UNIVERSITAT DE BARCELONA

A novel regulatory unit in the N-terminal region of c-Src

Miguel Arbesú Andrés

A novel regulatory unit in the N-terminal region of c-Src

Miguel Arbesú Andrés

A thesis presented for the degree of
PhD in Organic Chemistry

Supervised by:
Professor Miquel Pons

Organic Chemistry section
Inorganic and Organic Chemistry department
Faculty of Chemistry

UNIVERSITAT DE
BARCELONA

January 2018

I, Miguel Arbesú Andrés, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

In re ~~mathematica~~ scientifica ars proponendi
quaestionem pluris facienda est quam solvendi.

Adapted from Georg Cantor's doctoral thesis, 1867.

Summary

c-Src is a central player in several cellular signaling pathways. It controls important cellular processes like cellular proliferation, survival or motility. Therefore, a number of tumoral diseases have been related to abnormal c-Src activity. Among them, colorectal cancer stands out, as c-Src deregulation correlates tumor with progression and clinical outcome.

This tyrosine kinase is part of a larger group of functionally and structurally related proteins termed Src Family Kinases. These proteins share the same domain architecture: a cassette formed by a catalytic domain (SH1), two regulatory domains, SH2 and SH3, and a variable intrinsically disordered region (the Unique domain) that ultimately anchors to the inner face of the cellular membrane via the N-terminal SH4 domain, also disordered. The sequence and structure of the cassette are highly conserved, and thus unsurprisingly Src Family Kinases perform closely related and often overlapping functions. However, the role of intrinsically disordered regions has remained unclear, although they are known to be functionally relevant.

In this work, the structural and functional relationship between the intrinsically disordered SH4 and Unique domains with the neighboring folded SH3 domain in c-Src is explored. Interactions between disordered and ordered proteins are often characterized by the formation of complexes that are specific and functional but structurally heterogeneous. Moreover, conformational plasticity is a fundamental feature for function. These assemblies are known as fuzzy complexes. Here this theoretical framework, usually applied to isolated partners, is extended to the intramolecular interface between covalently bound domains instead of isolated pairs. The concept of fuzzy binding is also used in order to describe interactions based on sets of dynamic, transient, and promiscuous contacts between ill-defined sets of interactors.

In order to characterize the system, an integrative strategy using short and long range Nuclear Magnetic Resonance techniques and Small Angle X-ray Scattering is applied to several constructs containing different combinations of bound or isolated domains. It is demonstrated that the folded SH3 domain acts as a scaffold for the disordered region, which interacts in a specific manner with its partner. Both disordered domains, SH4 and Unique, are involved in the process albeit they contribute differently. Additionally, it is shown that the Unique domain is not a random coil, but contains a significant degree of pre-arrangement that is independent of the scaffold.

Sequence determinants are then searched by comparison of the sequences of different Src Family Kinases. Four conserved phenylalanine residues are found and their implication in Unique domain pre-organization and Unique:SH3 domain interaction tested. All these amino acids are found to favor compaction of the intrinsically disordered region, and at the same time to perturb close contact with the scaffold. In addition, mutations in the interacting zones of the SH3 domain are also studied to test reciprocity. In all, the fuzzy complex model is proven for the SH4:Unique:SH3 system.

Then, the results are extrapolated to the full-length c-Src to test its biological relevance. A co evolutionary analysis suggests that the fuzzy model may be a general feature for the whole Src Family, so the closest member of the family, Yes, is also tested experimentally. The initial results on long-range contacts suggests a similar arrangement between the scaffold and the disordered region.

In all, it is suggested that plastic, fuzzy interfaces between ordered and disordered domains may be a relevant mode for the transmission of functional information within multidomain proteins.

Finally, a first approach for a structural study of the c-Src fuzzy complex in a native-like lipid environment, including natural co-translational modifications, is presented. A protocol for sample preparation is developed and Dynamic Nuclear Polarization solid state NMR is shown to be an adequate tool for further analysis.

Preface

This thesis was developed at the Biomolecular NMR group of the University of Barcelona under the supervision of Professor Miquel Pons, within the project *Src Unique domain signaling in colorectal cancer* funded by the Fundació Marató TV3. The PhD internship at the Leibniz-Forschungsinstitut für Molekulare Pharmakologie in Berlin was funded by Instruct, a Landmark ESFRI project.

The document is organized as follows:

- An introduction to the topics of c-Src, intrinsically disordered proteins, and fuzzy complexes, followed by a bullet point list of objectives.
- Five sections of results sharing the same general structure: exposition of the results, and brief discussion. The first one, 2.1, contains a small introduction on previous results. The last section, 2.4, summarizes the results of my internship and can be regarded as an independent unit, with its own introduction to the matter and to the experimental approach used.
- An overall discussion in terms of the concepts developed along the introduction is presented, with a final list of conclusions.
- A Methods & Materials section with details on the experimental aspects and data treatment.
- An Appendix containing the original data from which some excerpts shown in the results section were derived.

The results derived from this work have been published under references Maffei et al. (2013), Maffei (2015) and Arbesú et al. (2017):

> Maffei, M. et al., 2013. Lipid Binding by Disordered Proteins. Protocol Exchange.

Maffei, M. et al., 2015. The SH3 Domain Acts as a Scaffold for the N-Terminal Intrinsically Disordered Regions of c-Src. Structure, 23(5), pp.893–902.

Arbesú, M. et al., 2017. The Unique Domain Forms a Fuzzy Intramolecular Complex in Src Family Kinases. Structure, 25(4), pp.630–640.e4.

Finally, I acknowledge that this document has been elaborated using the Markdown thesis template developed by Pollard et al. (2016).

Academic acknowledgements

I would like to start with the academic acknowledgments in reverse historical order. Thus, my thesis advisor Prof. Miquel Pons is the first on the list.

Miquel, I want to thank the trust you deposited in me for carrying on this project. I have sincerely done my best to meet your expectations. On that task I acknowledge your help, for your door and your brain have always been open. Your advise has definitely shaped me as a scientist. Not only that, but your human quality is also a standard I will try to keep up to. For all that, I will always be grateful.

In the second place, I want to thank all current and past Pons lab members whom have had the pleasure to work with and befriend. One follows his own path through his thesis, but it is the company of the others what makes it enjoyable and rewarding. Special words go to Dr. Mariano Maffei and Dr. Irene Amata, who handed me the c-Src project and got me started into molecular biology and NMR, respectively; Dr. João M.C. Teixeira for sharing his knowledge and thoughts; Dr. Tiago N. Cordeiro, and Dr. Ildefonso Marín who have been always available for advise and helpful discussion; and Laareb Irrem Mohammad Jabeen, the best heiress I could imagine for the myristoylated future of c-Src in the lab (best of luck with it!).

I would also like to mention here Dr. Margarida Gairí, from the NMR facility at the University of Barcelona for her expertise and kindness. Thanks too to Prof. Hartmut Oschkinat from the FMP in Berlin and all the team there (specially Michel-Andreas Geiger and Dr. Wing Ying Chow) for the worthwhile stay I enjoyed between them and the time and attention they dedicated to me.

Finally, I would like to thank all the persons involved in my past academic training, from high and primary school time at home to the B.Sc. at the University of Oviedo: Flor, Rafaela, Paco, Eusebio, Fernando, Nacho, Dolores, José Manuel, and so many others. It is back there where all started and it is because of them, all servants of the public education system as Miquel, that I came this far. It is my duty as their pupil to proudly acknowledge and defend their work.

Personal acknowledgements

In the personal aspect, I can only start by thanking my parents Javier and María Cruz, my sister Estela, and grandfather Lucio: this would not be possible without your love and unwavering support. You made me the person I am today. I sincerely hope this thesis will make you proud.

I would also like to remember here those who would have enjoyed to see this work done but, unfortunately, did not have the time: grandparents Manuel and Ángeles, father-in-law Jorge, and friend and fellow chemist Ramón. You are always in my thoughts.

I also want thank my dearest friends without simply falling in an enumeration that would result ironically unpersonal. To those from home, for making me feel as I never left whenever I am back, and to those I have met during these years in Barcelona and made it a new home for me: you know who you are and how much you are loved. Thanks for always being there.

And last, but not least, to my wife Sara. You have lived every day of this thesis from the inside, being my peace and my strength. Thanks for your love, patience, encouragement, and inspiration. I can not express how fortunate I am to go hand in hand with you.

I love you.

Contents

List of Figures

List of Tables

Abbreviations

ATP	**A**denosine **t**ri**p**hosphate
BSA	**B**ovine **S**erum **A**lbumine
CaM	**C**al**m**odulin
CE	**C**ross **E**ffect
CP	**C**ross **P**olarization
CS	**C**hemical **S**hift
CSP	**C**hemical **S**hift **P**erturbation
DARR	**D**ipolar **A**ssisted **R**otational **R**esonance
DLS	**D**ynamic **L**ight **S**cattering
DMPC	1,2-**D**imyristoyl-*sn*-**g**lycero-3-**p**hosphoryl**c**holine
DMPG	1,2-**D**imyristoyl-*sn*-**g**lycero-3-**p**hosphoryl**g**lycerol
DNP	**D**ynamic **N**uclear **P**olarization
DSA	**D**oxyl **S**tearic **A**cid
DSS	4,4-**d**imethyl-4-**s**ilapentane-1-**s**ulphonic acid
EOM	**E**nsemble **O**ptimization **M**ethod
GST	**G**lutathione **S**-**T**ransferase
HEPES	4-(2-**h**ydroxy**e**thyl)-1-**p**iperazine**e**thane**s**ulphonic acid
HMM	**H**idden **M**arkov **M**odel
HMQC	**H**eteronuclear **M**ultiple **Q**uantum **C**oherence
HSQC	**H**eteronuclear **S**ingle **Q**uantum **C**oherence
IDP	**I**ntrinsically **D**isordered **P**rotein
ITC	**I**so**t**hermal **C**alorimetry
INEPT	**I**nsensitive **N**uclei **E**nhanced by **P**olarization **T**ransfer
IDR	**I**ntrinsically **D**isordered **R**egion
LC-MS	Coupled **L**iquid **C**romatography and **M**ass **S**pectrometry
LUV	**L**arge **U**nilamellar **V**esicle
MTSL	(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl **me**thane**sul**fonothioate
MD	**M**olecular **D**ynamic

MW	**M**icro **W**ave
NMR	**N**uclear **M**agnetic **R**esonance
NMT	**N**-**M**yristoyl **T**ransferase
NOE	**N**uclear **O**verhauser **E**ffect
NUS	**N**on **U**niform **S**ampling
MAS	**s**olid **s**tate **N**uclear **M**agnetic **R**esonance
nRTK	non-**R**eceptor **T**yrosine **K**inase
PIC	**P**rotease **I**nhibitor **C**ocktail
PKA	**P**rotein **K**inase **A**
PMSF	**P**henyl**m**ethyl**s**ulphonyl**f**luoride
PPII	**P**oly**P**roline type **II** helix
PRE	**P**aramagnetic **R**elaxation **E**nhancement
RDC	**R**esidual **D**ipolar **C**oupling
pRDC	paramagnetic **R**esidual **D**ipolar **C**oupling
RSV	**R**ous **S**arcoma **V**irus
RTK	**R**eceptor **T**yrosine **K**inase
SAXS	**S**mall **A**ngle **X**-ray **S**cattering
SD	**S**tandard **D**eviation
SDS-PAGE	**S**odium **d**odecyl **s**ulphate **p**oly**a**crylamide **G**el **E**lectrophoresis
SFK	**S**rc **F**amily **K**inases
SH	**S**rc **H**omology domains
SH4-UD	Construct formed by **SH4** and **U**nique **D**omains
SPR	**S**urface **P**lasmon **R**esonance
TEV	**T**obacco **E**tch **V**irus
TRIS-HCl	**tris**(hydroxymethyl)aminomethane **h**ydro**c**hloride
ULBR	**U**nique **L**ipid **B**inding **R**egion
USH3	Construct formed by **S**H4, **U**nique and **SH3** domains

Chapter 1

Introduction

In this section I first introduce the subject of this thesis, the human tyrosine kinase **c-Src** and the related **Src Family Kinases (SFK)**. I also briefly review the role of c-Src in cancer and its relevance as a drug target. Then, I present the topic of **Intrinsically Disordered Proteins (IDPs)** with a special emphasis on IDP protein binding and protein **fuzzy complexes**, relevant for the study of the N-terminal region of c-Src which includes ordered and disordered regions. Finally, I discuss the **biophysical characterization of IDPs**, focusing on the Nuclear Magnetic Resonance (NMR) techniques used in this thesis.

1.1 A historical overview of Src

Nowadays Src is a scientific *hot topic*, mostly because of its implication in tumoral diseases, with ~4 peer-reviewed research papers steadily published everyday during the last 15 years[1]. However, Src history goes a century back in time, well before the advent of modern structural biology (figure 1.1).

As extensively reviewed in Martin (2001) and Martin (2004), the first findings date from the 1910s, when Peyton Rous observed an avian spindle-cell sarcoma that could be transplanted and induce the disease to healthy chickens (Rous 1910). The finding of cell filtrates able to provoke tumors (Rous 1911) led to the identification of the Rous Sarcoma

[1]Source: PubMed historical track on "Src" topic.

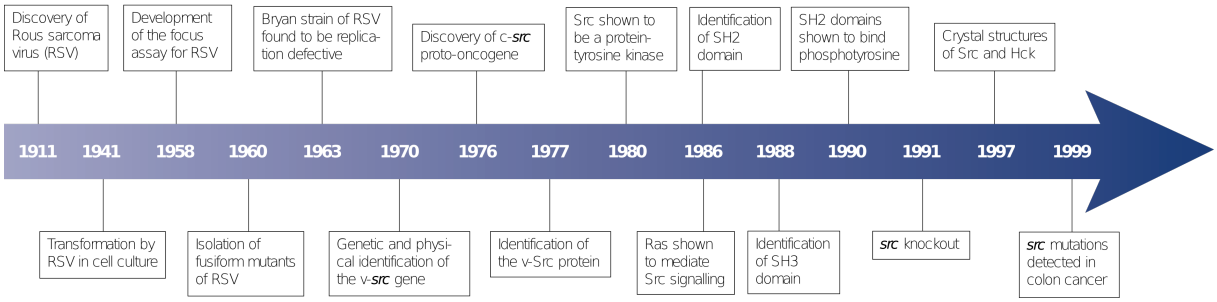


Figure 1.1: Key events in Src historical timeline until 2000. Reproduced with permission from Martin (2001).

Virus (RSV), which granted him the Nobel prize in Medicine in 1966 for the discovery of tumor-inducing viruses.

During the 1950 - 1960s, extensive research was done on a large number of viral strains. Special interest was raised by the capacity of RSV and other related RNA retroviruses to induce cell transformation in several types of mammalian cells, a fact that defied the established knowledge about oncogenic viruses at the time. These efforts led to the identification of the key transforming factor in the 1970s: the viral Src (*v-Src*) gene and its protein product (Wang et al. 1976; Brugge et al. 1978). Also in the late 1970s a cellular analogue was discovered: **c-Src** (Stehelin, Guntaka, et al. 1976; Stehelin, Varmus, et al. 1976). Most importantly, evidences showed that the Src gene is not viral but cellular in origin, and is widely spread among vertebrates. These findings led to the second Nobel prize derived from Src research, awarded to J. Michael Bishop and Harold E. Varmus in 1989. Decades of research in RSV and other oncogenic viruses thus permeated to the field of cancer biology. The term oncogene was first coined by Huebner & Todaro (1969) referring to the endogenous *viral information* carried by cells that, upon activation, could cause cancer. The name **proto-oncogene** was then defined to emphasize that the cellular precursors of the oncogenes lack transforming capacity unless mutated or over-expressed.

The ground-breaking discovery of this first proto-oncogene and its role in cancer sparked a whole research area (Bister 2015). By the early 1980s, Src was characterized as a tyrosine kinase (Hunter & Sefton 1980), revealing a new enzyme class[2] later found to be key in cell signaling. It was soon observed that the viral forms of Src had a much higher activity than c-Src, consistently with a lower transforming capacity of the latter and a number of variations in sequence. Therefore, during the following three decades, much endeavor was put on c-Src domain architecture, modulation, and function. I discuss those aspects in more detail later in this section.

[2]At the time, only serine and threonine kinases were known. In fact, Src was initially thought to phosphorylate threonine residues.

Significant contributions to the field of structural biology stem from Src research. Novel domains were recognized in Src and other related proteins, and were thus termed **Src Homology domains (SH#)**. Besides the catalytic **SH1** domain, two protein-binding domains related to Src activity were described, namely **SH2** and **SH3** (Sadowski et al. 1986; Mayer et al. 1988). These common functional elements are widespread among signaling proteins, which led to the concept of modular protein interaction domains that is now in the basis of cell signaling (Sudol 1998; Pawson 2004).

The N-terminal region remained obscure until recent years due to its **intrinsically disordered** nature. The concept of intrinsic disorder (Dunker et al. 2001) will be further detailed in following sections, but in simple terms it refers to proteins (or regions) that do not adopt a single, stable conformation, but remain highly dynamic. That conformational heterogeneity of intrinsically disordered regions (**IDRs**) precludes crystallization, and therefore they are excluded from X-ray studies. It was found that the very N-terminal ~20 residues were important for c-Src to bind lipids via clusters of basic residues and co-translational myristoylation, so determining cellular location and activity (Patwardhan & Resh 2010). Thus the region was termed **SH4 domain** (Silverman 1992; Buser et al. 1994). More details on c-Src structure, regulation, lipid binding and distribution are given in the following sections.

The rest of the intrinsically disordered region was observed to be poorly conserved among Src-related proteins (see section 1.4 below) both regarding sequence and length, and was therefore named **Unique domain**. The birth and explosive growth of the study of intrinsically disordered proteins from the beginning of the 2000s turned the focus towards these N-terminal regions. Although the Unique domain was initially considered a simple tether between the folded core and the lipid anchoring SH4 domain, functional elements have been described during the last years by our group and others. Gingrich et al. (2004) found that the Unique domain of c-Src was capable to establish protein-protein interactions by binding the NMDA receptor. Later, our group confirmed this capacity by describing the interaction between the Unique domain and calcium-loaded calmodulin (Pérez et al. 2013). In the same work, it was also found that the Unique domain could also bind to membranes through a specific region, which was termed **Unique Lipid Binding Region (ULBR)**. Besides that, both the SH4 and Unique domains contain multiple phosphorylation sites (Amata et al. 2013; Amata et al. 2014). While some have been functionally characterized, such as S17, which inhibits SH4 lipid binding upon phosphorylation, most of them remain undefined.

Finally, after more than 100 years of accumulated knowledge since the discovery of the RSV, proto-oncogenes (and c-Src among them) are nowadays key targets for the treatment

of several human tumors, as reviewed in the following sub-section.

1.2 c-Src function, cell signaling, and cancer

Protein phosphorylation is the most widespread post-translational modification. The transfer of a phosphate group from an ATP molecule to the hydroxyl group of serine, threonine or tyrosine side chains is catalyzed by kinase enzymes, while the reverse reaction is carried out by phosphatases.

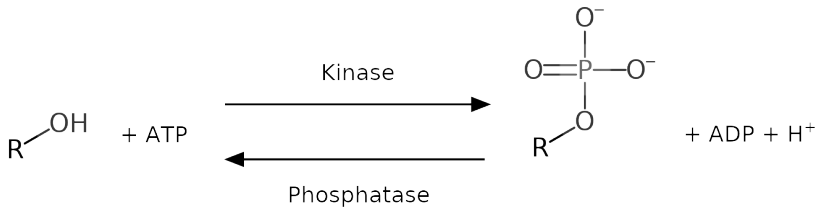


Figure 1.2: Phosphorylation/dephosphorylation reactions.

This modest modification can however prompt important structural changes in the substrate, via interactions of the new phosphate dianion with positively charged groups or H-bond formation. The typical effect on substrate activity is *on/off* switching upon phosphorylation or dephosphorylation. Reaction specificity and reversibility allows for selective, time dependent response to external stimuli. Further control on kinase and phosphatase expression and activity provides means for tight control and emergence of complex patterns and networks.

Thus unsurprisingly, one of the major biological functions of phosphorylation is **cell signaling and cycle control**. Signaling pathways have evolved along with organism complexity and are central regulatory elements for correct and coordinated cell behavior, specially in superior organisms. Aberrant regulation is most critical in neoplasia, defined by deregulated cell proliferation and, if malignant, capacity to invade and survive (Blume-Jensen & Hunter 2001). In those cases, perturbations in the delicate balance between cell division, growth and apoptosis can lead to cancer.

Introduced by Manning (2002), the term *kinome* corresponds to the protein kinase complement of the human genome. It comprises more than 500 kinases accounting for ~2 % of the genome. The most common phosphorylations occur in serines by far (86.4 %), followed by threonines (11.8 %) and, last but not least, tyrosines (1.8 %) (Johnson 2009). It is remarkable that the tyrosine kinase family underwent an explosive growth with the

appearance of Metazoa, and are thus associated with complex organisms (Manning et al. 2002).

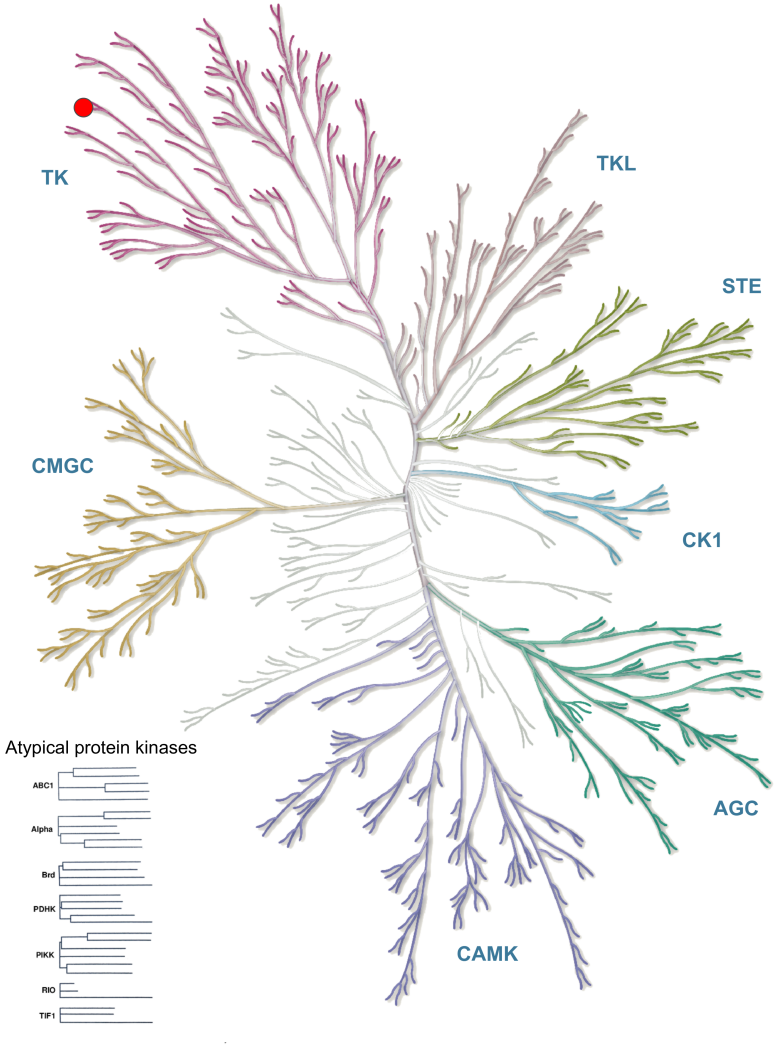


Figure 1.3: Human kinome map courtesy of Cell Signaling Technology, Inc. c-Src is indicated as a red circle. **AGC** Containing PKA, PKG, PKC families; **CAMK** Calcium/calmodulin-dependent protein kinase; **CK1** Casein kinase 1; **CMGC** Containing CDK, MAPK, GSK3, CLK families; **STE** Homologs of yeast Sterile 7, Sterile 11, Sterile 20 kinases; TK Tyrosine kinase; TKL Tyrosine kinase–like.

Albeit the >90 tyrosine kinases encoded in the human genome are a minority of the total kinases and generate a small fraction of the total phosphorylated substrates in cells, they are specially involved in signal transduction (Paul 2004). More specifically, c-Src and the closely related **Src Family Kinases** (later reviewed in this section) occupy a critical place in the signaling network and control important cellular functions such as proliferation, differentiation, survival or motility (Thomas & Brugge 1997). Thus, abnormal response to stimuli induces downstream effects in pathways that prevent uncontrolled growth and receive apoptotic signals.

Being the first proto-oncogene discovered and an ubiquitously expressed signaling hub, human c-Src correlation with cancer was soon sought. As reviewed by Irby & Yeatman (2000), since the early 1980s and through the 1990s numerous papers highlighted c-Src enhanced activity in a range of tumoral pathologies in breast, colon, lung, prostate, bladder, brain, and pancreas among others. Surprisingly, it turned out that activating mutations are seldom found, the first one reported only by 1999 (Yeatman et al. 1999). Instead, high c-Src protein levels and, most importantly, up-regulated activity has been consistently linked to tumor initiation and progression, but specially the latter. Src-dependent mechanisms have also been reported for latent bone metastasis in breast cancer (X. H.-F. Zhang et al. 2009).

Furthermore, not only c-Src but many of its substrates are also linked to tumoral processes. In most cases, it is in collaboration with other cellular effectors that c-Src oncogenic potential develops. The most important cooperating partners associated to Src are EGF receptors, FAK tyrosine kinase and sex steroid hormone receptors (Ishizawar & Parsons 2004).

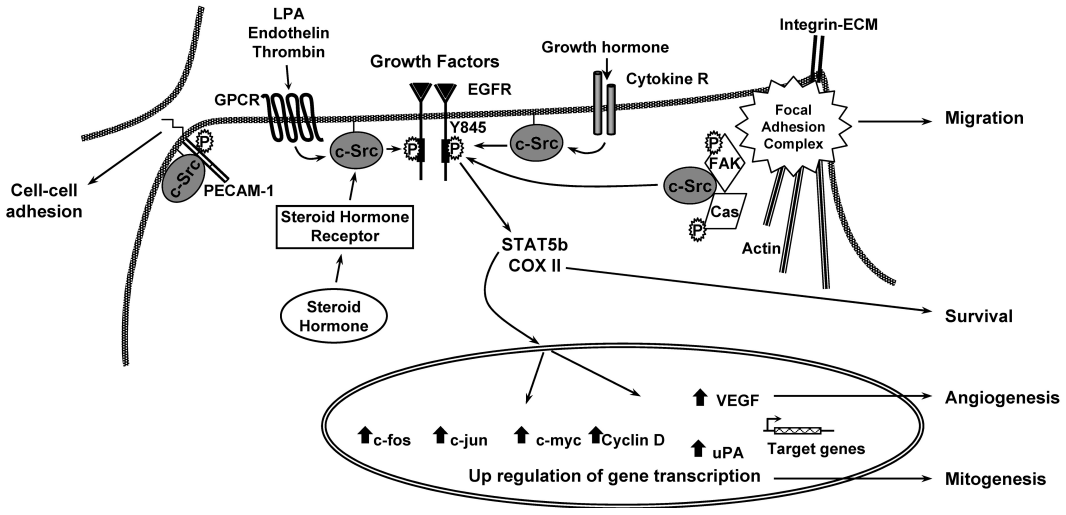


Figure 1.4: Oncogenic processes regulated by c-Src. Reproduced with permission from Ishizawar & Parsons (2004).

One of the most studied disease models is **colorectal cancer**. Upregulated c-Src activity has been reported to happen in the very first stages of the disease, to increase with progression and is associated with poor prognosis (Cartwright & Eckhart 1990; Anon 1993). Because of its importance and occurrence, this thesis has been developed within the project *"c-Src Unique domain signaling in colorectal cancer"*, funded by the Fundació Marató TV3.

Given their close connection with cancer, kinases have become important therapeutic targets during the last 20 years (J. Zhang et al. 2009), and constitute a good example

of modern targeted molecular pharmacology. One of the first successful tyrosine kinase inhibitors was Imatinib, developed to target the Bcr-Abl oncogenic kinase in chronic myelogenous leukemia, which supposed a paramount hallmark (Capdeville et al. 2002). As of 2016, the FDA has approved the use of 28 kinase inhibitors, most of them targeting tyrosine kinases associated to tumoral diseases (Wu et al. 2016). The vast majority of them are ATP competitors that bind the nucleotide binding site of the catalytic domain.

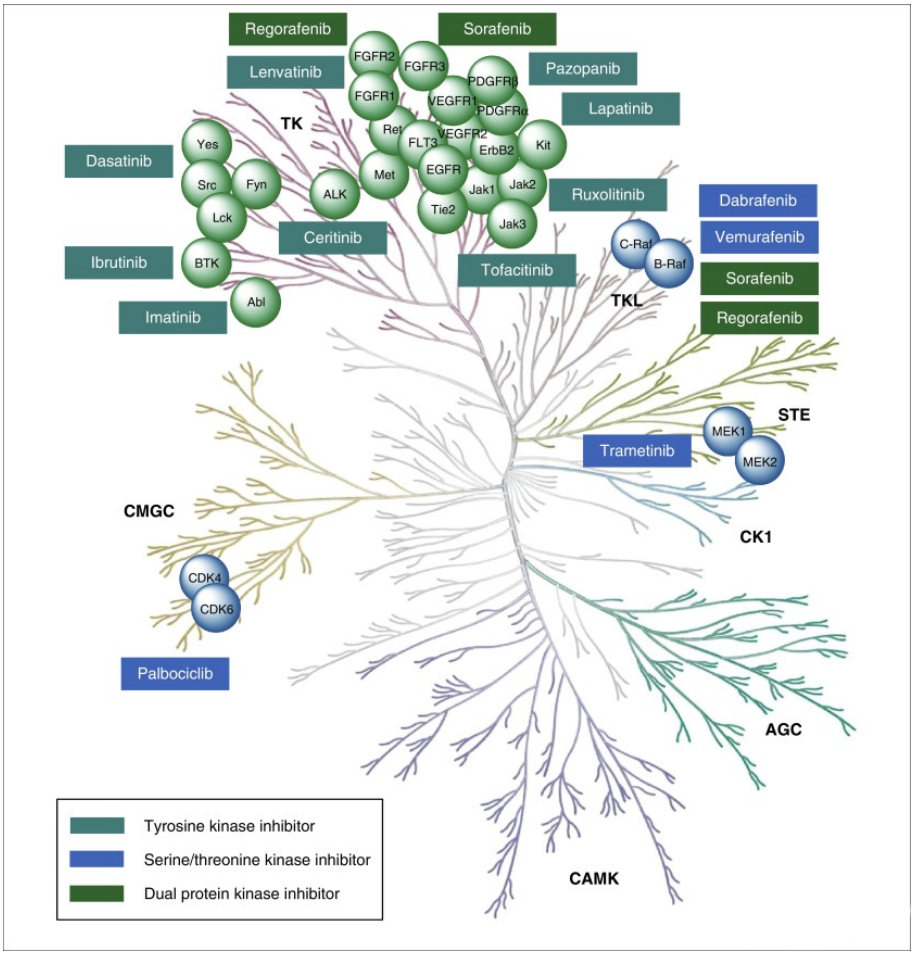


Figure 1.5: Kinase inhibitors approved as of 2016 plotted over the human kinome map shown in figure 1.4 with the names of targeted enzymes. Reproduced with permission from Wu et al. (2016).

ATP mimicking drugs have however some important drawbacks, mostly derived from their selectivity for a very common motif in Nature. The first is one is limited selectivity in some cases, which may rise unexpected side effects by interfering with the activity of other kinases related to the primary target. Signal pathway rewiring only but hinders the prediction of unexpected toxicities derived from this phenomenon. In addition, besides the >500 human tyrosine kinases, some other ~2 000 proteins contain nucleotide binding sites, which adds more potential off-targeting problems. Finally, there is the problem of acquired drug resistance, mostly by mutations in the catalytic domain. This complication typically arises due to the strong selective pressure induced by the cytotoxic effects of the

treatment.

Due to these issues, alternative strategies are being developed, such as exploiting the synergistic effect of drug combinations (several inhibitors, or inhibitors and other drugs such as rapamycin) or allosteric targeting. The latter is an interesting option with some examples of success, such as GNF-2, a drug that binds the myristate binding site in Bcr-Abl (Fallacara et al. 2014). Thanks to our increasing understanding of the structural characteristics of kinases, their mechanisms and interactions, it is expected that this field will provide new efficient drugs in the future (J. Zhang et al. 2009; Cowan-Jacob et al. 2014).

In the particular case of Src, success has been limited due to these difficulties despite the strong body of knowledge accumulated over the years. According to the Open Target database (Koscielny et al. 2017), three Src inhibitors have reached phase IV clinical research to date: Dasatinib, Bosutinib, and Vandetanib. Another four compounds, Saracatinib, Ilorasertib, KX2-391, and XL-228, are already in phase III. All of them are ATP competitors except KX2-391, which binds to the peptide binding region of the catalytic domain.

The close structural resemblance between the c-Src and other tyrosine kinases such as the Src Family Kinases (**SKFs**, see section 1.4) makes drug specificity a specially relevant matter if one desires to design precise therapeutic inhibition strategies. Thus, given the importance of the target, Src drug development is an open and active field with an special emphasis on mechanisms other than ATP agonism, such as targeting allosteric networks, to improve efficacy and reduce off-target toxicity (Dar & Shokat 2011).

1.3 c-Src structure, cellular location and regulation

As introduced in the previous section, the discovery of the Src Homology domains led to the concept of modular protein interaction domains: functional building blocks that are pervasively found in Nature in different arrangements. Although there may be a certain degree of sequence variability which determines particular details, e.g. specificity or affinity, the fold and basic function are highly conserved.

There are 1 306 tyrosine kinases deposited in the Pfam database (Finn et al. 2016) sharing the same domain arrangement as c-Src: A variable disordered N-terminal region, followed by a cassette formed by SH3 and SH2 protein-binding domains and a catalytic SH1 domain connected by relatively short flexible coils, and finally a short regulatory C-terminal tail.

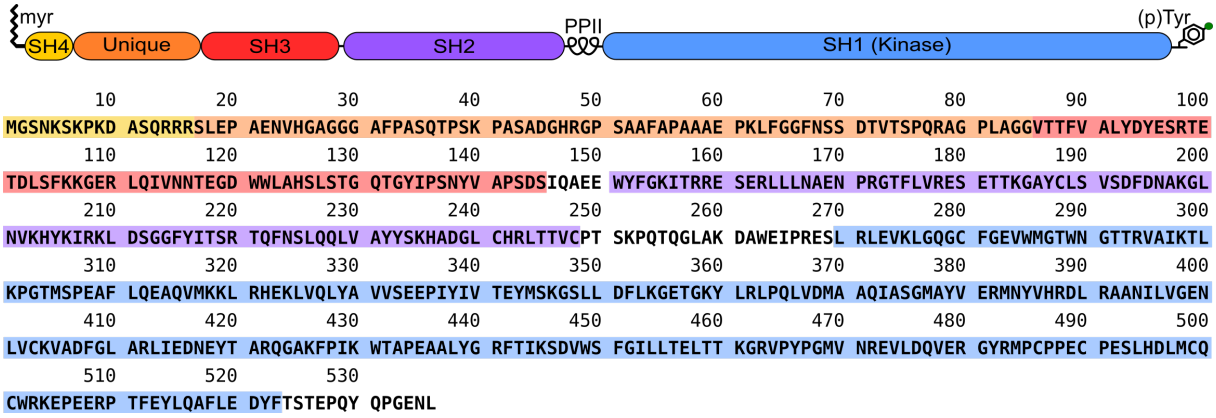


Figure 1.6: c-Src domain architecture and sequence.

The N-terminal intrinsically disordered region makes a difference between two main groups of tyrosine kinases: receptor (**RTKs**) and non-receptor (**nRTKs**). Both types associate to the cellular membrane for signal transduction. However, while RTKs have transmembrane and extracellular receptor domains coupled to the intracellular kinase domain, nRTKs are exclusively intracellular. The specific membrane anchoring mechanism is a key determinant for nRTK activation, transport, cellular location and, ultimately, activity (Resh 2006; Patwardhan & Resh 2010). In the case of nRTKs sharing c-Src domain architecture, the N-terminal intrinsically disordered region is responsible for binding the inner face of the membrane.

1.3.1 The SH4 domain

As introduced in the former section, the membrane anchoring mechanism of c-Src involves the initial 20 amino acids, which form a disordered but conserved region termed **SH4 domain**. SH4 lipid binding is controlled by a cooperative two-signal mechanism that involves protein lipidation and electrostatic interactions (Resh 1999).

Myristoylation is a co-translational modification that takes place directly in the protein nascent chain at the ribosome by the N-myristoyl Transferase enzyme (**NMT**). NMT recognizes an *MGxxxZ* consensus sequence, where x is any amino acid and Z is S or T, and cleaves the initial methionine to further create an amide bond between a myristate group and the now N-terminal glycine.

c-Src remains in the perinuclear region until it is activated upon phosphorylation of its C-terminal regulatory tail (see below). Then, it is transported to the cellular membrane. About 10 of the 14 myristate moiety carbons penetrate in the membrane driven by hydrophobicity. However, only-myristate lipid binding is weak, with a K_d around 10^{-4},

insufficient to provide a stable anchor.

The second c-Src membrane-binding signal is a cluster of basic residues of the SH4 domain, organized in two sets: ^{5}KSKPK^{8} and ^{14}RRR^{16}. The positively charged residues electrostatically interact with the inner membrane leaflet, enriched in acidic phospholipids (Buser et al. 1994; Sigal et al. 1994). The additional binding sums cooperatively to that of myristate, leading to a 3 000-fold increase in affinity. Remarkably, the SH4 domain is depleted of hydrophobic residues so, unlike for the myristoyl moiety, there is no insertion in the membrane whatsoever. The anchor is therefore formed by a *stitch and a plaster* that efficiently fasten the rest of the protein to the surface.

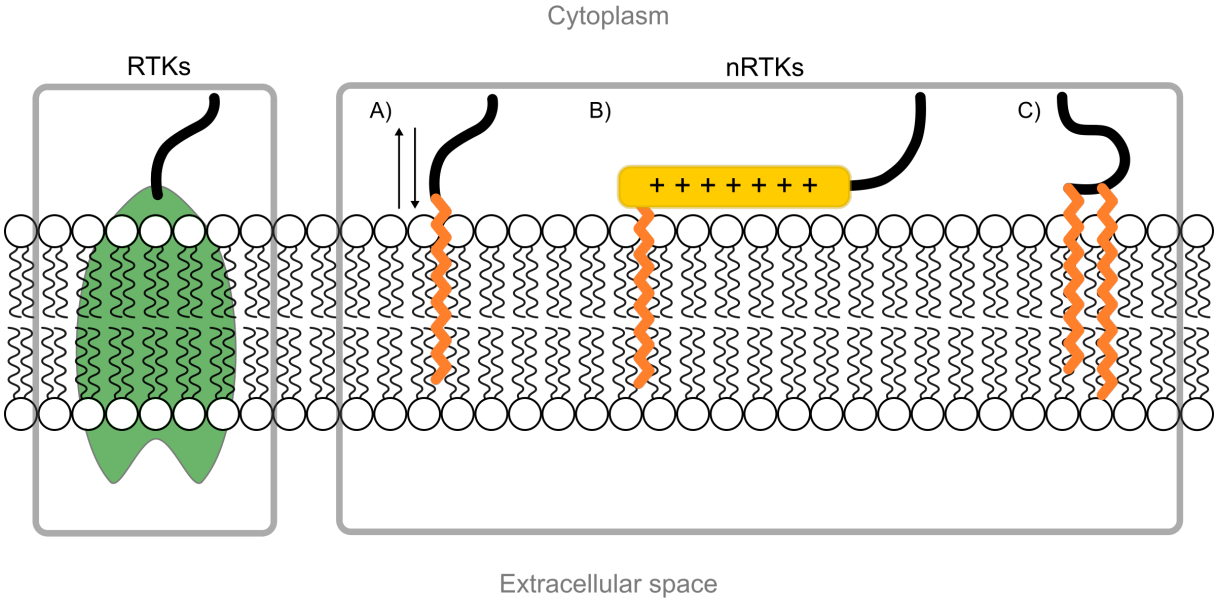


Figure 1.7: Cellular membrane binding mechanisms of receptor (left) non-receptor (right) tyrosine kinases. A) Only myristoylation (labile); B) myristoylation + electrostatics; C) myristoylation + palmitoylation.

1.3.2 The Unique domain

The next 65 residues following the SH4 domain form a lowly conserved region both in terms of length and sequence present in all c-Src related SFK members. In consequence, it has been termed the Unique domain. All SFK Unique domains are predicted to be intrinsically disordered regions (Santos & Siltberg-Liberles 2016). In the case of c-Src, our group was the first to provide experimental evidence of this property by NMR and **Small Angle X-ray Scattering (SAXS)** studies (Pérez et al. 2009). It was also shown that, although disordered, the Unique domain displays a certain degree of pre-organization - i.e. it is not a random coil (see a more detailed description in sub-section 2.1.1).

As mentioned in the former section, because of the challenging properties of IDRs, the

functional relevance of c-Src Unique domain remained obscure until recent years when, besides other protein-protein interactions and phosphorylation events described above, an additional Unique domain lipid binding region was described (Pérez et al. 2013). Unlike the SH4 domain, this membrane binding motif does contain hydrophobic residues like L62, F64 and F67, but neither penetrates in the lipid layer.

1.3.3 The SH3 domain

Adjacent to the Unique domain is the **SH3 domain**, the first folded domain forming the core cassette shared by all SFKs (Yu et al. 1992). It is a small folded domain, only 65 amino acid long arranged in a β sandwich made of five antiparallel strands. Between strands β1 - β2, and β2 - β3, are two prominent and flexible loops, named RT and nSrc, respectively. The latter has an isoform present in neural tissue in which T117 changes for TRKVDVR. A smaller loop termed distal extends between β3 - β4, and finally a short 3_{10} helix kink is located between strands 4 and 5. The β sheet formed by strands 2, 3 and 4 forms an hydrophobic surface that is responsible for the main SH3 function: recognition of polyproline II helices (**PPII**). The RT and nSrc loops lay at one of the edges, and participate in additional, weaker interactions with the ligand. SH3-PPII binding serves a dual function: it binds the enzyme ligand or other substrates related with localization, and also stabilizes an inactive conformation via intramolecular interactions (see below).

Two additional roles have been recently discovered by our group: lipid binding and **interaction with the intrinsically disordered SH4 and Unique domains** (Pérez et al. 2013). The RT and nSrc loops show weak interaction with lipids as shown by NMR experiments but, most importantly, they are also sensitive to the presence of the N-terminal IDR. These results led to further studies in which it was demonstrated that the SH3 domain acts as a scaffold for the IDR via multiple weak contacts (Maffei et al. 2015).

1.3.4 The SH2 domain

Towards the C-terminus, it follows the **SH2 domain**, a 100 amino acid long phosphotyrosine (pTyr) binding module (Filippakopoulos et al. 2009). While pTyr constitutes the main recognition feature, adjacent residues -2 to +4 provide further affinity and selectivity. The tertiary structure is formed by a three stranded β sheet and two α helices, one packed against each face. The pTyr binding site is located at one border of the sheet, in the cleft formed by the extremes of the helices. As the SH3 domain, SH2 also fulfills a bi-

nary role, one as an auto-regulatory switch (see below), and another recruiting substrates. A short but functionally important 22 amino acid PPII helix segment links the SH2 and SH1 catalytic domain, whose role I comment below regarding c-Src regulation.

1.3.5 The SH1 (catalytic) domain

The **SH1 domain** is 250 amino acids long, and is organized in two parts: a small N-terminal lobe, and a large C-terminal lobe, connected by a flexible hinge. The N-lobe is formed by a five-stranded β sheet and an important α helix (αC) between β3 and β4, while the larger C-lobe consists on five tightly packed α helices. The key ATP binding site sits in the cleft between lobes (Xu et al. 1997), where two conserved hydrophobic patches, one from each lobe, form the catalytic *spine* that holds the adenosine group. An additional hydrophobic *spine* further stabilizes the catalytic site (Foda et al. 2015).

Finally, at the C-terminal end of the SH1 domain is the regulatory tail (12 amino acids long). This short segment contains a residue that is crucial for c-Src regulation: Y530. Moreover, the lack of this residue in the constitutionally active v-Src is the main reason of its transforming capacity (Tanaka & Fujita 1986). The work of Matsuda et al. (1990) together previous studies observed that c-Src activity strongly depended both on the SH2 and SH3 domains, and phosphorylation of Y530, suggesting that the folded cassette had a built-in *lock*. The subsequent X-ray structures of c-Src active and inactive conformations [Xu et al. (1997); Xu et al. (1999); Cowan-Jacob et al. (2005); figure 1.8] provided high detail on the mechanism, which involves large scale domain rearrangement (Huse & Kuriyan 2002). Harrison (2003) eloquently described the main components of c-Src's auto-inhibitory mechanism as *a latch,a clamp and a switch*, which I pass to describe in the next sub-section.

1.3.6 The auto-inhibitory mechanism of c-Src

Y530 is mostly found phosphorylated in cells (Bjorge et al. 2000), typically due to the action of the CSK or CHK kinases[3]. This makes the SH2 domain target the C-terminal regulatory tail and pack tightly against the C-lobe of the SH1 domain, at the opposite side of the catalytic site. As a consequence, conformational rearrangement on the αC helix of the lobe blocks the ATP binding pocket thus abrogating activity, hence the denomination: the *latch*.

[3]The reference provided comprehends a review on the diverse kinases and phosphatases responsible for c-Src activation/deactivation, a subject out of the scope of this thesis.

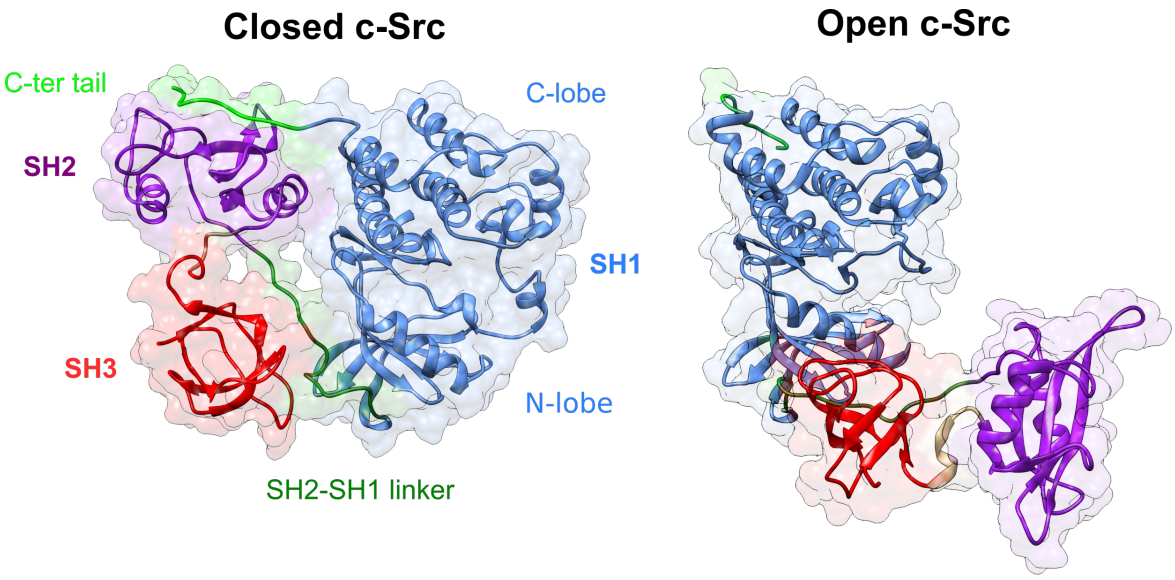


Figure 1.8: c-Src X-ray structures of the closed (PDB:1SRC) and open (PDB:1Y57) states. Notice that the IDR is missing in both. The domain color legend is the same as in figure 1.6.

The large rearrangement also facilitates the interaction between the SH3 domain and the SH2-SH1. Although it does not contain the exact SH3 recognition motif *(K/R)xxPxxP* (where x is any amino acid), the linker adopts a PPII helix conformation which binds to the SH3 hydrophobic binding patch. Additionally, the linker is in contact with the N-lobe of the SH1 domain, allowing further interactions between the SH3 domain and the catalytic domain. Therefore, although the SH3-linker interaction is not completely specific, is is effectively sandwiched between the SH3 and SH1 domains. The whole set of contacts further packs the folded cassette in a locked conformation, and destabilizing mutations in this triple interface lead to activation (Gonfloni et al. 1997). The tight packing of the SH2 and SH3 domains against the catalytic SH1 is therefore *the clamp* referred by Harrison (2003) (see figure 1.9).

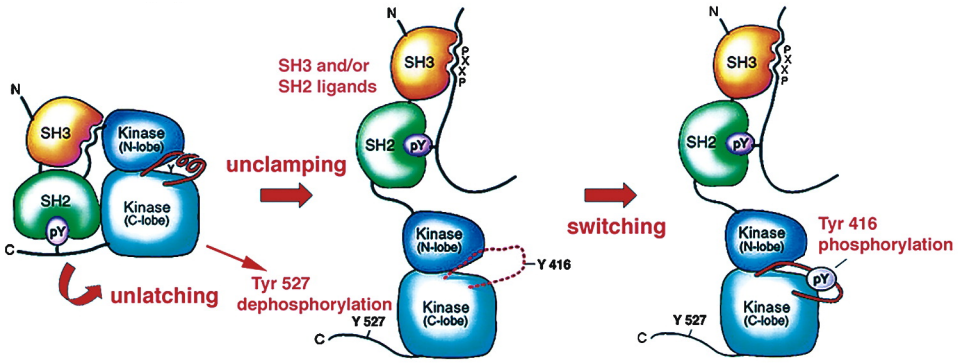


Figure 1.9: c-Src activation mechanism. Reproduced with permission from Harrison (2003).

When pY530 is dephosphorylated by the action of phosphatases, the *latch and clamp* are released: the SH2 and SH3 domains detach from the SH1 domain and swing outwards

(see figure 1.9), while the catalytic site rearranges into an open, active conformation. The latter consist mainly in a displacement of the αC helix and breakage of the catalytic spine.

Still, subsequent phosphorylation of Y419 in the so-called activation loop of the C-lobe enhances the enzymatic activity by further stabilizing the ATP binding site. This additional activation step is *the switch*.

The role of inter-domain dynamics is of paramount importance regarding c-Src activity. Using SAXS to study a constitutionally active c-Src mutant (Y520F) in solution, Bernadó et al. (2008) demonstrated that, even when the latch is fully released the SH2-SH3 clamp remains at an equilibrium shifted towards the closed form (85 %). This fact further evidences the need for a finely tuned control mechanism in c-Src, since such a small conformational population have dramatic biological effects. Indeed, the multi-domain cassette comprising the SH3, SH2 and SH1 domains exquisitely depends on relative domain traslation and orientation and several intramolecular allosteric mechanisms (Fajer et al. 2017), all components being tightly interdependent (Gonfloni et al. 2000). The network of allosteric connections that serves as a signal relay between the regulatory domains and the catalytic site of c-Src is described and experimentally tested in Foda et al. (2015).

The reader may have noticed that the IDR is not mentioned in this canonical autoinhibitory mechanism. The hinge between the SH3 and Unique domains lies at the opposite side of the *PxxP* binding site, so it was classically assumed that the disordered N-terminal region was an extravagant leash to anchor the cassette to the membrane. However, the results obtained by our group in recent years (Pérez et al. 2013; Maffei et al. 2015) showed that the IDR including the SH4 and Unique domains interact with functionally important zones of the SH3 domain. These observations therefore raise the question of whether the IDR is somehow functionally connected with the cassette.

1.4 The Src Family Kinases (**SFKs**)

The Src Family Kinases is a group of closely related tyrosine kinases comprising eight members in humans: **Src**, **Yes**, **Fyn**, and **Fgr** as the **SrcA subfamily**; and **Blk**, **Hck**, **Lck**, and **Lyn** as the **SrcB subfamily** (Thomas & Brugge 1997). The more distantly related **Frk** is considered by some authors as a separate subgroup on its own. As c-Src, all of them are involved in cell signaling pathways, but they differ in expression, cellular location and specificity. For example, c-Src, Yes and Fyn are ubiquitously expressed (although particularly high levels are specific to some cell types), whereas Blk, Fgr, Hck, Lck, and Lyn are mostly found in hematopoietic cells. The variety of cellular processes

they control is broad, as described in the case of c-Src, but there is a remarkable level of functional overlap between them, both regarding upstream and downstream partners.

As previously commented, all SFKs display a large sequence and therefore structural homology, except for the intrinsically disordered domains (see figures 5.1 to 5.3 in the Appendix). The SH4 domain of c-Src is particular since the rest of SFKs has, in addition to myristoylation, palmitoylation as the second membrane binding signal instead of a polybasic amino acid cluster. The Unique domains vary widely in length, from the 90 amino acids in Yes to only 57 amino acids in Blk. Sequence alignment shows an extreme divergence, and only some short (3-4 amino acids) coincident patterns are detected at first sight. Only their amino acid composition, enriched in polar residues and depleted of hydrophobics as is typical of IDPs (see next section) is common and denotes their nature. Only the Unique domain of Lck has a defined structural function nicely reviewed by Boggon & Eck (2004): it contains a *CxxC* motif together with the complementary *CxCP* motif in the CD4 or CD8α T-cell coreceptors (also located in an IDR), form a stable zinc finger. Beyond that, phosphorylation is a common phenomenon in all SFK Unique domains, also related with the nature of IDRs. These modifications and their biological effects have been reviewed by Amata et al. (2014), although the structural basis of the mechanism behind is unknown.

The folded modules, SH3, SH2 and SH1 domains, share an extremely similar structure among all SFKs, specially regarding the secondary structure elements. The general descriptions for each module done in the previous sub-section apply for all members, so I will not make a detailed comparative description here. However the interested reader can find exhaustive reviews in Sicheri & Kuriyan (1997) (focused on Hck vs Src), Boggon & Eck (2004) (a structural approach, includes the auto-regulatory mechanisms), Engen et al. (2008) (a review of biological, biophysical, and computational studies).

I will however discuss a more general aspect that is relevant for this thesis: the conservation of structural disorder and flexibility among SFKs. Santos & Siltberg-Liberles (2016) have performed a computational analysis of tyrosine kinase sequences (not only SFKs) sharing the SH3-SH2-SH1 cassette present in c-Src, focusing on the dynamic regions of the cassette in order to get insight on the conservation of allosteric networks involved in regulation[4]. In their approach sequences of numerous orthologues and paralogues were used, so evolutionary dynamics could be examined.

The secondary structure conservation patterns were consistent among all kinases (also non c-Src related) since structure is typically more conserved than sequence, in agreement with

[4]Another interesting aspect of this work that I will not discuss here is the conservation of phosphorylation patterns among these kinases.

the resolved structures available in the literature.

The most interesting result regarding this thesis were the results about disorder quantification, distribution and conservation. The authors argue that, in general, the location of functional disordered regions is clade-specific, suggesting that they serve particular fine-tuning functions in each case. In the case of SFKs, the phylogenetic tree shows that the whole group is closely related, and the regions share location: beyond the SH4 and Unique domains, most of the disorder concentrates in the SH3, SH2 and SH2-SH1 linker (the *clamp*). The main difference between SFKs lays in the amount of disorder.

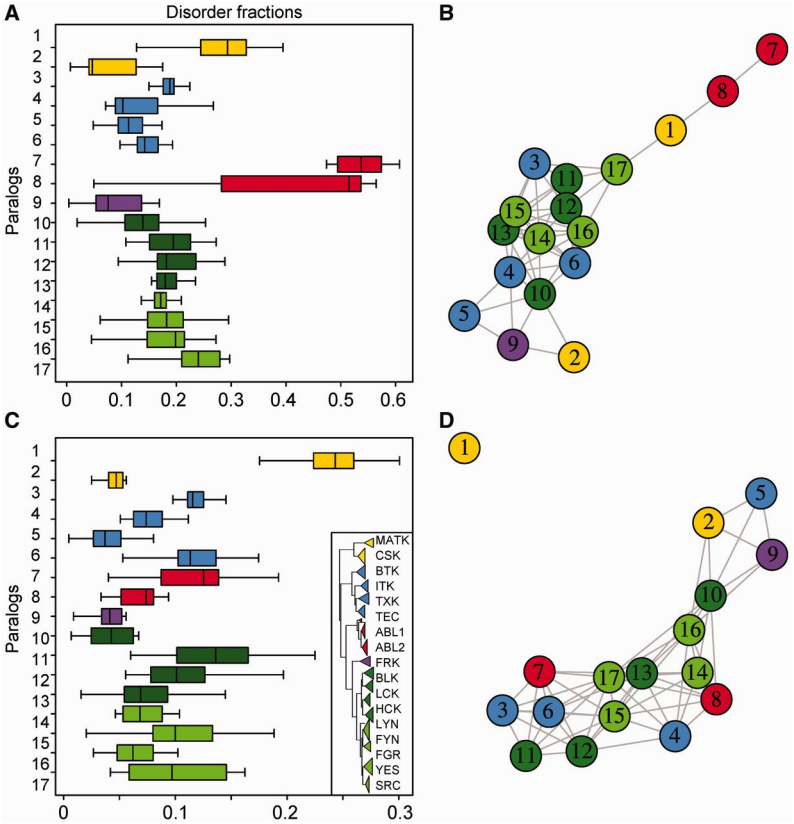


Figure 1.10: Fraction of structural disorder among tyrosine kinases based on IUPred scores (Dosztanyi et al. 2005), where 0 means ordered and 1 fully disordered, for A) the full length proteins or C) the cassette domains. On the right, the respective network similarity analysis (B, D). SFKs, including Frk, correspond to numbers 9 - 17 (see legend). Adapted with permission from Santos & Siltberg-Liberles (2016).

Considering full length proteins including the SH4 and Unique domains, c-Src is the SFK member containing more disorder, followed by Yes and Lck. Blk is in contrast significantly devoid of disorder, while the rest of SFKs follow a similar distribution. The trend is less clear when considering the SH3-SH2-SH1 cassette exclusively. The absolute amount of disorder is obviously reduced and Yes falls back significantly, but c-Src remains among the most disorder-containing SFK.

Interestingly, the statistical significance networks of the full length sequences clusters to-

gether c-Src, Lck, Hck, and Yes, leaving apart Fgr or Lyn, which share the SrcA subfamily. Lyn (head of SrcB), Fyn and Fgr group together, while Blk is an outlier. When only the folded cassette is considered the nodes are more scattered and loosely related and only c-Src, Lyn, and Fgr are together. These variations show that the specific contributions of the respective IDRs suppose an important fraction that should be accounted for.

Although Santos & Siltberg-Liberles (2016) focus on the cassette and mostly exclude the N-terminal disordered regions, it is relevant in the context of this thesis that these long time ignored segments comprise an important potential regulatory element not only in c-Src but also for the whole SFK, although the mechanisms are yet unknown and may be diverse. Recent studies (Kathiriya et al. 2014) also suggest that IDRs drive kinase-kinase interactions which are fundamental in signaling pathways and pathologies derived from their deregulation. Their results show that 90 % of kinase-kinase interactions involve at least one IDR-containing partner. Therefore, the role of the SFK N-terminal SH4 and Unique domains may extend further than self-regulation of the activity.

1.5 Intrinsically Disordered Proteins (**IDPs**)

Up to this point I have mentioned the presence of functionally relevant but obscure N-terminal intrinsically disordered regions not only in c-Src, but in all SFK members. It is now necessary to introduce the characteristics of this kind of proteins/regions, their biological functions and relevance (specially regarding protein-protein interactions), and the methods used for their study.

For the sake of clarity, I selectively cover a few aspects of IDPs relevant for the work presented here. However, the interested reader can find more detailed discussions on topics only briefly mentioned here in the special issue of Chemical Reviews on IDPs from 2014 (Uversky 2014) or in the introductory book by Tompa (2009).

As a clarification, I would like to state that the differentiation between Intrinsically Disordered Protein (IDP) and Region (IDR) is somewhat undefined. IDRs are considered significantly long disordered stretches, the size threshold typically being >20 - 30 amino acids. Similarly, IDPs are not to be considered proteins that are completely disordered from end to end, but rather proteins containing a significant degree of intrinsic disorder. We can again establish a proportion threshold of >20 - 30 % of amino acids in the sequence being disordered. Thus, when discussing generalities on structural disorder, both terms can often be exchangeable.

1.5.1 Defining intrinsic disorder

From a purely chemical point of view, a protein is just a polymer made of amino acids via amide bonds and, potentially, disulfide bridges between cysteines (aside from co or post-translational modifications). Such a simple definition however contains an astronomical number of possibilities: From the standard 20 amino acid dictionary, a 100 amino acid chain allows for 20^{100} theoretical sequences. To give a sense of scale, this is a far larger figure than the number of atoms in the known universe, estimated to be in the order of 10^{80}.

Regarding the possible structures a given sequence of N amino acids can adopt, a simplified model where only the ϕ and ψ Ramachandran angles of the backbone are considered illustrates that the freedom degrees grow exponentially with the number of residues. Thus, a layer of potential structural diversity adds on top of the possible sequence combinations.

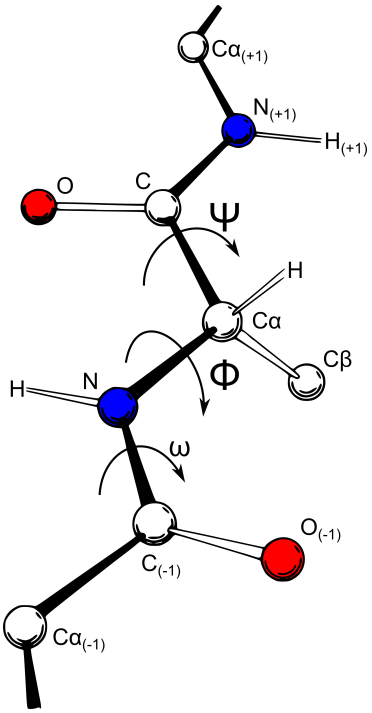


Figure 1.11: Protein backbone dihedral angles. Adapted with permission from original drawing by Dcrjs, vectorised by Adam Rędzikowski (Wikimedia).

However, proteins are not just polymers, but biological machinery: tools that organisms have found to provide an advantageous function for their survival, and are selected according to their performance to do so. Hence, as biological systems, the key concept we need to keep in mind to analyze them is **evolution** (Dobzhansky 1973). In the vastness of the protein sequence space available, Nature has extremely narrowed its evolution-driven search (Smith 1970). A large fraction of the potential sequences correspond to peptides that would uncontrollably aggregate under generic physiological conditions. These polymers

are, in general, functionally useless (moreover, potentially pathogenic and / or lethal for an organism) and are not selected by evolution. Consequently, only some sparse clusters in the immense chemical space are populated.

One of them groups what we know as folded proteins: peptide chains that can collapse in a controlled manner to a particular conformation. Doing so, a certain amino acid sequence adopts a stable (although dynamic) structure with the ability to perform a specific function, which can thus be selected by evolution. The respective correlations between sequence and spatial configuration, and the latter with functionality, are the so called **Anfisen's dogma** (Anfinsen 1973) and **structure-function paradigm** (Edsall 1995).[5] These concepts were early recognized in structural biology and have been keystones in the field for decades, to the extent that today it is possible to calculate a close approximation to a protein structure exclusively from its sequence (Moult 2005) and also predict its function (Whisstock & Lesk 2003).

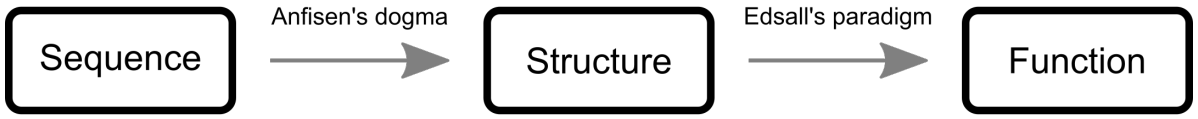


Figure 1.12: Classical structural biology univocal sequence-structure-function relationships.

Another evolutionarily selected region of the sequence space is that of proteins that do neither aggregate nor collapse, but are still functional. Instead, they remain extended in solution, continuously inter-converting between a vast number of conformations. These are known as **Intrinsically Disordered Proteins/Regions** (**IDPs/IDRs**)[6]. The particular amino acid usage conferring these properties determines a *non-folding* code (Williams et al. 2001) for whose identification and prediction many reliable methods have been developed (He et al. 2009; Li et al. 2015).

The fact that IDPs escape the structure-function criterion succinctly defined above, and present characteristic sequence composition statistics, has served as a classification principle, but does not clarify what functional consequences derive from intrinsic disorder or, ultimately, its evolutionary advantages.

[5]These concepts are here enunciated in their most simple form, assuming univocal correlations, as a plain introduction to the initial *fixed-structure-centric* view in the field. However, both notions have been superseded and refined to include dynamic phenomenons such as allosterism, as it is later explained.

[6]The evolutionary constrains through which protein ensembles lacking a well defined equilibrium structure are selected are matter of further discussion beyond the scope of this introduction, but have been reviewed elsewhere (Siltberg-Liberles 2011; Siltberg-Liberles et al. 2011; Wei et al. 2016).

Indeed, defining intrinsic disorder is not a trivial task. The most comprehensible approach is to make a negative definition based on the safe ground of classic structural biology. As noted by Tompa (2009), we can contrast various odd experimental properties with the ones a protein scientist would expect, and use that as an indicator for disorder. Examples include, but are not limited to: enhanced sensitivity to proteolysis, resistance to denaturation (heat, acid, alkali or chaotropic agent-induced), poor signal dispersion in NMR spectra, larger than expected hydrodynamic dimensions… These and other signs score for a *disorder diagnosis*, and are in fact reliable. However, not all these features are mandatory and they may be present to different extents in a system-dependent fashion. Hence, such an diffuse and heterogeneous definition is not sufficiently precise and sound to put the researcher's mind at ease.

The main difference between ordered and disordered proteins lays in the respective relationships between the different regions of sequence space and the geometry of their associated **free energy landscapes** (Konrat 2010)[7] (see figure 1.13). Folded proteins are not absolutely rigid, they possess flexibility and most often need coordinated spatial rearrangements to function[8]. However, it is nowadays accepted that, after a certain stochastic stage in which primary intramolecular contacts are established, the chain finally collapses towards an equilibrium configuration, and then remains fluctuating close to it (Dill & Chan 1997). The underlying principle behind this behavior is the rugged energy landscape of the system, where the deep wells and flatter regions alternate giving place to a Boltzmann-weighted population of conformations that is heavily shifted towards the equilibrium configuration(s).

In contrast, IDPs continuously fluctuate between a large number of energetically similar configurations as a consequence of the relatively featureless and flat free energy landscape determined by their sequence (see 1.13). This conformational meandering has been suggested to be a chaotic regime (Uversky 2016), as is the initial step of protein folding according to **Molecular Dynamics (MD)** simulations (Braxenthaler et al. 1997). The resulting behavior is thus dominated by complex dynamics with different timescales and amplitudes.

As previously introduced, in the case of folded proteins the general correspondence between sequence and free energy landscape and therefore spatial configuration distribution

[7]Free energy landscapes are representations of the system's free energy in terms of its degrees of freedom. Since there is a colossal number of them in a protein, the choice of degrees used to meaningfully represent the hyperdimensional free energy surface is arbitrary (e.g. the Ramachandran angles).

[8]Proteins may have more than one stable conformation depending on the particular functional mechanism, but it typically involves discrete transitions between a reduced number of possibilities. As an example, see the case of myoglobin described in the seminal work on energy landscapes and protein dynamics by Frauenfelder et al. (1991).

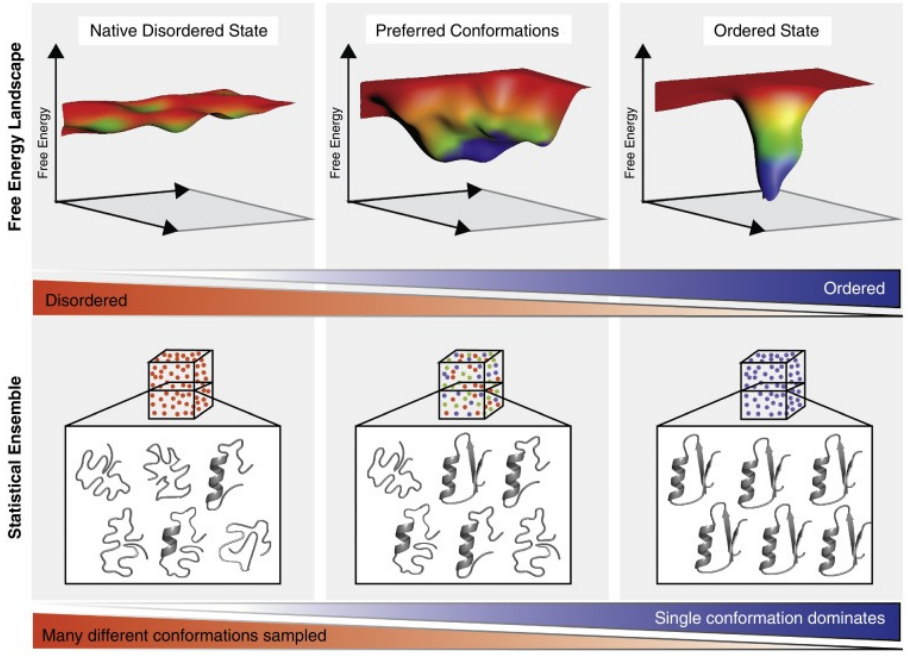


Figure 1.13: Free energy landscapes associated to different contents of disorder. Reproduced with permission from Flock et al. (2014).

is well characterized thanks to the massive body of knowledge accumulated over the years. So, we know the amino acid rules that encode specific energetically stable spatial configurations in a certain sequence (Koga et al. 2012), and even the reciprocal problem of designing novel functional templates and sequences has been successfully tackled (Chevalier et al. 2017). However, that is not the case for intrinsically disordered proteins.

Since the same building blocks - the 20 standard amino acids - and the same laws of physics apply in both cases, it may seem that the problem with IDPs is a lack of sufficient experimental evidence from which general principles can be extracted[9]. However, it is not the only reason. The ever-increasing information we have from disordered systems only but emphasizes the wide variety of situations involving this phenomenon: from IDPs that form compact but dynamic *molten globules* to almost featureless random coils, including all the range in between, and even the possibility of context-dependent transitions. The same variety applies to IDP functionality.

It is the diverseness of disordered systems combined with the inherent difficulty of characterizing moderately to highly heterogeneous conformational ensembles due to their flat free energy landscapes, either experimentally or computationally, what has made the elemental rules of disorder (if there are any) elusive so far. Therefore, a large part of our current knowledge is limited to a phenomenological classification of disparate cases based

[9]As it is related in the following sub-section, the field of IDPs only bloomed by the mid-late 1990s, while the first structures from ordered proteins date back from the late 1950s.

on sequence features, function, cellular location, etc. (Uversky 2002; Vucetic et al. 2003; Lee et al. 2014), despite great advances in theoretical frameworks have been done, such as the concept of metastructure (Konrat 2009; Naranjo et al. 2012) or the poly-electrolyte approach (Mao et al. 2013; Das et al. 2015).

1.5.2 Towards a new structure-function paradigm

The cornerstone of protein structural biology was laid by Emil Fischer with the *lock and key* model he proposed to explain enzyme stereospecificity (Fischer 1894). His idea relied on strictly geometric constrains and assumed that the interaction between a specific protein and its substrate takes place *only if the shoe fits*. Up to the the mid-20$^{\text{th}}$ century, experiments reporting the correlation between protein denaturation and enzymatic activity, such as the work of Mirsky & Pauling (1936) with pepsin, experimentally supported the notion that function needs structure (Edsall 1995).

Later on in the late 1950s, the first X-ray crystallography protein structures were obtained (Kendrew et al. 1958) and the success of these detailed static snapshots would definitely settle the structure-function paradigm. However, indications that a strict structure may be the main but not the only determinant of protein functionality had already been reported.

In a foreseeing work on the capacity of albumin to specifically recognize diverse small molecules with the same binding sites (8 years before the first protein structure was solved!), Karush (1950) had suggested **configurational adaptability** as the underlying mechanism. In his words:

> "By [configurational adaptability] is meant that there exist a number of sites on the protein, each associated probably with several side chains, which to a varying extent can assume a large number of configurations in equilibrium with each other and of approximately equal energy." (Karush 1950).[10]

The role of dynamics in protein-substrate recognition and activity started to be formulated in the 1960s after experimental observations that could not be explained by single static structures. The pioneering works of Koshland and Monod (Koshland 1958; Monod et al. 1965; Koshland et al. 1966) described how some proteins and/or ligands necessarily experimented conformational changes upon binding events, leading to the recognition of the *conformational selection* and *induced fitting* models, which still prevail and are discussed today (Gianni et al. 2014).

