1 E3	1 - 3 - 5	19.3 ± 0.6	74.0	0.81
	MH2-E3	1.3 - 3.8 - 6.8	17.0 ± 0.3	66.0	0.69
	L	1 - 2.2 - 4.5	29.9 ± 0.4	111.0	0.60

Also the disorder propensity predictions and the conservation, increased and decreased, respectively, when compared to the surrounding E3 residues (see figure 4.1A).

The distance distribution function in figure 4.10C, indicates that both constructs are globular with the S2MH1E3 being slightly bigger. Regarding the Kratky plot, both constructs show similar flexibility profiles with the S2MH1E3 showing a somewhat slightly increased flexibility given by the higher values at s.Rg > $\approx$4 and the non-convergence to values approaching zero [40, 137]. The latter could be explained by the presence of the E3 exon and its flexible nature. Nonetheless it is of note that for a highly flexible E3 one would expect possibly a higher divergence in the Kratky plot for s.Rg > $\approx$2. This non-verified hypotheses could indicate that the exon is not highly flexible (see figure 4.10 A and B), maybe interacting transiently with other protein elements, namely the DNA binding β-hairpin.

4.0.5 Smad2 is an oligomeric protein, in a monomer-dimer-trimer equilibria, shaped by phosphorylation

After characterization of the solution behaviour of the Smad2 MH1 domain and its linker, we followed the same strategy as for Smad4. Having established the IDP-like behaviour of S2L we randomized its positions and acquired SAXS data to characterize Smad2 full-length in solution. Compared to Smad4, the Smad2 experimental workflow had another layer of complexity: the oligomerization propensities of R-Smads. The oligomeric nature of R-Smads, namely Smad2 and Smad3, has been established through many published works [4, 138] (see section 1). The exact stoichiometry of the complexes has been a matter of debate; the basal state (pre-phosphorylation) is thought to be manly monomeric [139] or dimeric with almost no trimeric species formed [125, 140], with phosphorylation possibly shifting the equilibrium to the trimeric state [21, 141]. Some authors do not detect dimeric species while others refer variable stoichiometries with the equilibrium being concentration-dependent. To our knowledge no solution data has been reported for a full-length recombinant R-Smad protein.

to address this issue we acquired SAXS data on Smad2 full-length wild-type (S2FLWT) and a phosphomimetic (S2FLEEE) to reproduce a phosphorylated Smad2 (see section 3) [4]. SAXS data at different concentrations, 13.6, 24.3, 46.8 and 56.1 µM, for S2FLWT (see figure 4.11A, left panel) and 9.4, 18.7 and 28.1 µM for S2FLEEE (see figure 4.11B, left panel). The χ^2, in increasing concentration order, was 0.77, 0.67, 0.75 and 0.70 for S2FLWT and 0.8, 0.76 and 1.03 for S2FLEEE. All χ^2 values, for the EOM-derived ensembles (see section 3) are in excellent agreement with the experimental data.

For the S2FLWT construct at 13.8 µM, the monomer state predominates with 99.1% in contrast to the S2FLEEE that at 9.4 µM has a monomer population of 40.1%, reaching at the 26.1 µM maximum a 19.5% monomer fraction. From the analysis above we conclude that the oligomerization, for the full-length proteins is concentration-dependent and trimer formation is favoured when S2FLWT is phosphorylated at the C-terminus. In S2FLWT, dimer formation reaches a top of 18.5% at 55.1 µM while for S2FLEEE dimer populations are almost invariant with concentration, with a constant fraction of ≈37% (see figures 4.11A, B and A.8A, B). This concentration independence suggests that dimer formation is an intermediate important for trimer formation, a species thought not to be formed as reported in previous studies [5] [140].

To gain further insights into the self-association mechanism we produced a deletion mutant, S2LMH2WT lacking the MH1 domain. SAXS data analysis for this construct revealed that trimer formation is more favoured when compared to the full-length proteins. At 60.2 µM trimer formation was almost complete, reaching a 99.9% fraction (see figure A.7A and 4.11C). Dimer formation was also concentration independent with fractions ≈26%, while the monomer was undetectable at 60.2 µM (see figure A.8A). The χ^2 statistics also reflected the agreement of the EOM-ensemble with the experimental SAXS data (see figure A.8A, B). The most plausible reason for a favourable shift towards trimer formation is the deletion of the MH1 domain; it can interfere with trimer formation either by intra or inter-molecular interactions. For Smad2, auto-inhibited conformations can't be discarded, as reported for Smad4 (see section 4.0.3). Previous studies on a equivalent deletion mutant of Smad3, showed that trimer formation at 18-29 µM was 50% complete and that phosphorylation at the C-terminal tail increased the self-association affinity 32-35 fold [140]. For S2LMH2WT we obtained similar results with the 50% fraction being reached at ≈15 µM. The obvious step would be to produce an equivalent phosphorylated protein for this Smad2 mutant (e.g. S2LMH2EEE). However, we didn't manage to establish a successful production protocol, with the SAXS data constantly reporting profiles characteristic of an aggregation-prone protein.

[4] Although mimetizing phosphorylation with glutamic acids is not always a good approximation, for R-Smads it has been shown to activate TGFβ, in a similar fashion, as the phosphorylated species, in an in vivo scenario [20].

[5] The authors reported that dimer formation was insignificant but for a deletion mutant of Smad3, not including the MH1 domain. Human Smad3 has 92% sequence similarity to Smad2.

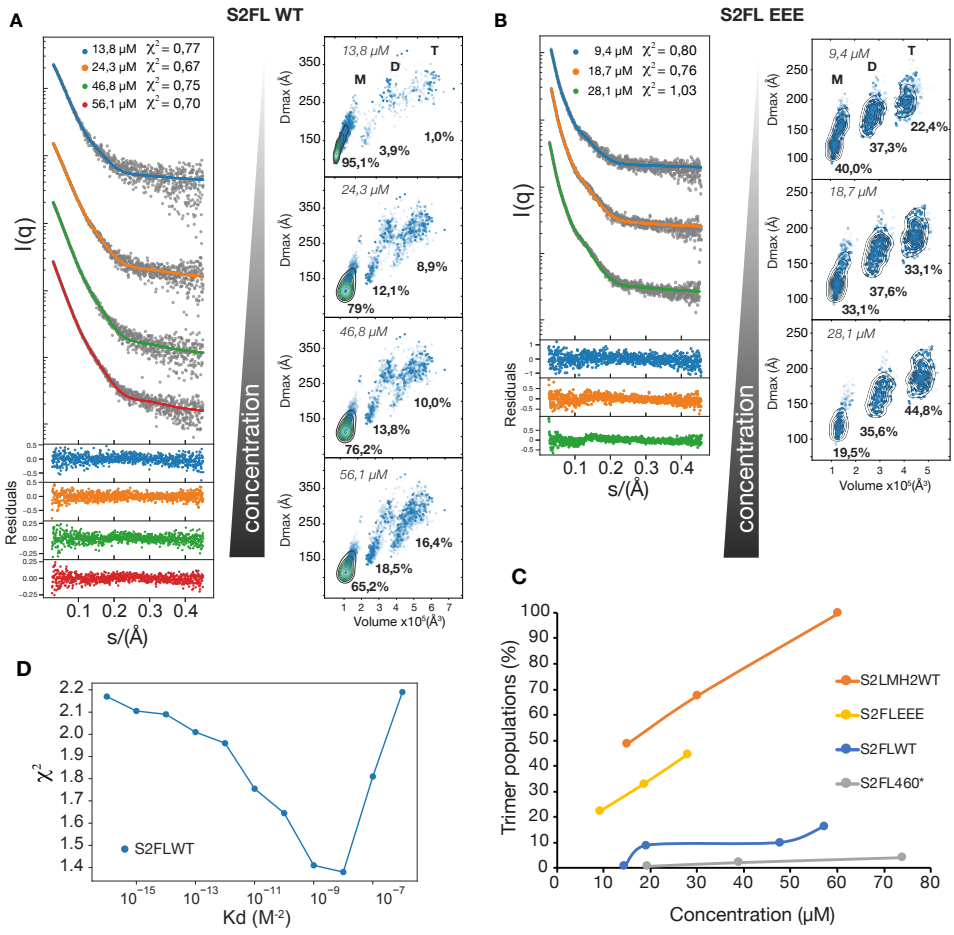


FIGURE 4.11: Smad2 full-length SAXS data analyses. **A**: *Left panel*: EOM-derived SAXS profiles for the S2FLWT construct in blue, orange, green and red, with corresponding residuals below. Experimental profiles are in grey. *Right panel*: Population distribution from the EOM ensembles in **A**. Dmax is the maximum distance for each selected conformation and plots are arranged in increasing concentration from top to bottom. M,D,T are monomer dimer and trimer. Below each oligomeric state is depicted its calculated population. **B**: EOM-derived SAXS profiles for the S2FLEEE construct. Figure layout is identical to **A**. **C**: Trimer populations derived from **A**, for S2FLWT, **B**, for S2FLEEE and figure A.7, for S2LMH2WT and S2FL460*. **D**: Estimated affinity for S2FLWT using best-fit combinations for simulated Kds, calculated using equations 4.4 to 4.7. Concentrations used for Kd simulations are the same as in **A**.

Building upon the previous results, we produced another deletion mutant without the C-terminal phosphorylatable tail. For this mutant, S2FL460*, no trimer formation was observed at almost 80 µM, indicating the C-terminal tail as an essential element for trimer formation. The used concentrations of 19.5, 38.9 and 73.9 µM showed reasonable χ^2 statistics of 0.64, 0.67, and 0.76, respectively (see figure A.7B, left panel). Even though that for this construct trimer formation is almost abolished, dimer formation increased up to ≈20%. The COSMIC database [142] list mutants with a stop codon at the C-terminus of Smad2. Abolishing trimer formation and consequently TGFβ activation, could indicate that tumours where this mutation is detected, inactivate TGFβ and its tumour-surpressor phenotype [1, 143].

To have a quantitative metric that could describe further these self-associating mechanisms and also to validate the EOM ensembles analyzed above, we established a workflow for calculating affinities using explicit models. Using the retrieved populations described in figure 4.11 and solving the analytical equations for the equilibrium, would allow us to describe in a more rigorous way the Smad2 self-association models, as stated below.

The monomer-dimer-trimer equilibrium is described by the following equations: in equation 4.4, the general equilibrium, with a dimeric species as intermediate is given, where K_2 and K_3 are the constants for dimer and trimer formation, respectively.

$$3M \underset{k_{d2}}{\overset{k_2}{\rightleftharpoons}} D + M \underset{k_{d3}}{\overset{k_3}{\rightleftharpoons}} T \tag{4.4}$$

Subsequently the equilibrium constants K_{d2}, K_{d3} and the total constant K_d are defined in equation 4.5.

$$\begin{aligned} K_{d3} &= \frac{[M][D]}{[T]}, \\ K_{d2} &= \frac{[M]^3}{[D][M]} = \frac{[M]^2}{[D]}, \\ K_d &= K_{d2}K_{d3} \end{aligned} \tag{4.5}$$

The total concentration, $[M]_{tot}$ the one used for SAXS experiments, is given by the mass balance in equation 4.6.

$$[M]_{tot} = [M] + 2[D] + 3[T] \tag{4.6}$$

As no analytical expression exists for this equilibrium, the system given by equation 4.7 has to be solved algebraically, combining equations 4.5 and 4.6. For the total SAXS concentration, M_{tot} and the simulated constants K_{d2} and K_{d3}, it is trivial to find the theoretical oligomeric concentrations by solving the system in equation 4.7.

$$\begin{cases} [M] + 2[D] + 3[T] - [M]_{tot} = 0 \\ [D]K_{d2} - [M]^2 = 0 \\ [T]K_{d3} - [D][M] = 0 \end{cases} \tag{4.7}$$

Having calculated the theoretical oligomeric concentrations, $[M]$, $[D]$ and $[T]$ we retrieve the populations Y_M, Y_D and Y_T for the monomer, dimer and trimer species, respectively (see equation 4.8).

$$Y_M = \frac{[M]}{[M]_{tot}}, Y_D = \frac{2[D]}{[M]_{tot}}, Y_T = \frac{3[T]}{[M]_{tot}} \tag{4.8}$$

With the theoretical populations we can thus use them as lineal coefficients for obtaining the lineal combination, of the average per-species EOM-derived SAXS ensembles, with be best χ^2 statistics [6]. If a convergence is observed to a reasonable χ^2, we find our global affinity constant K_d, for this equilibrium [7]. Using SAXS to infer affinity constants have been recently used for other systems, namely protein-RNA associations and one-state protein self-assembly [144–146].

In figure 4.11D the estimated affinity for S2FLWT is $Kd \approx 10^{-8}M^{-2}$ with a minimum χ^2=1.38, while for the S2LMH2WT, without the MH1 domain, $Kd < 10^{-11}M^{-2}$ and χ^2=1.20 (see figure A.8B). For S2LMH2WT we didn't observe a relative minimum, as for S2FLWT, with higher values giving also reasonable χ^2 statistics. The reported affinities, using analytical ultra-centrifugation, for an equivalent deletion construct of Smad3 were $Kd \approx 3.3 \cdot 10^{-10}M^{-2}$ and phosphorylation [8] increased it by 32-35 fold [140]. Both results indicate that possibly S2LMH2WT has a higher self-associating affinity than an equivalent construct of Smad3. For S2LMH2WT the affinity, when compared to S2FLWT, is higher by a one to two orders of magnitude, indicating a possible role for the MH1 domain in interfering with the oligomerization mechanism.

Phosphorylation of S2FLWT is expected to increase the affinity by more than an order of magnitude to values approaching S2LMH2WT (see figure 4.11C). Following the same strategy as for S2FLWT, yielded a higher affinity for S2FLEEE by about an order of magnitude. In spite of the latter, the χ^2 statistic was not satisfactory, possibly because during the averaging procedure for the monomer, dimer and trimer curves, some conformational characteristics were not captured and/or the SAXS profiles could also be affected by some residual aggregation that impairs this type of analysis.

[6] Usually reasonable means χ^2 values lower than 1.5 [41].

[7] All calculations were made using in-house scripts built using Bash and Python programming languages.

[8] The authors also used a phosphomimetic variant with three glutamic acids.

Going a step further we also devised a strategy, similar to Smad4, to probe the MH1-MH2 inter-domain distances of Smad2, for S2FLWT and S2FLEEE (see figure 4.4C). The inter-domain distances appear to be concentration-dependent; as we increase concentration the inter-domain distance seems to shift to lower distances, for all oligomeric states. The concentration dependence is more pronounced for the S2FLWT protein, with a shift to lower distances for the peak at ≈50Å. This dependence seems to be less pronounced for S2FLEEE for the dimeric [9] and trimeric species and even less so for the monomeric species (see figure 4.12).

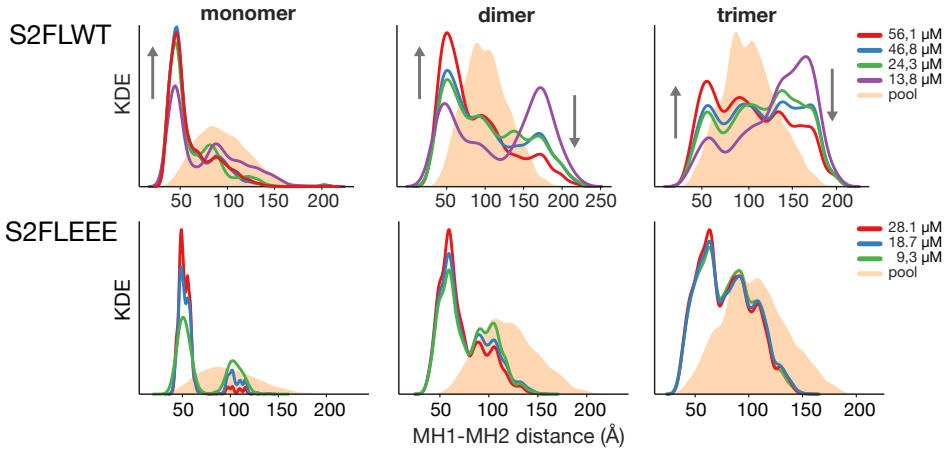


FIGURE 4.12: Smad2 full-length center-of-mass distance distributions from the random pool and the EOM ensembles. Center-of-mass distributions, between the MH1 and MH2 domains, for the S2FLWT and S2FLEEE constructs. In orange is the random pool distribution for each oligomeric state and in red, blue, green and purple the concentrations from the SAXS-derived EOM ensembles. KDE is the kernel density estimate. The grey arrows indicate a decrease or decrease of the KDE.

Remarkably the peak with the distance ≈50Å reports similar inter-domain distances as seen for Smad4 (see figure 4.4C, cluster 1), in spite of its shorter inter-domain linker (≈130 residues for Smad4 and ≈90 for Smad2), as apparent for the monomer state in both S2FLWT and S2FLEEE. The shits observed for the 120-250Å interval, when compared to Smad4 distribution, could be mainly due to intra-molecular steric hindrance and to the presence of other oligomeric species.

