STRUCTURAL INVESTIGATION OF THE STAR PROTEIN SAM68

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

Jaelle Nicola Foot MSc (University of Leicester)
BSc (University of Southampton)
Department of Biochemistry
University of Leicester

April 2015

Jaelle N. Foot

Structural investigation of the STAR protein Sam68

Abstract

Sam68 is a member of the STAR family of proteins, linking post-transcriptional gene regulation with signal transduction pathways. Sam68 has been shown to specifically regulate the alternative splicing of many genes through interactions with the pre-mRNA and spliceosomal machinery. Selection of particular isoforms of several of these genes has been shown to contribute to neoplastic transformation and aberrant Sam68 expression and function has been implicated in the development of various genetic diseases and cancers. It is therefore important to understand Sam68 RNA recognition at the molecular level in order to design next generation drug therapies.

This thesis describes the structural and biophysical techniques used to define the bipartite RNA consensus sequence specifically recognised by Sam68 and the mechanisms of interaction. This data provides a model of Sam68 contribution to alternative splicing regulation. Splice site selection is also influenced by post-translational modifications of Sam68 including serine and threonine phosphorylation. In several cases, the phosphorylation state of Sam68 directly influences the outcome of splicing and leads to cancer development. Identification of phosphorylation sites in the STAR domain of Sam68 by NMR and radiolabelled kinase assays reveals how this post-translational modification may affect RNA binding at the molecular level.

Acknowledgements

First and foremost I would like to express my deepest appreciation for my supervisor Dr Cyril Dominguez. I have thoroughly enjoyed my PhD experience and that is entirely down to him for creating a wonderful working environment, full of support, helpful discussion and passion for science. I could not have imagined having a better advisor for my PhD. I must also extend my gratitude towards the other members of the Dominguez group; particularly Mikael and Carika, and more recently Ayesha, Oksana, Sarah, Adam and Eby for the friendly advice and discussion.

I am indebted to Dr Raj Patel and Professor John Schwabe for giving me such an excellent start to my scientific career during my masters, and for the discussions and advice given by my committee members Professor Ian Eperon and Professor Richard Bayliss.

I would especially like to thank Dr Peter Watson, for his time and patience throughout my masters project and beyond. His continuous support and advice, both in science and life have been invaluable over these last few years. I have been very lucky to have had you as a mentor and friend, and wish you and your new family all the very best for the future.

I must also extend my gratitude to the other members the Schwabe group, particularly Dr Louise Fairall for her proofreading of this thesis and Dr Gregg Hudson for helpful suggestions and advice.

I was able to visit the lab of Professor Michael Sattler at Technische Universität München during my PhD and would like to thank him for the opportunity. It was a great experience and his lab members, particularly Dr Hyun-So Kang and Dr Ralf Stehle, were very welcoming and helpful.

Finally, I wish to thank my family and friends for their making this a very happy few years for me. Thank you Charlie for the laughs and japes. My sister Chloe for being a princess and putting a smile on my face with pictures of cakes, cats and crossfit during thesis writing. I'm so pleased for you that you've joined us science nerds and wish you all the best in your studies. Our four-legged family members; Indy, Molly and Tiggs for the cuddles. And my wonderful parents, Paul and Diane, without their constant support and encouragement none of this would have been possible. I will be eternally grateful and will always read the question, check the answer.

Dedicated to

Margaret Brown

Forever Near

List of Contents

Table of Contents

List of Figures

List of Tables

Abbreviations

APC	Adenomatous polyposis coli
AS4	Alternatively spliced segment 4
ASF/SF2	alternative splicing facotr 1/pre-mRNA-splicing factor 2
ATP	Adenosine triphosphate
BLI	Bio-layer Interferometry
BRK	Breast cancer kinase
BSA	Bovine Serum Albumin
CBP	CREB-binding protein
Cdc2	Cell division cycle protein 2
CDK	Cyclin dependent kinase
cDNA-RDA	cDNA representational difference analysis
CLIP	Cross-linking Immunoprecipitation
CLK	CDK-like kinases
Co-IP	Co-immunoprecipitation
CRM1	Chromosome region maintenance 1
CTE	Constitutive transport element
DNA	Deoxyribonucleic Acid
DTT	Dithiothreitol
EGF	Epidermal growth factor
EMT	Epithelial to mesenchymal transition
ERK1	Extracellular signal-regulated kinases
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
FBP2	FUSE binding protein 2
FP	Fluorescence Polarisation
FSH	Follicle stimulating hormone
FXTAS	Fragile X-associated tremor/ataxia syndrome
GSK	Glycogen synthase kinases
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HEK	Human embryonic kidney
HGF	Hepatocyte growth factor
HIV	Human immunodeficiency virus
hnRNP	Heteronuclear ribonucleoprotein
HSQC	Heteronuclear Single Quantum Coherence
IPTG	Isopropyl-β-D-Thiogalactopyranoside
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
ITC	Isothermal Titration Calorimetry
KHDRBS	KH Domain containing RNA Binding Signal transduction associated 1
KSRP	K-homology splicing regulator protein

LH	Luteinizing hormone
MAPK	Mitogen activated pathway kinase
MEF	Myeloid Elf-1-like factor
MLL-EEN	Mixed lineage leukaemia – extra eleven nineteen
MPD	2-Methyl-2, 4-pentanediol
mTOR	Mammalian target of rapamycin
NLS	Nuclear Localisation Signal
NMR	Nuclear Magnetic Resonance
Nrxn	Neurexin
OD	Optical Density
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDB	Protein database
PEG	Polyethylene glycol
PIAS1	Protein inhibitor of activated STAT-1
PRMT1	Protein arginine methyltransferase 1
PTM	Post-translational modification
QKI	Quaking
Ras-GTP	Rat sarcoma-Guanosine-5'-triphosphate
RBD	RNA binding domain
RBP	RNA binding protein
RHKO	Random homozygous knockout
RNA	Ribonucleic Acid
RRE	Rev response element
RRM	RNA recognition motif
RS domain	Arginine-Serine rich domain
RT-PCR	Reverse transcription PCR
Sam68	Src-associated in mitosis (68kDa)
SAXS	Small-angle X-ray scattering
SEC-MALS	Size exclusion chromatography-multi-angle light scattering
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SGCE	epsilon sarcoglycan
SH3	Src homology domain 3
SIA	Scaffold independent analysis
SNB	Sam68 nuclear bodies
snRNP	Small nuclear ribonucleoproteins
SR	Serine-arginine
SRPK	SR protein kinase
STAR	Signal transduction and activation of RNA
SUMO	Small ubiquitin-like modifier
TEV	Tobacco Etch Virus
TGCT	Testicular germ cell tumours
WT	Wild-type
XPO1	Exportin 1

Chapter 1. Introduction

1.1. The human genome

Discovery of the structure of DNA by James Watson and Francis Crick in 1953 brought to light the genetic code and thus the transfer of information from DNA to RNA through transcription, and subsequently to proteins by translation (Watson, Crick 1953). Published in full some 40 years later, the human genome project revealed the presence of 20,500 genes within our three billion base pair sequence (International Human Genome Sequencing Consortium 2001). This figure was relatively small compared to initial estimates that were based on the estimated size of the human proteome, approximately 100,000 proteins.

This huge diversity in protein production from a limited number of genes is due in a large part due to alternative splicing, allowing more than one protein isoform to be produced from a single gene. It was in early 1977, 22 years after the original identification of RNA, that splicing was first observed using electron microscopy (Berget, Moore & Sharp 1977). Originally thought to be rare, it has now been shown to occur in plants, animals and fungi (Wang, Brendel 2006, Kim, Magen & Ast 2007, Okazaki, Niwa 2000), and to be tissue specific and essential for correct cellular differentiation and development (as reviewed in (Javier Lopez 1998)). The discovery of the ribonucleoproteins that orchestrate splicing and the depth of RNA chemistry involved in pre-mRNA processing took another few decades (Hinterberger, Pettersson & Steitz 1983). Our current knowledge of the process and regulation of splicing still falls short of completing the entire picture of this complex process and its link to disease.

It is however, now clear that RNA is not just a passive molecule, transferring genetic information from the nucleus to cytoplasm for protein translation. The ENCODE project, for example, has identified that between 20% and 80% of the human genome is transcribed, although only 1.5% encodes protein, suggesting that by no means all of the RNA transcribed is set on a path of protein production (http://encodeproject.org). Accordingly, in order to understand the role of different RNAs in the cell, and the processes by which RNA is matured towards translation, the hunt for RNA binding proteins (RBPs) is on. In the last decade over 800 human RBPs have been predicted and identified predicted, harbouring recognised RNA binding domains (RBDs) such as

RNA recognition motifs (RRMs) and hnRNP K homology (KH) domains, and some with novel RNA binding ability (Castello et al. 2012, Baltz et al. 2012).

It is essential to understand the relationship between the RNA and RBPs that regulate alternative splicing, since mutations in both are well established to initiate the onset of various genetic diseases (Singh, Cooper 2012, Lukong et al. 2008).

1.2. Mechanism of splicing

Eukaryotic genes are characterised by long, non-coding introns and rather short protein-coding exons. During and following transcription, and production of these long pre-mRNAs, a series of processing steps are carried out before export of the mature mRNA to the cytoplasm for translation by the ribosomal machinery. One key stage of maturation requires the removal of intronic sequences and ligation of the remaining exons in a process termed splicing. This is operated by the spliceosome, a large and dynamic multiprotein and RNA complex composed of five main small nuclear ribonucleoproteins; snRNPs U1, U2, U4, U5 and U6, along with more than 100 additional proteins (Figure 1.1). This macromolecular machinery assembles in a series of splicing factor dependent arrangements, in a tightly regulated fashion.

The spliceosome binds to specific, conserved regions at the intron/exon junction called splice sites. The 5’ splice site is located at the 5’ end of the intron and is characterised by a GU dinucleotide located within a larger, less conserved sequence. The 3’ splice site is located at the 3’ end of the intron and is composed of an AG dinucleotide, preceded by the branchpoint and a polypyrimidine tract. Binding of particular elements of the splicing machinery to each of these sites drives the splicing process, which occurs in two transesterification steps. The first step is cleavage of the intron from the 5’ exon, and the ligation of the 5’ end of the intron to the branch point, leaving the 5’ exon detached and an intron-3’ exon fragments in the form of a lariat. In the second step, the 5’ end of the detached exon is ligated to the 3’ end of the exon attached to the lariat, which releases the intron, still in the form of a lariat (Black 2003).