[10]Note the astonishing coincidence with the definition of intrinsic disorder given in sub-section 1.5.1.

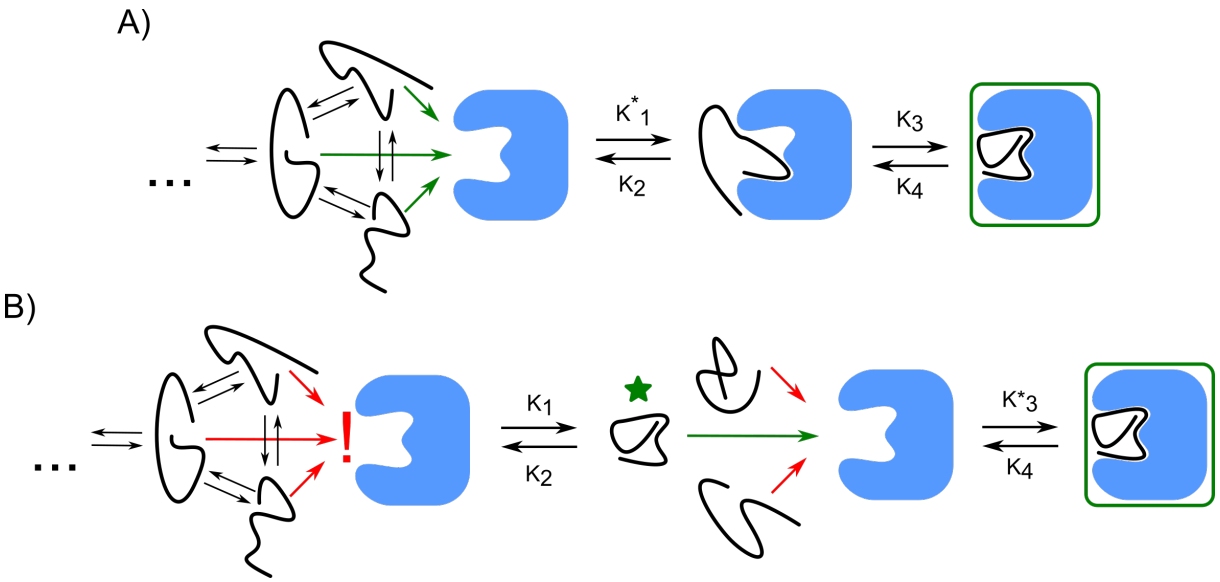


Figure 1.14: A) Induced fit vs B) conformational selection models. Asterisks indicate pseudo-first order steps.

While conformational selection postulates the pre-existence of a population distribution states with different binding affinities, induced fitting assumes that the adoption of the most competent configuration takes place upon binding. These concepts are directly connected to **allosterism**, the process by which an event in a specific part of a biomolecule (typically binding) induces a change in a distant site, altering its activity (e.g. binding affinity to the same or other ligand).

Thus, molecular recognition and regulation, and therefore protein activity were early known to depend on the system's dynamics, but the approach would remain structure-centric. Further work on the kinetics and thermodynamics of protein mechanisms was hence mainly based on transitions between few fixed structures and, paralleled by the growing number of structures being resolved, resulted in a wide and deep understanding of cellular process at the molecular level.

The fact that disordered regions fail to crystallize or, at best, present very poor electron densities in X-ray crystallography further kept them underlooked, even though most often disordered tracts alternate with folded regions as in modular proteins like SFKs. Dunker et al. (2001) argue that the classic protein structure hierarchy (primary, secondary, tertiary and quaternary levels) has also contributed to the historical understatement of functional disorder, even when its importance for protein function was often recognized. A structure-function *culture* thus became predominant in the field for decades, and intrinsically disordered proteins/regions were considered outliers to the general trend rather than a class on their own until the late 1990s (Dunker et al. 2001).

The advent of biomolecular NMR (Wüthrich 1995), specially from the late 1970s on, would put the focus on dynamics. This powerful technique (see sub-section 1.9.2 for further details) also allowed to resolve the structure of proteins not amenable to X-ray crystallography, but most importantly, gave unprecedented insight on their internal motions (Wüthrich & Wagner 1978). An example of its application is the characterization of the allosteric structural relays through which information is transmitted between distant regions of a protein (Tompa 2014).

An aspect of structural biology research that resulted instrumental in re-shaping the structure-function concept is **protein folding**[11]. Although IDPs have been also termed *unfolded* proteins, there is a subtle but capital detail: IDPs are not proteins yet to fold, since their thermodynamics just do not eventually lead to such situation. On the contrary, denatured proteins can most often be refolded and regain activity by adopting the same native structure that was lost.

As summarized in Dill & Chan (1997), the resolution of the Levinthal paradox (Levinthal 1969) - i.e. how an amino acid chain is able to find in the order of seconds (or less) the most stable configuration among the vast *hay-stack* of potential conformations - first led to postulate the existence of pre-existing thermodynamic pathways and then to a statistical view of protein structures as populations arising from the free energy landscapes previously introduced. The concept of **conformational ensembles** for the characterization of dynamic phenomena has been thence adopted (Tsai et al. 1999). The conceptual basis therefore shifted from that of a static structure over which dynamics are then incorporated, to a thermodinamically populated conformational space which ultimately defines function.

The ensemble approach was obviously applied in the field of IDPs as a fundamental concept to tackle their intrinsic conformational heterogeneity (Dunker et al. 2001), and has been further extended to better understand processes like molecular recognition or allosterism by disordered systems (Boehr et al. 2009; Motlagh et al. 2014). Thus, although the structure-function paradigm holds in many relevant cases, it has been recognized that the structural continuum from strictly ordered to fully disordered systems, their interplay, and the mixed use Nature does of them, require a re-evaluation of the dogma (Wright & Dyson 1999). Nowadays, an increasing number of ensemble models of IDPs is being developed and becoming available to the community (Varadi et al. 2014).

Finally, an aspect that is particularly important for the study of IDPs is the dynamics of the free energy landscapes themselves. Regarding molecular recognition by IDPs,

[11]The important role of chaperones in protein folding is not covered in this introduction, but they also constitute examples of highly dynamic and structurally plastic proteins.

Boehr et al. (2009) discuss that most often, conformational selection and induced fit work in synergy. Structural heterogeneity is already implicit in the conformational selection model, and relates with the dynamic nature of IDPs in a straightforward manner. However, the authors interpret induced fit as a **population shift** on the conformational ensemble, rather than a well defined conformational transition following the binding of a unique competent configuration. Free energy landscapes in IDPs are weakly rugged, with landmarks in the order of $k_B T$, thus providing an enhanced sensitivity to environmental cues arising from fine tuning of the ensemble statistical properties.

1.5.3 A brief history of intrinsic disorder protein and its relevance

The field of intrinsically disordered proteins has settled and witnessed a huge expansion in the last 20 years, and is now an integral part of structural biology. However, although not systematically established yet, the phenomenon of intrinsic disorder was early recognized in the 1950s, as occasional reports of odd behaviors in proteins demonstrate (Dunker et al. 2001).

An early and clear example of protein structural disorder is that of milk caseins, proteins that help digestion and are therefore functional (reviewed in Tompa 2009). In his work, McMeekin (1952) compared spectroscopic properties of caseins and denatured globular proteins and deduced that caseins adopted compact extended conformations in solution[12]. Further research would demonstrate that these proteins almost lacked secondary structure elements. Therefore, caseins were soon recognized as an exception, fully random coil proteins without any of the classic structural motifs.

As I previously introduced, artificially denatured proteins were used as a model to study these proteins that are in a dynamic, structurally heterogeneous state. Concepts like the Flory random coil model (Flory & Volkenstein 1969) were borrowed from polymer theory in order to model their behavior (Tanford et al. 1966), and therefore also applied to IDPs. It would be later proven that IDPs are qualitatively different from denatured proteins regarding hydrodynamics, showing more compact conformations than random coils (Kohn et al. 2004).

Specially after the 1970s, more experimental evidence of structural disorder in relevant proteins was found. Such were the cases of myelin basic protein, MBP (involved in diseases

[12] An everyday life experimental evidence of the disordered nature of caseins is their resistance to heat: one can boil milk without signs of precipitation, whereas folded proteins like egg albumin usually aggregate upon denaturation by heat, since their *sticky* hydrophobic cores become exposed.

affecting the nervous system), microtubule associated protein MAP2 (a neuronal protein involved in microtubule stability), the functional tail of DNA-binding histone H5, and others like fibrinogen or glucagon. As commented in the previous sub-section, protein NMR permitted for the first time the observation of dynamic regions and proteins, and the amount of detailed evidence of structural disorder rapidly accumulated, specially during the 1990s. Relevant examples include the widely studied α-synuclein, related with amyloidogenic neuronal pathologies as Alzheimer's disease (reviewed by Uversky 2003), and the IDRs in the p53 tumor suppressor protein (reviewed by Joerger & Fersht 2008) or PTEN (phosphatase and tensin homolog protein) (Malaney et al. 2013).

Another important aspect derived from of this new knowledge that helped to recognize intrinsic disorder as a quantitatively distinct phenomenon is **function**. It was noted that structural disorder was something different from the flexibility often found in the short loops of folded proteins. It affected more extended regions with specific functionality (Schlessinger et al. 2011), specially translation and transcription regulation. The paradigmatic case of p21, which was found to bind and inhibit several complexes between Cdk and different cyclins (Kriwacki et al. 1996) offered an insight on why disorder may be a desirable or even indispensable feature for protein activity (Plaxco & Groβ 1997).

Despite of the inertia of the structure-function paradigm in the field, the overwhelming evidence showing that intrinsic disorder is a common phenomenon presented in the pioneering works of Romero et al. (1998) and Garner et al. (1998) settled a solid basis for the recognition of IDPs. Using neuronal networks trained with experimental information from X-ray and NMR structures, the authors were able to predict IDRs > 40 amino acids in over 15 000 proteins from the Swiss Protein Database, 1 000 of them being completely disordered IDPs. Following work would further detail the wide spread of intrinsic disorder in Nature. According to Ward et al. (2004):

> "Putative long (>30 residue) disordered segments are found to occur in 2.0% of archaean, 4.2% of bacterial, and 33% of eukaryotic proteins." (Ward et al. 2004).[13]

Wright & Dyson (1999) and Dunker et al. (2001) published the nowadays considered *foundational reviews* that helped to definitely solidify and disseminate the study of intrinsic disorder. These works summarized the known examples and work done so far, vowed for the reformulation of the structure-function paradigm, and also posed questions that

[13]The exceptional presence of IDRs in eukaryotic organisms is tightly related with their role in signaling pathways and organism complexity. Remarkably, in Dunker et al. (1998) the authors would link their predictions with the implications of structural disorder in molecular recognition. These points are further discussed in section 1.6.

have only recently been answered as, for example, if disorder is retained in the crowded intracellular environment (F.-X. F. Theillet et al. 2014).

Another important aspect of intrinsic disorder that has drawn attention to IDPs is the correlation between structural heterogeneity and human disease (Uversky et al. 2008; Tompa 2009; Uversky et al. 2014), most importantly tumoral (Iakoucheva et al. 2002) and neurodegenerative (Uversky 2009), but also cardiovascular pathologies (Cheng et al. 2006), or diabetes. Although some of the most prominent IDPs involved in disease (p53, PTEN, α-synuclein, etc.) have been extensively characterized, the intrinsic difficulty to target extremely dynamic proteins with small molecules with acceptable affinity and selectivity has severely hampered clinical application (Chen & Tou 2013). Only a few cases of satisfactory binding have so far been reported (e.g. Iconaru et al. 2015; Neira et al. 2017), but mostly as a proof of concept for IDR drugability (Heller et al. 2017). Thus, the current efforts mostly focus on targeting the protein-protein interactions and complexes involving IDRs.

The major prevalence of structural disorder in cancer arises from the fact that IDPs are often part of important signaling pathways controlling cell cycle and other key functions, as in the case of c-Src. In the following sub-sections I will thus discuss in more detail this aspect of the disorder-disease relationship, emphasizing on protein-protein interactions, which most often are at the molecular basis of IDP misbehavior.

1.6 IDP functions, cellular signaling and disease

Whole proteome disorder predictions highlight that IDPs/IDRs are more prevalent in certain cellular functions, specially transcription and translation regulation, and cell signaling, whereas others as catalysis or structural roles are depleted from it. This uneven distribution originates from the unusual properties of IDPs due to their dynamics (Wright & Dyson 2014), and is most often related with **molecular recognition**.

Enhanced plasticity permits IDPs to engage in transient interactions with multiple partners (simultaneously or not), or multiple interactions with a single partner, with significant affinity. Selectivity is often conferred by the presence of small interactors embedded in the sequence, termed **Short Linear Motifs** (**SLiMs**), **Eukaryotic Linear Motifs** (**ELMs**) (Tompa et al. 2014) or **Molecular Recognition Features** (**MoRFs**) (Kotta-Loizou et al. 2013), depending on the definition criterion. These short stretches (typically < 10 amino acids) are modular recognition elements which provide specificity and promiscuous multivalency (Cumberworth et al. 2013), since IDPs often harbor a number of them

and can thus interact with a variety of partners. **Multifunctionality, plasticity, and modularity** therefore emerge as a common functional traits of IDPs

The modification of a motif, either via post-translational modification - phosphorylation, glycosylation, acetylation, etc. - or alternative splicing can also adjust the IDP activity by selectively suppressing or enhancing the role of a single or several motifs either permanently or transiently, so providing environment-dependent **dynamic tunability**. Since IDPs tend to sample extended conformations, their enhanced exposure makes them preferential substrates for modification.

The ability to undergo disorder-to-order (or vice versa) transitions, structural preformation, local propensities to adopt more ordered configurations, or just shift the conformational ensemble, all of which can also be triggered by post-translational modification (Iakoucheva 2004), allows for different degrees of kinetics and affinities. **Adaptative dynamics** is therefore another novel feature, which also relates with fine control over IDP function.

All these traits are specially useful in the context of cell signaling. Signaling networks are tightly controlled systems at all levels, from protein expression to degradation, since their activity requires high spatiotemporal precision and fast responsiveness to a large number of possible inputs. Any deregulation in the cascades may lead to anomalous transmission of the information and pathological cell behavior as cancer. IDPs/IDRs have been found to be consistently present in signaling pathways, usually mediating protein-protein interactions and very often as important hubs (Iakoucheva et al. 2002; Wright & Dyson 2014).[14] The role of IDPs in cellular regulation, and its further implication in disease (Babu et al. 2011) is hence and active area of research. In the last years, it has been even demonstrated that pathogens (specially viruses) mimic short linear motifs to hijack and subvert the host signaling networks (Davey et al. 2011; Via et al. 2015), a phenomenon termed molecular mimicry.

An important aspect regarding the topic of this thesis is the relationship between ordered and disordered tracts in signaling proteins. Besides the functional advantages for signaling of IDPs by themselves, the cooperativity between regions with different degrees of structural heterogeneity brings novel regulatory modes in signaling. The conformational ensemble approach has thus been extended to tackle the role of structural disorder in modular signaling proteins as c-Src, which alternate ordered domains with flexible IDRs and recognition motifs.

[14]Some IDPs even play antagonistic roles in different pathways, such as GSK3, involved in Insulin and Wnt signaling routes.

Recent works from Tompa (2014) and Csizmok et al. (2016) have stressed that allosterism is an energetic coupling between distant regions, not necessarily based on structural changes, because it can also rely on dynamics. The *new ensemble view* challenges the structure-centric approach that considers allosteric pathways to be strictly mechanic and enthalpy driven, based on sets of coupled conformational changes between adjacent residues. The authors review different experimentally characterized examples, such as the DNA binding Ets-1 transcription factor (Pufall 2005), the Sic1 cell cycle protein (Mittag et al. 2008), or the *Drosophila* Ultrabithorax transcription factor (Liu et al. 2008), in which allosteric information is propagated to distant regions along IDRs without concomitant structuration. In these cases, allosteric remodeling of the protein free energy landscape means a shift in the conformer populations, which results in modified dynamics in the long-range distance and a specific functional output. Therefore, it is not necessarily the development of new, more stable conformations upon signal reception, but a precise change in the averaged set of intramolecular transient contacts what acts as a conduit for functional information along the molecule.

Tompa (2014) further argue that all cases of allosterism, including auto-inhibition (as the mechanism in c-Src, previously exposed in sub-section 1.3.6), and other unexplained mechanisms, are part of the more general concept of **multisterism**.

The functional consequences of entropy-driven multistery are variate, ranging from ultrasensitivity to external stimuli to establishment of very precise thresholds to trigger a response. The common novelty is that modular proteins become much more complex signaling machinery. Classic allostery models were developed to explain the non linear responses observed, for example, on protein activity towards increasing concentration of a unique allosteric effector - e.g. a ligand. These mechanisms therefore introduce the possibility of different conformations with different responses, but always in the context of **one input - one output**. This is, allosteric - or, in general, functional information transmission - mechanisms in ordered proteins are non-linear switches, in the sense that they respond to a single input although in a complex manner. In contrast, the inclusion of structural disorder gives place to **signal integrators**: modular proteins are able to respond in a precise and dynamic way to varying combinations of inputs with specific outputs (see figure 1.15). Therefore, IDRs permit Nature to program algorithmic behaviors in a single node of a signaling network, thus enabling fine control of much more complex systems by adding a new layer of sophistication.

Given the foremost role of IDR protein-protein interactions in these multisteric mechanisms of which c-Src is a clear example, I will next discuss the details and specifics of these contacts, with an emphasis on entropy-driven processes.

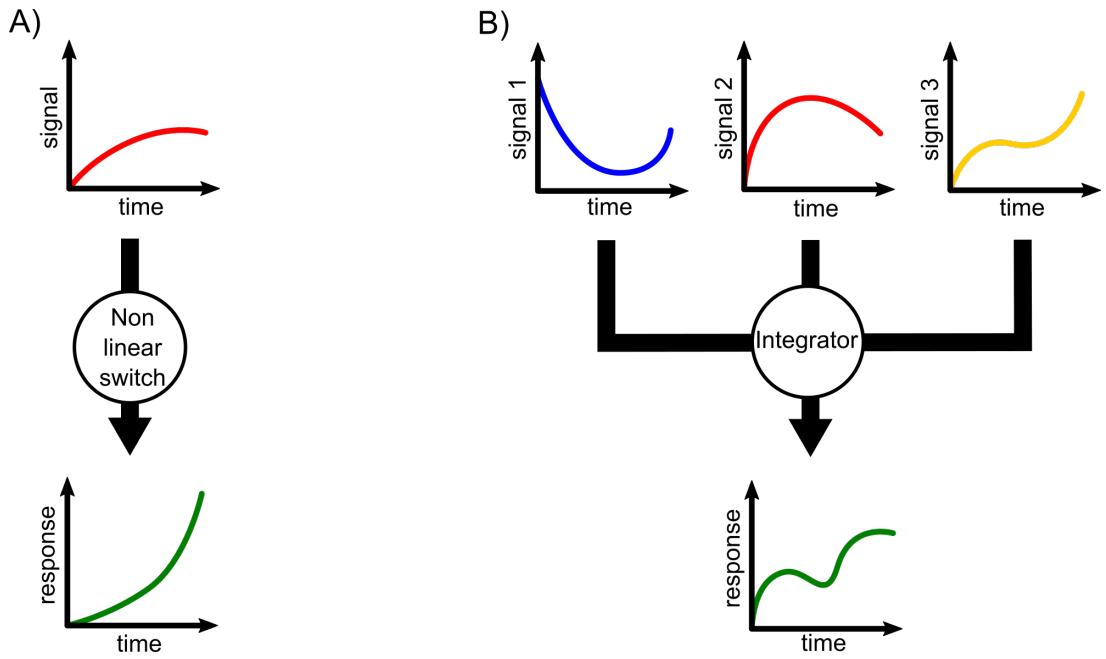


Figure 1.15: Schematic representation of A) a non-linear switch, and B) a signal integrator.

1.7 Protein binding by disordered proteins

It has been commonly accepted that protein-protein complexes involving IDPs/IDRs are characterized by high selectivity, often towards a variety of unrelated partners, combined with low affinity and short-lived association (Zhou 2012). The image is very intuitive: one can imagine how a protein or region which does not fold but remains dynamic (Liu & Huang 2014) can establish a number of alternative transient contacts that cooperatively build up a specific but weak association.

However, the protein order-disorder continuum present in Nature leads to an ample variety of situations when protein:protein interactions are considered, making difficult to establish a common framework. An array of molecular recognition phenomena involving structural disorder has been so far described (Uversky 2002; Tompa 2009; Uversky 2011; Dogan et al. 2014; Mollica et al. 2016; Schad et al. 2017), ranging from large multi-domain proteins and folded protein assemblies functionally depending on short disordered linkers (as is the case of the c-Src cassette domain) to IDP:IDP complexes; from femtomolar to milimolar affinities; from extreme specificity to entropic bristles indiscriminately engaging in promiscuous contacts; from long, low complexity, repetitive sequences to IDRs harboring a plethora of specific interaction motifs (SLiMs); from fast to slow association or dissociation, and thus from extremely short to relatively long-lived complexes; or from IDPs that fold upon binding to cases of a high degree of retained disorder. IDPs are thus recognized as *interaction specialists* able to perform multiple roles ranging from

entropic chains to specific wrappers, as reviewed in Tompa et al. (2015) and illustrated in figure 1.16.

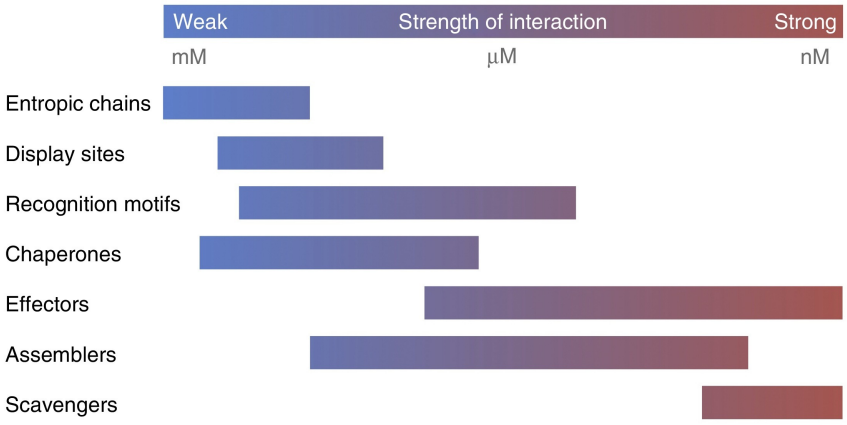


Figure 1.16: Different affinity ranges of IDPs accomplishing different functions. Reproduced with permission from Tompa et al. (2015).

The defining feature common to all these situations despite the functional diversity is that a significant degree of **conformational heterogeneity** (and thence consequent dynamics) **is present, at least in one stage of the recognition event, and at least in one of the partners**.

This fact obviously influences on the thermodynamics and kinetics of the process, for which we can define:

$$(1)\quad \Delta G^\circ = \Delta H^\circ - T\Delta S^\circ,$$

where ΔG° is the Gibbs free energy variation, ΔH° is the enthalpy change, T is temperature, and ΔS° is entropy variation, all of them in the standard state. From here, the reaction isotherm equation can be derived:

$$(2)\quad \Delta G^\circ = -RT \ln K_{eq},$$

where R is the ideal gas constant, and K_{eq} is the equilibrium constant. Finally, for the kinetics we can establish:

$$(3)\quad K_d = \frac{k_{off}}{k_{on}},$$

where K_d is the equilibrium dissociation constant, and k_{on} and k_{off} are the respective association and dissociation rate constants.

In their review on experimental thermodynamic data from binary protein complexes involving IDPs[15], Teilum et al. (2015) report remarkable differences between ordered:ordered and disordered:ordered assemblies, but also common features (see figure 1.17). The average ΔG° values of their data sets involving IDPs are in average 2.5 Kcal mol^{-1} less stable than the values from complexes between ordered proteins. Most interestingly, the ΔH° values are statistically equal, so all the destabilizing contribution has an entropic origin ($-T\Delta S^\circ > 0$). Also interestingly, isothermal enthalpy-entropy compensation holds in both cases, and also similar interaction surfaces exist in both sets in terms of amino acid composition and size. The authors therefore conclude that loss of conformational entropy by the disordered ligand is the main source of the lower stability of IDP-folded protein complexes, which is noticeable but not large.

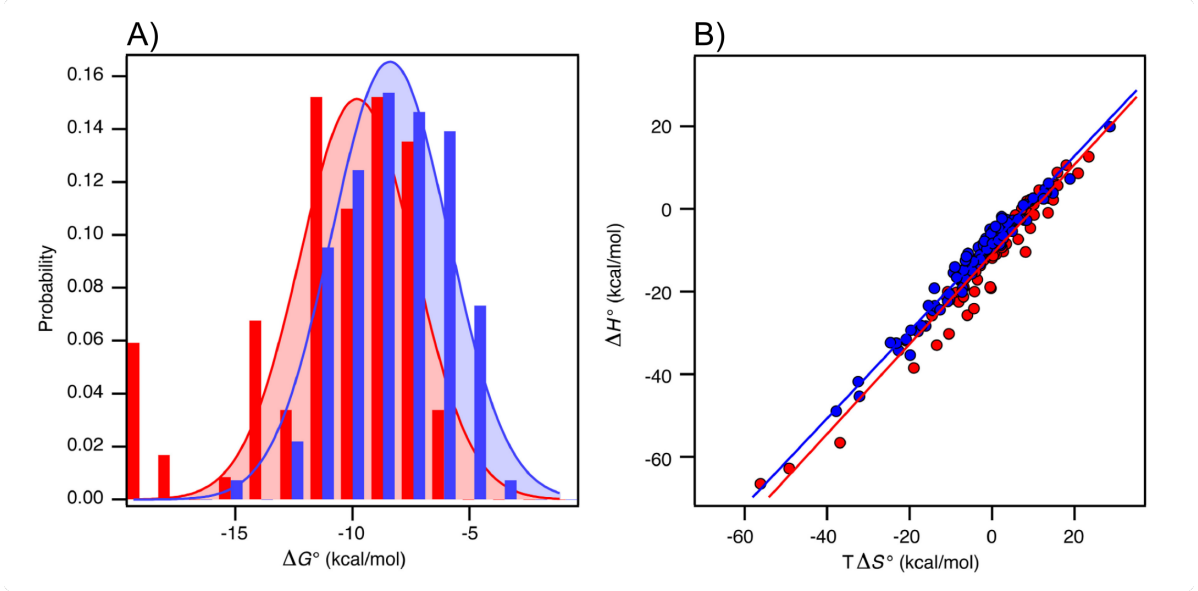


Figure 1.17: A) ΔG° histograms and Gaussian fitting for complexes between ordered proteins with ordered (red) and disordered (blue) partners reported in Teilum et al. (2015); B) ΔH° and $T\Delta S^\circ$ correlation in both cases, with the corresponding linear regressions. Adapted with permission from Teilum et al. (2015).

Regarding kinetics, Motlagh et al. (2014) and Teilum et al. (2015) have observed that whereas similarly fast k_{on} rate constants are observed when IDPs are involved, large k_{off} values are the general trend and the main reason for the transient trend in IDP complexes. Shoemaker et al. (2000) introduced the possibility that structural plasticity may enhance k_{on} rate constants by expanding the sphere of influence of the protein and enable early transient contacts that facilitate recognition (*fly-casting* mechanism), by factors between 1 - 3 with respect to rigid proteins. Zhou et al. (2012) argue that this mechanism has a quite modest importance and structural disorder is rather a form to avoid too low rate

[15]The authors recognize that, although both sets are reasonably large to be significant, the limitation to binary complexes and the fact that allosterism is not considered may imply further differential features of complexes involving intrinsic disorder.

constants due to orientational restraints during partner alignment. Finally, the K_d value statistics in the curated DIBS database of IDR-folded protein complexes (Schad et al. 2017) reveal the previously mentioned variety of binding situations, from sub-nanomolar to milimolar regimes (see figure 1.18).

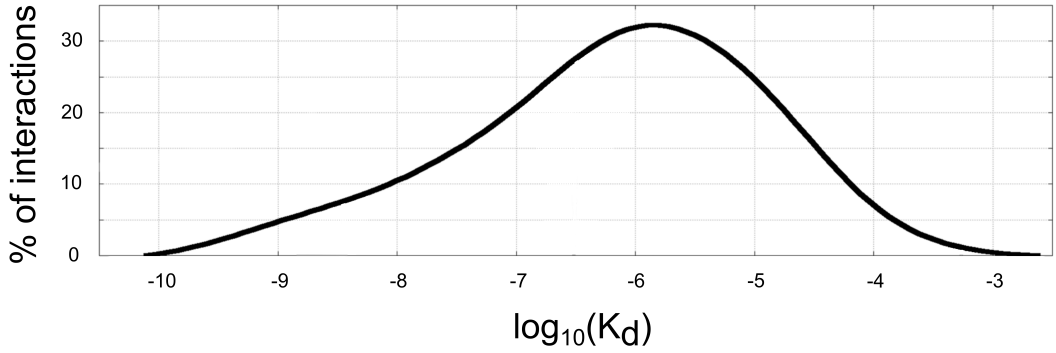


Figure 1.18: Distribution of K_d values for complexes between intrinsically disordered and ordered proteins in the DIBS database. Adapted with permission from Schad et al. (2017).

How ΔG° remains in a negative range that is effective for molecular recognition in spite of the, in principle, unfavorable change in entropy, is one of the most interesting aspects of binding by IDRs, and ultimately integrates dynamics with function (Flock et al. 2014; Dogan et al. 2014). It also contributes to the capacity of disordered systems to establish competent and promiscuous interactions with a large number of partners, even simultaneously, and to rapidly respond to environmental changes by rapid shifts in the statistical ensemble that reconfigure contacts on the fly (Liu & Huang 2014; Berlow et al. 2015).

Based on equation 1, three possible regimes can be envisaged for productive interaction involving IDRs - i.e. $\Delta G^\circ < 0$:

1. $\boldsymbol{\Delta H^\circ \ll 0}$: Maximal enthalpic gain.

   According to the isothermal enthalpy-entropy compensation often observed in biomolecular recognition processes (Wereszczynski & McCammon 2012; critically reviewed in Chodera & Mobley 2013), achieving a large favorable binding enthalpy brings an equally large loss of entropy due to massive reduction of structural freedom degrees. This model would correlate with a strict *folding upon binding* model, in which a large interaction interface is obtained through IDR structuration, as in the example of the bacterial phd:doc toxin/antitoxin system (Garcia-Pino et al. 2010). The enthalpic contribution supersedes the entropic loss due to the IDR

folding that is not compensated by the changes in solvation, and drives binding. Since the final state of the complex is well defined, these kind of complexes involving IDRs were among the first to be described and are in many cases well characterized (Dyson & Wright 2002).

2. $\boldsymbol{-T\Delta S^\circ \ll 0 \implies \Delta S^\circ > 0}$: Maximal entropic gain.

   This situation is less common, and involves a order-disorder transition upon which a dynamic interacting region is gained. It is also very likely that the solvent contribution due to hydration of the novel IDR is a significant contribution to both entropic and enthalpic terms. An example is the partial unfolding of two α helices of the anti-apoptotic BCL-2 on its complex with p53 upon binding to PUMA, with subsequent release of p53 (Follis et al. 2013). Another important factor in this particular example would be the favorable $\Delta S^\circ_{traslational}$ and $\Delta S^\circ_{rotational}$ gained upon detachment of p53.

3. $\boldsymbol{-T\Delta S^\circ \approx 0 \implies \Delta S^\circ \approx 0}$: Minimal entropic loss.

   Given the existing disorder-order continuum and the alternate use Nature makes of regions with different degrees of flexibility, this intermediate regime is likely where most systems lay. Given a favorable, moderate or modest ΔH°, there exist two non exclusive strategies to reduce the entropic cost of the binding: to pre-pay, and not to pay.

   In the first case, the IDR can possess a preexisting residual structuration, such as local secondary structure propensity, or long range contacts acting as a soft scaffold. In the second situation, a similar degree of structural heterogeneity is retained after the recognition event so $S^\circ_{final} \approx S^\circ_{initial}$, and therefore $\Delta S^\circ \approx 0$.

   In both cases, if it is assumed that solvation and IDR rotational freedom degrees will remain in the same order upon partner association, it is the combination of two components of the ΔS° term what ultimately determines the entropic penalty of the binding:

   $$(4)\ \Delta S^\circ = \Delta S^\circ_{conformational} + \Delta S^\circ_{configurational}$$

   where $\Delta S^\circ_{conformational}$ is the entropy change associated to the variation in the **structural heterogeneity of the IDR ensemble**, and $\Delta S^\circ_{configurational}$ reflects the contribution stemming from the different forms in which the IDR and its partner (folded, in general) can associate - in other words, the **structural heterogeneity of the complex interface**. Baxa et al. (2014) have estimated that the entropic conformational penalty in protein folding is three- to fourfold larger for the backbone than for

the side chains. Additionally, the authors state that both components are largely uncoupled, so it is safe to assume that the same holds in the context of IDP binding. For this reason, it can be conceived that, in a simple model, $\Delta S^{\circ}_{conformational}$ upon IDR binding is mostly determined by the backbone dynamics, while $\Delta S^{\circ}_{configurational}$ incorporates an additional contribution arising from the side chains and how they entangle with the partner[16].

Conformational and configurational plasticity associated to molecular recognition is generalized as **fuzziness**, a concept defined by Tompa & Fuxreiter (2008) and Fuxreiter (2012) that I discuss in detail in the following sub-section.

Finally, and in order to illustrate some of the different situations described above, I show the studies by Frederick et al. (2007) on the IDP **calmodulin (CaM)** binding to a variety of high affinity peptides. Figure 1.19.A shows how the different peptides bind in a similar range of ΔG[17], but the ΔH and ΔS values are case-dependent. Since the $\Delta S_{conformational}$ approximately correlates with the total entropic contribution (figure 1.19.B), the authors conclude that CaM is able to use its internal dynamics to tune binding affinity depending on the ligand identity. This gallery of real examples underlines the importance of $\Delta S^{\circ}_{conformational}$ in IDP multi-functionality and plasticity.

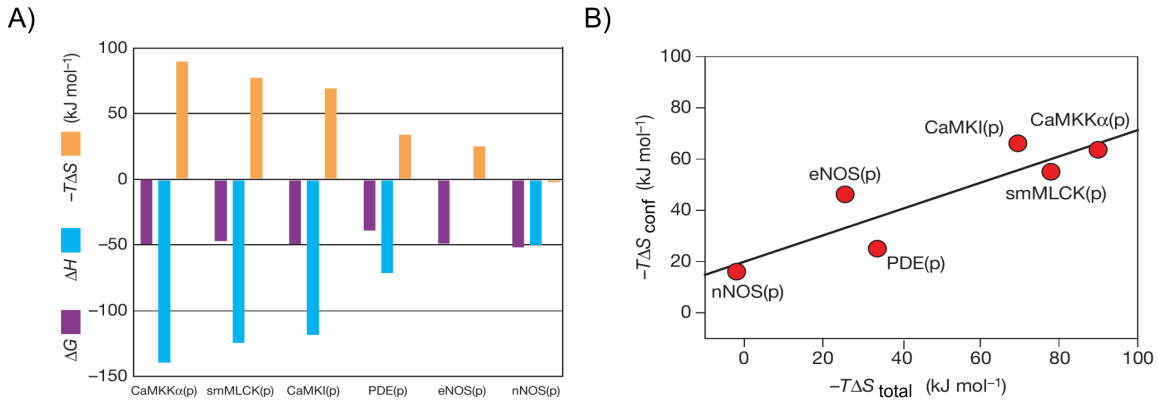


Figure 1.19: A) Calmodulin ITC-derived thermodynamic parameters with different peptides; B) contributions of $\Delta S_{conformational}$ to total ΔS_{total}. Adapted with permission from Frederick et al. (2007).

[16]In the case of multivalent IDPs it is possible that $\Delta S^{\circ}_{configurational}$ depends strongly on backbone dynamics. For example, if competitiveness or any other mechanism that can impose a configurational restrain is to occur, the potential entropic gain from degeneracy in the association will be hampered.

[17]ITC measurements were performed at 35 °C.

1.8 The concept of fuzziness in protein complexes

Lofti Zadeh started the field of *fuzzy logic* in order to mathematically describe sets with ill-defined boundaries (Zadeh 1965). This important kind of objects are best exemplified as the definitions we use in everyday life: our coffee may be just *warm enough*, or a meal be *quite good*. Binary descriptions (0 - 1 in logic terms) are not useful to describe states that incorporate a variable margin for diversity. Hence, Zadeh developed the framework for a new kind of logic based on variables having an associated distribution of truth values between 0 and 1 instead of one of the two binary options `True` or `False`. This formulation goes beyond Boolean logic, allowing the use of *vagueness* for evaluating the truthness of propositions.

Tompa & Fuxreiter (2008) borrowed this term in order to describe structural heterogeneity in protein complexes. Instead of fixed structures as well defined initial and final states, fuzzy complexes consist on an initial conformational ensemble (an IDP/IDR) that establishes contact with a partner and still retains a degree of heterogeneity in the bound state, which determines the functional outcome. Fuzzy complexes are thus in a stochastic regime dictated by dynamics, in contrast with the classic deterministic view based on the structure-function paradigm. The influence of retained disorder on the functional outcome is the essential element that discerns a fuzzy complex from a random connector in non-specific contact with a partner.

In their pioneer work introducing the concept of **protein fuzzy complexes**, Tompa & Fuxreiter (2008) argue that structural disorder in protein-protein interactions involving IDPs represents a continuum, from complexes displaying *static disorder*, in which the bound IDR can adopt only a few alternative conformations, to different degrees of *dynamic disorder*, where a large portion of the conformational space remains available. In further research works and reviews (Fuxreiter 2012; Fuxreiter & Tompa 2012; Sharma et al. 2015; Miskei, Antal, et al. 2017) the authors have expanded the repertoire of available examples of these fuzzy complexes, and also categorized topological and functional classes of them based of the degree of heterogeneity (see figure 1.20) and type of activity they exert (as classified in figure 1.16).

The four reference topological classes[18] are, in order of increasing conformational heterogeneity:

1. Polymorphic complexes

[18]The authors define these classes as reference situations, as most fuzzy complexes may represent intermediate cases in the continuum of structural heterogeneity.

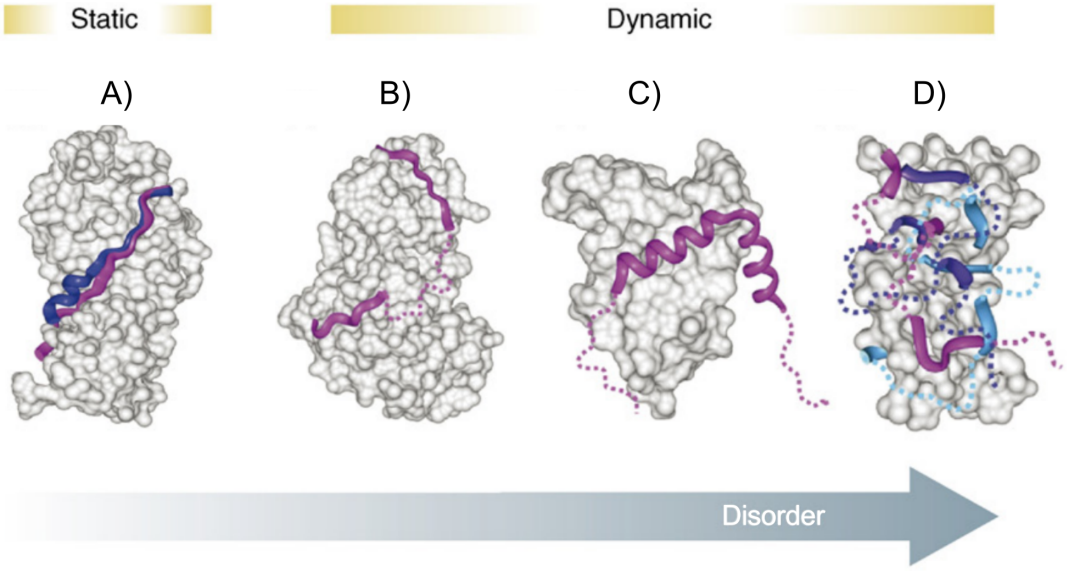


Figure 1.20: Topological classes of fuzzy complexes: A) polymorphic; B) clamp; C) flanking; D) random. Solid ribbons represent well defined bound stretches, whereas dotted ribbons are non-resolved but functional regions. Adapted with permission from Tompa & Fuxreiter (2008).

The IDR can bind its partner through a relatively small variety of stable conformations. These complexes are representative of static disorder.

2. Clamp complexes

   The IDR acts as a linker between two folded domains responsible for the binding, but does not engage with the substrate.

3. Flanking complexes

   Disorder is retained in the margins of the binding site, and at the same time contributes to establish additional contacts.

4. Random complexes

   Interacting short linear motifs are embedded in an IDR.

In the context of the theoretical IDP binding scenarios I outlined in the previous sub-section, a direct correlation can be drawn along these topological categories and $\Delta S^{\circ}_{conformational}$. As structural heterogeneity increases from polymorphic to random complexes, the proportion of binding areas that can provide $\Delta H^{\circ} < 0$ at the cost of conformational entropic cost decreases, while the fraction of flexible regions that can remain dynamic escalates. Because of the dominating structure-centric view of protein:protein interactions is often assumed that largely disordered final states can not

achieve stable complexes, although evidence suggests otherwise as demonstrated by the works of Teilum et al. (2015) or Frederick et al. (2007) above mentioned.

However, the particular features of IDRs can turn the tables and retain significant $\Delta H^\circ < 0$ even in the presence of disorder. Since IDPs are dynamic and often adopt extended conformations, their interacting interfaces have a larger effective area per residue than those of ordered proteins (Gunasekaran et al. 2003). This permits short motifs to establish a variety of interactions through virtually any element of their backbone or side chains. Thus, the loss of the favorable enthalpy associated to a large, rigid interface is compensated by, maybe weaker, but many sparse anchoring elements, specially if the IDR is multivalent.

A favorable enthalpic contribution can be thereby retained by means of binding degeneracy, which directly correlates with $\Delta S^\circ_{configurational}$. The fact that backbone and side chain contributions can be largely uncoupled (Baxa et al. 2014) further reinforces this mechanism. Additionally, with increasing levels of disorder, potential structuration upon association can be restricted to the binding motifs, thus reducing the entropy loss. Furthermore, linear motifs can be locally pre-structured to some extent also minimizing the penalty. Thence, synergistic trade-off between conformational and configurational entropy to retain enthalpy is definitely an emerging property characteristic of fuzzy complexes (Flock et al. 2014), and it is this delicate balance what ensues controlled functionality.

A recent experimental example is the detailed study of the thermodynamics of the fuzzy complex between the C-terminal IDR of antitoxin CcdA, which adopts α helical structure at the time of binding the toxin dimer CcdB (Hadži et al. 2017). The authors perform a series of mutations that affect contacting and non-contacting residues. Their results show that mutations in residues not directly involved in protein:protein interaction reduce the degree of structuration both in the bound and free forms ($\Delta\Delta S^\circ_{conformational} \approx 0$), but also promote alternative isoenergetic configurations ($\Delta\Delta S^\circ_{configurational} > 0$ and $\Delta\Delta H^\circ \approx 0$) thus minimizing the particular ΔG° of the mutant complex.

Finally, the case-specificity of the framework presented so far brings our attention to the details of the very boundary between IDRs and their ordered partners. Olsen et al. (2017) have recently postulated that fuzziness should also be considered at this level, introducing **fuzzy binding**. This refinement and extension of the concept is specially relevant in the context of multivalent IDPs.

In the classic static view, the interfaces between structured proteins are rather rigid, the degrees of freedom mostly limited to side chain motions or restricted backbone wiggling. In the cases of static disorder, alternative binding configurations are available and therefore contribute to a limited extent. However, in many cases of dynamic disorder it is found

that several IDR recognition motifs target different regions of the folded partner and may interact in complex ways between them. Olsen et al. (2017) thus differentiate four types of modes of association within structurally heterogeneous complexes:

- **Two-state binding**

  The IDR and the folded scaffold contain each one a half of a specific couple of interactors, as in the case of c-Src SH3 domain and the PPII motif in the SH2-SH1 linker.

- **Avidity**

  More than one pairs of interacting motifs are split between the IDR and its partner. Since the first encounter between an interacting pair temporarily restricts the meandering of the other, if the length of the linker is appropriate then cooperativity emerges. Clamp fuzzy complexes as that of the scaffold Ste5p with the Fus3p MAP kinase complex (Bhattacharyya 2006) are clear examples.

- **Allovalency**

  In this model, several binding sites (identical or not) in the IDR compete for the same binding region in the structured scaffold. A beautiful example is the complex between yeast proteins Sic (IDR) and Cdc4 (Mittag et al. 2008; Mittag et al. 2010). Sic contains six phosphorylation sites which sub-optimally bind the same region of Cdc4[19].

- **Fuzzy binding**

  This mode relaxes the definition of binding sites in the IDR or the folded partner. Instead of speaking of pairs of well defined motifs, interacting units are broken down to sets of functional groups, or even atoms, that interact with each other as **collections of sub-sites** that serve as weak interactors. Another aspect introduced at this more detailed level is promiscuity: interactors not longer form exclusive pairs. Instead each sub-site may be able to transiently associate with a variety of complementary partners, giving place to a plastic network of interactors which is highly dynamic. The suitability of the term *fuzzy* is utterly justified if we revisit the initial motivations of Zadeh when he developed fuzzy logic: to describe vaguely defined sets with diffuse boundaries.

[19] Additionally, in this example increasing Sic phosphorylation creates binding sites which then interact between them, forming a functional dynamic web driven by electrostatic interactions around the folded domain. This mechanism that emerges from dynamics illustrates the importance of structural plasticity.

In order to summarize and to put in context the results that will be presented in the second part of this thesis, I would like to stress the importance of random fuzzy complexes with fuzzy binding modes. The contribution of $\Delta S^{\circ}_{configurational}$ to protein:protein association is fully exploited in fuzzy binding, where the set of possible permutations between interactors - i.e., complex configurations - is maximized and thus able to potentially generate a large, distributed, favorable enthalpic term.

Next, I will detail some of the difficulties inherent to the characterization of IDPs in general and dynamic, structurally heterogeneous systems as fuzzy complexes, along with an introduction on the approaches and techniques used in this work.

1.9 Tools and methods for the characterization of IDPs

The foremost obstacle for the biophysical characterization of IDPs and fuzzy complexes is ironically the source of their fascinating functional properties: the structural heterogeneity (and conspicuous dynamics) arising from a relatively flat free energy landscape. Since every single IDP molecule in a sample independently probes a wide variety of accessible conformations over time, there is a large number of contributions to empirical measurements. Hence, the experiments used to obtain information on structural and/or dynamic aspects of IDPs provide **time and/or space-averaged observables**, depending on the underlying physical principles of the method used.

An additional problem is that the dynamics that dominate IDPs span along a wide range of time scales, covering from extremely fast side chain motions in the picosecond scale to slower, large extent motions in the order of milliseconds. Therefore, the experimental requirements in order to fully characterize an ensemble are high, in terms of time, know-how, and equipment, plus adding the problem of correct integration of data from different sources and phenomena. The actual amount of experimental information used in studies on disordered systems is, in any case, sparse.

Characterization of IDPs and their free energy landscapes is a vast topic due to the variety of observables and parameters that can be measured and derived in order to describe different structural and dynamic features, and furthermore because of the diversity of techniques available, which ranges from different spectroscopies to hydrodynamic methods. For these reasons, I refer the interested reader to the chapters devoted to experimental methods for IDPs in Tompa (2009), Uversky & Longhi (2010), and the book by Felli & Pierattelli (2015) on the application of NMR to the study of IDPs. The latter is probably the most extensive and varied subtopic in IDP characterization because of the power

and versatility of the technique towards dynamic systems, as explained in the following section.

Thus, instead of an endless enumeration of diverse techniques, observables and parameters, I will classify the type of information we can collect from IDPs by the spatial ambit they report on: **short-range, long-range, or global**.

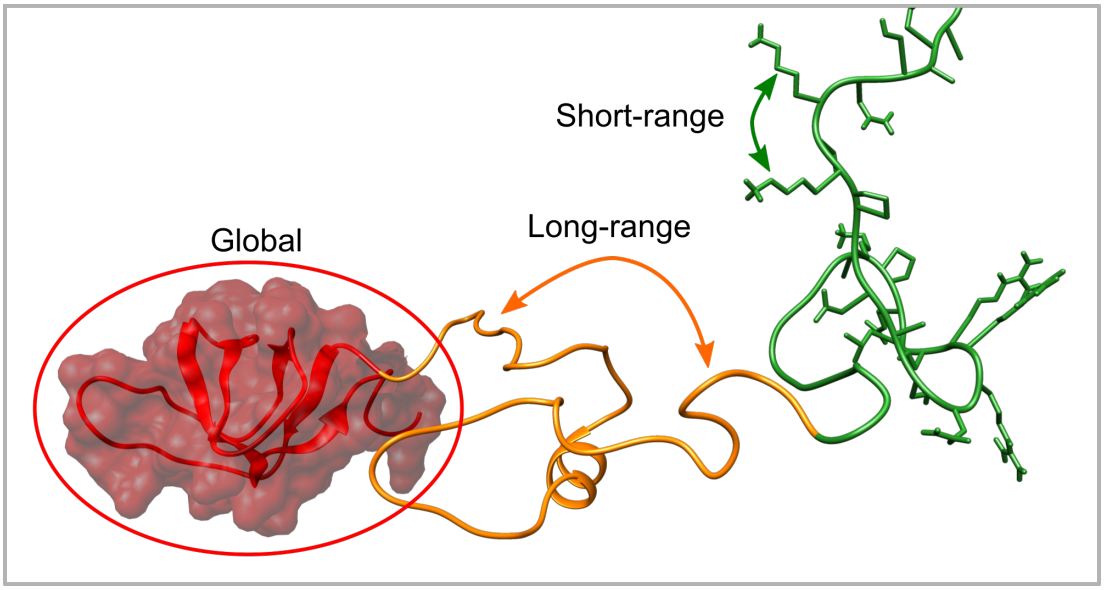


Figure 1.21: Different spatial scales represented over c-Src residues 1-150 (IDR + SH3 domain)

At the most local level we have parameters that reveal close contacts involving specific zones of the IDR. These contacts can be either intramolecular (e.g. secondary structure elements or proximal tertiary structure in ordered proteins) or intermolecular, as in binding processes. As a reference threshold, we can use the maximum range of ~6 Åreported by **Nuclear Overhauser Effect (NOE)** NMR spectroscopy, the main method used to obtain protein structure spatial restrains.

Next in the scale, we may be interested in longer range spatial relationships that report on proximity between regions distant in sequence that approach each other, but do not establish close contact. An example may be that of inter-domain distances. The range here may extend up to 25 - 30 Å.

Finally, it is also interesting to have information on the overall size and/or shape of a protein or a complex. This has special relevance for IDP studies, since IDR ensembles may undergo very significant size variations upon association as, for example, in folding upon binding events, or upon post-translational modification. The range of this scale may span from 10^1 - 10^2 Å.

The reader may have noticed that the above exposed types of information is completely

structure-based, without any reference towards dynamics. Mostly based on NMR techniques, characterization of protein dynamics is a huge field of research by itself. A thorough analysis of the dynamic modes in a protein typically involves a large number of experimental observables at different magnetic fields and, most recently, a range of temperatures (Abyzov et al. 2016). In the case of IDPs, modeling and interpreting the results is an additional issue far from trivial, since a well established, general physical framework is not yet available (Salvi et al. 2017). Hence, due to the fact that the main tools I use along this work are time-averaged and for the sake of clarity and extension, the reader with interest on protein dynamics (IDPs specially) is referred to the many excellent books (Keeler 2010; Levitt 2013), and reviews available (Palmer et al. 1996; Palmer et al. 2001; Salvi et al. 2017).

Another aspect that is important for the study of IDP relaxation (Salvi et al. 2016) among other phenomena and I will only mention here is **Molecular Dynamics (MD)** simulation (Schor et al. 2016; Best 2017). Although the development of novel methods and the ever-increasing computational power keep *in silico* simulation of biomolecules steadily advancing towards more complete studies of more complex systems, IDPs still remain challenging. Fast dynamic modes have been successfully incorporated in IDP MD simulations, slow motions dominating long-range contacts are not amenable due to the long simulation times it would require to cover them (Salvi et al. 2016). Additionally, the enhanced water exposure of IDPs and low energy barriers between conformations create a strong dependency on the force field used, which has been proven critical for accuracy and proper sampling of the conformational space (Rauscher et al. 2015).

1.9.1 Ensemble modeling of IDPs

As it has been already introduced, conformational ensembles are a powerful approach to model the behavior of IDPs. Although ideally capable to capture the features and consequences of conformational heterogeneity, the construction, analysis, and visualization of these tools also poses its own challenges asides from the inherent difficulties of studying highly dynamic systems.

In the first place, constructing ensembles able to capture the vast structural diversity of IDPs is not a trivial problem (Fisher & Stultz 2011). The ensemble consists on an enumeration of conformations with an assigned weight. Average properties are computed from the ensemble as weighted averages. It is obviously impossible to enumerate all accessible conformations of a large IDP, but it has been shown that good predictions of SAXS data can be achieved with random ensembles in the order of 10 000 conformations

or less. Thus, it is generally assumed that the deviations from a random coil of a particular IDP can be modeled as changes in the weights of the populations forming the ensemble.

Deconvoluting averaged information in order to describe the individual components arising from the system heterogeneity is, mathematically speaking, an ill posed inverse problem (Ravera et al. 2016). The amount of experimental information is typically limited, so the problem is heavily underdetermined. Instead, the direct problem - this is, back-calculation of theoretical structural parameters for a synthetic ensemble - is a straightforward question for many observables, such as SAXS scattering curves (Svergun et al. 1995) or chemical shift values (Shen & Bax 2010). Notwithstanding, in many cases the physical model, averaging procedure, and uncertainties or errors must be carefully taken care of.

Two main approaches exist for constructing ensembles that reproduce experimental observations: Maximum Entropy (**ME**) and Maximum Weight (**MW**) (Ravera et al. 2016). Since the possible conformational ensembles that can fit the experimental data - i.e. solutions - are virtually infinite due to indetermination, both strategies differ in the kind of solution they seek. ME methods try to find large collections of conformations that satisfy the empirical boundaries while minimizing the spread of the respective weighting values of each conformer. Instead, MW approaches search for the minimal set of conformations with large statistical weights that can reproduce the experimental observations. In summary, each strategy is convenient depending on the characteristics of the structural heterogeneity and the type of experimental information to be reproduced. For example, strong preferences for particular conformations will be better represented through MW methods, whereas more diffuse subjacent distributions are better fit by ME modeling. Therefore, most practical applications implement a balanced mix of both approaches, being this choice a crucial issue. An updated relation of available methods is provided in table 1.1. In any case, it must be kept in mind that these ensembles are theoretical constructs that recast the empirical experimental information in a different representation frame, which may be read and processed more easily by the human mind. Thus, ensemble *modeling* is more similar to a change of coordinates than to the generation of a structural model for a folded protein.

Table 1.1: Selected list of available methods for structural ensemble determination as of 2017. Adapted with permission from Bonomi et al. (2017).

Approach	Method name	Reference
Maximum Entropy	Maximum entropy restraints	(Pitera & Chodera 2012)
		(Roux & Weare 2013)
	Replica-averaged metadynamics	(Camilloni et al. 2013)
	Maximum entropy restraints	

Approach	Method name	Reference
	for distance histograms	(Roux & Islam 2013)
	Ensemble-biased metadynamics	(Marinelli & Faraldo-Gómez 2015)
	Experiment directed metadynamics	(White et al. 2015)
	EROS	(Różycki et al. 2011)
	COPER	(Leung et al. 2016)
	ENSEMBLE	(Choy & Forman-Kay 2001)
	Bayesian ensemble refinement	(Hummer & Köfinger 2015)
	BE-SAXS	(Antonov et al. 2016)
	EISD	(Brookes & Head-Gordon 2016)
	BELT	(Beauchamp et al. 2014)
	Integrated Bayesian approach	(Xiao et al. 2014)
	Sethi *et al.*	(Sethi et al. 2013)
	Reference ratio method	(Olsson et al. 2013)
	BICePS	(Voelz & Zhou 2014)
	Metainference	(Bonomi et al. 2016)
Maximum	EOM	(Bernadó et al. 2007)
Weight	ASTEROIDS	(Nodet et al. 2009)
	SES	(Berlin et al. 2013)
	SAS	(Chen et al. 2007)
	MaxOcc	(Bertini et al. 2010)
	MES	(Pelikan et al. 2009)
	BSS-SAXS	(Shaw et al. 2010)
	BioEM	(Cossio & Hummer 2013)
	BW	(Fisher et al. 2010)
	Multi-state Bayesian modeling	(Molnar et al. 2014)

1.9.2 Nuclear Magnetic Resonance and Intrinsically Disordered Proteins

As it has already been introduced, solution NMR is arguably the most complete tool to study protein structures and dynamics. The arsenal of NMR experiments that has been developed over decades allows to study their structural features and characterize dynamics in a broad range virtually using the same samples and equipment. Enhanced dynamics provide favorable spectral properties to IDPs in terms of relaxation, thus alleviating the

classic *large size curse* that limits solution NMR[20] and providing, in general, sharp signals. At the same time, it has been mentioned that many other mainstream techniques like X-Ray crystallography are severely limited when facing IDRs. These aspects have made NMR the foremost tool for the characterization of IDPs since the very beginning of the field and, together with different methods of ensemble modeling, exhaustive pictures of IDP structural, energetic, and functional traits can be obtained (Schneider et al. 2012; Jensen et al. 2013; Jensen et al. 2014).

NMR is based on the interaction between active nuclear spins subject to a strong magnetic field with electromagnetic radiation. The coupled spin networks can be characterized by inducing and transferring magnetization through it in a precisely controlled manner, and finally recording how that magnetization decays - the NMR signal. Because energy levels are extremely low, the whole process is non-destructive and can be performed in physiologically meaningful conditions. Recent advances even allow real time observation of isotopically labeled proteins or metabolites in living cells (Luchinat & Banci 2017).

Another important aspect is that NMR provides information at the atomic level, so it permits to study proteins in an extremely selective and complete way, observing specific backbone or side-chain reporters that can provide per-residue data. A selection of relevant NMR observables is illustrated in figure 1.22.

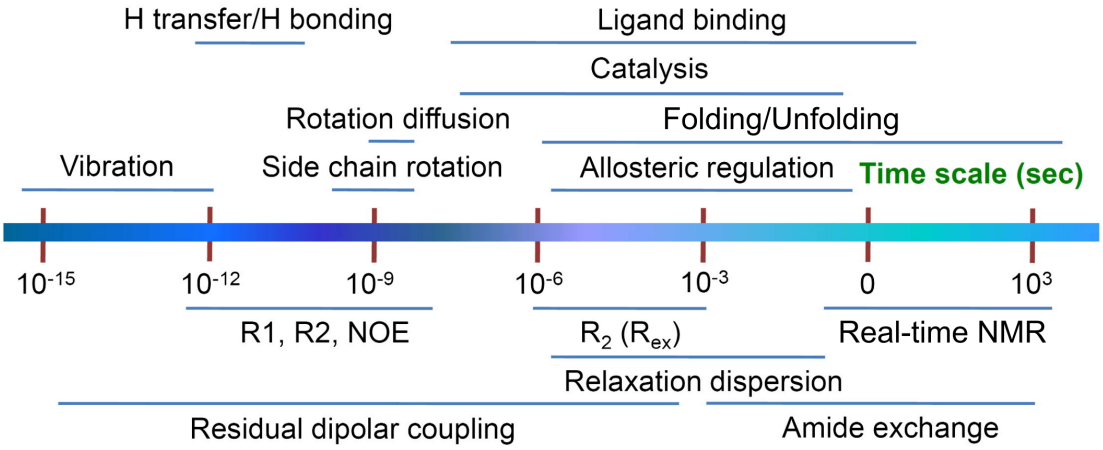


Figure 1.22: Dynamic regimes and timescales. Adapted with permission from Kumar & Balbach (2015).

In the next sub-sections I will briefly introduced the two NMR methods I used during the elaboration of this thesis: **Chemical Shift Perturbation (CSP)** and **Paramagnetic Relaxation Enhancement (PRE)**. As it will be exposed, the combination of these tools fulfills the requirements for fuzzy complex characterization by providing comple-

[20]See the introductory paragraphs on sub-section 1.9.2 for more details.

mentary information on short and long range structural parameters for the whole protein. Introduction of selected mutations and comparison with wild type references permits to withdraw functional information.

1.10 Chemical Shift Perturbation (**CSP**)

The **Chemical Shift (CS, δ)** is an adimensional magnitude directly proportional to the resonance frequency of an NMR active nucleus subject to a fix magnetic field.

$$(5)\ \delta_{nuc} = \frac{\omega_{nuc} - \omega_{ref}}{\omega_{ref}} \cdot 10^6,$$

where δ_{nuc} is the chemical shift of a particular nucleus, ω_{nuc} is its resonance frequency at a given field, and ω_{ref} is the resonance frequency of an arbitrary reference compound.

Beyond this operational definition, the chemical shift is dominated by the nucleus identity through the specific gyromagnetic ratio γ_{nuc}, which determines the Larmor precession frequency $\omega_{0,nuc}$ at a given field B_0 (the magnetic field applied by the spectrometer).

$$(6)\ \omega_{0,nuc} = -\gamma_{nuc} B_0$$

In real molecules, a small, local contribution, B_{loc}, that adds to the main magnetic field B_0 emanates from the surrounding nuclei and electrons, thus making the nucleus chemical shift *unique* in a sense[21]. The spatial and time averaged magnetic properties of the nucleus surroundings is hence termed the *chemical environment*. Some of the most important contributions come from anisotropic electron distribution due to the presence of neighboring heavy atoms with a strong inductive effect (in proteins, mainly nitrogen and oxygen), aromatic rings, which create important magnetic fields due to ring currents, or ionizable nuclei. Since magnetic field dependency with distance is $\frac{1}{r^3}$ and the these local fields are extremely weak, chemical shifts are sensitive only in the **short range** distance regime previously defined.

Although other nuclei like C', $C\alpha$, and $C\beta$ are the best probes to characterize secondary structure (Kragelj et al. 2013; Berjanskii & Wishart 2017), the **backbone amide ^{1}H-^{15}N pairs** are probably the most commonly observed reporters in protein NMR (Jensen et al.

[21]The *uniqueness* of the chemical shift of a particular nucleus in a simple molecule depends on the connectivity and symmetry of the spin system it belongs to. The basic concepts of chemical and magnetic equivalence can be found in NMR manuals like Keeler (2010).

2014) through 2D ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) experiments and related variants. Some advantages of 2D ^{1}H-^{15}N experiments are:

1. All amino acids (except prolines) provide a single signal, thus providing a per-residue fingerprint for the protein given the assignment has been performed.

2. 2D ^{1}H-^{15}N correlation experiments are sensitive (in terms of NMR) and fast to acquire - about 30 minutes are typically enough for a $>$ 100 mM sample - with a reasonable equipment (e.g. a 600 MHz spectrometer with a cryoprobe, as it is our case). More recently, fast pulsing and sparse sampling techniques allow for even lower sample concentration requirements or lower acquisition times (Solyom et al. (2013); Kazimierczuk & Orekhov (2015); see sub-section 2.4.4 in the results chapter for more details)

3. Amide backbone H-N pairs are sensitive to primary (amino acid identity and sequence), secondary, and tertiary protein structure levels or intermolecular contacts if they are close enough. Engagement in H-bonds by the H-N pair itself or nearby groups (e.g. backbone carbonyls, sidechains, etc.), or preferential side chain conformations greatly affect the chemical shifts.

4. In IDPs, the lack of stable secondary structure exposes amide protons to the bulk aqueous solution (with which they can be exchanged), and also to the dynamic network of electrostatic interactions that can be established between nearby or distant residues (Pujato et al. 2005). These factors make H-N pairs sensitive to environmental variables that can modify the ensemble of interactions, such as pH or ionic strength.

The **Chemical Shift Perturbation (CSP)** method is the study of how the chemical shift of the signals evolve upon controlled changes in experimental conditions, shall they be sequence mutations, addition of a ligand, or changes in the physicochemical properties of the medium. In order to do so, a reference spectrum - let it be the wild type form of the protein under study, in some standard reference conditions - is recorded, and the cross-peaks are selected and labeled with the corresponding residue identity obtained from assignment. Then, an otherwise identical sample in which a single experimental has been precisely modified is prepared, and the corresponding correlation spectra is acquired and assigned. In the case of 2D ^{1}H-^{15}N correlation experiments, the changes in both ^{1}H and ^{15}N chemical shifts are merged as an empirically weighted (hence the 0.2 factor for $\Delta\delta_N$ in equation 7) Cartesian distance between the initial signal coordinates and the new peak position:

(7) $CSP = \sqrt{\frac{1}{2}(\Delta\delta_H^2 + 0.2 \cdot \Delta\delta_N^2)}$

This methodology can be applied to study the effect of discrete changes, such as mutations, or to monitor and fit how modification of a continuous variable such as ligand concentration or pH affects the CSP (a classic titration) (Williamson 2013). Plotting the calculated CSP versus the protein sequence allows an almost per-residue quantitative mapping of the effects induced.

There also exist some drawbacks to take into account when acquiring 2D ^{1}H-^{15}N correlation spectra on IDPs. The most recognizable signature of protein intrinsic disorder in these experiments is the low amide ^{1}H chemical shift dispersion (7.5 - 8.5 ppm, whereas the typical range is 7.0 - 10.0 ppm for folded proteins) due to the homogeneity of their chemical environment in absence of secondary structure. The Unique domain of c-Src is a perfect example of this behavior (see figure 1.23).

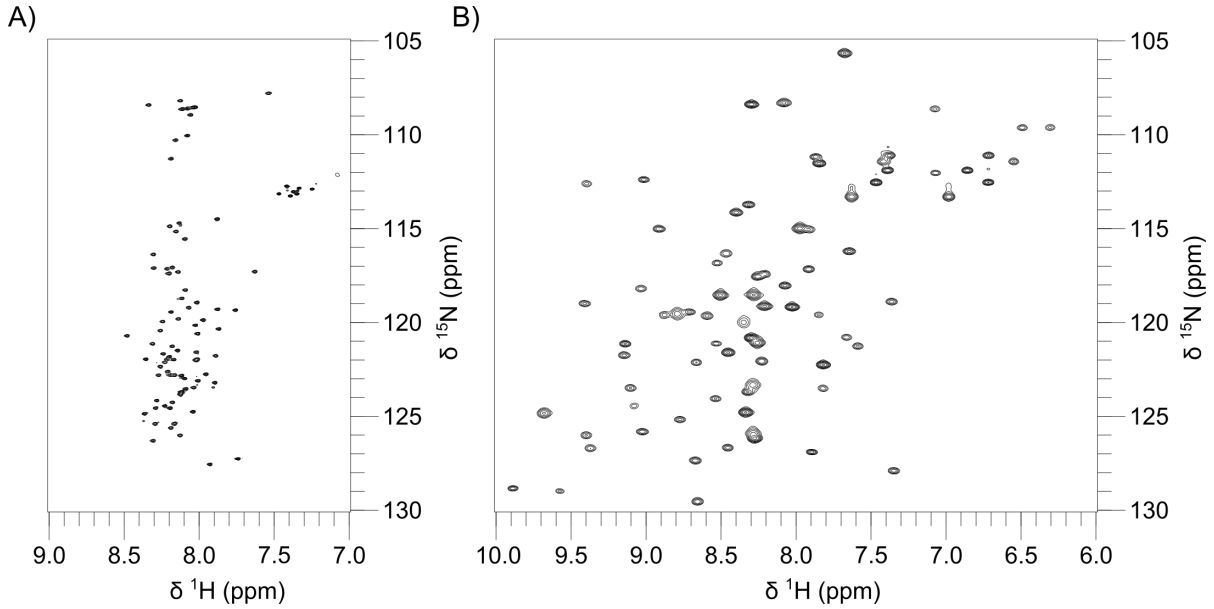


Figure 1.23: $^{1}H - ^{15}N$ HSQC spectra of A) the N-terminal IDR of c-Src formed by the SH4 and Unique domains (acquired at 278 K), and B) the ordered c-Src SH3 domain (acquired at 298 K).

Since spectral dispersion is low, and IDRs tend to adopt low complexity amino acid usage statistics, it is often found that signal overlap is a problem, specially for long sequences. The use of NMR experiments that provide sharper signals and best resolution can help mitigate this issue, but in extreme cases, selective labeling or higher dimensionality experiments may be required. Another issue that relates with the lack of stable structure and subsequent solvent exposure is that amide protons can easily exchange with the bulk water, leading to signal loss. This problem can be circumvented by slowing down exchange using a lower the sample temperature, or moderately acidic media.

Sample preparation is a critical step, specially when performing CSP studies on IDPs. The responsive nature of disordered proteins to small changes of any kind, such as sample concentration, temperature, pH, or ionic strength, and the fact that very often IDPs are sensitive to degradation makes the sample preparation process very delicate. Therefore, robust protocols, careful manipulation, and strict sample control quality are mandatory in order to avoid undesired interference.

Finally, it should be stressed that, being the CSP a time and space averaged parameter that integrates multiple contributions in terms of source of the observed chemical shifts, its interpretation may not be trivial in IDPs. Local modifications such as sequence mutations will surely provoke important CSP in its closest vicinity, and maybe also in residues further away in sequence. Not only that, but a subset of transient interactions within the protein may be reconfigured, so exaggerating some interactions or masking others. It is thus important to keep in mind the ensemble view of the dynamic nature of IDRs and not to assume a deterministic view on the interpretation of CSP: causality does not always apply straightforwardly.

1.11 Paramagnetic Relaxation Enhancement (**PRE**)

1.11.1 Brief introduction to paramagnetic NMR

The use of paramagnetic species as spin relaxation sources for biomolecular NMR studies has long provided an important set of tools for the structural biologist (Bertini & Luchinat 1986). Paramagnetism originates from the presence of unpaired electrons[22] and, because of their strong magnetic moment (almost 10^3 larger than that of proton), their distance-dependent dipolar effects have a sensibly larger reach in space (up to 35 Å) than those found in diamagnetic molecules, like NOE. Additionally, while NOEs provide only one-to-few sets of inter-proton distances, paramagnetic methods report information for all observed signals at once referenced to a selected position where the paramagnetic is introduced. This is also an advantage over other long range methods like FRET (Förster Resonance Energy Transfer) or Electron Paramagnetic Resonance (EPR), where single distances between fixed pairs of probes are measured. These characteristics have thus made paramagnetic NMR a powerful complementary approach to study long range contacts, specially for IDPs (Schneider et al. 2012).

[22]By definition, a paramagnetic system contains at least one unpaired electron. In contrast, diamagnetic systems have all their electrons paired. For simplicity, here I identify paramagnetism with the presence of unpaired electrons, disregarding other situations.

Three main paramagnetic NMR methods have been developed over the years: paramagnetic **Residual Dipolar Coupling (pRDC)**, **Pseudo Contact Shift (PCS)**, and **Paramagnetic Relaxation Enhancement (PRE)**. Among these, PRE has proven to be a particularly useful and easy to implement method to probe long range transient interactions, and minor states in protein ensembles (Otting 2010) for practical reasons. In short, PRE induces signal broadening and therefore a decrease in intensity that is distance dependent. More details are given later.

1.11.2 Paramagnetic spin labels

Paramagnetic species can be classified in two types, according to their magnetic susceptibility tensor, χ: isotropic or anisotropic - i.e., their magnetic moment can be constant, or vary depending on its orientation relative to the applied magnetic field. While all paramagnetic centers induce PRE due to dipolar interaction between the electron and the nuclear spins, only anisotropic ones give place to RDC and PCS because of their orientation dependency. Thus, although anisotropic spin labels can potentially provide more geometrical information, they are also more complex to describe. Additionally, the large signal shifts due to PCS very often make re-assignment necessary, a process that may be specially difficult due to signal loss by PRE. For these reasons, isotropic spin labels are typically preferred to exclusively measure PRE (Clore & Iwahara 2009; Otting 2010).