[9] No structural information is available for a dimeric Smad intermediate state, so we assumed the putative dimer interface as derived by the trimeric Smad2 structure (see section 3), by retrieving one unit and performing molecular dynamics simulations for stability assessment. During the 85ns trajectory the dimeric interface appeared to be stable (see figure A.7C), with this being the dimer state for the SAXS data analysis. The simulation protocol was identical to the one described in section 3.

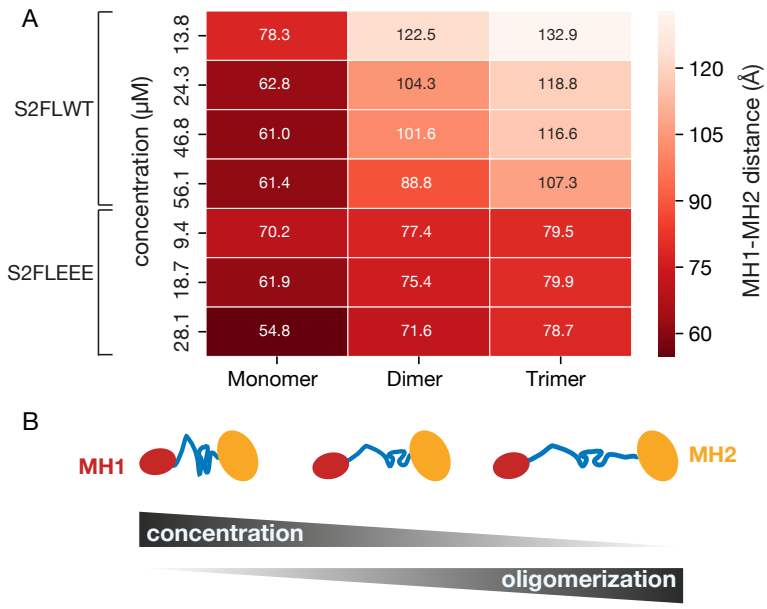


FIGURE 4.13: **A**: Smad2 full-length average center-of-mass distance distributions derived from figure 4.12. Smaller to higher distances are represented in a dark-to-bright red gradient, respectively. **B**: Model resuming the results obtained in panel A. The colouring scheme obeys to figure 4.1A layout. The greys bars depict, from left to right, decreasing concentration and increasing oligomerization (monomer-dimer-trimer).

Taking the above results together and considering the average inter-domain distances from figure 4.12 (see figure 4.13A), we can establish a possible mechanism for Smad2 activation. The inter-domain distances increase with the oligomerization state and decrease with concentration as schematized in figure 4.13B. The observed changes are more pronounced for S2FLWT and less apparent for the monomeric state, both for S2FLWT and S2FLEEE.

4.0.6 Smad7 protein production and stability analysis - an unstable multi-domain protein

As a chronological note, the first trials for obtaining pure full-length Smads started with Smad7 because no structural information was published regarding the full-length or any of its individual domains. As referred in section 1.1.1, Smad7 belongs to a different class of Smad proteins, the I-smads, and our objective was to obtain structural information on this sub-family. The ultimate goal was to lay

the foundations for a, structurally based, mechanism of TGFβ inhibition by the inhibitory Smads.

Our first attempts started with the Smad7 MH2 domain. To increase the probability of obtaining soluble and homogeneous protein we cloned multiple constructs in multiple fusion partners, as described in table 4.4. Construct design was based firstly in multiple sequence alignments/disorder predictors, trying to maintain conserved/ordered regions and varying variable/disordered regions, and secondly, fusing our protein to different solubility enhancing tags [147, 148]. Looking at the Smads protein sequence alignment in figure A.1 we defined four boundaries: D257-R426, encompassing the conserved MH2 domain. The V229-R426 construct, encompassing the MH2 plus a linker segment and the V229-S413 construct encompassing the linker, the MH2 domain and a deletion of a divergent C-terminal tail. We also cloned the full-length protein, M1-R426, and a deletion mutant where a disordered and divergent stretch of 65 residues was deleted (M1-(Δ20-85)-R426).

TABLE 4.4: Smad7 constructs tested for protein production. All plasmids have the 3C protease cleavage site, except for the SUMO constructs that are cleaved with the SUMO protease, Ulp1. Thio is thioredoxin. Exp and Sol is the expression and solubility, respectively. Stab is the stability. The Δ20-85 is a deletion of residues 20 to 85. The -/+ signs describe a negative or positive result, respectively.

Construct	**Plasmid**	**Fusion** tag	**Exp./Sol.**	**Stab.**
Smad7 D257-R426	pPEU10	N-term His-thio.	-/-	
Smad7 D257-R426	pOPINE	C-term His-tag	-/-	
Smad7 D257-R426	pOPINS	N-term His-SUMO	+/+	-
Smad7 D257-R426	pOPINM	N-term His-MBP	+/+	-
Smad7 D257-R426	pPEU11	N-term His-Z	-/-	
Smad7 V229-S413	pPEU10	N-term His-thio.	-/-	
Smad7 V229-S413	pPEU11	N-term His-Z	-/-	
Smad7 V229-R426	pPEU10	N-term His-thio.	-/-	
Smad7 V229-R426	pOPINE	C-term His-tag	-/-	
Smad7 V229-R426	pOPINS	N-term His-SUMO	+/+	-
Smad7 V229-R426	pOPINF	N-term His-tag	-/-	
Smad7 V229-R426	pPEU11	N-term His-Z	-/-	
Smad7 V229-R426	pOPINM	N-term His-MBP	-/-	
Smad7 M1-R426	pOPINF	N-term His-tag	+/+	-
Smad7 M1-(Δ20-85)-R426	pOPINF	N-term His-tag	+/+	-

Looking at table 4.4 the constructs that yielded positive expression and solubility tests were the ones fused to the SUMO (D257-R426, V229-R426) and MBP

(D257-R426) tags, regarding the MH2, with the full-length construct and deletion mutant also yielding positive results with a His-tag. In spite a successful protein production for all constructs, the stability was compromised were all proteins aggregated within hours or minutes after purification. The latter timespan was not always reproducible and was purification dependent. An exception was construct D257-R426 (MH2 domain) with a N-term His-MBP fusion; this protein showed an increased stability profile but when the MBP was cleaved the construct precipitated and/or was bound to the MBP and impossible to separate by using ion exchange or affinity chromatographies. The MBP tag is known to dramatically increase protein solubility, but also to yield false positives: it can bind to unfolded or aggregated proteins with high affinity. Also posteriorly to protease cleavage the passenger proteins can sometimes precipitate when fused to MBP [149, 150]. The buffers used for purification protocols followed established protocols described in section 3.1 and table B.3.

To increase protein stability we followed a high-throughput approach by employing a stability screening [10], in 96 well plates, for the two constructs that yielded the best results, in the preliminary trials reported above. For the M1-R426 construct (full-length) the maximum temperature obtained was around 45 ºC (see figure 4.14) and all conditions gave similar or lower Tm values, to the initial preliminary trials.

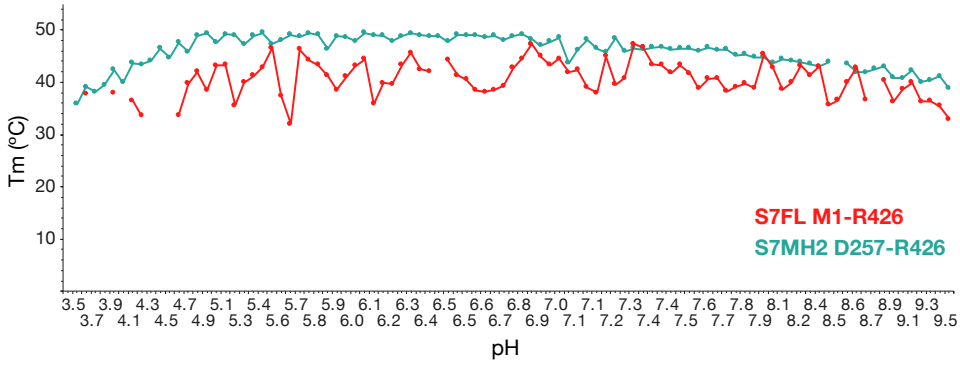


FIGURE 4.14: Melting temperatures, obtained by differential scanning fluorimetry, for S7FL M1-R426 and S7MH2 D257-R426 using the buffer screening described in table A.1 in 200 mM NaCl. Gaps represent the Tm values that couldn't be determined.

The melting temperatures seemed to be pH independent with a value always below 50 ºC. For the D257-R426 construct the values were slightly higher, than the

[10] This screening assumes that an increase in melting temperature reports a favourable buffer condition [95–97]

full-length construct, with the maximum value around 48 ºC. This profile seemed to be slightly pH dependent with lower Tm for pH's below 5 and higher than 8.

Following the previous results we hand-picked a few conditions, where the Tm were identical to the preliminary trials, (e.g. HEPES and MES pH 7) as alternatives for protein production. We were not able to increase protein stability with all conditions giving qualitative similar results.

For the D257-R426 construct we set up protein crystallisation trials with no avail; almost all plates showed protein aggregation with some conditions giving phase separation. None of the trials gave satisfactory results and we decided not to pursue further experiments. Regarding the full-length constructs we acquired SAXS data with the profiles reporting an aggregation-prone proteins.

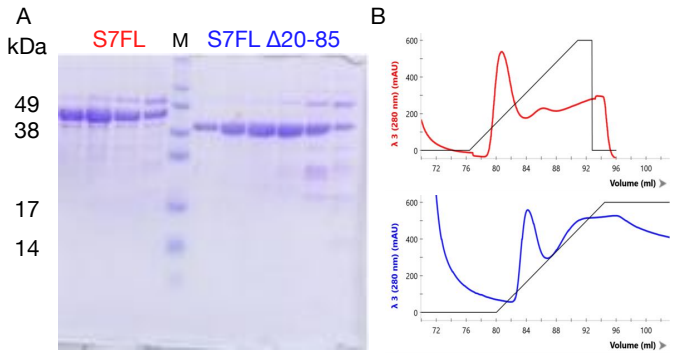


FIGURE 4.15: **A**: SDS-PAGE gel of Smad7 soluble fractions. S7FL is Smad7 full-length, corresponding to residues M1-R426. To the left is represent the marker sizes in kilodaltons, corresponding to the marker, M. **B**: Nickel affinity chromatography profiles for the S7FL construct, in red, and the Smad7 M1-(Δ20-85)-R426, in blue. The gradient, in brown, was from 0 to 400 mM of imidazole in 15 column volumes using a 1 ml HisTrap™ column (GE Healthcare).

As seen in figure 4.15A, both the full-length and the deletion mutant, gave satisfactory SDS-PAGE bands from the chromatography profiles in figure 4.15B. In spite of the latter both constructs were very aggregation-prone and precipitation was observed from within minutes to hours after purification.

Our initial goal was to obtain sufficient amounts of a folded and stable Smad7 construct for biophysical and biochemical experiments. Our objective was partially accomplished, by laying the bases for future construct optimization experiments and reinforcing high-throughput Tm determination, as a useful tool in recombinant protein production.

4.0.7 Implications for TGFβ signalling

Smad proteins are multi-domain proteins connected by disordered linkers. For human Smad4 the observed population distribution, in a predominantly non-autoinhibited state, suggests a crucial role for its inter-domain linker. The linker of Smad4 is a target of multiple post-translational modifications (see section 1) and also, as put forward by this work, an enforcer of the solution conformations of Smad4. The linker of Smad4 is highly conserved in metazoans [4], an unusual trend for disordered inter-domain linkers. The latter suggests its important role is not only effected by being a target of post-translational modifications, but also as to provide mechanical leverage for the Smad4 conformational landscape. Also if the auto-inhibited state was predominant, probably, the high entropic costs would be to high for maintaining this state, with such a long linker.

When we tried to model the linker by simple polymer scaling laws, namely worm-like and Gaussian chains, no values agreed with the experimental data, suggesting a more complex structural landscape, opposing a strict random coil ensemble only encoded by sequence length, as suggested for other multi-domain proteins [132, 151, 152].

Domains connected by flexible linkers usually increase encounter rates, of the tethered domains, by several orders of magnitude. More studies are needed to really identify sequence determinants that govern the Smad4 landscape by, for example, changing the linker length and altering its sequence properties (e.g. proline content, charge partitioning) and observing effective concentration changes [153].

Multi-domain proteins are advantageous. Local concentration increase, compared to isolated domains, could shorten the timescales of cellular response to stimuli [154]. The Smad4 linker is slightly more compact in S4FL than in S4L; transient interactions of the linker with Smad4 domains are not discarded and should also be investigated.

Curiously the closest conformations (see figure 4.4C, cluster 1) are mainly at the opposite interface, of the MH2 domain of Smad4, where the transcription factors usually bind. As a speculative note, we could extrapolate that linker conformations probe Smad4 for protein partner interactions, by not occluding the protein binding interface (see section 1.)

Regarding Smad2, the same reasoning stated above for multi-domain proteins is applied, when comparing to its non-ligated counterpart.

In our experimental conditions we detected dimeric intermediates that seemed to be important for trimer formation and a subsequent TGFβ activation. Dimer formation increases with concentration only for S2FLWT and S2FF460* and a possible role in inhibiting trimer formation should be further investigated.

A trimer population of ≈16%, for S2FLWT at 56.1 µM and almost zero for S2FL460*, stressed the importance of the C-terminal tail not only when it is phosphorylated. The reported trimer formation for S2FLWT is not completed disregarded to interact with S4FL and its functional role is currently not clear.

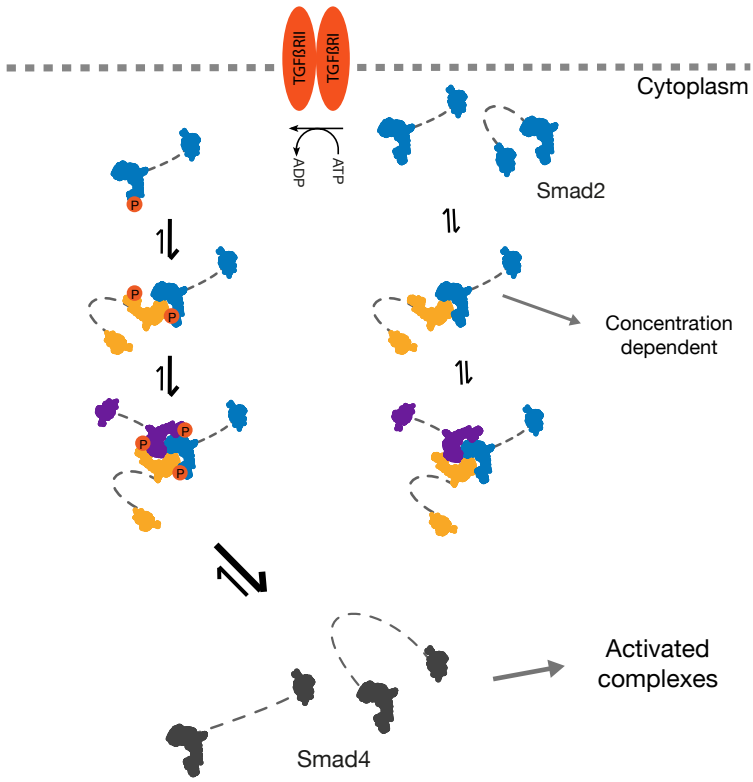


FIGURE 4.16: Smads general activation mechanism. Human Smad2 full-length protein exists in a monomer-dimer-trimer equilibrium. Upon phosphorylation, by the TGFβ receptor TGFβI, the equilibrium is shifted towards the trimeric species, with the dimeric state as an intermediate. The inter-domain distances are concentration-dependent and increase, on average, with the oligomerization state. For both Smads, open and closed conformations seem plausible. Deletion of the MH1 domain and the C-terminal tail increases or depletes trimer formation, respectively. Smad4 could interact with the phosphorylated species and also to a lesser extent with the non-phosphorylated counterpart. A myriad of combinations could be formed between other R-Smads and transcription factors to activate or repress target genes in the cell nucleus, as pointed out by a recent study [155].

One can speculate that as Smad2 has a shorter linker than Smad4, the possible auto-inhibited state for Smad2 is more plausible due to the lower entropic costs, as seen for the monomeric state.

For Smad7 we were largely unsuccessful in establishing a satisfying purification protocol for this protein. One possible explanation might be that Smad7 needs to be co-purified in the presence of other cofactors or proteins partners, to increase its stability. Possible candidates could be Smurf2 [12], Smad4 [156] or even the TGFβ receptor [157].