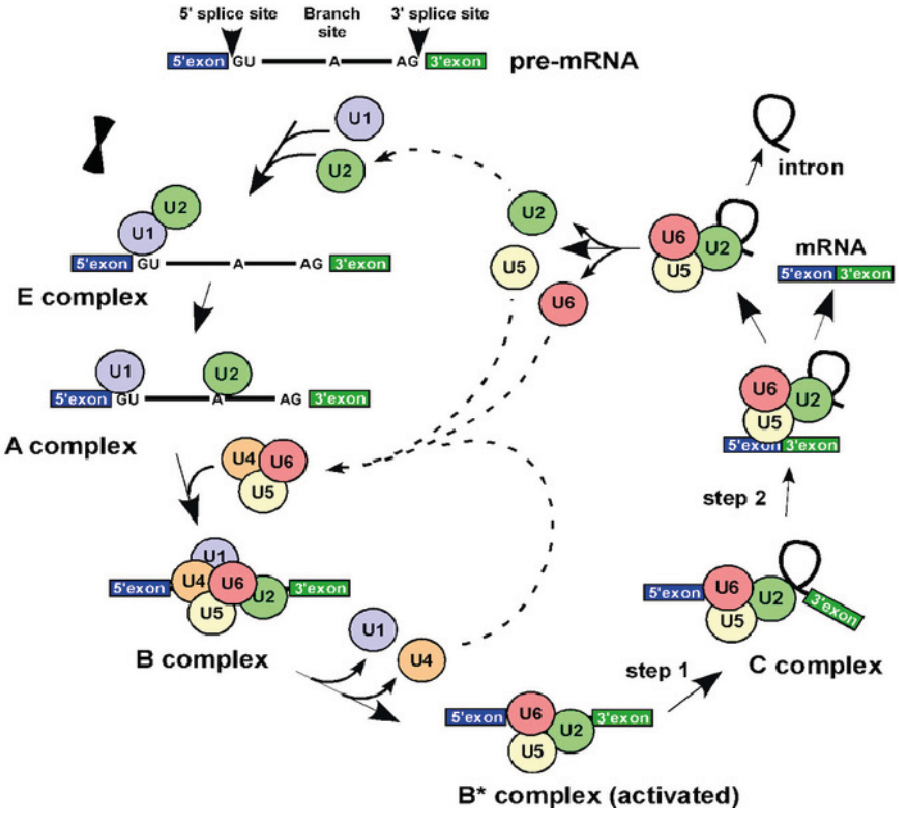


Figure 1.1: **Mechanism of Splicing**
Splice regulatory elements within the pre-mRNA are recognised by specific members of the spliceosomal machinery. Sequential recruitment of different snRNPs to each splicing complex in turn catalyses the two transesterification reactions to remove the intronic RNA. This process is tightly regulated by many other ribonucleoproteins and splicing factors such that splicing occurs in a tissue and cell cycle specific manner.
(Taken from https://www.mpibpc.mpg.de/289605/07-spliceosome)

Each stage of the splicing process is tightly regulated by snRNPs and many other ribonucleoproteins and splicing factors. Figure 1.1 gives a brief overview of the role of snRNPs in splicing. During the early formation of the spliceosome, U1 binds directly to the 5' splice site and then recruits U2 to form the E complex. The snRNA of U2 recognises the intronic branchpoint and base-pairs to form the A complex. The B complex is then formed by recruitment of snRNPs U4, U5 and U6, which then becomes the catalytically active C complex via rearrangement of the snRNPs and replacement of U1 by U6 at the 5' splice site and loss of U4 from the complex. At this stage, both transesterification steps can be conducted, the exons ligated together and the intronic lariat released with the three remaining snRNPs, which are then recycled (Black 2003).

1.3. Mechanism and regulation of alternative splicing

Coordination of splicing is important as in most cases there are several different arrangements by which the exonic sequences can be ligated together (Figure 1.2). The

majority of exons are constitutive and are always retained in the mature mRNA. Others, known as cassette exons, can either be included or excluded depending on the particular protein isoform required. In some cases, two or more of these cassette exons are mutually exclusive and only one can be retained following alternative splicing. Through use of alternative 5' and 3' splice sites, exons can be lengthened or shortened, and in some cases introns can be retained.

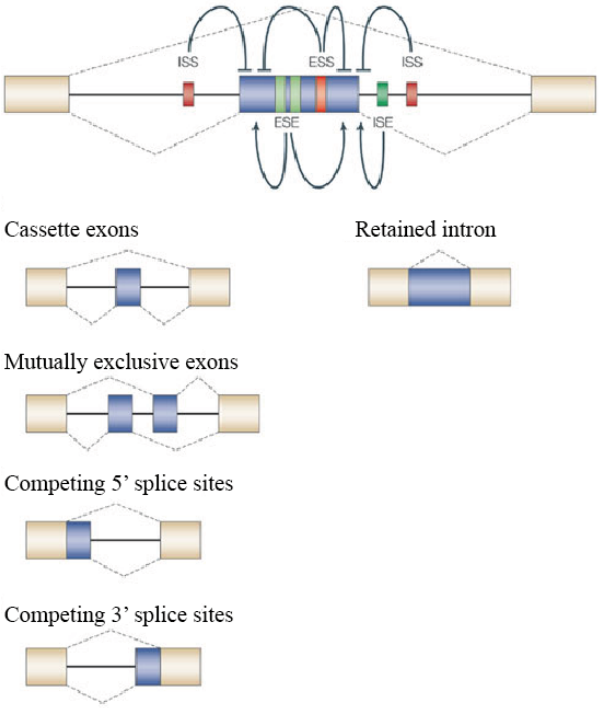


Figure 1.2: **Mechanisms of alternative splicing**
Splicing factors specifically recognise conserved sequences within the pre-mRNA to regulate differential inclusion of intronic and exonic RNA (reproduced from (Matlin, Clark & Smith 2005)).

Most eukaryotic pre-mRNA transcripts encode for more than one mature mRNA and as many as 95% of human genes undergo alternative splicing, resulting in increased proteome diversity from a limited number of genes (Black 2003). Differential assortment of exons often results in protein products with differing function in terms of protein-protein interactions, protein-ligand interactions, catalytic activity and localisation within the cell (Tazi, Bakkour & Stamm 2009). Therefore, alternative splicing must be tightly regulated in order to ensure correct expression of each protein isoform as it is required in the cell, indeed specific isoforms need to be produced in a cell-type and cell cycle specific manner. Aberrant splicing has been implicated in up to 50% of all human genetic diseases, as reviewed in (Tazi, Bakkour & Stamm 2009). In addition, many different cancers, several neurological, metabolomic and aging disorders

have been attributed to defects in splicing. Splicing errors have been shown to arise due to single and multiple point mutations in exonic and intronic regulatory elements and splice sites, loss of splicing factor function and alterations in the ratios of each protein isoform that is expressed.

Splice site selection is an important decision that is made early in spliceosomal assembly, involving RNA binding proteins that recognise distinct sequences in the pre-mRNA. There are several families of RBPs that are known to have an impact on this process through interactions with other members of the spliceosomal machinery and through binding to the pre-mRNA in a sequence specific manner.

There are four types of RNA sequences recognised by RBPs. Exonic splicing enhancers (ESEs) that often recruit serine-arginine (SR) proteins (section 1.3.2.2), exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs) that often recruit heteronuclear ribonucleoproteins (hnRNPs) (section 1.3.2.1) and intronic splicing enhancers (ISEs) (reviewed in (Black 2003)).

1.3.2. The role and function of RBPs in alternative splicing

It is a balancing act between positive and negatively acting regulatory proteins to determine the selection of a splice site and ultimately the protein isoform that is expressed. Furthermore, binding of RBPs to their ssRNA target sequences is modulated by other protein-RNA, protein-protein and RNA-RNA interactions as well as post-translational modifications of splicing factors. These allow alternative splicing to be regulated in a cell cycle and tissue specific manner.

Often, both SR and hnRNP proteins are present in spliceosomal subunits, interacting with overlapping ESEs/ESSs and exerting counteracting effects on splicing. Ultimately, splice site selection and binding affinity of the spliceosome is determined by the overall contribution of both hnRNP and SR proteins and characterisation of these protein families has been undertaken in order to understand regulation of alternative splicing and how changes in this process affect disease development (reviewed in (Smith, Valcárcel 2000)).

1.3.2.1. hnRNPs

The most studied and largest RBP family are the heterogeneous nuclear ribonucleoproteins (as reviewed in (Jean-Philippe, Paz & Caputi 2013)). Ranging in size from 30kDa to 120kDa, this family is characterised by the presence of one or more

RNA binding domains and various auxiliary domains (Dreyfuss et al. 1993). RNA recognition motifs (RRMs) are the most common RBD found in hnRNPs but also in other RBP families. RRMs are composed of a 90 amino acid sequence that is able to recognise specific nucleic acid sequences (as reviewed in (Maris, Dominguez & Allain 2005)). In addition, some hnRNPs contain alternative RBDs such as RGG boxes and KH domains, as well as other auxiliary domains which function in protein-protein interactions and affect functional specificity and subcellular localisation, such as glycine-rich and proline-rich motifs (Burd, Dreyfuss 1994).

hnRNPs are ubiquitously expressed and are one of the most abundant nuclear proteins, with total levels varying between cell type (Kamma, Portman & Dreyfuss 1995). In general, this family of proteins is found to be diffuse within the nucleus, but is also able to remain bound to target mRNA and translocate to the cytoplasm via nuclear pores to interact with translational machinery (Pinol-Roma, Dreyfuss 1992). This nuclear-cytoplasmic shuttling, as well as functional regulation of hnRNPs is in part mediated by their post-translational modifications such as phosphorylation (Blanchard, Brunel & Jeanteur 1977), ubiquitination (Laroia, Schneider 2002), SUMOylation (Li et al. 2004) and methylation (Rajpurohit, Paik & Kim 1992). Furthermore, hnRNP-RNA complexes are highly dynamic and constantly subject to addition and loss of hnRNPs and other RBPs on a pre-mRNA, whose activity levels are also variable. Therefore, hnRNPs exhibit multiple functions in pre-mRNA splicing, translation, microRNA processing, mRNA trafficking and telomere maintenance.

1.3.2.2. SR Proteins

Serine-Arginine rich (SR) proteins are a group of splicing factors that are involved in the operation and modulation of alternative splicing through facilitating spliceosome assembly and splice site selection (reviewed in (Shin, Manley 2004, Änkö 2014)). This family is generally characterised by an N-terminal RNA binding domain formed of one or two RNA recognition motifs (Krainer, Conway & Kozak 1990) and a C-terminal arginine-serine rich domain (RS domain) of varying length enriched in arginine-serine repeats that function in protein-protein interactions. SR proteins have been shown to exert their effect on alternative splicing through binding to the pre-mRNA and recruiting RS-domain containing proteins to initiate spliceosome formation (Wu, Maniatis 1993). This is achieved through specific interaction between RRM domains

and pre-mRNA in addition to RS domain contacts with the 5' splice site and branchpoint.