Although initially restricted to metalloproteins natively able to coordinate metal ions, paramagnetic NMR methods were then further expanded through the use paramagnetic moieties (also termed spin labels) that harbor unpaired electrons and can be selectively attached to proteins. Paramagnetic agents need to accomplish two main requisites:

1. Electrochemically stable in standard experimental conditions (nearly neutral aqueous solutions, etc.).
2. Easily transformable into a diamagnetic equivalent (a diamagnetic reference is always mandatory, as it is later explained).

There exist two types of spin labels: chelators that bind paramagnetic ions (mainly Mn^{2+},Fe^{2+},Co^{2+},Ni^{2+},Cu^{2+}, and most of the lanthanide ions), or nitroxide derivatives (see figure 1.24). From all these, only Mn^{2+}, Gd^{3+}, and nitroxide spin labels are isotropic spin labels. Chelators exist in a variety of flavors depending on the metal they coordinate, rigidity, number and type of linker groups to bind the protein, etc. However,they tend to be bulkier than nitroxides and require replacement of the metal ion for a diamagnetic substitute of similar size in order to obtain a diamagnetic reference to compare

with. Nitroxide spin labels instead can be readily reduced to their corresponding diamagnetic hidroxylamines in mild conditions without significantly disturbing the protein or the medium. An ample variety of nitroxides is available thanks to their use in EPR, and many are cheap and commercially available.

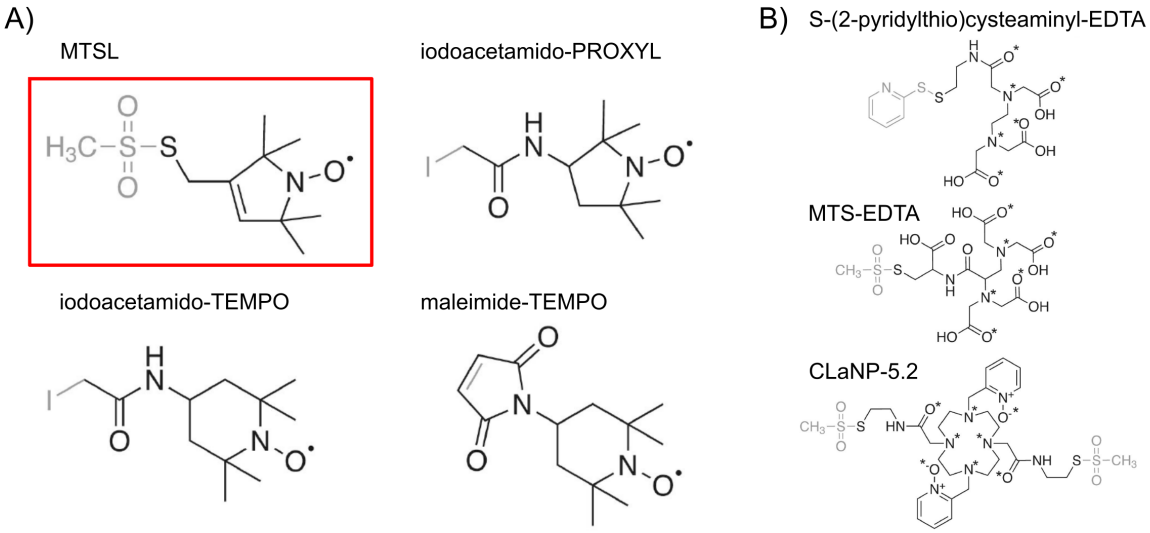


Figure 1.24: Examples of A) nitroxide and B) metal chelating spin labels. Leaving groups for coupling are depicted in grey. Adapted with permission from Clore & Iwahara (2009).

The most common approach for protein functionalization with spin labels is the formation of disulfide bonds with cysteines, either pre-existing or introduced *ad hoc* via site directed mutagenesis (although alternatives exist to target lysine, histidine, tyrosine, or methionine side chains, Kosen (1989)). For this reason, the common scaffold of nitroxide spin labels consists on: a substituted[23] 5 or 6-member ring bearing the paramagnetic N-O moiety, a flexible linker that does not restrict mobility, and a sulfur atom with a good leaving group as methanetiosulfonate for S-S bonding. During this thesis, I have used **MTSL** (*S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate, highlighted in figure 1.24), a commercial spin label commonly used in protein NMR.

Practical use of spin labels is relatively simple (Kocherginsky & Swartz 1995). Given a protein sample containing one single cysteine (which, in the case of an IDP will be accessible), the first step is label attachment. MTSL readily reacts with free thiol groups as shown in figure and the adduct is stable in absence of strong reducing agents.

After acquiring the spectrum of the paramagnetic sample, MTSL is reduced to the diamagnetic hydroxylamine to record the reference spectrum. Ascorbate is a mild reducing

[23]Ring functionalization ensures radical stability by providing steric hindrance, and can also improve water solubility of the spin label.

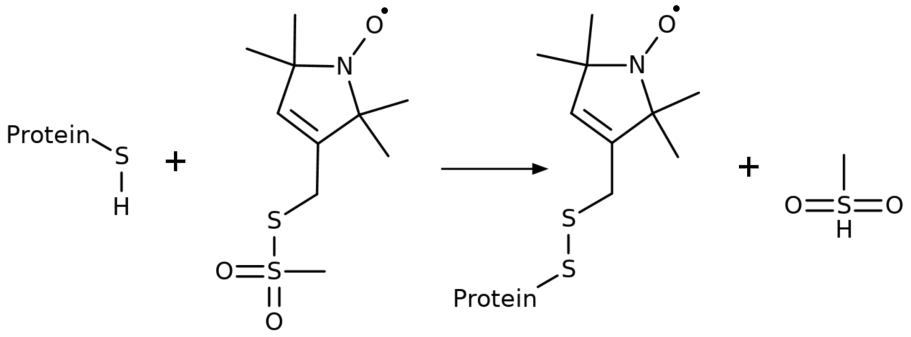


Figure 1.25: Spin labeling of a cysteine side chain with MTSL.

agent that can carry out the reaction without cleavage of the disulfide bond and consequent label detachment. The radicalary mechanism at neutral pH is shown in figure 1.26.

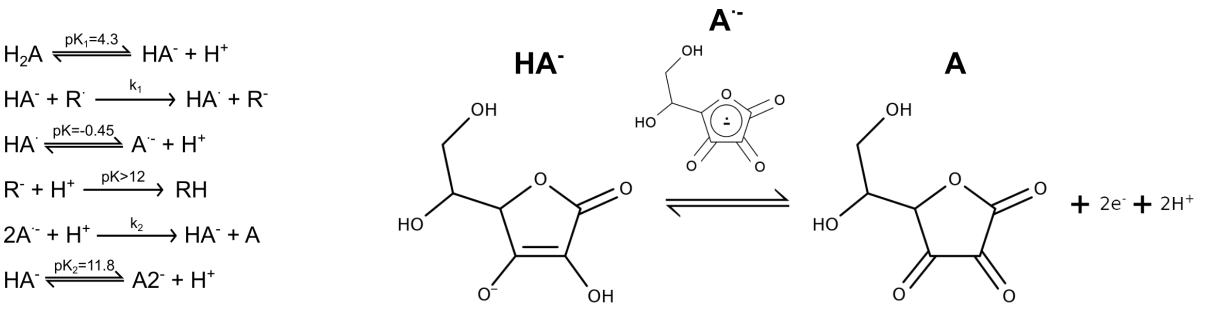


Figure 1.26: Equation of nitroxide reduction by ascorbic acid. A simplified version showing the reducing activity of ascorbate is also shown with the structures of the relevant species (HA^-: ascorbate anion; $HA^{\cdot -}$: ascorbate radical; A: dehydroascorbic acid)

1.11.3 PRE theory and application in short

For the sake of simplicity, here I will only introduce some of the relevant concepts to understand and interpret PRE in a semi-quantitative way. The reasons are latter discussed, but the reader interested in further insight can find excellent reviews on the theoretical aspects of PRE and their use on IDPs in Clore & Iwahara (2009), Gillespie & Shortle (1997a), and Gillespie & Shortle (1997b).

The local magnetic fields generated by a paramagnetic center in a molecule exert an important effect on the longitudinal and transverse relaxation rates (R_1 and R_2, respectively) of nearby NMR active spins via dipolar interactions. The Solomon - Bloembergen formalism used to describe spin relaxation rates (Solomon 1955), later extended by Bloembergen & Morgan (1961) to systems including effect of electron spin.

For practical and theoretical reasons detailed by Iwahara et al. (2007), the most common

approach to determine PRE in proteins is to measure the ^{1}H transverse relaxation rates, R_2, for the amide H-N pairs in uniformly ^{15}N labeled samples. When one measures the $R_{2,para}$ relaxation rates on a paramagnetic sample, the observed value contains both the native diamagnetic term $R_{2,dia}$, plus an additional paramagnetic contribution, Γ_2.

$$(8)\ \Gamma_2 = R_{2,para} - R_{2,dia}$$

Hence the reason for the need of a diamagnetic reference. Once the diamagnetic term is canceled out by subtracting the reference, the purely paramagnetic contribution[24] of a nitroxide spin label can be simplified (Kosen 1989; Battiste & Wagner 2000) to:

$$(9)\ \Gamma_2 = \frac{K}{r^{-6}}\left(4\tau_c + \frac{4\tau_c}{1+\omega_H^2\tau_c^2}\right),$$

where K is a parameter integrating system-specific physical constants, r^{-6} is the proton-electron distance, τ_c is the correlation time for the dipolar interaction, and ω_H is the proton Larmor frequency.

Noticeably, it can be observed that the distance appears as a single value. This is actually imprecise because of the spin label mobility. For this reason, the Solomon - Bloembergen formalism was complemented with the *model-free* approximation of Lipari & Szabo (1982) (not shown here) to account for the internal mobility of the paramagnetic center provide the theoretical basis to quantitatively interpret PREs and extract distance restraints. Therefore, an averaged $\langle r^{-6}\rangle$ is obtained.

When dealing with IDPs - recall the ensemble model - the $\langle r^{-6}\rangle$ spatial averaging is extended from the unpaired electron position to the positions of all affected nuclear spins, which makes modeling more complex (Ganguly & Chen 2009). Also, the exponential dependency redounds in relevant uncertainties in the distance constrains derived from the experimental error in R_2 measurements. In addition, this dependency tends to overestimate the weight of compact conformations with large contributions. An only-PRE based analysis is therefore a complex task that requires a lot of time and experimental work, since some authors have pointed out that datasets of PRE every ~15 amino acids are needed for reliable modeling (Silvestre-Ryan et al. 2013).

For these motifs, a simplified, semi-quantitative approach based on 2D ^{1}H-^{15}N correlation experiments is often used in order to undertake studies on IDPs focused on function rather

[24]For isotropic spin labels with long electron spin correlation times like nitroxides, the Curie relaxation mechanism is negligible and paramagnetic relaxation is dominated by the dipole-dipole Solomon mechanism described here.

than structurally intensive (Battiste & Wagner 2000). Since R_2 relates directly to signal line width and therefore inversely to peak intensity, it can be assumed that the ratio between peak intensities in paramagnetic and diamagnetic samples satisfactorily depicts long range contacts between distant regions.

$$\frac{I_{para}}{I_{dia}} = \frac{R_{2,dia}\exp(-\Gamma_2 t)}{R_{2,dia} + \Gamma_2}, \tag{10}$$

where I_{para} and I_{para} are the respective peak intensities of the paramagnetic sample and the corresponding diamagnetic reference, and t the relaxation time, that includes the acquisition time and additional delays in which transverse magnetization is present. If the spin label positions are intelligently chosen based on other structural or functional observations, a few PRE experiments can provide a useful image of the long range contacts present in the protein.

Objectives

Previous results from the research group had demonstrated the functional importance of the intrinsically disordered region of c-Src and suggested a connection between this region and the neighbor SH3 domain.

The objectives of this thesis are:

1. Characterization of the intrinsically disordered SH4 and Unique domains in the context of the SH3 domain to assess its potential scaffolding role and to identify conformational features in the disordered domains potentially affecting the interaction between folded and unfolded regions in c-Src.

2. Characterization of disorder in the interface between ordered and disordered regions in c-Src.

3. Identification and characterization of sequence elements in the disordered regions of c-Src potentially encoding conformational features.

4. Exploring the generality of the interactions between the SH3 domain and its preceding disordered regions in other members of the Src Family of Kinases.

5. Setting up the conditions for the future structural characterization by solid-state NMR of the SH4-Unique and SH3 domains anchored to lipid membranes.

Chapter 2

Results

The following figure 2.1 represents the sequence and secondary structure of the c-Src region studied in this thesis. It also contains a comprehensive list of the wild type and derived constructs used. I present it here so it can serve as a guide for the reader along this chapter.

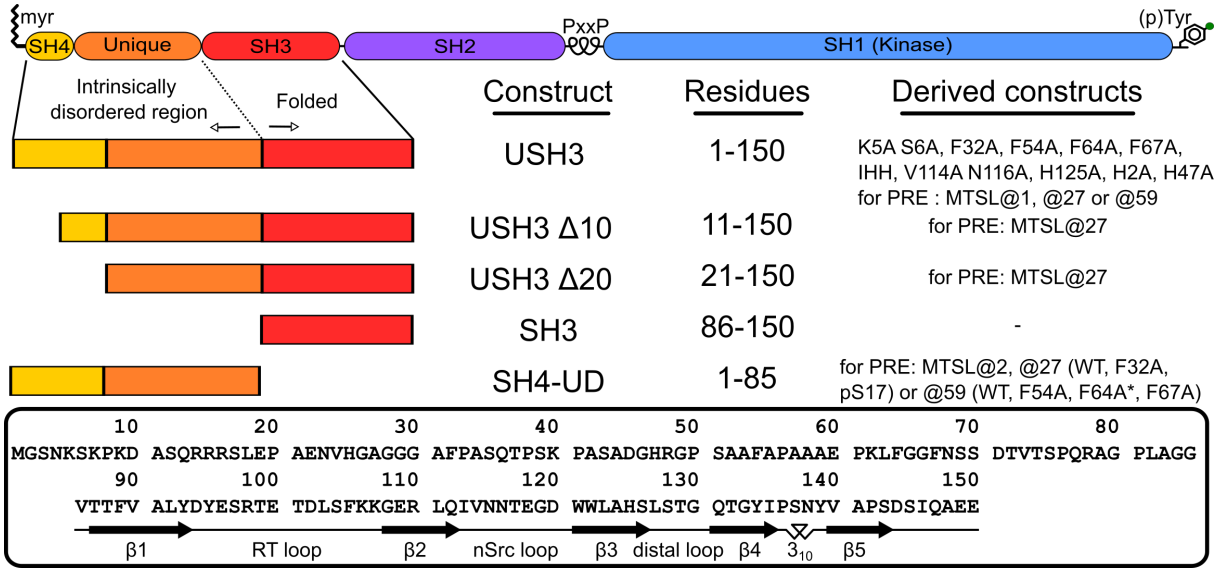


Figure 2.1: c-Src construct guide.

Also to provide a useful reference, all CSP and PRE plots contain a string at the upper edge of the y axis in which secondary structure is represented as text, according to the following convention:

- -: Random coil or intrinsically disordered.
- /: α helix.
- =: β strand.

- ^: 3_{10} helix

Additionally, prolines are indicated as asterisks near the baseline, to difference these positions from unassigned or lost signals.

Finally, I would like to note that all NMR spectra for IDRs (SH4-UD constructs and SH4 and Unique domains of USH3 constructs) were acquired at 278K, while spectra of the folded SH3 domain, either isolated or in USH3 forms, were acquired at 298K. Thus, some CSP and PRE plots for the same molecule presented here are split. The technical reasons for the use of different temperatures are explained in sub-section 2.4.4.

2.1 Scaffolding of the Intrinsically Disordered Region induced by the SH3 domain

2.1.1 Context

The interaction of the intrinsically disordered SH4 and Unique domains and the adjacent, folded SH3 domain had been previously reported by our group using PRE and CSP NMR experiments (Pérez et al. 2009; Pérez et al. 2013; Maffei 2015; Maffei et al. 2015). Using a construct containing c-Src **residues 1 - 150**, which comprise the SH4, Unique and SH3 domains (from now on **USH3**) with a cysteine residue inserted in position 59 and an attached MTSL paramagnetic tag (see section 1.11 in the introduction for details), Pérez et al. (2013) observed both intra- and inter-domain long range contacts.

The SH3 domain signals report long range contacts, most importantly in regions 95 - 105 (RT loop), 112 - 127 (nSrc loop and β3 strand) and 133 - 142 (β4 and 3^{10} helix). Intramolecular interactions also take place within the disordered region. Residues 13 - 22, 30 - 36, 42 - 49 and 73 - 79 are most affected by the relaxation induced by the MTSL at residue 59, as indicated by the decreased intensity ratios (I_{para}/I_{dia}). Comparison with the theoretical curve of the unbiased model generated with Flexible Meccano (Ozenne et al. 2012) denotes a departure from the random coil reference. A construct containing only the disordered domains, termed **SH4-UD (residues 1 - 85)**, with the paramagnetic probe in the same position A59C, shows an almost identical PRE profile.

These results suggest that the transient contacts from position 59 within the IDR observed in USH3 are mostly preserved in absence of the SH3 domain, and therefore there is a notable degree of pre-organization. This is consistent with the positive Residual Dipolar Couplings displayed by residues 63 - 70, which encompass the core of the Unique Lipid

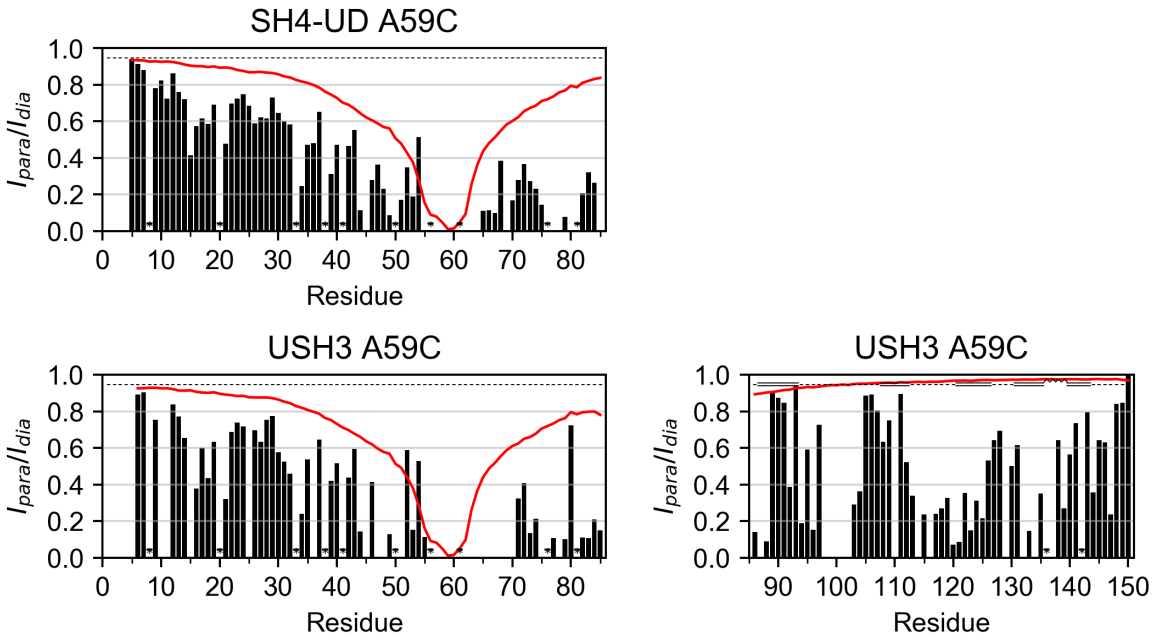


Figure 2.2: PRE of USH3 and SH4-UD A59C (Pérez et al. 2009; Pérez et al. 2013) (black bars). The red lines represent the theoretical random coil profile as a reference (see sub-section 2.2.2).

Binding Region (Pérez et al. 2009). This region had also been theoretically predicted to have a small extent of residual structure using different algorithms.

The CSP of wild type USH3 versus truncated constructs comprising either the isolated IDR (SH4-UD) or only the SH3 domain (residues 86 - 150) were calculated in order to detect short range interactions between them (Pérez et al. 2013; Maffei et al. 2015) (figure 2.3). Besides the obvious perturbation in the hinge region between the Unique and SH3 domains, residues S6, R15, and R16 in the SH4 domain and T37, D45, A55, A57, E60, G65, N68, ^{72}TVT^{74}, and G80 in the Unique domain show significant CSP in the IDR. In the SH3 domain, the largest changes correspond to Y95, S97, R98, E100, D102, and L103 in the RT loop; R110 in β2 strand; ^{113}IVNN^{116}, E118 in the nSrc loop; W121 and H125 in β3; T132 and Y134 in β4 ; S137, and N138 in the 3^{10} helix, and S145, I146 and A148 at the C-terminal tail. These results are coherent with the observed PRE contacts and evidence close interaction between the IDR and the globular domain.

SH3 binding to a synthetic, high affinity peptide (VSL12: *Ac-VSLARRPLPPLP-OH*) (Rickles et al. 1995) containing a canonical *PxxP* recognition motif had been reported to allosterically modulate the SH3-Unique interaction, preventing most of the contacts (Pérez et al. 2013). CSP of the isolated IDR and SH3 versus the full-length construct, all in presence of the PxxP peptide, indicated that only the residues in the SH4 domain and H25, H47 and their respective neighbors in the Unique domain, and in the nSrc loop remain perturbed upon binding (figure 2.3). This suggests that even in the absence of

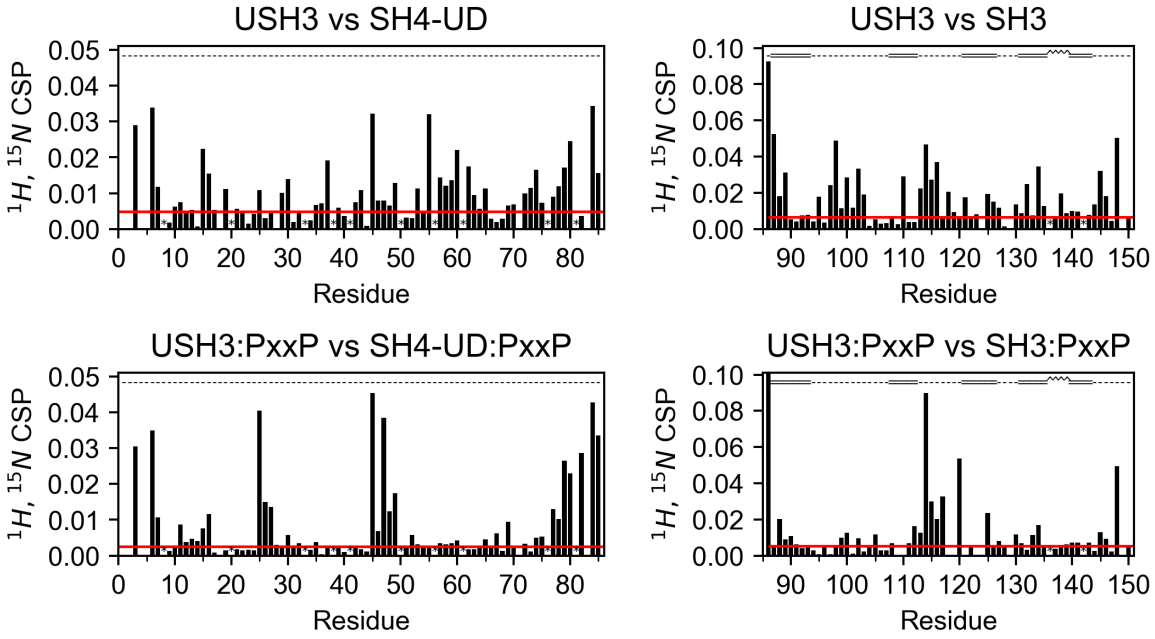


Figure 2.3: CSP of USH3 WT vs the isolated SH4-UD and SH3 domains alone (top) and in presence of the PxxP peptide (bottom). The red line represents a significance threshold defined in Methods and Materials.

the Unique:SH3 interactions the SH4 domain still interacts with SH3 via nSrc[1]. It also follows that the SH3 RT loop is the main responsible for most contacts with the Unique domain in the *apo* form. The fact that the interactions of the SH4 and Unique domain are differently affected by VLS12 indicate that they contain separate binding regions and raises the question of their mutual interplay.

Based on these results and considering the highly dynamic nature of intrinsically disordered systems and their particular properties, we can imagine two extreme models. In the first, the Unique domain conformational space would be heavily biased, being the main force driving the SH4:SH3 interaction even in the absence of direct Unique:RT loop contacts. Alternatively, a specific SH4:SH3 interaction could be the major factor, forcing the Unique domain into constrained conformations that approach the RT loop. The real scenario may however be an intermediate situation, or even more complex if there were other unrecognized aspects - e.g., other interactions, competition or cooperativity, etc. In order to gain insight, I studied the individual contributions of each disordered domain (SH4 and Unique) by removing or mutating them individually, using the SH3 domain as a reporter for CSP.

[1]The chemical shifts of histidines are highly environment-sensitive and often display non systematic perturbations, even between identical samples. They are further studied in sub-section 2.3.5.

2.1.2 The isolated SH4 domain interacts with multiple sites of the SH3 domain

I first used CSP to map short range interactions between the SH4 and SH3 domains in absence of the linking Unique domain (figure 2.4). ^{1}H-^{15}N HSQC spectra of the isolated SH3 domain were acquired both alone and in the presence of a synthetic peptide with the sequence of the SH4 domain (residues 2 - 19). Because the interaction is weak, up to 10-fold peptide to protein excess was added. The plot of CSP per residue (figure 2.4) revealed that, in the absence of the Unique domain, addition of the isolated SH4 domain strongly perturbed the RT loop (^{98}RTETDL^{103}). The distal loop and adjacent residues (H125, G126, G130, T132, and Y134) were perturbed to a lesser extent and the effect over the nSrc loop was subtler (Q112, N116, and E118). The experiment was then performed in the presence of the PxxP peptide. The RT loop perturbation disappeared in the presence of the ligand, while the rest of minor interactions were mostly retained.

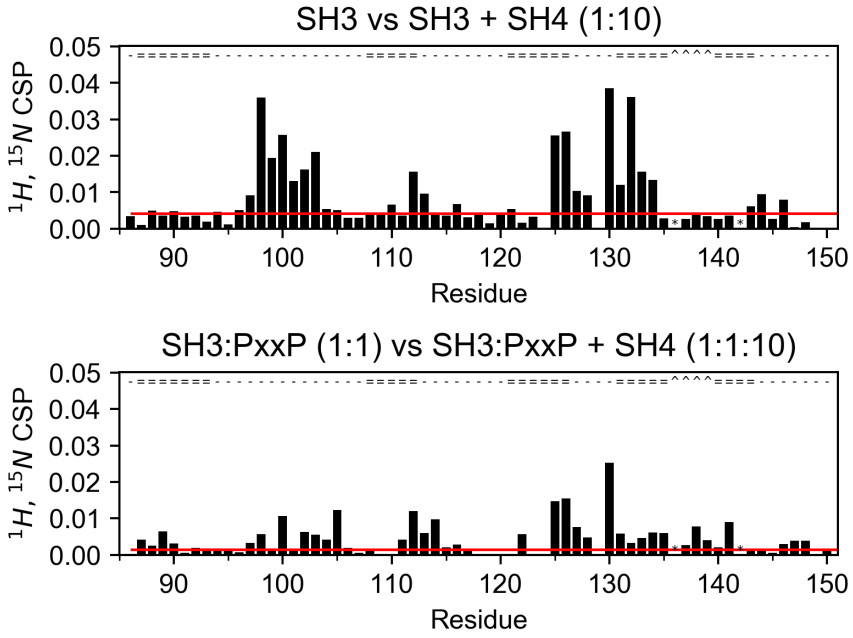


Figure 2.4: CSP of SH3 upon addition of 1:10 excess SH4 peptide, both in the *apo* (top) and PxxP ligand bound (bottom) forms. The red line represents a significance threshold defined in Methods and Materials.

Compared with the CSP of USH3 vs the isolated SH3 domain (figure 2.3), it can be observed that perturbations induced by SH4 alone are similar for the RT (except for E100 and D102, comparatively less affected) and distal loops, but smaller for the nSrc[2]. Also,

[2]Of course, the different entropic contributions and SH3:SH4 concentrations impede direct comparison of CSP magnitudes with the USH3 vs SH3 CSP. However, the relative intensities of the changes in the different regions can be contrasted.

as in full-length USH3, addition of the PxxP peptide to the SH3-SH4 mixture abolished the interactions involving the RT loop while retaining the rest.

These results show that the isolated SH4 domain preferentially interacts with the SH3 domain RT loop and less intensely with the distal and nSrc loops. Binding between both domains is in any case weak (mind the 10-fold excess of SH4 used). The fact that all three loops are contacted by SH4 precludes from establishing a specific correspondence between those sub-sites and the interacting elements along the IDR. Instead, the CSP between isolated and bound domains described in the previous chapter stem from a complex combination of SH4 and Unique promiscuous contacts with overlapping zones of the SH3 domain. If so, potential competition or cooperativity between the interactions of the SH4 and Unique with the SH3 domain or alternative contacts, even within the IDR itself, should be considered as well. Next, I analyzed the effect of removing or mutating the SH4 domain in the context of USH3, with the Unique domain present.

2.1.3 Absence or mutation of the SH4 domain reveals a complex scenario

In order to better understand the contribution of each domain, I studied three USH3 constructs in which the SH4 domain was modified. In the first mutant, USH3 Δ20 (21 - 150), the whole SH4 domain was removed, along with S17, which is known to abolish SH4 lipid binding when phosphorylated by PKA (Pérez et al. 2013). Therefore, only the Unique:SH3 contacts were left. In the second one, USH3 Δ10 (11 - 150), only the first ten residues of SH4 were removed, so the second positively charged patch (^{14}RRR^{16}) and S17 were conserved. Finally in the third mutant only residue S6, which remains perturbed when USH3 is bound to the PxxP peptide, and the neighboring K5, were replaced by alanine. This mutation partially disrupts the first polybasic cluster in SH4, ^{5}KSKPK^{9}.

First, CSP versus the wild type, full-length USH3 construct were calculated for all SH4-mutated forms (figure 2.5).

Regarding the Unique domain and apart from the trivial CSP in residues close to the truncation points, changes were small in all cases. Most remarkably, the Unique domain residues reporting on interactions with the SH3 domain (see figure 2.3) remained unperturbed in all cases, except for D45 and also V73 in USH3 K5A S6A. Perturbed residues were similar in both the Δ10 and Δ20 constructs: D45, H47, and R48 were the most affected. In the Δ10 mutant, H25, G30, S35, and T37 were also perturbed. Interestingly, the K5A S6A mutation, being a much less radical modification, had smaller but

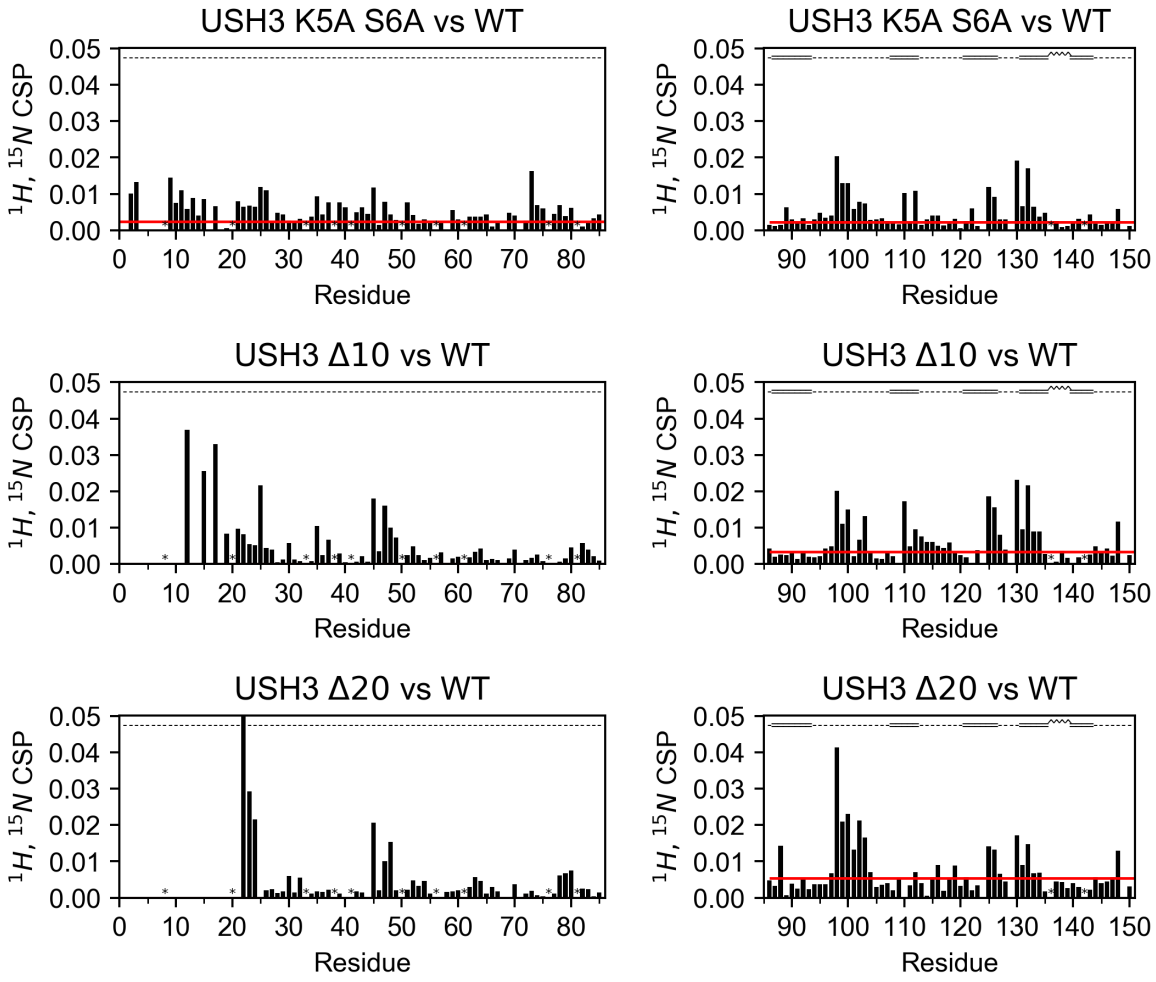


Figure 2.5: CSP of SH4 mutant USH3 constructs vs the wild type reference. The red line represents a significance threshold defined in Methods and Materials. When not shown, it is assumed to be at the noise baseline level.

more spread effects along the Unique domain, being H25, G26, D45, and V73 the most perturbed residues.

Analysis of the CSP induced in the SH3 domain signals using USH3 WT showed differential effects for the various modifications of the SH4 domain. In the case of K5A S6A and Δ10 constructs, both the RT (^{98}RTE^{100}) and distal loops (G130 and T132), along with residues in β3 and β4 strands flanking the latter, were moderately affected. R110 and Q112 in strand β2 at the beginning of the nSrc loop were also slightly perturbed, specially in USH3 Δ10. In contrast, complete removal of the SH4 domain (USH3 Δ20) induced larger perturbations in the RT loop (^{98}RTETDL^{103}), similar ones in the distal and surrounding regions, and negligible changes in the nSrc loop.

These results show that all three SH3 loops sense, although differently, the deletion or mutations of the SH4 domain in the USH3 constructs. From the different effects provoked

by the two truncations, it can be deduced that residues 11-20 are responsible for most of the CSP in the RT loop. This is in agreement with the results from the previous subsection were it was shown that, in absence of the Unique domain, SH4 alone preferentially interacts with the RT loop of isolated SH3. The K5A S6A mutation is enough to induce modest CSP in the RT and distal loops, whereas only removal of residues 1-10 provoked noticeable changes in the nSrc loop, the most insensitive one.

Perturbations in wild type USH3 caused by changes in the SH4 domain may reflect direct (SH4:SH3) effects, but could also result from indirect (SH4:Unique, Unique:SH3) consequences. To investigate these possible effects, I compared CSP against the isolated SH3 domain for USH3 WT and the SH4 variants (figure 2.6).

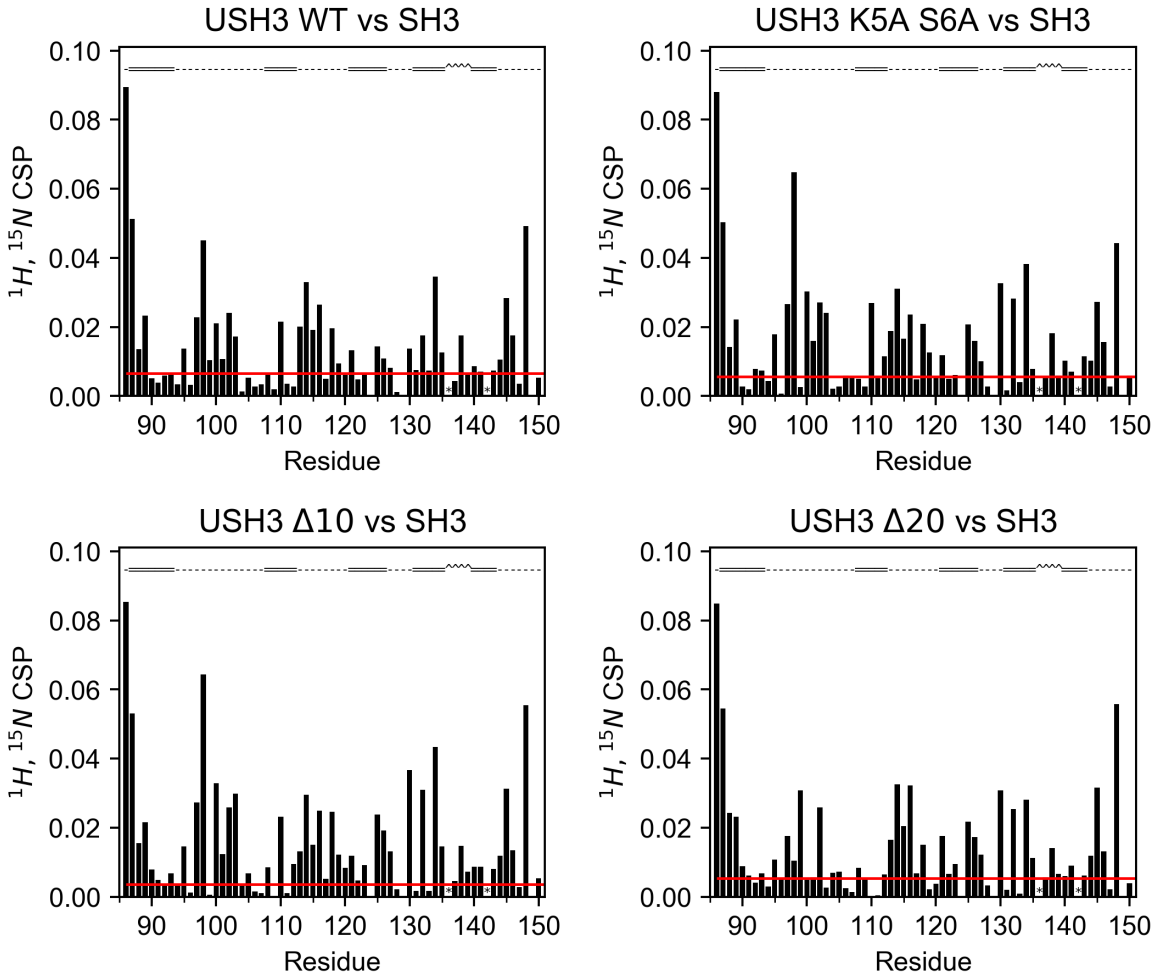


Figure 2.6: CSP of SH4 mutant USH3 constructs vs isolated wild type SH3 reference. The red line represents a significance threshold defined in Methods and Materials.

Compared to the CSP of the full-length form vs isolated SH3, a large part of the affected areas in the RT, nSrc and distal loops were still perturbed in all SH4 mutants. This confirms that both the SH4 and Unique disordered domains interact with the same zones

of SH3 in absence of each other. Importantly, the Δ20 construct showed a significantly smaller perturbation in the RT loop (95 - 101, with the exception of T99). The major role of the SH4 domain in the contacts involving the RT loop is thus manifest also when the Unique domain is present. Also interestingly, R98 and E100 showed larger CSP upon K5A S6A and Δ10 mutations.

Finally, CSP were calculated for the USH3 Δ20, Δ10 and K5A S6A mutants in presence of the PxxP peptide and compared to the PxxP ligand bound SH3 domain (figure 2.7). Doing so, the measurements would be devoid of interactions involving the RT loop.

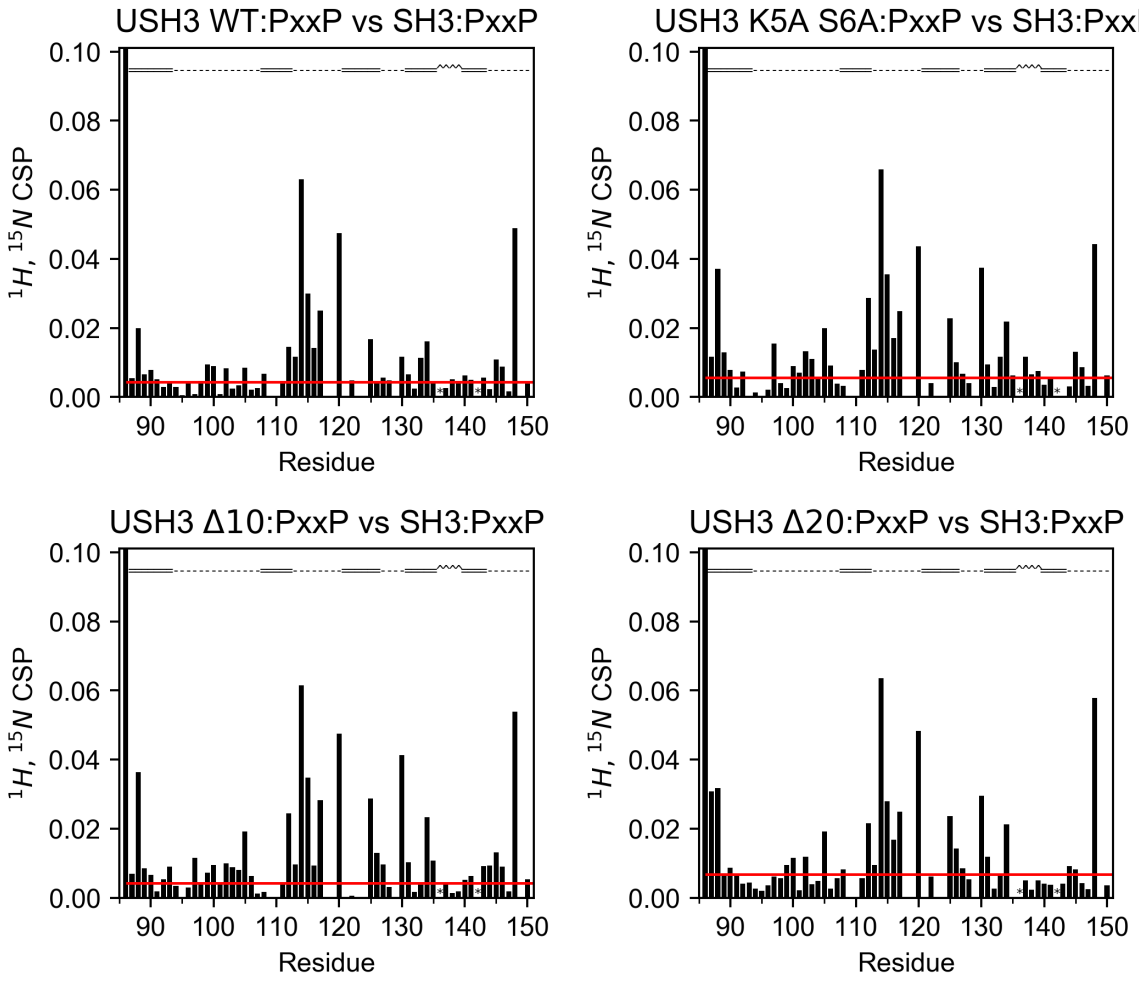


Figure 2.7: CSP of SH4 mutant USH3 constructs vs isolated wild type SH3 reference, both complexed with the PxxP peptide. Only respective SH3 domains shown. The red line represents a significance threshold defined in Methods and Materials.

Indeed, most of the perturbations in the RT loop were lost in all cases. Large CSP were observed for USH3 Δ20 in the nSrc loop and sparse residues close to the distal loop, comparable to the ones in the wild type construct (figure 2.7). Similar effects were also observed in the K5A S6A and Δ10 mutants. This means that not only SH4 but also the

Unique domain contacts the nSrc loop of the ligand-bound SH3.

In such a complex and dynamic context the interpretation of combined ^{1}H-^{15}N CSP is limited, so the overall effect of the SH4 mutations is not amenable and an alternative approach is needed.

2.1.4 CSP mapping of the effect of SH4 domain modifications

In order to disentangle the perturbations induced by the changes in the SH4 domain I first did a comparative integrated CSP analysis, using the SH3 domain as the reference frame and merging the CSP values observed in different zones of interest (Table 2.1).

Table 2.1: Integrated CSP values for different regions of the SH3 domain for USH3 WT and the respective relative changes (%) for the different SH4 variants. Central: 95 - 143; RT loop: 94 - 107; nSrc loop: 111 - 120.

Region	USH3 WT CSP (ppm)	% Change USH3 WT CSP		
		K5A S6A	Δ10	Δ20
Central	0.6771	15	15	-8
RT loop	0.2161	23	18	-27
nSrc loop	0.1975	-6	-11	-11

Defining residues 95 - 143 as the central region, which concentrated most of the changes, the overall perturbation was increased by 15 % for both the K5A S6A and Δ10 variants with respect to the wild type, whereas it was reduced by ~8 % in the Δ20. An independent analysis of the RT (94 - 107) and nSrc loops (111 - 120) revealed that these differences derive from the RT loop, for which the trend was even more evident (+23 % and +18 % for K5A and Δ10; -27 % for Δ20). In contrast, the nSrc loop showed similar small perturbation losses for all three constructs (-6, -10 and -11 %, respectively).

These results confirm that the nSrc loop is only slightly interacting with residues 1 - 10 in the SH4 domain. On the contrary, the RT loop responds more sensibly and differentially. It turns to be an intricate issue to extract further information from combined CSP, specially for such a dynamic system, in which experimental evidences point to inter-correlated interactions. The combined ^{1}H,^{15}N CSP merely reports on the variation of the local chemical environment, but do not allow to specify the nature of the change. Nonetheless, representation of the single components (i.e. the $\Delta\delta$ for each nucleus, as when comparing overlapped spectra) potentially allows to discern environments with equal CSP but different chemical shifts and hence classify them, provided we have references.

In the case of the USH3 system, we can establish an *open-closed* model. The chemical shifts of the isolated SH3 domain would represent an open state, as an hypothetical USH3 in which the IDR was completely non-interactive but (maybe) with itself. On the opposite, the chemical shifts of the wild type USH3 construct equates the closed state, having all the native interactions disregarding their specific nature. Using such references, we can plot the relative positions of each peak for both references and analyze for each mutant if the peak position resembles an open or a closed state. The CSP mapping is shown for all residues in the RT and nSrc loops in figures 2.8 and 2.9, respectively ,and the complete mapping is available in figures 5.4 to 5.7 in the Appendix.

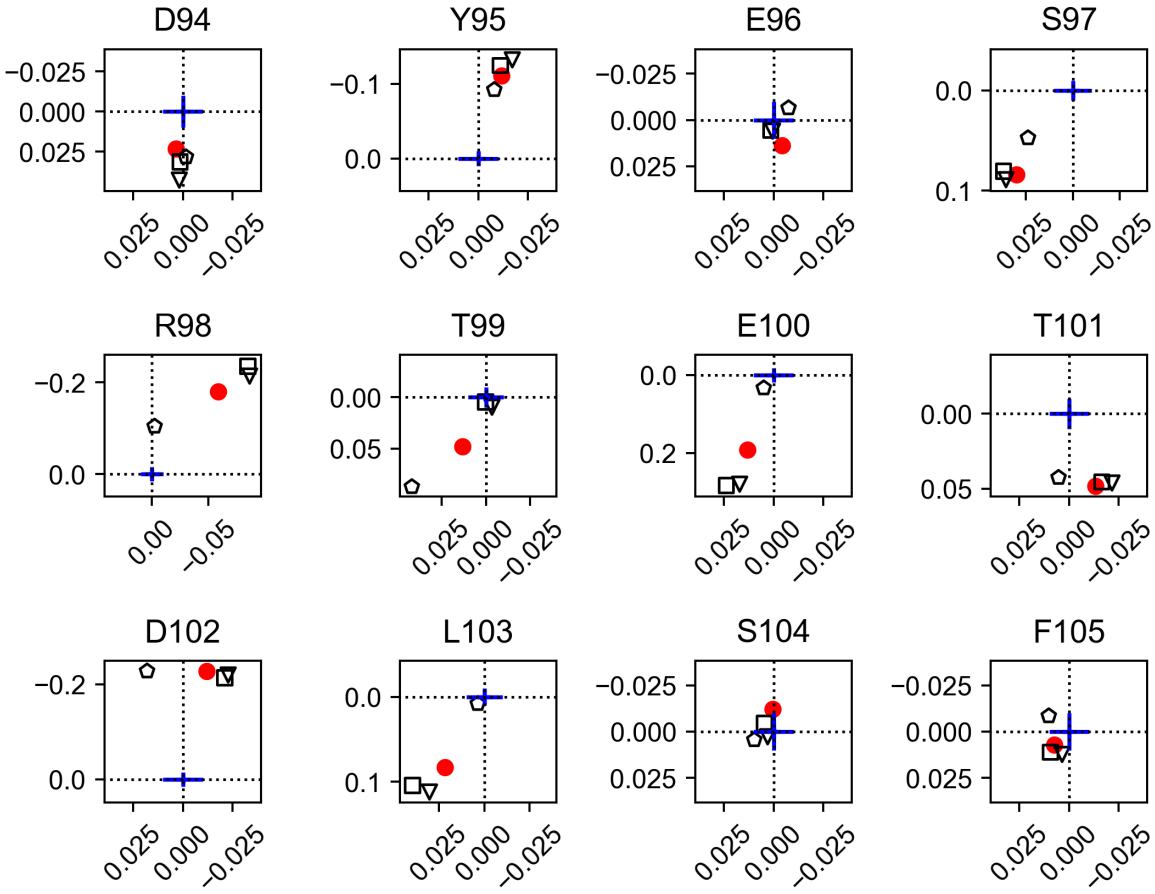


Figure 2.8: RT loop CSP mapping for USH3 WT (red dot), K5A S6A (triangle), Δ10 (square) and Δ20 (pentagon). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^1H$ (x axes) and $\Delta\delta^{15}N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

The relative insensitivity of the nSrc loop to the modification of the SH4 domain is evident, since the signals for all three mutants lay very close to the USH3 wild type reference. The RT loop signals show that removal of the SH4 domain (Δ20) makes the system shift towards an *open-like* state resembling the free SH3. A notable exception is T99 in the tip of the RT loop: whereas in K5A S6A and Δ10 the signals are *SH3-like*, the Δ20

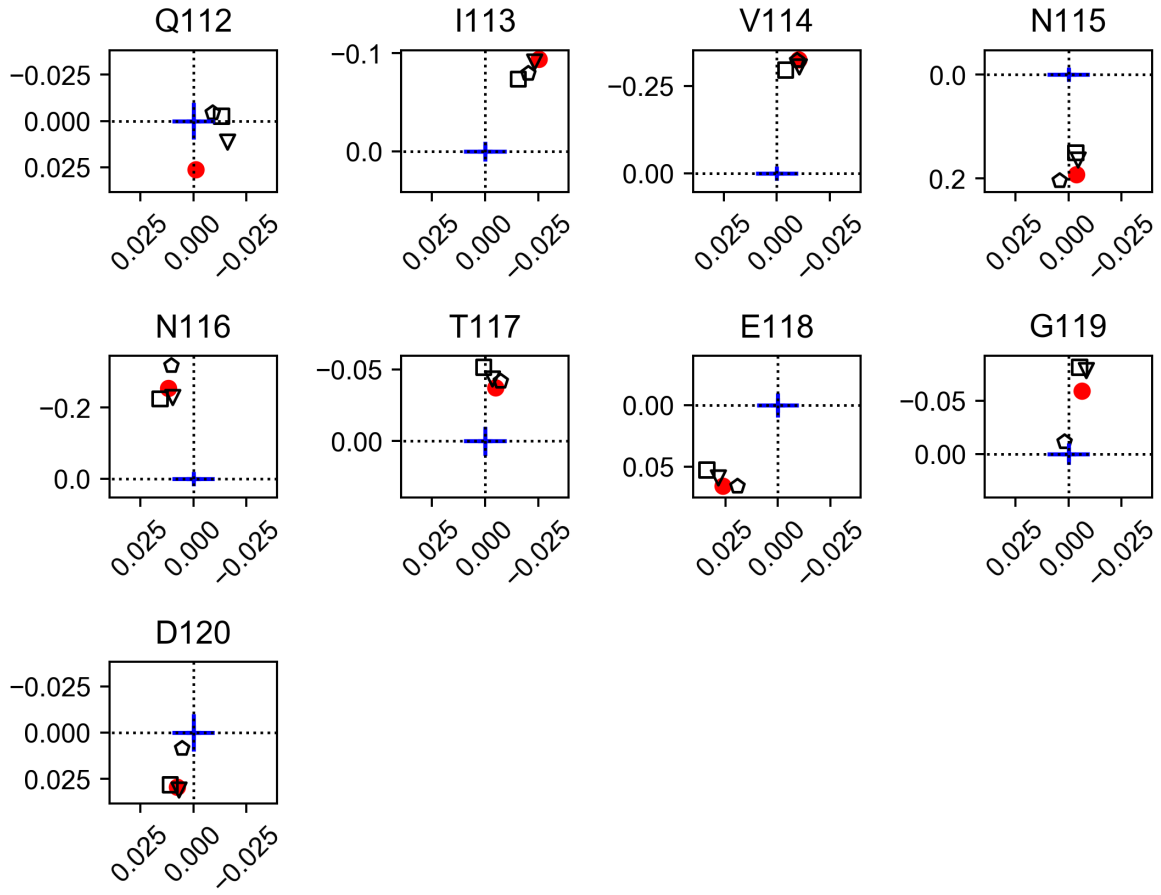


Figure 2.9: nSrc loop CSP mapping for USH3 WT (red dot), K5A S6A (triangle), $\Delta 10$ (square) and $\Delta 20$ (pentagon). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^1 H$ (x axes) and $\Delta\delta^{15} N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

T99 shifts further than the wild type reference. This suggests that residues 11-20 impair alternative interactions of T99 with the Unique domain.

On the contrary, truncation of only the first ten residues or substitution of K5 and S6 by alanine shifted the signals towards a state that would be even more *closed* than the wild type. From this we can deduce that amino acids 11 - 20 in the SH4 domain favor intimate interaction of the IDR with the RT loop. On the contrary, residues 1 - 10 partially inhibit these contacts, probably as a consequence of competing interactions, either with the rest of the IDR or with the SH3 domain.

Finally, distal loop and adjacent β sheet signals from amino acids H125, G130 and Y134 were also sensitive to perturbations in the SH4 domain. Curiously, while Y134 showed the same behavior as the RT loop, H125 and G130 were perturbed in a *more-closed* fashion in all cases.

2.1.5 Discussion

The initial results about the contacts between the IDR comprising the SH4 and Unique domains and the folded SH3 domain (Pérez et al. 2013; Maffei et al. 2015) showed that the RT, nSrc and distal loops in SH3 are sensitive to the presence of the disordered region (1 - 85). Following the CSP results of USH3 vs SH3, one may be tempted to do a straight interpretation, assuming that the Unique domain interacts mostly with the RT loop, unless detached by addition of the PxxP ligand. In such situation only the SH4 domain and nSrc loops would remain perturbed, thus suggesting a lasting, specific contact between them. However, systems including IDRs can be deceptive when studied by CSP, as discussed in section 1.10. Indeed, further experiments presented here define a more complex model.

In the first result subsection, using CSP from the isolated SH4 and SH3 domains, I find that the isolated SH4 mainly contacts with the RT loop, and more modestly the nSrc and distal loops. When SH3 is complexed with the PxxP peptide, only the SH4:RT loop contacts are lost.

At this point, a discussion on the mechanism behind the effects induced by PxxP peptide binding is needed. The association has been reported to globally affect SH3 backbone dynamics, with ligand-induced strains propagating via the network of H-bonds that hold together the β sheet core, across the small domain to distant regions (Cordier et al. 2000). Since the core of the *PxxP* binding site does not overlap the SH3 regions sensitive to the presence of the IDR, evidence supported an allosteric mechanism (Pérez et al. 2013). The CSP of VSL12 binding to isolated SH3 shown in figure 2.10 evidences large perturbations (notice that the CSP magnitude is 10-fold towards those of SH4-UD binding or SH4 mutation) in almost all the domain, specially in the RT and nSrc loops and the 3^{10} helix.

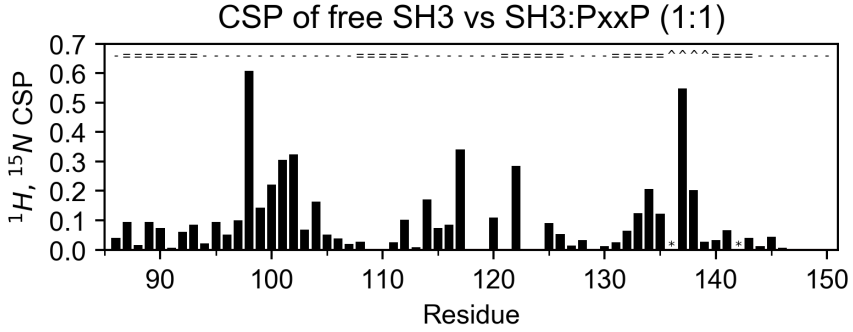


Figure 2.10: CSP induced by PxxP ligand VSL12 binding to the isolated SH3 domain. The significance threshold lays at the noise baseline.

However, the high resolution X-ray structure of c-Src SH3 bound to the VSL12 pep-

tide used here (PDB:4RTZ) was simultaneously deposited by Bacarizo & Camara-Artigas (2013). The structure revealed that the ligand could play a role in the stabilization of the RT loop. Thus, we revisited the work of Feng et al. (1995), Weng et al. (1995), and Erpel et al. (1995), (PDB:1QWF) which reported interactions of amino acids flanking the *PxxP* motif beyond the core of the binding site in SH3 being relevant for ligand affinity and specificity.

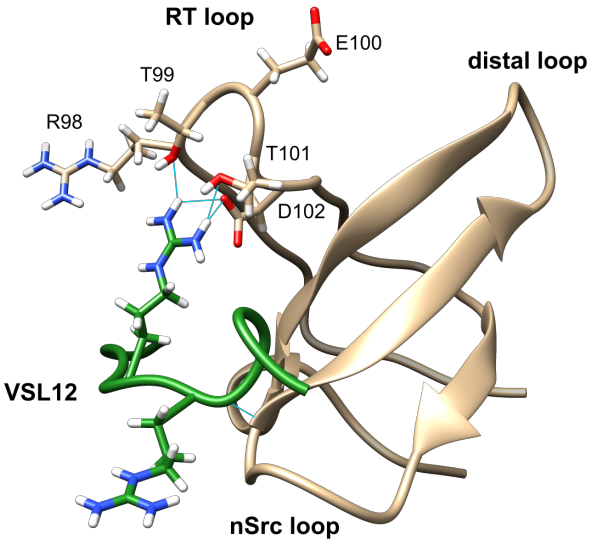


Figure 2.11: SH3 domain complexed with VSL12 PxxP peptide (PDB:1QWF). Intermolecular H-bonds are indicated in light blue.

SH3-binding ligands are classified in *class I or II*, depending if the basic residue precedes or follows the core *PxxP* motif [*(K/R)xxPxxP* and *xPxxPx(K/R)*, respectively]. They differ in the orientation of the PPII helix upon binding. In this case, the VSL12 PxxP peptide is a *class I* ligand. The peptide orientation makes the Arg side chain flanking the *PxxP* motif to insert below the RT loop and coordinate with T99 and D102, while R98 side chain points outwards, as shown in figure 2.11. These interactions probably leave the RT loop unavailable for contacting the disordered domains.

Stabilization of the RT loop by the PxxP peptide is conceivably more relevant for the interaction with SH4, given its high fraction of Lys and Arg residues (6/18 = 33 %). Moreover, the theoretically most stable configuration for the ^{14}RRR^{16} motif in the SH4 domain maximizing the distance between positively charged guanidino groups would be a helical arrangement with 120° turns. Interestingly, the two consecutive Arg in the VSL12 peptide complexed with the SH3 domain adopt a similar conformation. This suggests competition between the high affinity ligand and the weakly interacting SH4, obviously shifted in favor of the ligand. The nSrc loop would still remain available for the SH4

domain.

This discussion implies that, although the use of the VSL12 PxxP peptide for our structural studies has proven to be useful, we need to be careful in the interpretation of our results. It is likely that the SH3 domain with its internal dynamics is more sensitive than a simple rigid platform for the IDR. Therefore, we updated the affirmation that PxxP ligand binding *prevents interactions with the Unique domain* to *impairs interactions with the IDR mediated by the RT loop.*

In the second subsection, I show how the effect of mutating or removing the SH4 domain in presence of the Unique domain is also sensed by all three SH3 loops, but specially by the RT. Evidence suggests that residues 11 - 20 are a main component of the interaction with the RT loop, supporting the above-mentioned ^{14}RRR^{16} motif hypothesis. It is also demonstrated that the Unique domain interacts with the three loops in absence of SH4. In addition, the PxxP peptide-bound USH3 Δ20 mutant still displays interactions between the Unique domain and the nSrc loop. These results highlight that both disordered domains independently contact the very same regions of SH3, the RT and nSrc loops, in absence of each other. It is usual in IDPs that one binding site displays low affinity for several partners (Tompa & Fuxreiter 2008) and also a necessary condition for fuzzy binding complexes (Olsen et al. 2017).

Finally, I use CSP mapping to further detail the observations from the second sub-section. Comparison of the SH4 mutants to the isolated SH3 and the wild type USH3, taken as references in the context of an *open-closed* equilibrium leads to conclude that, in absence of SH4, open-like states (i.e., reduced IDR:SH3 contacts) are favored. This means that the whole SH4 domain and its interactions, specially with the RT loop as shown previously but also with the other loops, have an overall compacting effect. However, not all the SH4 domain contributes to closed states. Instead, I show how both the K5A S6A and Δ10 variants adopt forms even more *closed* than the wild type (i.e., display enhanced IDR:SH3 interactions). It follows that these first ten residues of the SH4 domain (and particularly K5 and/or S6) inhibit the closed state, but the effect is superseded in the full-length construct. The different consequences of the various SH4 domain alterations also endorse the possibility of crossed interactions within the SH4 domain itself or between the SH4 and Unique domains. The particular case of T99, able to establish alternative contacts with the Unique domain in absence of the SH4 domain, further confirms this point.

In all, the results presented in this section shed light on the nature of the interactions between the IDR and the folded SH3 domain. It is demonstrated that the Unique and SH4 domains promiscuously interact with specific regions of the SH3 domain, specially the RT and nSrc loops. The contact sites are scattered along the whole IDR and show a

complex interrelated pattern with the SH3 loops and between them. Therefore, the SH3 domain acts as a scaffold for the whole IDR, being the RT and nSrc loops (and the distal to a lesser extent) hubs for short range interactions with the disordered domains. This is best illustrated by the CSP mapping of the RT and nSrc signals in the SH4 mutants. Nevertheless, the interaction between residues 11-20 of the SH4 domain and the RT loop seems to be particularly important and specific, contributing to more closed conformations in the context of USH3.

Alternative and transient interactions of an IDP/IDR trough several low affinity motifs with a folded partner, combined with retained disorder by the IDR, are defining characteristics of the so called fuzzy complexes of the random type, as discussed in section 1.8 of the introduction. Although the conventional definition involves isolated partners, the c-Src N-terminal region discussed here has key attributes to be considered an **intramolecular fuzzy complex**. The entropic consequence of having both components bound to each other can be the emergency of interactions. Thus, an increase in the effective local concentration by restricted diffusion can promote very low affinity contacts that otherwise would not take place if separated[3].

Long range, non-random transient contacts within an IDR evidence a constrained conformational space, an extremely common feature in IDPs to different extents. The interactions between distant regions observed by PRE in both SH4-UD and USH3 constructs (Pérez et al. 2009; Pérez et al. 2013) suggest that pre-organization is independent of the SH3 domain scaffolding contribution. However, due to the promiscuity of the IDR-SH3 interactions, these CSP experiments do not completely clarify if there is an interdependence between these interactions and the Unique domain pre-organization. Thus, a more in depth analysis integrating long and short range structural information is required in order to fully characterize this fuzzy complex. That is the purpose of the following section.

[3]As it happens with the isolated SH4, no interaction was detected by NMR in 1:1 mixtures of isolated SH4-UD and SH3 (Pérez et al. (2009), supplementary material).

2.2 Characterization of an intramolecular fuzzy complex using SAXS and PRE

In the previous section I showed how the SH3 domain acts as a scaffold for the intrinsically disordered SH4 and Unique domains of c-Src via multiple, promiscuous contacts concentrated in the RT and nSrc loops. It was also shown that the SH4 domain residues 11 - 20 preferentially interact with the RT loop. Hence, I proposed an intramolecular, random type fuzzy complex model. Here, I will go into the details of the interface between the structured and disordered components and provide a more detailed description on the interactions and IDR pre-organization mentioned before, to further test the fuzzy complex hypothesis.

2.2.1 SAXS reveals compaction of the IDR only in presence of the SH3 scaffold

Contribution statement: Sample preparation, data acquisition and modeling were done by Dr. Yolanda Pérez, Dr. Pau Bernadó and Dr. Tiago N. Cordeiro, respectively. My contribution here is the analysis and interpretation of the experimental data in combination with the NMR results, along with Dr. Tiago N. Cordeiro, and the development of the conformer ensemble visualization method.

The first studies on the N-terminal region of c-Src done in our group (Pérez et al. 2009; Pérez et al. 2013) included unpublished SAXS measurements on the full length USH3 and the SH4-UD wild type constructs. As discussed in the introduction, the dynamic nature of our system makes the solution samples to be highly structurally heterogeneous and hence not suitable for plain modeling, so an ensemble approach to fit the experimental data was used (Bernadó et al. 2007). The scattering curves ($logI(s)$ vs s, where I is the scattering intensity and s is the momentum transfer) of both molecules show the characteristic monotonically decreasing profile typical of polydisperse mixtures, further confirming that conformational heterogeneity is retained in the fuzzy complex (figure 2.12).

The Kratky plots ($s^2 \cdot I(s)$ vs s) allows to discriminate the folded domain contribution at low angles in the USH3 profile and the increasing tail due to the disordered region (figure 2.13).

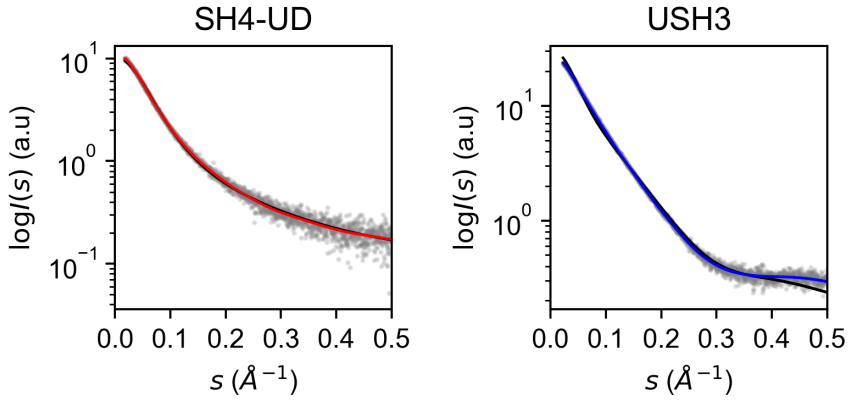


Figure 2.12: Scattering curves of SH4-UD and USH3 constructs. Experimental values are represented as grey dots, random coil fittings as black lines, and EOM fittings as colored lines (see below).

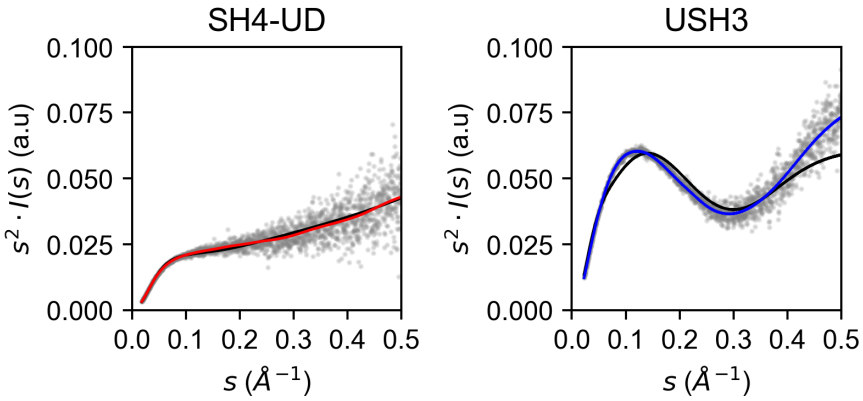


Figure 2.13: Kratky plots of SH4-UD and USH3 constructs. Experimental values are represented as grey dots, random coil fittings as black lines, and EOM fittings as colored lines (see below).

A random pool of 10 000 conformers was generated for the SH4-UD using Flexible Meccano (Ozenne et al. 2012). Flexible Meccano's conformational space exploration is based on random selection of amino acid-specific ϕ/ψ values from a curated featureless database for the backbone with steric clash as the only exclusion criterion. Afterwards, side-chains are added and their potential steric clashes alleviated, and finally the structure energy is refined in a water environment simulation using GROMACS (Hess et al. 2008) and the TIP3P water model (Jorgensen et al. 1983).

For USH3, the X-ray structure PDB:4HXJ (Xiao et al. 2013) was used as a rigid core to model the domain residues 87 - 141 and then random coil conformations were attached to its ends for the IDR (residues 1 - 86) and the mobile C-terminus end (142 - 150). In both cases, the ϕ/ψ propensities of residues 60 - 75 were the ones reported to reproduce the experimental RDC mentioned in the previous section (Pérez et al. 2009). Thus, the model is refined based on both global (SAXS) and local (RDC) structural features.

The theoretical scattering profile of each conformer was then computed using CRYSOL (Svergun et al. 1995). Then, EOM (Bernadó et al. 2007) was used to select a sub-ensemble of conformers in such a way that the average theoretical scattering curve fitted the experimental one in each case. The resulting fitting curves of USH3 and SH4-UD for both the random and EOM-selected are shown in figure **xx**. In both cases, the goodness of each fitting confirms the departure from the random coil observed by PRE mentioned in results section 2.1.

The structural consequence of the conformational bias is reflected in the distribution of the R_g values in each conformer pool (2.14). The results for each construct go in different directions. The R_g histograms for the USH3 random and optimized ensembles clearly show significant compaction. While the random coil model shows a single, quasi-Gaussian population centered at ~30 Å, the EOM-selected ensemble has a bimodal distribution in which a new, more compact population comprising ~30 % of the conformations appears centered a ~24 Å. This substantial compaction is in agreement with the previous results that revealed the many interactions between the IDR and the SH3 domain and supports the scaffolding role of the latter.

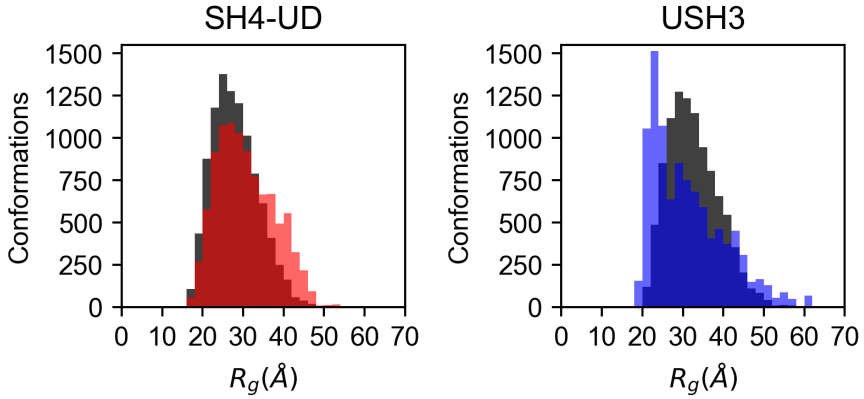


Figure 2.14: R_g histograms for the random coil (black) and EOM-selected (color) fitting ensembles for SH4-UD and USH3 constructs.