Inter-domain distances increase with oligomerization (see figure 4.16), perhaps to prepare Smad2 to interact with other protein partners and DNA. On the other hand if Smad2 local concentration is increased, the decrease in inter-domain distances, could mean an inhibition for TGFβ with MH1 being unavailable to interact with target genes. Protein local concentration variations, working as an activator or repressor of signalling pathways has been recently reported to occur in DNA-interacting proteins and other transcription factors, both in vivo and in vitro [158–161]. As a corollary, local protein concentration could act as a regulator of Smad2 activation, by interfering with its inter-domain distances.

4.1 Mutational landscape analysis of Smads - the MH1 of Smad4

The work presented in this chapter was done in collaboration with master student Àngela Vea Badenes.

4.1.1 Mutations mainly affect charged and hydrophobic residues

By looking at figure 4.17 it is apparent that Smad4 MH1 cancer mutations primordially affect hydrophobic and charged residues, by decreasing the fraction of the former and increasing the latter. These residues are mainly involved in maintaining the protein fold and by establishing charged interactions that are important for protein function.

4.1.2 Smad4 MH1 cancer mutations affect protein stability

We have selected 58 tumour mutations of Smad4 MH1 domain identified in patients, deposited in the COSMIC and cBioPortal databases [142, 162]. Of them, we selected a subset of 25 to sample the potential effects that these mutations can cause in protein function measured as: aggregation propensities of the mutated protein and/or alterations in DNA binding capacity and protein stability (see table 4.5). We chose mutated positions in elements of secondary structure or in loops, in the DNA binding hairpin, in and around the residues coordinating the

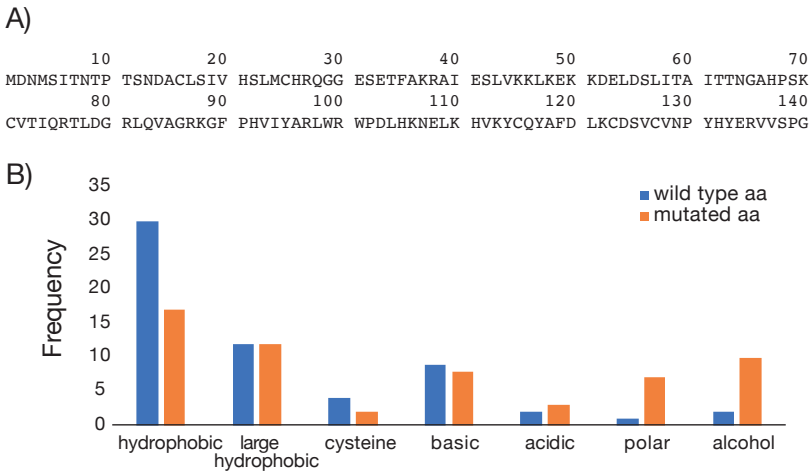


FIGURE 4.17: **A**: Human Smad4 MH1 amino acid sequence. **B**: Frequency of mutation retrieved from the COSMIC database [**forbes:2015**]. Amino acid classification followed guidelines from the cBioPortal [162].

zinc atom, and also previous information available in the literature regarding the role of some of these mutations in patients (see table 4.5).

We cloned and expressed the wild-type protein and mutants as depicted in section 3.1. Twenty-five mutants were overexpressed but only five were soluble (E41K, K45N, G65E, G86C, K106R and R135Q). These soluble mutations were located at the α1, α2, α3, β3, α4 and at the C-terminal region of the MH1 domain, respectively (see figures 4.18A, 4.18B), displaying Tm profiles characteristic of folded proteins (see figure A.9A) [95, 96], with the exception of K106R, which was purified in the soluble fractions but with a profile characteristic of an unstable/unfolded protein (see figure A.9A). The soluble proteins correspond to mutations involving charged residues or glycine, located mainly in solvent exposed areas (see figure 4.18B), away from the protein core.

In general, all mutants studied displayed a decrease in melting temperatures with respect to the wild-type protein but with values in the range of folded samples. The order of stability according to the Tm values is, WT (69.4 ºC), K45N (68.3 ºC) E41K (63.1 ºC), G65E (56.8 ºC) G86C (53.43 ºC) and R135 (53.96 ºC), as seen in figure 4.19.

Also, the melting temperatures decreased by around 10 ºC for all mutants,

[11] Figure 4.19A was produced in collaboration with the author and was previously published in the master thesis of Àngela Vea Badenes with slight changes. It is showed here for illustrative purposes.

TABLE 4.5: Mutants analyzed for establishing the mutational landscape. Reason for selection, cancer type and the database of data retrieval are showed.

Mutant	Database	Cancer type	Reason for selection
E41K	COSMIC	-	Stability. DNA binding
L43F	COSMIC	Pancreas	Stability
K45N	COSMIC	Large intestine	DNA binding
L57V	COSMIC + cBioPortal	Pancreas, colorectal	Stability
G65E	cBioPortal	Colorectal	Bibliography. DNA binding
G65R	COSMIC	Large intestine	Bibliography. DNA binding
R76I	COSMIC	Large intestine	Stability
G86C	COSMIC	Lung	DNA binding
H92Y	COSMIC, cBioPortal	Cervical, pancreas large intestine	Stability
Y95H	COSMIC	-	Stability
R97H	COSMIC, cBioPortal	Colorectal, stomach large intestine, uterus	Bibliography. DNA binding
L98F	COSMIC, cBioPortal	Pancreas, large intestine	Stability
W99R	cBioPortal	-	Bibliography
R100T	COSMIC, cBioPortal	Large intestine, pancreas	Bibliography
R100W	COSMIC, cBioPortal	Lung	Bibliography
R100G	COSMIC	Large intestine	Bibliography
L104F	COSMIC, cBioPortal	Colorectal, skin large intestine	Multiple cancers
K106R	COSMIC + cBioPortal	Lung, soft tissue	DNA binding
C115R	COSMIC	Large intestine	Zinc coordination
A118V	COSMIC, cBioPortal	Pancreas, oesophagus large intestine	Multiple cancers. Stability
V128M	COSMIC	Thyroid	Methionine mutation
N129K	COSMIC	Large intestine	DNA binding
P130L	COSMIC, cBioPortal	Pancreas	From bibliography. DNA binding
Y133N	cBioPortal	Bladder	Stability
R135Q	COSMIC	Biliary tract	Stability

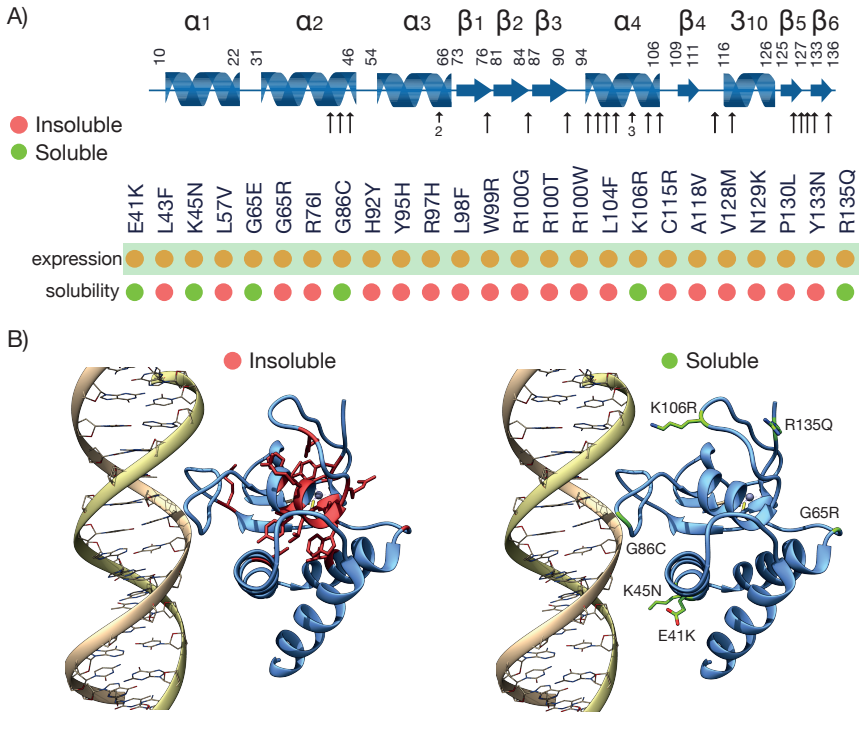


FIGURE 4.18: **A**: Smad4 MH1 secondary structure elements retrieved from its x-ray structure (PDB:3QSV). At the bottom panel are the mutants studied during the course of this work. Soluble, insoluble and positive expression are depicted in green, red and yellow respectively. The E. Coli strain was used for mutant production. **B**: Soluble and insoluble mutants, from **A**, represented in the Smad4 MH1-DNA complex. The zinc atom is in sphere representation.

when they were recorded in the presence of EDTA, a metal chelating agent. Titrations were performed with increasing amounts of EDTA up to 0.32 mM and revealed the same relative order of stability for the proteins, with values ranging from 56 ºC for the WT to 49 ºC for the R135Q mutant (see figure 4.20). From the previous results, we could conclude that the differential effects in protein stability are not primarily leveraged by zinc coordinating effects.

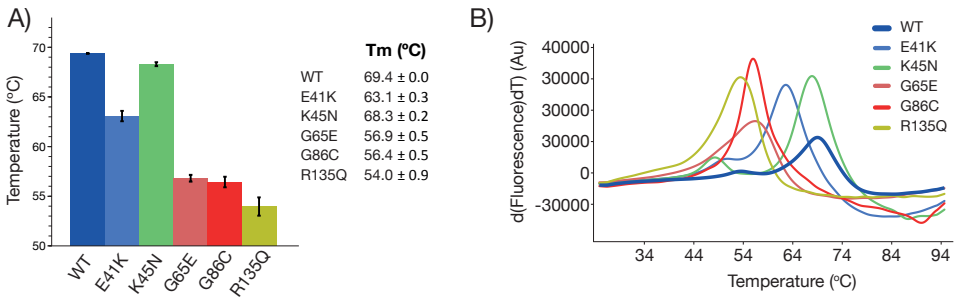


FIGURE 4.19: **A**: Melting temperatures for the soluble mutants. **B**: Derivatives for the differential scanning fluorimetry profiles depicted in 4.19A [11].

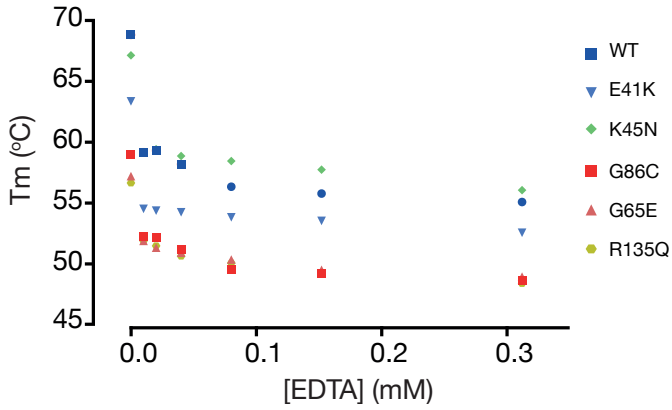


FIGURE 4.20: DSF derived melting temperatures for Smad4 MH1 soluble mutants, in the presence of increasing concentrations of EDTA.

4.1.3 Mutants maintain DNA binding functionality to the SBE canonical DNA motif

Smad proteins can function as tumour suppressors or enhancers. Generally, in oncogenes, mutations tend to affect functional sites altering their regulation, while in tumour suppressors they cluster mainly in the protein core, in areas that destabilise the protein fold [163, 164]. In Smad proteins, tumour mutations are distributed along the sequence, covering both, the MH1 domain involved in DNA recognition, the linker and the MH2 domain protein interacting sites [4].

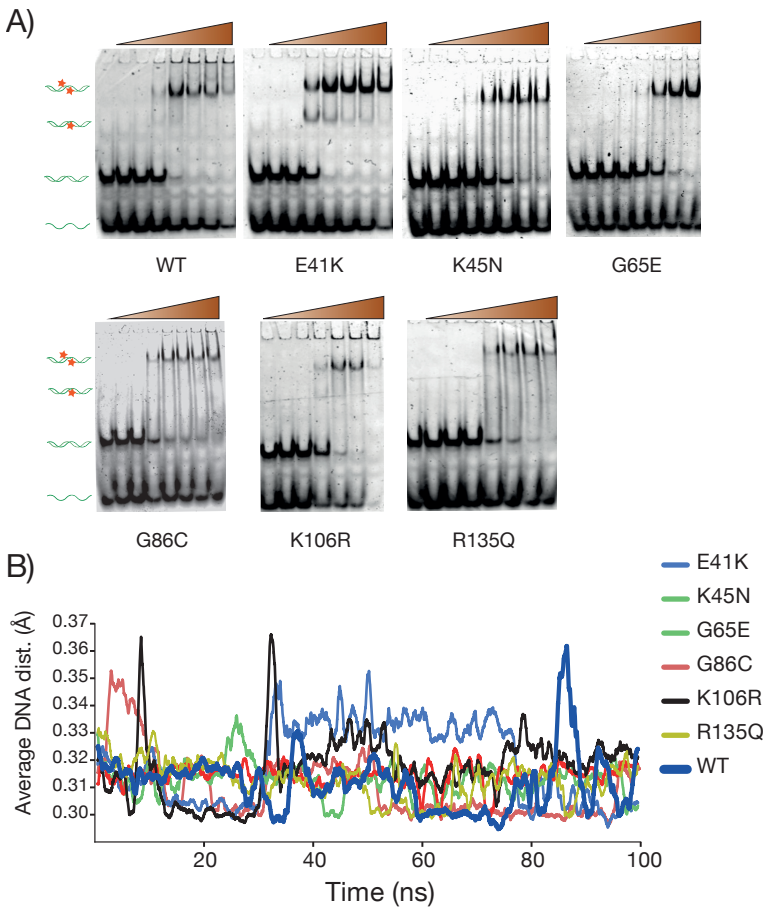


FIGURE 4.21: **A**: EMSAs binding assays for the Smad4 MH1 WT and soluble mutants. The cartoons to the left of the gels represent, from bottom to top: single stranded DNA, double stranded DNA, monomer binding to double stranded DNA and dimer binding to single stranded DNA [12]. **B**: Molecular dynamics simulation analysis for the average DNA binding distance regarding the binding interface for the Smad4 MH1-DNA complex (PDB:3QSV) as previously depicted [4, 7].

To investigate if the soluble mutants were compatible with the DNA binding

[12]Figure 4.21A was produced in collaboration with the author and was previously published in the master thesis of Àngela Vea Badenes. It is showed for illustrative purposes.

properties of Smad4, we performed EMSA assays using the canonical Smad Binding Element described in the literature [4, 7, 165]. As displayed in figure 4.21A, all soluble mutants bind DNA, within the same affinity range, as that of the WT, given by the saturation at similar protein concentrations. Albeit their stability profiles revealed a difference in melting temperatures of approx. 10ºC (see figure 4.19A), even the G86C mutation located at the DNA binding interface, maintained DNA binding functionality.

4.1.4 SAXS analysis of mutants, does not reveal major conformational changes, when compared to the WT

To determine if the mutations affected the global shape of the proteins, we produced samples of the five soluble mutants for SAXS data acquisition. Of the five mutants, only the ones with the Tm closest to the WT (E41K and K45N), gave SAXS data of sufficient quality for further data processing (see figure 4.22A). The other three mutants showed signs of aggregation even at 1mg/ml (≈55 µM), given by the non-linear dependence of the q2 vs ln(I) at low scattering angles [137]. The E41K and K45N mutants, also showed concentration dependent aggregation but at higher concentrations, around 4 mg/ml (≈220 µM).

The three curves are very similar, with the exception for q below ≈0.15 Å, where the mutants have a higher Dmax at 61 Å and an Rg of 20.79 ± 0.2 (E41K) and 19.8 ± 0.01 Å (K45N). The WT has a Dmax of 56 Å and a Rg of 15.9 ± 1 Å (see figure 4.22B). The Dmax and Rg differences, between the WT and the mutants, are due to the impossibility to cleave completely the His-tag from the mutants, maybe due to some tag interactions or aggregation propensities. The differences observed for the Pr distribution (see figure 4.22B) are mainly also due to the presence of the His-tag, since the center of the distribution is roughly at the same distance, for the three proteins, only diverging at higher distances around 32-34 Å.

The above results reinforce the melting temperature data, reporting that the E41K and K45N variants were the most stable mutants.

4.1.5 *In Silico*, equilibrium and non-equilibrium molecular dynamics simulations reflect the complexity of the Smad4 MH1 mutational landscape - the importance of salt-bridges

To gain insights into the atomic mechanisms that shape mutant stability, we decided to perform *in silico* energy calculations and long molecular dynamics simulations of all the soluble mutants and some of the insoluble ones for comparison, as controls.

In Silico stability calculations

The logic behind this approach was to compare the *in silico* results with experimental data and use this knowledge to predict the behaviour of other mutations.