SR protein function is regulated through post-translational modifications such as arginine methylation, lysine acetylation and serine phosphorylation (as reviewed in (Zhou, Fu 2013). Recently, it was shown that SRSF1 contains three sites of Arginine methylation between its two RRM domains, promoting its translocation to the nucleus (Sinha et al. 2010). Additionally, lysine acetylation of SR proteins through specific histone acetyl transferases (HATs) and lysine deacetylation by specific histone deacetylases (HDACs) may provide a complex mechanism of splice site selection control (Edmond et al. 2011). Finally, and most extensively studied, SR proteins are subject to phosphorylation on their serine residues by the CMGC family of kinases. This group of specific proteins is comprised of cyclin-dependent kinases (CDKs), mitogen-activated protein kinases (MAPKs), glycogen synthase kinases (GSKs) and CDK-like kinases (CLKs) (Kannan, Neuwald 2004). SR proteins have been shown *in vivo* to be phosphorylated by CDKs and SRPKs (Gui, Lane & Fu 1994, Colwill et al. 1996). It is likely that these kinases work in a "relay" fashion, with SRPKs predominantly located in the cytoplasm as well as the nuclear compartment, and CLKs colocalised with SR proteins in nuclear speckles, dynamic nuclear structures that contain an abundance of splicing factors (reviewed in (Lamond, Spector 2003)). Overexpression of both families of kinases has been shown to alter the cellular localization of SR proteins, in particular from nuclear speckles to the nucleoplasm (Colwill et al. 1996). Interestingly, both function as "polymerizing kinases", whereby after each kinase reaction, the kinase remains bound and transfers multiple phosphates to adjacent SR dipeptides (Ghosh, Adams 2011).

1.3.3. Regulation of alternative splicing through signal transduction

The function of both hnRNPs and SR proteins is regulated by post-translational modifications, some of which are dictated through signalling pathways. An example of the regulation of splice site selection through cell signalling and its effect on cell fate is in the alternative splicing of the *Bcl-x* transcript. Choice of alternative 5' splice site in exon 2 results in the translation of a short protein isoform, Bcl-x_S, or a long isoform, Bcl-x_L (Figure 1.3). Interestingly, these protein isoforms have antagonistic functions, the shorter is pro-apoptotic and induces cell death, and the longer is anti-apoptotic and inhibits cell death pathways (Boise et al. 1993). Therefore, expression of a particular

isoform affects the susceptibility of a cell to apoptosis, and cancer cells have been shown to upregulate Bcl-x_L which promotes their survival. It is likely that this shift in the ratio of isoforms in cancer cells is driven by alterations in splice site selection. Figure 1.3 shows the *Bcl-x* transcript and the balance between more than ten different splicing factors that dictate splice site selection. Furthermore, the function of these splicing factors is regulated by many different kinases and signalling pathways. A genome-wide siRNA experiment identified approximately 160 proteins that influence the outcome of *Bcl-x* splicing. Many of these were RBPs but also many kinases and signalling molecules (Moore et al. 2010). This emphasises the complexity of alternative splicing events and highlights the importance of understanding these systems in order to elucidate their full potential as an avenue for therapeutics.

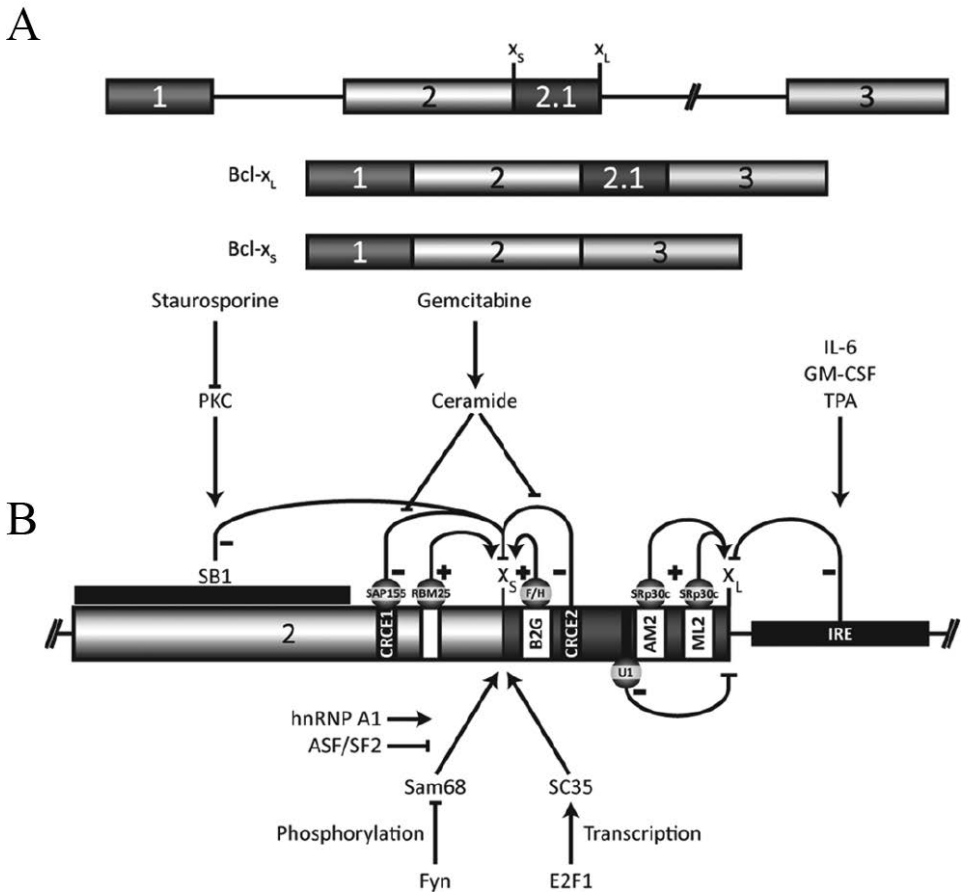


Figure 1.3: **Alternative splicing of Bcl-x**
Two alternative splice sites within exon 2 of the *Bcl-x* gene result in two potential protein isoforms, Bcl-x_S and Bcl-x_L (A). The selection of each splice site is positively and negatively regulated by a large group of different RBPs including Sam68 and by signalling pathways (B) (reproduced from (Revil et al. 2009)).

1.4. STAR proteins

A family of RNA binding proteins that is implicated in the alternative splicing of Bcl-x is the STAR (Signal Transduction and Activation of RNA) family of proteins. Although these proteins are not functionally related, they are linked by an evolutionarily conserved domain of approximately 200 amino acids that harbours RNA binding properties. Since the identification of its founding member, Sam68 (Src-associated in mitosis 68kDa) in 1994, subsequent STAR family members were discovered in plants, yeast, nematodes, flies and mammals (Courtneidge, Fumagalli 1994) (reviewed in (Vernet, Artzt 1997).

Sam68, more formally known as KHDRBS1 (KH Domain containing RNA Binding Signal transduction associated 1), was initially identified as a 62kDa protein binding to Ras-GTP and was later found to be a downstream target of Src and Fyn (Taylor, Shalloway 1994). It is the best-characterised member of the family and two paralogs were subsequently discovered, resulting from a gene triplication event 520-610 million years ago (Ehrmann et al. 2013), forming one of three subfamilies. These are referred to as SLM1 (KHDRBS2) and SLM2/T-STAR (KHDRBS3) (Sam68-Like Mammalian proteins) (Venables et al. 1999) (Di Fruscio, Chen & Richard 1999). A second subfamily that is not functionally related but also contains a central STAR domain is Quaking (QKI) in mouse and human, Gld-1 in *C.elegans* and HOW in *Drosophila*. The final subgroup consists of SF1 in mammals and more distant relatives such as SPIN1 in rice (Figure 1.4).

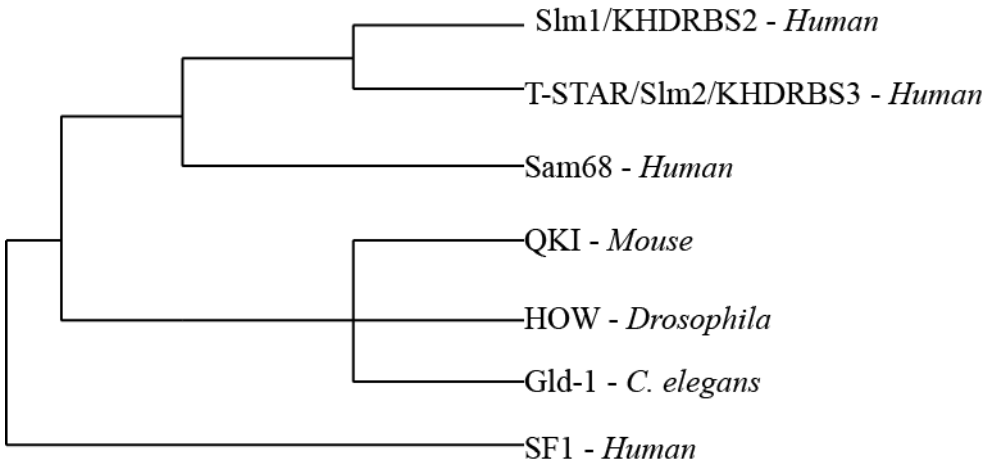


Figure 1.4: **Family of STAR proteins**
Simplified family tree of the STAR family, showing the two evolutionarily distinct subfamilies of Sam68, SLM1 and SLM2 and QKI, Gld-1 and HOW and the species from which each protein was first identified (reproduced from (Volk, Artzt 2010)).

As their name suggests, STAR proteins are involved in signal transduction cascades and are regulated with respect to RNA metabolism by post-translational modifications, placing this family as a potential key link between signalling and splicing. In addition to alternative splicing, these proteins have been shown to affect other aspects of RNA metabolism, such as transcription, mRNA export, mRNA stability and translation (reviewed in (Volk, Artzt 2010)). This thesis will focus on the function of Sam68.