In the case of SH4-UD, containing only the disordered domains, the SAXS-optimized ensemble slightly favors more extended conformations than the random ensemble (more population in the 35-45 Å range). This can be explained because scattering derived R_g values are root mean square averages weighted by the electron density, and thus they tend to slightly overestimate extended conformations.

The fact that significant compaction is observed in the presence of the SH3 domain confirms its scaffolding role. Moreover, since the PRE contacts are very similar both in presence or absence of the SH3 domain, it is reasonable to assume that the scaffold *syn-*

chronizes pre-existent transient approximations between distant regions, boosting compaction. Correspondingly, in order to probe the details of long range contacts between the intrinsically disordered SH4 and Unique domains with the SH3 domain, and also within the IDR itself, it is necessary to rely on PRE.

A new approach to ensemble model visualization

The EOM modeling method uses the following basic workflow, which is a common strategy as mentioned in sub-section 1.9.1 of the introduction:

1. Measure an experimental population-averaged magnitude (e.g. the SAXS curves shown before).
2. Generate a featureless pool of conformations (a random coil model).
3. Back-calculate the theoretical individual values of the measured magnitude for each conformation.
4. Select a subset of the pool such that the averaged theoretical measurement fits the experimental data.
5. Comparison of the initial and selected ensembles based on averaged theoretical structural parameters.

The constructed ensembles, collections of coordinates for each atom, are always kept *under the hood*. The reason is that the initial random coil pool is an idealization, not a strict representation of the system's conformational space. Moreover, the selected sub-ensemble is nothing but a linear combination of conformations from the random pool, weighted to optimize fitting goodness to one or more experimental parameters. This means that the conformer pools are not necessarily real conformations adopted by the physical molecule in solution, but only an approximation or the solution to an equation, as previously introduced in the introduction.

Yet, if we assume ergodicity[4], a large collection of random coil conformations is a good estimation of the conformational space accessible to the system when no conformational

[4]As explained in Shannon & Weaver (1949), when a stochastic process is ergodic any reasonably large sample represents the statistical properties of the system as a whole. This property yields safe and regular statistical models. So, regarding the conformational sampling of an ideal random IDP, any extensive random set of conformations is as representative as successive frames over time - i.e. an MD simulation.

preferences exist and sampling is stochastic. On the other hand, EOM modeling reflects the deviations from that ideal behavior arising from the intramolecular interactions, regardless of their nature - this is, it reflects the conformational heterogeneity. Therefore, we have both an abstract but consistent reference (the random pool), and a representation of the system reflecting, in average, its structural features.

The USH3 random model constructed with Flexible Meccano (and the EOM-selected subensemble) can be envisaged as a solid sphere with a long tail that spreads arbitrarily from a specific point (residue 86) with the only boundary of steric clash with itself and the scaffold. The image would be that of a *cloud* surrounding the rigid surface. Since SAXS reflects global features, my purposes were:

- To obtain a realistic sense of scale of the relative dimensions of the IDR and the scaffold in the complex.
- To have a visual representation of how the EOM-selected subset diverges from the random coil pool (i.e., which specific residues originate the differences)

Thus, I first used BioPython (Cock et al. 2009) to extract the respective C_α coordinates from all 10 000 PDB files of USH3 random coil conformations and the 1 462 EOM-selected conformations. Since all structures were aligned and oriented along the major axis of the SH3 domain and the IDR conformers spanned randomly along the same axis (arbitrarily assigned as x, see figure 2.15), my initial approach was to plot 2D projections (XY, XZ, YZ) of all C_α as dots, thus reflecting a mass density map based on backbone traces.

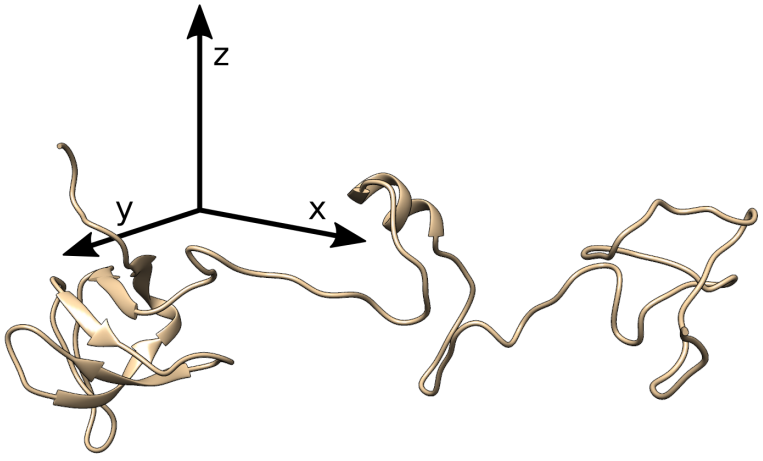


Figure 2.15: A random USH3 conformer and the reference axes.

However, the major drawback when displaying very large data sets like this (124 270 spread C_α positions from the IDR EOM model, since the SH3 is static) is *oversampling*.

Common 2D plotting software will return a featureless blob formed by overlapping points, and 3D plotting tools are more computationally intensive and perform badly such large data sets, while still suffering oversampling if the data set is very large. The alternative strategy is to aggregate the data without loosing information. Thus, I used Datashader (Continuum Analytics 2015), a Python package designed to efficiently plot big data. The pipeline is composed of three simple steps:

1. Projection: Data points are projected into 2D bins. The process was repeated for all Cartesian planes. In order to obtain comparable sets, the random coil ensemble was sub-sampled to the size of the EOM sub-ensemble.

2. Aggregation: The projected parameters are computed bin-wise as desired. In this case, the projected Cartesian dimensions were turned into counts so the bin aggregated value represented atom density.

3. Transformation: Bins are turned into pixels according to defined parameters. In this case, a blue-red color map was used to represent atom counts per pixel. A cube root relationship between atom count and color value was used in order to account the approximately spherical distribution. The result was a C_α density map normalized in volume for each projection of the cloud.

These aggregated coordinate projections satisfactorily depict the relative dimensions of the scaffold and its surrounding *cloud*. It is evident now that the IDR can sample a volume much larger than the one occupied by the scaffold. This new perspective is conceptually convenient when thinking of a fuzzy complex, because it highlights how all the weak interactions involved are transient and dynamic due to large conformational heterogeneity, but still specific as shown in section 2.1.

However, the projections fail to evidence significant differences between ensembles. Despite the significant compaction observed in the R_g histograms from each conformer set in figure 2.14, the clouds look the same size and densities are likewise similar, except in the hinge zone closer to the scaffold. Indeed, in some of the projections, the random model is slightly more dense in the region closer to the SH3 domain.

The reason is that these density plots lack sequence information. Because of the *ball and chain* shape of the individual conformations and despite of the cube root color scale correction, the atom density of the IDR cloud is dominated by a quasi-spherical distribution. Thus, I deduced that compaction may be most evident by depicting and comparing the sequential per-residue coordinate distributions on each axis.

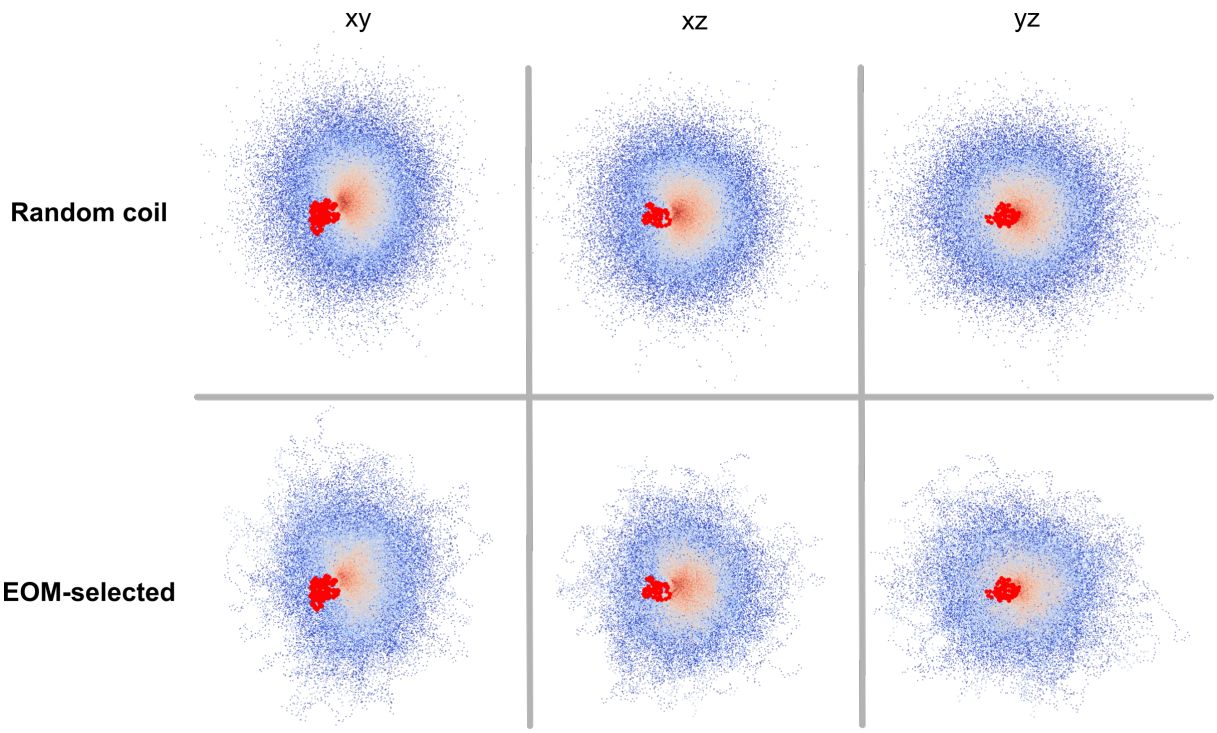


Figure 2.16: $C\alpha$ position aggregated projections for all Cartesian planes, both for the random coil and EOM ensembles. C_α density is represented in an increasing blue - red color scale, while SH3 positions are depicted in bright red.

In order to quantitatively compare the residue specific distributions of the random and selected ensembles, I used the Kolmogorov-Smirnov two sample test (Pratt & Gibbons 1981; Jones et al. 2001), a statistical tool that permits to test if two one-dimensional distributions differ.

Let x be a variable with a cumulative distribution function $F(x)$. Being $F_n(x)$ the empirical cumulative distribution of a sample of n observations of variable x, the statistic D_n is defined as:

$$(11)\quad D_n = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

If $F_n(x)$ belongs to a sample from distribution $F(x)$, then D_n tends to zero. Thus, in order to test if two samples with empirical cumulative distributions $F_{1,n}$ and $F_{2,n}$ and sizes n, m belong to the same distribution, $D_{n,m}$ is defined as follows:

$$(12)\quad D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Roughly speaking, $D_{n,m}$ (here D for simplicity) is the maximum absolute difference between the empirical distribution functions of two given samples, as depicted in figure 2.17.

Large D values therefore represent changes either on the shape or the position of the distribution.

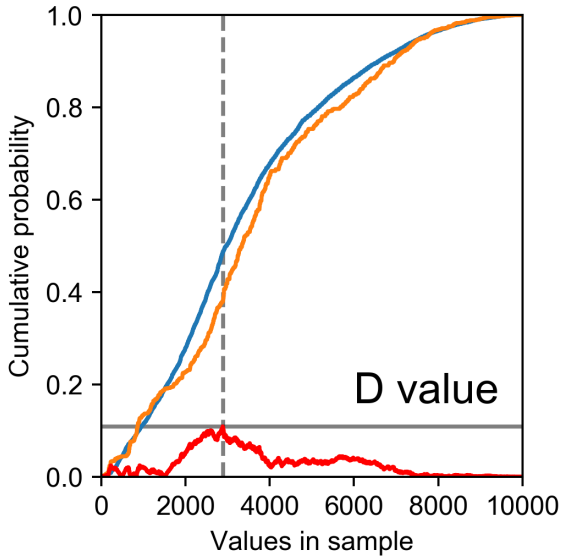


Figure 2.17: Graphical representation of the D statistic between two arbitrary samples (blue and orange). The red line represents the absolute difference between their respective cumulative probabilities.

I individually applied the Kolmogorov-Smirnov test to all normalized coordinate distribution pairs on each axis for each residue forming the IDR. Sample sizes n, m remained constant, as they corresponded to the number of reference and EOM-selected conformers. The results were then plotted as D value per residue for each individual Cartesian axis ($x = \{X, Y, Z\}$) as shown in figure 2.18). Since this method is robust with respect to the shape of the distribution and the broadening due to larger available space as residues are further from the rigid scaffold equally affects both ensembles, D values are directly comparable.

The plot clearly shows that the distributions on the Y and Z axes are very similar for both the random and optimized ensembles and the deviations are small. The X axis however evidences a significant, localized deviation. The maximum D values correspond to residues 70 - 80, while 55 - 70 show moderate values, and residues 40 - 55 have lower D. On the contrary, residues 1 - 40 and 80 - 85 display the same non-deviated behavior in all three axes. Thus, EOM has selected a subset of conformations that fit the SAXS curves in which the IDR becomes increasingly more compact as it gets closer in sequence to the scaffold, except for the five hinge residues directly adjacent to it. The selection of these precise regions being more compact may arise from the fact that the initial random pool used here had been designed to reproduce experimental RDC data, as mentioned above. Tuning the ϕ/ψ propensities randomly introduced small populations (5 - 10 %) of α helical tracts along the region 60 - 75 (Pérez et al. 2009). In the context of an intrinsically

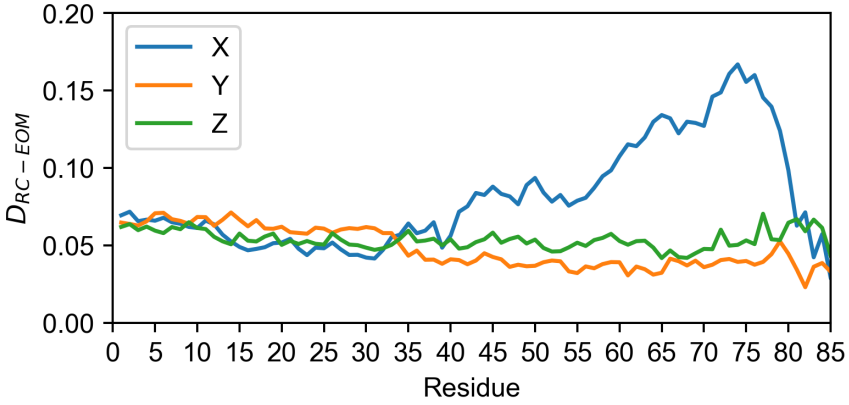


Figure 2.18: D statistics between coordinate distributions of the random coil and EOM ensembles for residues forming the IDR, over each Cartesian axis.

disordered region, such segments are more compact than the extended random coil. Since the EOM genetic algorithm follows the Maximum Weight principle (see sub-section 1.9.1 in the introduction), it is reasonable that during the weighting process EOM picked up and favored these locally compact conformers.

2.2.2 Mapping of intramolecular long range contacts in presence of the SH3 domain

As commented in the introduction, if one is to consistently characterize a disordered system using PRE, several data sets with the paramagnetic probe in different positions are needed. So, in order to gain insight on the details of long range contacts in the USH3 construct, I prepared two new mutants, A1C and A27C in addition to the already reported USH3 A59C (Pérez et al. 2013).

The USH3 A1C form would map the regions visited by the very N-terminal end, that is, the start of the SH4 domain. Position 27 was chosen because it lays in a region that is approached by the MTSL in position 59 at the beginning of the ULBR, both in the SH4-UD and the USH3 constructs. Therefore, besides mapping the contacts from that position, the PRE profile could also show if the contacts within the IDR are reciprocal. The resulting experimental PRE profiles from the IDR of USH3 constructs A1C, A27C and A59C are shown in figure 2.19, along with the corresponding random coil simulations.

Analysis of the long range contacts using PRE profiles is straightforward for the SH3 domain. Since it has a stable 3D structure, the regions approached by the paramagnetic tag in the different positions are well defined and the intensity ratio reductions can be readily plotted over its solved structure (PDB:4HXJ) as shown in figure 2.20.

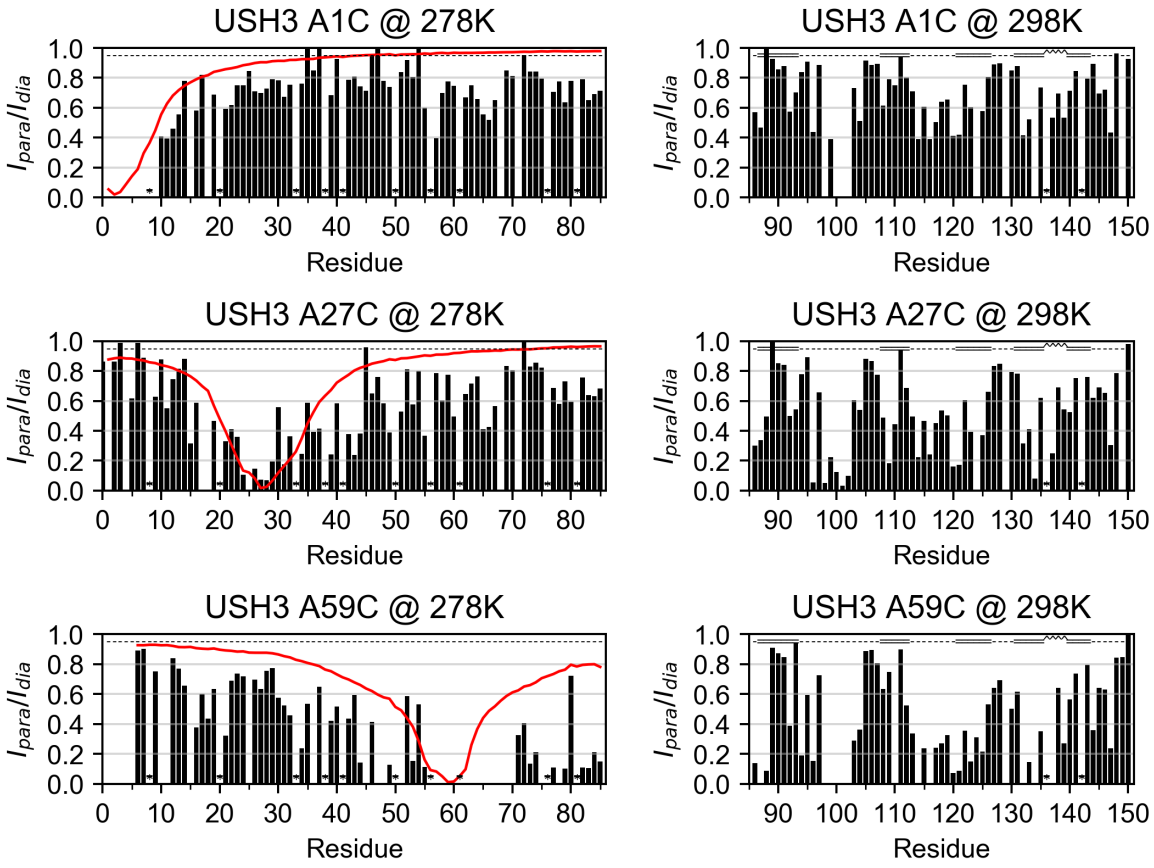


Figure 2.19: PRE profiles for USH3 constructs A1C, A27C, and A59C (Pérez et al. 2013). Cysteines indicate the respective MTSL spin label positions. The Flexible Meccano random coil simulations are shown in red for the IDRs.

The PRE induced over the SH3 signals from each MTSL position was most intense for the RT loop (96 - 104), the nSrc loop and the residues of β strands 2 and 3 immediately neighboring it (109 - 122); and residues 132 and 133 in the C-terminal end of the $\beta 4$ strand. These residues define a region that lays between both loops. In the A27C and A59C constructs, there were additional contacts affecting the whole $\beta 3$ and nSrc loop, as well as the $\beta 4$ and the following 3^{10} helix. This concurs with the increase of PRE effect with diminishing average distance to the tag position. These results therefore designate the groove formed by the RT and nSrc loops and strands $\beta 3$ and $\beta 4$ as a hot spot consistently approached by very distant regions of the whole IDR. The long range interaction mapping is thus in agreement with the CSP analyzed in the previous section, which pointed to both loops as main contact sites sensitive to the SH4 and Unique domains.

A thorough analysis of the PRE data from the intrinsically disordered SH4 and Unique domains is more intricate. In that case, both the tag and the approached residues lay in an extremely mobile segment. The effect of IDR dynamics in conjunction with the $\langle r^{-6} \rangle$ distance dependency of the PRE effect can induce large variations for intensity

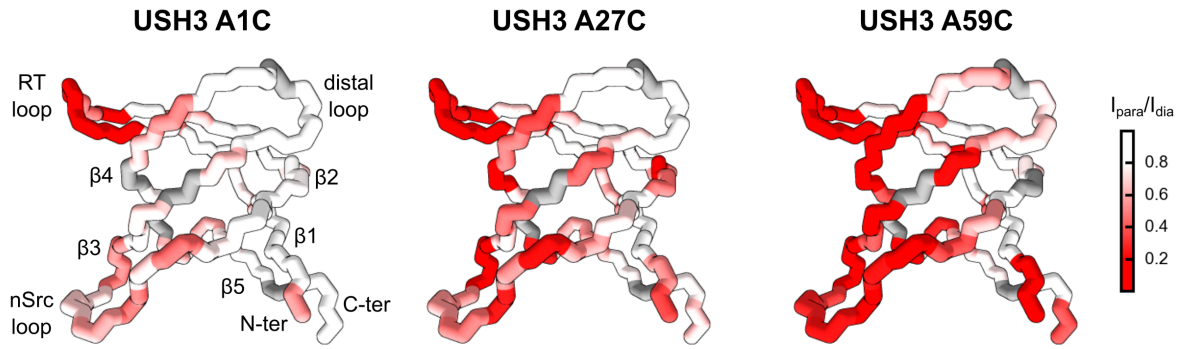


Figure 2.20: PRE profiles for SH3 domains of USH3 constructs A1C, A27C, and A59C plotted over PDB:4HXJ as a red-white color scale. Unassigned residues are colored in grey.

ratios of contiguous residues. This, together with the typical spectral overlap of IDRs and the common presence of proline residues, which do not possess amide protons and are thus undetectable in ^{1}H-^{15}N HSQC spectra, make the experimental PRE profiles of IDRs seesaw and incomplete due to lost, unassigned, or not visible signals. Following the classic approach where random coil simulations are plotted with the experimental data (figure 2.20), general trends can be appreciated and large contact regions easily identified. Still, subtle variations between profiles are easily overlooked and difficult to quantify. An alternative approach for PRE data analysis is therefore desirable for the IDR.

Analysis of PRE-detected contacts in IDRs using the novel ΔPRE mapping

Classic PRE profiles of IDRs are still a valuable tool, but their reach is limited to quantitative and low resolution information unless painstakingly modeled as explained in subsection 1.11.1 of the introduction. Such an analysis would also require a large number of PRE data sets, and presumably from complementary techniques. However, a study like this thesis already requires a large number of mutants and constructs and hence the amount of experimental work would scale dramatically. So, there is a trade-off between the model accuracy and the scope of the study. The optimal point here is to maximize the quality of the information obtained from relatively simple experiments in a wide variety of constructs.

As in the previous sub-section, the random coil model is taken as an abstract but robust featureless reference. The analysis done here many times depends on comparing low resolution PRE profiles from different samples, with an emphasis in how they differ from the random coil simulation. Conveniently, the Flexible Meccano (Ozenne et al. 2012) soft-

ware includes a module for theoretical PRE calculation that generates random backbone ensembles and accounts for the experimental variables and MTSL spin label mobility (see sub-section 1.11.3 in the Introduction)

Hence, I developed a data processing method that would permit to focus on the deviations from the reference and facilitate contrast. The approach was to calculate the differences between predicted and experimental intensity ratios (here termed **ΔPRE**) to focus on bias, and then apply a smoothing function to reduce noise while keeping the general trends.

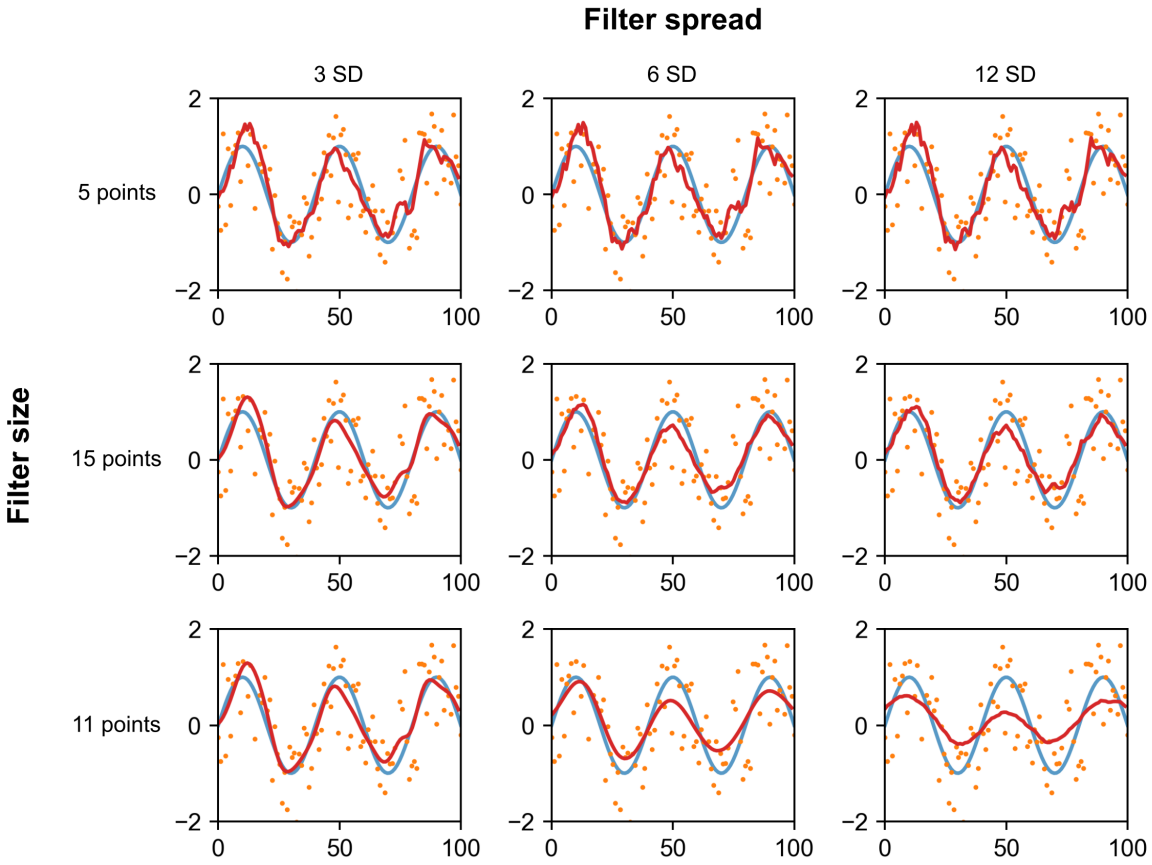


Figure 2.21: Effect of Gaussian filter size and spread (SD of the Gaussian distribution) over a dummy data set. The original sinusoidal signal is drawn as a blue line, over which random noise is added (orange dots). The different Gaussian-filtered signals are the red lines.

Gaussian smoothing is typically used in image processing to reduce noise at the cost of detail[5]. Mathematically, the procedure is a convolution of the sequence-ordered ΔPRE values with a Gaussian kernel. The advantage of Gaussian-shaped over other kind of filters such as a box (i.e., a rolling mean average) is that they have optimal frequency response, therefore noise elimination is more consistent and high frequency features are preserved.

[5]Accordingly, the method is also termed Gaussian blurring.

Filter size and spread are used-defined and permit adjustment of noise elimination and level of detail left (see figure 2.21). Because of the size and characteristics of the ΔPRE data sets I decided to use a kernel of 7 points with a spread of 1 **Standard Deviation (SD)**.

I chose the convolution routines included in the Astropy Python package (Robitaille et al. 2013). An advantage of this particular implementation is that missing values can be interpolated (up to certain extent, depending on filter size), so the curve obtained is more complete than the original set. Although purely aesthetical, since no new information is introduced, this contributes to a better visualization.

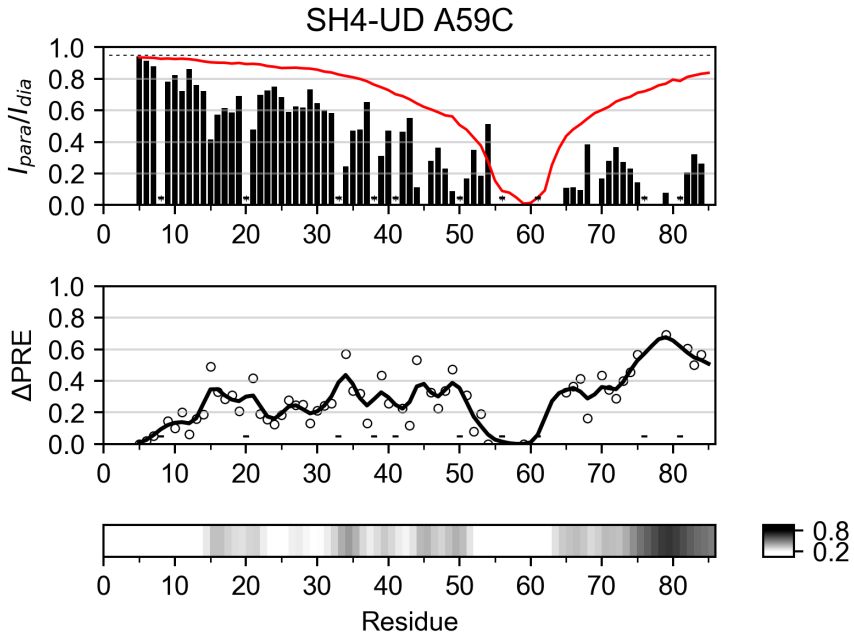


Figure 2.22: Example of how ΔPRE profiles are constructed from experimental PRE profiles and random coil simulations. The data set used corresponds to SH4-UD A59C. Raw ΔPRE values are black circles, while the smoothed profile is shown as a black solid line.

The resulting plots showing the raw ΔPRE values and the smoothed curve are called **ΔPRE profiles**. These plots directly report on long range contacts with the paramagnetic tag in the form of conformational bias from the random coil reference. They can be easily transformed into 1D heat maps that display the contact sites in a intensity color scale along the sequence. This representation mode is specially useful when comparing a large collection of ΔPRE profiles from different constructs. An example on how these plots are constructed is shown in figure 2.22. Based on the median ΔPRE values in different data sets, I established an arbitrary threshold 0.2 to consider a deviation statistically significant.

Back to the USH3 A1C, A27C and A59C constructs, all three ΔPRE plots showed de-

viations from the random ensemble simulation (figure 5.17). The paramagnetic probe at position 1 displayed significant contacts with residues 55 - 68, which encompass the ULBR and were shown to retain residual structure (Pérez et al. 2009) and 73 - 85, neighboring the SH3 domain. The construct with the MTSL tag at C27 shows most intense interactions with residues 36 - 44, 46 - 70 and 76 - 85. All these contacts are indeed reciprocal with those observed in the USH3 A59C form, in which the regions 12 - 29 and 31 - 44 close to 27 and also including part of the SH4 domain departed from the random coil simulation.

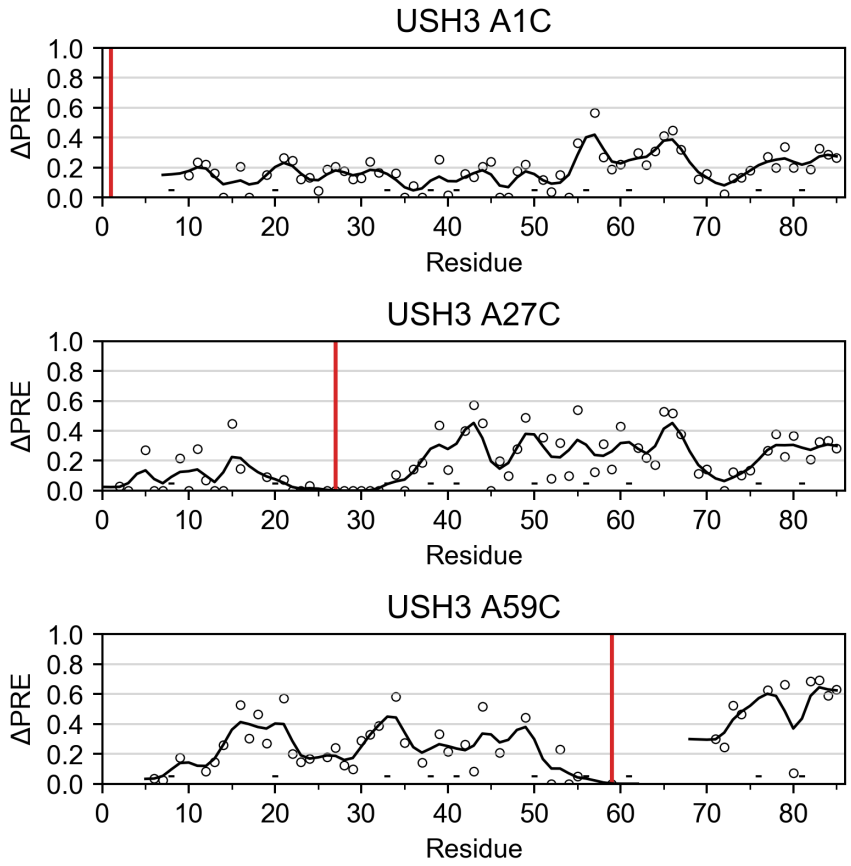


Figure 2.23: ΔPRE profiles for the IDRs of USH3 A1C, A27C and A59C. The respective spin label positions are indicated with red vertical lines.

Another important feature disclosed by these ΔPRE plots were the shapes of the curves obtained with the tag at positions 27 and 59. Thanks to the Gaussian smoothing, an oscillatory pattern could be readily identified along residues 15 - 75, with maxima consecutively proceeding with a frequency of 5 - 6 residues. Each profile has a *blind spot* in the signals neighboring the MTSL positions, due to the bleaching effect of paramagnetic relaxation in very short distances. Overlay of the two paramagnetic *points of view*, 27 and 59, provided a complete image on the interactions along the SH4 and Unique domains, as

shown in figure 2.24. Maxima and minima in the oscillations of both sets were coherent, except for amino acids 40 - 45.

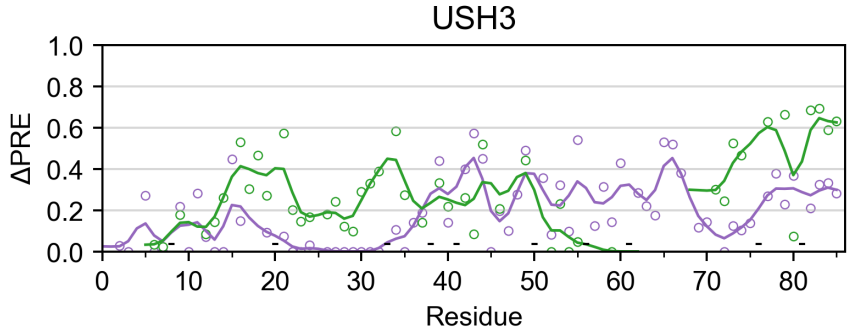


Figure 2.24: Overlapped ΔPRE profiles for the IDRs of USH3 A27C (purple) and A59C (green).

Since the ΔPRE plots directly report on averaged approximations of each residue to the paramagnetic center, the conserved oscillatory pattern can be interpreted as concerted departures from the mean distance. This puts some light on the pre-arrangement of the IDR, suggesting that besides long range contacts, shorter scale conformational restrictions may also contribute to the overall compaction. In order to further test if pre-organization is inherent to the Unique domain, I then studied in detail the PRE contacts in absence of the scaffold.

2.2.3 Long range contacts are mostly retained in absence of the SH3 domain

The PRE profiles of the SH4-UD with a MTSL radical in positions 2 and 59 had already been reported (Pérez et al. 2009). Here I present those data with improved random coil simulations and I add a new SH4-UD A27C construct that completes the set and permits comparison with the forms that include the SH3 domain. The derived ΔPRE mappings for all sets were then calculated for all previous and new constructs and shown in figure 2.25. The complete data sets can be found in the Appendix.

The PRE and ΔPRE profiles from positions 27 and 59 were practically identical regardless of the presence of the SH3 domain shown in as seen in the compared heat maps of figure 2.25, both in terms of the extent and location of the transient contacts. According to the results from the USH3 constructs, interactions observed from positions 27 and 59 are again reciprocal. The main difference was observed when the MTSL in located in position 2[6]. While in the full-length form there was a contact with the ULBR (residues 55 - 70), it

[6]The difference between MTSL tag positions between USH3 (residue 1) and SH4-UD (residue 2)

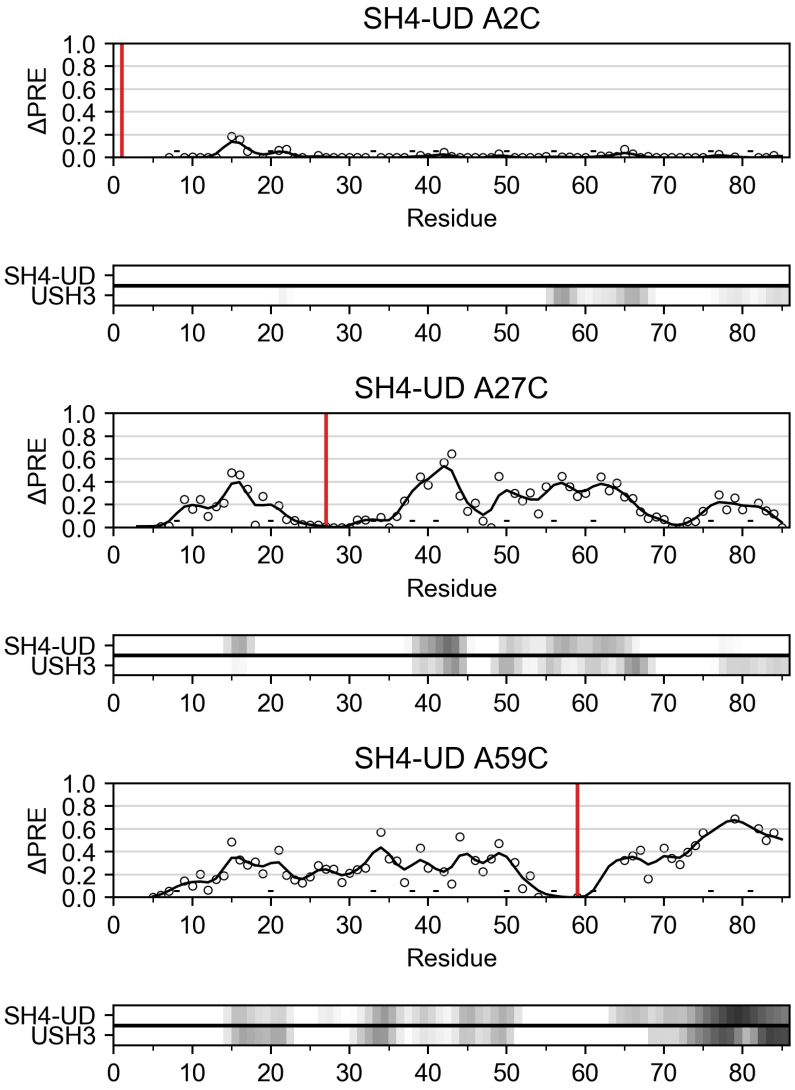


Figure 2.25: ΔPRE profiles of SH4-UD A2C, A27C and A59C. The respective spin label positions are indicated with red vertical lines. The respective heat maps are also provided (lower plots, top), along with those from the equivalent IDRs of USH3 constructs (lower plots, bottom).

was completely lost in the SH4-UD form. Only a small contact with the C-terminal end of the SH4 (14 - 19) was detected.

Also remarkably, the smoothed ΔPRE curves of SH4-UD 27 and 59 still displayed the same oscillatory patterns as in the USH3 analogues (figure 2.26). Overlay of the SH4-UD profiles showed how the oscillations observed from positions 27 and 59 were coherent with those from USH3, only with occasional ±1 shifts in the extreme positions between residues 51 - 70. Thus, the local conformational restraints are not induced by the scaffold but inherent to the IDR.

constructs is due to different cloning strategies, see Methods and Materials chapter.

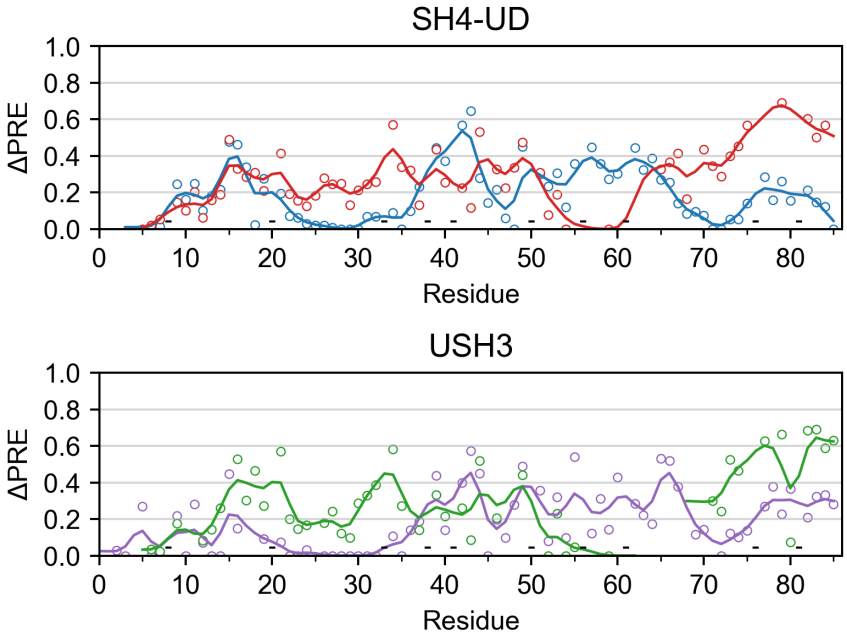


Figure 2.26: Overlapped ΔPRE profiles for: top) SH4-UD A27C (blue) and A59C (red); bottom) IDRs of USH3 A27C (purple) and A59C (green).

Finally, in order to discard intermolecular PRE effects due to non specific interactions between SH4 and/or Unique domains of different molecules (or even with the MTSL or the Streptag used for purification, see Methods and materials section), I did a control experiment consisting on specifically searching for intermolecular crossed PRE. To do so I mixed, in 1:1 proportion and to a total concentration equal to that used in the PRE experiments, ^{15}N labeled SH4-UD as an observable reporter, with non isotopically labeled (hence invisible in a ^{1}H-^{15}N HSQC) SH4-UD with the paramagnetic probe attached to C59 as the relaxation source. The maximum range of the PRE effect induced by MTSL in ~25 Å. If two different SH4-UD molecules were to approach each other closer than that threshold, statistics dictate that there is a 50 % chance that it will be a reporter-paramagnetic pair. Due to the large gyromagnetic constant of the free electron, even very modest associations would be evident. As shown in figure 2.27, the control experiment did not show any statistically significant intermolecular contact. This ensures that all observed contacts stem from conformational restraints of the disordered domains.

As a conclusion, the conservation of all the network of transient contacts within the Unique domain confirms the significant degree of pre-arrangement of the IDR and its relevance to promote interaction with the SH3 domain. Conservation of the oscillatory patterns further supports the importance of these intrinsic constraints. Finally, the fact that the SH4 engages in contact with the ULBR only in presence of the SH3 domain reinforces the scaffolding function of the folded component.

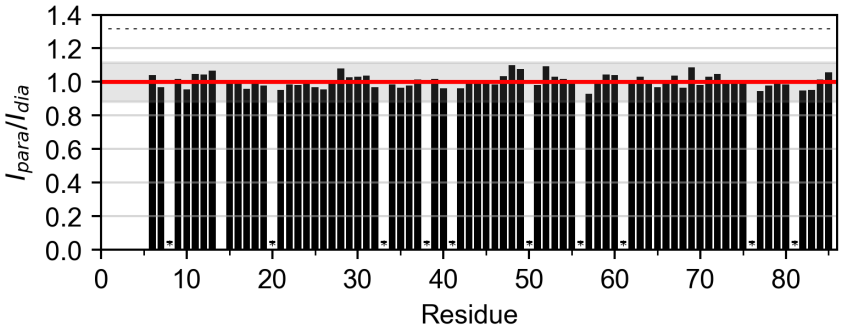


Figure 2.27: Intermolecular PRE control between spin labeled SH4-UD A59C and ^{15}N SH4-UD. The red line indicates the expected no interaction intensity ratio value. The grey area represents a confidence interval of ± 3 SD of the experimental PRE values.

2.2.4 Unique domain pre-organization is independent of the SH4 domain

In addition to the previous USH3 constructs and in order to test the potential contribution of the SH4 domain to the long range contacts, I made two truncated forms of USH3 A27C based on those described in the previous section: Δ10, missing the first 10 residues of the SH4 domain; and Δ20, lacking the whole domain. The ΔPRE in the IDRs of each construct and the PRE contacts over each SH3 domain are shown in figures 2.28 and 2.29, respectively.

Interestingly, the regions showing the maximum ΔPRE in the Unique domain for both truncated forms were the same as in the full-length: residues 36 - 44, 46 - 70 and 76 - 85. Also, the intensity of the ΔPRE was only slightly reduced. This means that the intramolecular contacts within the Unique domain of USH3, and therefore the degree of pre-organization, are independent of the SH4 domain.

The PRE profiles of the SH3 domains in the truncated Δ10 and Δ20 mutants pointed to the very same regions as in the full-length, namely the RT and nSrc loops and the groove between them (figure 2.29). The effect of partial or total removal of the SH4 domain was a progressive reduction of PRE, less noticeable in the loops. Remarkably, the Unique domain kept approaching the same region also in the Δ20 construct, although the short range SH4:SH3 interactions shown by CSP in the previous section were abolished. Hence, it follows that the SH4 domain helps keeping the Unique domain closer to the face of SH3 where contacts happen but it is not the only contribution.

These PRE experiments demonstrate that, when connected to the SH3 domain, the Unique domain is conformationally restrained to a large extent independently of the SH4 domain. SH4:SH3 interactions help keeping the dynamic *cloud* formed by the highly

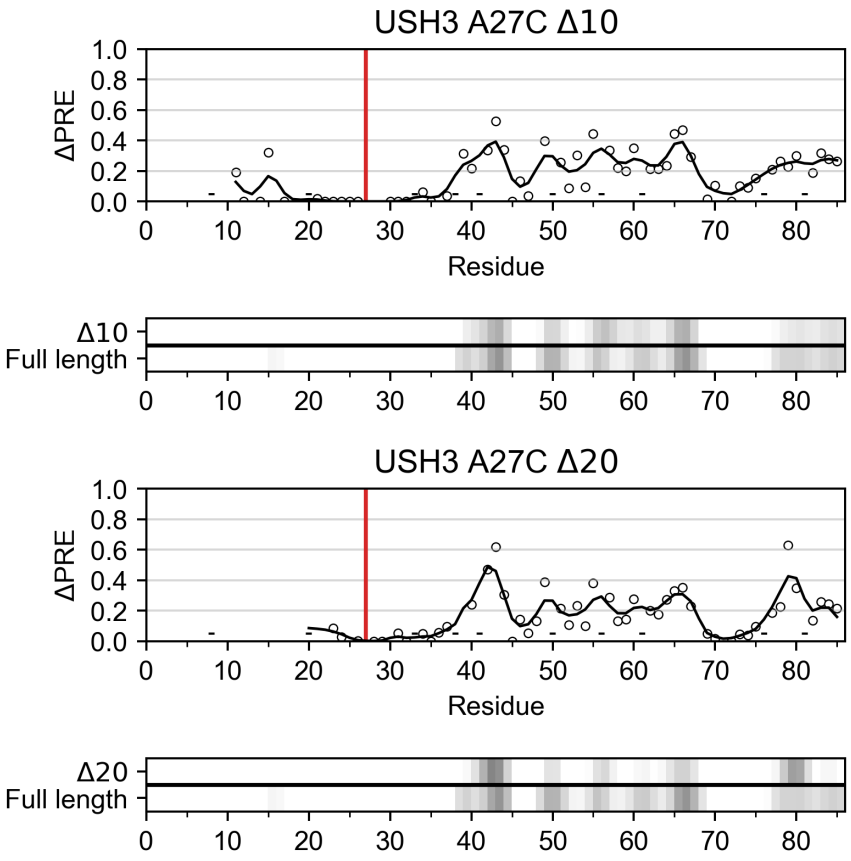


Figure 2.28: ΔPRE profiles for the IDRs of USH3 A27C Δ10 and Δ20. The spin label position is indicated with red vertical lines. The respective heat maps are also provided (lower plots, top), along with those from full length IDR (lower plots, bottom).

dynamic Unique domain closer to the scaffold surface.

2.2.5 Discussion

In this section I use experimental data from two complementary techniques, namely SAXS and PRE NMR, to detail the inter-domain long range contacts of the disordered SH4 and Unique domains with the scaffolding SH3 domain, and also the pre-organization of the IDR.

In the first section, I present experimental SAXS data and ensemble modeling, done by collaborators, for SH4-UD and USH3 samples. Data from the isolated IDR do not show any particular feature but a slight tendency towards more extended conformations. The R_g quadratic dependency on the averaged electron density in scattering measurements accounts for the apparent *conformational swelling* observed in the SH4-UD sample.

This observation may seem to be in conflict with the presence of the long range con-

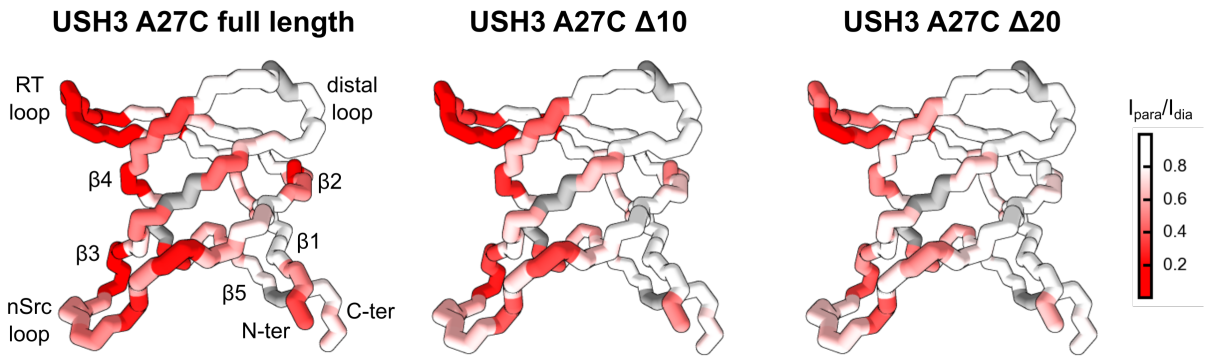


Figure 2.29: PRE profiles for SH3 domains of USH3 A27C full length, Δ10, and Δ20 plotted over PDB:4HXJ as a red-white color scale. Unassigned residues are colored in grey.

tacts observed by PRE. Intuition dictates that, if interactions between distant regions are present in the ensemble, this should be directly reflected in more compact populations than those of a purely random coil model with no specific contacts. This discrepancy has however been already reported both for intrinsically disordered and chemically unfolded proteins (Kohn et al. 2004). One factor for disagreement is the different nature of the structural parameters calculated from SAXS and PRE experiments. The spatial restraints determined by the strong paramagnetic spin relaxation mechanisms depend on $\langle r^{-6} \rangle$, being r the distance between the relaxation source and the observed nucleus, and tend to overweight the contribution of compact, minor sub-populations. Another aspect to account for is that residual structured elements are rare and transient. Kohn et al. (2004) have estimated that, at any instant, >3% of the residues of a denatured protein populates significantly compact structures. Hence, it is not surprising that the long range PRE intramolecular contacts in the isolated IDR are not detected by SAXS.

Most importantly, the full-length construct including all three domains shows a significant deviation from the random coil modeling (figure 2.14). The discrepancy was only fitted by a ~30 % population of compact conformations (~24 Å vs 30 Å). The compaction observed in presence of the SH3 domain is explained by the scaffolding induced over the SH4 and Unique domains by the folded element, due to synchronization of the transient interactions described in the previous section, while keeping a dynamic equilibrium.

In the second section I use the already published PRE data from USH3 with the MTSL tag attached to position 59 in addition with two new USH3 constructs with cysteines introduced in positions 1 and 27 respectively. Plotting the PRE ratios of the SH3 signals over the PDB structure of the folded domain permits a precise mapping of the region being approached by the different disordered regions. All long range contacts turned out to concentrate in the same face of the SH3 domain, encompassing the RT and nSrc loops

(specially approached from all tag positions) and the β3 and β4 strands laying between them. It is remarkable that the hydrophobic core that binds the canonical SH3 proline-rich ligands is located at the other face of the core β sandwich of SH3, marginally limiting with the hot spot described here. Potential implications in Src function or regulation are discussed later in sub-section 2.4.1.

It is interesting to keep in mind the projections of the SAXS-optimized ensemble cloud (figure 2.16) while inspecting the PRE and CSP detected interactions between the IDR and the SH3 domain. The volume available for the IDR to explore is much larger than excluded volume occupied by the scaffold. Paramagnetic tag positions 1, 27 and 59 are 84, 58 and 26 amino acids away in sequence from SH3. This would theoretically allow PRE contacts with any part of the SH3 surface from all MTSL positions used. However, the specificity of both short and long range interactions is evident, being all of them restricted to the RT and nSrc loops and the groove between them. The combination of specific transient interactions with a highly heterogeneous framework is a key notion behind fuzzy complexes.

The study of long range interactions within the IDR, both isolated and in the context of the USH3 construct is then addressed. Given the limitations of the classic PRE profiles, I develop the novel ΔPRE plots. The smoothed difference between experimental and simulated intensity ratios is calculated for the USH3 constructs mentioned above (A1C, A27C, and A59C) and the analogous SH4-UD forms. The results show that the interaction pattern is basically conserved for positions 27 and 59. Additionally, contacts with the tag in position 1 are observed in residues 55 - 70 only in presence of the SH3 domain, confirming its scaffolding role. It is probable that the approximation between the SH4 domain and the ULBR arises from both elements transiently interacting with the SH3 RT loop, hence shortening the average distance between them, rather than from specific interaction. Only modest contacts are observed in residues 10 - 20 from MTSL positions 27 and 59, so SH4:Unique interactions are likely weak in comparison, although enhanced by scaffolding.

Using the ΔPRE profiles I identify an oscillatory pattern in the contact intensities of both SH4-UD and USH3 constructs with the MTSL tag at positions 27 and 59. Oscillations suggest concerted alternative variations on the average distances of residues 15 - 75 to the paramagnetic center, with an 5 - 6 amino acid frequency between maxima. These joint fluctuations are a sign of local conformational restriction. As the same kind of regularly fluctuating ΔPRE curves are obtained regardless of the presence of the SH3 domain, the locally correlated dynamics originating them must reside in the Unique domain.

When searching for sequence determinants that can induce restricted motions in IDRs,

the usual suspect is proline. As it is well known, proline has a particular Ramachandran plot and also do the residues preceding it due to the locked ϕ angle (Ramachandran et al. 1963; Ho & Brasseur 2005). IDRs are typically enriched in proline residues as they are the most disordered-promoting residues due to their capacity to disrupt secondary structure elements (such as α helices) and thus avoid collapse into a stable configuration (Campen et al. 2008; Cheng et al. 2010; F.-X. Theillet et al. 2014). Thus the presence of prolines increases flexibility in structured proteins, but can induce stiffening in the highly dynamic disordered context, where its restricted ϕ angle reduces the degrees of freedom.

The SH4 and Unique domains contain 9 prolines: P8, P20, P33, P38, P41, P50, P56, P61, and P81. When proline positions (with a ± 1 range) are plotted along ΔPRE (figure 2.30), it becomes evident that most of the maxima sit next to proline residues. Their abundance and spacing can help to extend correlated motions beyond their adjacent neighbors. The only non proline-associated maxima are found in the SH4 (residues 14 - 19) and the ULBR (63 - 68). The outliers can be respectively explained by the rigidity induced by the ^{14}RRR^{16} motif, derived from the need to accommodate the positively charged side chains and the residual structure detected by RDC in the 60-75 stretch.

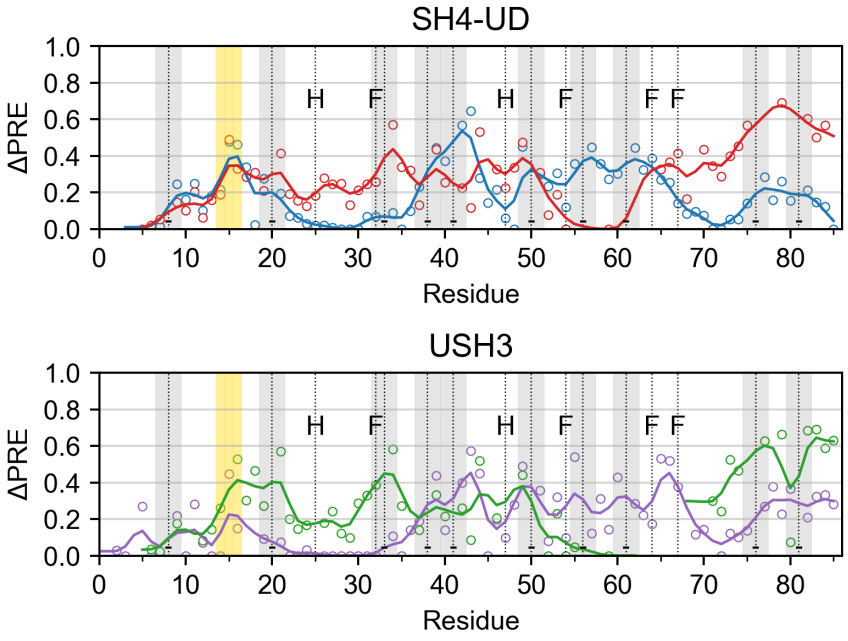


Figure 2.30: Overlapped ΔPRE profiles for: top) SH4-UD A27C (blue) and A59C (red); bottom) IDRs of USH3 A27C (purple) and A59C (green). Phenylalanines and histidines positions are marked with vertical lines and text. Prolines are also marked, with a grey area delimiting positions ± 1 around them. The ^{14}RRR^{16} motif in the SH4 domain is highlighted in yellow.

A consequence of these collective motions is the alleviation of the entropic cost for far-off regions to mutually approach (Baxa et al. 2014). Therefore, the oscillatory behavior of the ΔPRE is yet another aspect of the IDR preconditioning that further highlights

its importance. Also, the existence of *proline brackets* intercalated between short motifs has already been proposed and successfully applied to sequence-based identification of functional elements in proteins (Kini & Evans 1995; Kini 1998). More recently, Crabtree et al. (2017) have reported that mutation of prolines flanking IDR segments with helical propensity upon partner binding affects the lifetime of the complex between the disordered MLL protein and the folded KIX domain of CBP. IDRs are hubs for protein-protein interactions, such as recognition by kinases or calmodulin binding in the case of the SH4 and Unique domains. A delimiting *proline mark-up* may therefore tune partner recognition by locally promoting exposition or hindrance in a dynamic environment.

The new ΔPRE plots hence allow to disclose the dynamic pre-organization of the IDR in a new, more detailed level than the stark observation of diffuse contacts in the classic PRE profiles, while keeping relatively simple experimental requirements. As a summary, the ΔPRE values are used to chart the ensemble of interactions within the IDR with and without the scaffold from all MTSL positions (figure 2.31). This plot provides an unequivocal visualization of the contact network within the disordered domains and of these with the scaffold and help to put in context the PRE data.

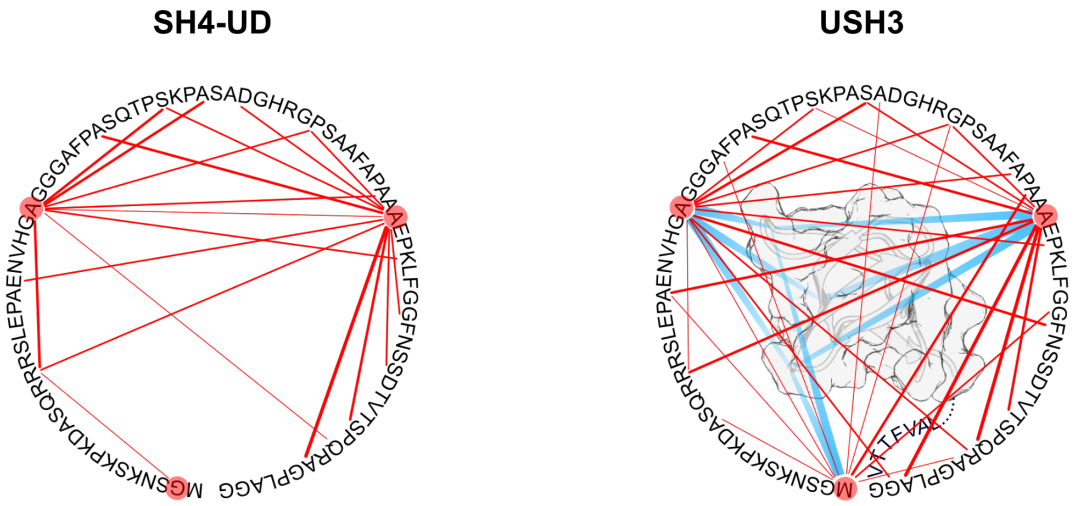


Figure 2.31: Cartoons depicting the network of long range contacts detected by PRE in SH4-UD and USH3 from all constructs. Spin label positions are highlighted in red in the sequences. Intra-IDR interactions are quantitatively represented as red lines, while SH4-UD:SH3 contacts are qualitatively indicated with blue lines.

The effect of partial or total truncation of the SH4 domain on the long range interactions was also tested with USH3 constructs Δ10 and Δ20 with a paramagnetic probe in position 27. In both cases only a small progressive reduction of the PRE contacts was observed, but not changes in the approached residues. Thus, the overall effect of the SH4 domain in terms of long range contacts is to help holding the IDR closer to the SH3 domain RT and

nSrc loops. It is also demonstrated that specific Unique:SH3 interactions independent of the SH4 domain contribute to keep the IDR cloud in the vicinity of the RT and nSrc loops.

The fact that the $\Delta 10$ mutant, which showed increased short range contacts by CSP, specially in the RT loop (sub-section 2.1.4), did not lead to more intense PRE effect leads to think that probably the first ten residues of SH4 are inhibiting alternative interactions involving other parts of the IDR. Those competing contacts would be released upon $\Delta 10$ removal but not affect the global compaction towards the SH3.

In conclusion, there exist two distinct contributions to the USH3 compaction determined by SAXS: one dependent on the scaffold, and another intrinsic to the Unique domain. The first includes the SH4:RT and Unique:RT/nSrc short range contacts studied by CSP in the previous section, here shown to help holding the *IDR conformer cloud* closer to the scaffold. The specific contact between residues 11-20 in SH4 and the RT loop is a prevailing component. The second contribution depends exclusively on Unique domain pre-conditioning, reported as non-random PRE contacts and locally restricted dynamics associated to prolines.

Interplay between both mechanisms is evident in one sense: The reduced entropic cost endowed by intrinsic conformational restriction facilitates low affinity interactions with SH3. More specifically, Unique domain pre-organization can promote the contacts of the SH4 domain with the loops by restricting its excursions to the proximity of the SH3 face displaying PRE and CSP, thus providing a tethering function.

Correlation in the opposite direction has not been observed so far, at least regarding the SH4 domain. PRE results show that Unique domain pre-organization is independent of its presence or interactions with SH3. Besides, Unique:SH3 interactions are mostly unaffected as observed by CSP and PRE. Nonetheless, emergent behavior is expected in random type fuzzy complexes, specially if undergoing ensemble allostery (Tompa & Fuxreiter 2008; Motlagh et al. 2014; Sharma et al. 2015). Thus, if the threshold for productive interactions is lowered by conformational restriction, alternative low affinity associations may gain function.

The next relevant questions to answer are: where does the pre-arrangement of the intrinsically disordered Unique domain stem from? As commented in the introduction, the functional information of a peptide chain is ultimately concealed in its primary structure, the amino acid sequence (Anfinsen 1973), although in the case of IDRs the *code* is more cryptic than in folded proteins. Prolines have been pointed out as locally stiffening elements from the ΔPRE plots, but concerted dynamics alone do not explain compaction.

Therefore, there must be other particular sequence determinants encoding for the short and long range contacts described here and in the previous section. Identifying and testing those key elements is the topic of the next section.

2.3 Search and assessment of sequence determinants for Unique domain interactions

In the last section I exhaustively explored the long range intramolecular contacts that arise in USH3 in the context of a random type fuzzy complex. Results emphasized the importance of the conformational pre-configuration in the intrinsically disordered Unique domain. The open question now is: which specific elements in the sequence of the IDR induce pre-organization and enable those contacts? In order to find out, I present here the results of sequence alignment and functional mutations of USH3 of specific, conserved residues important for the formation of the fuzzy complex.

2.3.1 Sequence alignment of the SFKs

As discussed in the introduction (section 1.5), intrinsically disordered sequences have distinctive amino acid usage statistics (Dunker et al. 2001). For example, charged and polar amino acids that enhance chain solvation are over-represented in contrast with bulky hydrophobics which, on the contrary, favor its collapse. Another related feature is low sequence complexity (Romero et al. 2001), which sometimes results in stretches of repeats of a reduced number of amino acids. Also, IDR sequences usually display increased evolutionary rates (Brown et al. 2002). Finally, functional elements in IDRs tend to be *peptide motifs*, sequence patterns smaller than 10 amino acids, like the SH3 recognition motif described in section 2.1 (Davey et al. 2012; Pancsa & Fuxreiter 2012; Tompa et al. 2014). This is the case of the Unique domain in the Src Family Kinases, for which even the closest members in the cladogram display very low conservation (see figures 5.1 to 5.3 in the Appendix). These features (reduced amino acid dictionary, sequence degeneracy, high variability, small patterns) hinder homology studies on IDPs, specially because algorithms find problems on aligning small elements and distributing gaps in such lowly conserved sequences.

However, I already had a considerable amount of structural information that could be exploited to interpret alignments. So, I decided to use the MUSCLE algorithm (Edgar 2004), which searches and aligns small sub-sequences as a first approach (*k-mers*), to align the sequences of:

- The closest SFK members to Src (SrcA subfamily): Yes, Fyn and Fgr.
- Src homologues of biologically relevant models: *Homo sapiens*, studied here, *Mus musculus*, *Gallus gallus* and *Xenopus laevis*.

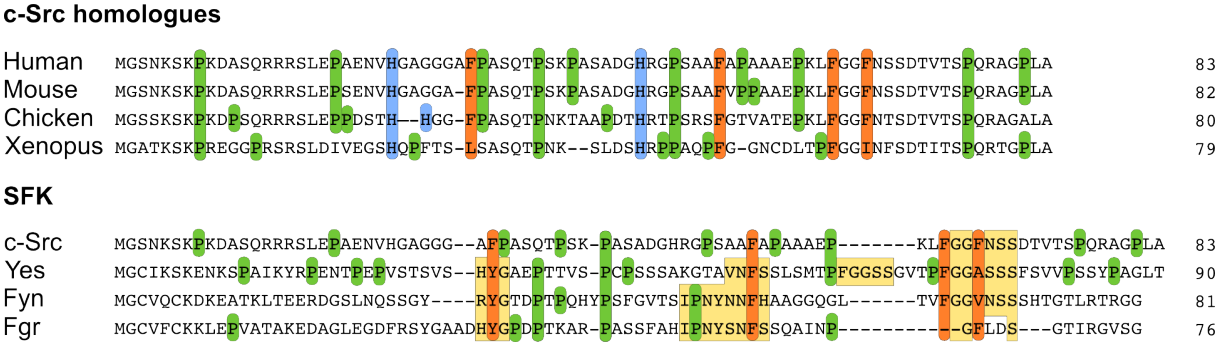


Figure 2.32: Alignment of c-Src homologues and closest SFKs from the SrcA subfamily. Prolines are highlighted in green, phenylalanines in orange, and histidines in blue. Particularly conserved motifs containing hydrophobic residues are marked in yellow.

The first appealing feature in the light of our previous knowledge is that the ULBR core *FGGF* motif, which is key for lipid binding by the Unique domain as previously reported and functionally relevant (Pérez et al. 2013), was highly conserved among all homologues, and also conserved with some variations (*FGGx* being x an hydrophobic) in other SKFs. Hydrophobic residues in general and aromatics in particular are uncommon the IDP sequences (Dunker et al. 2001) and are, when present, often involved protein-protein interactions locally forming hydrophobic-enriched interfaces (Mészáros et al. 2011).

Hence, the conservation and potential functionality of these uncommon residues drew my attention. Moreover, other similar patterns containing aromatics were found in the SFK family: $\Phi_1 xx \Phi_2$, where Φ_1 is Phe or Tyr, x are Gly, Ser or Gln (turn-promoting residues) and Φ_2 is an aromatic or hydrophobic. The motifs found are ^{50}YNNF^{53} and ^{64}FGGV^{67} in Fyn, ^{54}FSSL^{57} in Yes, and ^{53}YSNF^{57} in Fgr. Another conserved pattern involving aromatics , *IPSNYxNFx* being x = Ser, Asn or His, was found in Fyn (46 - 53) and Fgr (48 - 56) and partially (^{52}VNFS^{55}) in Yes. Only F54 was conserved in Src. A last motif was *xYG* (x= His or Arg), which is conserved in all members except for Src, in which only a Phe in position 32 is found in place of Tyr. Interestingly, some of these Tyr are known phosphorylation sites in SFKs (Amata et al. 2014), thus connecting the observation of conserved aromatic residues with one of the most common post-translational regulatory systems on IDPs (Bah & Forman-Kay 2016).

Conserved aromatic residues F32, F54, F64 and F67 in the Unique domains of c-Src happen to lay in the region that concentrates most of the long range contacts detected by PRE (figure 2.30). This suggests that the pre-organization may derive from hydrophobic contacts between phenylalanine residues, not sufficiently intense to collapse and fix a stable structure but enough to hold a dynamic hydrophobic core, forcing the Unique domain into loop-like conformations. The next section is dedicated to the experimental testing of this working hypothesis.

A second interesting feature is the distribution of the proline residues associated with the oscillatory patterns observed in the ΔPRE profiles (also figure 2.30). Although exact positions are not conserved among SKFs, the number of prolines is similar between the conserved F32 and F54 and their equivalent positions: 4 in Src, Fyn and Fgr and 3 in Yes. Only P41 in conserved and present in all members of the family. Coupling between the number of prolines between conserved interacting hydrophobics may be a signature of a pre-configuration mechanism common among Unique domains.

2.3.2 Conserved aromatic residues mediate long range contacts within the Unique domain

As introduced in the previous sub-section, a number of conserved residues, namely F32, F54, F64 and F67, are located along the Unique domain of c-Src and are candidates to drive long range interactions. F64 and F67 are already known to be functionally relevant, forming the core of the Unique Lipid Binding Region. Mutation of F64 and its two immediately flanking residues to Ala (^{63}LFG^{65} to ^{63}AAA^{65}) resulted in abolished lipid binding, aberrant phenotypes in *X. laevis* oocytes and reduced invasiveness in *in vitro* assays with Src-dependent cancer cells (Pérez et al. 2013; Maffei 2015).