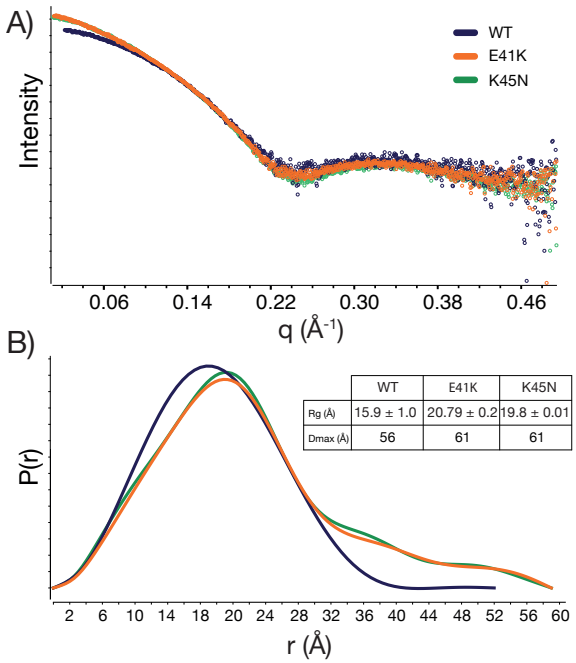


FIGURE 4.22: **A**: SAXS profiles for the Smad MH1 WT, E41K and K45N mutants. **B**: Pair distance distribution function for the proteins described in panel **A**. Rg is the radius of gyration and Dmax is the maximum distance observed in the distribution.

We first used a simplified approach aimed at predicting approximate values for free energy calculations using the rosetta software suite [104, 166]. Using another state of the art predictor, the foldX software [167], yielded similar results. Our calculations indicated that most of the mutations were destabilising with respect to the wild-type sequence, as revealed by high values above 1.5 REU (see section 3.9). While melting temperatures and free energy differences ($\Delta\Delta G$) have to be compared with care, both are correlated with changes in protein stability and share enthalpy (ΔH) and entropy (ΔS) changes, that could be compared qualitatively [168]. Also, protein in vitro thermostability has been recently shown to be an accurate metric for protein misfolding and/or aggregation propensities in a cellular context [169]. The C115R mutation (one of the zinc coordinating residues) displayed the highest value of $\approx$16 REU (see figure 4.23). According to the simulations, some mutations might have a marginal stabilising effect, (e.g. E41K and R76I) or nearly neutral as for the L43F and R76I mutants. However, when we tried to produce the R76I mutant with recombinant techniques we found it insoluble. All of the soluble mutants, with the exception of the E41K, were predicted to be

destabilising. The G65R and G65E mutants proved to be a paradigmatic case, *in silico* calculations predicted both mutations to be highly destabilising. However, when we expressed the proteins, the G65E mutant was soluble and folded and the G65R was insoluble. Given the non-apparent correlation between the calculations and the experimental information, we decided to perform molecular dynamics simulations to gain insights into the possible mechanisms for the mutational landscape.

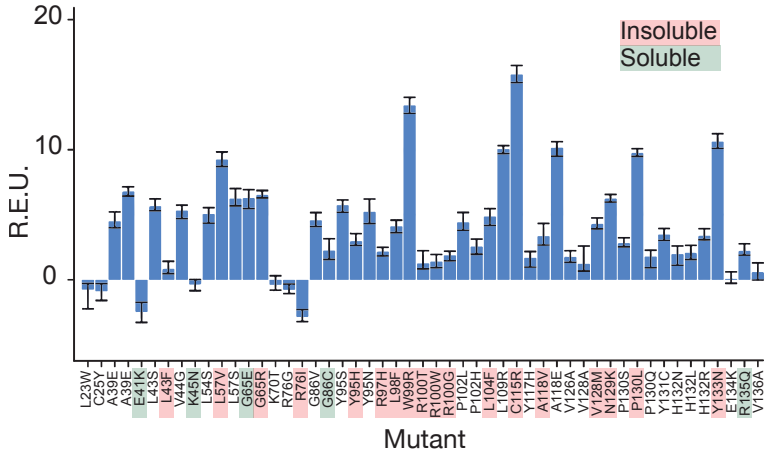


FIGURE 4.23: *In silico* stability calculations for the Smad4 MH1 domain mutants, in the COSMIC database, using the rosetta modelling suite, as referred in 3.9. R.E.U stands for rosetta energy units and approximates to the free energy change upon amino acid mutation.

Molecular dynamics simulations

Residue mean square deviation (RMSD) and the Radius of Gyration (Rg) are considered two metrics by which one could observed atomic level effects on protein stability, when analysing molecular dynamics simulations [170]. Higher RMSD and Rg values, usually correlate with a higher propensity of the mutation to affect protein stability and a subsequent role in malignant processes [170, 171]. We also performed native contact analysis (Q), which measures the loss of native contacts along a MD trajectory when compared to a reference frame (the first frame of the trajectory) with values in the range from 1 (all native contacts) to 0 (zero native contacts) [102]. Another metric that correlates with protein instability is a large hydrophobic SASA (hSASA), related to the exposure of the protein hydrophobic core, often affected in cancer mutants [170]. To a more in-depth description of molecular dynamics simulations the reader is encouraged to review section 1.2.4.

We decided to perform the simulations for the six soluble mutants as well as for three other mutants that were insoluble or highly unstable (G65R, R100T and A118V). With this we expected to sample both extremes of the protein stability landscape, for a more meaningful analysis, balancing the needed computational power.

Regarding the residue RMSD metric at 300K the most stable mutants correspond to the E41K, G65E and G86C with an average RMSD of 2.41, 2.53 and 2.76 Å respectively; followed by the WT (3.11 Å), G86C (2.76Å), K106R (2.9 Å), A118V (2.9 Å) and G65R (2.9 Å). The most flexible ones are the K45N (3.5 Å), R100T (3.8 Å) and R135Q (3.66 Å) mutants (see figure 4.24A). The average RMSD seems to oscillate more in these mutants, reflecting their higher flexibility propensity. A similar trend is observed for the Rg metric (see figure 4.24E).

Looking at the Tm and RMSD values and trying to compare them quantitatively, one would expect a negative correlation between RMSD vs Tm; lower stability usually correlates with higher RMSD values [172, 173]. In our case we observed that the RMSD of the WT (highest Tm) is similar to mutants with lower Tm values. It has been shown that some cancer missense mutations, that decrease protein stability, could decrease protein flexibility as a loss-of-function mechanism [163, 174]. Looking at the RMSF at 300K (see figure 4.25A), overall, the residues showing the highest fluctuations are located at the loops between α1 and α2, α2 and α3 helixes and β1, β2 and β4 β-sheets. The areas showing significant increased RMSF, compared to the WT, are the α3 helix for the G65E, G65R and R100T mutants (see figure 4.25A); the region comprising the β1 and β2 sheets corresponding to the R100T, A118V and to a less extent the K106R mutant; and finally, the area in proximity to the C115 residue, one of the zinc coordinating residues for this protein family fold [4, 7]. This effect was markedly increased, for the K106R mutant, with a higher RMSD, possibly affecting zinc coordination and explaining its intractability for structural and biochemical studies. Also, the R100T mutant showed higher overall RMSF values, a mutant previously reported to affected protein stability [24].

Regarding Q analysis and in agreement with the RMSD analysis, we observed that the WT loses roughly 12% of native contacts at 100 ns at 300K (see figure 4.25C), stabilising at $\approx$88%. All the mutants retain more native contacts at 100 ns, namely the E41K and G65E. At the end of the simulation the mutants retaining less contacts correspond to the insoluble mutants R135Q, K106R followed by the WT; all the other mutants cluster at around 92%. Despite this apparent loss of contacts during the simulation, possibly indicating instability, we observe that the WT is the one that maintains the most number of total salt bridges, during the 500 ns simulation (see figure 4.26A), at 300K. Also, the WT maintains the lowest hydrophobic SASA, at 300K, indicating the highest core compaction, critical for protein stability. The mutants with the highest hSASA are the insoluble R100T, K106R, A118V and the less stable G86C (see figure 4.24C). Also, the A118V has an unstable cation-Pi interaction at 300K. This mutation is located next to the F119,

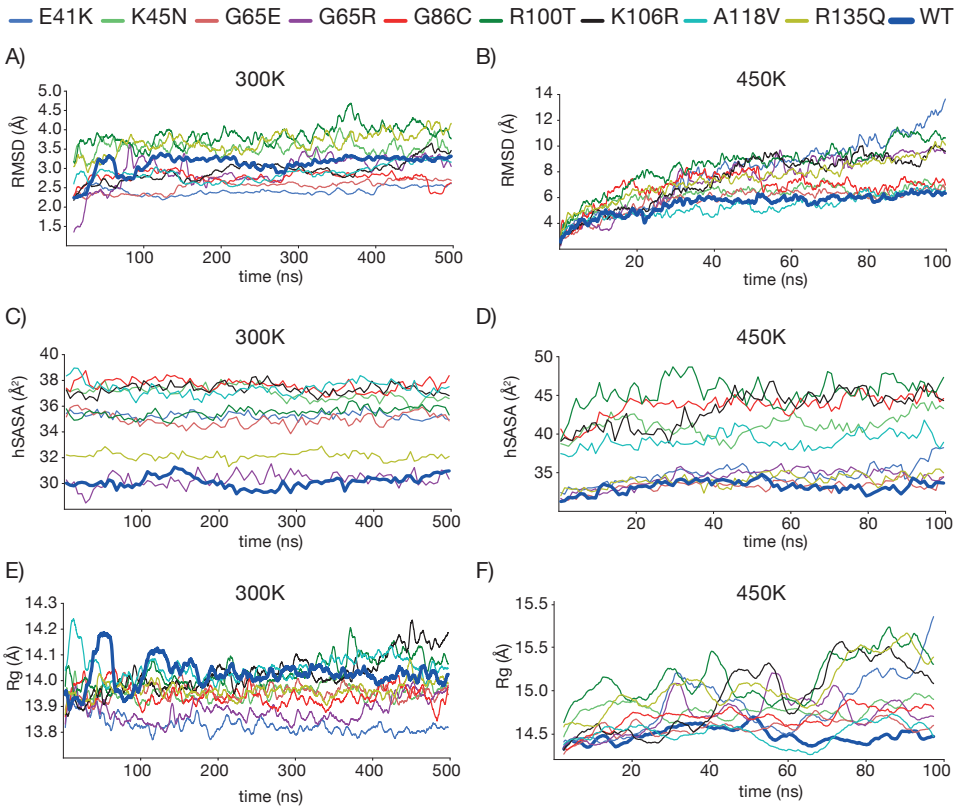


FIGURE 4.24: Molecular dynamics simulation analysis. Total time was 500 and 100 ns for the 300K and 450 K simulations respectively. **A** and **B**: Root mean squared deviation. **C** and **D**: Hydrophobic solvent accessible surface. **E** and **F**: Radius of gyration.

participating in this interaction, possible destabilising it with the bulkier Valine side-chain. These interactions are known to be critical for maintaining protein folding and stability [175].

We also performed MD simulations at a higher temperature of 450K to stress the mutants and to allow us to probe if the inherent flexibility observed for the WT (higher RMSD and Rg, as seen in figures 4.24B, 4.24F) is an intrinsic mechanism for its function. The rationale behind this was the following; by increasing the temperature we planned to simulate protein instability, establishing a parallel with an in vivo scenario, namely changes in protein concentration, increasing aggregation propensities or other mechanism that would introduce protein instability inside the cell. High temperature MD is an established metric for probing

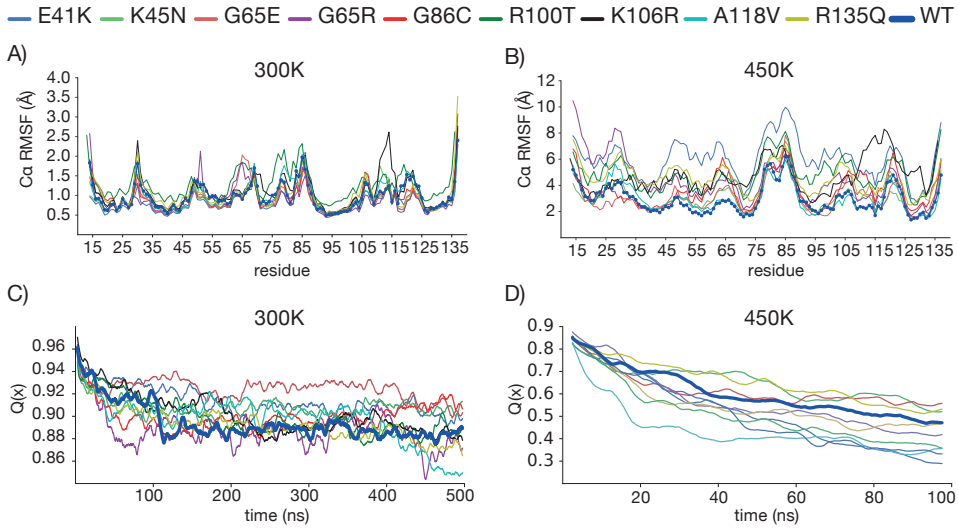


FIGURE 4.25: Molecular dynamics simulation analysis. Total time was 500 and 100 ns for the 300K and 450 K simulations respectively. **A** and **B**: Alpha carbon root mean squared fluctuation. **C** and **D**: Fraction of native contacts.

unfolding events [172, 176, 177].

The simulations at 450K showed that the WT is one of the most stable with a lower RMSD and Rg, than the rest of the mutants (see figures 4.24B, 4.24F), when compared with the 300K simulations. The A118V mutant also maintains a similar RMSD trend, as the WT, but has an unstable cation-Pi interaction and large hydrophobic areas both at 300K and 450K (see figure 4.24D). This mutant also loses more than half of its native contacts before 20 ns of simulation (see figure 4.25D). Also this residue is currently the most mutated one, in the Smad4 MH1 domain, at the COSMIC database. At 450K higher RSMF values correlate approximately to the same areas as in the 300K simulation, but with higher values as expected. Here the WT maintains the lowest per-residue variation, but the differences become more apparent at 450K, with all the mutants shifting to higher values. Interestingly, the E41K mutant that has a lower RMSD and RMSF at 300K, shows the highest RMSF at 450K (see figure 4.25B). Regarding the protein hydrophobic core, given by the hSASA in figures 4.24C and 4.24D, the WT is the one were the hydrophobic core is less exposed both at 300 and 450K. Taking the above results together we can assume that the WT protein, when destabilised, is able to decrease its flexibility and maintain its hydrophobic core intact, when compared to the mutants.

We also performed the molecular dynamics simulations in the presence of the

SBE dsDNA (double stranded DNA) for the mutants and for the WT protein. The WT S4MH1 bound to DNA did not show any major distance fluctuations at the protein-DNA interface, represented by residue pairs R81-G135, Q83-A155, K88-G156 and K88-A157. These residues have been shown to be critical for maintaining the protein-DNA complex [4, 7]. With respect to the mutated proteins, some minor fluctuations were observed for the pair K88-A157, in mutants K106R and E41K, that seem to be non-significant in the overall dynamics distance fluctuations (see figure 4.21B). While the distances between DNA binding key residues does not change and these mutants are not directly involved in DNA binding, albeit in the vicinity of the protein DNA-interface, they seem to introduce an overall small destabilization away from the β-hairpin.

Trying to reconciliate the experimental protein stability results with the molecular simulations, we performed regression tests of the average metrics, with respect to the melting temperatures. The only metric that gave a strong correlation was the salt bridges one, both at 300 and 450K (see figure 4.26B) with a correlation of 0.93 and 0.84, respectively. Salt bridges increase in thermophilic proteins versus their mesophilic counterparts [178] and changes in their geometry can lead to tumorigenic processes.

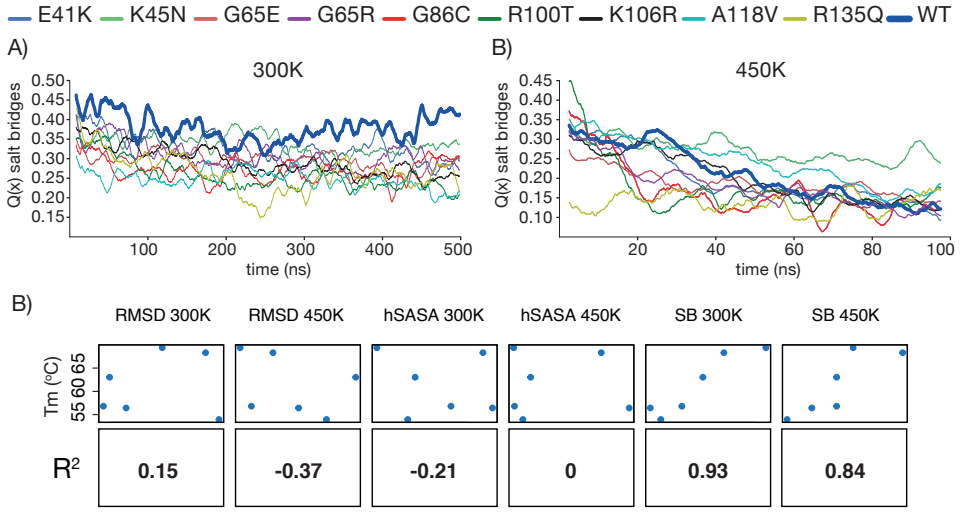


FIGURE 4.26: **A**: Molecular dynamics analysis for the salt bridges fraction of native contacts. **B**: *Top panel*: Average RMSD, hSASA and salt bridges (SB) at 300 and 400K versus the melting temperature for the WT and soluble mutants. *Bottom panel*: Coefficient of determination for the data described above.