1.4.2. Role of Sam68 in RNA metabolism

1.4.2.1. Alternative Splicing

In 1999, five years after its discovery, Sam68 was the first STAR protein to be shown to be involved in alternative splicing. Hartmann *et al.* demonstrated using a yeast-two-hybrid screen that Sam68 interacts with YT521B, a tyrosine phosphorylated protein located in nuclear dots (Hartmann et al. 1999). Sam68 was shown to affect alternative splicing in a concentration dependent manner using *in vivo* splicing assays. Hence this protein is a potential mediator of signal transduction events and splice site selection. In addition to associating with protein members of the spliceosomal machinery, Sam68 was subsequently shown to directly bind to pre-mRNA. Figure 1.5 lists many of the splicing targets of Sam68 that have been identified. Interestingly these gene targets encode proteins that are involved in a wide variety of cellular processes. For example, *mTOR* splicing is implicated in adipogenesis (Huot et al. 2012), *Neurexin* (*Nrxn*) in synapse function and differentiation (Iijima et al. 2011), *Cyclin D1* in cell cycle progression (Paronetto, Cappellari & Busa 2010), *Bcl-x* in apoptosis (Paronetto et al. 2007), *Sgce* in spermatogenesis (Paronetto et al. 2011) and neurogenesis (Chawla et al. 2009) and *CD44* in cell migration (Locatelli, Lange 2011). The outcome of alternative splicing of several of these genes impacts the development of many diseases such as splicing of the androgen receptor (*AR*) in prostate cancer (Rajan et al. 2008), *ASF*/*SF2* in epithelial to mesenchymal transition (EMT) (Valacca et al. 2010) and *SMN2* in spinal muscular atrophy (Pedrotti et al. 2010). Furthermore, in the case of *CD44*, *Bcl*-x and *ASF*/*SF2* splicing, the outcome is affected by the phosphorylation state of Sam68. Therefore given the impact on disease progression, the contribution of Sam68 to splicing is diverse and important to understand.

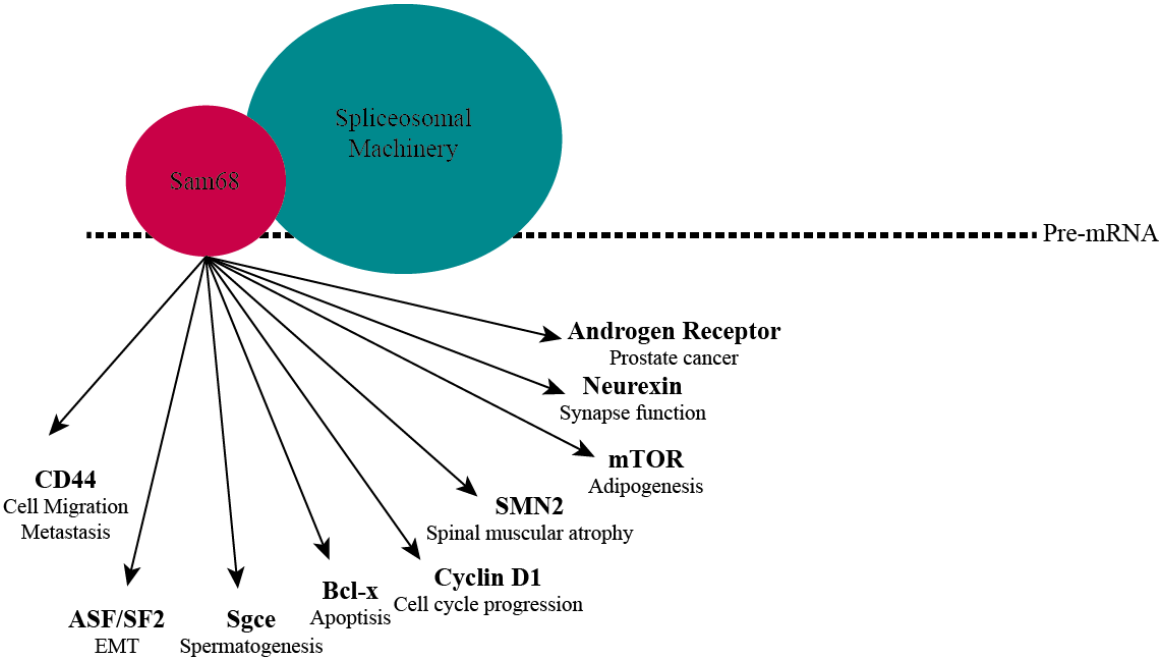


Figure 1.5: **Alternative splicing function of Sam68**
San68 interacts with members of the spliceosomal machinery and binds directly to pre-mRNAs to regulate splice site selection. Gene targets include *CD44*, *ASF/SF2*, *Sgce*, *Bcl-x*, *Cyclin D1*, *SMN2*, *mTOR*, *Neurexin* and the androgen receptor, demonstrating that Sam68 contributes to the regulation of many different cellular processes.

1.4.2.2. Retroviral Transport

In addition to alternative splicing, Sam68 has also been shown to facilitate RNA export from the nucleus to the cytoplasm. In particular, RNAs containing a constitutive transport element (CTE) and unspliced retroviral RNAs containing a human immunodeficiency virus Rev response element (HIV RRE) (Reddy et al. 1999).

In the case of HIV RRE, Sam68 has been shown to enable export of viral RNAs through formation of a tripartite complex with Rev. However it is not yet fully understood how each of these components interact with each other and whether export is dependent on chromosome region maintenance 1 (CRM1)/exportin1 (XPO1) (Li et al. 2002) (Reddy et al. 1999). Reduction of Sam68 expression in astrocytes, expression of a cytoplasmic mutant of Sam68 and tyrosine phosphorylation of Sam68 have all been shown to affect the localisation of Rev, inhibit HIV RNA export and viral production (Li et al. 2002)(Reddy et al. 1999, Soros et al. 2001)(Najib et al. 2005b). These data place Sam68 as a key player in the function of Rev with respect to HIV RNA nuclear export and HIV1 production. In fact, Sam68 has since been shown to be absolutely required for HIV1 protein production (Modem et al. 2005), through enhancement of 3' end processing of HIV RNA (Mclaren, Asai & Cochrane 2004), export and translation (He, Henao-Mejia & Liu 2009).

Clearly Sam68 is essential for processing of retroviral RNAs, however more research is required to understand the full complexity of its involvement and elucidate its potential as a therapeutic target.

1.4.3. Role of Sam68 in signal transduction

Outside its central STAR RNA binding domain, Sam68 contains various regulatory motifs that are sites of post-translational modifications. These are regulated by different signal transduction pathways and affect Sam68 localisation and function. Through its direct involvement in RNA metabolism and signal transduction, Sam68 exerts effects in various different cellular processes, as outlined below.

1.4.3.1. Cell cycle

Given its initial identification as a substrate of Src during mitosis, it was immediately apparent that Sam68 may affect cell cycle progression. Confirmation of this was demonstrated by Pillay et al in 1996, through use of Radicol, a Src kinase inhibitor, on mouse fibroblasts. This inhibits tyrosine phosphorylation of free and Src-associated Sam68 and prevents cells from exiting mitosis. Since cells were able to enter mitosis, Sam68 appears to be involved only in the later stages of cell division (Pillay, Nakano & Sharma 1996). A year later Barlat *et al*. discovered a natural isoform of Sam68 that lacks the KH domain and is expressed in normal cells that are in growth arrest (Barlat et al. 1997). Cells transfected with this ΔKH isoform were unable to enter S-phase due to lack of expression of Cyclin D1, placing Sam68 RNA binding function as key in cell cycle progression. In fact, Sam68 regulates splice site selection of *Cyclin D1* pre-mRNA (Barlat et al. 1997) (Taylor, Resnick & Shalloway 2004), and it was later shown that SUMO modification of Sam68 modulates its repression of *Cyclin D1* transcription, leading to G1 arrest (Babic, Cherry & Fujita 2006). Furthermore, Sam68 appears to affect other stages of the cell cycle, as retarded growth was observed in Sam68 deficient DT40 cells due to elongation of the G2-M phase (Li et al. 2002). These data suggest that Sam68 is required for cell cycle progression and that its specific function is regulated by post-translational modifications (PTMs) such as tyrosine phosphorylation and SUMOylation.

The effect of Sam68 on cell cycle progression has been shown to be a potential target for cancer therapy. In breast cancer cell lines, stimulation by epidermal growth factor (EGF) induces tyrosine phosphorylation of Sam68 within its nuclear localisation signal

(NLS) by breast cancer kinase (BRK), inducing Sam68 localisation into Sam68 nuclear bodies (SNBs) (Lukong et al. 2005). Overexpression of Sam68 was shown to induce G1 cell cycle arrest and phosphorylation of Sam68 on tyrosine residues by BRK inhibits this anti-proliferative effect. This suggests that BRK-induced tumourigenesis may be attenuated by overexpression or inhibition of tyrosine phosphorylation of Sam68.

Sam68 is also upregulated in prostate cancer cells, and depletion of Sam68 by RNAi resulted in a reduction in the rate of proliferation of prostate cancer cell line LNCaP at G1 phase (Busa et al. 2007). Once again, this effect on the cell cycle appeared to be through Cyclin D1 expression.

Whether Sam68 acts as a tumour suppressor or oncogene through its effects on cell cycle progression remains to be clearly understood. Further studies, with particular focus on the involvement of PTMs on Sam68 function must be conducted.

1.4.3.2. Development

The involvement of Sam68 in crucial developmental processes was revealed by the generation of Sam68 knockout (-/-) mice. These animals survive to adulthood, with just 20-30% dying perinatally (Richard et al. 2005). This suggests that Sam68 is not essential for development, and this is likely due to the existence of two, highly related Sam68 subfamily members, SLM1 and T-STAR that are able to compensate for the loss of Sam68. Sam68$^{-/-}$ mice harbour several key phenotypes which implicate this RBP in fertility, bone development, adipogenesis, neurological development and tumourigenesis. Male Sam68 null mice are infertile, through severe alterations in spermatogenesis, as reviewed in (Ehrmann, Elliott 2010). Sam68 is highly expressed in the testes and this effect on germ cell development in the male is partly through aberrant translational controls (Paronetto et al. 2009).

Fertility is also compromised in female mice, with smaller litter sizes, longer breeding times required for successful pregnancy and a delay in sexual maturity. These observations resulted from reductions in the number of ovulated oocytes and changes to oestrus cycles, which are regulated in part by the follicle-stimulating hormone (FSH) and the luteinizing hormone (LH), both of which were downregulated in Sam68$^{-/-}$ female adults. Sam68 was shown to bind to the pre-mRNA for both of these hormones through crosslinking-immunoprecipitation experiments, suggesting that it is necessary for correct development and ovulation of ovarian follicles (Bianchi et al. 2010).