The PRE profiles of these USH3 and SH4-UD AAA mutants with an MTSL radical in position 59 had been published (Maffei et al. 2015) and showed somewhat distinct profiles suggesting an influence on the restricted conformational space of the Unique domain. The exact details remained to be explained. However, the hypothesis of an hydrophobic cluster mediated by conserved aromatic residues may establish a new functional layer including the ULBR. The first step was to test the effect of individual replacement of each phenylalanine to alanine on the transient contacts within SH4-UD. The absence of SH3 would allow us to focus exclusively in the contribution to the long range contacts and pre-organization of the disordered domains. Because of the *blind spot* due to the paramagnetic, we alternatively chose MTSL positions 27 or 59 (our *wild type* references) in such a way that the F#A mutation would be as close as possible to it and most of the long range effects remained observable. In order to facilitate a comparative analysis, ΔPRE profiles and heat maps for the wild type and F#A mutants were calculated (figure 2.33, PRE profiles can be found in the Appendix)

The ΔPRE profiles from all SH4-UD F#A mutants retained a similar shape compared to the respective references, but long range contacts were significantly reduced in all cases. The F32A mutation dramatically reduced contacts of the paramagnetic probe with the extensive region 50 - 70 (-52 % lower ΔPRE values in average). However, it did not affect

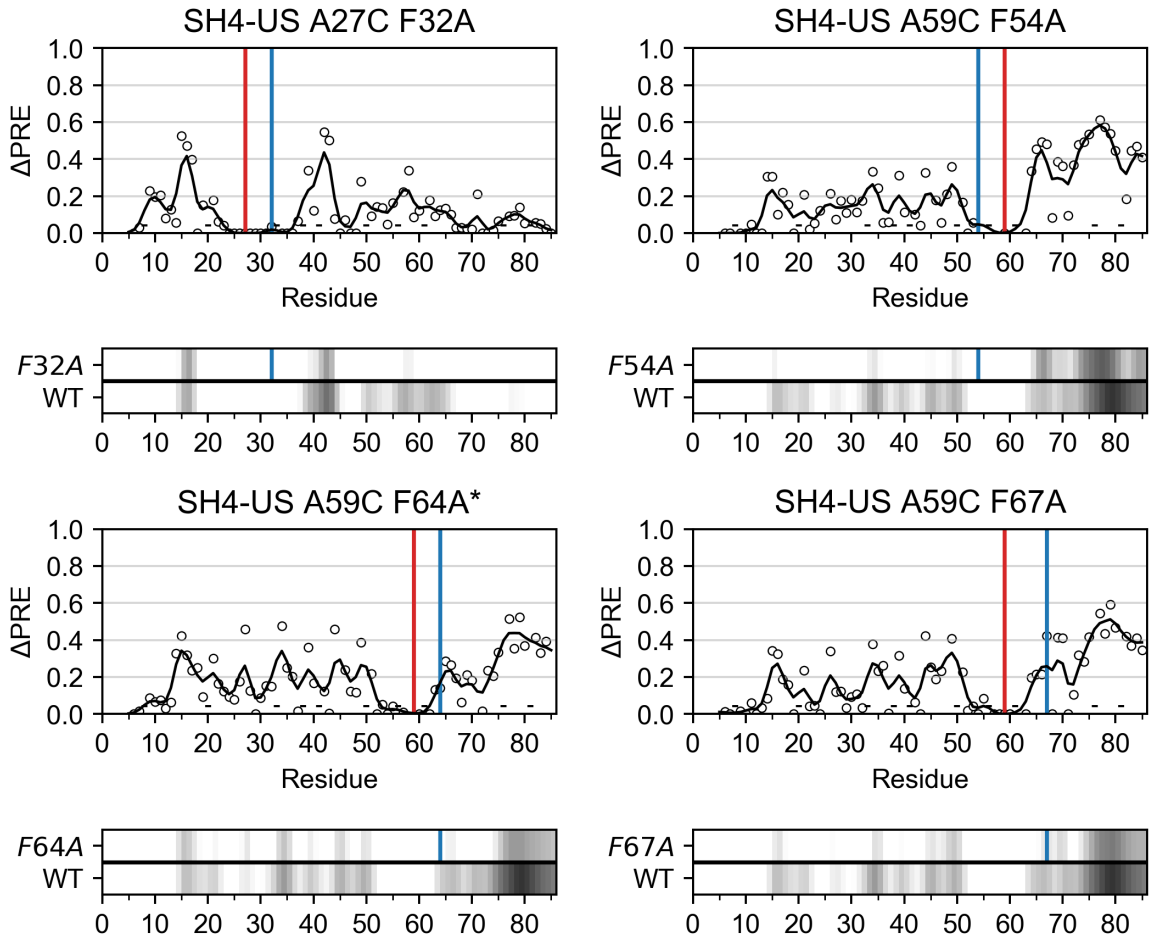


Figure 2.33: ΔPRE profiles of all SH4-UD F#A mutants. Spin label positions are indicated with red vertical lines, while F#A substitutions are drawn in blue. The respective heat maps are also provided (lower plots, top), along with those from the wild type reference (lower plots, bottom). *: The F64A construct corresponds to the triple ^{63}LFG^{65} to ^{63}AAA^{65} mutant above mentioned, from Pérez et al. (2013).

contacts with residues in the SH4 domain. Substitution of F54 had a corresponding effect, diminishing interactions in region 30 - 50 (-43 % ΔPRE reduction in average) but also in 15 - 20, at the end of the SH4 domain. Phenylalanine residues F64 and F67 in the core of the ULBR also showed to mediate long range interactions, but to a lesser extent than F32 and F54. Their respective quenching effect on the ΔPRE in the 30 - 50 stretch was -34 and -36 %, respectively. Their effects on the contacts with the C-terminal residues of the Unique domain (75 - 85) were instead more intense.

Thus, the mutational analysis of ΔPRE confirmed the differential contribution by aromatic residues to Unique domain compaction. It can be interpreted that F64 and F67, known to bind lipids and closer in sequence to the SH3 domain (85 - 150), also contribute to maintain long range contacts, but F32 and F54 are the ones help the most to bring together the N-terminal half of the Unique domain into loop-like conformations, having

a more noticeable effect. Particularly, the SH4 domain was only sensitive to replacement of F54. Either a side effect of disrupted hydrophobic contacts between F54 and F32, that may bring SH4 close enough to sense F54 substitution, or a very weak direct interaction, this supports that Unique domain pre-arrangement is independent from SH4.

Secondly, although the intensity of the interactions was partially quenched, the oscillatory patterns were retained in all cases. Therefore, local correlation may be an intrinsic property of the Unique domain of Src independently of transient hydrophobic contacts.

2.3.3 Aromatics also affect short range inter-domain interactions

Next, I tested the effect of individual phenylalanine replacement in the context of USH3 constructs. The idea was to check if the conserved aromatics forming the hydrophobic core that compacts the Unique domain would have a dual role, being also involved in inter-domain interactions - i.e., Unique:SH3.

In order to dissect pre-organization and specific contacts as much as possible, I used CSP instead of PRE to focus on changes in the short range contacts upon the different F#A mutations. The CSP of each F#A construct were calculated with USH3 wild type as reference (figures 2.35 and 2.35).

Substitution of F32 for alanine lead to extensive CSP in the region 20 - 50 in the Unique domain. H25, G26, T37, A42, S43, H47, and G49 were specially affected, besides A31, adjacent to the mutation point. The effects over the SH3 domain were very significant in β2 (R110, Q112) and the distal loop and surrounding β strands 3 and 4 (H125, S126, G130, and G133).

F54A mutation also induced CSP in a wide section of the Unique domain, 25 - 30 and 44 - 66. The intensity was however smaller than in USH3 F32A, residues H25, A44, and L63 being specially affected. The affected signals from SH3 residues were the same as by F32 substitution, but in this case the CSP magnitude was 2 - 3 times smaller.

F64 and F67 have been mentioned to form a highly conserved *FGGF* motif with known lipid binding capacity that can be abolished by mutation of ^{63}LFG^{65} to AAA. Interestingly, both F#A substitutions displayed a very similar CSP pattern. Regarding the IDR, large CSP were observed in the stretch 55 - 85 in both cases. For USH3 F64A, the most affected residues beyond those neighboring the mutation were ^{72}TVT^{74}, L82, and S6 in the SH4 domain. Interestingly, S17 was also perturbed. F67A substitution also affected ^{72}TVT^{74}

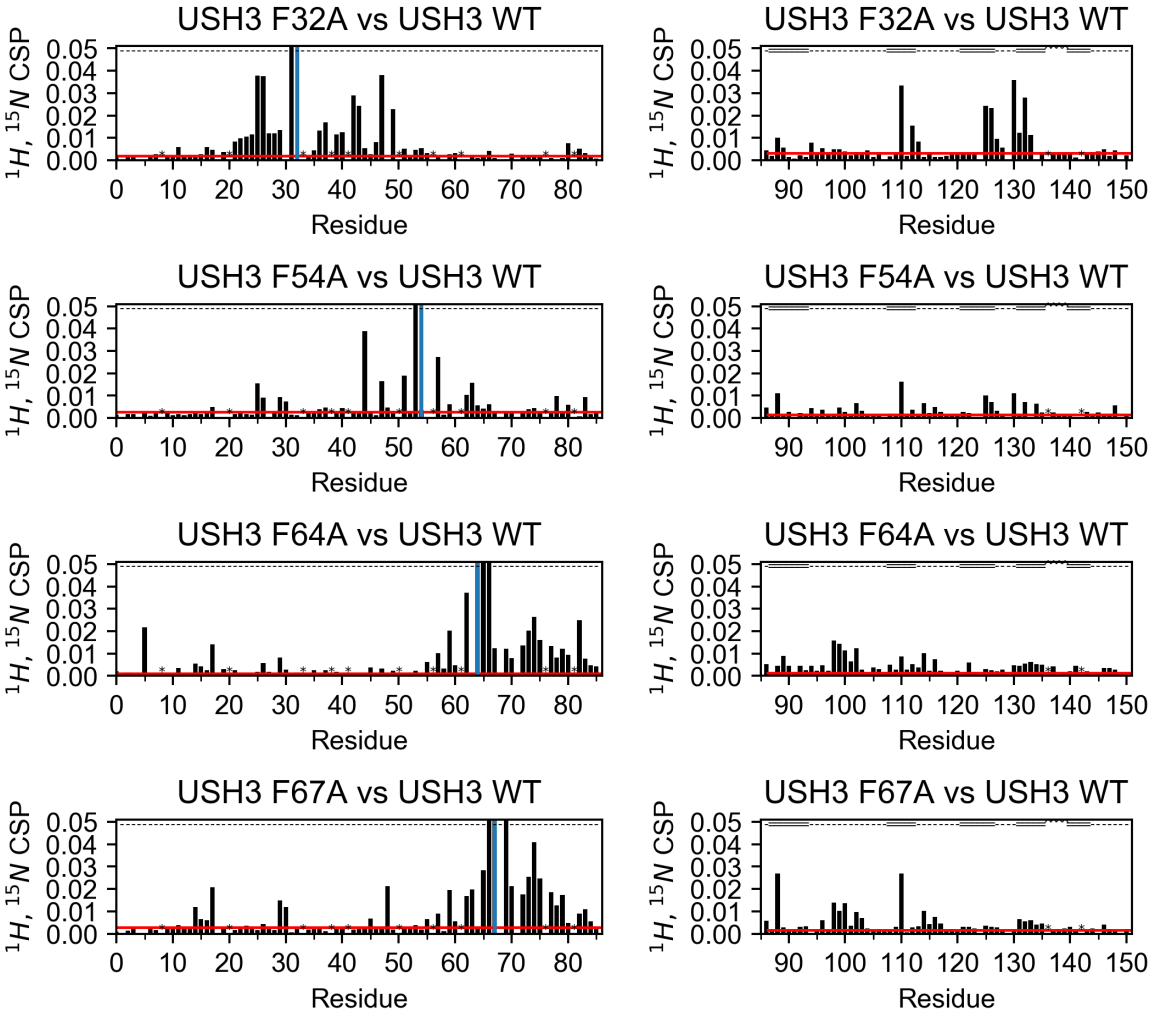


Figure 2.34: CSP induced by each F#A mutation in USH3. F#A substitutions are indicated with blue lines. The red line represents a significance threshold defined in Methods and Materials.

and the nearest residues specially, and also R48, G30, G29, and again S17 but more intensely than F64A, as well as its preceding *RRR* motif more modestly. Regarding the SH3 residues, CSP were modest but concentrated specially in the RT loop (^{98}RTETD^{102}) and affected less importantly the nSrc loop (^{113}IVNN^{116}) and β4 (^{130}GQTGYI^{135}). In addition, only upon F67 mutation T88 and R110 displayed large CSP.

It should be remembered here that the CSP observed in the SH3 domain do not necessarily stem from direct short range contacts involving the specific mutated residues, but also from a possible rearrangement of the conformational ensemble. This may affect the ability of the interacting groups embedded in the dynamic mesh that is the Unique domain (e.g., residues T37, D45, A55, etc. described in sub-section 2.1.1) to reach the RT and nSrc loops and other secondary binding zones.

As in the ΔPRE analysis of the role of the conserved aromatics in the Unique domain

compaction, it is again evident that the phenylalanines can be classified in two distinct pairs.

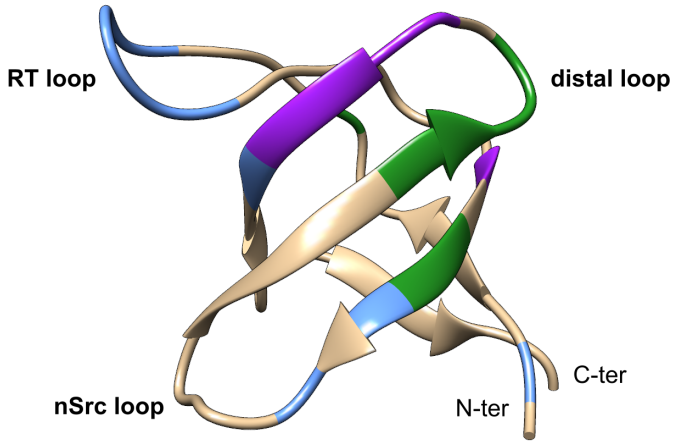


Figure 2.35: SH3 domain residues affected by USH3 F#A mutations plotted over PDB_4HXJ. Green: affected by F32A and/or F54A substitution; blue: affected by F64A and/or F67A replacement; purple: affected by both mutation pairs (F32-F54 and F64-F67).

On the one hand, F32 and F54 induce CSP in the N-terminal half of the Unique domain but not in the SH4. Their effect over residues in the SH3 domain is also particular, not being affected the RT and nSrc loops, main hubs for short and long range contacts, but the β strands laying between them, and the distal loop.

On the other hand, F64 and F67 substitutions mainly affect the C-terminal region of the Unique domain. As this particular stretch including the ULBR had been reported by RDC to retain a small extent of residual structure or rigidification, it is likely that the core *FGGF* motif plays a central role on it. Disruption of such particular pattern, with two bulky hydrophobic residues separated by small and steric-hinder free glycines seems to transmit changes further away than the immediate local environment. Also regarding the CSP over the SH3 domain signals, the behavior is different. Both F64 and F67 substitution to alanine induce changes in the loops, most importantly in the RT.

The small but noticeable perturbations observed in the SH4 domain, particularly in S17, may be a side effect due to scaffolding, as interactions between the ULBR and the RT loop, and the SH4 domain with the same site of SH3, take place alternatively, thus reducing the average distance.

In conclusion, it has been proven that the conserved aromatic residues in the Unique domain of c-Src play dual roles regarding intramolecular interactions in the context of USH3, both modulating Unique domain compaction and affecting its short range contacts with the diverse binding sites of the SH3 domain.

2.3.4 Effect of Unique domain induced loops on inter-domain interactions

Contribution statement: This section was done in collaboration with Diana Navarro as a part of her M.Sc. thesis under my practical supervision. My exclusive contribution was NMR experiment acquisition, whereas sample preparation and data analysis were done together.

The fact that a binding event between a host and a flexible ligand typically benefits from ligand pre-structuring due to reduced entropy penalty is well known in supra-molecular chemistry (Wittenberg & Isaacs 2012). Despite lacking the classic biunivocal relation between binding sites, fuzzy complexes may also benefit from this effect by further restriction of the search space for productive encounters. Moreover, IDPs/IDRs that fold upon binding tend to display conformational selection to different extents (Dogan et al. 2014). Therefore, an alternative approach to further explore the relationship between Unique domain compaction and Unique:SH3 interaction was undertook.

I hypothesized that, if the SH4 and Unique domains natively tend to populate loop-like conformations, it may be possible to capture them just by allowing the system to spontaneously form disulfide bonds. Thus, we prepared a SH4-UD construct (the isolated IDR) with cysteines introduced in the positions that were observed to establish mutual long range contacts by PRE, 27 and 59, (sub-section 2.1.2) thus potentially enclosing a 32 amino acid loop including F32 and F54. Additionally, we prepared a less strained variant (50 amino acid loop) by inserting cysteines at positions E22 and T72, respectively at the extremes of the Unique domain bordering the SH4 domain and the ULBR, and encompassing all four conserved phenylalanines.

Using a mild oxidizing conditions (1.1 equivalents of $K_3[FeCN_3]$, atmospheric oxygen, vigorous stirring) and high dilution (dropwise addition over 7 mL of oxidant solution), we were able to produce both cyclized forms C27-C59 and C22-C72 with no dimerization or scrambling products, as confirmed by LC-MS.

We analyzed the CSP between the cyclized and wild type forms of SH4-UD, overlooking the obvious CSP induced in the vicinity of cysteine insertion points (figure 2.36). We found that larger perturbations concentrated within the loop of the less strained cyclized form, C22-C72, whereas in the C27-C59 construct, CSP were smaller and spread both inside and outside the cycle. In SH4-UD C22-C72, the largest changes were observed between amino acids 39-65 and all phenylalanines were similarly perturbed. However, the more strained SH4-UD C27-C59 showed relevant CSP only for D10, D45, H47 and less

importantly T74, as well as other minor changes.

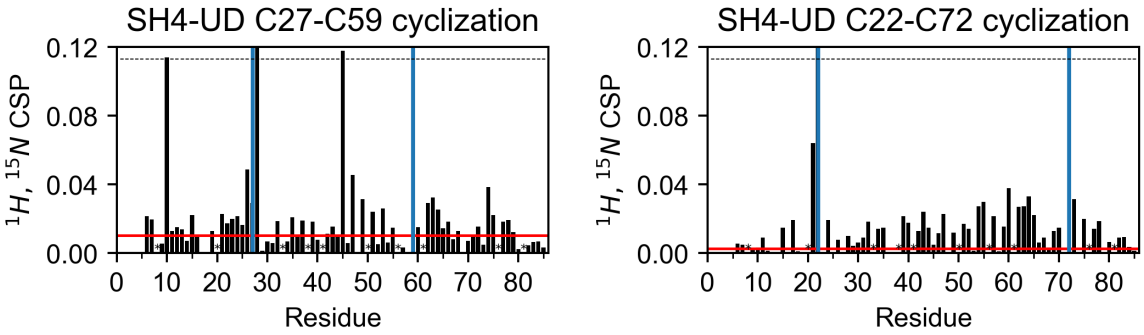


Figure 2.36: CSP induced by cyclization of SH4-UD A27C A59C and E22C T72C constructs. Cysteine positions are indicated with blue lines. The red line represents a significance threshold defined in Methods and Materials.

The fact that the less strained loop displayed larger CSP may seem counter intuitive. However, it must be accounted that the larger loop encompasses all four aromatic residues, whereas C27-C59 only contains F32 and F54. Since ring currents induced by the magnetic field contribute very significantly to the chemical environment of nearby or approached nuclei, it is not strange that conformational restriction of the four results in larger CSP than only two being more tightly held.

Next, we tested the effect of induced restrictions in the Unique domain on its interaction with the SH3 domain (figure 2.37). In order to do so, we added one equivalent of isolated SH3 domain to ^{15}N labeled samples of SH4-UD C22-C72 and C27-C59, and monitored the CSP induced in the latter by the presence of the scaffold.

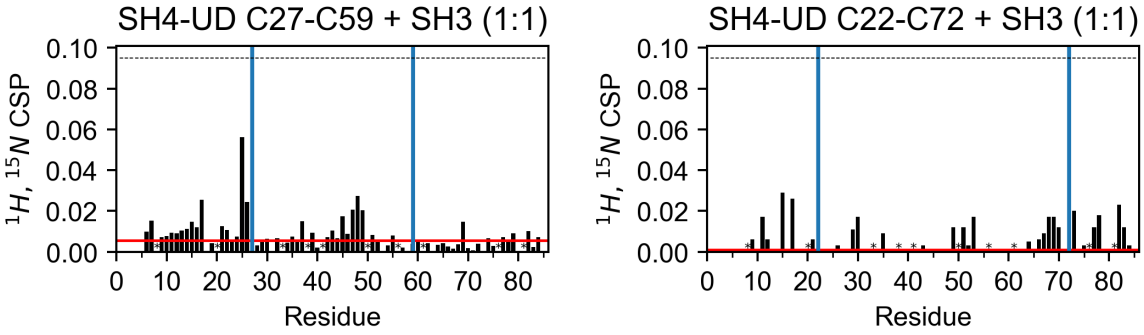


Figure 2.37: CSP induced by addition of 1:1 SH3 to cyclized SH4-UD C27-C59 and C22-C72 constructs. Cysteine positions are indicated with blue lines. The red line represents a significance threshold defined in Methods and Materials.

Accordingly to the initial hypothesis, upon addition of free SH3, the shorter C27-C59 cycle suffered significant CSP in residues K7, ^{14}RRRS^{17}, H25, and G26 outside the loop, and T37, D45, ^{47}HRG^{49} and S69 within it. The latter residues showed a consistent pattern,

and are part of the set known to contact SH3 (see sub-section 2.1.1). The less constrained loop instead showed smaller CSP, many of them outside the cyclized region (A11, R15, S17, V73, R78, and L82 vs G29, G30, G49, S51, A53, N68, and S69).

2.3.5 The role of histidines in the Unique domain

> **Contribution statement**: All mutant USH3 samples used in this section were produced by Farman Ali Khan during his internship in our group. I supervised him, acquired the NMR spectra and analyzed the data.

The Unique domain contains two histidine residues, H25 and H47. The histidine imidazolic ring is an aromatic conjugated system, but also an amphoteric moiety. The pK_a of the protonated histidine (conjugate acid of the neutral form) is ~6.0, depending on its particular environment. Therefore, in near-physiological conditions histidines can be sensitive to changes in pH or ionic strength in the medium. Since in IDRs side chains are generally exposed to the solvent, histidine responsiveness to small changes on its average environment is magnified. Additionally, the strong influence of ring currents in the chemical shift (as explained in the case of phenylalanines in the previous section) can induce strong perturbations upon slight variations in the bulk solution when doing NMR experiments.

During the different experiments I performed either with wild type or mutated constructs of c-Src SH4-UD and USH3, I had already observed histidine hypersensitivity for both H25 and H47. All the protein samples prepared during this thesis were dissolved in a 50 mM sodium phosphate buffer ($I = 0.11$ M at room temperature) at pH = 7.0. The buffer was carefully prepared fresh, the pH controlled using a pH-meter previously calibrated each time, and pH checked again when differences were observed. Still, the two histidine residues have been observed to sometimes show slight CSP even between samples of the same construct (figure 5.18 in the Appendix).

For this reason, and the fact that none of them is conserved among other Src Family Kinases, I had excluded them from the study on the role of aromatic-driven Unique domain compaction or Unique:SH3 interaction.

However, both H25 and H47 were shown to be sensitive to the presence of the SH3 domain, substitution of F32 and F54, and the introduction of conformational restraint via disulfide bonds (sub-sections 2.1.3, 2.3.3 and 2.3.4). Therefore, I decided to explore the effects of both H#A replacements in USH3 constructs using CSP.

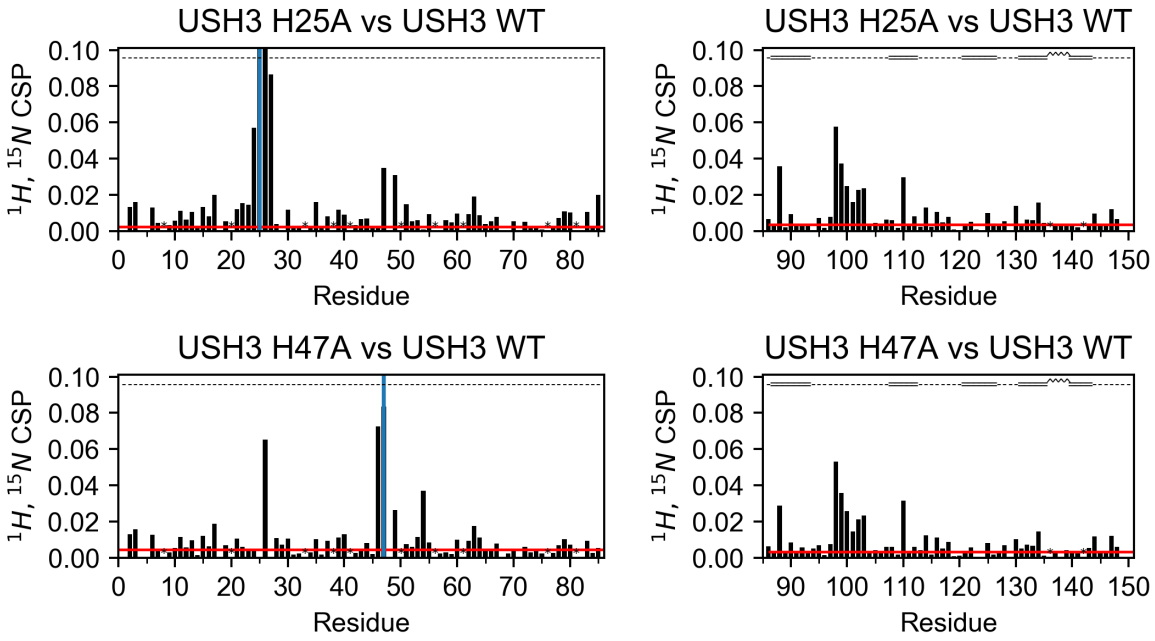


Figure 2.38: CSP of USH3 H25A and H47A vs USH3 WT. H#A substitutions are indicated with blue lines. The red line represents a significance threshold defined in Methods and Materials.

Regarding the intrinsically disordered regions, it was found that many scattered, mainly polar, residues showed modest CSP upon both mutations. Examples include S6, R15 and S17 in the SH4 domain and S35, S39, K40, L63 or A79. Most importantly, I observed that substitution of either H25 and H47 was sensed by the remaining histidine and surrounding residues. Regarding contacts with the conserved aromatics, it was interesting to observe that only F54 showed significant CSP in the H47A mutant, but no other effect was observed.

The SH3 domains of both mutated construct displayed basically the same CSP pattern. Important perturbations arose in the RT loops in both mutants (^{98}RTETDL^{103}), being R98 notably more affected with a CSP 50% higher than the second largest. The nSrc loop was also affected to a lesser extent, specially V114, N118 and L118. Other isolated residues were also particularly perturbed, specially T88 in the hinge region, H125, G130, Y134, and A144 in the distal loop and β strands 3 and 4.

In conclusion, given the response of the highly polar RT loop to both substitutions, an underlying network of electrostatic interactions between the Unique and SH3 domain should not be ruled out. The CSP results for the IDR only confirm close contact between F54 and H47, but if histidines contribute to Unique domain compaction can not be confirmed without the corresponding ΔPRE experiments. In any case, it is tempting to speculate with the idea of having electrostatic-sensitive interactors embedded in the ensemble and, at the same time, participating in a network interactions between aromatics that constrains

the dynamics of the whole.

2.3.6 Other functional mutations: SH3 loops

Contribution statement: All mutant USH3 samples used in this section were produced by Farman Ali Khan during his internship in our group. I supervised him, I acquired the NMR spectra and analyzed the data.

Finally, in order to obtain complementary information about determinants in the IDR for the Unique:SH3 contacts, I decided to mutate key residues in the interacting *hot spots* detected in the scaffold: the RT, nSrc and distal loops. Being SH3 a small, folded domain, mutations in the middle of β strands 3 or 4 were likely to disrupt the core β sandwich motif, so the groove between the RT and nSrc loop was not explored.

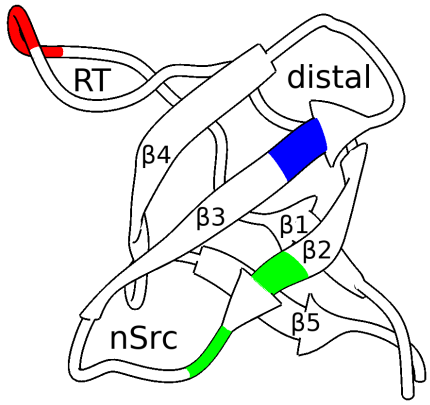


Figure 2.39: SH3 cartoon displaying the loop mutations

CSP versus the wild type USH3 were calculated for the following loop mutants:

- RT loop: key residues ^{98}RTE^{100} at the apex of the loop were changed to IHH. That particular sequence corresponds to that found in the SH3 domain of tyrosine kinase Hck, which has a RT more flexible than c-Src (Arold et al. 1998).
- nSrc loop: the most perturbed residues in this loop by removal of the IDR were V114 and N116 (sub-section 2.1.1), located at the end of β2 and the beginning of nSrc. Therefore, I decided to mutate ^{114}VNN^{116} to ANA.
- distal loop: H125, located at the end of β3 bordering the loop, has been shown to respond to the presence of the IDR. Also, as it happened with H25 and H47 in the Unique domain, H125 was observed to be sensitive to small changes, varying even

between identical samples (see figure 5.18 in the Appendix). Thus, it was mutated to alanine.

As expected, all three mutations induced relevant CSP along the whole SH3 domain (figure 2.40). The nSrc loop mutation particularly affected amino acids ^{131}QTGY^{134} located deep in the inter-loop groove, at the C-terminal end of β4, and marginally to R98 and T99 in the RT loop. H125A mutation induced considerable CSP in residues S104, K106, G108, R110, ^{112}QIVN^{115}, and E118, along the C-terminal end of the RT loop, β2, and the nSrc loop, specially. RTE to IHH mutation instead provoked spread, smaller CSP, being only noteworthy the perturbations on I135 in β4. Interestingly, residues T87 and F89, in the hinge region between the folded and disordered, were perturbed by nSrc and distal loop mutations, but not by the IHH substitution in the RT loop.

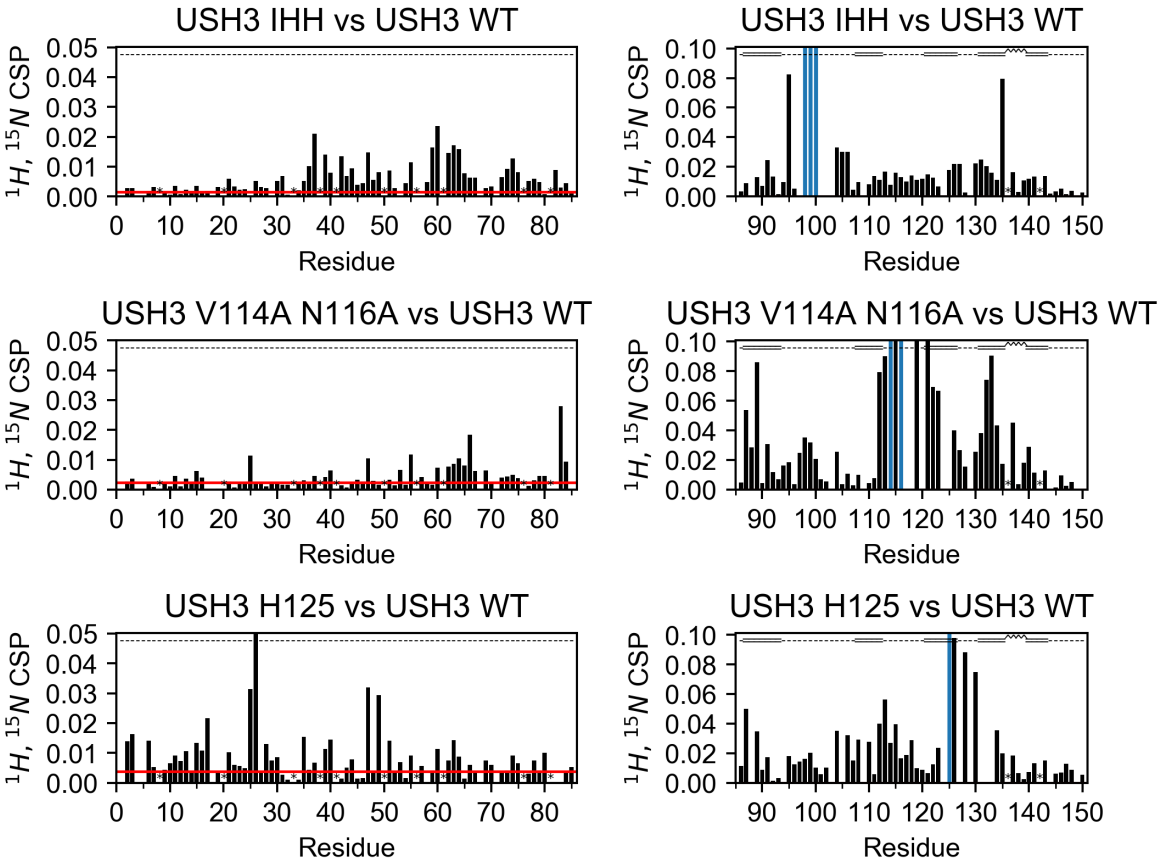


Figure 2.40: CSP of USH3 loop mutants vs USH3 WT. Mutations are indicated with blue lines. The red line represents a significance threshold defined in Methods and Materials. Notice that the CSP scale for the IDRs is half of that for the SH3 residues.

Regarding the CSP of the corresponding SH4 and Unique domains, interesting information arose. The RT and nSrc loops have been identified as *interaction hubs* which concentrate most of the short and long-range contacts with both the SH4 and Unique domains and

therefore were expected to display the largest changes. The RT loop mutant did in fact affect T37, T39, H47, A55, A59, ^{59}AE(P)KLFG^{65}, and ^{72}TVTS^{75}, all between those pointed out in previous sections to contact SH3. Curiously, no significant CSP were observed in the SH4 domain. In contrast, the nSrc loop mutant displayed more modest perturbations affecting H25 and H47 and residues in the ULBR: L53, A55, ^{60}E(P)KLFGGFNS^{70}. A83 in the very C-terminal end was the most perturbed signal. Only small CSP were observed in the SH4 domain.

Interestingly, it was H125A substitution in the distal loop the only modification that sparked changes in the SH4 domain, specially for S3, S6, R15, and S17. Also remarkably, H25 and H47 showed important differences, along with other mostly polar residues of the Unique domain: G26, T35, S39, K40, G49, S51, E60, L63, and T74. These results suggest that the set of IDR interactors dominated by electrostatic interactions is particularly sensitive to changes in the distal loop. The fact that H125 is also hypersensitive to environmental conditions and seems to be heavily coupled to H25 and H47 suggests an underlying regulatory mechanism.

Finally, I would like to remark again that the analysis of CSP in proteins containing IDRs can be deceptive, specially with limited data sets. Therefore, the unresponsiveness the SH4 domain for the RT and nSrc loop mutations should not be taken as an absolute negative proof. As commented in section 2.1, it may happen that interactions within the IDR re-arrange, resulting in null or low CSP, even if alternative contacts are indeed taking place.

2.3.7 Discussion

Along this section I search, find and test specific sequence determinants ruling both the intramolecular transient contacts leading to Unique domain pre-arrangement and the interactions between the disordered Unique domain and the scaffolding SH3 domain.

In the first place I focus on phenyalanine residues in the Unique domain, namely **F32, F54, F64 and F67**. The typical IDP/IDR depletion on aromatic residues, which in general favor hydrophobic collapse, and the conservation of patterns among SFKs suggested potential participation on either intra- or inter-domain interactions.

I calculate the ΔPRE profiles of four SH4-UD constructs, each one incorporating a single F#A substitution, and compare them to their corresponding wild type references. Results demonstrate that each one of the four phenylalanines contributes to transient long range contacts boosting compaction. Attending to the details, analysis of the individual

contributions permit to differentiate two distinct pairs: F32 and F54 contribute the most in the compaction of the middle section of the IDR, while F64 and F67, at the core of the ULBR, have a smaller although still relevant effect. The fact that these results were obtained in absence of the scaffold only but underlines the importance of these conserved aromatic residues in the pre-conditioning of the Unique domain. The subsequent CSP study of phenylalanine substitutions in USH3 constructs further confirms the divergence between F32/F54 and F64/F67. While mutation of the former affects amino acids 20 - 70 of the Unique domain, the latter only perturbs positions 60 - 85 at the C-terminal end. The more intense and localized effects of F64 and F67 mutations are consistent with the residual structure initially detected by RDCs in the region 55 - 70 (Pérez et al. 2009), since a more restrained environment may limit the reach of the perturbations, while structure disruption would them more noticeable.

Thus, the Unique domain is dynamically held together at least by an hydrophobic cluster formed by the two pairs of conserved phenylalanines.

The Unique:SH3 inter-domain short range contacts are likewise affected differently by both pairs of phenylalanines. While F32 and F54 mutations to alanine preferentially affect the distal loop and $\beta 2$, $\beta 3$, and $\beta 4$ strands, both F64A and F67A replacements mainly perturb the RT loop, and the nSrc and distal loops to a lesser extent. These results bring a more detailed image of the inter-domain interactions, being the ULBR the main competitor with SH4 for contacting the RT loop.

These conclusions are most relevant because they imply that the same sequence determinants code for different, although coupled, functions: Unique domain compaction of the and its interactions with the rigid scaffold. The fact that both aspects are driven by conserved aromatic residues is in agreement with the current knowledge of protein-protein interactions involving IDPs, and fuzzy complexes in particular. On the one hand, the particular balance between conformational freedom and pre-formed structure (this is, the $S_{conformational}$ aspect), so far thought to be case-specific, is known to be of paramount importance in IDP/IDR interactions, as reviewed in section 1.8 of the Introduction. In this particular case, the proline *code* leading to correlated local motions described in section 2.2 and the presence of conserved, order-promoting aromatics are, at least, some of the key elements determining the plasticity of the system.

On the other hand, there is the role of hydrophobic residues in protein-protein interactions. There are two important aspects to discuss in this sense. In the first place, although aromatics are not common in IDRs (Dunker et al. 2001), it has been observed that IDP/IDR interfaces are more hydrophobic than the average composition (Mészáros et al. 2011). Furthermore, aromatic residues are also associated to large-scale assemblies

involving IDRs (Wu & Fuxreiter 2016). In line with this, it has been proposed that IDRs have specifically evolved to provide much larger contact interfaces than globular proteins (Gunasekaran et al. 2003). Therefore, spaced but available aromatics partially stabilized in an hydrophobic cluster seems a nice solution to conceal a potential hydrophobic binding site without generating a problematic (i.e., aggregation-prone) localized apolar stretch.

Secondly, IDR-contacting regions in folded proteins use to be extended, hydrophobic-enriched shallow grooves, often surrounded by charged amino acids which can accelerate binding rates (Ganguly et al. 2012). This seems to be the case for the small and promiscuous SH3 domain. A look at the Coulombic surface potential of the scaffold confirms that the SH3 contact region fulfills that description, being neutral most of the grove between the RT and nSrc loops, extending to the distal, with small negatively charged patches at the top of the RT and towards the canonical PPII binding site (see figure 2.41).

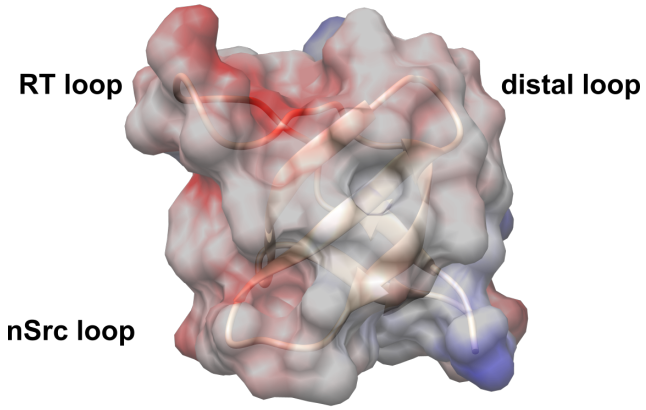


Figure 2.41: Coulombic potential of the SH3 domain over PDB:4HXJ. Calculated using the Coulombic surface coloring routine included in Chimera 1.11 (Pettersen et al. 2004).

The effect of Unique domain conformational restriction is then addressed by studying two SH4-UD constructs in which non-native disulfide bonds are created between distant points, C22-C72 and C27-C59, respectively. It is shown that the CSP versus the linear, wild type form are larger for the longer loop, probably due to the contribution of all four phenylalanines and their corresponding ring currents to the chemical shift. However, the interaction with isolated SH3 is observed to be larger for the most constrained loop, C27-C59. Remarkably, this agrees with the previous results that identify F32/F54 and F64/F67 as functionally different. Thus, enclosing the first pair would further enhance their compacting role, and restrict the SH4 excursions. At the same time, the second pair at the core of the ULBR would also be constrained, but remains outside the loop, unhindered for interaction with the SH3 loops. These results further highlight the functional importance of pre-structuring and dynamics of the IDR for its interaction with the

scaffold.

Next, I tackle the role of the two histidines present in the Unique domain, H25 and H47. These particular residues are extremely sensitive to small variations in the experimental conditions and add yet another layer of complexity to the USH3 fuzzy system. CSP studies in USH3 constructs reveal that both histidines are mutually sensitive to changes in each other. Most importantly, it is observed that both H#A substitutions mainly affect polar residues in the SH4 and Unique domains, and the RT loop in SH3. While the CSP within the IDR are harder to interpret due to its intrinsic dynamics, the effect induced in the RT loop (and specially R98) suggest that electrostatics also contribute to inter-domain contacts. This goes in line with the role of charged residues facilitating protein-protein interactions suggested by Ganguly et al. (2012) and commented earlier.

The role of electrostatics in IDR function and ensemble properties has been extensively studied and is thought to be one of the main determinants ruling the nature of the ensembles (Mao et al. 2013, Das et al. (2015)). Bridging polymer physics and electrostatics, Pappu and collaborators have approached this issue and found that net charge per residue and linear sequence distributions of oppositely charged amino acids modulate IDP conformational properties (Mao et al. 2010, Das & Pappu (2013)). The CIDER software (Holehouse et al. 2017) provides an exploratory analysis by classifying IDR sequences according to these parameters.

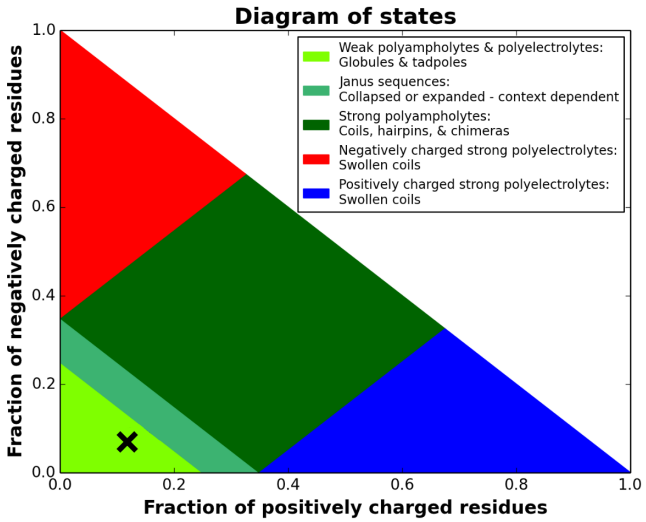


Figure 2.42: Das-Pappu diagram obtained using the CIDER software (Holehouse et al. 2017). The black X denotes the position of c-Src SH4-UD region.

The results for the c-Src N-terminal IDR (residues 1 - 85) shows that, having low fraction of charged residues and net charge per residue (both <0.25), this IDR classifies as a weak poly-electrolyte (slight excess of positive charges, due to the SH4 domain), but bordering

context dependency[7]. Thus, changes in the equilibrium between the protonated (positively charged) and deprotonated (neutral) forms of the histidine may alter the ensemble conformational properties.

In any case, an scenario in which histidines play a role in compaction by means of hydrophobicity or either electrostatics, and simultaneously affect the interactions between the IDR and its scaffold is plausible and merits further exploration. There is also possibility of feedback between Unique domain compaction, and therefore histidine side chain average environment, and protonation state, which would permit more sophisticate regulation mechanisms. I suggest to quantitatively measure, using NMR, the effects of pH and ionic strength on compaction and Unique:SH3 contacts as an opening wedge to this aspect.

Finally, I put the focus on the scaffolding element of the fuzzy complex, the SH3 domain. Since folded domains are more delicate to mutate than disordered regions, and most of the interactions concentrate in the RT, nSrc and distal loops, I individually mutate each one of them thus avoiding structure disruption.

The RT loop mutation to resemble that of c-Src's paralogue Hck (^{98}RTE^{100} to IHH) affects most of the Unique domain, but specially those residues responsible for contacts with the scaffold. Surprisingly, no significant CSP are observed the SH4 domain, although amino acids ^{98}RTE^{100} were specifically perturbed by addition of isolated SH4 domain to SH3 or SH4 deletions in USH3 forms (sub-sections 2.1.1, 2.1.2). However, as explained in section 2.1.2, the network of transient inter- and intra-domain contacts is intricate and dynamic. For example, T99 was observed to engage alternative interactions with the Unique domain in absence of SH4. On top of that, the RT loop is extensive and highly dynamic as well. The large CSP of I135 in β4 suggests a change in the accessibility to the deeper part of the groove between the RT and nSrc loops. In addition, signals from the hinge between the IDR and the scaffold are mostly unperturbed, discarding large changes in the conformational excursions of the IDR. Taken altogether, these results suggest rearrangement of pre-existing interactions rather than major loss of contacts.

Mutation of key residues V114 and N116 to alanine does induce some changes in those residues in the boundary between the Unique and SH3, but instead modest and localized CSP are observed between residues 55 - 67. This may be a sign of specificity between the nSrc loop and the ULBR in the Unique domain. Also interestingly, the mutation also provokes CSP along β4, the one most buried in the groove between the RT and nSrc loops

[7]In Holehouse et al. (2017), the authors state that "It is worth emphasizing that the boundary between R1 and R2 is rather *ad hoc*". Therefore, the sequence of SH4-UD may already fall in the *context dependent* category.

(see figure 2.39). Two non-exclusive mechanisms can explain that perturbation. On the one hand, in line with the model proposed by Cordier et al. (2000), it is possible that the structural consequences of the mutation propagate across the rigid β sheet core. On the other hand, as it may happen with the RT loop, altered nSrc loop dynamics can affect groove accessibility, thus redirecting preexistent nSrc-IDR contacts.

At last, the single H125A mutation in the distal loop is observed to specifically affect H25 and H47 in the Unique domain, suggesting that the set of histidines in USH3 is interrelated and maybe another functional element with an electrostatic component. Also interestingly, this mutation, although further away from the *hot spot* than the others, is adjacent to β 2. This explains how its effects spread from the end of the RT loop to the beginning of the nSrc (see figure 2.39), stressing the ability of the SH3 to transmit perturbations between distant regions and therefore its natural potential for allosterism.

The conclusion to remark here is that not only the intrinsically disordered partner of the fuzzy complex is important, but the scaffold is something else than a simple platform. I have shown that there also exist sequence determinants in the SH3 that affect the overall complex formed with the SH4 and Unique domains. Some, as those in the RT loop, are also involved in classic proline-rich substrate recognition (Feng et al. 1995) and in stabilizing the inactive form of c-Src (Erpel et al. 1995; Gonfloni et al. 1997), whereas others as those in the nSrc or distal loops, seem to be unrelated[8] or, at least, involves non-overlapping zones of SH3. Unsurprisingly, the complexity enclosed in such a modest domain (remember that c-Src's SH3 domain is **only 65 amino acids long**) is an object of debate (Mayer 2001).

Thus, it would not be surprising that the ability to entangle into fuzzy complexes is an unrecognized function of (at least some) SH3 domains. This idea opens the following questions:

- Are N-terminal fuzzy complexes involving SH4, Unique and SH3 domains a general feature among SFKs?
- How would this new functionality fit with the classic open-closed model of full-length c-Src?
- How could environmental cues be transmitted across the fuzzy interface?

These and other questions are the matter of the next and final result section.

[8] A future **Isothermal Calorimetry (ITC)** study is planned to test the possible effect of these mutations on the recognition of the VSL12 *PxxP* peptide.

2.4 Beyond Src

In this section I will expose four blocks of results that support the generality of the finding of the N-terminal fuzzy complex in Src among the SFK. The two first connect this newly found functional element with some of the established knowledge about c-Src: the canonical open-closed model for the full-length construct and the effect of Unique domain phosphorylation in the fuzzy complex.

The next two blocks show a co-evolutionary analysis of the SFK and distant homologues, and also experimental NMR results from the closest member to c-Src, Yes, which further support the generality question raised by sequence alignment.

2.4.1 The fuzzy complex model in the context of full-length Src

The canonical open-closed auto-inhibitory model of c-Src, reviewed in sub-section 1.3.6 of the Introduction, establishes that in the inactive, closed form, the SH3 domain packs against the 14 amino acid segment linking the SH2 and SH1 domains via its PPII recognition surface. These interactions can be inspected in detail in the high resolution X-ray model of the full-length protein, PDB:2SRC (Xu et al. 1997; Xu et al. 1999). Importantly, the SH4 and Unique domains are missing from this model since IDRs impede crystallization and are typically cut out from the constructs.

The SH2-SH1 linker does not contain the standard *(K/R)xxPxxP* class I motif, but has a glutamine instead of the second proline. However, it still adopts a poly-proline II helix conformation that is recognized by SH3. Additional contacts take place between the RT and nSrc loops and the SH1 domain N lobe. While nSrc loop interactions are modest, residues R98 and T99 extensively contact the $\beta 2 - \beta 3$ loop of SH1 (figure 2.43).

Figure 2.43: Interface between the SH3 (red) and SH1 (blue) domains with the SH2-SH1 linker (green) in A) closed c-Src (PDB:2SRC), and B) open c-Src (PDB:1Y57). H-bonds are highlighted in orange.

Thus, this close contact in the inactive form raises the question of whether the IDR-interacting interface of SH3, identified by using USH3 constructs lacking the SH2 and SH1 domains, would be accessible in the full-length protein or not. This issue is most important since it determines the biological relevance of the intramolecular fuzzy complex formed by the SH4, Unique, and SH3 domains in absence of the rest of the protein.

Hence, I plotted the PRE results from the USH3 construct with the paramagnetic probe at position 59 (the one showing the most intense PRE effects in the SH3 domain) over the surface of the closed full-length Src (figure 2.44).

It can be observed in the crystal structure that, although the RT and nSrc loops do contact the SH1 domain and clinch the SH2-SH1 linker in between, most of the IDR interface region is still exposed. The relative position of the SH3 domain makes the groove including $\beta 2$, $\beta 3$ and $\beta 4$ accessible even if residues in the RT and distal loops engage in inter-domain contacts. Additionally, the Unique domain pre-organization has been demonstrated to be independent of the scaffold. Therefore, it is reasonable to presume that the N-terminal fuzzy complex is conserved also in the full-length c-Src. There is also the possibility of additional contacts between the SH4 and/or Unique domains with the SH2 and/or SH1 domains which may further tune the conformational properties of the complex.

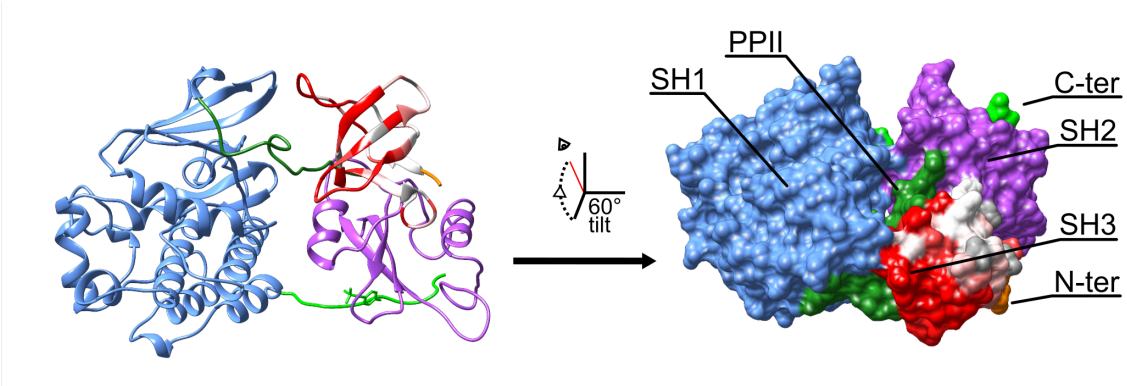


Figure 2.44: USH3 A59C PRE as shown in 2.19 plotted over the SH3 domain in c-Src structure PDB:2SRC.

Finally, c-Src switching to the open conformation supposes a major domain rearrangement, specially regarding the SH2 domain which translates ostensibly once unbound from its target phosphotyrosine in the catalytic SH1 domain (figure 1.8). However, the displacement of the SH3 domain is much smaller. In the PDB:1Y57 (Cowan-Jacob et al. 2005) crystal structure representing the active open c-Src, it is observed that the RT loop still contacts linker and SH1 residues, albeit the interactions are different (figure 2.43). Therefore, not only the closed but also the open active conformation, some important IDR-contacting SH3 sites will be involved in additional, potentially competing interactions.

2.4.2 Post-translational modifications affect transient contacts

As commented in the introduction, phosphorylation is the most common post-translational modification. It allows signaling networks to rapidly adapt to environmental changes in a reversible manner. Protein phosphorylation and intrinsic disorder are intimately correlated, since their enhanced exposure makes them an ideal target for kinases/phosphatases (Pejaver et al. 2014). Additionally, the inherent conformational plasticity of IDPs/IDRs confers them the ability to respond differently to a variety of external cues. Some of the possible outputs include recognition of the phosphorylated residue by new partners, changes in the aggregation state (even to the point of phase separation from the bulk), stabilization/destabilization of secondary structure elements or longer range contacts, and even trigger disorder-to-order transitions (Bah et al. 2015; Bah & Forman-Kay 2016).

More specifically, the phosphorylation of SFK Unique domains has been reviewed by our group in Amata et al. (2014). As SFKs are important nodes in several signaling networks, anomalous phosphorylation levels in their Unique domains are often associated with disease. Such are the cases of S69 in c-Src, Y32 in the p56 form of Lyn, Y34 in Fgr, Y46 in Frk, or S21 in Fyn.

Regarding c-Src, multiple phosphorylation sites has been reported or predicted in the IDR: S17, T37, S39, S43, S51, S69, S70, T72, T74, and S75. Our group has studied c-Src Unique domain phosphorylation both *in vitro*, and in cellular extracts and *X. alevis* oocytes using real-time NMR and Mass Spectrometry (Pérez et al. 2013; Amata et al. 2013). A complex cross-talking pattern was observed between different phosphorylation sites in those studies, thus suggesting a global functional significance.

Given the group's background on lipid binding, one of the most interesting sites is S17 in the C-terminal end of the SH4 domain. As part of an *RxxS* motif, that particular serine targeted by the cAMP-dependent PKA kinase. Although highlighted as functionally relevant (Obara 2004), its biological relevance remains obscure. However, Pérez et al. (2013) demonstrated that phosphorylation of S17 inhibits lipid binding by the SH4 domain in non-myristoylated SH4-UD constructs[9]. Therefore, this major player of the N-terminal fuzzy complex is a strong candidate to be a sensor able to modify the conformational ensemble and thus transduce information to the ordered cassette.

[9]In the same work, T37 and S75 were shown to be concomitantly phosphorylated by Cdk5 and to inhibit lipid binding by the ULBR in the Unique domain. However, since the *in vitro* reaction yields a mix of mono and di-phosphorylated products difficult to separate, I decided to use PKA, which quantitatively phosphorylates pS17.

Pérez et al. (2013) also reported CSP and RDC measurements of SH4-UD pS17 (figure 2.45). Those results indicated that no significant changes take place beyond the local environment. Only minor CSP and RDC variations were observed in amino acids 45 - 50.

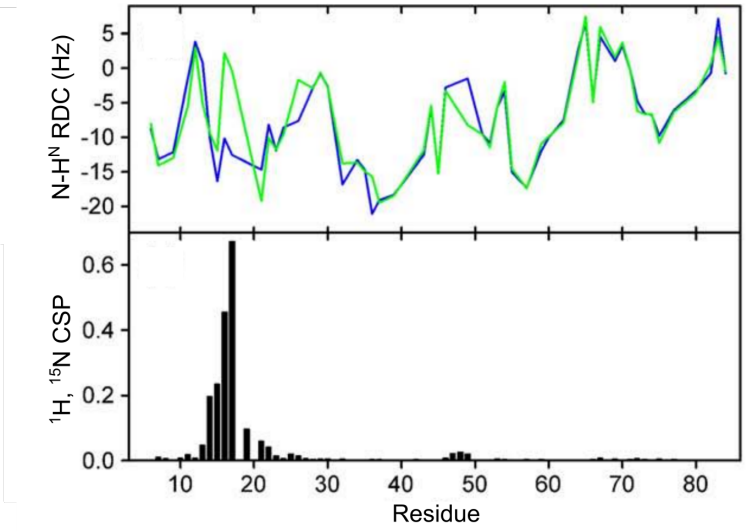


Figure 2.45: RDCs (top) and CSP (bottom) of SH4-UD pS17 (green line) vs the unmodified form (blue line). Adapted with permission from Pérez et al. (2009).

However, I intended to probe the possible effects on transient long-range contacts within the IDR, and therefore I measured PRE and calculated the corresponding ΔPRE of the SH4-UD pS17 phosphorylated *in vitro*. I decided to use the construct with the MTSL paramagnetic tag inserted in position 27 in order to have a reference to compare with, and place the *blind spot* induced by relaxation as close as possible to the modification site, to focus on distant effects. It should be stressed that intramolecular contacts measured as ΔPRE are those reported by the MTSL attached to C27 with the rest of the protein and therefore do not reflect direct interactions of the phosphoserine.

In vitro phosphorylation of S17 in the SH4-UD A27C construct was visible in the ^{1}H-^{15}N HSQC spectra, as the original signal was missing, and a new signal with the typical δ values of phosphoserines appeared instead, as shown in figure 2.46. The product was also checked with LC-MS, which confirmed total conversion.

The differences between the ΔPRE profiles (figures 2.47, and 5.19 in the Appendix) of the phosphorylated and unphosphorylated forms evidence that, although the overall shape of the profile was retained upon modification, there was a loss of long-range interactions in the N-terminal half of the IDR. The ΔPRE maximum corresponding to the ^{14}RRR^{16} pattern preceding S17 was significantly reduced. A minor local maximum centered on P20 was likewise affected. Most importantly, the large ΔPRE interaction observed for amino acids 37 - 44 also decreased between ~30 % to ~50 %. The oscillatory pattern between

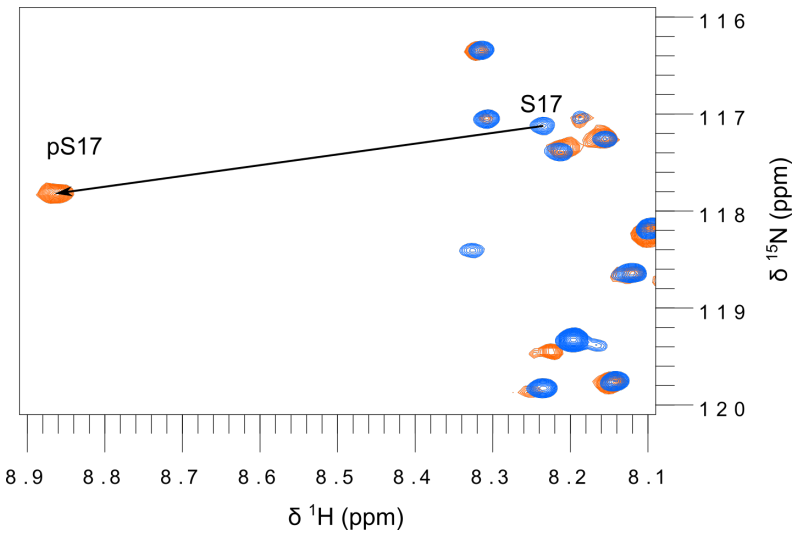


Figure 2.46: $^{1}H - ^{15}N$ SOFAST HMQC spectra of reduced (diamagnetic) SH4-UD A27C tagged with MTSL before (purple) and after (green) *in vitro* phosphorylation of S17.

50 - 70 was conserved, with a new maximum centered in the ULBR *FGGF* motif. Also remarkably, slightly enhanced contacts were observed for H47 and its neighbors and from S70 to G80.

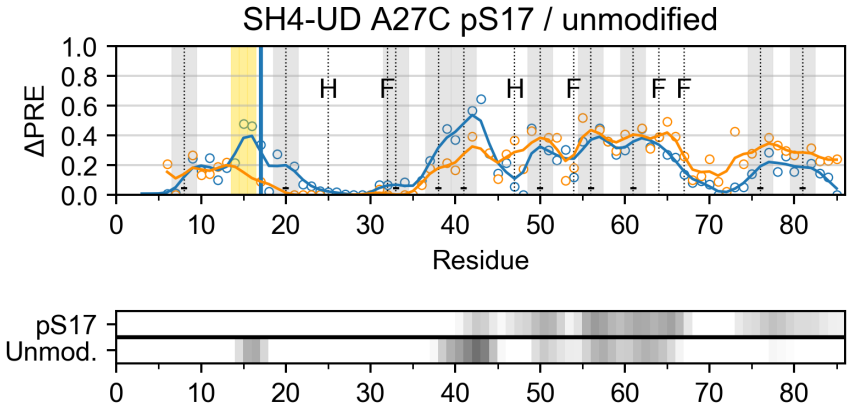


Figure 2.47: ΔPRE of SH4-UD A27C pS17 (orange) and the unmodified reference (blue). The vertical blue line indicates the position of S17. Phenylalanines and histidines positions are marked with vertical lines and text. Prolines are also marked, with a grey area delimiting positions ±1 around them. The ^{14}RRR^{16} motif in the SH4 domain is highlighted in yellow.

The most evident mechanism through which introduction of a phosphate group can perturb a conformational ensemble is the change on the native charge distribution of the sequence. Besides that, the oxygen atoms of the phosphate, carrying a partial negative charge, are able to establish new hydrogen bonds. There is also the steric hindrance, but the size of the phosphate moiety and the extended average conformations of the disordered region make it less relevant, specially regarding long-range contacts.

The concurrent contact decrease in the C-terminal end of the SH4 domain and the 37 -

44 region, both of which flank the paramagnetically induced *blind spot*, means that the ensemble shifts towards conformations in which the MTSL excursions between residues 15 - 45 are reduced. Since the PRE effects are averaged, that includes either spatial (larger average distance) or time (shorter average residence) restrains in the intramolecular contacts. This may be explained either by disruption of a preexisting contact between SH4 and 37 - 44 leading towards more extended conformations or, on the contrary, formation of a new interaction between those regions that leaves the MTSL unable to approach them. In any case, the fact that CSP are small and RDCs were only locally affected indicate that the interaction, either created or broken, must be transient and weak and does not provoke a major transition on the ensemble state or its dynamics. Nevertheless, as it was demonstrated for the F#A mutations in the Unique domain in sub-sections 2.3.2 and 2.3.3, small modifications in the degree of compaction of the IDR can affect the overall properties of the fuzzy complex.

The local ΔPRE changes observed in the ^{14}RRR^{16} arginine patch of the SH4 domain is probably due to the introduction of a negative charge right beside it. There is the possibility of salt bridge formation between the positively charged guanidino groups in the arginine side chains and the phosphate. This could induce a kink in the second half of the SH4 domain that could result in the loss of lipid binding capacity above-mentioned. The sharp, local increase in the RDC for these residues suggest local stiffening consequent to phosphorylation.

2.4.3 Coevolution analysis suggests that long range interactions are conserved in SFKs

Molecular co-evolution can be defined as:

> "[…] coordinated changes that occur in pairs of […] biomolecules, typically to maintain or refine functional interactions between those pairs" (Juan et al. 2013).

Its application to study interactions in proteins have their ground in the thermodynamic mutant cycle analysis (Hidalgo & MacKinnon 1995), used to experimentally determine specific energetic coupling between amino acids, either from the same of different molecules. With the advent of large scale sequence databases and the development of new statistical methods in bioinformatics, the method was generalized to permit complete mapping of energetic coupling based on evolutionary data (Lockless 1999). Co-evolutionary analysis

has been typically applied to study intra- (folding, structure) and intermolecular protein interactions (complex formation), but also allosterism (Süel et al. 2003). Thus, the tool has been incorporated to the structural biologist toolbox (Juan et al. 2013).

However, application of co-evolutionary methods to IDPs has been addressed with limited success (Jeong & Kim 2011), and mainly used as a mean to predict whether a sequence is intrinsically disordered or not. The main reason is that sequences encoding IDRs/IDPs typically display low conservation due to enhanced evolutionary rates in comparison with globular proteins. This phenomenon, known as *flexible disorder*, correlates with IDPs associated to cell signaling and interaction hubs, such as c-Src (Bellay et al. 2011). Still, there exist particular cases in which functionally relevant co-evolutionary couplings have been detected in proteins containing both intrinsically disordered and folded domains, as in the work of Chemes et al. (2012) with the E7 viral oncoprotein.

It was commented in the introduction (section 1.3) that the Unique domain is highly divergent among members of the SFK, while the SH# domains are mostly conserved in comparison. However, some regions as the flexible loops of the SH3 domains do show enhanced variability (Larson & Davidson 2000). Interestingly, in the context of globular proteins, a positive correlation between functional regions with increased mobility and higher number of co-evolutionary couplings has been described (Jeon et al. 2011). Besides, our group had already highlighted the occurrence of coupled mutations in the *FGGF* motif of the ULBR in the c-Src Unique domain, and the RT loop of SH3 in viral forms of Src with different transforming capacities (Maffei et al. 2015).

Therefore, I decided to search for co-evolutionary couplings, not only within the intrinsically disordered SH4 and Unique domains, but in the context of USH3. I chose the GREMLIN software, developed by Kamisetty et al. (2013), to do so. In short, this tool first uses the HHBlits algorithm (Remmert et al. 2011) and a pre-clustered version of the Uniprot database (The UniProt Consortium 2017) to find homologues, orthologues and/or paralogues related to a query sequence. Thereafter, a **Hidden Markov Model (HMM)** is used to create a multiple sequence alignment, from which sequences displaying long gaps are trimmed out so a solid sequence set is obtained. This methodology has been reported to provide ortholog-enriched sequence collections based on domain architecture (Hegyi & Gerstein 2001). This feature is particularly useful, since the more conserved SH4 and SH3 domains help to bypass the low conservation of the Unique domain. From there, pairwise evolutionary couplings are finally computed and classified according to a statistical score, so a complete contact map is obtained.

The search returned 151 sequences related to c-Src USH3[10], from which couplings were

[10]For the sake of space and because of the uselessness of just printing sequences in a format that can

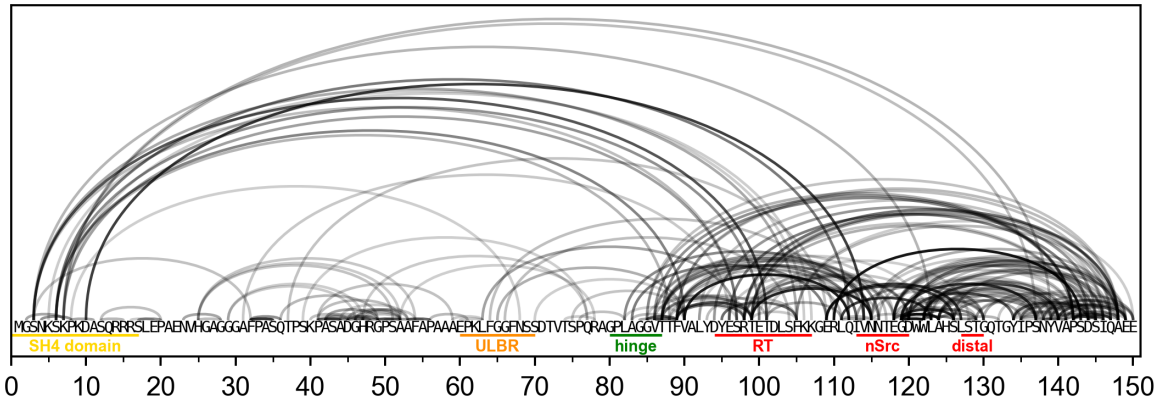


Figure 2.48: Co-evolutionary couplings within c-Src USH3 displayed in gray scale

calculated and displayed here in the form of an arch graph, the color of the arch between a residue pair representing the score in gray scale (see the Appendix for sequences and raw scores). Most of the highest scored pairs belonged to the SH3 domain, an expected result since the modular domain is well conserved. The RT and nSrc loops, and all β strands strongly correlate between them. The network of couplings involving the loops probably reflects the SH3 domain diversity regarding substrate specificity towards proline-rich motifs and the different adaptations they suffered along evolution.

Also, relevant correlations were also found between residues in the SH4 and SH3 domains. Most importantly, S3, ^{5}KSK^{8}, and D10 are connected with the RT and nSrc loops. From these, positions S3, S6 and K7 would correspond with the palmytoylation motif (*MGCxxSy*, where x is any amino acid and y is A, K or S) present in all SFKs but c-Src. Since those three regions are able to bind lipids, it is plausible that their co-evolution relates at least partially to membrane interaction by the fuzzy complex.

Lower score contacts are also predicted between the ULBR and both loops, supporting the observations from v-Src sequences. Nevertheless, the most evident feature within the Unique domain is the dense collection of couplings between residues in the region 40 - 55 despite the low sequence conservation (see figures in section 5.1 of the Appendix). These conserved contacts support the generality of the Unique domain pre-organization characterized in c-Src in sections 2.1 and 2.2.

These results suggest that the structural and functional connection between the SH4 and Unique domains and the scaffolding SH3 described in the previous sections - i.e. the fuzzy complex - is a general feature of the whole SFK family.

not be readily parsed, I refer the reader to the Supporting Information of Arbesú et al. (2017), where all sequences and GREMLIN scores can be found.

2.4.4 The case of Yes: Experimental evidences of a common mechanism

Contribution statement: All Yes USH3 samples used in this section were produced by Montserrat López during her M.Sc. at our group. I supervised her, acquired the NMR spectra and assigned the backbone (H_N, N, C', Cα and Cβ) of the protein.

Yes is the SFK member most closely related to c-Src (see figure 1.3). Their homology is such that they perform overlapping cellular functions (Hoey et al. 2000) and they can practically compensate the absence of each other (Stein et al. 1994). Most importantly, Yes is also associated with tumoral processes (Dubois et al. 2015), and in particular with colorectal cancer (Sancier et al. 2011).

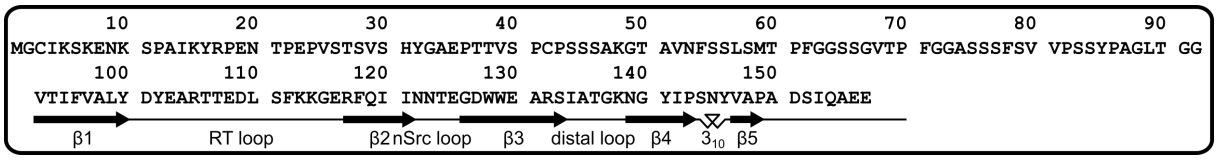


Figure 2.49: Sequence Human Yes USH3 (IDR on top, SH3 domain at the bottom), with the corresponding secondary structure elements.

The sequence divergence between both kinases is concentrated in the N-terminal IDR encompassing the SH4 and Unique domains. Regarding SH4, the most obvious difference is that Yes contains both myristoylation and palmytoylation motifs. Besides that, the first cluster of basic residues is also present in the first ten amino acids. These differences make a significant difference in trafficking and cellular location (Sato et al. 2009).

Nevertheless, a study using chimeric Src/Yes construct demonstrated that swapping both the SH4 and Unique domains from Yes to constitutionally active c-Src is sufficient to inhibit its transforming potential in chicken embryo fibroblasts (Summy et al. 2003), but not if the domains are individually exchanged. This suggests that the whole IDR works cooperatively, as shown here for the fuzzy complex in c-Src. Concerning the Unique domain, despite the low degree of conservation, all the aromatic residues present in c-Src and shown to induce IDR compaction also exist in Yes, as commented in sub-section 2.3.1. Additionally, the number of proline residues enclosed between conserved aromatics is similar (6 in c-Src, 5 in Yes) although only positions 41 and 61 (Yes numbering) are conserved.

All these hints suggested that Yes was a good candidate to experimentally test the generality of the N-terminal fuzzy complex found in c-Src. In order to do so, I intended to use

NMR as well. Since the sequence of Yes contains two native cysteine residues, C3 (palmitoylation site, SH4 domain) and C42 (Unique domain), and in order to avoid dimerization or scrambling, I mutated C3 to serine. After cloning, expression and purification of the Yes USH3 C3S construct, it was first necessary to assign the NMR signals of protein backbone. The standard battery of complementary experiments (HNCO, HNcaCO, HNCA, HNcoCA, HNCACB, and HNcoCACB) was acquired both at 278K and 298K. As in the case of c-Src, the reason is that solvent exposure of the IDR makes amide proton exchange with the bulk very efficient, and therefore at room temperature the signal intensity much smaller than for the folded moiety. Reducing the temperature slows down exchange while the IDR remains dynamic, so sharp peaks are obtained. Conveniently, at 278K signals from the SH3 domain are reduced due to slower tumbling of the folded domain because of increased solvent viscosity, thus alleviating signal overlap.