Maintaining a network of stable salt bridges both at 300K and 450K (e.g K45N) seems to be the most important metric (see figure 4.26B), even if other critical metrics are affected (e.g K45N hSASA), when comparing the simulations data to the melting temperatures. Even if E41 and R135 are the only mutant residues directly involved in salt bridges, allosteric disruption, of a network of salt bridges, is a mechanism through which malignant processes could be developed, as postulated for other systems [170].

Taking all of the above results together, the mutants seem to exert their effects in different, and not immediately apparent ways. The ones we could study, maintain DNA biding functionality while introducing instability through different mechanisms, either by destabilising the protein core (e.g. R100T, K106R), critical cation-Pi interactions (A118V), affecting overall flexibility (R135Q, R100T) and disrupting salt bridges (G86C, R100T, K106R, R135Q). A paradigmatic case of the Smad4 MH1 mutational landscape complexity is revealed by the G65E, G65R mutant pair. The G65E mutant was soluble and the G65R insoluble. This glycine is located at a loop between $\alpha 2$ and $\beta 1$ and some flexibility is expected. The G65E mutant is more flexible, than the G65R and maintains more native contacts (see figure 4.25A), possibly due its less bulky side-chain. Even with the analysis presented above, the mechanism of the dramatic stability shifts observed, for the G65E and G65R mutants, is not completely clear.

4.1.6 Implications for TGFβ signalling

Stressing the importance in the analysis of different metrics for accessing mutational effects, cancer mutations can affect proteins in different ways; they can affect stability, protein-protein interactions, allosteric regulations and protein post-translational modification sites [179]. These effects could be amplified when a destabilization is introduced into the system, here represented by an increased temperature of 450K.

The decreased stability accompanied by an increased flexibility, in the performed non-equilibrium molecular dynamics simulations, could suggest that the Smad4 MH1 mutants may be involved in a wider network of protein interactions, including for example the protein-ubiquination pathway, that depends on a particular and very well controlled cellular environment, were protein destabilization plays a part. As shown above the direct binding of Smad4 to DNA is not affected, so, probably, altering DNA binding functionality is not the principal mechanism by which these mutants alter TGFβ functionality.

Recently a genome wide survey revealed the most frequently mutated residues in the COSMIC database [180]. Analysing 23 cancers simultaneously, the top dominant mutations were: *E-K*, R-H, *R-Q*, R-C, *A-V*, A-T, D-N, *P-L*, *R-W*, and *G-R*, in decrescent order of frequency. More than half of the analyzed mutations in this

work (in italic) were present in the studied system and all were insoluble or destabilizing, reinforcing that the findings reported in this work could be applicable to other systems. Residues affected are usually hydrophobic and/or charged, stressing the importance of hydrophobic interactions and salt bridges in maintaining protein stability, as stated above. As recently established and also stated above, post-translation modifications could also affect DNA binding [181].

The Smad4 MH1 domain melting temperature seems to be in the higher spectra of a recently reported Tm distribution for the human proteome, with a maximum of ≈70ºC [169]. Misfolding, induced by a defaulting translation, could be partially prevented by increasing protein tolerance to mutations [182]. Due to the high to medium Tm values observed, even for the cancer mutants, the high sequence conservation in Smad4 and the differential effects observed in our simulations, variations in local Smad4 protein concentration and/or post-translational modifications (e.g. oxidation), as recently postulated [169, 183], could be another possible mechanism by which cancer mutations modulate TGFβ activity. This could decrease protein half-life, a reporter of protein stability, as previously shown for the R100T insoluble mutant [24, 184]. Also the effect of the mutations could be beneficial or detrimental depending on the cellular context and the tumour suppressor/activator activity of Smad4.

This work is the first, to our knowledge, atomic level mutational landscape information for this class of zinc finger proteins. Deciphering the mechanisms by which cancer mutants affect TGFβ activation will be an ongoing topic of research. The work presented here could lay the base for future structural and biochemical studies.

Chapter 5

Conclusions

The work presented in this thesis is mainly divided in two parts, the first part inquired the conformational landscapes of human full-length Smad proteins, while the second part analyzed the effect of cancer mutations in the Smad4 MH1 domain. For the first part we developed and/or modified software for data analysis and established production protocols for each representative of the Smad protein family and for the second part, by coalescing biochemistry experiments with multi-temperature molecular dynamics simulations, we found that:

- The inter-domain linkers of Smad2 and Smad4 behave as intrinsically disordered proteins.
- Smad4 full-length is a monomeric and flexible protein in an equilibrium between elongated and more compact conformations.
- In solution Smad4 is not predominantly in an auto-inhibited conformation and the MH1 and MH2 are independently functioning domains.
- Smad2 is an oligomeric protein populating a monomer-dimer-trimer equilibria shaped by phosphorylation. Phosphorylation and MH1 deletion shifts the equilibrium towards trimer formation, while deletion of the C-terminal phosphosite tail abolishes trimer formation. The inter-domain association is concentration-dependent.
- Smad7 is an unstable multi-domain protein.
- Smad4 MH1 mutations mainly affect charged and hydrophobic residues.
- Cancer mutations seem to affect protein stability while maintaining DNA binding functionality.
- By comparing melting temperature analysis and molecular dynamics simulations, we proposed a mutational landscape mechanism for the Smad4 MH1 domain that exerts its effects by disrupting salt-bridge networks.

Overall we established a framework for describing Smad protein structure from an intra- and inter-molecular perspective.

Appendix A

Results

TABLE A.1: Plate layout for the buffer optimisation experiments, using the Slice pH™ kit from Hampton research, for the Smad7 constructs.

Well	Buffer	pH
A01	Citric acid	3.5
A02	Citric acid	3.8
A03	Citric acid	4.1
A04	Citric acid	4.4
A05	Sodium citrate tribasic $2H_2O$	3.6
A06	Sodium citrate tribasic $2H_2O$	3.9
A07	Sodium citrate tribasic $2H_2O$	4.2
A08	Sodium citrate tribasic $2H_2O$	4.5
A09	Sodium acetate $3H_2O$	3.7
A10	Sodium acetate $3H_2O$	4.0
A11	Sodium acetate $3H_2O$	4.3
A12	Sodium acetate $3H_2O$	4.6
B01	Sodium acetate $3H_2O$	4.9
B02	DL-Malic acid	4.7
B03	DL-Malic acid	5.0
B04	DL-Malic acid	5.3
B05	DL-Malic acid	5.6
B06	DL-Malic acid	5.9
B07	Succinic acid	4.8
B08	Succinic acid	5.1
B09	Succinic acid	5.4
B10	Succinic acid	5.7
B11	Succinic acid	6.0
B12	Sodium cacodylate $3H_2O$	5.2
C01	Sodium cacodylate $3H_2O$	5.5
C02	Sodium cacodylate $3H_2O$	5.8
C03	Sodium cacodylate $3H_2O$	6.1

C04	Sodium cacodylate $3H_2O$	6.4
C05	MES H_2O	5.3
C06	MES H_2O	5.6
C07	MES H_2O	5.9
C08	MES H_2O	6.2
C09	MES H_2O	6.5
C10	Bis-tris	5.7
C11	Bis-tris	6.0
C12	Bis-tris	6.3
D01	Bis-tris	6.6
D02	Bis-tris	6.9
D03	ADA	5.8
D04	ADA	6.1
D05	ADA	6.4
D06	ADA	6.7
D07	ADA	7.0
D08	Imidazole	6.2
D09	Imidazole	6.5
D10	Imidazole	6.8
D11	Imidazole	7.1
D12	Imidazole	7.4
E01	Bis-tris propane	6.4
E02	Bis-tris propane	6.7
E03	Bis-tris propane	7.0
E04	Bis-tris propane	7.3
E05	MOPS	6.5
E06	MOPS	6.8
E07	MOPS	7.1
E08	MOPS	7.4
E09	MOPS	7.7
E10	HEPES Sodium	6.6
E11	HEPES Sodium	6.9
E12	HEPES Sodium	7.2
F01	HEPES Sodium	7.5
F02	HEPES	6.8
F03	HEPES	7.1
F04	HEPES	7.4
F05	HEPES	7.7
F06	Tris hydrochloride	7.2
F07	Tris hydrochloride	7.5
F08	Tris hydrochloride	7.8
F09	Tris hydrochloride	8.1
F10	Tris	7.3
F11	Tris	7.6

F12	Tris	7.9
G01	Tris	8.2
G02	Tris	8.5
G03	Tricine	7.4
G04	Tricine	7.7
G05	Tricine	8.0
G06	Tricine	8.3
G07	Tricine	8.6
G08	Bicine	7.5
G09	Bicine	7.8
G10	Bicine	8.1
G11	Bicine	8.4
G12	Bicine	8.7
H01	Bis-tris propane	8.5
H02	Bis-tris propane	8.8
H03	Bis-tris propane	9.1
H04	Bis-tris propane	9.4
H05	Glycine	8.6
H06	Glycine	8.9
H07	Glycine	9.2
H08	Glycine	9.5
H09	AMPD	8.7
H10	AMPD	9.0
H11	AMPD	9.3
H12	AMPD	9.6

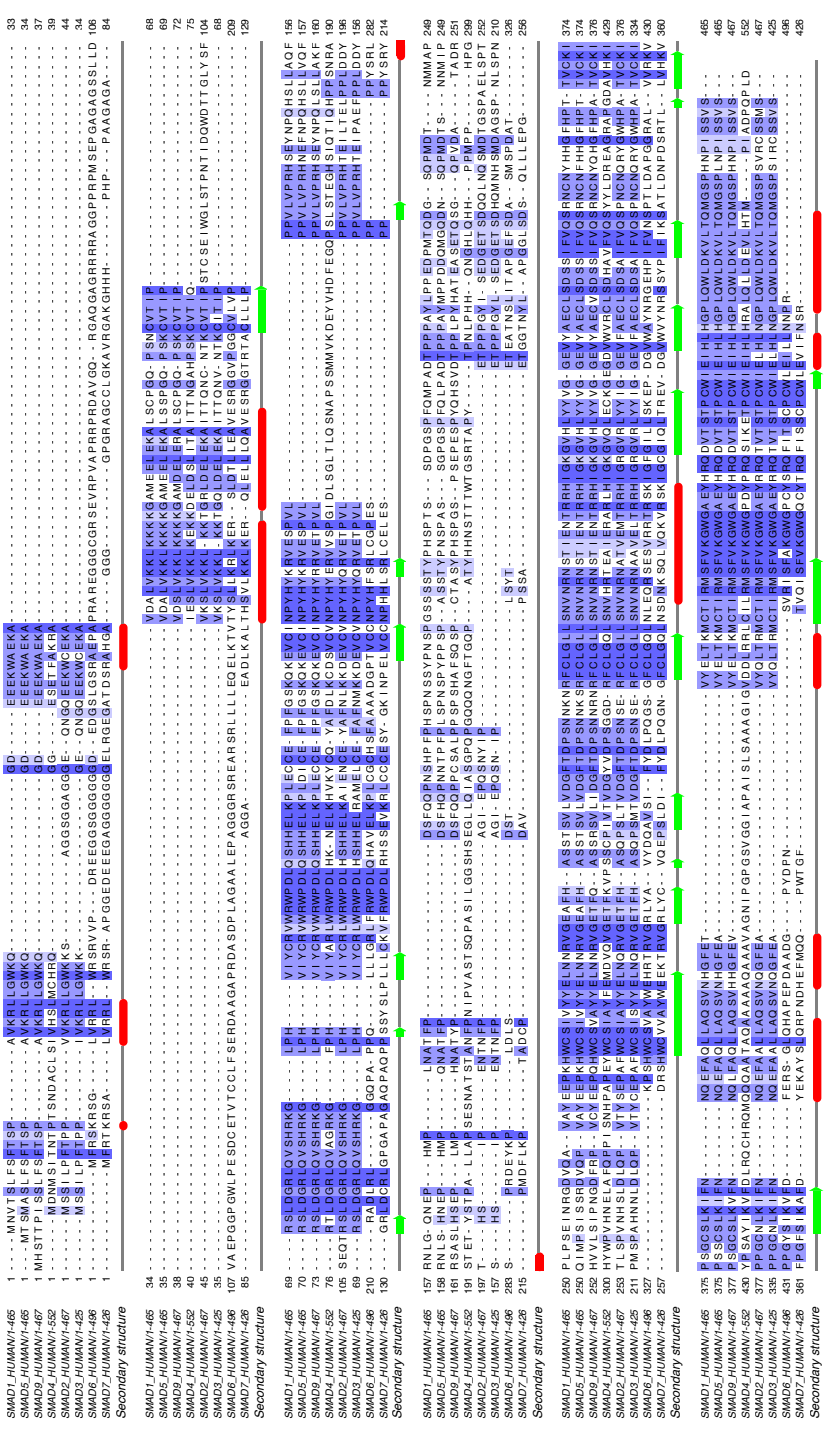


FIGURE A.1: Smads sequence alignment. Similarity is represented by shades of blue with a cutoff of 30%. The sequence names obey to the following rule: Uniprot id/sequence numbering. Secondary structure was calculated with jpred [185]. Sequence representation was done using jailview [186].

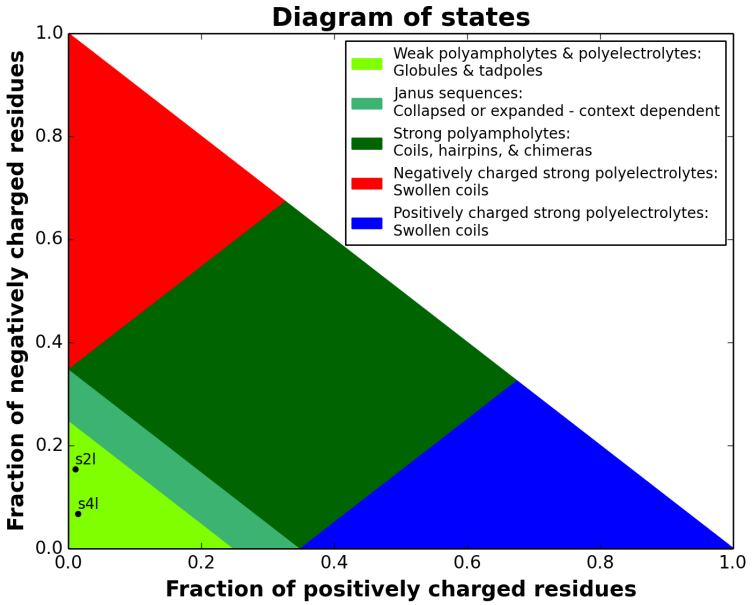


FIGURE A.2: Phase diagram for Smad2 and Smad4 inter domain linkers calculated using CIDER [115].