Another characteristic of these mice is resistance to age-related bone loss. It is well established that loss of bone density is inversely proportional to bone marrow fat, through preferential differentiation of mesenchymal cells into adipocytes rather than osteocytes. Sam68 is now thought to be a key player in the regulation of bone marrow mesenchymal differentiation, favouring adipocyte differentiation over osteoblast differentiation (Richard et al. 2005). Consistently, whilst Sam68$^{-/-}$ mice retain their bone mass over time compared to wild-type littermates, they possess fewer adipocytes within the bone marrow and exhibit a lean phenotype in general including resistance to obesity, insulin and glucose intolerance on a high fat diet (reviewed in (Huot, Richard 2012)). This was found to be a result of aberrant splicing of the *mTOR* gene, lowering the overall amount of stable mTOR being expressed in white adipose tissue (WAT), a change directly regulated by Sam68 (Huot et al. 2012).

Sam68 has also been implicated in neuronal development as knockout mice exhibit poor motor control and behavioural deficits (Lukong, Richard 2008). Specifically, Sam68 has been identified as a key regulator of signal-dependent splicing of *Neurexin1* (*Nrxn1*) pre-mRNAs in the mouse brain (Iijima et al. 2011). This family of synaptic cell surface receptors are required for correct assembly of pre-synaptic terminals and contribute to plasticity processes in the brain (Ushkaryov et al. 1992). The incredible diversity throughout this protein family gives neuronal cells the ability to fine tune their molecular repertoire in response to external stimuli, despite being encoded by only three distinct genes, *Neurexin 1-3*. Alternatively spliced segment 4 (AS4), which is evolutionarily conserved between all *neurexin* pre-mRNAs, has been particularly well studied and inclusion of exon 20 within this region has been shown to influence protein-protein interactions and guide formation of synapses (Boucard et al. 2005) (Ehrmann et al. 2013).

Interestingly, increased expression of Sam68 in neuronal progenitor cells and mouse germ cells has been shown to enhance the skipping of exon 8 of the *epsilon sarcoglycan* (*sgce*) mRNA (Chawla et al. 2009)(Paronetto et al. 2011). This protein is required for anchorage of the dystrophin complex to the plasma membrane of muscle cells which connects the cytoskeleton to the extracellular matrix (Ozawa et al. 2005). There are several protein isoforms of Sgce that lack exon 8 that are found in the brain and in the testes and mutations in this gene have been linked to psychiatric symptoms and male infertility through aberrant differentiation.

1.4.4. Sam68 in cancer

Through its involvement in various cellular processes, Sam68 has been shown to affect the development of conditions and diseases including; cancer, osteoporosis, spinal muscular atrophy, Fragile X-associated tremor/ataxia syndrome (FXTAS), obesity, fertility and cardiac hypertrophy. Cancer has been most widely studied of these diseases, with upregulation of Sam68 contributing to poor prognosis in; breast cancer (Lukong et al. 2005, Locatelli et al. 2011), prostate cancer (Rajan et al. 2008, Busa et al. 2007, Cappellari et al. 2014), colorectal cancer (Liao et al. 2013), glioblastoma (Modem et al. 2011), non-small cell lung cancer (Zhang et al. 2014), renal cancer (Zhang et al. 2009), T-lymphoblastic leukaemia (Lazer et al. 2007) and cervical cancer (Li et al. 2012) (Rajan et al. 2008, Busà, Geremia & Sette 2010).

It is not yet clearly understood how Sam68 function contributes to neoplastic transformation. Initially Sam68 was thought to be a tumour suppressor following a 25% reduction of Sam68 expression in NIH3T3 cells by random homozygous knock-out (RHKO) compared to WT cells (Jones, Schedl 1995). These transformed cells exhibited defective contact-independent inhibition and anchorage independent growth, as well as the ability to drive the growth of metastatic tumours in nude mice, independent of Src activity and in a partially reversible manner (Liu et al. 2000). However, no changes in cell growth were observed and this was not the case when Sam68 haploinsufficiency was induced in the same cells (Liu et al. 2000). Subsequent studies rather identified Sam68 as primarily involved in the progression of neoplastic transformation in a variety of tissues (Table 1.1). This picture is incredibly complex and to elucidate the full potential of Sam68 as a therapeutic target more research is required. Thus far, the contribution of Sam68 in tumourigenesis has been attributed to its function in signal transduction pathways such as Mek/ERK1 (Matter, Herrlich & Konig 2002), HGF/Met (Locatelli, Lange 2011, Locatelli et al. 2011) and Akt/FOXO3a (Song et al. 2010) as well as alternative splicing.

In prostate cancer cells Sam68 was shown to regulate splicing of *Cyclin D1*, such that overexpression of Sam68 induced selection of Cyclin D1b, the more oncogenic isoform of the protein (Paronetto, Cappellari & Busa 2010). Furthermore, SUMOylation of Sam68 has been shown to reduce the expression of Cyclin D1, affecting cell cycle progression and inhibiting the initiation of apoptosis, although this has not been associated yet with cancer (Babic, Cherry & Fujita 2006). Also in prostate cancer cells, Sam68 has been shown to have separable effects on the alternative splicing of the

androgen receptor, and its activity during transcription, which may drive prostate cancer phenotypes (Rajan et al. 2008).

Additionally, Sam68 has been shown to influence the outcome of splicing of *CD44*, encoding a cell-surface glycoprotein implicated in cell adhesion, cell-cell interactions and cell migration that are required during development, the immune response and tumourigenesis. Aberrant splicing, and in particular inclusion of variable exon 5 (exon v5), has been shown to be important in tumour progression and be regulated by Sam68 as part of the Ras signalling cascade (Matter, Herrlich & Konig 2002).

It has also been suggested that this signal transduction pathway affects the involvement of Sam68 in the epithelial to mesenchymal transition (EMT) and the reverse process, MET, which is a crucial process during embryogenesis in vertebrates (Valacca et al. 2010) but also importantly drives the migratory potential of epithelial cancer cells resulting in invasion and metastasis of tumours (reviewed in (Thiery 2002)). The speed at which this process occurs is enhanced in cells that express a constitutively active splice variant of the Ron proto-oncogene. Selection of this isoform is, in part, regulated by the splicing factor ASF/SF2 which promotes skipping of Ron exon 11 when overexpressed (Valacca et al. 2010). Sam68 has been shown to influence EMT/MET programs through regulation of ASF/SF2 expression and splicing by nonsense-mediated decay (AS-NMD), through the ERK1/2 signal transduction pathway (Valacca et al. 2010).

Furthermore, Sam68 has been shown to affect apoptosis (Taylor, Resnick & Shalloway 2004) and has been identified as a key regulator of *Bcl-x* alternative splice site selection (Paronetto et al. 2007), with tyrosine phosphorylation of Sam68 by Src-like kinases regulating its choice between the short, pro-apoptotic, isoform and the long, anti-apoptotic isoform in live cells (Paronetto et al. 2007). Although this has not been identified in particular cancers, it is likely that this contributes to neoplastic transformation in some way.

Cancer	Effect	Mechanism	Refs
Breast	Upregulation of Sam68 - Increased proliferation and metastasis (decreased affects following downregulation of Sam68)	Link to HGF/Met pathway and breast cancer kinase	(Lukong et al. 2005)(Locatelli et al. 2011)
Cervical	Sam68 expression level and localisation affect lymphatic metastases and prognosis	Unknown	(Li et al. 2012)
Colorectal	Sam68 upregulation and nuclear localisation - Increased proliferation, invasion and migration	Unknown	(Liao et al. 2013)
Glioblastoma	Increased Sam68 expression – Increased proliferation	Linked to ratio with Hsp22, which induces Sam68 expression	(Modem et al. 2011)
Melanoma	Potential Biomarker for disease	Unknown	(Chunyun et al. 2014)
Non-small cell Lung	Increased Sam68 expression - poor prognosis	Unknown	(Zhang et al. February 2014)
Pancreatic	Decreased expression and phosphorylation of Sam68 – cell survival	Associated with tyrosine phosphorylation by Fyn and promotion of Bcl-x(L) splice site selection	(Chen et al. 2011)
Prostate	Increased Sam68 expression - Increased cell survival and proliferation	Alternative splicing of androgen receptor Alternative splicing of CD44 following genotoxic stress and DNA damage Alternative splicing of cyclin D1 Interaction with SND1 (transcriptional coactivator)	(Busa et al. 2007) (Rajan et al. 2008) (Cappellari et al. 2014)
Renal	Increased expression and cytoplasmic localisation of Sam68 – poor prognosis	Unknown	(Zhang et al. 2009)
T-lymphoblastic Leukaemia	Overexpression of Sam68 – increased tumourigenesis	Association with Vav1	(Lazer et al. 2007)

Table 1.1: Sam68 in neoplastic transformation.
Sam68 affects the development of many different cancers. In some cases the mechanism of action is known and involves Sam68 alternative splicing function.

1.4.5. Post-translational modifications

1.4.5.1. Tyrosine Phosphorylation

Sam68 was originally identified in its tyrosine phosphorylated state by western blot analysis of NIH 3T3 cells transfected with Src (Wong et al. 1992). Although initially mis-identified as p62, a GTPase-associated protein, it was soon discovered to be a

substrate of Src during mitosis, and that this interaction was dependent on the SH2 and SH3 domains of Src (Lock et al. 1996) (Taylor, Shalloway 1994). Given the RNA-binding properties of Sam68, it was postulated that Src may be a regulator of RNA processing through this interaction, in a cell-cycle dependent manner. Sam68 was subsequently found to interact with and be phosphorylated by other Src-family kinases (SPKs), such as Fyn, in a similar manner (Richard et al. 1995).

Identification of Sam68 as an adapter molecule in signal transduction pathways was established when it was observed that Sam68 tyrosine phosphorylation is induced by cell surface receptors, such as insulin (Najib et al. 2005), leptin (Martín-Romero, Sánchez-Margalet 2001) and the T-cell receptor (Fusaki et al. 1997). Furthermore, this post-translational modification allows association of Sam68 with SH2-domain containing proteins including Sik/Brk (Derry et al. 2000), Grb2 (Trüb et al. 1997), GRAP (Trüb et al. 1997), Nck (Lawe, Hahn & Wong 1997), PLCγ-1 (Maa et al. 1994), PI3K p85α (Taylor et al. 1995) and members of the Src and Itk/Tec family kinases (Andreotti et al. 1997). Phosphorylation occurs on a tyrosine-rich region at the C-terminus of Sam68 (Figure 1.6), however due to the close proximity of these residues to each other, it has been challenging to identify which particular residues of Sam68 are targeted.