It is noteworthy that the spectra acquisition was done following a strategy combining **fast pulsing sequences** (**BEST-TROSY**, Solyom et al. (2013)) and **Non Uniform Sampling** (**NUS**, reviewed in Kazimierczuk & Orekhov (2015)). These relatively novel methods are being increasingly adopted by the NMR community as standard tools since they provide important advantages for specific purposes, such as acquisition of multidimensional spectra (specially $n > 2$). Without diving into technical details, the use of BEST-TROSY techniques is specially well suited for IDPs because of their short experimental time, sensitivity and, most importantly, enhanced spectral resolution. Fast and selective pulse schemes yielding transverse relaxation optimization, magnetization recovery, and selection of specific transitions exploiting H-N cross relaxation result in sharper signals and better resolution compared to classic pulse sequences. In the case of heavily crowded spectra such as those from IDPs, resolution can be determinant for a successful assignment.

NUS is instead a scheme for experiment acquisition, virtually applicable to any NMR experiment. It consists in the acquisition of only a fraction of the time domain points in the indirect dimension(s) and subsequent reconstruction of the whole frequency spectrum. The concept behind NUS it is that spectra are mostly devoid of signals and therefore the solution for the reconstruction is *sparse*, and can be obtained by methods other than the classic Fourier transform. This paradigm shift circumvents the classic limitation imposed by the Nyquist theorem and, depending on the case, permits to acquire only as little as <5 % of the total indirect data points. Since the experimental time scales exponentially with dimensionality, using NUS dramatically reduces the time spent acquiring spectra. Depending on the system to study, the time reduction can be traded for sensitivity by increasing the number of accumulations, or either resolution, by adding points in the indirect dimensions.

The assignment covered 83 % of the IDR due to proline abundance, which by lacking the amide proton break the *assignment walk*, and the presence of multiple stretches of consecutive serines resulting in non resolvable peaks despite the methodology used. As for the SH3 domain, 91 % of the residues were unambiguously assigned. Then, I acquired PRE experiments on the Yes USH3 C3S construct with a paramagnetic MTSL tag attached to the native C42, both at 278 K and 298 K.

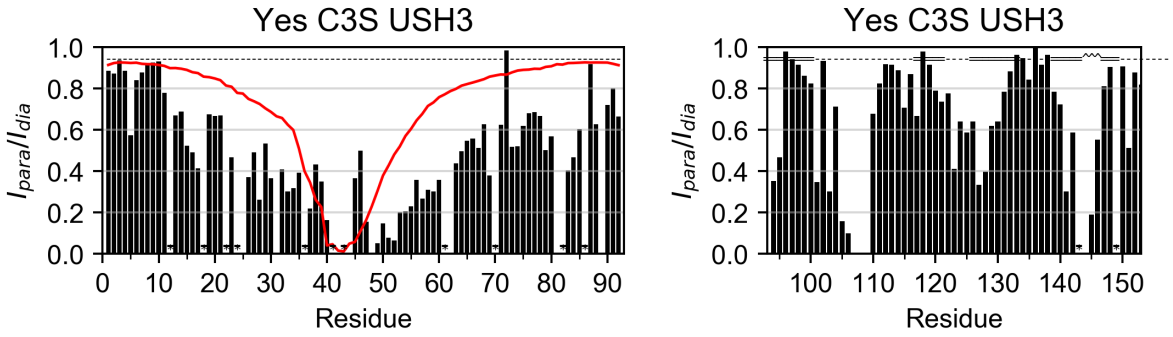


Figure 2.50: PRE profile of Yes USH3 with a MTSL paramagnetic tag at native C42.

The PRE and ΔPRE profiles (figures 2.50 and 2.51) of the SH4 and Unique domains vs the corresponding random coil simulation evidence significant deviations from the featureless model, as in the case of c-Src. While in the first ten residues of the SH4 domain only K5 showed enhanced relaxation, residues 10 - 20 displayed significant contacts. The next 20 residues up to the MTSL position were affected likewise. Residues 50 - 87 showed consistent long-range contacts with ΔPRE values ~0.4, being G72 the only exception. Smaller deviations were observed in the last 5 residues in the hinge region.

Also notably, the ΔPRE profile exhibits alternating maxima and minima. Although the pattern is not as evident as in c-Src, it should be stressed the only one data set is shown here, in comparison with the complementary constructs used in sub-section 2.2.2. From the eight maxima over the statistical threshold, five are associated to neighboring prolines (P18, P22/P24, P61, P70, and P82). The one centered around position 33 is close to Y32; to one around 52 corresponds to the conserved ^{52}VNFSS^{56}. Finally, the small maximum between 73 - 75 comprehends the center of the ^{71}FGGASSSFSV^{80} stretch, containing two phenylalanines.

The SH3 domain also showed important contacts partially resembling those observed in c-Src in the RT and nSrc loops. Although part of the RT loop could not be assigned, important PRE effects were observed in residues D101, E103, R105 and T106, indicating a strong interaction pattern. Interestingly, Y100, and Y102, which are intercalated between them and are involved in *PxxP* ligand recognition, remained mostly unaffected. The contacts in the nSrc loop were slightly less intense and extended to residues in the

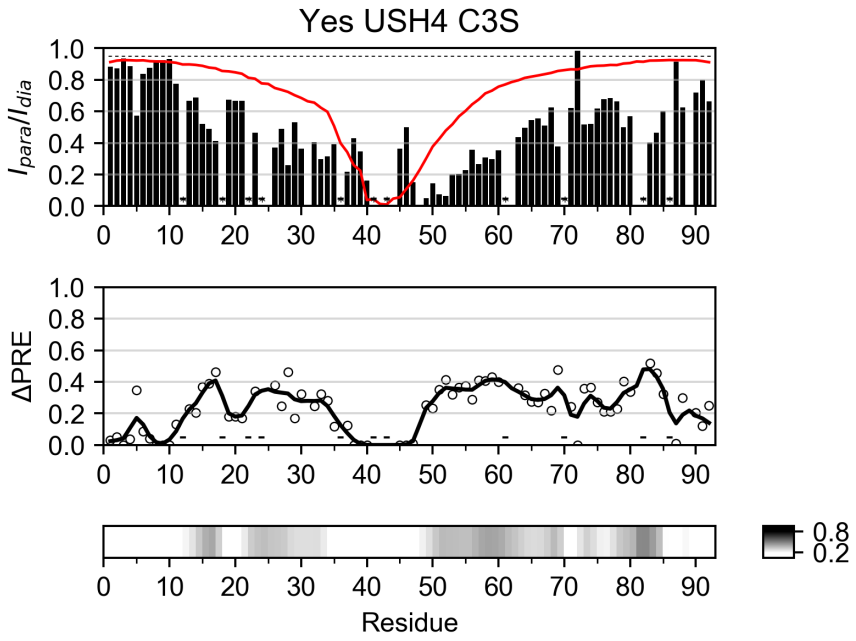


Figure 2.51: ΔPRE profile of Yes USH3 with a MTSL paramagnetic tag at C42.

beginning of $\beta 3$ (^{123}NTEGDW^{128}). In this case, W128 is involved in poly-proline motif recognition. The end of $\beta 4$ and the short 3_{10} helix connecting with $\beta 5$ also displayed significant PRE. The latter includes Y146, also involved in *PxxP* binding and not affected in c-Src. Additional contacts not observed in c-Src were detected at the beginning of $\beta 1$.

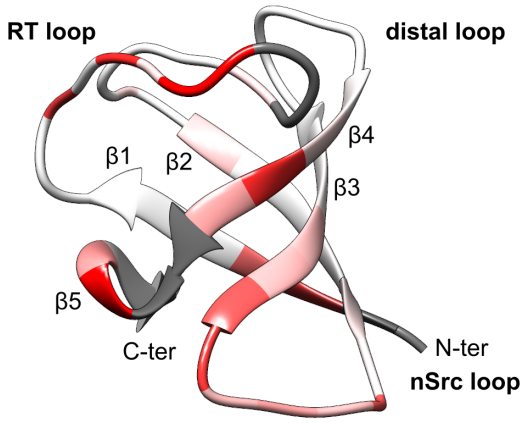


Figure 2.52: PRE profile of Yes USH3 C42-MTSL mapped over Yes SH3 crystal structure PDB:2HDA. Notice that the PPII binding area is now facing the reader.

The overall picture in comparison with the results obtained for c-Src is that the loops remain being *hot spots*, but instead of the groove between them, the IDR tends to preferentially contact part of the recognition surface for *PxxP* ligands. There exist two particular *PxP* motifs in the Unique domain of Yes: ^{22}PEP^{24}, and ^{41}PCP^{43}. Although none of them

is predicted to be a SH3 recognition site (Dinkel et al. 2016), whether the Unique domain can transiently adopt PPII conformations in these or other places and therefore interact with SH3 in a non-canonical fashion remains to be explored.

Besides the small differences in sequence between Yes or c-Src SH3 domains the fold architecture is perfectly conserved among all SFKs, but there is a significant difference unique to Yes. The beginning of the nSrc loop is particularly dynamic and the loop itself is stabilized in a different conformation from SH3 domains of all other SFKs (PDB:2HDA; Martín-García et al. 2007). The only differences in sequence are conservative, only two changes from hydrophobic to hydrophobic: ^{111}**L**QI**V**N^{115} in c-Src to ^{118}**F**QI**I**N^{122} in Yes. This difference may be related to the characteristic substrate recognition specificity of Yes.

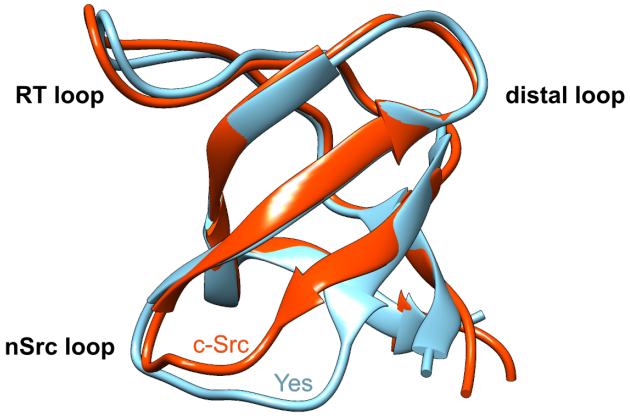


Figure 2.53: Comparison of Yes (blue) and c-Src (orange) SH3 domains (PDBs:2HDA;4HXJ)

In any case, these PRE and ΔPRE mappings are consistent with the fuzzy complex hypothesis. Future studies are planned to further characterize the Yes N-terminal region and explore its functionality. The results obtained from the c-Src fuzzy complex and the sequence analogies and divergences will be a useful guide to undertake this project.

2.4.5 Discussion

In the first sub-section I analyze how the PRE mapping for the IDR-SH3 interaction from section 2.2, fits in the full-length c-Src context. Based on the observation of X-ray structures of the closed (inactive) and open (active) conformations of c-Src, which originally miss the intrinsically disordered SH4 and Unique domains, it is safe to assume that the IDR-contacting region is mostly available. Despite the fact that both the RT and nSrc loops engage in interactions with the adjacent SH1 domain and the linker segment

between SH2 and SH1 in both situations, the groove between the loops always remains accessible.

These models, together with the SH3-independent Unique domain pre-organization described in section 2.2, suggest that the fuzzy complex is retained in the full-length form of the protein. However, the different availability of the RT and nSrc loops will surely affect the conformational ensemble properties of the fuzzy complex. As an example, the SH3 loop mutants studied in sub-section 2.3.6 showed that the unavailability of interactors promotes alternative contacts.

It is then plausible that, since the loops participate in different contacts with the rest of the protein in the open or closed forms, the average structural features are likewise modified upon c-Src activation. Thus, the fuzzy complex conformational ensemble (and therefore its functionality) may be tuned by the activation state of the complete protein.

Inversely, it could occur that the IDR affects the equilibrium between the open and closed conformations in the full-length protein by competing with the sites binding SH1 or the SH2-SH1 linker. Bernadó et al. (2008) demonstrated that the equilibrium of constitutionally active c-Src in solution[11] is heavily shifted towards the close inactive state (comprising an 85 % of the population) while the open active form is a minor specie. Therefore, only a minor fraction of c-Src is responsible for the catalytic activity, so even a small deviation in the equilibrium could be sufficient to exert an effect on function. A possible scenario may be the competition of the SH4 and Unique domains for the RT and nSrc loops.

Next, I study if phosphorylation can induce changes in the transient contacts within the IDR ensemble. In order to do so, I apply ΔPRE profiling to a SH4-UD construct that I could quantitatively phosporylate *in vitro* in a single, functionally characterized position, S17. While previous CSP and RDC experiments had only reported small and local changes, ΔPRE unveil broad longer range effects in the N-terminal half of the IDR including the SH4 domain and part of the Unique domain.

This is a proof of concept of the fuzzy complex conformational tunability by post-translational modifications, and a possible mechanism through which environmental sensitivity can be incorporated into it. In addition, it evidences correlation between the SH4 and Unique domains. So, although Unique domain pre-arrangement has been shown

[11] All this discussion has been done considering c-Src in solution. As noted in the introduction, c-Src is co-translationally myristoylated at the N-terminus of the SH4 domain in Nature, and so binds the inner face of the cellular membrane upon activation. Although unpublished results indicate that the fuzzy complex is retained in lipid bound myristoylated USH3, it is evident that the spatial restriction induced by the surface and further interactions with it will also have an effect on the native c-Src ensemble.

to be independent of the SH4 domain, it does not mean that both regions are functionally disconnected from each other. Considering the scaffolding role of the SH3 domain for SH4 and the Unique domain, and the complex network of weak contacts between all of them, these results reinforce a view of the complex as **a functional whole**, rather than chained modules.

In the following section I use a bioinformatic approach to probe the generality of the N-terminal intramolecular fuzzy complex of c-Src among related proteins. As already suggested by the sequence alignments, some sparse but important functional elements (i.e. the aromatic residues) seem to be conserved in other SFKs. Hence, in a more general and systematic approach, I perform a co-evolution analysis of the USH3 region of c-Src and 151 related sequences (orthologues, paralogues, etc.) and find significant coupling between distant regions. Interestingly, the search returned sequences belonging to a single Pfam family (Finn et al. 2016), SH3_1, from the 14 families forming the SH3 clan in the pre-clustered database.

Besides couplings arising within the folded and highly conserved β sandwich motif, many others relate the IDR with the RT and nSrc loops. Most importantly, meaningful co-evolutionary relationships are also found between the SH4 domain and the ULBR of the Unique domain with the RT and nSrc loops, in agreement with the experimental CSP and ΔPRE observations previously presented in this work. Additionally, the central region (40 - 55) of the Unique domain also displays a collection of shorter range couplings within itself, related to conserved interactions that participate in Unique domain pre-organization.

The power of co-evolutionary coupling analysis resides in the fact that it is not needed that the same kind of interaction between a pair applies for all the proteins analyzed, since the statistical relationships found are adaptative and thus *mechanism agnostic*. This means that, even if the hydrophobic contribution to pre-organization in c-Src would be unique to it or its close relatives, an alternative interaction mode in the same location - let it be a salt bridge, for example - causing the same effect in other proteins, is equally detected. This more abstract generality principle is hence more appropriate for the poorly conserved Unique domains than classic sequence alignment.

The agreement between the experimental data and the sequence-derived co-evolutionary connections so supports the idea of a fuzzy complex being a common arrangement between a subset of the SH3 domain universe and their adjacent N-terminal IDRs in Src-related proteins. Thus, the use of this tool helps to confirm a novel SH3 domain function, IDR binding. Identification of a mechanism involving such a wide variety of substrates in not trivial and endorses the applicability of this method to IDPs and IDRs.

In the final subsection, I complement the fuzzy model theoretical predictions for other SFKs with experimental observations on human Yes, the closest SFK to c-Src despite the sequence divergence between their Unique domains. After applying state-of-the-art NMR methodology in order to assign the signals from the analogous Yes USH3 construct, I perform PRE experiments and ΔPRE analysis in order to observe potential long-range contacts.

The ΔPRE plots of the IDR of Yes USH3 with a paramagnetic tag at C42 (Unique domain) evidence significant deviations from the random coil model and suggest IDR pre-organization. The intramolecular contacts are extensive and affect both the Unique and SH4 domains. The conservation of the aromatic residues characterized in section 2.3 for c-Src and the similar number of prolines are hints of a common mechanism for c-Src and Yes.

Important PRE effects are observed in the RT and nSrc loops of the SH3 domain, in line with those found in c-Src. Interestingly, the IDR contacts in Yes partially overlap with the *PxxP* binding site, shifting from the groove formed by β strands 3 and 4 that was contacted in c-Src. These results suggest that the RT and nSrc loop are the key elements for establishing a fuzzy IDR-SH3 interface, whereas the contribution from neighboring regions may be more system-specific.

In conclusion, these first experimental observations on Yes come to support the generality of N-terminal fuzzy complexes in SFKs, formed by intrinsically disordered but pre-organized SH4 and Unique domains delicately cast around SH3 domains through weak interactions mainly with their RT and nSrc loops. The RT and nSrc loops are also determinants for substrate recognition and specificity (Kay 2012) and so have a dual role. The differences in sequence among the IDRs and loops of different SFKs may determine the functionality of the system by influencing target specificity and not only activity directly.

2.5 Solid state NMR studies on the lipid-bound myristoylated fuzzy complex

Contribution statement: The work shown here is the result of a PhD stay at the NMR-supported structural biology group, led by Prof. Hartmut Oschkinat, at the Leibniz-Forschungsinstitut für Molekulare Pharmakologie (**FMP**) in Berlin, Germany. I acknowledge the support and the use of resources of INSTRUCT, a Landmark ESFRI project, which funded the stay. Sample optimization and preparation was done in collaboration with Laareb Irrem Mohammad; acquisition of the ssNMR experiments was done under the supervision and assistance of Wing Ying Chow and Michel-Andreas Geiger.

In this final section, I show the results obtained during my INSTRUCT funded 3-month stay at FMP-Berlin. There I applied DNP-enhanced solid state NMR to study a myristoylated USH3 construct in the presence of membrane-mimicking lipids. These preliminary studies permit the observation of the fuzzy complex in a native-like environment and pave the way to structural characterization of functionally relevant phenomena.

2.5.1 Review on previous results from myristoylated Src constructs

As commented in the introduction, *in vivo* c-Src is immediately myristoylated at G2 (SH4 domain) in a non-reversible manner after expression[12]. This modification is essential for its correct activity in several ways. Two pools of myristoylated c-Src exist in cells: cytoplasmic or membrane-bound. When in solution, c-Src remains in the perinuclear region in an inactive state. Upon activation, it is translocated to the inner face of the cellular membrane, where it develops its function in a tightly regulated manner (Patwardhan & Resh 2010). Both activation and transport are myristoylation dependent.

It has been already introduced that, although lipid binding is a common feature of all SFKs, the way c-Src attaches to the cellular membrane is unique and determines its targeting, activation and regulation. While combined myristoylation and palmitoylation result in a strong, persistent binding in most SFKs (Gottlieb-Abraham et al. 2016), only myristoylation and the electrostatic contribution of the poly-basic cluster in c-Src SH4 lead to a more labile association. Spatial activation and membrane delivery are also

[12]From now on, any myristoylated construct derived from c-Src will be preceded by the suffix **myr-**.

determined by these different binding modes, as demonstrated for the c-Src/Lyn pair by Sandilands et al. (2007).

In consequence, myristoylation and lipid binding are key functional elements of the N-terminal fuzzy complex here characterized that can not be overlooked. Characterization of lipid binding by myristoylated c-Src constructs including both the IDR (SH4 or SH4-UD) or the fuzzy complex (USH3) has been an important line of research in our group. Recent results from **Surface Plasmon Resonance (SPR)** and single-molecule Fluorescence Microscopy studies (A.-L. Le Roux, Castro, et al. 2016; A.-L. Le Roux, Busquets, et al. 2016) have revealed that, while most of myr-Src constructs participate in reversible binding, a minor fraction engages in persistent association. Analysis of the minor form showed that the persistently bound fraction is composed of self-associated myr-Src, mostly as a dimer. Though receptor-associated kinases are known to form functional dimers, it is not the case neither for c-Src nor any other SFK (Sicheri & Kuriyan 1997). This novel finding can therefore have major implications in the tightly regulated activity of c-Src *in vivo*.

Solution PRE NMR studies conducted with myrUSH3 AAA mutant in presence of paramagnetic lipid vesicles doped with 1 % **5-doxyl stearic acid (DSA)** (see figure 2.54) gave a valuable insight on the association. DSA is a spin probe which contains a nitroxide radical attached to the fatty acid chain, that inserts in the vesicle membrane at ~12 Åfrom the polar head and induces relaxation in the vicinity of the membrane surface (up to ~15 Å) (Chu et al. 2010). Since the radical in this paramagnetic tag is embedded in the membrane, it can not be reduced with a water soluble reagent. Therefore, the diamagnetic reference is the spectrum in presence of non-doped vesicles.

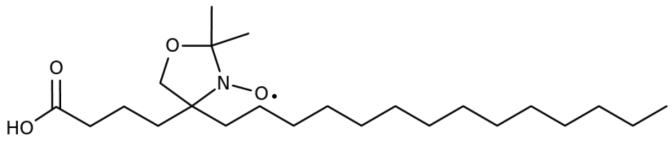


Figure 2.54: Structure of the 5-doxyl stearic acid spin probe. Notice the nitroxide radical attached to the fatty acid chain.

The PRE profile thus obtained highlights the regions binding strongly to lipids, which display lower I_{para}/I_{dia} ratios, as observed for the SH4 domain. However, as PRE is an extremely sensitive technique, also regions with low lipid affinity or brought closer to the surface are equally discernible. Such is the case of the ULBR which, although harboring the AAA mutation, still has residual binding capacity. In contrast, residues 20 - 40 in the Unique domain do not sense paramagnetic relaxation. Interestingly, although the RT and nSrc loops of the SH3 domain bind lipids (Pérez et al. 2013), the whole domain displays general strong PRE (~0.3 intensity ratio) comparable to that of the SH4 domain. With

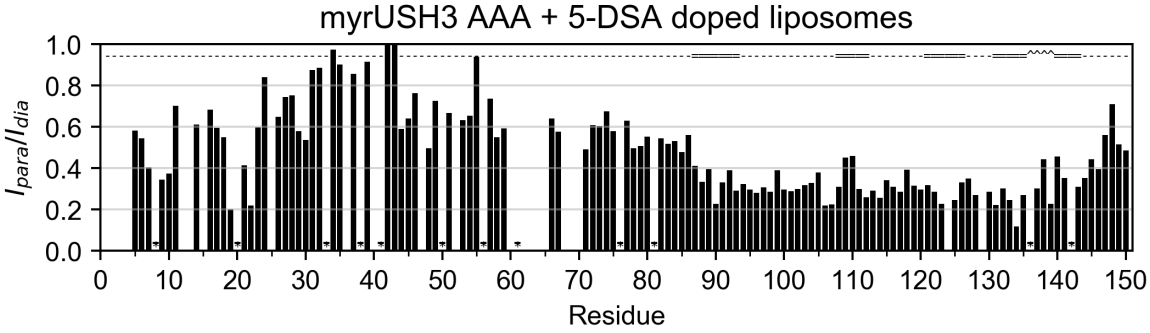


Figure 2.55: PRE profile of the myrUSH3 AAA construct using 5-DSA doped paramagnetic vesicles at 298 K.

the fuzzy complex in mind, these results suggests that in the lipid bound myrUSH3 AAA construct, the SH3 domain is retained in the vicinity of the surface while it is still free to tumble locally.

The image provided by this experiment illustrates a characteristic of the membrane-bound fuzzy complex: *conformational heterogeneity*. The myristoylated SH4 domain is strongly anchored to the membrane, while the Unique domain can adopt a variety of conformations around the SH3 domain, leaving the latter trapped close to the membrane but able to adopt different orientations. The network of transient contacts within the IDR is likely to be retained thus partially restricting conformational excursions.

Thus, we decided to start the structural characterization of these minor self-associated forms, working with myrUSH3 constructs attached to lipid vesicles as membrane models. Due to the particular challenges posed by these systems, which will be explained in the following, we used **Dynamic Nuclear Polarization (DNP) solid-state Nuclear Magnetic Resonance (ssNMR)**, also briefly introduced.

2.5.2 ssNMR, DNP and membrane-bound proteins

While solution NMR is an exceptional tool for structural studies, it has a fundamental limitation: slow or null molecular tumbling. The frequency of molecular tumbling motions affect transverse relaxation by favoring spin-spin interactions. Faster transverse relaxation leads to line broadening and ultimately, signals become undetectable. This is most recognizably reflected as an upper limit in molecular size. With increasing dimensions, the tumbling motions of the molecule become increasingly slower and relaxation grows faster. However, it also applies to non-soluble membrane proteins that can only be isolated embedded in hydrophobic supports, such as membranes. Therefore, although

some ingenious tricks have been developed and pushed the limit (Foster et al. 2007; Frueh et al. 2013), some systems are not accessible neither to solution NMR, nor to X-ray crystallography if they fail to crystallize.

Since only the myristoyl moiety inserts in the lipid layer, the SH4 and unique domains are disordered and the SH3 domain is small, the particular dynamics of the system make it observable by solution NMR even if bound to large, slowly tumbling lipid models. However, the self-associated minor species has only be observed in immobilized lipid supports after washing the labile species (A.-L. Le Roux, Castro, et al. 2016; A.-L. Le Roux, Busquets, et al. 2016), so we turned to a technique that could allow us to focus on the membrane-bound forms: **solid state NMR**[13].

ssNMR is the implementation of Nuclear Magnetic Resonance to study solid samples in which molecules do not tumble isotropically (or nearly) as they do in solution. This simple difference makes detectable the anisotropic components of the interactions between nuclear spins, electrons, and the magnetic field, which are otherwise averaged if the molecule tumbles freely, and lead to extremely broadened signals.

This draws some technical differences compared with solution-NMR, most remarkably the use of **Magic Angle Spinning** (**MAS**, Hennel & Klinowski (2005)) in order to suppress the anisotropic contributions by eliminating the orientation dependence. The rotor containing the sample is spun at a specific angle from the magnetic field ($\arctan\sqrt{2} = 54.74°$) at frequencies close to that of the interactions to suppress. Since dipolar couplings between ^{1}H nuclei require very fast spinning to be canceled (>100 000 Hz, due to their large gyromagnetic ratio, $\gamma_{^1H}$), ^{13}C detection is the typical approach for biomolecules. Notwithstanding, ultrafast MAS for ^{1}H detection is a growing field in the latest years, able to yield signals with line widths not far from those obtained in solution NMR (Zhang et al. 2017).

High power decoupling pulses are also usually employed in combination with MAS to suppress scalar couplings. Other common approaches in MAS ssNMR include sample deuteration, which reduces the network of ^{1}H-^{1}H dipolar couplings, and selective isotopic labeling schemes, which reduce spectral complexity.

MAS ssNMR also allows the user to selectively reintroduce some of the anisotropic interactions. Dipolar coupling is the standard method of transferring and manipulating magnetization, instead of scalar coupling as it is done in solution NMR. The currently available ssNMR toolbox developed over the years gives access to structural information

[13]ssNMR is a field way too vast and complex to write here a complete and rigorous presentation and, additionally, it is not my area of expertise. For these reasons, I refer the reader to Duer (2004) as an introductory lecture on the subject for those with a previous background in solution NMR.

allowing for protein backbone assignment and structure resolution (Castellani et al. 2002) of membrane proteins, fibrils, or large multi-protein assemblies among other challenging systems. Regarding the system studied here, MAS ssNMR has been successfully applied to folded domains of similar size anchored to different lipid models, including bicelles and vesicles (Nomura et al. 2014).

A particular challenge for these ssNMR experiments on myrUSH3 was sensitivity. First, as it will be explained in the next section, samples are difficult to obtain in large amounts. Secondly, in order to mimic a native-like environment, the real sample amount in the rotor will depend on how much protein attaches to the lipid support[14]. Finally, the persistent assemblies are minor species, therefore the number of associated molecules will be extremely low.

As a reminder, due to the physical principles ruling the interactions between the spins and the magnetic field, NMR is an intrinsically low-sensitivity method. The transitions from which NMR signals arise depend on the gap between spin energy levels, ΔE. For a single spin with $S = 1/2$ under the influence of a magnetic field depends of strength, B_0:

$$\Delta E = -\hbar\gamma_{spin}B_0, \quad (13)$$

where $\hbar$ is the Planck constant and γ_{spin} is the gyromagnetic ratio of the spin. For an ensemble of spins, the Boltzmann distribution permits to estimate the ratio between excited (N) and basal state populations (N_0), which accounts for the degree of polarization, i.e., the magnetization that is manipulated and ultimately detected as an NMR signal:

$$\frac{N}{N_0} \propto \exp(\frac{-\Delta E}{k_B T}), \quad (14)$$

where k_B is the Boltzmann constant, and T is the temperature. In the case of ^{1}H, the most common nucleus with higher γ, polarization under a field of 14 T (600 MHz ^{1}H Larmor frequency) at 80 K is a mere 0.02 % (see figure 2.56). The largest static magnetic fields currently available for NMR are around 28 T (a factor of 2); measuring at lower T is neither practical nor efficient enough; and γ_{spin} is an intrinsic property of the spin being observed.

To overcome the sample limitation by boosting the NMR signal, we decided to use **DNP enhanced MAS ssNMR**. Dynamic Nuclear Polarization is the process of transferring

[14]In ssNMR, the rotor is usually filled with protein sample exclusively, if possible: micro-powder, aggregate, fibrils, etc. In our case, most of the rotor was be full with lipid vesicles.

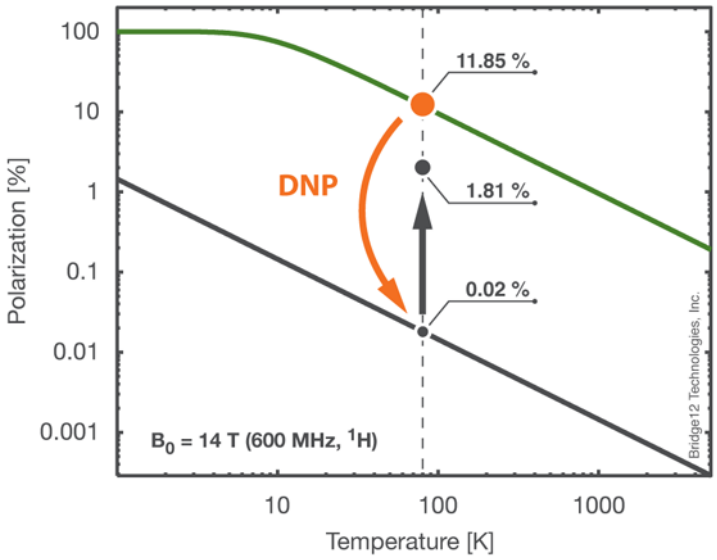


Figure 2.56: Polarization vs temperature curves at $B_0 = 14$ T for electron (green line) and ^{1}H (black line). The practical DNP polarization gain for ^{1}H is indicated with a black arrow. Courtesy of Bridge12 Technologies, Inc.

magnetization from a highly polarized electron spin reservoir to the nuclei of interest. Since the electron gyromagnetic ratio, γ_{e^-}, is 660 times that of ^{1}H, its polarization level increases proportionally, being 4 % in the same example conditions above mentioned.

Although proposed by Overhauser (1953) and demonstrated experimentally by Carver & Slichter (1953), DNP remained in the realm of solid-state physics until the early 1990s, as related by Slichter (2014). The reason is that polarization transfer to nuclei relies on **microwave (MW)** radiation to saturate electron spin transitions. It was not until the early 1990s that adequate continuous wave MW sources to induce DNP at high magnetic fields were developed (Becerra et al. 1993). At the same time, progress was made to obtain stable radicals as unpaired electron sources. These and other advances finally made possible the application of DNP to ssNMR. Although not an standard, DNP MAS ssNMR systems are commercially available (see figure 2.57), and the technique has found a niche in structural biology (Akbey & Oschkinat 2016) due to its unique capacity to study low concentration biomolecules and/or drastically reduce experimental time.

The instrument available at FMP-Berlin is a Bruker 9.4 T (400 MHz ^{1}H Larmor frequency) wide bore spectrometer coupled to a continuous wave cyclotron operating at 9.65 T, able to generate stable 263.6 GHz microwaves with a power of 25 W. The 3.2 mm MAS probe used is connected to a temperature control unit able to keep the sample at cryogenic values (80 - 170 K) while it spins at rates up to 15 000 Hz.

Among a variety of DNP mechanisms, this particular set-up makes use of **Cross-Effect**

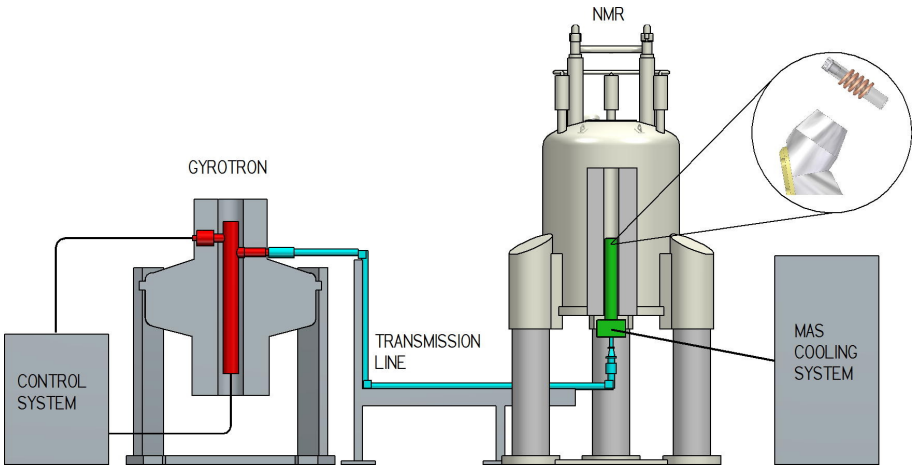


Figure 2.57: DNP MAS solid state NMR instrument scheme. Reproduced with permission from Rosay et al. (2016)

(CE) DNP (Mentink-Vigier et al. 2012). It consists on a three spin system (two electrons and one nucleus) for generating nuclear hyperpolarization, which is then transferred to other nuclei via spin diffusion across a homogeneous glass matrix. Specific water soluble biradicals exist for this purpose (Hu et al. 2004), with the advantage that CE only requires low radical concentrations (in the order of μM). The specifics for sample preparation are discussed in the following subsection. Any reader interested in further technical details on DNP MAS ssNMR and its application to the study of biomolecules is referred to Akbey & Oschkinat (2016).

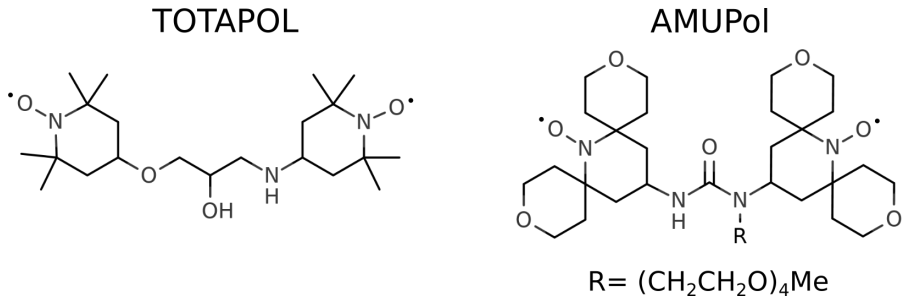


Figure 2.58: Structures of TOTAPOL and AMUPol biradicals

An important disadvantage in DNP-enhanced ssNMR coming from the combined use of cryogenic temperatures and radicals that induce PRE in their vicinity is line broadening, which limits spectral resolution (Barnes et al. 2010). Adequately enough, the ssNMR group at FMP had experience on DNP-enhanced ssNMR of SH3 domains, and knew the system's limitations beforehand.

Another important aspect to account is the above-mentioned conformational heterogeneity of the myrUSH3 construct. Whereas in solution NMR the conformational ensemble is dynamic and continually interconverts, in DNP conditions the sampling is frozen (the-

oretically, at the glass transition temperature, at around -80 °C, Hu & Tycko (2010)). Therefore, the observed signals come from a collection of static conformations. The lack of fast dynamics makes that, instead of detecting the average δ values, contributions from all conformations are equally observed and hence provoke further line broadening.

Nonetheless and in spite of the complexity that stems from conformational heterogeneity, ssNMR has been applied to the study of non-aggregated IDPs and IDRs (Zhong et al. 2007, Ahmed et al. (2009)) and, specially DNP thanks to its capacity to amplify weak signals from minor species.

2.5.3 Obtention of isotopically labeled, lipid-bound, myristoylated samples

The expression and purification of native-like myristoylated samples is far from trivial. Great effort has been dedicated by our group and collaborators to obtain myristoylated samples of the SH4-UD and USH3 constructs of c-Src at a scale and quality sufficient for structural characterization (Flamm et al. 2015). This is most critical for NMR studies, for which large amounts of isotopically labeled sample is required. Besides, the preparation of myristoylated samples suitable for ssNMR, and particularly its DNP-enhanced variant, was novel to our group and, to my knowledge, no protocols were available.

For these reasons, although the protocols are summarized in the Methods & Materials section, I will review here the particular problems faced during sample preparation and the contributions I made to solve the issues.

Expression

Recombinant USH3 over-expression in *E. Coli* is well established and optimized, so good yields are routinely obtained. The limitation of this approach is that bacteria lack the molecular machinery for myristoylation, so the fatty acid has to be introduced in some other way. In our group, the chosen method is the co-expression of the USH3 construct of interest with yeast N-myristoyl transferase (**NMT**). Using a bi-cystronic vector, we over-express both the substrate and the enzyme, the latter in a smaller amount, upon induction with IPTG. Addition of sodium myristate dissolved with BSA to facilitate uptake by bacteria allows concurrent myristoylation and co-expression.

As discussed in Flamm et al. (2015), an important side reaction is the generation of the N-lauroylated protein substrate. The byproduct is undesired because the shorter lauroyl

moiety (12 vs 14 carbon atoms) confers different, non-natural lipid binding properties to the protein construct. Besides, separation by size exclusion and reverse-phase chromatography were found to be problematic and inefficient. N-lauroylation is due to β-oxidation of the myristic acid in the medium by bacteria, as an alternative energy source, and the lack of NMT specificity between the 14 carbon atoms long chain of myristic acid and shorter species.

The problem had been partially solved by Flamm et al. (2015) by:

1. Using the Marley strategy for cell culture and limiting growth: Bacteria are growth in rich medium to a limited extent (O.D. ~ 0.5 and always < 0.7) and then transferred to labeled medium only for expression of ^{15}N labeled samples. This reduces starvation and limits β-oxidation.

2. Adding palmitic acid as a *lipid bait*: β-oxidation of palmitic acid leads to myristic acid, replenishing the reagent in the culture. Additionally, NMT can not incorporate the longer fatty acid chain (16 carbon atoms) and therefore side products are not generated.

Still, the extent of lauroylation is specially critical when expressing ^{13}C isotopically labeled samples. Since nitrogen and carbon sources need to be controlled, so called *minimal media* are used to culture the bacteria. These media are basically buffered solutions of salts and vitamins to which isotopically labeled carbon and/or nitrogen sources are supplied. For uniform carbon labeling, the most commonly used source is fully labeled D-glucose ($^{13}C_6$). Being the only energy source in the medium, it is metabolized by the bacteria to synthesize amino acids and thus and incorporated in the over-expressed protein. Since ^{13}C labeled reagents are expensive, a minimal but sufficient amount is typically added[15]. Therefore, when culturing bacteria in unlabeled rich medium the energy derived from carbon hydrates is virtually unlimited in the timescale of growth and expression, but when working on minimal media, budget economy imposes a limit. As a consequence, the alternative carbon and energy source in the medium for myristoylation (the lipid moiety), turns even more attractive for bacteria.

My solution consisted on reducing as much as possible the amount of lipids added to the medium to not induce β-oxidation, passing from 200 mM of 1:1 palmitic and myristic acid in the standard protocol to 50 mM. Expression tests showed that N-lauroylation was suppressed, while good yields were still obtained.

[15]The approximate cost of $^{13}C_6$ D-glucose is around 200 €/g. In a standard protocol, 3 grams are added per liter of culture, and depending on the protein to be expressed, culture volumes for a sample batch may range from 1 to 4 liters.

Although partial or total deuteration is a common practice in ssNMR (Akbey & Oschkinat 2016), the delicate balance and control needed for satisfactory expression levels led us to produce only fully protonated samples, since ^{2}D cultures require D_2O media, which perturbs bacterial growth.

Purification

The next bottleneck was purification. In the first place, protein stability is a major issue, since IDPs are typically sensitive to degradation. Our experience showed that the myristoylated forms are even more degradation prone than the unmodified forms. The problem was solved by strictly working in cold conditions after cell lysis and minimizing the protocol time. Additionally, we decided to start these structural studies with the myristoylated form of the USH3 ^{63}LFG^{65} to AAA mutant described in section 2.3.2. This mutation removes a major cleavage site as detected by MS, making the corresponding myrUSH3 resistant to degradation and thus a useful *work horse*.

Another issue is the peculiar physicochemical properties of the myristoylated constructs. The fatty acid moiety boost the affinity for lipids which, in some cases, can be problematic. Flamm et al. (2015) took profit of the membrane binding properties of the myristoylated constructs to separate them from the cell lysate. It was found that the fraction that remained in solution upon mechanical lysis and ultracentrifugation was partially bound to the NMT enzyme and problematic to purify. However, a significant, pure fraction was observed to bind the pellet. Therefore, they re-solubilized the myristoylated protein by washing the membrane pellet with 1 % v/v Triton X100, a common detergent in molecular biology.

Further removal of the detergent is mandatory, since it would disturb lipid binding (myrUSH3 is trapped in Triton X100 micelles) and disrupt any lipid model used. Therefore, during the subsequent affinity purification step with a Ni-NTA resin cartridge[16], the coordinated protein was washed and eluted with detergent-free buffers. However, elution was found to be problematic, and required doubling the standard amount of imidazole (from 400 mM to 800 mM) and using as much as 10 column volumes of elution buffer. Since elution was less problematic with the AAA mutant of myrUSH3 (with the ULBR inhibited for membrane binding), I speculated that it was a matter of high local concentration inducing a hydrophobic aggregate in the resin surface.

Hence, I decided to add 0.05 % of Triton X100 the elution buffer, which was shown to be

[16] A His-tag was present at the C-terminus of each myrUSH3 construct, see Methods & materials section.

sufficient for complete elution using standard volumes and imidazole concentration (see figure 2.59). After elution, I further diluted the sample with an equivalent volume of detergent and imidazole-free buffer, lowering Triton concentration to 0.025 % w/v. Since the Critical Micelle Concentration (CMC) for Triton X100 is 0.02 % w/v (OpenWetWare 2017), the further dilution due to sample injection in the size-exclusion chromatography column was enough to reach sub-CMC detergent concentration. Therefore, the Triton X100 molecules no longer formed micelles and eluted apart from the myrUSH3 sample.

It is remarkable that myrUSH3 constructs remain soluble after detergent removal and can be later concentrated at least up to 200 μM without signs of aggregation whatsoever.

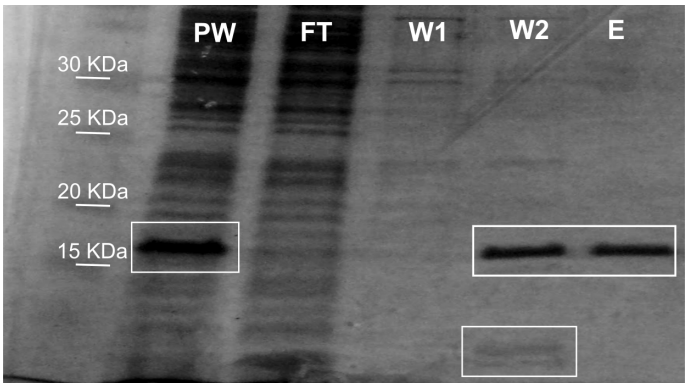


Figure 2.59: SDS-PAGE gel from myrUSH3 WT purification. **PW**: Pellet wash with Triton before affinity purification; **FT**: Flow-through from Ni-NTA affinity cartidge; **W1**: Wash with 10 mM imidazole; **W2**: Wash with 200 mM imidazole; **E**: Elution with 400 mM imidazole and 0.05 % Triton. Notice the degradation in W2. Non-degraded protein loss in W2 was latter reduced using 100 mM imidazole. See Methods and materials for further details on buffer composition.

Large unilamellar vesicles as lipid models for ssNMR samples

Large Unilamellar Vesicles (LUVs) are widely used lipid models used to study membrane-associated proteins (Szoka & Papahadjopoulos 1980). They are composed of a single, spherical bilayer formed by the lipid of choice. In our case, since c-Src shows higher affinity for negatively charged than neutral lipids due to the contribution of the SH4 basic patches, we used a 1:1 mix of **1,2-dimyristoyl-*sn*-glycero-3-phosphorylcholine** (**DMPC**, neutral) and **1,2-dimyristoyl-*sn*-glycero-3-phosphorylglycerol** (**DMPG**, negatively charged) (see figure 2.60). I chose LUVs over larger multilamellar bicelles because they offer maximum homogeneous surface for the same amount of lipids, but are still large enough to be sedimented by ultracentrifugation and packed in a ssNMR rotor.

Lipid vesicles are easily by hydrating a dry lipid film previously deposited in a glass flask by rotaevaporation of the organic solvents. Afterwards, the solution is extruded several times (18-20 passes) though a fixed-size filter and a nearly monodisperse distribution of

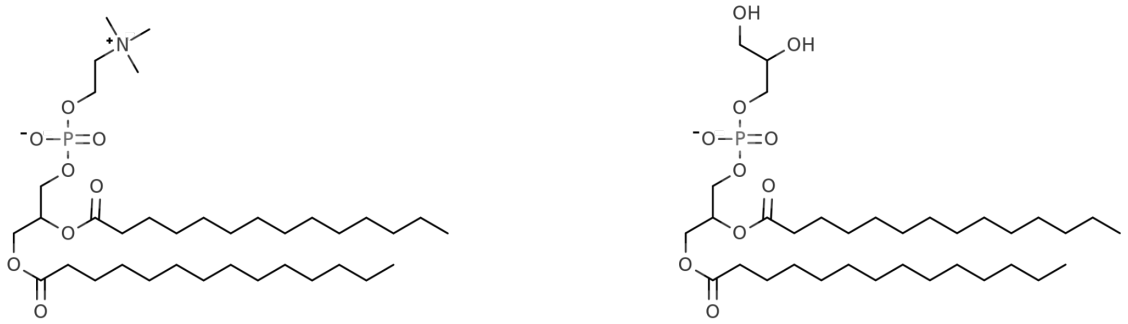


Figure 2.60: Structures of DMPC and DMPG. Notice that DMPC is a zwitterion, whereas DMPG is negatively charged.

vesicles of diameter close to the pore size is obtained (Mui et al. 2003). In this case, I used a 200 nm filter, which led to a sharp vesicle size distribution centered at 152 nm, as checked by **Dynamic Light Scattering (DLS)**. DMPC/DMPG vesicles in neutral pH phosphate buffer solution are stable for weeks kept at 4 °C, not precipitating neither showing aggregation (see figure 2.61).

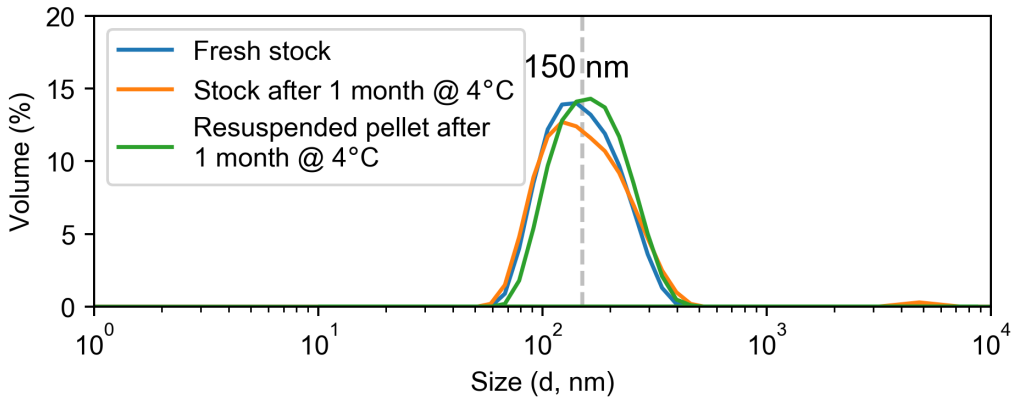


Figure 2.61: DLS stability test for the DMPC/DMPG LUV stock and pellet.

Once the lipid support was ready, the next step consisted on populating its surface with the myristoylated, isotopically labeled protein. The lipid and protein stock solutions were mixed and incubated for 90 minutes at 37 °C while shaking. The temperature was above the DMPC/DMPG phase transition temperature to facilitate the insertion of the myristoyl moiety in the vesicles.[17] Immediately after, the mix was ultracentrifuged for 2 hours at 150 000 g and 4 °C. The process is illustrated in figure 2.62.

A gel-like, highly viscous pellet formed by packed vesicles, spotted with myrUSH3 molecules on their surfaces was thus obtained. Spectrophotometric determination of the protein concentration in the supernatant allows to estimate the fraction of the initial

[17]The main phase transition, from gel to liquid crystal state, is at +23 °C for pure vesicles.

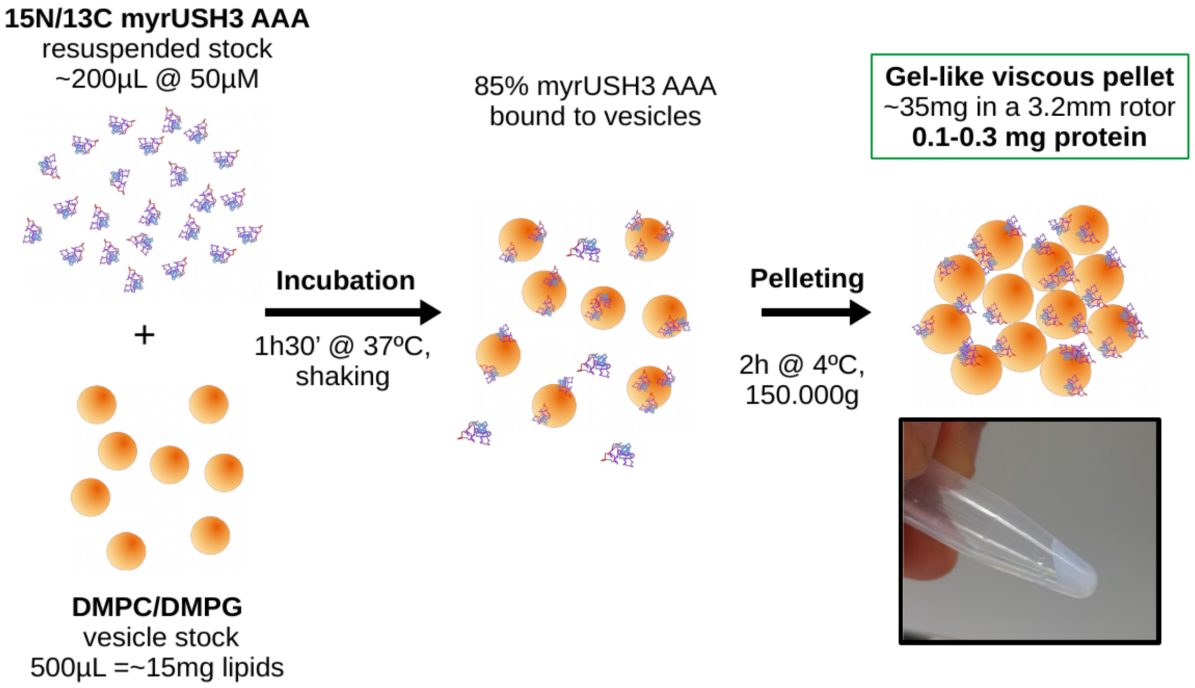


Figure 2.62: LUV-bound myrUSH3 AAA sample preparation scheme.

stock that bound to the lipids. It was found that ~85 % of the protein was captured in the vesicle gel.

Remarkably, DLS measurements showed that the gel ultracentrifugation process does not disrupt the vesicles. Upon gentle resuspension of a gel aliquot it was observed that the average size and dispersion of the distribution remained unchanged (2.61).

The initial volume of lipid stock solution was chosen so a 3.2 mm ssNMR rotor would be completely filled with the final gel sample. Thus, the final sample weight was ~35 mg, from which only 15 mg correspond to dry lipids, **0.15 - 0.3 mg** to the myristoylated protein, and the rest would be interstitial and encapsulated water forming the vesicles. It should be stressed that the sample of interest comprises a merely 0.4 - 0.9 % in weight. Together with the fact that self-associated species are minor, the need for DNP enhancement is evident.

Using these figures, it is possible to estimate the number of protein molecules per liposome. First, the number of lipid molecules per liposome is calculated: (Encapsula NanoSciences LCC 2009):

$$(15)\quad N_{\text{lip mol}} = \frac{[4\pi(\frac{d}{2})^2 + 4\pi(\frac{d}{2} - h)^2]}{a}$$

where d and h are the liposome diameter (experimentally determined) and thickness

(estimated in 5 nm); and a is the area of the lipid head group (0.7 nm for DMPC and 1 nm for DMPG, I used the mean value, 0.85 nm). Thus, the number of lipid molecules per liposome is ~160 000.

Next, the estimated number of vesicles in the sample , N_{ves}, is calculated:

$$(16)\quad N_{ves} = V_{\text{lip sol}} \frac{C_{lip} N_A}{N_{\text{lip mol}}}$$

where $V_{\text{lip sol}}$ is the volume of stock lipid solution added, C_{lip} is the total lipid concentration in the stock, and N_A is the Avogadro constant. Using the protein mass in the final sample to calculate the number protein molecules, the average number of myrUSH3 molecules per vesicle can be estimated:

$$(17)\quad N_{\text{prot per ves}} = \frac{N_{\text{prot mol}}}{N_{\text{ves}}} = 145 \approx 10^2$$

With this estimation, we expect that the average vesicle in the pellet is recovered by 10^2 myrUSH3 molecules attached to its surface.

Radical addition for DNP

As previously introduced, CE DNP enhancement requires the homogeneous presence of a bi-radical as a polarizing agent. Thus, it is most common to add a cryoprotectant agent (DMSO or glycerol, usually) along with the radical, since water crystal formation can induce protein-radical separation. Due to the heterogeneous nature of these lipid-bound myrUSH3 samples, an optimal radical distribution is necessary in order to maximize sensitivity gain.

Following the results of Liao et al. (2016) regarding the choice of radical, cryoprotectant, and mixing method, the protocol I followed to prepare my samples for DNP was:

1. Use of AMUPol (Sauvée et al. 2013) as polarizing agent.
2. Use of glycerol as cryoprotectant.
3. Manual pipetting and mixing after sample ultracentrifugation.

AMUPol has better solubility in water than TOTAPol and a more favorable partition coefficient, resulting in higher sensitivity gains. Glycerol was reported to contribute to

a better radical distribution than DMSO. Manual mixing was recommended since the presence of the highly dense and viscous glycerol prior to ultracentrifugation impedes deposition of the lipid vesicles.

Thus, 10 μL of 5 mM AMUPol dissolved in 50:45:5 glycerol/H_2O/D_2O were added to the protein-rich vesicle gels obtained by ultracentrifugation, and carefully mixed. Sample consistency became less viscous upon addition of this *DNP juice*. The final samples were then spun down into a 3.2 mm rotor using a tabletop centrifuge and stored at -20 °C until measurement.

In total, three samples of 1:1 DMPC/DMPG LUV-bound myrUSH3 AAA were prepared from three lyophilized protein batches:

1. ^{13}C, ^{15}N-labeled myrUSH3 AAA.
2. ^{13}C-labeled myrUSH3 AAA.
3. ^{13}C-labeled myrUSH3 AAA.

2.5.4 ssNMR results

Here I add a table containing the amino acid composition of the myrUSH3 AAA construct (taking into account the six histidines from the C-terminal HisTag and the elimination of M1 upon myristoylation) and the individual IDR (SH4 and Unique domains) and the SH3 domains separately. It is intended to help interpret the signals in the ssNMR spectra, since these are often broad and overlapped.

Table 2.2: myrUSH3 AAA construct amino acid composition, full length (residues 2 - 156) and disaggregated for the IDR (2 - 85) and the SH3 domain (86 - 156).

Amino acid	Full length	IDR	SH3
Ala	22	18	4
Arg	7	5	2
Asn	7	4	3
Asp	6	2	4
Gln	6	3	3
Glu	9	3	6
Gly	15	11	4
His	9	2	7
Ile	3	0	3
Leu	7	2	5

Amino acid	Full length	IDR	SH3
Lys	7	5	2
Met	0	0	0
Phe	5	3	2
Pro	12	10	2
Ser	18	11	7
Thr	10	3	7
Tyr	4	0	4
Val	6	2	4
Trp	2	0	2
Cys	0	0	0

MAS ssNMR

We first acquired exploratory 1D and 2D carbon-detected MAS ssNMR experiments without DNP at high field (18.8 T, 800 MHz ^{1}H Larmor frequency spectrometer).

In the first place, we tested the effect of temperature in sample dynamics. We acquired pairs of 1D hC **Cross Polarization (CP)** and **Insensitive Nuclei Enhanced by Polarization Transfer (INEPT)** experiments, between -25 and +30 °C (figure 2.63). hC means that ^{1}H nuclei are first excited to exploit their higher polarization, which is then transferred via dipolar coupling to ^{13}C, the observed nuclei. CP is the most standard method to achieve polarization transfer in solid samples (Pines et al. 1973), whereas INEPT has the particularity of being efficient only in mobile regions. Acquisition of both experiment variants over a temperature range allows to quantitatively analyze how large scale dynamics are activated with increasing T.

It should be stressed that most of the sample mass corresponds to the hydrated DMPC/DMPG vesicles. Although the lipids are not uniformly labeled, there is a significant quantity of ^{13}C atoms due to the isotope natural abundance that make vesicles visible by NMR. The lipid matrix is in principle homogeneous (vesicles are unilamellar), but have a complex phase diagram and therefore complex ^{13}C spectra (Purusottam et al. 2015).

Starting from -25 °C, no signal was observed in the INEPT spectrum, indicating absence of detectable motions, neither in the protein nor in the lipids. In the CP spectrum instead, the signals from the lipids (two small, broad peaks at 15 and 24, and a sharper, more intense one at 33 ppm, corresponding to the abundant myristoyl methilene groups,

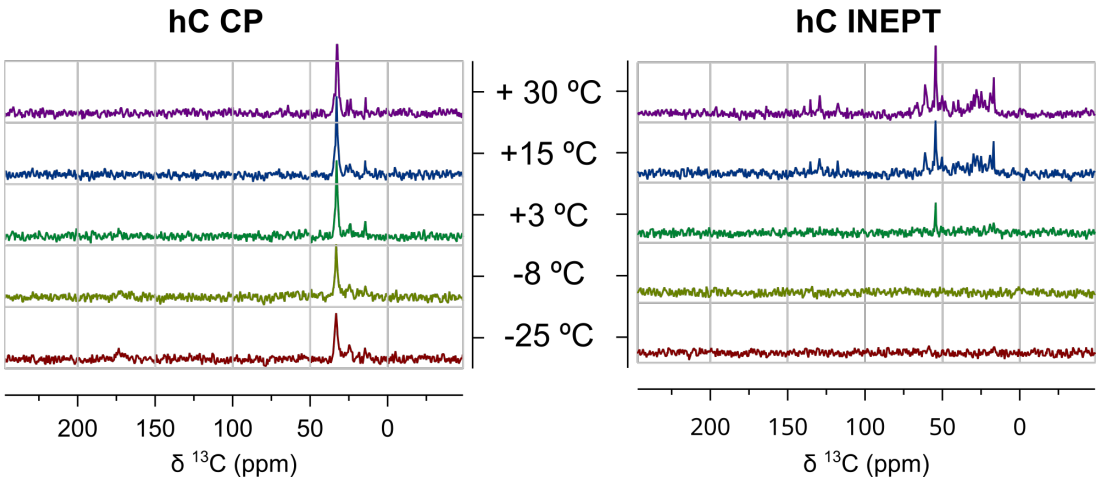


Figure 2.63: 1D hC CP (left) and INEPT (right) MAS ssNMR spectra without DNP at different temperatures. The °C scale is used here for practical reasons.

were observed. Additionally, weak signals from myrUSH3 AAA C' nuclei were visible between 160-190 ppm. This evidenced the need for DNP enhancement: even with 128 accumulated acquisitions, taking ~15 minutes acquisition time, signal to noise ratio was extremely poor, rendering impractical further 2D correlation experiments with regular MAS ssNMR. In any case, these results evidenced that LUV-myrUSH3 AAA dynamics are mostly suspended at -25 °C, which includes motions associated to vesicle fluidity, the individual flexibility of the IDR, and the global mobility of the USH3 moiety over the lipid surface.

At -8 °C, the weak signals from protein C' were further reduced in the CP experiment, while those corresponding to lipids between 0 - 50 ppm remain unchanged. No signals were yet detected in the INEPT variant.

In the CP spectrum at +3 °C, almost all signals in the C' region were already imperceptible. The lipid signal at 33 ppm sharpened and the rest remained mostly identical. In the dynamics-sensitive INEPT experiment, however, signals started to appear in the aliphatic region, most notably a sharp peak at 54 ppm. These peaks corresponded to the lipid head groups starting to gain mobility in the vesicle surface (Purusottam et al. 2015).

At +15 °C and +30 °C (over the vesicle transition temperature to liquid crystal phase), while no further changes where observed in the CP experiments, new signals gradually arose in the INEPT spectra between 10 - 70 and 110 - 140 ppm. While the most intense correspond to lipid molecules gaining mobility, the smaller peaks correspond to the peptide, probably to signals from the IDR.

An important issue for the DNP MAS experiments raised here by the effect of temperature on dynamics is sample heterogeneity and stability. Freezing the sample down to cryogenic

temperatures is mandatory for DNP. That implies that, if one is to acquire experiments with the same sample in different instrument shifts, the sample will temporarily heat when retired from the refrigerated stator. During this process, the integrity of the lipid vesicles could be compromised, as observed by the changes in mobility. Although it was shown that the transition from gel to liquid crystal and back did not affect vesicle stability, DLS curves of frozen and thawed pellet (-20 °C and +37 °C) showed that, at least upon resolubilization for the DLS measurement, the vesicles were disrupted and former larger assemblies (figure 2.64).

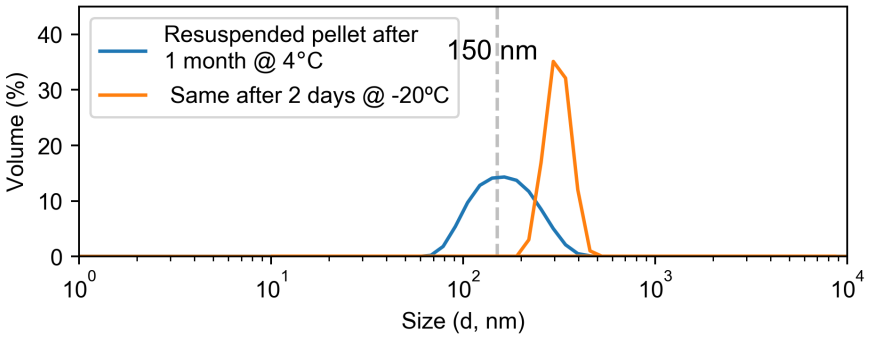


Figure 2.64: DLS stability test upon pellet deep freezing.

The drawback was taken care of by minimizing the temperature changes, storing the samples at -20 °C and transporting them ice-cold to and from the spectrometer. However, the issue needs to be further studied, for which we planned future electron microscopy experiments to test vesicle stability upon the glass transition at cryogenic temperatures.

DNP MAS ssNMR

The first experiments we acquired using DNP MAS ssNMR were 2D ^{13}C - ^{13}C correlation spectra using the **Dipolar Assisted Rotational Resonance (DARR)** implementation(Takegoshi et al. 2001). In short, spin polarization is transferred from ^{1}H to the ^{13}C nuclei by CP. Next, the magnetization is left to evolve under ^{13}C - ^{13}C dipolar coupling for a certain duration (**mixing time**), hence spreading to nearby nuclei and generating correlation that is finally recorded as cross-peaks in the spectrum[18]. Evidently, these homonuclear spectra are symmetrical about the diagonal.

Tuning the mixing time permits to select the reach of the correlations. Short mixing times of 10^1 ms only allow short range contacts - i.e. mostly intra-residue correlations -

[18]The pulse sequence is identical regardless the use of DNP. The only difference is that here, with simultaneous MW radiation, the electron spin magnetization is first transferred to the ^{1}H.

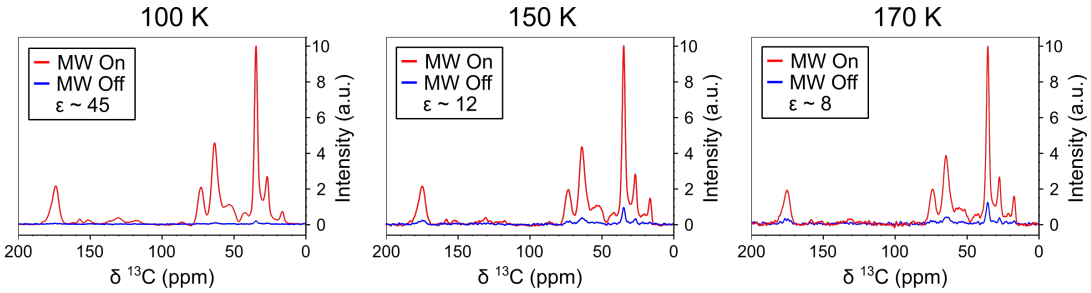
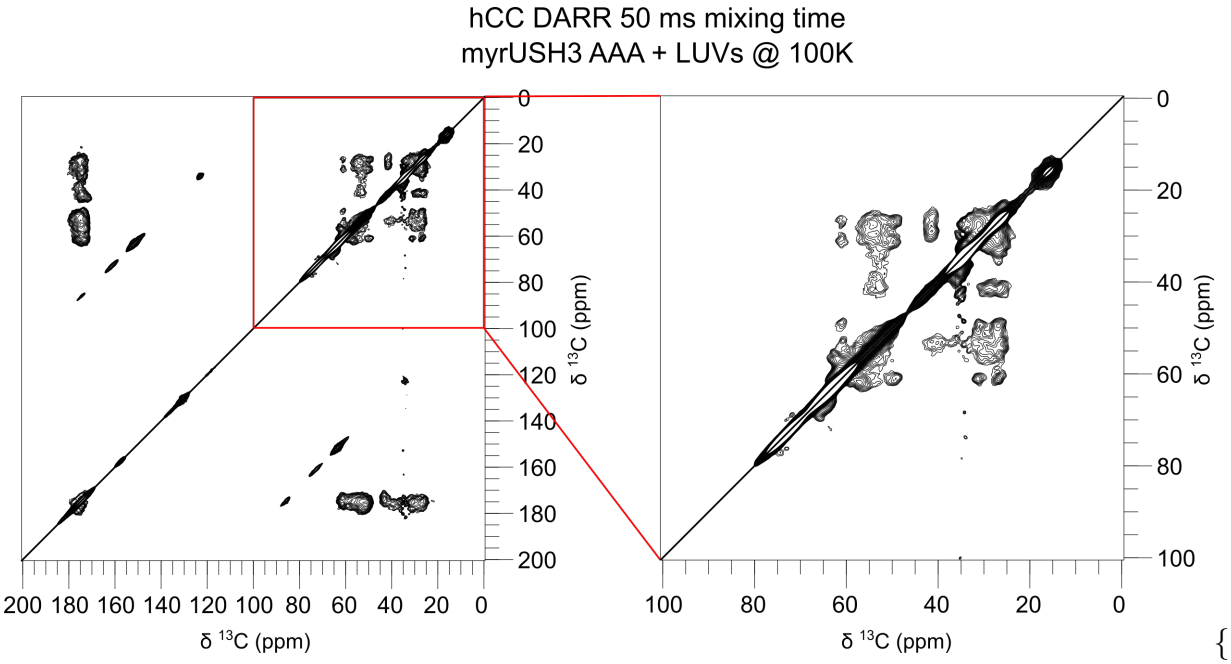


Figure 2.65: DNP enhancement (ϵ) estimation at different temperatures on ^{13}C labeled myrUSH3 AAA bound to lipid LUVs.

but thanks to the limited spin diffusion are well resolved. Instead, using longer evolution times in the order of 10^2 ms cross-peaks between more distant nuclei - e.g. inter-residue contacts - can build up at the cost of signal intensity and crowded spectra. Therefore, ^{13}C - ^{13}C correlation spectra are typically recorded with different mixing times.



{ width=75 % }

We estimated the DNP enhancement at different cryogenic temperatures in order to determine the working range (see figure 2.65). By measuring 1D hC spectra with and without MW radiation, and fitting peak intensities, it is possible to estimate signal enhancement. Our results were satisfactory, in the order of those reported by Liao et al. (2016).

We first acquired a series of C-C correlation spectra (hCC DARR) at 100 K in order to maximize signal gain with mid and long mixing times, 50 and 100 ms (figures **??** and 2.66).

The first spectra were promising, given the challenging sample and the intrinsic low res-

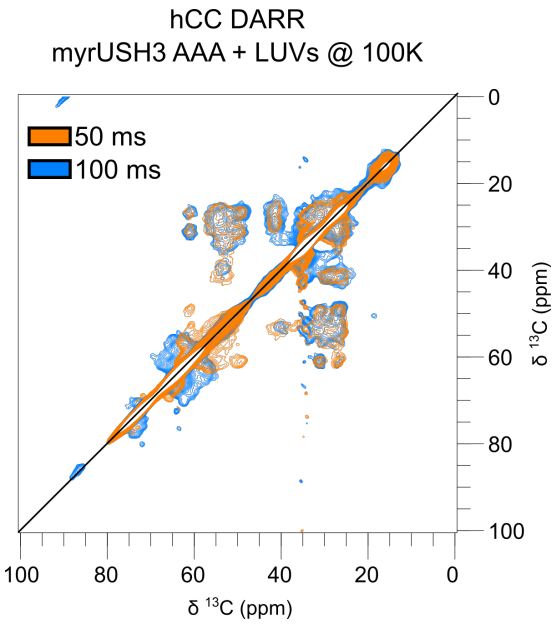


Figure 2.66: 50 and 100 ms mixing time hCC DARR of myrUSH3 AAA bound to lipid LUVs at 100 K.

olution typical of DNP ssNMR. Also noticeably, the experimental time was ~ 18 hours, meaning that not enhanced ssNMR is not applicable to this type of sample (recall that sample amount was < 0.5 mg). However, the spectral region corresponding to methyl correlations (0 - 25 ppm, including Ala, Thr, Leu, Ile, Val, etc.) was unpopulated. This is most likely due to relaxation by methyl group rotation at this particular temperature inducing severe line broadening. Thus, we next acquired hCC DARR spectra at 150K so methyl rotation would speed up and shift to the extreme narrowing regime, using only 50 ms mixing time to reduce magnetization dispersion (figure 2.67).

The experimental time remained reasonable in spite of the smaller DNP enhancement (~ 43 hours, since the number of acquisitions was increased from 24 at 100 K to 56 at 170 K), and methyl signals were now visible. Additionally, higher temperature provided better resolution due to less severe line broadening. Thus, we decided to finally explore the higher temperature, 170 K, and acquire 10 ms and 50 ms mixing times hCC DARR (figure 2.68). 128 acquisitions and ~ 3.5 days of experimental time were then required, which set the upper temperature limit in practice.