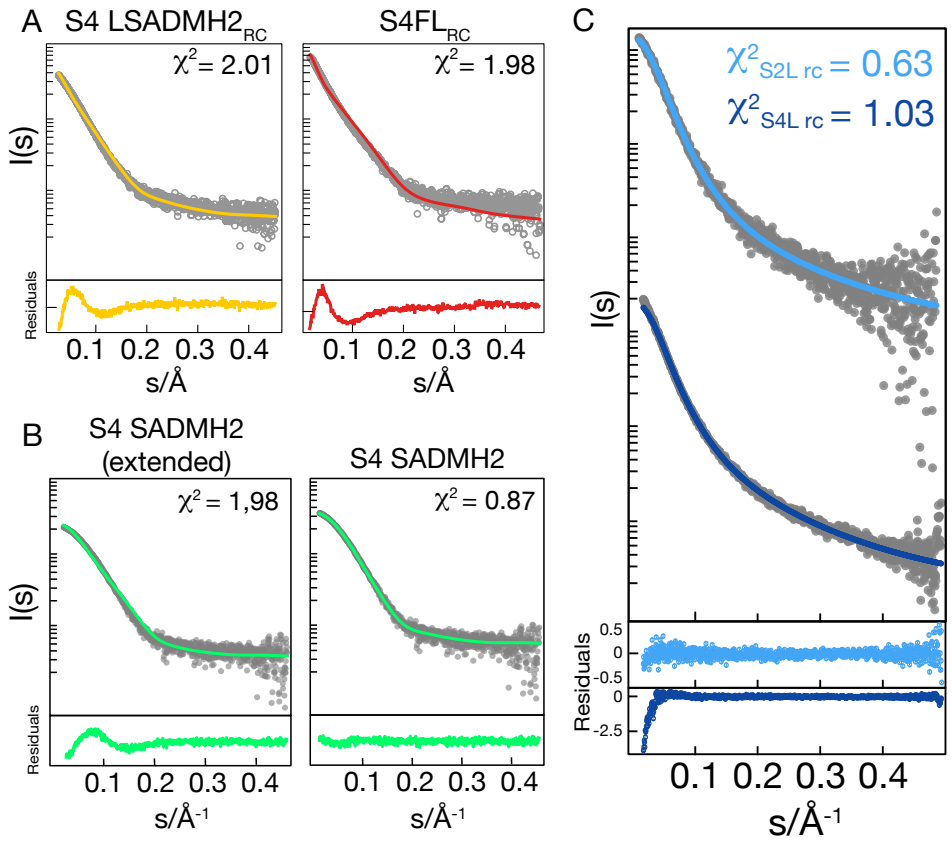


FIGURE A.3: **A**: SAXS random coil fits for the S4LSADMH2, in yellow and the S4FL, in red, constructs. **B**: SAXS fits for the extended and compact S4SADMH2 models. **C**: SAXS random coil fits for the Smad2 linker, in light blue and Smad4 linker in dark blue.

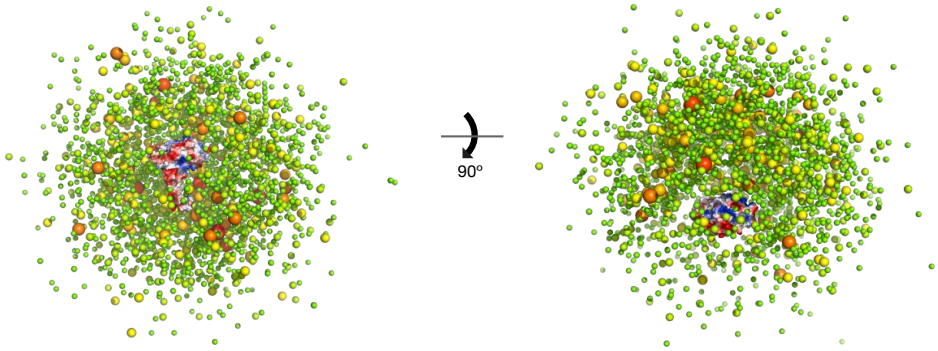


FIGURE A.4: Smad4 random pool ensemble, generated with flexible-meccano [59], for SAXS data filtering. In electrostatic representation is the MH2 domain and in sphere representation is centre-of-mass of the MH1 domain. The sphere radius is proportional to the number conformations was selected spanning from green (lowest) to red (highest).

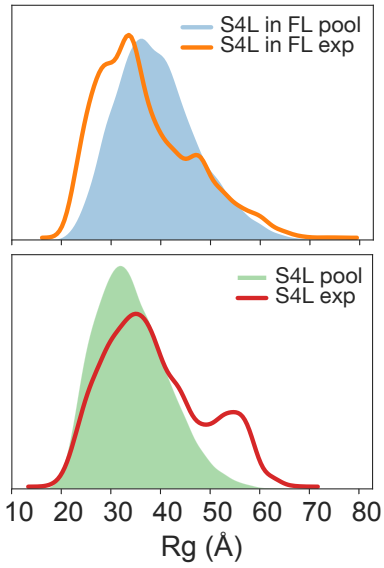


FIGURE A.5: Smad4 linker (S4L) radius of gyration in isolation or in full-length context. S4L pool and S4L exp SAXS data is the same as in figure 4.1. S4L in FL is the linker radius of gyration in a full-length context.

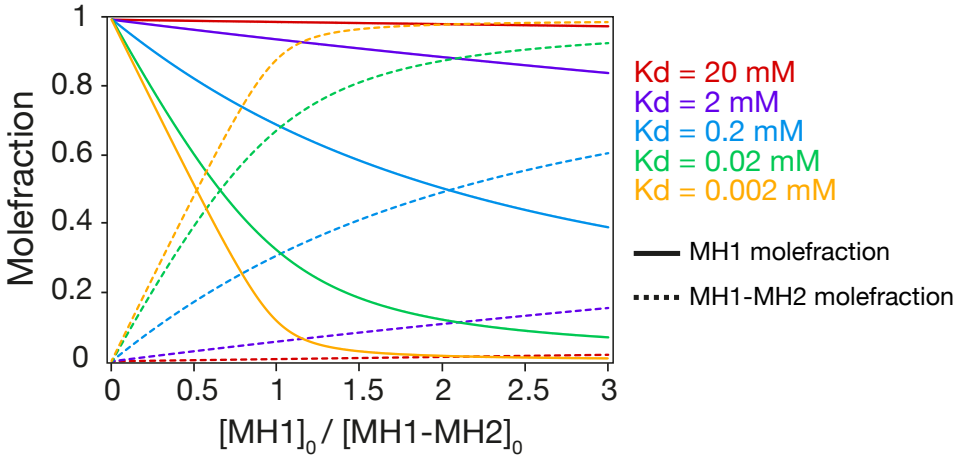


FIGURE A.6: NMR titration simulations for the Smad4 MH1-MH2 interaction. The one-site binding model was used and five binding constants were simulated, for a molar excess of 3 equivalents.

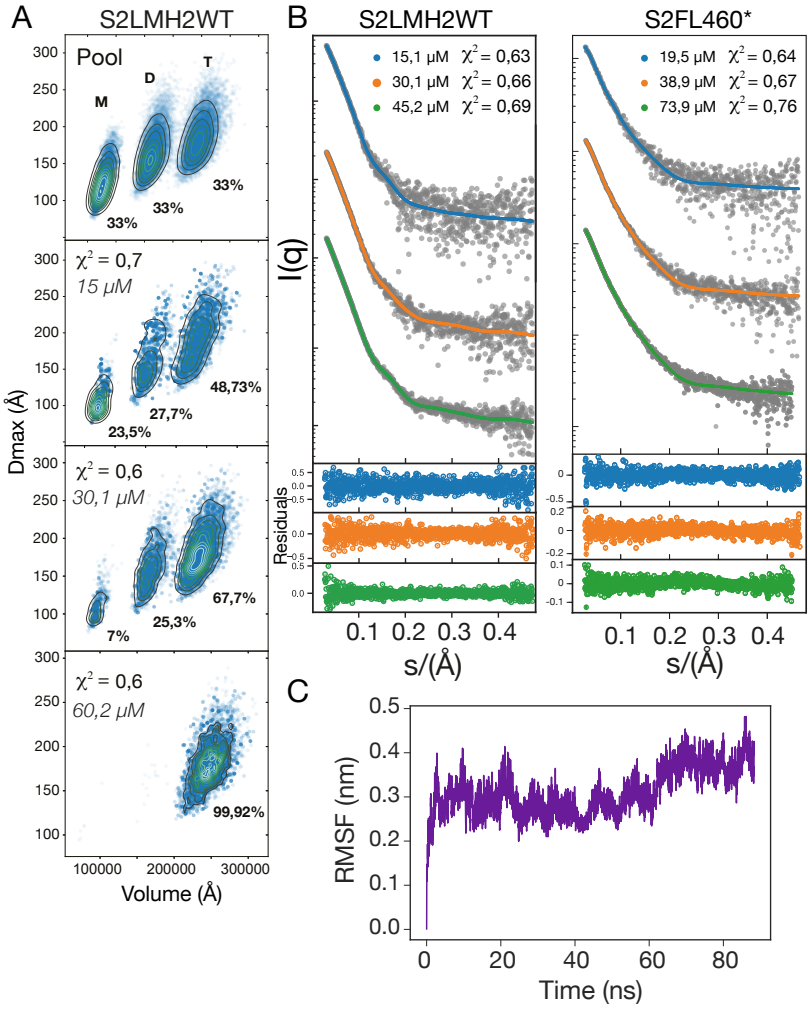


FIGURE A.7: SAXS data analyses of the S2LMH2WT and S2FL460* constructs. **A**: EOM-derived ensemble population analysis for the S2LMH2WT construct derived from **B**, left panel. M, D, T are monomer, dimer and trimer, respectively. Dmax is the maximum distance and volume is the volume for each conformation. **B**: EOM fits to the experimental SAXS profiles are in blue, orange and green with the corresponding residuals below. Experimental profiles are in grey. **C**: Molecular dynamics simulation of the S2MH2 (PDB:1KHX) putative dimer interface. RMSF is the root mean squared fluctuation and total time was 85 ns. Simulation protocol followed the one described in section 3.

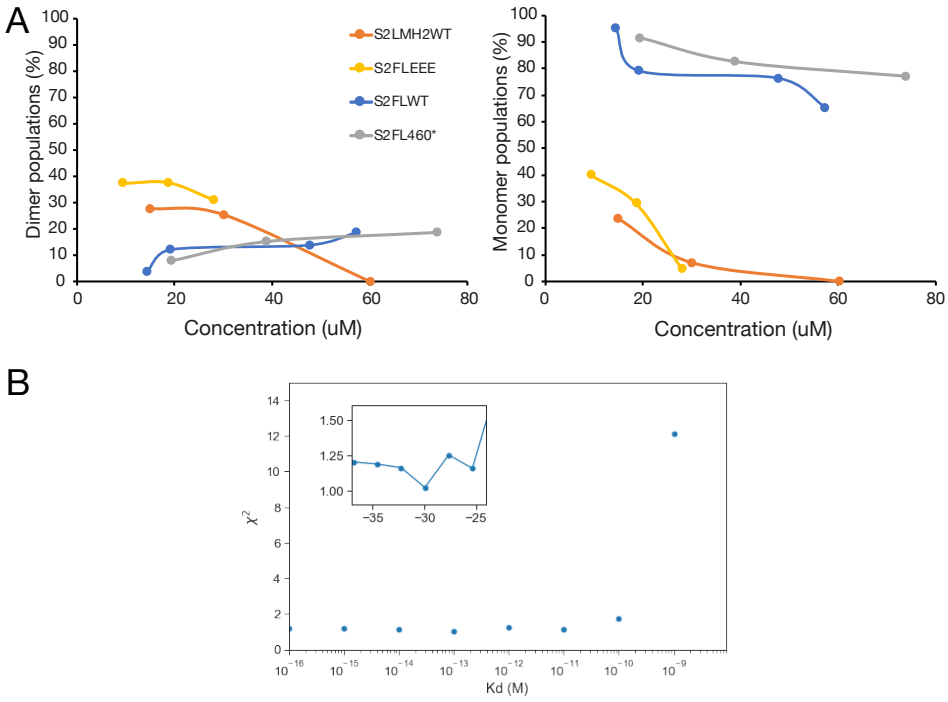


FIGURE A.8: Smad2 monomer and dimer populations and S2LMH2WT affinity estimates. **A**: Monomeric and dimeric fractions determined as in figure 4.11C. **B**: Affinity estimates for the S2LMH2WT protein determined as in figure 4.11D. The plot inlet is an augmented version of the last six points of the original plot, from left to right.

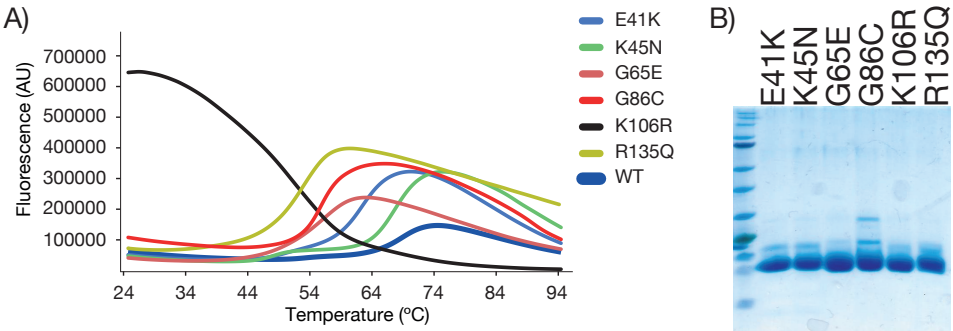


FIGURE A.9: **A**: Differential scanning fluorimery profiles for the Smad4 MH1 WT and soluble mutant proteins. Fluorescence is in arbritary units and temperature in Celsius. **B**: SDS-PAGE gels of the soluble mutants.

Appendix B

Materials and methods

TABLE B.1: Expression media.

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

SOC (0.5 L)	
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)	
LB Broth	25 g
Agar	15 g
Desired antibiotic:	
Ampicillin	100 μg/mL
Kanamycin	50 μg/mL

TABLE B.2: Expression media for labelled proteins.

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

M9 medium 1 L (10x)	
Na_2HPO_4	60 g
KH_2PO_4	30 g
NaCl	5 g
$^{15}NH_4Cl$	5 g

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

TABLE B.3: Protein purification buffers.

Lysis (1 L)	
TRIS	50 mM
NaCl	400 mM
Tween20	0.1 % (v/v)
Imidazole	400 mM
DNase	0.01 mg ml^{-1}
Lysozyme	0.25 mg ml^{-1}
SigmaFast (10x)	1x
PMSF	0.1 mM
TCEP	1 mM
Refolding buffer:	
TRIS	50 mM
NaCL	400 mM
pH	8
Urea	8 M
Imidazole	40 mM
TCEP	1 mM

Elution	
Smad2 (S2EB)	
Buffer A:	
TRIS	50 mM
NaCl	400 mM
Imidazole	40 mM
SigmaFast (10x)	1x
PMSF	0.1 mM
pH	7.5
TCEP	1 mM
Buffer B:	
TRIS	50 mM
NaCl	400 mM
Imidazole	1 M
pH	7.5
Smad4 (S4EB)	
Buffer A:	
TRIS	50 mM
NaCl	400 mM
SigmaFast (10x)	1x
PMSF	0.1 mM
Buffer B:	
TRIS	50 mM
NaCl	400 mM
Desthiobiotin	2.5 mM
SigmaFast (10x)	1x
PMSF	0.1 mM

Ion exchange	
Buffer A:	
TRIS	50 mM
NaCl	< 50 mM
pH	6 - 7.5
Buffer B:	
TRIS	50 mM
NaCl	1 M
pH	6 - 7.5

Gel filtration	
TRIS	50 mM
NaCl	150 mM
TCEP	1 mM
pH	6 - 7.5

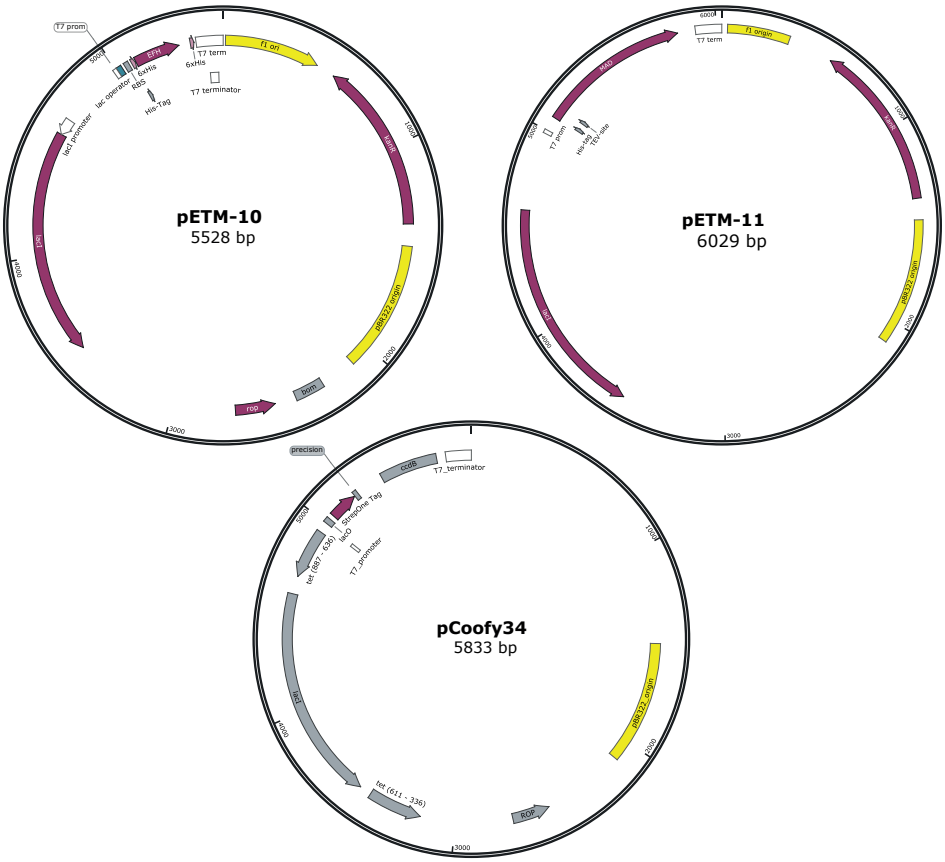


FIGURE B.1: Plasmids used for protein production for the Smad2 and Smad4 constructs.