The downstream effects of Sam68 tyrosine phosphorylation include regulation of alternative splicing function of this STAR protein. For instance, phosphorylation by Fyn was shown to influence splice site selection of the *Bcl-x* gene, with the longer, anti-apoptotic isoform being favoured by phosphorylated Sam68 (Paronetto et al. 2007). Additionally, leptin-mediated tyrosine phosphorylation of Sam68 in mouse C2C12 myoblasts was shown to decrease Sam68 RNA affinity using poly(U) RNA agarose beads followed by immunoblotting with anti-Sam68 antibodies (Maroni et al. 2009), inducing a self-regulatory loop of leptin via Sam68 splicing function and the Erk signalling pathway. Furthermore, P59 (Fyn) phosphorylation of Sam68 influences its intracellular localisation and association with YT251-B in nuclear dots, affecting Sam68 splice site selection (Hartmann et al. 1999).

Subsequently, aberrant regulation of Sam68 splicing function through signalling pathways has been shown to drive cancer progression. Tyrosine phosphorylation of Sam68 by Sik/Brk negatively regulates RNA binding affinity (Derry et al. 2000). This post-translational modification has been shown to be downstream of Met tyrosine kinase activation following hepatocyte growth factor (HGF) stimulation. This pathway

has been implicated in enhanced motility and invasiveness of breast cancer cells and is potentially through irregular splicing activity of Sam68 (Locatelli et al. 2011).The structure of the tyrosine-rich domain of Sam68 has been solved in complex with the armadillo repeat domain of the adenomatous polyposis coli (APC) tumour suppressor protein that is often mutated in colorectal cancer cells (Morishita et al. 2011). Phosphorylation of sites in this region of Sam68 affects the binding of APC, which impacts on the regulation of Wnt signalling, however the implications of this interaction with respect to cancer progression have not yet been identified (Morishita et al. 2011).
Association of Sam68 with Vav1, a GDP/GTP exchange factor, is regulated by tyrosine phosphorylation of Sam68 and affects tumourigenesis (Lazer et al. 2007). Finally, Sam68 relocalisation and interaction with Src at the plasma membrane during mitosis and cell attachment in mouse embryo fibroblasts (MEFs) modulates Src association and activation of RhoGTPases. Knockdown of Sam68 in these cells resulted in sustained Src activation and constitutive activation of these molecular switch proteins, resulting in loss of cell polarity and decreased motility. This places Sam68 in a self-regulatory pathway of Src, with potential effects in cancer (Huot et al. 2009).

1.4.5.2. Serine/Threonine Phosphorylation

Sam68 has also been shown to be phosphorylated on its serine and threonine residues by three different kinases; Cdc2 (Resnick et al. 1997), Erk1/2 (Matter, Herrlich & Konig 2002) and Nek2 (personal communication with Professor Claudio Sette).

1.4.5.2.i CDC2

Given that Sam68 was originally identified as a substrate of Src during mitosis, its potential functional role in cell cycle progression was investigated using nocodazole-treated (metaphase-arrested) HeLa and NIH3T3 cells (Resnick et al. 1997). Sam68 was found to be phosphorylated to a greater extent in these growth arrested cells over their unsynchronised counterparts. Furthermore, this STAR protein was phosphorylated on threonine residues during mitosis only, and serine residues during mitosis and interphase stages of the cell cycle. The kinase responsible for serine phosphorylation has not yet been identified; however Cdc2 was proposed to phosphorylate threonines during mitosis, although the precise position of phosphorylation is unclear. The consensus binding sequence of Cdc2 is (K/R)-S/T-P-(X)-(K/R), although this is a

proline-directed kinase and S/TP is the minimal binding sequence. This sequence does not exist within Sam68 STAR domain.

Cdc2 forms a complex with Cyclin B, which initiates mitosis and phosphorylates various regulatory and structural proteins, including Src. Phosphorylation and activation of Src is followed by phosphorylation on tyrosine residues of Sam68. This suggests that Cdc2 may indirectly regulate the function of Sam68 in cell cycle progression through Src. Discovery of Sam68 as a direct downstream substrate of Cdc2 places this kinase as a more direct regulator of Sam68, although the functional implications of this post-translational modification event have not yet been identified with respect to RNA binding and protein-protein interactions. It is also not known when this phosphorylation event occurs within the cell cycle. It is possible that phosphorylation of Sam68 by Cdc2 may occur at the onset of mitosis and have effects on the mitotic machinery and G2/M transition itself. It may also occur during metaphase and affect later and even post-mitotic processes. It is also possible that aberrations in this signalling cascade affect key cell cycle processes and checkpoints such as the G2/M transition and potentially contribute to neoplastic transformation (Resnick et al. 1997). Therefore understanding the implications of Sam68 phosphorylation on cell cycle regulation must be further investigated.

1.4.5.2.ii ERK1

Constitutive phosphorylation of Sam68 on tyrosine residues has been observed in transformed T cells (Fusaki et al. 1997), also identifying Sam68 as a player in cancer development. Deviation of regular splicing of *CD44*, and in particular inclusion of variable exon v5, is a feature of T lymphocytes and lymphoma cells and has been shown to be directly involved in tumourigenesis and cell migration. Inclusion of exon v5 is regulated by activation of the MAPK pathway (Weg-Remers et al. 2001). Sam68 was shown to bind to exonic splice-regulatory elements of *CD44* mRNA and to enhance exon v5 inclusion via the ERK pathway. Indeed, murine Sam68 was found to be phosphorylated by ERK and 8 proline-directed sites, outside of the STAR domain, were identified as potential sites of phosphorylation by this kinase. Indeed, S56, T71 and T84 were found to be essential for ERK-mediated inclusion of *CD44* exon v5 (Matter, Herrlich & Konig 2002). This makes Sam68 a direct link between signal transduction pathways and alternative splicing.

1.4.5.2.iii Nek2

Finally, Sam68 has also been shown to be phosphorylated by NIMA-related kinase, Nek2 (personal communication with Professor Claudio Sette). This protein is involved in centrosome duplication and separation and spindle formation during the G2/M transition (Fry 2002, Fry et al. 2012). Aberrant expression and activity of Nek2 drives aneuploidy and neoplastic transformation (Hayward, Fry 2006). Nek2 is primarily localised at the centrosome, and therefore outside of the nucleus where Sam68 is located. Recently, Nek2 has also been found within the nucleus in various cancer cells, and is associated with poor prognosis in myeloma (Naro et al. 2013). In patient samples, Nek2 was found to be nuclear in breast and lung tissues, and enriched in the nucleus in colon, prostate and cervix tissues. Furthermore, Nek2 is nuclear in cancer derived cell lines of breast, prostate, cervix, colon and seminomas. Importantly, Nek2 was found to colocalise in nuclear speckles with splicing factors including SRSF1, SRSF2, hnRNPA1, hnRNPF and Sam68. This suggests that Nek2 may specifically interact with splicing factors to affect the regulation of splicing (Naro et al. 2013).

Further unpublished work from the laboratory of Claudio Sette has shown that Nek2 phosphorylates Sam68 both *in vitro* and *in vivo*. Both proteins have been shown to be upregulated in breast and prostate cancer, and interestingly in testicular seminomas, but not other testicular germ cell tumours (TGCTs) (personal communication with Professor Claudio Sette). Furthermore, *in vivo* splicing assays using a *CD44* minigene demonstrated that when Sam68 and Nek2 were co-overexpressed, alternative splicing selection and inclusion of exon v5 was enhanced. Interestingly, overexpression of a kinase dead mutant of Nek2 with Sam68 did not affect splicing, although overexpression of active Nek2 alone did enhance inclusion of this exon. Depletion of endogenous Sam68 using RNAi demonstrated that Nek2 induced changes in splicing are dependent on Sam68. These data suggest that Nek2 phosphorylates and regulates Sam68 splicing activity. Further studies are required to understand the specific mechanism by which this PTM affects Sam68 function, and the implications on cancer development.

1.4.5.3. Lysine Acetylation

Sam68 has been shown to be acetylated *in vivo* in HEK293 cells and at higher levels in mammary cancer cell lines (Babic, Jakymiw & Fujita 2004). Acetylation of 21 of the 24 lysines was found within the STAR domain of Sam68 and poly(U) RNA binding was

shown to be positively correlated with acetylation. *In vitro,* CREB-binding protein (CBP) was shown to acetylate Sam68, and overexpression of this common non-histone acetyltransferase in 293T cells induced acetylation of Sam68 *in vivo*, although it is likely that other proteins are also responsible for this post-translational modification (Hong et al. 2002). Further studies must be undertaken to elucidate the level at which acetylation regulates Sam68 activity with respect to tumourigenesis.

1.4.5.4. Arginine Methylation

Sam68 contains two RG-rich motifs that have been shown to be targeted for arginine-methylation by PRMT1 *in vivo* (Côté et al. 2003). This post-translational modification is common for other RNA binding proteins, such as hnRNPs (Shen et al. 1998) and exerts effects on both the RNA binding potential and protein-protein interactions of Sam68.

Given the proximity of one asymmetrical dimethylarginine repeat to a proline-rich region of Sam68, it was shown that arginine methylation by PRMT1 inhibits the binding of Sam68 to SH3 (for example $p59^{fyn}$) but not WW domain proteins (for example FBP30), suggesting that this PTM selectively modulates protein-protein interactions of Sam68 (Bedford et al. 2000).

Furthermore, Sam68 was shown to interact with PRMT1 in a nuclear tripartite complex with the chimeric fusion protein mixed lineage leukaemia – extra eleven nineteen (MLL-EEN) commonly produced following chromosomal translocation in various cancers (Cheung et al. 2007). In particular, Sam68 was shown to interact with the SH3 domain of MLL-EEN via its P3 proline-rich motif. This tri-partite fusion protein complex enhances the self-renewal ability of hematopoietic cells, and knockdown of Sam68 or PRMT1 suppresses MLL-EEN-mediated transformation of these cells (Cheung et al. 2007).

Arginine methylation of Sam68 has also been shown to affect its RNA binding properties. It has been reported that the RG-rich regions harbour RNA binding ability to poly(U) sequences, and that methylation of the arginines in this region reduces RNA binding *in vivo* (Rho et al. 2007), since overexpression of PRMT1 in HEK293 cells reduced the affinity of Sam68 for poly(U) RNA.

Additionally, this post-translational modification prevents the export of unspliced HIV RNAs from the nucleus (Côté et al. 2003). This may be due to a change in intracellular localisation of Sam68 following arginine methylation as deletion of RG-boxes and use

of arginine methylase inhibitors resulted in Sam68 sequestration in the cytoplasm (Côté et al. 2003).