The spectra obtained had a reasonably good quality and allowed to discern particular sets of correlations. In order to obtain a first insight, I used PLUQin, an automated assignment software that queries chemical shifts to a curated database and returns scored tentative correlation identities (Fritzsching et al. 2013; Fritzsching et al. 2016). After a manual peak picking of significant maxima (the low resolution of the spectra make groups of broad peaks appear as large *blobs*), I passed the coordinates to PLUQin and then plotted the most relevant assignments as shown in figure 2.69 (score > 75 %, see

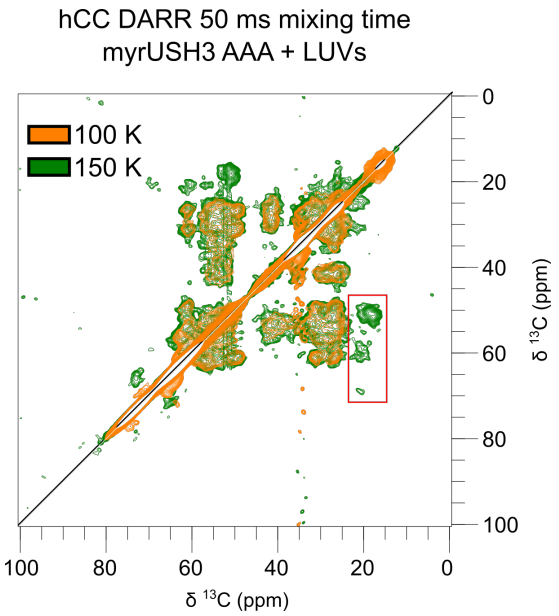


Figure 2.67: 50 ms mixing time hCC DARR of myrUSH3 AAA bound to lipid LUVs at 100 K and 150K. The methyl region with new signals is highlighted in red.

reference Fritzsching et al. (2016)).

Since the SH4 domain is most relevant for dimerization and it concentrates an important number of basic residues harboring amino groups, we also acquired a set of N-C correlation experiments at 100K: hNCA, hNCO, hNCACX and hNCOCX. In the first pair, magnetization is transferred to ^{15}N, and then selectively to nearby $C\alpha$ or C' atoms. In the CX variants, an additional C - C CP block follows, providing pseudo-3D correlations between the selected N - C signals and any other C in the vicinity. The reason to work in the lower temperature regime again is that these experiments are less sensitive due to imperfect successive magnetization transfers. The hNCACX with a 10 ms mixing time is shown here in figure 2.70.

Importantly, Arg and Lys side chain correlations could be detected, although low resolution precluded from clearly discerning peaks. Still, this opens the door to direct observation of the SH4 domain forming the intramolecular fuzzy complex in a native-like lipid environment.

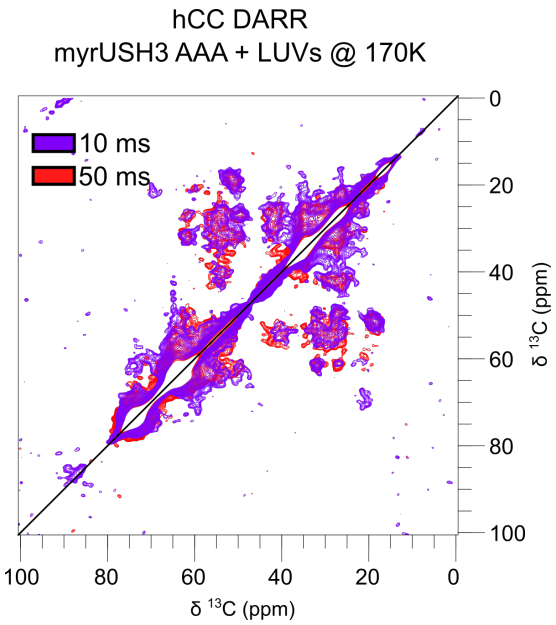


Figure 2.68: 50 and 10 ms mixing time hCC DARR of myrUSH3 AAA bound to lipid LUVs at 170 K.

2.5.5 Discussion

As a summary, I have been able to solve a series of fundamental issues that put the basis for a further structural study of myrUSH3 and the minor self-associated species. First, expression and purification accomplish the purity and quantity standards needed for a ssNMR-based approach. Secondly, LUVs have been shown to be a practical model for the lipid environment, although vesicle integrity should be further tested to ensure reproducibility. In the third place, DNP MAS ssNMR has been proven to overcome the major obstacle in this study: **sensitivity**. The sample preparation method has been optimized based on literature and the spectra obtained are satisfactory, although the suffer from low resolution. C - C and N -C correlations of different types of residues could be readily identified in feasible experimental times. To our knowledge, these are the first ssNMR spectra of natively myristoylated, membrane bound IDPs.

At this point, the main obstacle lays on the further steps to take in order to assign signals. While a purely ssNMR assignment is out of reach, a complete solution NMR assignment - including side chains - is neither a trivial task for a long IDR, nor is its transfer to ssNMR spectra. The strategy should therefore rely on comparing different samples with different lengths and/or isotopic labeling schemes. Thus, we could be able to distinguish between sets of signals from the IDR or the SH3 domain, or distinguish specific residues. For example, Ile, Trp and Tyr are exclusive reporters for the SH3 domain, whereas most Pro, Ala and Gly belong to the IDR.

We should also focus on Arg and Lys side chains because of their abundance in the SH4

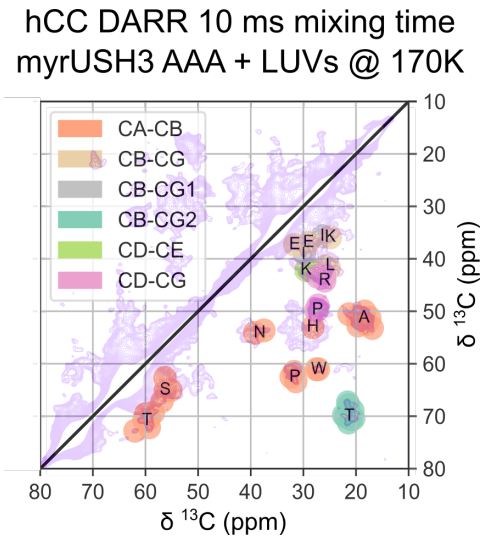


Figure 2.69: PLUQin automated assignment for a manually selected set of peaks of the hCC DARR spectrum with 10 ms mixing time shown in figure 2.68 (background).

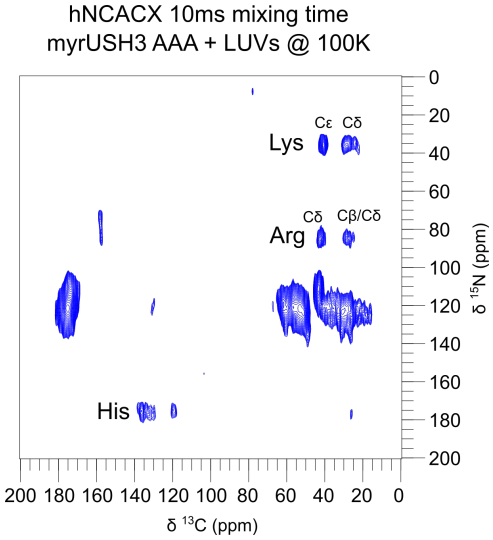


Figure 2.70: 10 ms mixing time hNCACX of myrUSH3 AAA bound to lipid LUVs at 100 K.

domain and probable implication in the self-association mechanism. In this regards, a search for inter-molecular interactions is feasible by using mixes of samples exclusively ^{13}C or ^{15}N labeled. N - C correlations detected with such set-up would be indicative of self-association.

Chapter 3

Discussion

In this final discussion I retrace the path drawn along the introduction, from c-Src and signaling proteins to fuzzy binding, in order to summarize how the intra-molecular fuzzy complex here described enables new functionality in the N-terminal region of c-Src, that experimental and theoretical data support to be a general feature in the Src Family of Kinases. Finally, I comment some perspectives based on the preliminary results obtained by ssNMR on myristoylated USH3 in a membrane-mimicking support.

3.1 An IDR and an ordered scaffold form an intramolecular fuzzy complex

Based on our previous knowledge of the system (Pérez et al. 2009; Pérez et al. 2013) and the conformational ensemble framework, I characterized the net of transient short and long-range contacts existing between the IDR comprising the SH4 and Unique domains, and the following ordered SH3 domain. The resulting model shows that the folded SH3 acts as a **scaffold** for the disordered domains with well defined *hot spots* that concentrate the contacts, thus forming an **intramolecular fuzzy complex**. I also delimited the location of **fuzzy binding sub-sites**: the whole SH4 domain and key scattered residues along the Unique domain (see section 2.1 for details) in the IDR; the RT and nSrc loops with β strands 3 and 4 between them in the SH3 domain, and the distal loop to a lesser extent, the latter specially involved in electrostatic interactions (see below).

In all, it was demonstrated that all three domains interact with each other both pairwise

and as a whole, basically through the same regions. Therefore, the N-terminal region of c-Src constitutes a deeply entangled **functional unit** regarding intramolecular contacts, rather than a series of connected but independent domains.

Mutational analysis revealed that the **SH4 domain is a short but complex interacting element** in the fuzzy complex. While its overall role in the IDR:scaffold interface is to keep the disordered ensemble close to the rigid surface, residues 11 - 20 promote close contact with SH3 mainly through interaction with the RT loop, whereas residues 1 - 10 partially preclude these interactions. Protein-protein interaction impairment by IDRs in fuzzy complexes has been reported (Gruet et al. 2016), and in this context adds complexity to the system. The fact that SH4 phosphorylation was able to perturb intra-IDR long range contacts suggests further potential to tune the conformational properties of the ensemble.

Regarding the **Unique domain**, a key feature besides its fuzzy interaction with the scaffold is its **pre-organization**. Analysis of long-range features in presence or absence of the SH3 or SH4 domains highlighted two interesting aspects:

- Conformational bias is an intrinsic property of the Unique domain, independently of SH4, SH3, and their interactions.

  The dynamic network of PRE-detected long-range, transient contacts within the IDR remained mostly unperturbed regardless of the presence of the scaffold. Additionally, the *hot spots* in SH3 did not vary with partial or total removal of SH4. These results underline the independence and functional relevance of the specific conformational bias of the Unique domain.

- Unique domain pre-arrangement and overall IDR interaction with the scaffold are related.

  SAXS profiles of the IDR indicated that the presence of the contiguous SH3 domain induces compaction in the USH3 conformational ensemble. Since IDR:SH3 interactions are weak, the entropic effect of being covalently connected - i.e., limited diffusion facilitating interconversion between fuzzy complex configurations - and the pre-organization of the Unique domain collaboratively enhance the dynamic set of contacts between them. Indeed, introduction of conformational restrain in the isolated IDR via Cys-Cys induced loops was shown to affect its capacity to interact with the scaffold.

3.2 Specific sequence determinants rule conformational heterogeneity and function

An important message extracted from the experimental results is that **there is specificity in conformational heterogeneity**. Confronting the projections of the USH3 SAXS ensembles with the cartoon resuming PRE contact mappings reflects that the dynamic IDR *cloud* loosely anchored to the scaffold is extremely free and diverse, but intra-molecular contacts are precise.

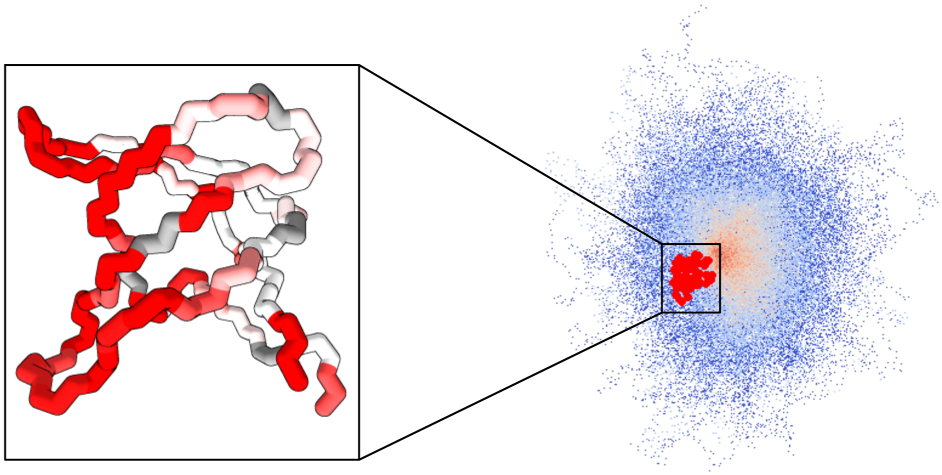


Figure 3.1: Comparison of the specificity in long range IDR:SH3 contacts as detected by PRE (see figure 2.20) vs the EOM-fitted IDR ensemble representing SAXS experimental data (see figure 2.16)

In the search for **sequence determinants** in the IDR, I first found a set of four phenylalanines in the Unique domain (F32, F54, F64, and F67, the latter two at the core of the ULBR) which happened to be mostly conserved in close c-Src homologues and paralogues. The results from each individual F#A substitution revealed their roles in the fuzzy complex. On the one hand, all four were found to promote long-range interactions in the isolated IDR, although the ones in the ULBR had a more local effect. On the other hand, all mutations induced short-range responses in the scaffold, also concentrated in the loops. Again, F64 and F67 differed in their effects, affecting specially the RT loop. Thus, the emerging model from these results is that a small number of **hydrophobic residues form a dynamic cluster via transient interactions** that partially restricts the conformational freedom of the IDR. Either as a consequence of the hydrophobic-driven compaction, or because these non-equivalent aromatics are also binding sub-sites, the IDR:scaffold interactions depend on these determinants. Although IDPs are commonly assumed to be mostly ruled by electrostatics, a similar mechanism has been reported by Ban et al. (2017). Furthermore, the authors demonstrate that the conformational population of the cluster can be tuned by a small molecule. Regarding

intermolecular interactions, large hydrophobic residues are often involved in forming high order assemblies via fuzzy interactions (Wu & Fuxreiter 2016).

Histidines H25 and H47 were also tested. Results **suggest a possible extra conformational regulatory mode based on electrostatic interactions** involving the SH4 domain, both histidines and other polar residues in the Unique domain, and the RT loop of SH3. This mechanism remains to be further characterized.

Besides long-range contacts, the oscillatory patterns in the ΔPRE profiles suggest the presence of locally concerted dynamics, in most of the cases associated with proline residues. These prolines may thus act as entropy moderators. First, prolines are secondary structure disruptors in IDRs and therefore increase $S_{conformational}$. Secondly, prolines may delimit functional motifs, thus imposing restrains to $S_{configurational}$.

Finally, I also tested mutations in all three loops of the SH3 domain in order to obtain a complementary view from the scaffold. Perturbations in the RT loop were observed to affect mostly the Unique domain, and the distal loop was observed to be related to the elements forming the electrostatic mechanism (SH4, histidines and polar residues in the Unique domain). These results underscore the importance of the flexible regions of the scaffold for fuzzy interactions.

3.3 Generality of the model in other SFKs

The **sequence alignment and coevolutionary couplings** suggest that some of the structural features and functions of the N-terminal region of c-Src may be conserved in other SFKs. Conservation of the phenylalanine residues under study and/or related motifs among most SFKs suggest that transient long-range driven by a dynamic hydrophobic cluster may be a common feature of the family. Additionally, the number of spaced proline residues does not vary significantly between them. Co-evolution indicates a strong correlation between the SH4 and SH3 domains, coherently with our structural observations on c-Src, as well as moderate intra-Unique domain couplings supporting pre-organization.

The PRE results obtained with the analogous USH3 form of **Yes confirmed the similarity** between the models: IDR conformational bias (also including oscillatory ΔPRE patterns), and contact IDR:scaffold specificity were observed. An interesting contrast was detected in the long-range, as the contact region also included the RT and nSrc loops like in c-Src, but in this case overlapped with the canonical PPII recognition site, instead of the following groove.

When I considered the N-terminal **fuzzy complex in the context of full-length** c-Src, it was evident that the model was not only plausible, but had the potential to connect the allosteric network existing within the cassette and the N-terminal IDR via SH3. Given the theoretical and experimental evidence exposed here, one may speculate that the intramolecular fuzzy complex model can be interpreted as an extension of the observations of Santos & Siltberg-Liberles (2016) on the importance of conserved structural disorder for allostery in SFKs. Thus, a common basis (IDR pre-arrangement, common sequence determinants, same scaffold) may serve to implement the nuances of each particular Unique domain in each member of the family.

3.4 Implications of fuzzy binding between ordered and disordered domains

As commented in the introduction (section 1.3), some functionalities of the N-terminal IDR of c-Src include: lipid binding ability of the SH4 domain (Sigal et al. 1994, Buser et al. (1994)) and the ULBR (including F64 and F67) (Pérez et al. 2013), calmodulin-binding (*idem*), recognition by serine and threonine kinases (Amata et al. 2014) and N-myristoyl transferase (Resh 1994), or binding to the NMDA receptor (Gingrich et al. 2004). To those, here are added its particular pre-arrangement and the ability to engage in a fuzzy intramolecular complex with the adjacent SH3 domain through an hydrophobic mechanism at least, and a probable electrostatic contribution. Evidently, all these functions are encoded in its primary sequence, either as specific, well defined motifs (as is the case of phosphorylation sites) or more vague determinants (IDR:SH3 interactions are an example). All this collection of functions embedded in only 85 amino acids illustrates the current image of **IDRs as multivalent tools**, able to accommodate a number of functionalities thanks to their evolutionary sequence malleability (Brown et al. 2011). However, in the context of fuzzy complexes, the concept goes beyond: a diversely functional disordered region forms a very particular[1] *cloud* loosely associated to a mostly rigid scaffold, enabling it with new features and regulatory mechanisms. The image is specially powerful since it tears down the binary distinction between protein order and disorder by embodying how, in Nature, both aspects not only mix but are deeply interlaced.

The SH3 domain is classically a docking site for proline-rich motifs. This building block is not only found in tyrosine kinases, but also in many other unrelated proteins and is considered a standard example of protein interaction modular domains (Mayer 2001).

[1]The term *Unique* (domain) acquires a new structural meaning here.

There is also a solid body of research on the specificity and affinity for a variety of substrates (Ladbury & Arold 2000), the importance of its flexible parts (i.e., the RT, nSrc, and distal loops) on substrate recognition (Feng et al. 1995, Hiipakka & Saksela (2007)) or its potential for allosterism (Zafra Ruano et al. 2016). Mayer (2001) has highlighted the challenge of disentangling the particular functional effects of SH3 domains in protein-protein interaction networks from the probability distribution for multiple contacts they pose.

It is interesting that the most flexible (and therefore conformationally heterogeneous) zones of SH3 are the most important interactors in the scaffold. This depicts the interface between the IDR and the ordered platform not as an accurate border between the rigid solid and free space for conformational exploration by the attached IDR, but as a boundary region sampled also by the scaffold. In fuzziness terms (sub-section 1.8), although loop extension and flexibility is negligible when compared with the IDR in absolute terms, the potential contribution to $S_{configurational}$ and H can be extremely relevant. Given the vast structural diversity available for the IDR, a modest increase in the alternative conformations of the scaffold can benefit from the potentially exponential growth of possible configurations, even if only a small fraction of them is productive.

Therefore, **fuzzy protein-protein interactions** should not be regarded only in terms of the degree of structural disorder brought about by each partner and its variation before and after association, but also as the **plasticity of the new boundary** created between both. It is ultimately this type of interface what enables the transmission of information between multivalent, environmental sensitive IDRs (Miskei, Gregus, et al. 2017) and modular assemblies of folded domains. The relevance of these fuzzy interfaces extends further, modulating how chained folded domains interact with each other, as highlighted by Santos & Siltberg-Liberles (2016).

A recent NMR study (Tong et al. 2017) on the effects of the kinase inhibitor Dasatinib in the folded cassette of c-Src has revealed that a) the drug induces changes in the dynamics of the kinase SH2-SH3 binding sites, and b) weak contacts are retained in the drug-inhibited form. The authors suggest that SH2-SH3 contacts with the SH1 domain may thus be modulated by the drug. Therefore, knowing of the details of these subtle but important fuzzy interfaces may provide a more complete view towards rational drug design.

3.5 Myristoylated USH3 results and perspectives

As a result of my INSTRUCT-funded internship at the FMP-Berlin, I was able to establish the experimental set-up for the next step on the characterization of the c-Src N-terminal fuzzy complex. Myristoylated USH3 samples bound to membrane-mimicking supports were prepared and spectra were acquired. Despite the challenges posed by the system, DNP MAS ssNMR showed to be the tool of choice for the characterization of this heterogeneous system and the functionally important minor species previously observed.

Programmed experiments include the preparation of selectively labeled samples (Higman et al. 2009) in order to obtain simpler spectra and focus on specific amino acids. Regarding the minor associated species, mixed samples of only ^{15}N and only ^{13}C myrUSH3 will be prepared in order to test inter-molecular correlation. Finally, the absence of size limitations in ssNMR could permit us to include more of the c-Src domains, up to working with the full length system in a lipid environment.

Conclusions

1. The N-terminal region of c-Src can be described as an intramolecular fuzzy complex between the intrinsically disordered SH4 and Unique domains, and the ordered SH3 domain.

2. Specific sequence determinants - most importantly, four phenylalanine residues - have been found to affect Unique domain pre-organization and IDR:scaffold interactions.

3. Experimental results from Yes together with theoretical predictions support the generality of this model among other members of the Src Family Kinases.

4. DNP MAS ssNMR is an adequate tool to study the native myristoylated form of c-Src in membrane-mimicking environments.

Chapter 4

Methods and Materials

A complete list of buffers and their compositions for the respective constructs is available at the end of this chapter.

I would also like to notice here that, since the IDR is sensitive to degradation to different extents in the distinct constructs, **keeping the sample cold as long as possible was a general rule for all protocols**. This includes buffers, resins and other materials used. This tip is specially important for the myristoylated forms, since degradation is not easily removed and is thus a potential source of sample contamination. In addition, **all spectra were acquired within 3 days after sample preparation at most** (or resuspension if lyophilized). After this period, new samples were used. This ensured absence of degradation in all experiments.

4.1 Protein cloning and expression

The following table contains the plasmid in which each construct (and derived mutants) was cloned and other details.

Table 4.1: Table of constructs containing cloning details and termini modifications respect the wild type sequences.

Constructs	Plasmid	Other	N-terminal mod.	C-terminal mod.
SH4-UD	pET-14b	Strep-tag	None	SAWSHPQFEK
USH3	pETM-30	His-tag	GAMA(GSN…)	None
		GST-fusion protein		

Constructs	Plasmid	Other	N-terminal mod.	C-terminal mod.
		TEV cleavage		
SH3	pETM-30	His-tag	GAMA(VTT…)	None
		GST-fusion protein		
		TEV cleavage		
myrUSH3	pETDuet	His-tag	None	HHHHHH
Yes USH3 C3S	pETM-30	His-tag	GA(MG**S**…)	None
		GST-fusion protein		
		TEV cleavage		

All plasmids were amplified in *E. coli* XL-10 Gold cells, purified by miniprep and sequenced. The corresponding mutations were inserted with standard commercial site directed mutagenesis kits, sequenced and their expression products confirmed by MS.

E. coli BL21 Rosetta (DE3) pLysS strains were used for expression in all cases following the same protocol, except for the myristoylated construct which is explained in section 4.4.

4.1.1 Standard expression: c-Src SH4-UD, USH3 and SH3, and Yes USH3.

Plasmids were transformed into *E. coli* BL21 Rosetta (chloramphenicol resistant) by 60'heat shock at 42 °C, followed by 2' ice bath, then left for 60' at 37 °C and finally streaked in agar plates with the corresponding antibiotics (see table 4.2)

Table 4.2: Plasmids used and their respective antibiotic resistances used in all culture media.

Plasmid	Resistance
pET-14b	Ampicillin
pETM-30	Kanamycin
pETDuet	Ampicillin

50 mL LB pre-cultures were grown overnight, then centrifuged for 10' at 1 000 g and resuspended in isotopically labeled M9 culture medium (1 L). $^{15}NH_4Cl$ (0.5 g/L) was used as ^{15}N source, and ^{13}C-Glucose for ^{13}C labeled samples (3 g/L). 1 mL of solution Q (a mixture containing traces of diverse metals, see buffer list below) and 10 mL of

Kao-Michayluk vitamin mix.

Cultures were then grown in 3 L baffled Erlenmeyer flasks at 37 °C and shaking at 150 rpm until an **Optical Density (OD_{600})** of 0.5 - 0.7. Then, 0.7mM IPTG was added for induction, cultures cooled down at 4°C for 20' and expression was left overnight (< 18 hours) at 25 °C shaking at 150 rpm. Typically, 2 L of culture were enough to obtain 2 - 4 protein samples for NMR purposes.

Cells were finally harvested by centrifugation for 30' at 5 000 g, the pellets resuspended in lysis buffer including PMSF and commercial Protein Inhibitor Cocktail (PIC), and stored at -20 °C.

4.1.2 Myristoylated protein expression

As commented in section 2.5, the NMT enzyme and the USH3 substrate are co-expressed in the pETDuet bi-cystronic vector for co-translational myristoylation. The expression protocol was optimized to reduce undesired byproducts and maximize yield.

The Marley method involved a first growth of the bacteria in rich LB medium (2 x 1 L cultures in flat bottom 3L Erlenmeyers) for approximately 1.5 hours at 37 °C and 140 rpm shaking until an OD of ~ 0.6 (**strictly below 0.7**) was reached. Cultures were then harvested by centrifugation for 30' at 1 000 g. In parallel, 1 L of isotopically labeled M9 medium was prepared with the usual amounts of isotope sources. 40 mL of deionized water were heated to 85 °C and 1.25 mM of both sodium myristate and palmitate and one equivalent of NaOH to lipids are dissolved for 30'. Then, temperature is lowered to 50 °C, the pH adjusted to 8.0, and 600 µM of BSA are dissolved.

The lipid solution was then added to the minimal medium to yield a total lipid concentration of 50 mM. The harvested bacteria from the two LB cultures were carefully resuspended and added too. The concentrated culture is immediately induced with 1 mM IPTG and expression carried out for 5 hours at 28 °C and 130 rpm. Bacteria were finally harvested and stored as described before.

4.2 Standard protein purification

Cells were defrozen in a water bath for 15' and then sonicated in an ice bath for 5' using 30''bursts followed by 30''pauses. DNaseI was then added and left act for 20' at 4 °C. The

lysated culture was then ultracentrifuged for 45' at 75 000 g and 4 °C. The protein-rich supernatant was collected for subsequent affinity purification.

4.2.1 Strep-tag affinity purification

In the case of SH4-UD constructs containing a Strep-tag, 2 mL of 1:1 Strep-Tactin resin suspension were added to a gravity column, equilibrated with 2 column volumes (CV) of lysis buffer, and mixed with the lysate. Binding was done overnight at 4 °C under gentle shaking.

The flow-through was then discarded, the resin washed with 5 CV of wash buffer. Finally, 1/2 CV of elution buffer was added, left shaking gently for 15' and recovered. The process was then repeated 4-5 times without shaking.

4.2.2 His-tag affinity purification

In the case of c-Src and Yes constructs containing a His-tag (except myristoylated forms, see below), 5 mL Ni-NTA agarose cartridges were used. A peristaltic pump was utilized to load the cell lysate at low speed (2 mL/min), then wash the resin with ~ 2 CV of wash buffer and medium speed (5 mL/min), and finally elute the GST-fusion protein of interest.

Thereafter, an size exclusion PD-10 column was used to exchange the elution buffer for lysis buffer again. 1mM DTT, 0.5mM EDTA, and 1:100 His-tagged TEV protease were added. TEV cleavage was left overnight at 4 °C rocking gently. The product was then briefly centrifuged (10' at 5 000 g) and passed through the regenerated Ni-NTA agarose cartridge. In this case, the flow-through containing the protein fragment of interest was collected, while His-tagged GST and TEV were trapped and removed.

After affinity purification, a size exclusion **Fast Protein Liquid Chromatography (FPLC)** step followed. In all cases a Superdex 75 26/60 column was used. The respective FPLC buffers were used as mobile phases. The fractions of interest were pooled and their purity checked by **Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)**.

Rich fractions were then transferred to NMR buffer and concentrated by centrifugation using 5 KDa cutoff concentrators. Protein concentration was measured spectrophotometrically according to the equation:

(18) $C = \frac{A_{280} - A_{340}}{d \cdot \epsilon}$,

where A_{280} and A_{340} are the absorbances at the respective wavelengths (in nm), d is the path length through the sample, and ϵ is the molar extinction coefficient calculated from the protein sequence. Concentrations up to ~600 µM have been tested without signs of aggregation.

Samples were finally aliquoted and preferably used immediately, or either lyophilized and stored at -80 °C.

4.3 Myristoylated protein purification

Frozen pellets were lysated as described above and then ultracentrifuged for 45' at 75 000 g. The pellet obtained was resuspended in 8 mL of cold lysis buffer containing 1 % v/v Triton X100 and protease inhibitors, ultracentrifuged again, and the supernatant now containing solubilized myrUSH3 then collected. The pellet wash was repeated two times.

Since these constructs contain a C-terminal His-tag, a single affinity purification step is performed. For the AAA mutant presented here, the materials and protocol is the same as for the non-myristoylated construct. The wild type form presents some particularities already discussed in sub-section 2.5.3, so a second wash is done using 100 mM imidazole. The elution buffer contains 0.05 % Triton X-100 and is diluted with lysis buffer in 1:1 proportion prior to size exclusion purification.

The rest of the protocol is identical to that of non-myristoylated constructs. However, I would like to draw attention over a particularity of the purification process. Although the 5 mL Ni-NTA agarose cartridges used here have a large capacity and it should be possible to simultaneously purify two cell lysates, I observed that the yield drops drastically so it must be avoided; two sequential purifications are more efficient. This is probably due to hydrophobic interactions in the resin matrix and affects both AAA and wild type constructs.

4.4 Spin labeled sample preparation for PRE

All constructs used for PRE measurements contained a single cysteine residue. Consequently, all buffers used during purification included 5 mM DTT to avoid dimerization. FPLC fractions were concentrated to a volume of 2.5 mL and a concentration of ~ 100 mM. An MTSL solution stock (15 mg dissolved in 50 µL acetone) that was prepared in advance was kept nearby in dry ice, and a PD-10 column was used to remove DTT by eluting the protein in NMR buffer. An elution volume slightly shorter than the recommended (3.4 mL instead of 3.5 mL) was used to ensure DTT removal. Immediately after, a 16-fold excess of MTSL was added to the protein, and the the sample was left rocking overnight protected from light at 4 °C.

Next, a PD-10 column was used again to remove the excess of free MTSL, also using a safer elution volume, and the sample concentrated as usual while trying to protect it from light.

4.5 *In vitro* phosphorylation of SH4-UD A27C

The SH4-UD A27C was purified following the standard protocol. The fractions of interest from FPLC were pooled, concentrated and transferred to a buffer containing 100 mM HEPES, pH = 7.5, 100 mM NaCl, 25 mM $MgCl_2$, 10 mM DTT and 10 mM ATP. 400 units of recombinant PKA catalytic subunit from bovine heart (commercial) were added and incubated for 4.5 hours at 30 °C. Then, the enzyme was inactivated by heating the sample for 5' at 75 °C. The phosphorylated product was then repurified using Strep-Tactin resin and finally spin labeled with MTSL as described above. The yield in this case was lower than usual, so the final sample concentration was merely 30 µM.

4.6 Cyclization of double-Cys SH4-UD mutants.

As for the single-Cys mutants used for PRE, DTT is present at all purification steps. In this case when DTT is removed for activation of the thiol groups the objective is to obtain a fast reaction avoiding scrambling. Thus, a mild oxidizing solution was prepared adding 1.1 equivalents of $K_3[FeCN_3]$ in 7 mL of NMR buffer. The DTT-free protein was eluted from the PD-10 column dropwise directly over the solution under vigorous stirring to optimize diffusion and oxigenation. Then, another size exclusion buffer exchange was

performed in order to remove the reagent excess and reduction products. Cyclizations were controlled by LC-MS.

4.7 Solution NMR sample preparation

The final sample concentration range was 100 - 300 µM and the volumes were between 250 - 350 µL. If samples were to be used immediately after purification, 10 % D_2O was added and the sample transferred to a Shigemi tube. When lyophilized samples were used, deionized water was used to resuspend the powder, and then they were left to rehydrate for about 1 hour at 4 °C. Thereafter, concentration was checked and then the sample treated equally.

Importantly, the NMR buffer (50 mM NaP, pH = 7.00) was always prepared fresh at room temperature and its pH strictly controlled with a recently calibrated pH-meter. The proportions of phosphate salts used to prepare it were calculated in advance so corrections were minimal and did not perturb ionic strength.

4.8 Solid state NMR sample preparation

All myrUSH3 samples were expressed in Barcelona, lyophilized for safer transport and stored at -80 °C upon arrival.

Concentrated DMPC and DMPG stocks were prepared by dissolving 15 mg/mL of each in $CHCl_3$ and 60:35:5 $CHCl_3$:CH_3OH:H_2O, respectively. 2 mL of a 1:1 mix were prepared and rotaevaporated until dry. The film was then rehydrated in 0.5 mL of NMR buffer at 40 °C by vortexing. The resulting solution was then extruded manually by 18 - 30 through a 200 nm filter and the LUV size distribution checked by DLS.

Protein samples were resuspended by 1.5 hours at 37 °C, since shorter times or lower temperatures led to sample loss. The resuspended protein was mixed with the liposome stock and incubated for 1.5 hours at 37 °C. Thereafter, the sample was ultracentrifuged for 2 hours at 150 000 g and 4 °C until a viscous pellet was formed. The protein supernatant was collected in order to estimate the fraction of lipid bound protein.

The resulting pellet was mixed with 10 µL of 5 mM AMUPol dissolved in 50:45:5 glycerol/H_2O/D_2O, spun down into a 3.2 mm rotor, sealed and stored at 4 °C. After the first DNP experiment with the sample - i.e., first deep freezing - the samples were kept on ice when transported and stored at -20 °C.

4.9 Solution NMR acquisition and processing

All NMR experiments were conducted in a Bruker 600 MHz spectrometer equipped with a TCI cryoprobe and an Avance III console. Real acquisition temperatures used in all measurements were 278 K for observing the IDR and 298 K for the folded SH3 domain. Temperature changes on the same sample (e.g., USH3 constructs) were done in 5 K steps followed by a ~ 15' period for stabilization.

4.9.1 CSP experiment acquisition

Either ^{1}H-^{15}N sensitivity-enhanced HSQC, SOFAST HMQC, or BEST TROSY HSQC pairs of spectra were acquired for all CSP experiments, depending on the sample used. While the SOFAST variant provides best sensitivity, specially useful for low concentration samples, BEST TROSY has better resolution.

In the case of titrations with peptides (VSL12 PxxP and/or SH4), the reference sample was recovered from the Shigemi tube used an automatic syringe, transferred to an Eppendorf tube and mixed with a concentrated peptide solution. These stock solutions were typically dissolved in water in concentrations that allowed addition of < 1 µL to reach 1:1 protein:ligand concentrations, thus not changing significantly ionic strength.

4.9.2 PRE experiment acquisition

Immediately after sample preparation, the paramagnetic spectrum was first recorded. The number of acquisitions was estimated based on the intensities of known signals. Thereafter, the sample was recovered with an automatic syringe, transferred to an Eppendorf and a 5-fold excess of sodium ascorbate was added to reduce the spin label to the diamagnetic hydroxylamine. Again, the ascorbate stock is highly concentrated, so > 1 µL was added. It was checked that the buffering capacity of the NMR buffer kept pH stable. The reaction was left for 5' at room temperature and the sample transferred back to the tube. The diamagnetic reference was then acquired using the exact same experimental parameters.

4.9.3 Assignment acquisition

The assignments for c-Src SH4-UD had been first done in our group and deposited in BRMB:15563 (Pérez et al. (2009)), while the SH3 domain was already available and had

been transferred and confirmed with 3D correlation experiments. The same applies to the SH3:VSL12 complex.

Mutant assignments were transferred from the wild type reference and confirmed with 3D correlation experiments when necessary.

Regarding the assignment of Yes USH3 C3S, two sets of BEST-TROSY HNCO, HNCA, HNcoCA, HNCACB, and HNcoCACB at 278 K and 298 K. The non-uniform acquisition schedules were weighted according to transverse relaxation times of 28 ms for ^{15}N and 40 ms for ^{13}C. The sampling percentage was set to 40 % for the HNCO experiments and 10 % for the rest of experiments. All spectra were externally referenced to DSS.

4.9.4 NMR data processing

All 2D ^{1}H-^{15}N spectra were processed using Topspin 3.6 using standard parameters.

Yes assignments at 278 or 298 K were co-processed using the multi-dimensional decomposition routine in qMDD (Orekhov & Jaravine 2011) and NMRPipe (Delaglio et al. 1995) also using standard parameters.

4.10 Solid state NMR acquisition and processing

All DNP MAS ssNMR experiments were acquired in a Bruker 400 MHz wide bore spectrometer equipped with a 3.2 mm HCN probe, a MAS cooling unit, an Avance III console and a Bruker 9.65 T gyrotron operating at 263.6 GHz and 25 W. The MAS spin rate was 8 900 Hz.

All spectra were processed using Topspin 3.6 and analyzed using CCPN Analysis. I wrote a pipeline to feed PLUQin with CCPN Analysis peak lists and to parse and plot the results using Python 3.6.

4.11 NMR data analysis

Spectra were analyzed and assignments performed using CCPN Analysis 2.4 (Skinner et al. 2016).

All CSP, PRE and ΔPRE were calculated using the in-house written software package FarSeer-NMR (Teixeira J.M.C., Skinner S.P , Arbesú M., Breeze A.L., Pons, M. *Farseer-NMR: automatic treatment, analysis and plotting of large, multi-variable NMR data.* Submitted to Journal of Biomolecular NMR, 2017).

4.11.1 CSP calculation

CSP were calculated according to formula 7. The statistical significance threshold shown in all CSP plots was individually calculated for each set as the mean CSP for the lowest decile plus three standard deviations. This method allows to select unperturbed residues and establish a robust baseline regardless of how large CSP are.

4.11.2 PRE calculation

The noise baseline was set individually for the paramagnetic and diamagnetic spectra using CCPN Analysis for signal intensity normalization. The average intensity of an empty spectral region was selected, preferably at δ^1H values > 9.0 to avoid artifacts derived from imperfect water signal suppression. Peaks were then Gaussian fitted and their corresponding intensities exported to FarSeer-NMR. PRE values were finally calculated as the I_{para}/I_{dia} ratio.

4.11.3 Random coil PRE simulation

Flexible Meccano (Ozenne et al. 2012) was used to calculate theoretical random coil PRE profiles. Based on the construct sequence, a pool of 50 000 conformers was generated for each data set, for which theoretical profiles are individually calculated and then averaged. The average ^{1}H linewidth of the diamagnetic reference is used by FM as an estimation for $R_{2,dia}$, all values being between 15 and 30 Hz.

4.11.4 ΔPRE calculation

ΔPRE were calculated using FarSeer-NMR as the difference between the theoretical random coil value and the experimental intensity ratio. Negative values were set to zero, whereas prolines and unassigned peaks were treated as NaN values. A Gaussian filter of 7 points size and spread of 1 SD was used to convolve the raw ΔPRE values using

the Astropy routines (Robitaille et al. 2013). NaN values were interpolated up to four consecutive in the convolved profile, and otherwise left as NaN.

4.12 Co-evolutionary analysis

GREMLIN (Kamisetty et al. 2013) was fed with the sequence of the USH3 region of c-Src. The software then performed a search on a Uniprot database version previously pre-clustered by Pfam families using the HHblits algorithm (Remmert et al. 2011). An HMM alignment was then generated and sequences with > 75 gaps were excluded. A final set of 151 sequences was used for calculation of co-evolutionary coupling and can be found in the Supplementary Information of Arbesú et al. (2017).

4.13 SAXS data analysis and modeling

This work was done by collaborators, as stated in sub-section 2.2.1. Thus, the reader is referred to the Methods and Materials section of Arbesú et al. (2017).

4.14 Buffer list

As previously mentioned all lysis, wash, elution and FPLC buffers for cysteine mutants were added 5 mM DTT (fresh). Also importantly, competing reagents for affinity columns (imidazole for Ni-NTA agarose and desthiobiotin for Strep-Tactin) were always added fresh right before use.

- Solution Q:
  - 40mM HCl, 25 mM $FeCl_2$, 2.5mM $ZnCl_2$, 2.5 mM Na_2MoO_4, 1mM $CaCl_2$, 1mM HBO_3, 200 µM $MnCl_2$, 150 µM $CoCl_2$, 25 µM $CuCl_4$.
- M9 culture medium:
  - 22 mM $NaH_2PO_4 \cdot 7H_2O$, 22 mM KH_2PO_4, 8.5 mM NaCl, pH = 6.8.
- **SH4-UD constructs**:
  - Lysis buffer: 100 mM TRIS-HCl, 150 mM NaCl, 1 mM EDTA, 0.01% NaN_3, pH = 7.5.

  - Wash buffer: Same composition as lysis buffer, pH = 8.0.
  - Elution buffer: Same composition as wash buffer, plus 2.5 mM desthiobiotin.
  - FPLC buffer: 50mM NaP, 0.2 mM EDTA, 0.01% NaN_3, pH = 7.0.

- **USH3, SH3 and Yes USH3 constructs**:
  - Lysis buffer: 20 mM TRIS-HCl, 300 mM NaCl, 10 imidazole, 0.01 % NaN_3, pH = 8.0
  - Wash buffer: Same as lysis buffer.
  - Elution buffer: Same as lysis buffer, plus 400 mM imidazole.
  - FPLC buffer: 50mM NaP, 150 mM NaCl, 0.2 mM EDTA, 0.01% NaN_3, pH = 7.5.

- **myrUSH3 constructs**:
  - Lysis buffer: Same as for non-myristoylated.
  - Pellet wash buffer: Same as lysis buffer, plus 1 % v/v Triton X100.
  - Wash buffer I: Same as lysis buffer, plus 10 mM imidazole.
  - Wash buffer II (only for wild type): Same as lysis buffer, plus 100 mM imidazole.
  - Elution buffer: Same as lysis buffer, plus 400 mM imidazole.
  - Elution buffer for wild type form. Same as standard, plus 0.0 % v/v Triton X100.

- NMR buffer: 50 mM NaP, pH = 7.0.

Chapter 5

Appendix

5.1 SFK sequence conservation

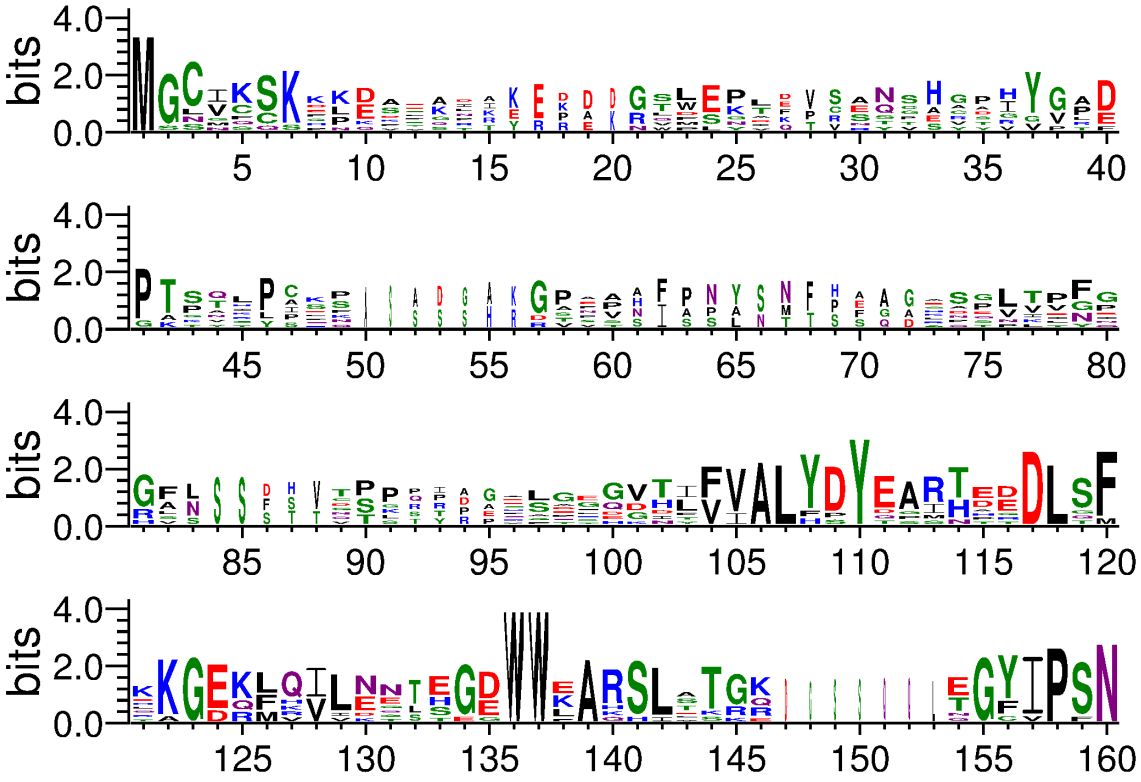


Figure 5.1: Sequence conservation logo for the human Src Family Kinases, including Fgr (part 1). Generated with WebLogo 3.6.0 (Crooks 2004).

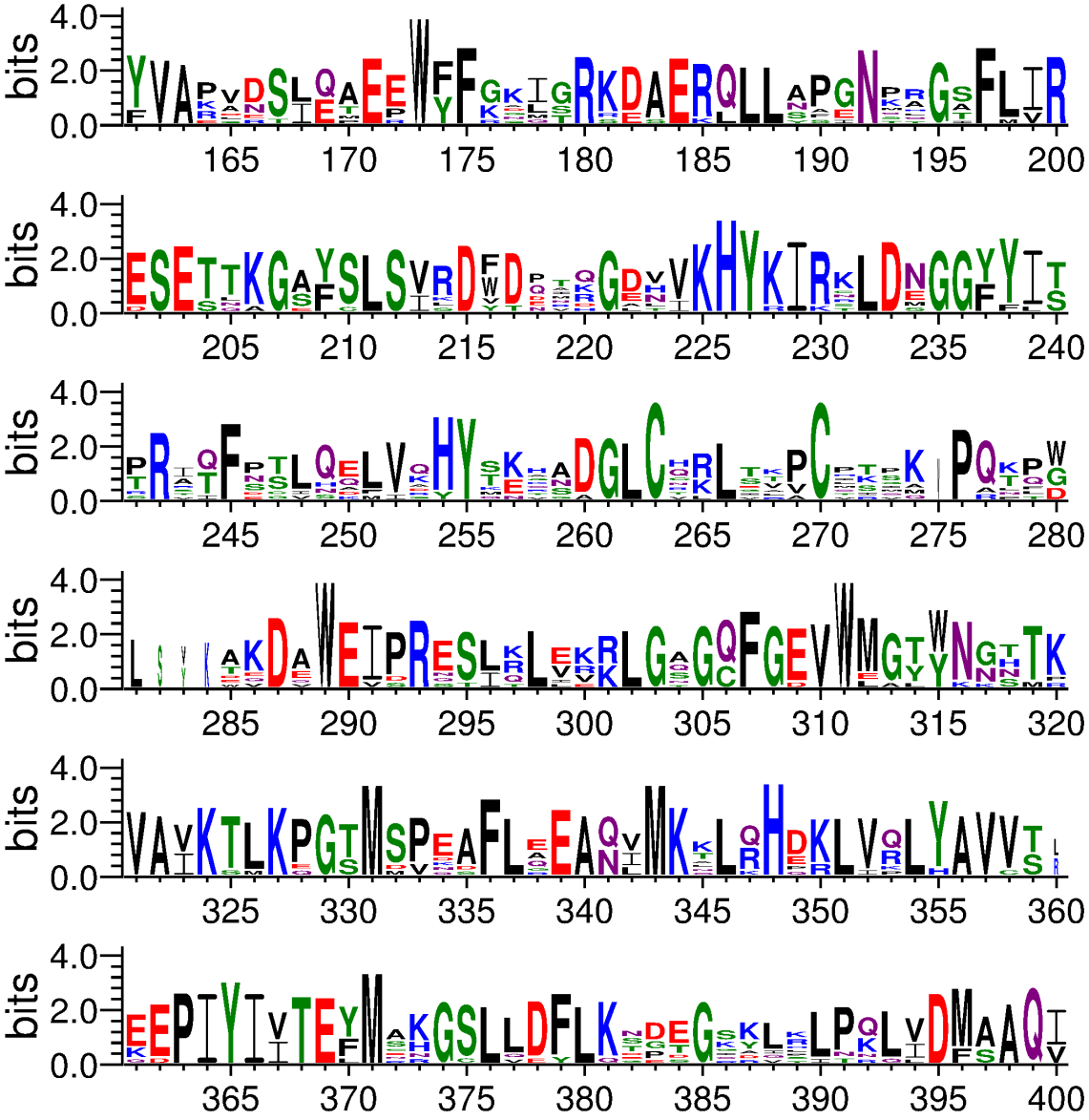


Figure 5.2: Sequence conservation logo for the human Src Family Kinases, including Fgr (part 2). Generated with WebLogo 3.6.0 (Crooks 2004).

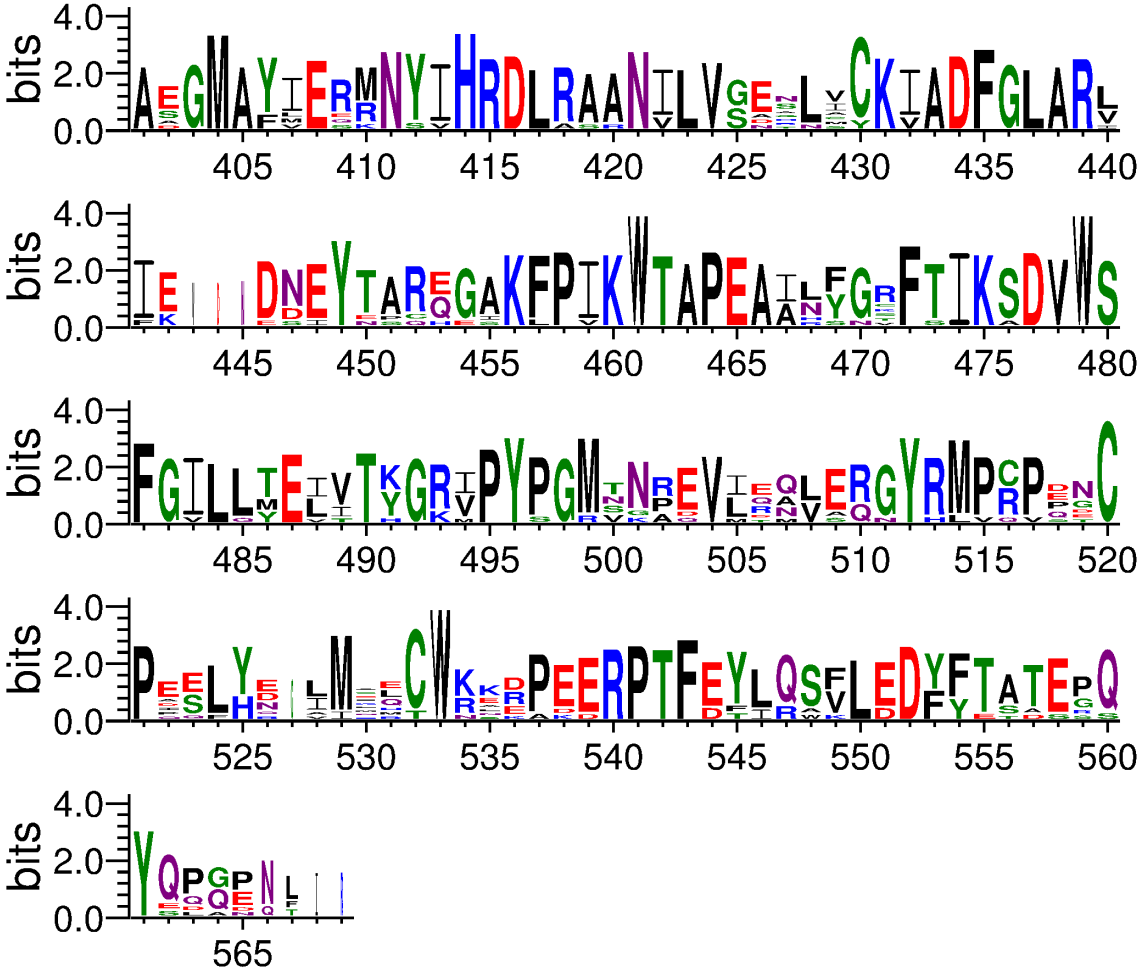


Figure 5.3: Sequence conservation logo for the human Src Family Kinases, including Fgr (part 3). Generated with WebLogo 3.6.0 (Crooks 2004).

5.2 Complete CSP mapping of the SH4 Δ mutants

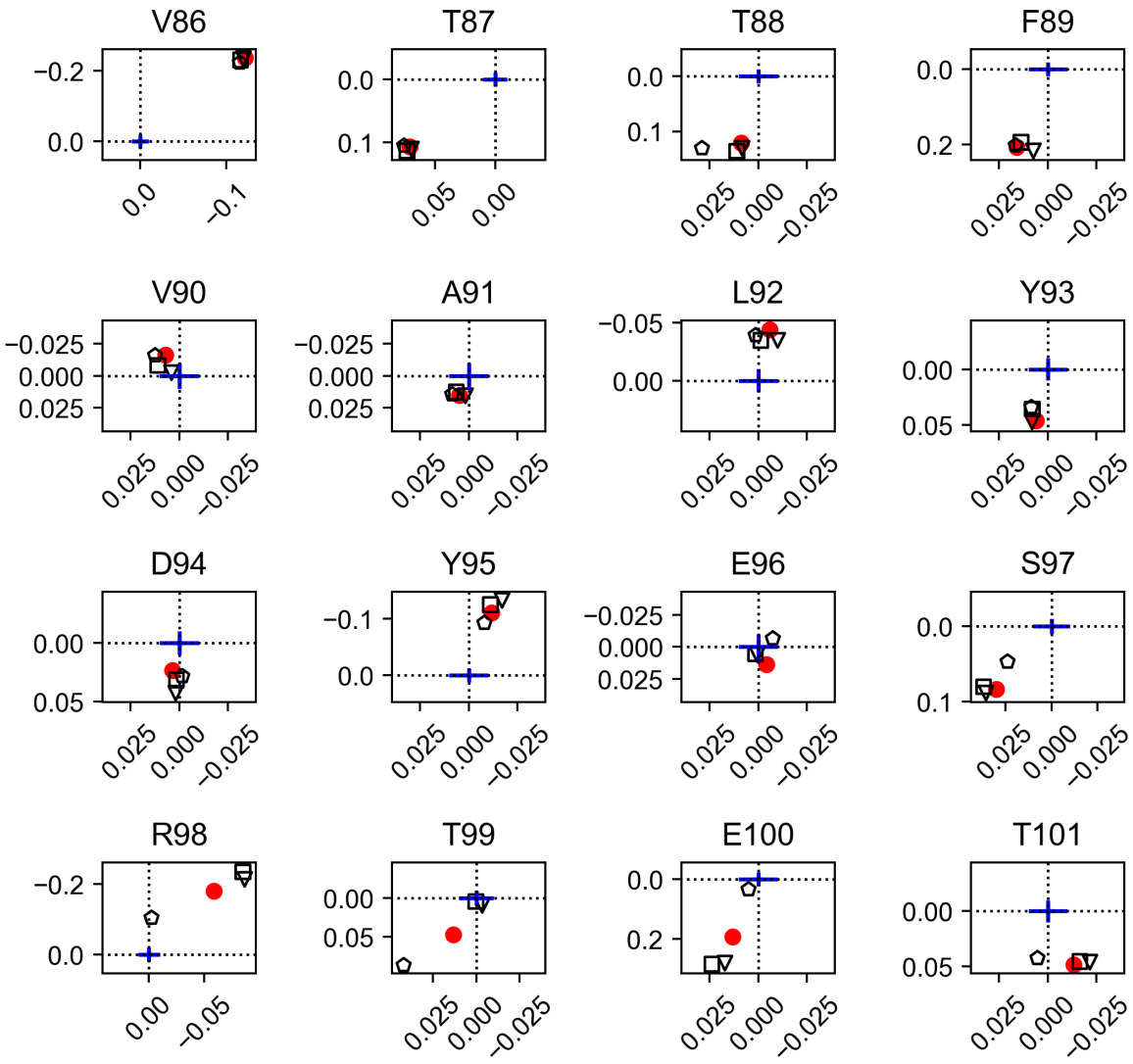


Figure 5.4: CSP mapping for USH3 WT (red dot), K5A S6A (triangle), $\Delta 10$ (square) and $\Delta 20$ (pentagon) (part 1). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^1 H$ (x axes) and $\Delta\delta^{15} N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

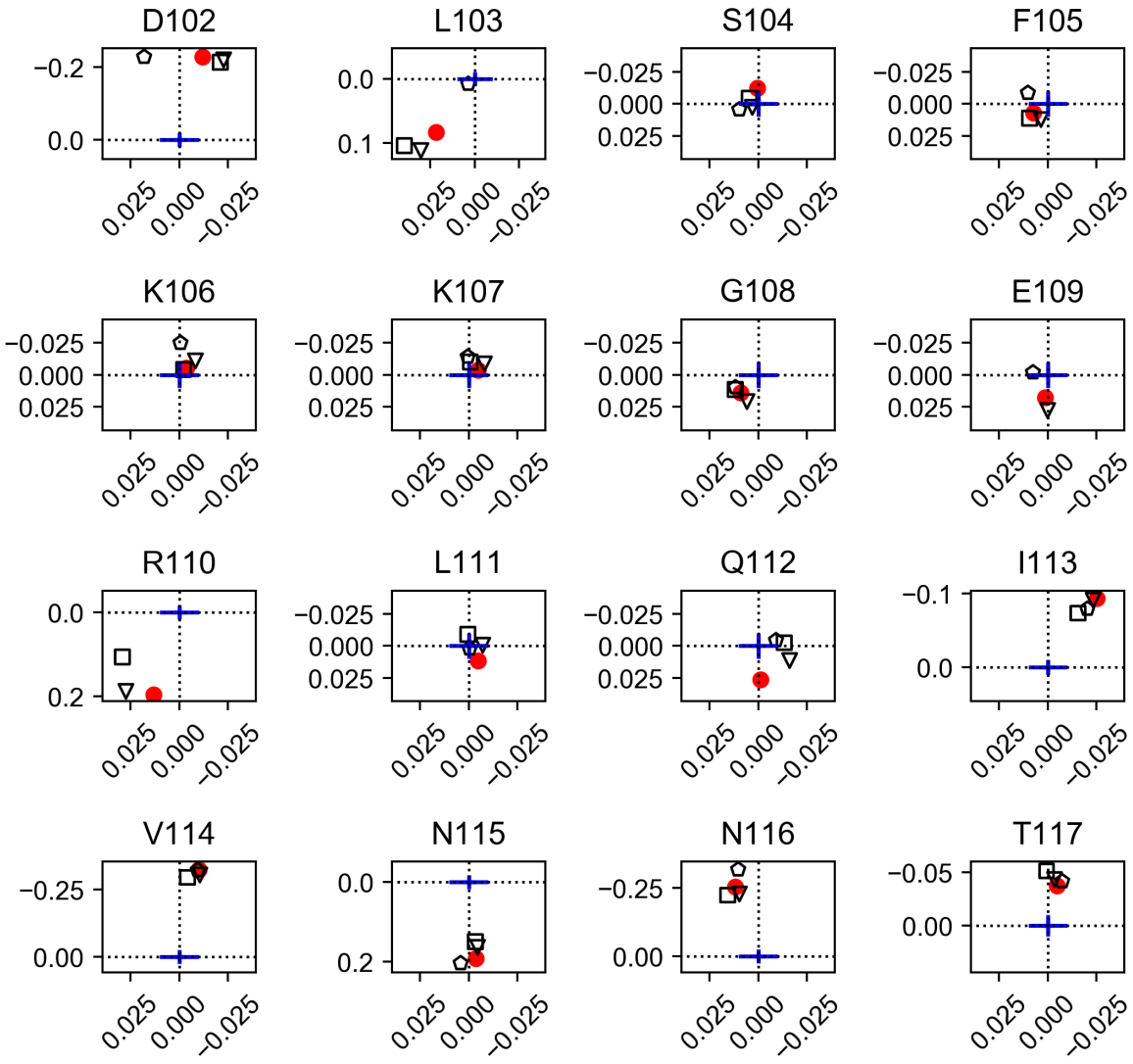


Figure 5.5: CSP mapping for USH3 WT (red dot), K5A S6A (triangle), $\Delta 10$ (square) and $\Delta 20$ (pentagon) (part 2). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^1 H$ (x axes) and $\Delta\delta^{15} N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

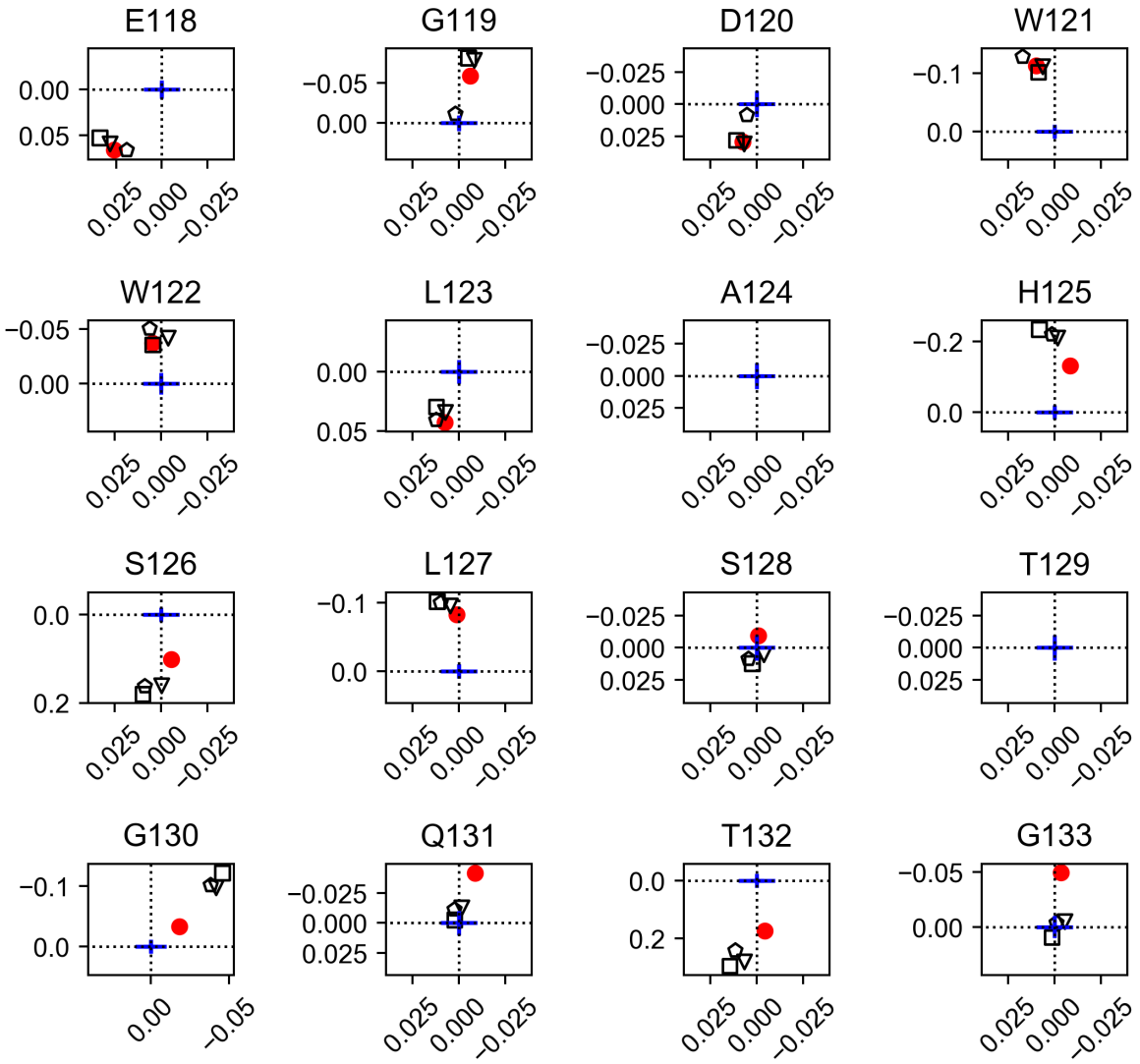


Figure 5.6: CSP mapping for USH3 WT (red dot), K5A S6A (triangle), $\Delta 10$ (square) and $\Delta 20$ (pentagon) (part 3). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^{1}H$ (x axes) and $\Delta\delta^{15}N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

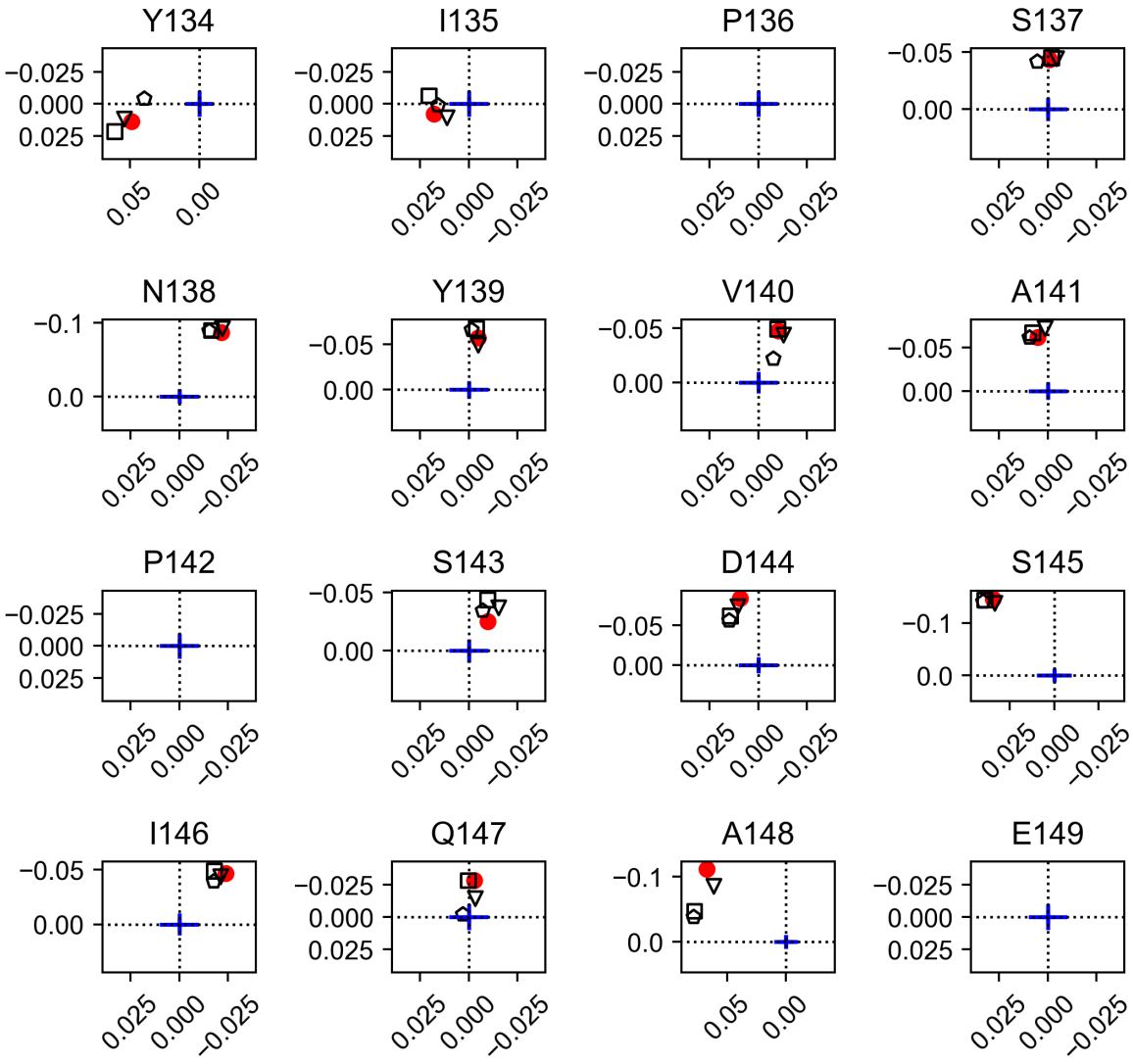


Figure 5.7: CSP mapping for USH3 WT (red dot), K5A S6A (triangle), $\Delta 10$ (square) and $\Delta 20$ (pentagon) (part 4). The origin positions correspond to the respective isolated SH3 signals. Relative scale between $\Delta\delta^1 H$ (x axes) and $\Delta\delta^{15} N$ (y axes) is indicated as a blue cross representing 0.01 ppm.

5.3 PRE and ΔPRE data sets of SH4-UD constructs

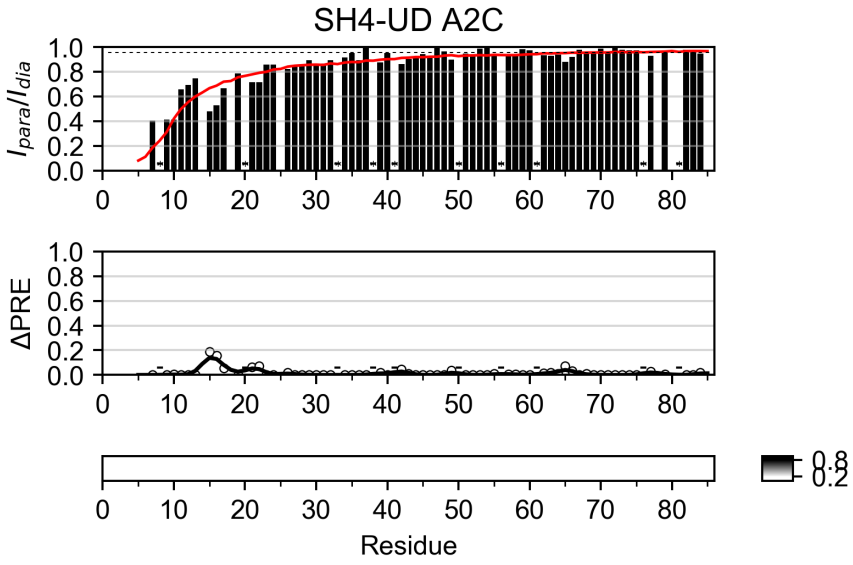


Figure 5.8: PRE and ΔPRE profiles and heat map for SH4-UD A2C.

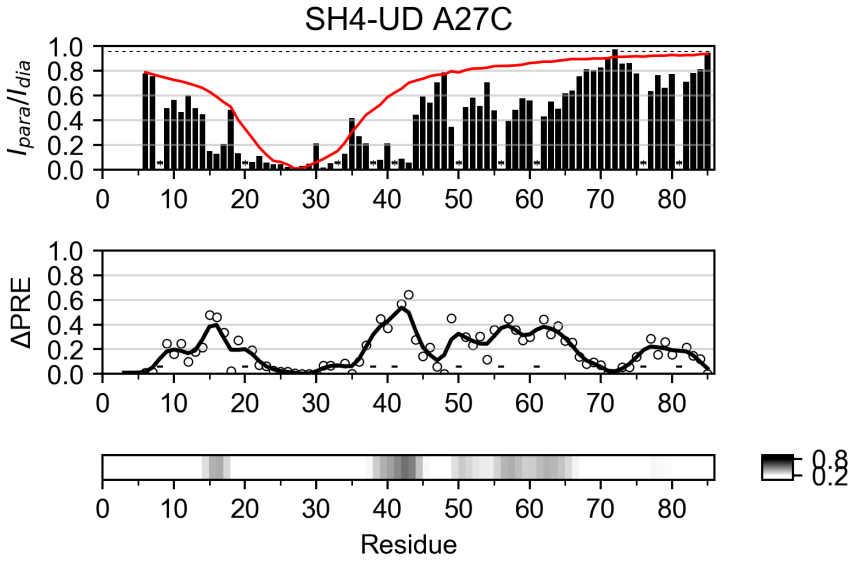


Figure 5.9: PRE and ΔPRE profiles and heat map for SH4-UD A27C.

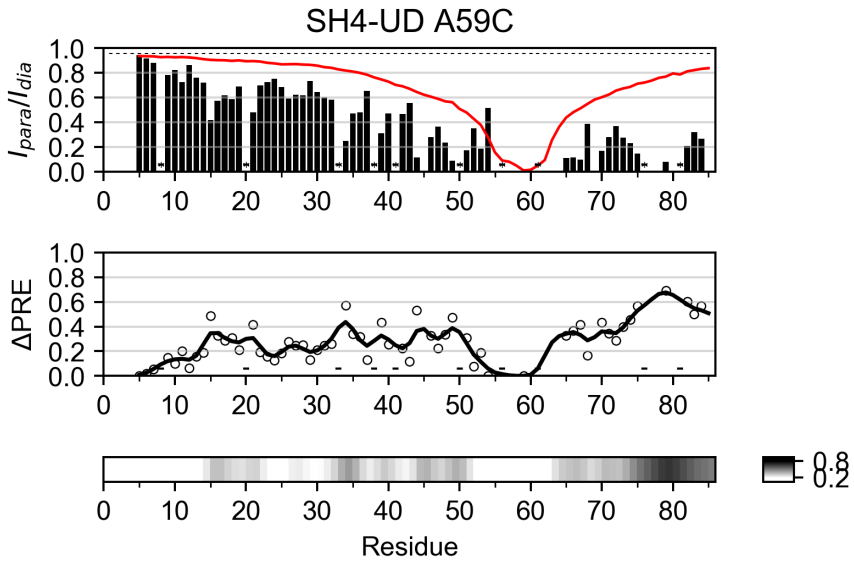


Figure 5.10: PRE and ΔPRE profiles and heat map for SH4-UD A59C.