Bibliography

[1] Charles J David and Joan Massagué. "Contextual determinants of TGF action in development, immunity and cancer." In: *Nature reviews. Molecular cell biology* 19.7 (July 2018), pp. 419–435.

[2] Pieter J A Eichhorn et al. "USP15 stabilizes TGF- receptor I and promotes oncogenesis through the activation of TGF- signaling in glioblastoma." In: *Nature medicine* 18.3 (Feb. 2012), pp. 429–435.

[3] Akiko Hata and Ye-Guang Chen. "TGF- Signaling from Receptors to Smads." In: *Cold Spring Harbor Perspectives in Biology* 8.9 (Sept. 2016).

[4] Maria J Macias, Pau Martín-Malpartida, and Joan Massagué. "Structural determinants of Smad function in TGF- signaling." In: *Trends in biochemical sciences* 40.6 (June 2015), pp. 296–308.

[5] Eric Aragón et al. "A Smad action turnover switch operated by WW domain readers of a phosphoserine code." In: *Genes & development* 25.12 (June 2011), pp. 1275–1288.

[6] Pinglong Xu, Jianming Liu, and Rik Derynck. "Post-translational regulation of TGF- receptor and Smad signaling." In: *FEBS letters* 586.14 (July 2012), pp. 1871–1884.

[7] Nithya Baburajendran et al. "Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers." In: *Nucleic acids research* 39.18 (Oct. 2011), pp. 8213–8222.

[8] E Gail Hutchinson and Janet M Thornton. "The Greek key motif: extraction, classification and analysis". In: *Protein Engineering Design and Selection* 6.3 (1993), pp. 233–245.

[9] G Wu et al. "Structural basis of Smad2 recognition by the Smad anchor for receptor activation." In: *Science (New York, N.Y.)* 287.5450 (Jan. 2000), pp. 92–97.

[10] Ken-ichi Miyazono et al. "Hydrophobic patches on SMAD2 and SMAD3 determine selective binding to cofactors". In: *Science signaling* (Mar. 2018).

[11] Shuting Bai and Xu Cao. "A nuclear antagonistic mechanism of inhibitory Smads in transforming growth factor-beta signaling." In: *The Journal of biological chemistry* 277.6 (Feb. 2002), pp. 4176–4182.

[12] P Kavsak et al. “Smad7 binds to Smurf2 to form an E3 ubiquitin ligase that targets the TGF beta receptor for degradation.” In: *Molecular cell* 6.6 (Dec. 2000), pp. 1365–1375.

[13] Xiaohua Yan, Ziying Liu, and Yeguang Chen. “Regulation of TGF-beta signaling by Smad7.” In: *Acta Biochimica et Biophysica Sinica* 41.4 (Apr. 2009), pp. 263–272.

[14] Akiko Hata et al. “Mutations increasing autoinhibition inactivate tumour suppressors Smad2 and Smad4”. In: *Nature* 388.6637 (July 1997), pp. 82–87.

[15] Xiaohua Yan et al. “Smad7 Interacts with R-Smads to Inhibit TGF-beta/Smad Signaling.” In: *The Journal of biological chemistry* (Nov. 2015).

[16] Joan Massagué. “TGF signalling in context.” In: *Nature reviews. Molecular cell biology* 13.10 (Oct. 2012), pp. 616–630.

[17] Caroline S Hill. “Transcriptional Control by the SMADs”. In: *Cold Spring Harbor Perspectives in Biology* 8.10 (July 2016), a022079.

[18] L Jayaraman and J Massagué. “Distinct oligomeric states of SMAD proteins in the transforming growth factor-beta pathway.” In: *The Journal of biological chemistry* 275.52 (Dec. 2000), pp. 40710–40717.

[19] Gareth J Inman and Caroline S Hill. “Stoichiometry of active smad-transcription factor complexes on DNA.” In: *The Journal of biological chemistry* 277.52 (Dec. 2002), pp. 51008–51016.

[20] B M Chacko et al. “The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization.” In: *Nature Structural Biology* 8.3 (Mar. 2001), pp. 248–253.

[21] Benoy Maramparambil Chacko. *The Structural Basis for the Phosphorylation-induced Activation of Smad Proteins*. 2004.

[22] Aristidis Moustakas and Carl-Henrik Heldin. “From mono- to oligo-Smads: the heart of the matter in TGF-beta signal transduction.” In: *Genes & development* 16.15 (Aug. 2002), pp. 1867–1871.

[23] Ewelina Guca et al. “TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF- signaling.” In: *Nucleic acids research* 46.17 (Sept. 2018), pp. 9220–9235.

[24] J Xu and L Attisano. “Mutations in the tumor suppressors Smad2 and Smad4 inactivate transforming growth factor beta signaling by targeting Smads to the ubiquitin-proteasome pathway.” In: *Proceedings of the National Academy of Sciences of the United States of America* 97.9 (Apr. 2000), pp. 4820–4825.

[25] Pau Martín-Malpartida et al. “Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors.” In: *Nature communications* 8.1 (Dec. 2017), p. 2070.

[26] Claudio Alarcón et al. "Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways." In: *Cell* 139.4 (Nov. 2009), pp. 757–769.

[27] Nicola Waddell et al. "Whole genomes redefine the mutational landscape of pancreatic cancer." In: *Nature* 518.7540 (Feb. 2015), pp. 495–501.

[28] Sang Cheul Oh et al. "Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype." In: *Nature communications* 9.1 (May 2018), p. 1777.

[29] Rosemary J Akhurst. "Targeting TGF- Signaling for Therapeutic Gain." In: *Cold Spring Harbor Perspectives in Biology* 9.10 (Oct. 2017).

[30] Daniele V F Tauriello et al. "TGF drives immune evasion in genetically reconstituted colon cancer metastasis." In: *Nature* 554.7693 (Feb. 2018), pp. 538–543.

[31] Masafumi Inui et al. "USP15 is a deubiquitylating enzyme for receptor-activated SMADs." In: *Nature Cell Biology* 13.11 (Sept. 2011), pp. 1368–1375.

[32] Daniela Rotin and Sharad Kumar. "Physiological functions of the HECT family of ubiquitin ligases." In: *Nature reviews. Molecular cell biology* 10.6 (June 2009), pp. 398–409.

[33] David A Jacques and Jill Trewhella. "Small-angle scattering for structural biology–expanding the frontier while avoiding the pitfalls." In: *Protein Science* 19.4 (Apr. 2010), pp. 642–657.

[34] Christopher D Putnam et al. "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution." In: *Quarterly Reviews of Biophysics* 40.3 (Aug. 2007), pp. 191–285.

[35] Igor N Serdyuk, Nathan R Zaccai, and Joseph Zaccai. *Methods in molecular biophysics: structure, dynamics, function*. Cambridge: Cambridge University Press, 2007.

[36] Dmitri I Svergun et al. *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford University Press, Aug. 2013.

[37] P Debye. "Zerstreuung von Röntgenstrahlen". In: *Ann Phys* 351.6 (1915), pp. 809–823.

[38] André Guinier. "La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques". In: *Ann Phys (Paris)* 11.12 (1939), pp. 161–237.

[39] Maxim V Petoukhov et al. "New developments in the ATSAS program package for small-angle scattering data analysis." In: *Journal of Applied Crystallography* 45.Pt 2 (Apr. 2012), pp. 342–350.

[40] Robert P Rambo and John A Tainer. "Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law." In: *Biopolymers* 95.8 (Aug. 2011), pp. 559–571.

[41] Robert P Rambo and John A Tainer. "Accurate assessment of mass, models and resolution by small-angle scattering." In: *Nature* 496.7446 (Apr. 2013), pp. 477–481.

[42] Nelly R Hajizadeh et al. "Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data." In: *Sci Rep* 8.1 (May 2018), p. 7204.

[43] D Svergun, C Barberato, and M H J Koch. "CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates". In: *J Appl Crystallogr* 28.6 (Dec. 1995), pp. 768–773.

[44] Petra Pernot et al. "Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution." In: *Journal of synchrotron radiation* 20.Pt 4 (July 2013), pp. 660–664.

[45] M P Williamson, T F Havel, and K Wüthrich. "Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry." In: *Journal of Molecular Biology* 182.2 (Mar. 1985), pp. 295–315.

[46] I I Rabi et al. "A new method of measuring nuclear magnetic moment". In: *Phys. Rev.* 53.4 (Feb. 1938), pp. 318–318.

[47] III Arthur G Palmer et al. *Protein NMR Spectroscopy: Principles and Practice*. 2nd. Academic Press, Nov. 2006.

[48] Quincy Teng. *Structural Biology: Practical NMR Applications*. 2nd ed. 2013. Springer, Sept. 2012.

[49] Gordon S Rule and T Kevin Hitchens. *Fundamentals of Protein NMR Spectroscopy (Focus on Structural Biology)*. 2006th. Springer, Dec. 2005.

[50] Yang Shen et al. "TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts." In: *Journal of Biomolecular NMR* 44.4 (Aug. 2009), pp. 213–223.

[51] Sjoerd J de Vries, Marc van Dijk, and Alexandre M J J Bonvin. "The HADDOCK web server for data-driven biomolecular docking." In: *Nature Protocols* 5.5 (May 2010), pp. 883–897.

[52] Francois-Xavier Theillet et al. "Cell signaling, post-translational protein modifications and NMR spectroscopy." In: *Journal of Biomolecular NMR* 54.3 (Nov. 2012), pp. 217–236.

[53] Steven M Pascal and Jennie M McKelvie. *NMR Primer: An HSQC-based Approach*. 1st. IM Publications LLP, June 2008.

[54] Michaeleen Doucleff, Mary Hatcher-Skeers, and Nicole J Crane. *Pocket Guide to Biomolecular NMR*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[55] M Sattler. "Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients". In: *Progress in nuclear magnetic resonance spectroscopy* 34.2 (Mar. 1999), pp. 93–158.

[56] Jaka Kragelj, Martin Blackledge, and Malene Ringkj Jensen. "Ensemble calculation for intrinsically disordered proteins using NMR parameters." In: *Advances in experimental medicine and biology* 870 (2015), pp. 123–147.

[57] Nicholas D Keul et al. "The entropic force generated by intrinsically disordered segments tunes protein function." In: *Nature* 563.7732 (Nov. 2018), pp. 584–588.

[58] Payel Das, Silvina Matysiak, and Jeetain Mittal. "Looking at the Disordered Proteins through the Computational Microscope." In: *ACS Cent Sci* 4.5 (May 2018), pp. 534–542.

[59] Valéry Ozenne et al. "Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables." In: *Bioinformatics* 28.11 (June 2012), pp. 1463–1470.

[60] Howard J Feldman and Christopher W V Hogue. "Probabilistic sampling of protein conformations: new hope for brute force?" In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (Jan. 2002), pp. 8–23.

[61] Mickaël Krzeminski et al. "Characterization of disordered proteins with ENSEMBLE." In: *Bioinformatics* 29.3 (Feb. 2013), pp. 398–399.

[62] Loïc Salmon et al. "NMR characterization of long-range order in intrinsically disordered proteins." In: *Journal of the American Chemical Society* 132.24 (June 2010), pp. 8407–8418.

[63] Massimiliano Bonomi et al. "Principles of protein structural ensemble determination." In: *Current opinion in structural biology* 42 (Jan. 2017), pp. 106–116.

[64] Daniel K Putnam, Edward W Lowe, and Jens Meiler. "Reconstruction of SAXS Profiles from Protein Structures." In: *Computational and Structural Biotechnology Journal* 8 (Nov. 2013), e201308006.

[65] Gunnar F Schröder. "Hybrid methods for macromolecular structure determination: experiment with expectations." In: *Current opinion in structural biology* 31 (Apr. 2015), pp. 20–27.

[66] Giancarlo Tria et al. "Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering." In: *IUCrJ* 2.Pt 2 (Mar. 2015), pp. 207–217.

[67] Pau Bernadó et al. "Structural characterization of flexible proteins using small-angle X-ray scattering." In: *Journal of the American Chemical Society* 129.17 (May 2007), pp. 5656–5664.

[68] L D Antonov et al. "Bayesian inference of protein ensembles from SAXS data." In: *Physical chemistry chemical physics : PCCP* (Nov. 2015).

[69] David H Brookes and Teresa Head-Gordon. "Experimental inferential structure determination of ensembles for intrinsically disordered proteins." In: *Journal of the American Chemical Society* 138.13 (Apr. 2016), pp. 4530–4538.

[70] Steen Hansen. "Bayesian methods in SAXS and SANS structure determination". In: *Bayesian methods in structural bioinformatics*. Ed. by Thomas Hamelryck, Kanti Mardia, and Jesper Ferkinghoff-Borg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 313–342.

[71] E Schrödinger. "Quantisierung als Eigenwertproblem". In: *Ann Phys* 384.4 (1926), pp. 361–376.

[72] Oren M Becker and Martin Karplus. *Guide to Biomolecular Simulations (Focus on Structural Biology)*. Softcover reprint of the original 1st ed. 2006. Springer, Aug. 2016.

[73] David A Case et al. "The Amber biomolecular simulation programs." In: *Journal of Computational Chemistry* 26.16 (Dec. 2005), pp. 1668–1688.

[74] Wendy D Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". In: *Journal of the American Chemical Society* 117.19 (May 1995), pp. 5179–5197.

[75] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids". In: *Journal of the American Chemical Society* 118.45 (Jan. 1996), pp. 11225–11236.

[76] J A McCammon, B R Gelin, and M Karplus. "Dynamics of folded proteins." In: *Nature* 267.5612 (June 1977), pp. 585–590.

[77] Michael W Mahoney and William L Jorgensen. "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions". In: *J. Chem. Phys.* 112.20 (2000), p. 8910.

[78] William L Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *J. Chem. Phys.* 79.2 (1983), p. 926.

[79] M Levitt and R Sharon. "Accurate simulation of protein dynamics in solution." In: *Proceedings of the National Academy of Sciences of the United States of America* 85.20 (Oct. 1988), pp. 7557–7561.

[80] David E Shaw et al. "Atomic-level characterization of the structural dynamics of proteins." In: *Science (New York, N.Y.)* 330.6002 (Oct. 2010), pp. 341–346.

[81] Adrià Pérez, Gerard Martínez-Rosell, and Gianni De Fabritiis. "Simulations meet machine learning in structural biology." In: *Current opinion in structural biology* 49 (Feb. 2018), pp. 139–144.

[82] Elisabeth Gasteiger et al. "Protein identification and analysis tools on the expasy server". In: *The proteomics protocols handbook*. Ed. by John M Walker. Totowa, NJ: Humana Press, 2005, pp. 571–607.

[83] F Delaglio et al. "NMRPipe: a multidimensional spectral processing system based on UNIX pipes." In: *Journal of Biomolecular NMR* 6.3 (Nov. 1995), pp. 277–293.

[84] Wim F Vranken et al. "The CCPN data model for NMR spectroscopy: development of a software pipeline." In: *Proteins: Structure, Function, and Bioinformatics* 59.4 (June 2005), pp. 687–696.

[85] Sven G Hyberts, Haribabu Arthanari, and Gerhard Wagner. "Applications of non-uniform sampling and processing." In: *Topics in current chemistry* 316 (2012), pp. 125–148.

[86] Zsófia Sólyom et al. "BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins." In: *Journal of Biomolecular NMR* 55.4 (Apr. 2013), pp. 311–321.

[87] Eran Eyal et al. "Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins." In: *Journal of Computational Chemistry* 25.5 (Apr. 2004), pp. 712–724.

[88] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1-2 (Sept. 2015), pp. 19–25.

[89] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field." In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (June 2010), pp. 1950–1958.

[90] Yifan Song et al. "High-resolution comparative modeling with RosettaCM." In: *Structure (London, England : 1993)* 21.10 (Oct. 2013), pp. 1735–1742.

[91] Adrian A Canutescu and Roland L Dunbrack. "Cyclic coordinate descent: A robotics algorithm for protein loop closure." In: *Protein Science* 12.5 (May 2003), pp. 963–972.

[92] A Sali and T L Blundell. "Comparative protein modelling by satisfaction of spatial restraints." In: *Journal of Molecular Biology* 234.3 (Dec. 1993), pp. 779–815.

[93] Argyris Politis et al. "A mass spectrometry-based hybrid method for structural modeling of protein complexes." In: *Nature methods* 11.4 (Apr. 2014), pp. 403–406.