1.4.5.5. SUMOylation

Sam68 has been shown to be modified by SUMO on Lysine 96 (Babic, Cherry & Fujita 2006). This is enhanced by SUMO E3 ligase, PIAS1. Mutation of this residue to arginine reduced the level of SUMOylation, and resulted in a decrease in Cyclin D1 expression, and induction of apoptosis. Conversely, a Sam68-SUMO fusion protein induced an increase in Cyclin D1 expression and reduction in apoptosis. Together, these suggest that the effect of Sam68 on cell cycle progression and apoptosis may be regulated by SUMOylation (Babic, Cherry & Fujita 2006).

1.5. Sam68 Structure

In order to fully understand the mechanisms by which Sam68 exerts its cellular functions, and subsequently how these are altered in cancer and other genetic diseases, it is necessary to understand its molecular structure, and how interactions with other molecules occur.

1.5.1. Domain organisation

The STAR family of proteins contain a conserved STAR domain (also referred to as the GSG domain as it was first identified in Grp33, Sam68 and Gld-1) (Figure 1.6). This domain can be divided into three sub-domains; a central heterogeneous nuclear ribonucleoprotein K homology (KH) RNA binding domain, and two flanking regions, the N-terminal (NK) and C-terminal (CK) domains. These latter two are often referred to in the literature as the Qua1 and Qua2 domains, respectively. The NK domain is responsible for dimerisation (Chen et al. 1997), and is lacking only in SF1, and the CK has been reported to contact the RNA in addition to the KH domain for several STAR family members (Daubner et al. 2014)(Teplova et al. 2013, Liu et al. 2001).

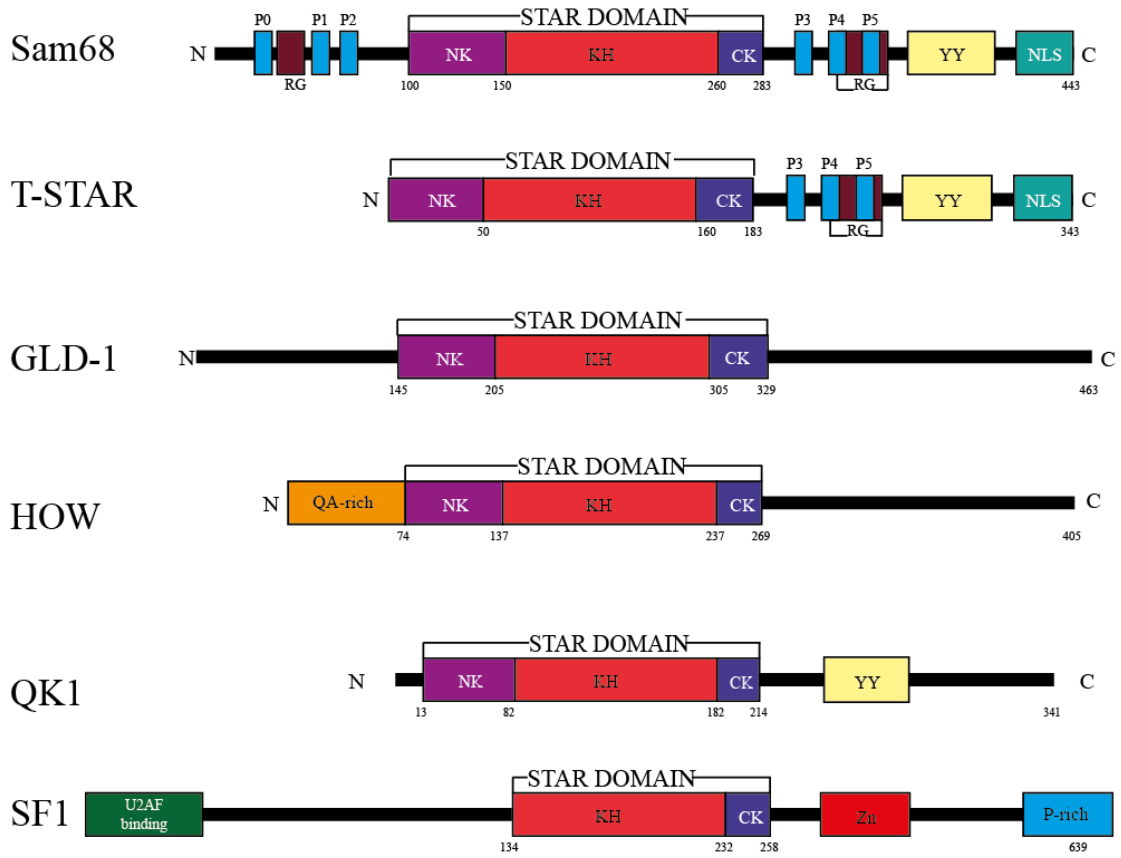


Figure 1.6: **Domain structure of the STAR family of proteins**
Each STAR protein contains an evolutionarily conserved STAR domain composed of a KH RNA binding domain, and two flanking domains; the NK dimerisation domain (except SF1) and the CK domain. Additional regulatory domains are located outside of the STAR domain, and give each protein distinct cellular functionalities. P0-P5: proline-rich regions, YY: tyrosine-rich region, RG: arginine-glycine-rich regions, NLS: nuclear localisation signal, QA: Glutamine-alanine-rich region and Zn: zinc finger domain.

Outside of the STAR domain each protein differs in regulatory regions (reviewed in (Lukong, Richard 2003)). Sam68 and T-STAR are very similar, with T-STAR lacking the first 100 amino acids. Both proteins contain proline-rich motifs (three on either side of the STAR domain in the case of Sam68). These bind SH3 and WW domains to facilitate tyrosine phosphorylation, which occurs at a tyrosine-rich motif at the C-terminus of the protein. Also at the very C-terminus is a non-conventional nuclear localisation signal (NLS) that drives the predominantly nuclear localisation of Sam68. Unlike common NLS motifs that include two distinct basic regions separated by around ten residues, Sam68 contains sparsely based residues. There are, however, two nuclear targeting motifs; a PXXR motif, which is found in T-STAR, SLM-1, Gld-1 and hnRNPC, U, K and M4, and a RXHPYQ/GR motif, which is present in T-STAR, SLM-1 and QKI homologues (Dormann et al. 2012)(Ishidate et al. 1997). Finally, Sam68 also

contains two RG-rich motifs, one at the N-terminus of the STAR domain and one of the C-terminus.

1.5.2. KH domains

The aim of this thesis is to understand the mechanisms by which Sam68 recognises its RNA targets. STAR proteins contain a conserved KH domain which is the main RNA binding domain. This common nucleic acid recognition motif was originally identified in hnRNPK (Siomi et al. 1993) and has since been found across archaea, bacteria and eukaryotes (Siomi et al. 1993). KH domains are generally around 70 amino acids in length and are most commonly found in multiple copies, for example there are two copies in fragile X mental retardation protein (FMRP) (Siomi et al. 1994), three in hnRNPK (Siomi et al. 1993), four in K-homology splicing regulator protein/FUSE binding protein 2 (KSRP/FBP2) (García-Mayoral et al. 2007) and 14 in vigilin (Dodson, Shapiro 1997). These are able to function independently and co-operatively, and often specificity and binding affinity is gained from having multiple sites of binding acting in coordination (Dejgaard, Leffers 1996). STAR proteins are quite unusual in that they only contain a single, larger KH domain, which may confer increased specificity and affinity for RNA through dimerisation of the NK domain. There are two types of KH domain, type 1 which is found in eukaryotes and type 2 which is found in prokaryotes. Both share a minimal RNA binding sequence, but the fold of this region is different.

Several structures of KH domains have been solved, the first of which corresponds to the 1st and 2nd KH domains of FMRP in 1996 by NMR (Valverde et al. 2008). The structure revealed a βααββα fold of the KH domain. This arrangement was also observed for NOVA-2, whose structure was solved by X-ray crystallography in complex with a short hairpin RNA, allowing the identification of a GXXG amino acid motif required for RNA recognition (Lewis et al. 2000, Lewis et al. 1999). For the various KH domain structures that have been solved, several common features have been reported. In each case, the nucleic acid ligand binds in a single-stranded conformation across one face of the KH domain within a conserved binding cleft comprised of α-helix 1 and 2 and β-strand 2. Interestingly this cleft is conserved between type 1 and type 2 KH domains and can accommodate four nucleic acid bases. The residues within the cleft tend to be hydrophobic in nature but a diverse set of interactions with nucleic acid bases are found. In some cases intermolecular π-π stacking interactions have been observed between aromatic side chains and nucleic acid

bases. For example, NusA KH domains 1 and 2, Nova and SF1, two hydrogen bonds are formed between an adenine base and the protein backbone in order to mimic a Watson-Crick base pair interaction.

In general, KH domains bind nucleic acids with low micromolar affinity, which is increased by coordination between one or more distinct KH domains within one protein. This provides increased specificity for the DNA or RNA sequence recognised and very little or no conformational changes occur to the KH domain upon ligand binding.

1.5.3. RNA recognition by STAR proteins

Various techniques have been used to define consensus binding sequences for several members of the family, including SELEX (systematic evolution of ligands by exponential enrichment), CLIP (cross-linking immunoprecipitation) and SIA (scaffold independent analysis) and to measure RNA binding affinities such as gel shift assay, isothermal titration calorimetry (ITC), fluorescence polarisation (FP) and mutagenesis. In addition to these techniques, structures of the STAR domain of Gld-1, QKI and SF1 have been solved in complex with RNA containing the sequence CUAAC by X-ray crystallography, giving further insight into the function of these proteins.

In general, KH domains have been shown to recognise up to four nucleotides and interact with five or more to form a stable complex (as reviewed in (Nicastro et al. 2015)). STAR proteins have been shown to have particular affinity for AU-rich RNA, and table 1.2 outlines the short RNA consensus sites identified by several members of the family (Volk, Artzt 2010). The STAR family of proteins fall into two distinct subfamilies (Figure 1.4) and this distinction is reflected in their RNA consensus sequences (reviewed in (Volk, Artzt 2010)). SELEX experiments with subfamily 1, comprised of Sam68, T-STAR and SLM-1, showed that RNA sequences that bound to Sam68 and T-STAR contained a conserved UAAA or UAA motif, respectively (Lin, Taylor & Shalloway 1997). The other subfamily; QKI and SF1 have been shown to bind a longer 6mer consensus sequence and Gld-1 a 7mer consensus sequence that include cytosine (Ryder et al. 2003, Ryder, Williamson 2004) (Berglund et al. 1997). This suggests that although the function of STAR proteins in binding AU-rich RNA is evolutionarily conserved throughout different species, they are not functionally redundant.