5.4 PRE and ΔPRE data sets of USH3 ΔSH4 mutants

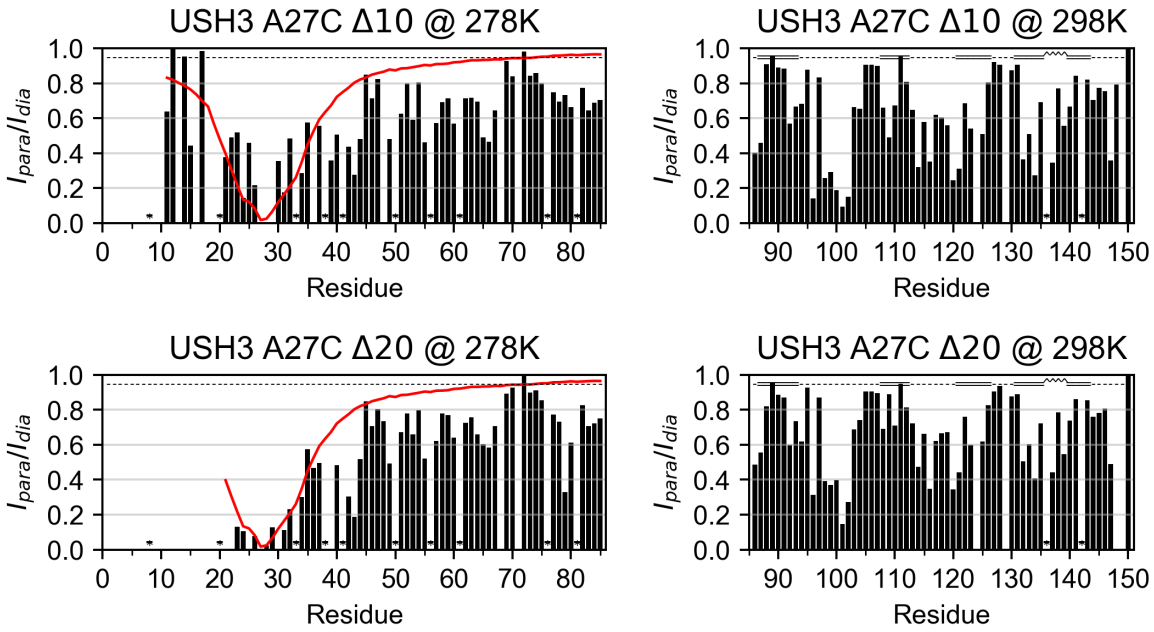


Figure 5.11: PRE profiles and heat map for USH3 A27C Δ10 and USH3 A27C Δ20.

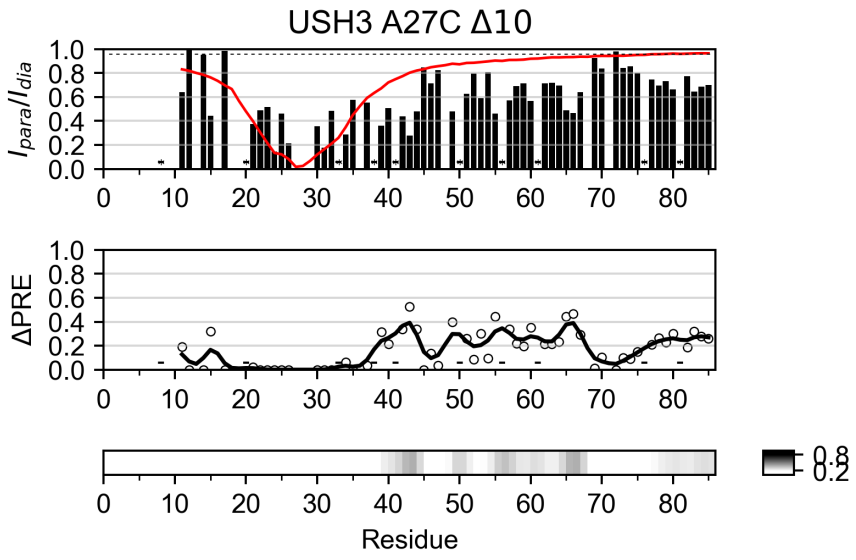


Figure 5.12: PRE and ΔPRE profiles and heat map for USH3 A27C Δ10 IDR.

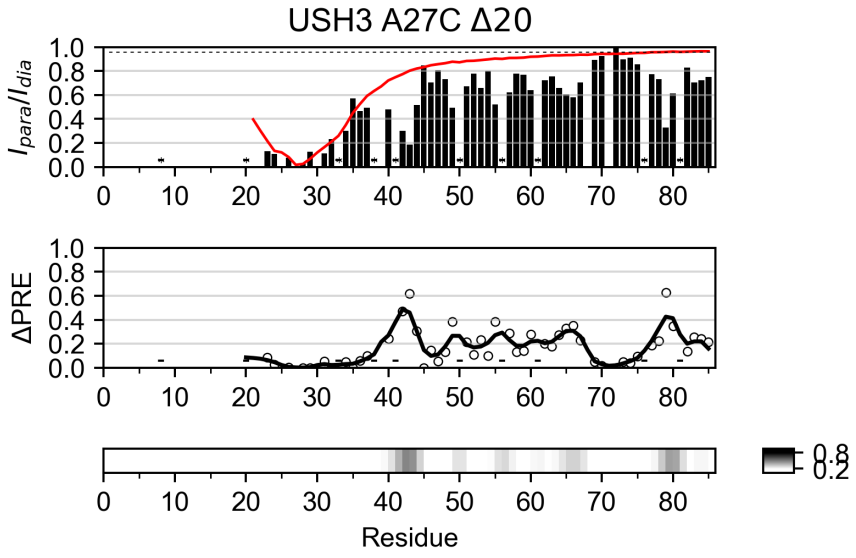


Figure 5.13: PRE and ΔPRE profiles and heat map for USH3 A27C Δ20 IDR.

5.5 PRE and ΔPRE data sets of SH4-UD F#A mutants

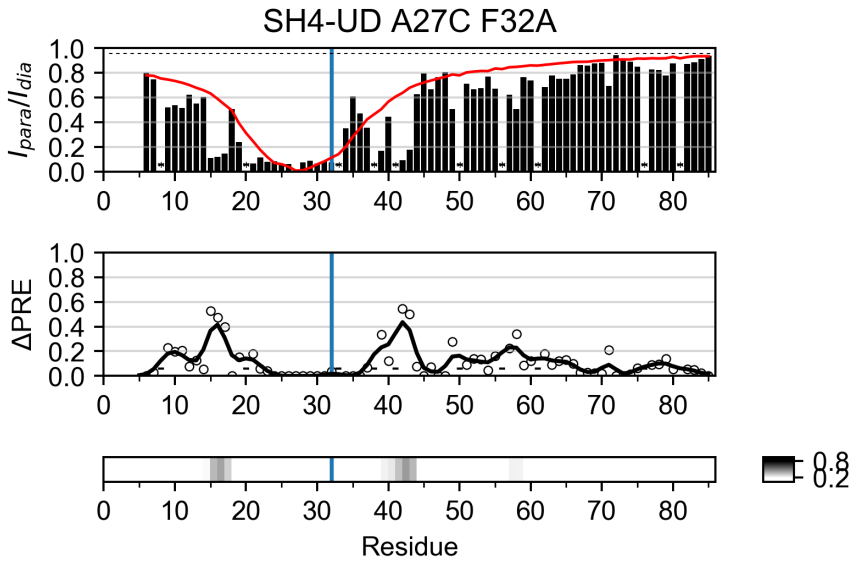


Figure 5.14: PRE and ΔPRE profiles and heat map for SH4-UD A27C F32A.

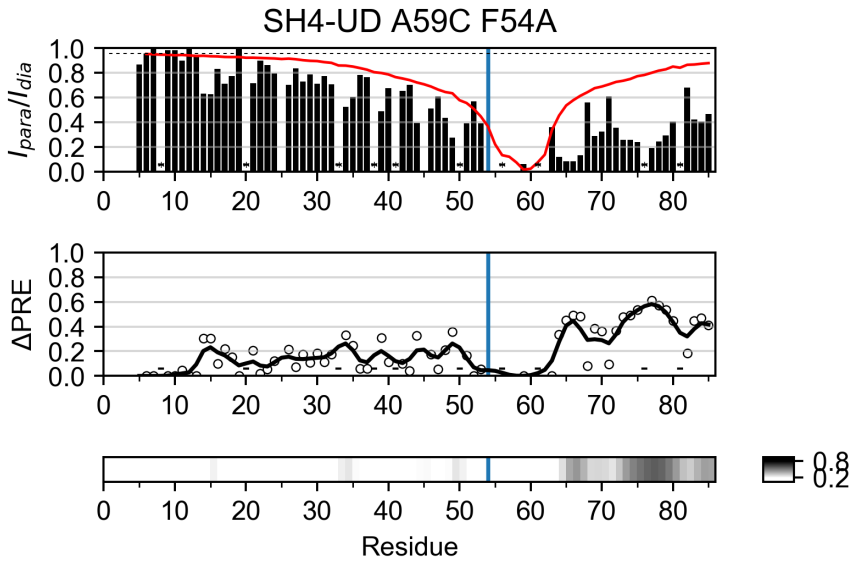


Figure 5.15: PRE and ΔPRE profiles and heat map for SH4-UD A27C F52A.

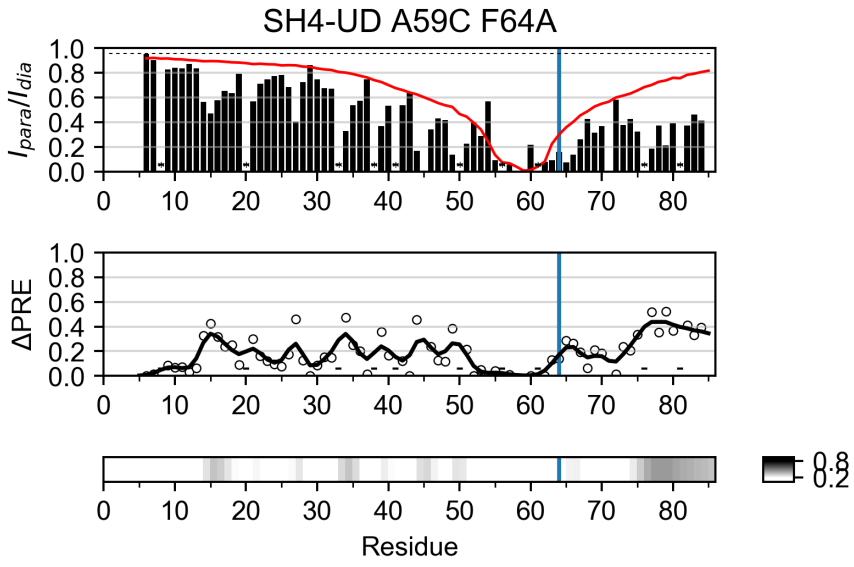


Figure 5.16: PRE and ΔPRE profiles and heat map for SH4-UD A27C F64A.

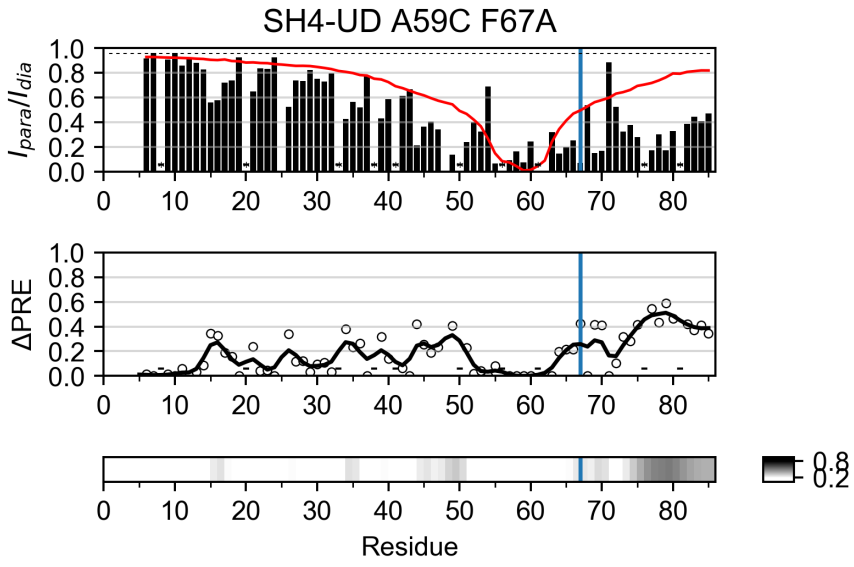


Figure 5.17: PRE and ΔPRE profiles and heat map for SH4-UD A27C F67A.

5.6 Histidine signal variability between identical samples

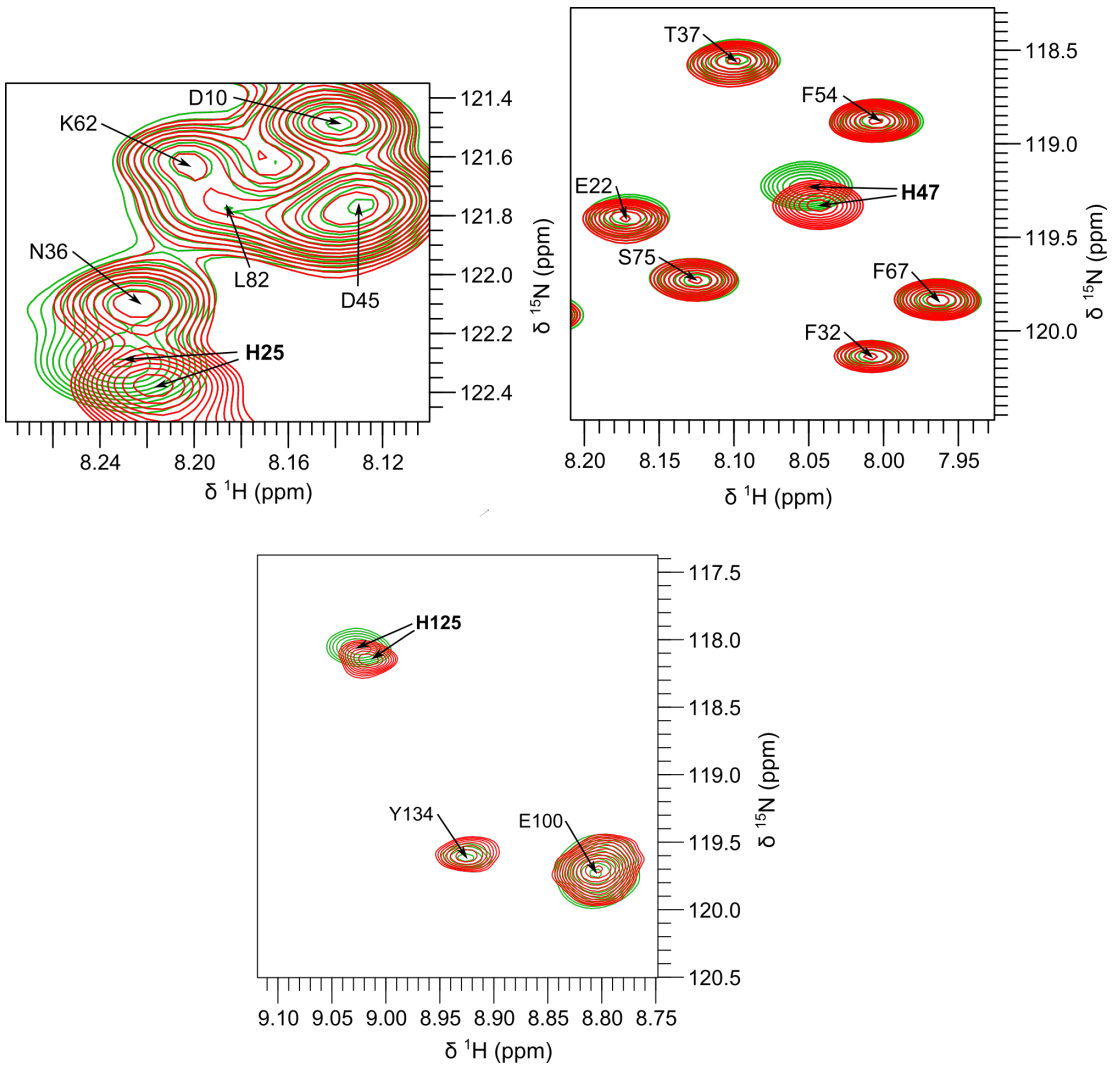


Figure 5.18: Comparison of H25, H47, and H125 between $^{1}H-^{15}N$ SOFAST HMQC spectra of two USH3 WT samples in nearly identical experimental conditions (temperature, buffer, concentration, etc., see Methods and Materials for details).

5.7 PRE and ΔPRE data sets of SH4-UD pS17

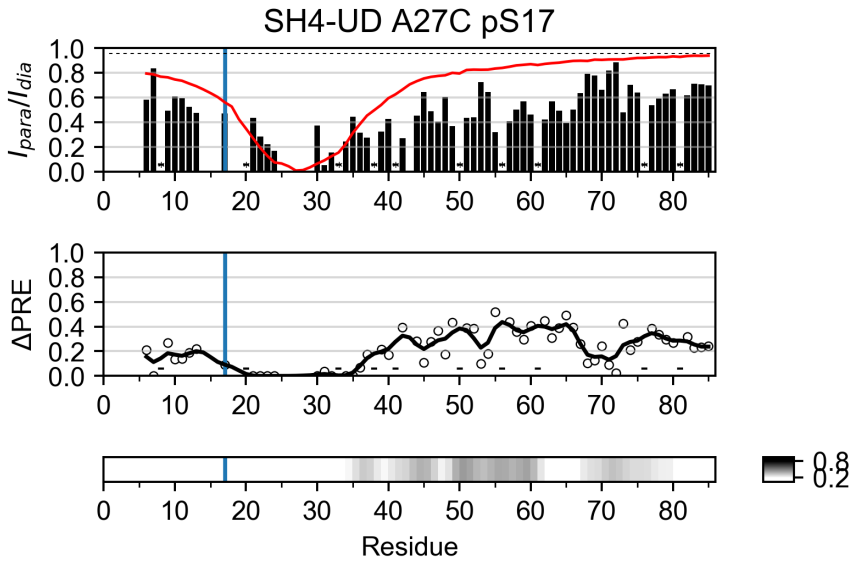


Figure 5.19: PRE and ΔPRE profiles and heat map for SH4-UD A27C pS17.

References

Abyzov, A. et al., 2016. Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation. *J. Am. Chem. Soc.*, 138(19), pp.6240–6251.

Ahmed, M.A. et al., 2009. Induced Secondary Structure and Polymorphism in an Intrinsically Disordered Structural Linker of the CNS: Solid-State NMR and FTIR Spectroscopy of Myelin Basic Protein Bound to Actin. *Biophys. J.*, 96(1), pp.180–191.

Akbey, Ü. & Oschkinat, H., 2016. Structural biology applications of solid state MAS DNP NMR. *J. Magn. Reson.*, 269, pp.213–224.

Amata, I., Maffei, M. & Pons, M., 2014. Phosphorylation of unique domains of Src family kinases. *Front. Genet.*, 5(June), pp.1–6.

Amata, I. et al., 2013. Multi-phosphorylation of the Intrinsically Disordered Unique Domain of c-Src Studied by In-Cell and Real-Time NMR Spectroscopy. *ChemBioChem*, 14(14), pp.1820–1827.

Anfinsen, C.B., 1973. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), pp.223–230.

Anon, 1993. Increase in activity and level of pp60c-src in progressive stages of human colorectal cancer. *J. Clin. Invest.*, 91(1), pp.53–60.

Antonov, L.D. et al., 2016. Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.*, 18(8), pp.5832–5838.

Arbesú, M. et al., 2017. The Unique Domain Forms a Fuzzy Intramolecular Complex in Src Family Kinases. *Structure*, 25(4), pp.630–640.e4.

Arold, S. et al., 1998. RT Loop Flexibility Enhances the Specificity of Src Family SH3 Domains for HIV-1 Nef. *Biochemistry*, 37(42), pp.14683–14691.

Babu, M.M. et al., 2011. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, 21(3), pp.432–440.

Bacarizo, J. & Camara-Artigas, A., 2013. Atomic resolution structures of the c-Src SH3 domain in complex with two high-affinity peptides from classes i and II. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 69(5), pp.756–766.

Bah, A. & Forman-Kay, J.D., 2016. Modulation of intrinsically disordered protein function by post-

translational modifications. *J. Biol. Chem.*, 291(13), pp.6696–6705.

Bah, A. et al., 2015. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*, 519(7541), pp.106–109.

Ban, D. et al., 2017. A Small Molecule Causes a Population Shift in the Conformational Landscape of an Intrinsically Disordered Protein. *J. Am. Chem. Soc.*, 139(39), pp.13692–13700.

Barnes, A.B. et al., 2010. Resolution and polarization distribution in cryogenic DNP/MAS experiments. *Phys. Chem. Chem. Phys.*, 12(22), pp.5861–5867.

Battiste, J.L. & Wagner, G., 2000. Utilization of Site-Directed Spin Labeling and High-Resolution Heteronuclear Nuclear Magnetic Resonance for Global Fold Determination of Large Proteins with Limited Nuclear Overhauser Effect Data †. *Biochemistry*, 39(18), pp.5355–5365.

Baxa, M.C. et al., 2014. Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci.*, 111(43), pp.15396–15401.

Beauchamp, K.A., Pande, V.S. & Das, R., 2014. Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. *Biophys. J.*, 106(6), pp.1381–1390.

Becerra, L.R. et al., 1993. Dynamic nuclear polarization with a cyclotron resonance maser at 5 T. *Phys. Rev. Lett.*, 71(21), pp.3561–3564.

Bellay, J. et al., 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.*, 12(2), p.R14.

Berjanskii, M.V. & Wishart, D.S., 2017. Unraveling the meaning of chemical shifts in protein NMR. *Biochim. Biophys. Acta - Proteins Proteomics*, 1865(11), pp.1564–1576.

Berlin, K. et al., 2013. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.*, 135(44), pp.16595–16609.

Berlow, R.B., Dyson, H.J. & Wright, P.E., 2015. Functional advantages of dynamic protein disorder. *FEBS Lett.*, 589(19PartA), pp.2433–2440.

Bernadó, P. et al., 2007. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 129(17), pp.5656–5664.

Bernadó, P. et al., 2008. Structural Characterization of the Active and Inactive States of Src Kinase in Solution by Small-Angle X-ray Scattering. *J. Mol. Biol.*, 376(2), pp.492–505.

Bertini, I. & Luchinat, C., 1986. *NMR of paramagnetic molecules in biological systems*, Benjamin/Cummings Pub. Co.

Bertini, I. et al., 2010. Conformational Space of Flexible Biological Macromolecules from Average Data. *J. Am. Chem. Soc.*, 132(38), pp.13553–13558.

Best, R.B., 2017. Computational and theoretical advances in studies of intrinsically disordered proteins.

Curr. Opin. Struct. Biol., 42, pp.147–154.

Bhattacharyya, R.P., 2006. The Ste5 Scaffold Allosterically Modulates Signaling Output of the Yeast Mating Pathway. *Science*, 311(5762), pp.822–826.

Bister, K., 2015. Discovery of oncogenes: The advent of molecular cancer research. *Proc. Natl. Acad. Sci.*, 112(50), pp.15259–15260.

Bjorge, J.D., Jakymiw, A. & Fujita, D.J., 2000. Selected glimpses into the activation and function of Src kinase. *Oncogene*, 19(49), pp.5620–5635.

Bloembergen, N. & Morgan, L.O., 1961. Proton relaxation times in paramagnetic solutions. Effects of electron spin relaxation. *J. Chem. Phys.*, 34(3), pp.842–850.

Blume-Jensen, P. & Hunter, T., 2001. Oncogenic kinase signalling. *Nature*, 411(6835), pp.355–365.

Boehr, D.D., Nussinov, R. & Wright, P.E., 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11), pp.789–796.

Boggon, T.J. & Eck, M.J., 2004. Structure and regulation of Src family kinases. *Oncogene*, 23(48), pp.7918–7927.

Bonomi, M. et al., 2016. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.*, 2(1), p.e1501177.

Bonomi, M. et al., 2017. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.*, 42, pp.106–116.

Braxenthaler, M. et al., 1997. Chaos in protein dynamics. *Proteins Struct. Funct. Genet.*, 29(November 1996), pp.417–425.

Brookes, D.H. & Head-Gordon, T., 2016. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J. Am. Chem. Soc.*, 138(13), pp.4530–4538.

Brown, C.J. et al., 2011. Evolution and disorder. *Curr. Opin. Struct. Biol.*, 21(3), pp.441–446.

Brown, C.J. et al., 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.*, 55(1), pp.104–110.

Brugge, J.S., Steinbaugh, P.J. & Erikson, R.L., 1978. Characterization of the avian sarcoma virus protein p60src. *Virology*, 91(1), pp.130–140.

Buser, C. a et al., 1994. Membrane binding of myristylated peptides corresponding to the NH2 terminus of Src. *Biochemistry*, 33(44), pp.13093–13101.

Camilloni, C., Cavalli, A. & Vendruscolo, M., 2013. Replica-Averaged Metadynamics. *J. Chem. Theory Comput.*, 9(12), pp.5610–5617.

Campen, A. et al., 2008. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.*, 15(9), pp.956–963.

Capdeville, R. et al., 2002. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug.

Nat. Rev. Drug Discov., 1(7), pp.493–502.

Cartwright, C.A. & Eckhart, W., 1990. Activation of the pp6Oc-sr protein kinase is an early event in colonic carcinogenesis. *Biochemistry*, 87(2), pp.558–562.

Carver, T.R. & Slichter, C.P., 1953. Polarization of Nuclear Spins in Metals. *Phys. Rev.*, 92(1), pp.212–213.

Castellani, F. et al., 2002. Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature*, 420(6911), pp.98–102.

Chemes, L.B. et al., 2012. Sequence Evolution of the Intrinsically Disordered and Globular Domains of a Model Viral Oncoprotein B. Xue, ed. *PLoS One*, 7(10), p.e47661.

Chen, C.Y.-C. & Tou, W.I., 2013. How to design a drug for the disordered proteins? *Drug Discov. Today*, 18(19-20), pp.910–915.

Chen, Y., Campbell, S.L. & Dokholyan, N.V., 2007. Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection. *Biophys. J.*, 93(7), pp.2300–2306.

Cheng, S., Cetinkaya, M. & Gräter, F., 2010. How sequence determines elasticity of disordered proteins. *Biophys. J.*, 99(12), pp.3863–3869.

Cheng, Y. et al., 2006. Abundance of Intrinsic Disorder in Protein Associated with Cardiovascular Disease. *Biochemistry*, 45(35), pp.10448–10460.

Chevalier, A. et al., 2017. Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674), pp.74–79.

Chodera, J.D. & Mobley, D.L., 2013. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu. Rev. Biophys.*, 42(1), pp.121–142.

Choy, W.-Y. & Forman-Kay, J.D., 2001. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.*, 308(5), pp.1011–1032.

Chu, S. et al., 2010. Solid-state NMR paramagnetic relaxation enhancement immersion depth studies in phospholipid bilayers. *J. Magn. Reson.*, 207(1), pp.89–94.

Clore, G.M. & Iwahara, J., 2009. Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. *Chem. Rev.*, 109(9), pp.4108–4139.

Cock, P.J.A. et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), pp.1422–1423.

Continuum Analytics, 2015. Datashader.

Cordier, F. et al., 2000. Ligand-induced strain in hydrogen bonds of the c-Src SH3 domain detected by NMR. *J. Mol. Biol.*, 304(4), pp.497–505.

Cossio, P. & Hummer, G., 2013. Bayesian analysis of individual electron microscopy images: Towards

structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.*, 184(3), pp.427–437.

Cowan-Jacob, S.W. et al., 2005. The Crystal Structure of a c-Src Complex in an Active Conformation Suggests Possible Steps in c-Src Activation. *Structure*, 13(6), pp.861–871.

Cowan-Jacob, S.W., Jahnke, W. & Knapp, S., 2014. Novel approaches for targeting kinases: allosteric inhibition, allosteric activation and pseudokinases. *Future Med. Chem.*, 6(5), pp.541–561.

Crabtree, M.D. et al., 2017. Conserved Helix-Flanking Prolines Modulate Intrinsically Disordered Protein:Target Affinity by Altering the Lifetime of the Bound Complex. *Biochemistry*, 56(18), pp.2379–2384.

Crooks, G.E., 2004. WebLogo: A Sequence Logo Generator. *Genome Res.*, 14(6), pp.1188–1190.

Csizmok, V. et al., 2016. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem. Rev.*, 116(11), pp.6424–6462.

Cumberworth, A. et al., 2013. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.*, 454(3), pp.361–369.

Dar, A.C. & Shokat, K.M., 2011. The Evolution of Protein Kinase Inhibitors from Antagonists to Agonists of Cellular Signaling. *Annu. Rev. Biochem.*, 80(1), pp.769–795.

Das, R.K. & Pappu, R.V., 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.*, 110(33), pp.13392–13397.

Das, R.K., Ruff, K.M. & Pappu, R.V., 2015. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 32, pp.102–112.

Davey, N.E., Travé, G. & Gibson, T.J., 2011. How viruses hijack cell regulation. *Trends Biochem. Sci.*, 36(3), pp.159–169.

Davey, N.E. et al., 2012. Attributes of short linear motifs. *Mol. BioSyst.*, 8(1), pp.268–281.

Delaglio, F. et al., 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, 6(3), pp.277–293.

Dill, K.A. & Chan, H.S., 1997. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.*, 4(1), pp.10–19.

Dinkel, H. et al., 2016. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, 44(D1), pp.D294–D300.

Dobzhansky, T., 1973. Nothing in Biology Makes Sense except in the Light of Evolution. *Am. Biol. Teach.*, 35(3), pp.125–129.

Dogan, J., Gianni, S. & Jemth, P., 2014. The binding mechanisms of intrinsically disordered proteins. *Phys. Chem. Chem. Phys.*, 16(14), pp.6323–6331.

Dosztanyi, Z. et al., 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16), pp.3433–3434.

Dubois, F. et al., 2015. YES oncogenic activity is specified by its SH4 domain and regulates RAS/MAPK

signaling in colon carcinoma cells. *Am J Cancer Res*, 5(6), pp.1972–1987.

Duer, M.J., 2004. *Introduction to solid-state NMR spectroscopy*, Blackwell.

Dunker, A. et al., 2001. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19(1), pp.26–59.

Dunker, A.K. et al., 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.*, pp.473–484.

Dyson, H. & Wright, P.E., 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12(1), pp.54–60.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5), pp.1792–1797.

Edsall, J., 1995. Hsien Wu and the First Theory of Protein Denaturation (1931). In *Adv. protein chem. vol. 46*. Elsevier, pp. 1–5.

Encapsula NanoSciences LCC, 2009. Liposome: Encapsula's Scientific Blog: The Number of Lipid Molecules per Liposome.

Engen, J.R. et al., 2008. Structure and dynamic regulation of Src-family kinases. *Cell. Mol. life Sci. C.*, 65(19), pp.3058–3073.

Erpel, T., Superti-Furga, G. & Courtneidge, S.A., 1995. Mutational analysis of the Src SH3 domain: the same residues of the ligand binding surface are important for intra- and intermolecular interactions. *Embo J*, 14(5), pp.963–975.

Fajer, M., Meng, Y. & Roux, B., 2017. The Activation of c-Src Tyrosine Kinase: Conformational Transition Pathway and Free Energy Landscape. *J. Phys. Chem. B*, 121(15), pp.3352–3363.

Fallacara, A.L. et al., 2014. Insight into the Allosteric Inhibition of Abl Kinase. *J. Chem. Inf. Model.*, 54(5), pp.1325–1338.

Felli, I.C. & Pierattelli, R., 2015. *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* I. C. Felli & R. Pierattelli, eds., Cham: Springer International Publishing.

Feng, S. et al., 1995. Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. *Proc. Natl. Acad. Sci. U. S. A.*, 92(26), pp.12408–124015.

Filippakopoulos, P., Müller, S. & Knapp, S., 2009. SH2 domains: modulators of nonreceptor tyrosine kinase activity. *Curr. Opin. Struct. Biol.*, 19(6), pp.643–649.

Finn, R.D. et al., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44(D1), pp.D279–D285.

Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft*, 27(3), pp.2985–2993.

Fisher, C.K. & Stultz, C.M., 2011. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 21(3), pp.426–431.

Fisher, C.K., Huang, A. & Stultz, C.M., 2010. Modeling Intrinsically Disordered Proteins with Bayesian

Statistics. *J. Am. Chem. Soc.*, 132(42), pp.14919–14927.

Flamm, A.G. et al., 2015. N-lauroylation during the expression of recombinant N-myristoylated proteins. Implications and solutions. *ChemBioChem*, 17(1), pp.82–89.

Flock, T. et al., 2014. Controlling entropy to tune the functions of intrinsically disordered regions. *Curr. Opin. Struct. Biol.*, 26, pp.62–72.

Flory, P.J. & Volkenstein, M., 1969. Statistical mechanics of chain molecules. *Biopolymers*, 8(5), pp.699–700.

Foda, Z.H. et al., 2015. A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nat. Commun.*, 6, p.5939.

Follis, A.V. et al., 2013. PUMA binding induces partial unfolding within BCL-xL to disrupt p53 binding and promote apoptosis. *Nat. Chem. Biol.*, 9(3), pp.163–168.

Foster, M.P., McElroy, C.A. & Amero, C.D., 2007. Solution NMR of Large Molecules and Assemblies †. *Biochemistry*, 46(2), pp.331–340.

Frauenfelder, H., Sligar, S. & Wolynes, P., 1991. The energy landscapes and motions of proteins. *Science*, 254(5038), pp.1598–1603.

Frederick, K.K. et al., 2007. Conformational entropy in molecular recognition by proteins. *Nature*, 448(7151), pp.325–329.

Fritzsching, K.J., Hong, M. & Schmidt-Rohr, K., 2016. Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. *J. Biomol. NMR*, 64(2), pp.115–130.

Fritzsching, K.J. et al., 2013. Practical use of chemical shift databases for protein solid-state NMR: 2D chemical shift maps and amino-acid assignment with secondary-structure information. *J. Biomol. NMR*, 56(2), pp.155–167.

Frueh, D.P. et al., 2013. NMR methods for structural studies of large monomeric and multimeric proteins. *Curr. Opin. Struct. Biol.*, 23(5), pp.734–739.

Fuxreiter, M., 2012. Fuzziness: linking regulation to protein dynamics. *Mol. BioSyst.*, 8(1), pp.168–177.

Fuxreiter, M. & Tompa, P. eds., 2012. *Fuzziness*, New York, NY: Springer US.

Ganguly, D. & Chen, J., 2009. Structural Interpretation of Paramagnetic Relaxation Enhancement-Derived Distances for Disordered Protein States. *J. Mol. Biol.*, 390(3), pp.467–477.

Ganguly, D. et al., 2012. Electrostatically Accelerated Coupled Binding and Folding of Intrinsically Disordered Proteins. *J. Mol. Biol.*, 422(5), pp.674–684.

Garcia-Pino, A. et al., 2010. Allostery and Intrinsic Disorder Mediate Transcription Regulation by Conditional Cooperativity. *Cell*, 142(1), pp.101–111.

Garner et al., 1998. Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite

Differing Structural Characterization. *Genome Inform. Ser. Workshop Genome Inform.*, 9, pp.201–213.

Gianni, S., Dogan, J. & Jemth, P., 2014. Distinguishing induced fit from conformational selection. *Biophys. Chem.*, 189, pp.33–39.

Gillespie, J.R. & Shortle, D., 1997a. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.*, 268(1), pp.158–69.

Gillespie, J.R. & Shortle, D., 1997b. Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J. Mol. Biol.*, 268(1), pp.170–184.

Gingrich, J.R. et al., 2004. Unique domain anchoring of Src to synaptic NMDA receptors via the mitochondrial protein NADH dehydrogenase subunit 2. *Proc. Natl. Acad. Sci.*, 101(16), pp.6237–6242.

Gonfloni, S. et al., 2000. Crosstalk between the catalytic and regulatory domains allows bidirectional regulation of Src. *Nat. Struct. Biol.*, 7(4), pp.281–286.

Gonfloni, S. et al., 1997. The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src. *EMBO J.*, 16(24), pp.7261–7271.

Gottlieb-Abraham, E. et al., 2016. The residue at position 5 of the N-terminal region of Src and Fyn modulates their myristoylation, palmitoylation, and membrane interactions. *Mol. Biol. Cell*, 27(24), pp.3926–3936.

Gruet, A. et al., 2016. Fuzzy regions in an intrinsically disordered protein impair protein-protein interactions. *FEBS J.*, 283(4), pp.576–594.

Gunasekaran, K. et al., 2003. Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.*, 28(2), pp.81–85.

Hadži, S. et al., 2017. The Thermodynamic Basis of the Fuzzy Interaction of an Intrinsically Disordered Protein. *Angew. Chemie Int. Ed.*, pp.1–5.

Harrison, S.C., 2003. Variation on an Src-like Theme. *Cell*, 112(6), pp.737–740.

He, B. et al., 2009. Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, 19(8), pp.929–949.

Hegyi, H. & Gerstein, M., 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, 11(10), pp.1632–1640.

Heller, G.T. et al., 2017. Sequence Specificity in the Entropy-Driven Binding of a Small Molecule and a Disordered Peptide. *J. Mol. Biol.*, 429(18), pp.2772–2779.

Hennel, J.W. & Klinowski, J., 2005. Magic-Angle Spinning: a Historical Perspective. In *New tech. solid-state nmr*. Springer, Berlin, Heidelberg, pp. 1–14.

Hess, B. et al., 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3), pp.435–447.

Hidalgo, P. & MacKinnon, R., 1995. Revealing the architecture of a K+ channel pore through mutant

cycles with a peptide inhibitor. *Science*, 268(5208), pp.307–310.

Higman, V.A. et al., 2009. Assigning large proteins in the solid state: a MAS NMR resonance assignment strategy using selectively and extensively 13C-labelled proteins. *J. Biomol. NMR*, 44(4), pp.245–260.

Hiipakka, M. & Saksela, K., 2007. Versatile retargeting of SH3 domain binding by modification of non-conserved loop residues. *FEBS Lett.*, 581, pp.1735–1741.

Ho, B.K. & Brasseur, R., 2005. No Title. *BMC Struct. Biol.*, 5(1), p.14.

Hoey, J.G., Summy, J. & Flynn, D.C., 2000. Chimeric constructs containing the SH4/Unique domains of cYes can restrict the ability of Src(527F) to upregulate heme oxygenase-1 expression efficiently. *Cell. Signal.*, 12(9-10), pp.691–701.

Holehouse, A.S. et al., 2017. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.*, 112(1), pp.16–21.

Hu, K.-N. & Tycko, R., 2010. What can solid state NMR contribute to our understanding of protein folding? *Biophys. Chem.*, 151(1-2), pp.10–21.

Hu, K.-N. et al., 2004. Dynamic Nuclear Polarization with Biradicals. *J. Am. Chem. Soc.*, 126(35), pp.10844–10845.

Huebner, R.J. & Todaro, G.J., 1969. Oncogenes of RNA tumor viruses as determinants of cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 64(3), pp.1087–94.

Hummer, G. & Köfinger, J., 2015. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.*, 143(24), p.243150.

Hunter, T. & Sefton, B.M., 1980. Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. *Proc. Natl. Acad. Sci. U. S. A.*, 77(3), pp.1311–1315.

Huse, M. & Kuriyan, J., 2002. The Conformational Plasticity of Protein Kinases. *Cell*, 109(3), pp.275–282.

Iakoucheva, L.M., 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32(3), pp.1037–1049.

Iakoucheva, L.M. et al., 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 323(3), pp.573–584.

Iconaru, L.I. et al., 2015. Discovery of Small Molecules that Inhibit the Disordered Protein, p27 Kip1. *Sci. Rep.*, 5(1), p.15686.

Irby, R.B. & Yeatman, T.J., 2000. Role of Src expression and activation in human cancer. *Oncogene*, 19(49), pp.5636–5642.

Ishizawar, R. & Parsons, S.J., 2004. C-Src and cooperating partners in human cancer. *Cancer Cell*, 6(3), pp.209–214.

Iwahara, J., Tang, C. & Marius Clore, G., 2007. Practical aspects of 1H transverse paramagnetic relax-

ation enhancement measurements on macromolecules. *J. Magn. Reson.*, 184(2), pp.185–195.

Jensen, M.R., Ruigrok, R.W. & Blackledge, M., 2013. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.*, 23(3), pp.426–435.

Jensen, M.R. et al., 2014. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.*, 114(13), pp.6632–6660.

Jeon, J. et al., 2011. Molecular Evolution of Protein Conformational Changes Revealed by a Network of Evolutionarily Coupled Residues. *Mol. Biol. Evol.*, 28(9), pp.2675–2685.

Jeong, C.-S. & Kim, D., 2011. Coevolved Rresidues and the Functional Association for Intrinsically Disordered Proteins. In *Biocomput. 2012*. WORLD SCIENTIFIC, pp. 140–151.

Joerger, A.C. & Fersht, A.R., 2008. Structural Biology of the Tumor Suppressor p53. *Annu. Rev. Biochem.*, 77(1), pp.557–582.

Johnson, L., 2009. The regulation of protein phosphorylation. *Biochem. Soc. Trans.*, 37(4), pp.627–641.

Jones, E. et al., 2001. SciPy: Open Source Scientific Tools for Python.

Jorgensen, W.L. et al., 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2), pp.926–935.

Juan, D. de, Pazos, F. & Valencia, A., 2013. Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, 14(4), pp.249–261.

Kamisetty, H., Ovchinnikov, S. & Baker, D., 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.*, 110(39), pp.15674–15679.

Karush, F., 1950. Heterogeneity of the Binding Sites of Bovine Serum Albumin. *J. Am. Chem. Soc.*, 72(6), pp.2705–2713.

Kathiriya, J.J. et al., 2014. Presence and utility of intrinsically disordered regions in kinases. *Mol. BioSyst.*, 10(11), pp.2876–2888.

Kay, B.K., 2012. SH3 domains come of age. *FEBS Lett.*, 586(17), pp.2606–2608.

Kazimierczuk, K. & Orekhov, V., 2015. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magn. Reson. Chem.*, 53(11), pp.921–926.

Keeler, J., 2010. *Understanding NMR spectroscopy*, John Wiley; Sons.

Kendrew, J.C. et al., 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610), pp.662–666.

Kini, R., 1998. Proline brackets and identification of potential functional sites in proteins: Toxins to therapeutics. *Toxicon*, 36(11), pp.1659–1670.

Kini, R. & Evans, H., 1995. A Hypothetical Structural Role for Proline Residues in the Flanking Segments

of Protein-Protein Interaction Sites. *Biochem. Biophys. Res. Commun.*, 212(3), pp.1115–1124.

Kocherginsky, N. & Swartz, H.M., 1995. *Nitroxide spin labels : reactions in biology and chemistry*, CRC Press.

Koga, N. et al., 2012. Principles for designing ideal protein structures. *Nature*, 491(7423), pp.222–227.

Kohn, J.E. et al., 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci.*, 101(34), pp.12491–12496.

Konrat, R., 2010. The Meandering of Disordered Proteins in Conformational Space. *Structure*, 18(4), pp.416–419.

Konrat, R., 2009. The protein meta-structure: a novel concept for chemical and molecular biology. *Cell. Mol. Life Sci.*, 66(22), pp.3625–3639.

Koscielny, G. et al., 2017. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, 45(D1), pp.D985–D994.

Kosen, P.A., 1989. Spin labeling of proteins. In *Methods enzymol.* Academic Press, pp. 86–121.

Koshland, D.E., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.*, 44(2), pp.98–104.

Koshland, D.E., Némethy, G. & Filmer, D., 1966. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1), pp.365–385.

Kotta-Loizou, I., Tsaousis, G.N. & Hamodrakas, S.J., 2013. Analysis of Molecular Recognition Features (MoRFs) in membrane proteins. *Biochim. Biophys. Acta - Proteins Proteomics*, 1834(4), pp.798–807.

Kragelj, J. et al., 2013. Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts. *ChemPhysChem*, 14(13), pp.3034–3045.

Kriwacki, R.W. et al., 1996. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci.*, 93(21), pp.11504–11509.

Kumar, A. & Balbach, J., 2015. Real-time protein NMR spectroscopy and investigation of assisted protein folding. *Biochim. Biophys. Acta - Gen. Subj.*, 1850(10), pp.1965–1972.

Ladbury, J.E. & Arold, S., 2000. Searching for specificity in SH domains. *Chem. Biol.*, 7(1), pp.R3–R8.

Larson, S.M. & Davidson, A.R., 2000. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.*, 9(11), pp.2170–2180.

Le Roux, A.-L. et al., 2016. Kinetics characterization of c-Src binding to lipid membranes: Switching from labile to persistent binding. *Colloids Surfaces B Biointerfaces*, 138, pp.17–25.

Le Roux, A.-L. et al., 2016. Single molecule fluorescence reveals dimerization of myristoylated Src N-terminal region on supported lipid bilayers. *ChemistrySelect*, 1(4), pp.642–647.

Lee, R. van der et al., 2014. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.*,

114(13), pp.6589–6631.

Leung, H.T.A. et al., 2016. A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.*, 12(1), pp.383–394.

Levinthal, C., 1969. How to fold graciously. In *Mössbaun spectrosc. biol. syst. proc. univ. illinois bull.* pp. 22–24.

Levitt, M.H., 2013. *Spin Dynamics : Basics of Nuclear Magnetic Resonance.*, Wiley.

Li, J. et al., 2015. An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014. *Int. J. Mol. Sci.*, 16(10), pp.23446–23462.

Liao, S.Y. et al., 2016. Efficient DNP NMR of membrane proteins: sample preparation protocols, sensitivity, and radical location. *J. Biomol. NMR*, 64(3), pp.223–237.

Lipari, G. & Szabo, A., 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, 104(17), pp.4546–4559.

Liu, Y., Matthews, K.S. & Bondos, S.E., 2008. Multiple Intrinsically Disordered Sequences Alter DNA Binding by the Homeodomain of the Drosophila Hox Protein Ultrabithorax. *J. Biol. Chem.*, 283(30), pp.20874–20887.

Liu, Z. & Huang, Y., 2014. Advantages of proteins being disordered. *Protein Sci.*, 23(5), pp.539–550.

Lockless, S.W., 1999. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286(5438), pp.295–299.

Luchinat, E. & Banci, L., 2017. In-cell NMR: a topical review. *IUCrJ*, 4(2), pp.108–118.

Maffei, M., 2015. *Structural and functional characterization of the intrinsically disordered Unique domain of c-Src.* PhD thesis. Universitat de Barcelona; Universitat de Barcelona.

Maffei, M. et al., 2015. The SH3 Domain Acts as a Scaffold for the N-Terminal Intrinsically Disordered Regions of c-Src. *Structure*, 23(5), pp.893–902.

Maffei, M. et al., 2013. Lipid Binding by Disordered Proteins. *Protoc. Exch.*

Malaney, P. et al., 2013. Intrinsic Disorder in PTEN and its Interactome Confers Structural Plasticity and Functional Versatility. *Sci. Rep.*, 3(1), p.2035.

Manning, G., 2002. The Protein Kinase Complement of the Human Genome. *Science*, 298(5600), pp.1912–1934.

Manning, G. et al., 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.*, 27(10), pp.514–520.

Mao, A.H. et al., 2010. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.*, 107(18), pp.8183–8188.

Mao, A.H., Lyle, N. & Pappu, R.V., 2013. Describing sequence–ensemble relationships for intrinsically

disordered proteins. *Biochem. J.*, 449(2), pp.307–318.

Marinelli, F. & Faraldo-Gómez, J.D., 2015. Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions. *Biophys. J.*, 108(12), pp.2779–2782.

Martin, G.S., 2001. The hunting of the Src. *Nat. Rev. Mol. Cell Biol.*, 2(6), pp.467–75.

Martin, G.S., 2004. The road to Src. *Oncogene*, 23(48), pp.7910–7917.

Martín-García, J.M. et al., 2007. Crystallographic structure of the SH3 domain of the human c-Yes tyrosine kinase: Loop flexibility and amyloid aggregation. *FEBS Lett.*, 581(9), pp.1701–1706.

Matsuda, M. et al., 1990. Binding of transforming protein, P47gag-crk, to a broad range of phosphotyrosine-containing proteins. *Science*, 248(4962), pp.1537–1539.

Mayer, B.J., 2001. SH3 domains: complexity in moderation. *J. Cell Sci.*, 114(Pt 7), pp.1253–1263.

Mayer, B.J., Hamaguchi, M. & Hanafusa, H., 1988. A novel viral oncogene with structural similarity to phospholipase C. *Nature*, 332(6161), pp.272–275.

McMeekin, T., 1952. Milk proteins. *J. Food Prot.*, (15), pp.59–63.

Mentink-Vigier, F. et al., 2012. Fast passage dynamic nuclear polarization on rotating solids. *J. Magn. Reson.*, 224, pp.13–21.

Mészáros, B., Simon, I. & Dosztányi, Z., 2011. The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins. *Phys. Biol.*, 8(3), p.035003.

Mirsky, A.E. & Pauling, L., 1936. On the Structure of Native, Denatured, and Coagulated Proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 22(7), pp.439–447.

Miskei, M., Antal, C. & Fuxreiter, M., 2017. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.*, 45(D1), pp.D228–D235.

Miskei, M. et al., 2017. Fuzziness enables context dependence of protein interactions. *FEBS Lett.*, 591(17), pp.2682–2695.

Mittag, T. et al., 2010. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure*, 18(4), pp.494–506.

Mittag, T. et al., 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. U. S. A.*, 105(46), pp.17772–17777.

Mollica, L. et al., 2016. Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Front. Mol. Biosci.*, 3(September), pp.1–18.

Molnar, K.S. et al., 2014. Cys-Scanning Disulfide Crosslinking and Bayesian Modeling Probe the Transmembrane Signaling Mechanism of the Histidine Kinase, PhoQ. *Structure*, 22(9), pp.1239–1251.

Monod, J., Wyman, J. & Changeux, J.P., 1965. On the nature of allosteric transitions: a plausible model.

J. Mol. Biol., 12(1), pp.88–118.

Motlagh, H.N. et al., 2014. The ensemble nature of allostery. *Nature*, 508(7496), pp.331–339.

Moult, J., 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15(3), pp.285–289.

Mui, B., Chow, L. & Hope, M.J., 2003. Extrusion Technique to Generate Liposomes of Defined Size. In *Methods enzymol.* pp. 3–14.

Naranjo, Y., Pons, M. & Konrat, R., 2012. Meta-structure correlation in protein space unveils different selection rules for folded and intrinsically disordered proteins. *Mol. Biosyst.*, 8(1), pp.411–416.

Neira, J.L. et al., 2017. Identification of a Drug Targeting an Intrinsically Disordered Protein Involved in Pancreatic Adenocarcinoma. *Sci. Rep.*, 7(November 2016), p.39732.

Nodet, G. et al., 2009. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. *J. Am. Chem. Soc.*, 131(49), pp.17908–17918.

Nomura, K. et al., 2014. Solid-State NMR Spectra of Lipid-Anchored Proteins under Magic Angle Spinning. *J. Phys. Chem. B*, 118(9), pp.2405–2413.

Obara, Y., 2004. PKA phosphorylation of Src mediates Rap1 activation in NGF and cAMP signaling in PC12 cells. *J. Cell Sci.*, 117(25), pp.6085–6094.

Olsen, J.G., Teilum, K. & Kragelund, B.B., 2017. Behaviour of intrinsically disordered proteins in protein–protein complexes with an emphasis on fuzziness. *Cell. Mol. Life Sci.*, pp.1–9.

Olsson, S. et al., 2013. Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data N. Fernandez-Fuentes, ed. *PLoS One*, 8(11), p.e79439.

OpenWetWare, 2017. http://www.openwetware.org/index.php?title=Main_Page&oldid=992714.

Orekhov, V.Y. & Jaravine, V.A., 2011. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog. Nucl. Magn. Reson. Spectrosc.*, 59(3), pp.271–292.

Otting, G., 2010. Protein NMR Using Paramagnetic Ions. *Annu. Rev. Biophys.*, 39(1), pp.387–405.

Overhauser, A.W., 1953. Polarization of Nuclei in Metals. *Phys. Rev.*, 92(2), pp.411–415.

Ozenne, V. et al., 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11), pp.1463–1470.

Palmer, A.G., Kroenke, C.D. & Patrick Loria, J., 2001. Nuclear Magnetic Resonance Methods for Quantifying Microsecond-to-Millisecond Motions in Biological Macromolecules. In *Methods enzymol.* Elsevier Masson SAS, pp. 204–238.

Palmer, A.G., Williams, J. & McDermott, A., 1996. Nuclear Magnetic Resonance Studies of Biopolymer Dynamics. *J. Phys. Chem.*, 100(31), pp.13293–13310.

Pancsa, R. & Fuxreiter, M., 2012. Interactions via intrinsically disordered regions: what kind of motifs?

IUBMB Life, 64(6), pp.513–520.

Patwardhan, P. & Resh, M.D., 2010. Myristoylation and Membrane Binding Regulate c-Src Stability and Kinase Activity. *Mol. Cell. Biol.*, 30(17), pp.4094–4107.

Paul, M.K., 2004. Tyrosine kinase – Role and significance in Cancer. *Int. J. Med. Sci.*, 52(7), p.101.

Pawson, T., 2004. Specificity in Signal Transduction. *Cell*, 116(2), pp.191–203.

Pejaver, V. et al., 2014. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.*, 23(8), pp.1077–1093.

Pelikan, M., Hura, G.L. & Hammel, M., 2009. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.*, 28(2), pp.174–189.

Pettersen, E.F. et al., 2004. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13), pp.1605–1612.

Pérez, Y. et al., 2009. Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: insights into the role of phosphorylation of the unique domain. *J. Mol. Biol.*, 391(1), pp.136–148.

Pérez, Y. et al., 2013. Lipid binding by the Unique and SH3 domains of c-Src suggests a new regulatory mechanism. *Sci. Rep.*, 3(1), p.1295.

Pines, A., Gibby, M.G. & Waugh, J.S., 1973. Proton-enhanced NMR of dilute spins in solids. *J. Chem. Phys.*, 59(2), pp.569–590.

Pitera, J.W. & Chodera, J.D., 2012. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.*, 8(10), pp.3445–3451.

Plaxco, K.W. & Groβ, M., 1997. The importance of being unfolded. *Nature*, 386(6626), pp.657–659.

Pollard, T. et al., 2016. Template for writing a PhD thesis in Markdown.

Pratt, J.W. & Gibbons, J.D., 1981. Kolmogorov-Smirnov Two-Sample Tests. In Springer New York, pp. 318–344.

Pufall, M.A., 2005. Variable Control of Ets-1 DNA Binding by Multiple Phosphates in an Unstructured Region. *Science*, 309(5731), pp.142–145.

Pujato, M. et al., 2005. pH Dependence of Amide Chemical Shifts in Natively Disordered Polypeptides Detects Medium-Range Interactions with Ionizable Residues. *Biophys. J.*, 89(5), pp.3293–3302.

Purusottam, R. et al., 2015. Probing the gel to liquid-crystalline phase transition and relevant conformation changes in liposomes by 13C magic-angle spinning NMR spectroscopy. *Biochim. Biophys. Acta - Biomembr.*, 1848(12), pp.3134–3139.

Ramachandran, G., Ramakrishnan, C. & Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1), pp.95–99.

Rauscher, S. et al., 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on

Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.*, 11(11), pp.5513–5524.

Ravera, E. et al., 2016. A critical assessment of methods to recover information from averaged data. *Phys. Chem. Chem. Phys.*, 18(8), pp.5686–5701.

Remmert, M. et al., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, 9(2), pp.173–175.

Resh, M.D., 1999. Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. *Biochim. Biophys. Acta - Mol. Cell Res.*, 1451(1), pp.1–16.

Resh, M.D., 1994. Myristylation and palmitylation of Src family members: The fats of the matter. *Cell*, 76(3), pp.411–413.

Resh, M.D., 2006. Trafficking and signaling by fatty-acylated and prenylated proteins. *Nat. Chem. Biol.*, 2(11), pp.584–590.

Rickles, R.J. et al., 1995. Phage display selection of ligand residues important for Src homology 3 domain binding specificity. *Proc. Natl. Acad. Sci. U. S. A.*, 92(24), pp.10909–10913.

Robitaille, T.P. et al., 2013. Astropy: A community Python package for astronomy. *Astron. Astrophys.*, 558, p.A33.

Romero, P. et al., 1998. Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.*, pp.437–448.

Romero, P. et al., 2001. Sequence complexity of disordered protein. *Proteins Struct. Funct. Genet.*, 42(1), pp.38–48.

Rosay, M., Blank, M. & Engelke, F., 2016. Instrumentation for solid-state dynamic nuclear polarization with magic angle spinning NMR. *J. Magn. Reson.*, 264, pp.88–98.

Rous, P., 1911. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J. Exp. Med.*, 13(4), pp.397–411.

Rous, P., 1910. A transmissible avian neoplasm. (Sarcoma of the common fowl.). *J. Exp. Med.*, 12(5), pp.696–705.

Roux, B. & Islam, S.M., 2013. Restrained-Ensemble Molecular Dynamics Simulations Based on Distance Histograms from Double Electron–Electron Resonance Spectroscopy. *J. Phys. Chem. B*, 117(17), pp.4733–4739.

Roux, B. & Weare, J., 2013. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.*, 138(8), p.084107.

Różycki, B., Kim, Y.C. & Hummer, G., 2011. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure*, 19(1), pp.109–116.

Sadowski, I., Stone, J.C. & Pawson, T., 1986. A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus

P130gag-fps. *Mol. Cell. Biol.*, 6(12), pp.4396–408.

Salvi, N., Abyzov, A. & Blackledge, M., 2017. Atomic resolution conformational dynamics of intrinsically disordered proteins from NMR spin relaxation. *Prog. Nucl. Magn. Reson. Spectrosc.*, 102-103, pp.43–60.

Salvi, N., Abyzov, A. & Blackledge, M., 2016. Multi-Timescale Dynamics in Intrinsically Disordered Proteins from NMR Relaxation and Molecular Simulation. *J. Phys. Chem. Lett.*, 7(13), pp.2483–2489.

Sancier, F. et al., 2011. Specific Oncogenic Activity of the Src-Family Tyrosine Kinase c-Yes in Colon Carcinoma Cells K. Anderson, ed. *PLoS One*, 6(2), p.e17237.

Sandilands, E., Brunton, V.G. & Frame, M.C., 2007. The membrane targeting and spatial activation of Src, Yes and Fyn is influenced by palmitoylation and distinct RhoB/RhoD endosome requirements. *J. Cell Sci.*, 120(15), pp.2555–2564.

Santos, H.G.D. & Siltberg-Liberles, J., 2016. Paralog-Specific Patterns of Structural disorder and phosphorylation in the vertebrate SH3-SH2-Tyrosine kinase protein family. *Genome Biol. Evol.*, 8(9), pp.2806–2825.

Sato, I. et al., 2009. Differential trafficking of Src, Lyn, Yes and Fyn is specified by the state of palmitoylation in the SH4 domain. *J. Cell Sci.*, 122(7), pp.965–975.

Sauvée, C. et al., 2013. Highly Efficient, Water-Soluble Polarizing Agents for Dynamic Nuclear Polarization at High Frequency. *Angew. Chemie - Int. Ed.*, 52(41), pp.10858–10861.

Schad, E. et al., 2017. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics*.

Schlessinger, A. et al., 2011. Protein disorder-a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.*, 21(3), pp.412–418.

Schneider, R. et al., 2012. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. BioSyst.*, 8(1), pp.58–68.

Schor, M., Mey, A.S.J.S. & MacPhee, C.E., 2016. Analytical methods for structural ensembles and dynamics of intrinsically disordered proteins. *Biophys. Rev.*, 8(4), pp.429–439.

Sethi, A. et al., 2013. Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics. *Chem. Phys.*, 422, pp.143–155.

Shannon, C.E. & Weaver, W., 1949. *The Mathematical Theory of Communication*, Univ of Illinois Press.

Sharma, R. et al., 2015. Fuzzy complexes: Specific binding without complete folding. *FEBS Lett.*, 589(19), pp.2533–2542.

Shaw, D.E. et al., 2010. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(6002), pp.341–346.

Shen, Y. & Bax, A., 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, 48(1), pp.13–22.

Shoemaker, B.A., Portman, J.J. & Wolynes, P.G., 2000. Speeding molecular recognition by using the

folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U. S. A.*, 97(16), pp.8868–73.

Sicheri, F. & Kuriyan, J., 1997. Structures of Src-family tyrosine kinases. *Curr. Opin. Struct. Biol.*, 7(6), pp.777–785.

Sigal, C.T. et al., 1994. Amino-terminal basic residues of Src mediate membrane binding through electrostatic interaction with acidic phospholipids. *Proc. Natl. Acad. Sci.*, 91(25), pp.12253–12257.

Siltberg-Liberles, J., 2011. Evolution of structurally disordered proteins promotes neostructuralization. *Mol. Biol. Evol.*, 28(1), pp.59–62.

Siltberg-Liberles, J., Grahnen, J.A. & Liberles, D.A., 2011. The evolution of protein structures and structural ensembles under functional constraint. *Genes (Basel).*, 2(4), pp.748–762.

Silverman, L., 1992. Lysine residues form an integral component of a novel NH2-terminal membrane targeting motif for myristylated pp60v-src. *J. Cell Biol.*, 119(2), pp.415–425.

Silvestre-Ryan, J. et al., 2013. Average conformations determined from PRE data provide high-resolution maps of transient tertiary interactions in disordered proteins. *Biophys. J.*, 104(8), pp.1740–51.

Skinner, S.P. et al., 2016. CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR*, 66(2), pp.111–124.

Slichter, C.P., 2014. The discovery and renaissance of dynamic nuclear polarization. *Reports Prog. Phys.*, 77(7), p.072501.

Smith, J.M., 1970. Natural selection and the concept of a protein space. *Nature*, 225(5232), pp.563–564.

Solomon, I., 1955. Relaxation Processes in a System of Two Spins. *Phys. Rev.*, 99(2), pp.559–565.

Solyom, Z. et al., 2013. BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR*, 55(4), pp.311–321.

Stehelin, D. et al., 1976. Purification of DNA complementary to nucleotide sequences required for neoplastic transformation of fibroblasts by avian sarcoma viruses. *J. Mol. Biol.*, 101(3), pp.349–365.

Stehelin, D. et al., 1976. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547), pp.170–173.

Stein, P.L., Vogel, H. & Soriano, P., 1994. Combined deficiencies of Src, Fyn, and Yes tyrosine kinases in mutant mice. *Genes Dev.*, 8(17), pp.1999–2007.

Sudol, M., 1998. From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene*, 17(11), pp.1469–1474.

Summy, J.M. et al., 2003. The SH4-Unique-SH3-SH2 domains dictate specificity in signaling that differentiate c-Yes from c-Src. *J. Cell Sci.*, 116(Pt 12), pp.2585–2598.

Süel, G.M. et al., 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10(1), pp.59–69.

Svergun, D., Barberato, C. & Koch, M.H.J., 1995. CRYSOL - A program to evaluate X-ray solution

scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, 28(6), pp.768–773.

Szoka, F. & Papahadjopoulos, D., 1980. Comparative Properties and Methods of Preparation of Lipid Vesicles (Liposomes). *Annu. Rev. Biophys. Bioeng.*, 9(1), pp.467–508.

Takegoshi, K., Nakamura, S. & Terao, T., 2001. 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.*, 344(5-6), pp.631–637.

Tanaka, A. & Fujita, D.J., 1986. Expression of a molecularly cloned human c-src oncogene by using a replication-competent retroviral vector. *Mol. Cell. Biol.*, 6(11), pp.3900–3909.

Tanford, C., Kawahara, K. & Lapanje, S., 1966. Proteins in 6-M guanidine hydrochloride. Demonstration of random coil behavior. *J. Biol. Chem.*, 241(8), pp.1921–1923.

Teilum, K., Olsen, J.G. & Kragelund, B.B., 2015. Globular and disordered—the non-identical twins in protein-protein interactions. *Front. Mol. Biosci.*, 2(June).

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45(D1), pp.D158–D169.

Theillet, F.-X. et al., 2014. The alphabet of intrinsic disorder: Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disord. Proteins*, 1(1), p.e24360.

Theillet, F.-X.F. et al., 2014. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.*, 114(13), pp.6661–6714.

Thomas, S.M. & Brugge, J.S., 1997. Cellular functions regulated by Src Family Kinases. *Annu. Rev. Cell Dev. Biol.*, 13(1), pp.513–609.

Tompa, P., 2014. Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.*, 114(13), pp.6715–6732.

Tompa, P., 2009. *Structure and function of intrinsically disordered proteins*, Chapman & Hall/CRC Press.

Tompa, P. & Fuxreiter, M., 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.*, 33(1), pp.2–8.

Tompa, P. et al., 2014. A Million Peptide Motifs for the Molecular Biologist. *Mol. Cell*, 55(2), pp.161–169.

Tompa, P. et al., 2015. Intrinsically disordered proteins: Emerging interaction specialists. *Curr. Opin. Struct. Biol.*, 35, pp.49–59.

Tong, M. et al., 2017. Survey of solution dynamics in Src kinase reveals allosteric cross talk between the ligand binding and regulatory sites. *Nat. Commun.*, 8(1), p.2160.

Tsai, C.-J., Ma, B. & Nussinov, R., 1999. Folding and binding cascades: Shifts in energy landscapes. *Proc. Natl. Acad. Sci.*, 96(18), pp.9970–9972.

Uversky, V.N., 2003. A Protein-Chameleon: Conformational Plasticity of α-Synuclein, a Disordered

Protein Involved in Neurodegenerative Disorders. *J. Biomol. Struct. Dyn.*, 21(2), pp.211–234.

Uversky, V.N., 2016. Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. *J. Biol. Chem.*, 291(13), pp.6681–6688.

Uversky, V.N., 2009. Intrinsic Disorder in Proteins Associated with Neurodegenerative Diseases. In *Protein fold. misfolding neurodegener. dis.* Dordrecht: Springer Netherlands, pp. 21–75.

Uversky, V.N., 2014. Introduction to Intrinsically Disordered Proteins (IDPs). *Chem. Rev.*, 114(13), pp.6557–6560.

Uversky, V.N., 2011. Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chem. Soc. Rev.*, 40(3), pp.1623–1634.

Uversky, V.N., 2002. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci.*, 11(4), pp.739–756.

Uversky, V.N. & Longhi, S. eds., 2010. *Instrumental Analysis of Intrinsically Disordered Proteins*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Uversky, V.N. et al., 2014. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.*, 114(13), pp.6844–6879.

Uversky, V.N., Oldfield, C.J. & Dunker, a K., 2008. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annu. Rev. Biophys.*, 37(1), pp.215–246.

Varadi, M. et al., 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, 42(D1), pp.D326–D335.

Via, A. et al., 2015. How pathogens use linear motifs to perturb host cell networks. *Trends Biochem. Sci.*, 40(1), pp.36–48.

Voelz, V.A. & Zhou, G., 2014. Bayesian inference of conformational state populations from computational models and sparse experimental observables. *J. Comput. Chem.*, 35(30), pp.2215–2224.

Vucetic, S. et al., 2003. Flavors of protein disorder. *Proteins Struct. Funct. Genet.*, 52(4), pp.573–584.

Wang, L.H. et al., 1976. Location of envelope-specific and sarcoma-specific oligonucleotides on RNA of Schmidt-Ruppin Rous sarcoma virus. *Proc. Natl. Acad. Sci. U. S. A.*, 73(2), pp.447–451.

Ward, J.J. et al., 2004. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.*, 337(3), pp.635–645.

Wei, G. et al., 2016. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.*, 116(11), pp.6516–6551.

Weng, Z. et al., 1995. Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions. *Mol. Cell. Biol.*, 15(10), pp.5627–5634.

Wereszczynski, J. & McCammon, J.A., 2012. Statistical mechanics and molecular dynamics in evaluating

thermodynamic properties of biomolecular recognition. *Q. Rev. Biophys.*, 45(01), pp.1–25.

Whisstock, J.C. & Lesk, A.M., 2003. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.*, 36(3), pp.307–340.

White, A.D., Dama, J.F. & Voth, G.A., 2015. Designing Free Energy Surfaces That Match Experimental Data with Metadynamics. *J. Chem. Theory Comput.*, 11(6), pp.2451–2460.

Williams, R.M. et al., 2001. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.*, pp.89–100.

Williamson, M.P., 2013. Using chemical shift perturbation to characterise ligand binding. *Prog. Nucl. Magn. Reson. Spectrosc.*, 73, pp.1–16.

Wittenberg, J.B. & Isaacs, L., 2012. Complementarity and Preorganization. In *Supramol. chem.* Chichester, UK: John Wiley & Sons, Ltd.

Wright, P.E. & Dyson, H.J., 2014. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, 16(1), pp.18–29.

Wright, P.E. & Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293(2), pp.321–331.

Wu, H. & Fuxreiter, M., 2016. The Structure and Dynamics of Higher-Order Assemblies: Amyloids, Signalosomes, and Granules. *Cell*, 165(5), pp.1055–1066.

Wu, P., Nielsen, T.E. & Clausen, M.H., 2016. Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug Discov. Today*, 21(1), pp.5–10.

Wüthrich, K., 1995. *NMR in structural biology : a collection of papers by Kurt Wüthrich*, Singapore River Edge, NJ: World Scientific.

Wüthrich, K. & Wagner, G., 1978. Internal motion in globular proteins. *Trends Biochem. Sci.*, 3(4), pp.227–230.

Xiao, R. et al., 2013. Structural framework of c-Src activation by integrin 3. *Blood*, 121(4), pp.700–706.

Xiao, X., Kallenbach, N. & Zhang, Y., 2014. Peptide Conformation Analysis Using an Integrated Bayesian Approach. *J. Chem. Theory Comput.*, 10(9), pp.4152–4159.

Xu, W. et al., 1999. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol. Cell*, 3(5), pp.629–638.

Xu, W., Harrison, S.C. & Eck, M.J., 1997. Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, 385(6617), pp.595–602.

Yeatman, T.J. et al., 1999. Activating SRC mutation in a subset of advanced human colon cancers. *Nat. Genet.*, 21(2), pp.187–190.

Yu, H. et al., 1992. Solution structure of the SH3 domain of Src and identification of its ligand-binding

site. *Science*, 258(5088), pp.1665–1668.

Zadeh, L., 1965. Fuzzy sets. *Inf. Control*, 8(3), pp.338–353.

Zafra Ruano, A. et al., 2016. From Binding-Induced Dynamic Effects in SH3 Structures to Evolutionary Conserved Sectors P. M. Kim, ed. *PLOS Comput. Biol.*, 12(5), p.e1004938.

Zhang, J., Yang, P.L. & Gray, N.S., 2009. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer*, 9(1), pp.28–39.

Zhang, R., Mroue, K.H. & Ramamoorthy, A., 2017. Proton-Based Ultrafast Magic Angle Spinning Solid-State NMR Spectroscopy. *Acc. Chem. Res.*, 50(4), pp.1105–1113.

Zhang, X.H.-F. et al., 2009. Latent Bone Metastasis in Breast Cancer Tied to Src-Dependent Survival Signals. *Cancer Cell*, 16(1), pp.67–78.

Zhong, L. et al., 2007. Solid-state NMR spectroscopy of 18.5 kDa myelin basic protein reconstituted with lipid vesicles: Spectroscopic characterisation and spectral assignments of solvent-exposed protein fragments. *Biochim. Biophys. Acta - Biomembr.*, 1768(12), pp.3193–3205.

Zhou, H.-X., 2012. Intrinsic disorder: signaling via highly specific but short-lived association. *Trends Biochem. Sci.*, 37(2), pp.43–48.

Zhou, H.-X., Pang, X. & Lu, C., 2012. Rate constants and mechanisms of intrinsically disordered proteins binding to structured targets. *Phys. Chem. Chem. Phys.*, 14(30), p.10466.