[94] Valérie Gabelica and Erik Marklund. "Fundamentals of ion mobility spectrometry." In: *Curr Opin Chem Biol* 42 (Feb. 2018), pp. 51–59.

[95] Shane A Seabrook and Janet Newman. "High-throughput thermal scanning for protein stability: making a good technique more robust." In: *ACS Combinatorial Science* 15.8 (Aug. 2013), pp. 387–392.

[96] Frank H Niesen, Helena Berglund, and Masoud Vedadi. "The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability." In: *Nature Protocols* 2.9 (2007), pp. 2212–2221.

[97] Mirella Vivoli et al. "Determination of protein-ligand interactions using differential scanning fluorimetry." In: *Journal of Visualized Experiments* 91 (Sept. 2014), p. 51809.

[98] Martin B Peters et al. "Structural survey of zinc containing proteins and the development of the zinc AMBER force field (ZAFF)." In: *Journal of Chemical Theory and Computation* 6.9 (Sept. 2010), pp. 2935–2947.

[99] Alan W Sousa da Silva and Wim F Vranken. "ACPYPE - AnteChamber PYthon Parser interfacE." In: *BMC Research Notes* 5 (July 2012), p. 367.

[100] Robert T McGibbon et al. "Mdtraj: A modern open library for the analysis of molecular dynamics trajectories." In: *Biophysical Journal* 109.8 (Oct. 2015), pp. 1528–1532.

[101] Naveen Michaud-Agrawal et al. "MDAnalysis: a toolkit for the analysis of molecular dynamics simulations." In: *Journal of Computational Chemistry* 32.10 (July 2011), pp. 2319–2327.

[102] Robert B Best, Gerhard Hummer, and William A Eaton. "Native contacts determine protein folding mechanisms in atomistic simulations." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.44 (Oct. 2013), pp. 17874–17879.

[103] Grant Thiltgen and Richard A Goldstein. "Assessing predictors of changes in protein stability upon mutation using self-consistency." In: *PloS one* 7.10 (Oct. 2012), e46084.

[104] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. "Role of conformational sampling in computing mutation-induced changes in protein structure and stability." In: *Proteins: Structure, Function, and Bioinformatics* 79.3 (Mar. 2011), pp. 830–838.

[105] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. "Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information." In: *elife* 3 (May 2014), e02030.

[106] Travis J Wheeler, Jody Clements, and Robert D Finn. "Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models." In: *BMC Bioinformatics* 15 (Jan. 2014), p. 7.

[107] Lukasz P Kozlowski and Janusz M Bujnicki. "MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins." In: *BMC Bioinformatics* 13 (May 2012), p. 111.

[108] M Madan Babu. "The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease." In: *Biochemical Society Transactions* 44.5 (Oct. 2016), pp. 1185–1200.

[109] Philippe Lieutaud et al. "How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe." In: *Intrinsically disordered proteins* 4.1 (Dec. 2016), e1259708.

[110] Tiago N Cordeiro et al. "Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression." In: *Structure (London, England : 1993)* (May 2019).

[111] Robert Konrat. "NMR contributions to structural dynamics studies of intrinsically disordered proteins." In: *Journal of Magnetic Resonance* 241 (Apr. 2014), pp. 74–85.

[112] Simone Kosol et al. "Structural characterization of intrinsically disordered proteins by NMR spectroscopy." In: *Molecules (Basel, Switzerland)* 18.9 (Sept. 2013), pp. 10802–10828.

[113] Veronique Receveur-Brechot and Dominique Durand. "How random are intrinsically disordered proteins? A small angle scattering perspective." In: *Current Protein & Peptide Science* 13.1 (Feb. 2012), pp. 55–75.

[114] Hagen Hofmann et al. "Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.40 (Oct. 2012), pp. 16155–16160.

[115] Alex S Holehouse et al. "CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins." In: *Biophysical Journal* 112.1 (Jan. 2017), pp. 16–21.

[116] Andrew Campen et al. "TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder." In: *Protein and Peptide Letters* 15.9 (2008), pp. 956–963.

[117] Erik W Martin et al. "Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation." In: *Journal of the American Chemical Society* 138.47 (Nov. 2016), pp. 15323–15335.

[118] Walter Basile et al. "Why do eukaryotic proteins contain more intrinsically disordered regions?" In: *PLOS Comput Biol* 15.7 (July 2019), e1007186.

[119] Kathleen A Burke et al. "Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II." In: *Molecular cell* 60.2 (Oct. 2015), pp. 231–241.

[120] Z Burkart-Solyom. *NMR methods for intrinsically disordered proteins: application to studies of NS5A protein of hepatitis C virus*. Tech. rep. 2014.

[121] B Qin, S S Lam, and K Lin. "Crystal structure of a transcriptionally active Smad4 fragment." In: *Structure (London, England : 1993)* 7.12 (Dec. 1999), pp. 1493–1503.

[122] Y Shi et al. "A structural basis for mutational inactivation of the tumour suppressor Smad4." In: *Nature* 388.6637 (July 1997), pp. 87–93.

[123] M P de Caestecker et al. "The Smad4 activation domain (SAD) is a proline-rich, p300-dependent transcriptional activation domain." In: *The Journal of biological chemistry* 275.3 (Jan. 2000), pp. 2115–2122.

[124] F Liu, C Pouponnot, and J Massagué. "Dual role of the Smad4/DPC4 tumor suppressor in TGFbeta-inducible transcriptional complexes." In: *Genes & development* 11.23 (Dec. 1997), pp. 3157–3167.

[125] Jia-Wei Wu et al. "Crystal structure of a phosphorylated smad2". In: *Molecular cell* 8.6 (Dec. 2001), pp. 1277–1289.

[126] Fátima Herranz-Trillo et al. "Structural Analysis of Multi-component Amyloid Systems by Chemometric SAXS Data Decomposition." In: *Structure (London, England : 1993)* 25.1 (Jan. 2017), pp. 5–15.

[127] Erik G Marklund et al. "Collision cross sections for structural proteomics." In: *Structure (London, England : 1993)* 23.4 (Apr. 2015), pp. 791–799.

[128] Mike P Williamson. "Using chemical shift perturbation to characterise ligand binding." In: *Progress in nuclear magnetic resonance spectroscopy* 73 (Aug. 2013), pp. 1–16.

[129] M Karplus and J Kuriyan. "Molecular dynamics and protein function." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.19 (May 2005), pp. 6679–6685.

[130] Vijay M Krishnamurthy et al. "Dependence of effective molarity on linker length for an intramolecular protein-ligand system." In: *Journal of the American Chemical Society* 129.5 (Feb. 2007), pp. 1312–1320.

[131] Huan-Xiang Zhou. "The Affinity-Enhancing Roles of Flexible Linkers in Two-Domain DNA-Binding Proteins†". In: *Biochemistry* 40.50 (Dec. 2001), pp. 15069–15073.

[132] Wade Borcherds et al. "Optimal affinity enhancement by a conserved flexible linker controls p53 mimicry in mdmx." In: *Biophysical Journal* 112.10 (May 2017), pp. 2038–2042.

[133] Maodong Li et al. "Disordered linkers in multidomain allosteric proteins: Entropic effect to favor the open state or enhanced local concentration to favor the closed state?" In: *Protein Science* 27.9 (Sept. 2018), pp. 1600–1610.

[134] Dale Stuchfield et al. "The use of mass spectrometry to examine idps: unique insights and caveats." In: *Methods in Enzymology* 611 (Nov. 2018), pp. 459–502.

[135] Antoni J Borysik et al. "Ensemble Methods Enable a New Definition for the Solution to Gas-Phase Transfer of Intrinsically Disordered Proteins." In: *Journal of the American Chemical Society* 137.43 (Nov. 2015), pp. 13807–13817.

[136] Łukasz Jaremko et al. "Fast evaluation of protein dynamics from deficient 15N relaxation data." In: *Journal of Biomolecular NMR* 70.4 (Mar. 2018), pp. 219–228.

[137] Soren Skou, Richard E Gillilan, and Nozomi Ando. "Synchrotron-based small-angle X-ray scattering of proteins in solution." In: *Nature Protocols* 9.7 (July 2014), pp. 1727–1739.

[138] Rik Derynck and Erine H Budi. "Specificity, versatility, and control of TGF-family signaling." In: *Science signaling* 12.570 (Feb. 2019).

[139] Ling Liu et al. "Smad2 and Smad3 have differential sensitivity in relaying TGF signaling and inversely regulate early lineage specification." In: *Scientific reports* 6 (Feb. 2016), p. 21602.

[140] J J Correia et al. "Sedimentation studies reveal a direct role of phosphorylation in Smad3:Smad4 homo- and hetero-trimerization." In: *Biochemistry* 40.5 (Feb. 2001), pp. 1473–1482.

[141] Caroline S Hill. "Transcriptional control by the smads." In: *Cold Spring Harbor Perspectives in Biology* 8.10 (Oct. 2016).

[142] Simon A Forbes et al. "COSMIC: exploring the world knowledge of somatic mutations in human cancer." In: *Nucleic acids research* 43.Database issue (Jan. 2015), pp. D805–11.

[143] Yigong Y Shi and Joan J Massagué. "Mechanisms of TGFB Signaling from Cell Membrane to the Nucleus". In: *Cell* 113.6 (June 2003), pp. 16–16.

[144] Tiago N Cordeiro et al. "Disentangling polydispersity in the PCNA-p15PAF complex, a disordered, transient and multivalent macromolecular assembly." In: *Nucleic acids research* 45.3 (Feb. 2017), pp. 1501–1515.

[145] Po-Chia Chen et al. "A General Small-Angle X-ray Scattering-Based Screening Protocol Validated for Protein-RNA Interactions." In: *ACS Combinatorial Science* 20.4 (Apr. 2018), pp. 197–202.

[146] Chris A Brosey and John A Tainer. "Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology." In: *Current opinion in structural biology* (June 2019).

[147] Eric J Steinmetz and Michele E Auldridge. "Screening Fusion Tags for Improved Recombinant Protein Expression in E. coli with the Expresso Solubility and Expression Screening System." In: *Current Protocols in Protein Science* 90 (Nov. 2017), pp. 5.27.1–5.27.20.

[148] Baolei Jia and Che Ok Jeon. "High-throughput recombinant protein expression in Escherichia coli: current status and future perspectives." In: *Open biology* 6.8 (2016).

[149] David S Waugh. "The remarkable solubility-enhancing power of Escherichia coli maltose-binding protein." In: *Postepy biochemii* 62.3 (2016), pp. 377–382.

[150] Sreejith Raran-Kurussi and David S Waugh. "A dual protease approach for expression and affinity purification of recombinant proteins." In: *Analytical Biochemistry* 504 (July 2016), pp. 30–37.

[151] Jo A Capp et al. "The statistical conformation of a highly flexible protein: small-angle X-ray scattering of S. aureus protein A." In: *Structure (London, England : 1993)* 22.8 (Aug. 2014), pp. 1184–1195.

[152] Huan-Xiang Zhou. "Quantitative account of the enhanced affinity of two linked scFvs specific for different epitopes on the same antigen." In: *Journal of Molecular Biology* 329.1 (May 2003), pp. 1–8.

[153] Charlotte S S and Magnus Kjaergaard. "Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics". In: *bioRxiv* (Mar. 2019).

[154] Buyong Ma et al. "Dynamic allostery: linkers are not merely flexible." In: *Structure (London, England : 1993)* 19.7 (July 2011), pp. 907–917.

[155] Philippe Lucarelli et al. "Resolving the combinatorial complexity of smad protein complex formation and its link to gene expression." In: *Cell systems* 6.1 (Jan. 2018), 75–89.e11.

[156] Toshiaki Mochizuki et al. "Roles for the MH2 domain of Smad7 in the specific inhibition of transforming growth factor-beta superfamily signaling." In: *The Journal of biological chemistry* 279.30 (July 2004), pp. 31568–31574.

[157] H Hayashi et al. "The MAD-related protein Smad7 associates with the TGFbeta receptor and functions as an antagonist of TGFbeta signaling." In: *Cell* 89.7 (June 1997), pp. 1165–1173.

[158] Jie Wang et al. "A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins." In: *Cell* 174.3 (July 2018), 688–699.e16.

[159] Denes Hnisz et al. "A phase separation model for transcriptional control." In: *Cell* 169.1 (Mar. 2017), pp. 13–23.

[160] Benjamin R Sabari et al. "Coactivator condensation at super-enhancers links phase separation and gene control." In: *Science (New York, N.Y.)* 361.6400 (July 2018).

[161] Mingjian Du and Zhijian J Chen. "DNA-induced liquid phase condensation of cGAS activates innate immune signaling." In: *Science (New York, N.Y.)* 361.6403 (Aug. 2018), pp. 704–709.

[162] Ethan Cerami et al. "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data." In: *Cancer discovery* 2.5 (May 2012), pp. 401–404.

[163] Henning Stehr et al. "The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors." In: *Molecular Cancer* 10 (May 2011), p. 54.

[164] Romain A Studer, Benoit H Dessailly, and Christine A Orengo. "Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes." In: *The Biochemical journal* 449.3 (Feb. 2013), pp. 581–594.

[165] J B Jones and S E Kern. "Functional mapping of the MH1 DNA-binding domain of DPC4/SMAD4." In: *Nucleic acids research* 28.12 (June 2000), pp. 2363–2368.

[166] Andrew Leaver-Fay et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." In: *Methods in Enzymology* 487 (2011), pp. 545–574.

[167] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." In: *J Mol Biol* 320.2 (July 2002), pp. 369–387.

[168] Hwee Ching Ang et al. "Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains." In: *The Journal of biological chemistry* 281.31 (Aug. 2006), pp. 21934–21941.

[169] Pascal Leuenberger et al. "Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability." In: *Science (New York, N.Y.)* 355.6327 (Feb. 2017).

[170] Tugba G Kucukkal et al. "Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins." In: *Current opinion in structural biology* 32 (June 2015), pp. 18–24.

[171] Anshuman Dixit et al. "Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: the impact on protein structure, dynamics, and stability." In: *Biophysical Journal* 96.3 (Feb. 2009), pp. 858–874.

[172] Ryan Day et al. "Increasing temperature accelerates protein unfolding without changing the pathway of unfolding." In: *Journal of Molecular Biology* 322.1 (Sept. 2002), pp. 189–203.

[173] Mariël G Pikkemaat et al. "Molecular dynamics simulations as a tool for improving protein stability". In: *Protein Engineering Design and Selection* 15.3 (Mar. 2002), pp. 185–192.

[174] Atanas Kamburov et al. "Comprehensive assessment of cancer missense mutation clustering in protein structures." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.40 (Oct. 2015), E5486–95.

[175] J P Gallivan and D A Dougherty. "Cation-pi interactions in structural biology." In: *Proceedings of the National Academy of Sciences of the United States of America* 96.17 (Aug. 1999), pp. 9459–9464.

[176] Chin Jung Cheng and Valerie Daggett. "Different misfolding mechanisms converge on common conformational changes: human prion protein pathogenic mutants Y218N and E196K." In: *Prion* 8.1 (Feb. 2014), pp. 125–135.

[177] Y Pan and V Daggett. "Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: free energy perturbation calculations using transition and denatured states from molecular dynamics simulations of unfolding." In: *Biochemistry* 40.9 (Mar. 2001), pp. 2723–2731.

[178] S Kumar, C J Tsai, and R Nussinov. "Factors enhancing protein thermostability". In: *Protein Engineering Design and Selection* 13.3 (Mar. 2000), pp. 179–191.

[179] Jay F Storz. "Compensatory mutations and epistasis for protein function." In: *Current opinion in structural biology* 50 (June 2018), pp. 18–25.

[180] Hua Tan, Jiguang Bao, and Xiaobo Zhou. "Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity." In: *Scientific reports* 5 (July 2015), p. 12566.

[181] Qianting Zhang et al. "ALK phosphorylates SMAD4 on tyrosine to disable TGF- tumour suppressor functions." In: *Nature Cell Biology* 21.2 (Jan. 2019), pp. 179–189.

[182] D Allan Drummond and Claus O Wilke. "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution." In: *Cell* 134.2 (July 2008), pp. 341–352.

[183] Gian Gaetano Tartaglia et al. "Life on the edge: a link between gene expression levels and aggregation rates of human proteins." In: *Trends in biochemical sciences* 32.5 (May 2007), pp. 204–206.

[184] A Hata et al. "Mutations increasing autoinhibition inactivate tumour suppressors Smad2 and Smad4." In: *Nature* 388.6637 (July 1997), pp. 82–87.

[185] Alexey Drozdetskiy et al. "JPred4: a protein secondary structure prediction server." In: *Nucleic acids research* 43.W1 (July 2015), W389–94.

[186] Andrew M Waterhouse et al. "Jalview Version 2–a multiple sequence alignment editor and analysis workbench." In: *Bioinformatics* 25.9 (May 2009), pp. 1189–1191.