Protein	RNA specificity (5'-3')	Method
Sam68	UAAA	SELEX
T-STAR	UAA	SELEX/CLIP
Gld-1	UACU(C/A)A	Gel shift/Mutagenesis
QKI	A(C/A)UAA	FP/Mutagenesis
SF1	UACUAAC	Gel Shift
HOW	A(C/U)UAA	Pull down

Table 1.2: RNA recognition by STAR proteins. RNA consensus sequences of five STAR proteins determined via different techniques demonstrate the difference in RNA targets of each subfamily.

1.5.4. STAR domain structures in complex with RNA

Since STAR proteins contain only one KH domain, they require an alternative mechanism to increase RNA binding specificity. This is achieved through the dimerisation of the NK domain, and also the RNA binding capability of the CK domain. The first structure of a STAR protein to be solved with RNA was that of the KHCK domain of SF1 with UAUACUAACAA by NMR (Liu et al. 2001) (Figure 1.7A). Interestingly, 5' to the sequence bound by the KH domain, the CK region was shown to specifically recognise the three nucleotides ACU. This domain was shown to be flexible in the absence of RNA, but α-helical in structure with RNA in solution. The precise recognition of RNA by the CK domain of Gld-1 was recently explored in more detail following solution structure determination of the KHCK domain of this STAR protein with the consensus RNA sequence identified in the *tra-2* gene. This allowed structural investigation of several previously studied mutations within the CK domain that cause severe phenotypes in *C. elegans*. These mutations can be divided into two groups. The first group; R328E, R314E, A321D, G227S, G227D and L313A all exert effects on the RNA binding capability of the CK domain. The second group; D310N, G308E. and P228S all interrupt the KH/CK interface and diminish RNA binding affinity. These data suggest that the CK domain is required for correct RNA recognition and modulation of protein function (Daubner et al. 2014). Furthermore, the crystal structures of SF1, Gld-1 and QKI in complex with RNA revealed an additional flexible loop between β- strands two and three that is not present in other KH domain containing proteins. Although not shown to directly interact with RNA, a later study of Sam68 revealed that RNA binding ability is disrupted when this loop is deleted (Lin, Taylor & Shalloway 1997).

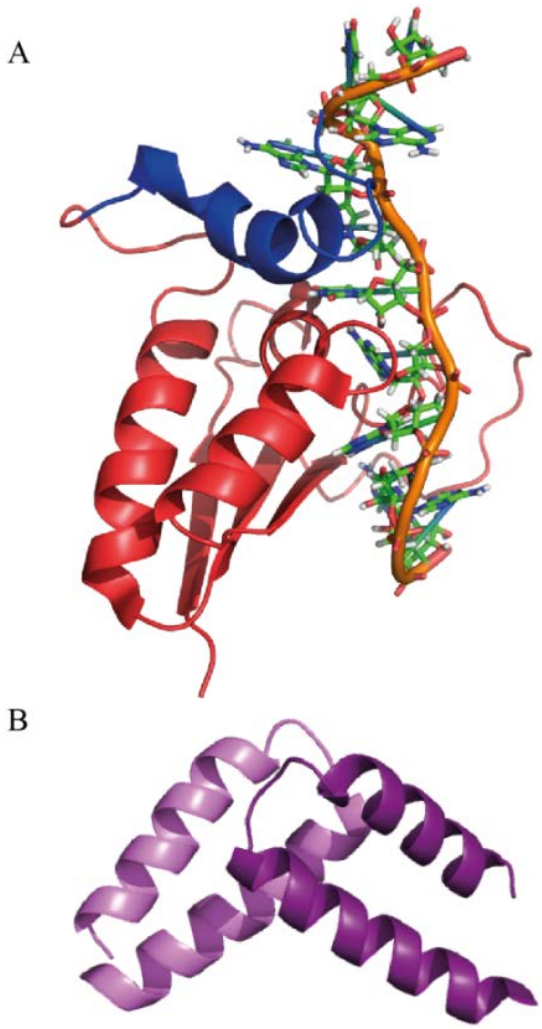


Figure 1.7: **STAR protein structure**
A) The KHCK domain of SF1 in complex with RNA was solved by NMR and demonstrates that both subdomains make contacts with the RNA (Liu et al. 2001). B) The only available structural information for the Sam68 subfamily of STAR proteins is the NMR structure of the NK dimerisation domain, which adopts the same helix-turn-helix motif as the NK domains of Gld-1 and QKI (Meyer et al. 2010), (Beuck et al. 2010) and (Beuck et al. 2012).

The structures of the NK dimerisation domain of Sam68 (Figure 1.7B), Gld-1 and QKI have been solved and reveal a conserved helix-turn-helix motif, stabilised through hydrophobic contacts (Meyer et al. 2010), (Beuck et al. 2010) and (Beuck et al. 2012). In all cases, the NK domain alone was shown to be sufficient for dimerisation and in the case of Sam68, mutation of Y103 to a serine was shown to inhibit *in vivo* splicing of a CD44 minigene, without a large overall effect on the structure of the NK domain and dimerisation *in vitro* (Meyer et al. 2010). This suggests that phosphorylation of this tyrosine residue may play a role in regulation of Sam68 RNA binding activity and that complete formation of a dimer is required for full splicing function of STAR proteins (Cukier, Ramos 2010).

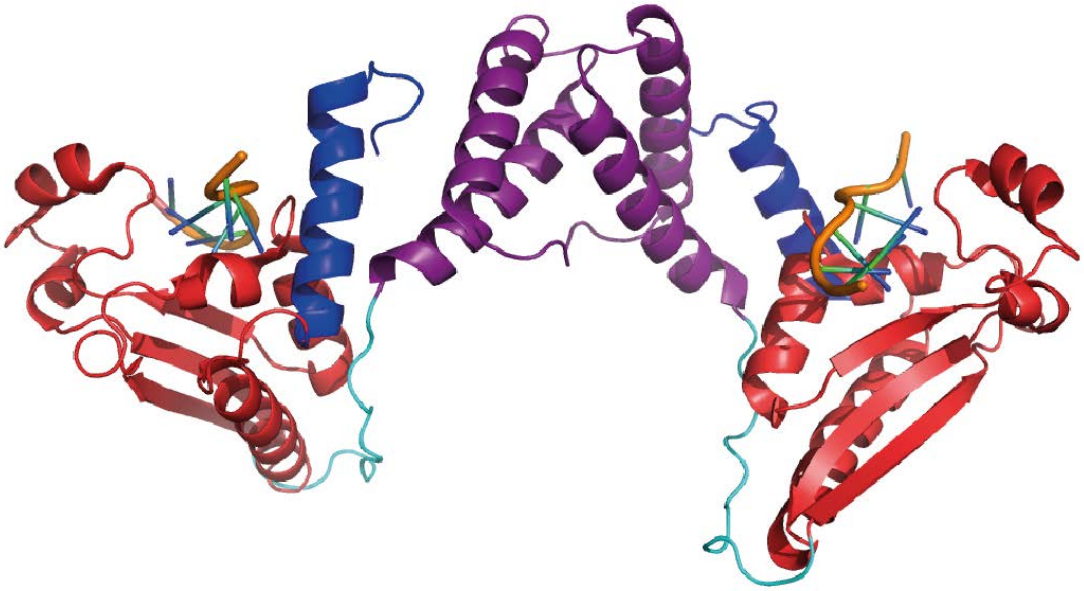


Figure 1.8: Gld-1 STAR domain structure
The structure of the STAR domain of Gld-1 in complex with RNA, with NK (purple), NK-KH linker (cyan), KH (red) and CK (blue) (Teplova et al. 2013).

Recently, the X-ray structures of the full STAR domains of Gld-1 (Figure 1.8) and QKI were solved in complex with CUAACAA and UUCACUAACAA, respectively (Teplova et al. 2013). In both proteins, the KH and CK domains are involved in binding, and the same conserved residues within these domains form the interface between the protein and the RNA. Additionally, these proteins recognise the same conserved UAAC sequence in a similar manner to SF1. These structures also show that the NK domain is sufficient for dimerisation and does not make contacts with the RNA. This domain was required for the precise orientation of the KH and CK domains through direct interactions with the CK domain, and places the two RNA binding domains at opposite ends of the dimer. Other contacts between helix 3 within the KH domain of Gld-1 and the flexible NK-KH linker also drive this orientation. However the linker is not visible in the QKI structure, which suggests that it may be more flexible in other STAR proteins. Finally, the conformation of the Gld-1 and QKI STAR domains indicate that these proteins are likely to bind to a single, long RNA with two protein recognition sites that are separated by at least 10 nucleotides.

1.6. AIMS and objectives

The aim of this thesis is to investigate the structural basis of RNA recognition by the STAR domain of Sam68, with respect to alternative splicing, and how this is regulated by post-translational modifications.

Chapters 3 and 4 will determine the specificity of RNA recognition by Sam68 using structural and biophysical techniques. This will be followed in Chapter 5 by investigations of the effect of serine and threonine phosphorylation on RNA binding using kinase assays and biophysical techniques.

Chapter 2. Methods and Materials

2.1. Materials

2.1.1. Plasmids

Plasmid vectors for bacterial protein expression were designed and available through the PROTEX cloning service at the University of Leicester. pLEICS-01 and pLEICS-03 are modified versions of pET-43.1a(+) and pET-47b(+) vectors, respectively. These modifications allow ligase-free ligation using homology tags and cloning using a BD in-fusion protocol and harbour two methods of selection. The first is antibiotic resistance, ampicillin and kanamycin, respectively; the second is through a *Bacillus Subtilis* (*SacB*) gene, inserted into the cloning region, since bacteria that express this gene cannot survive on agar plates containing sucrose. Furthermore, protein expression is regulated by a lac operon that is inducible using isopropyl-β-D-thiogalactopyranoside (IPTG), and proteins are expressed with an N-terminal tobacco-etch virus (TEV) cleavable 6-Histidine affinity tag. Vector details are seen in full in Figure 7.1.

2.1.2. DNA oligonucleotide primers

All oligonucleotides were purchased from Eurofins MWG Operon (Table 7.1).

2.1.3. RNA oligonucleotides

All unlabelled and fluorescein labelled RNAs were purchased from Dharmacon.

2.1.4. Bacterial strains

Rosetta BL21(DE3) and DH5α competent cell lines were both purchased from Novagen.

2.1.5. Standard chemicals and reagents

All standard chemicals and reagents were purchased from Fisher Scientific, Melford or Sigma-Aldrich unless otherwise stated.

Sam68 antibodies were purchased from Santa Cruz Biotechnology Inc.

Protein crystallography sparse matrix screens were purchased from Molecular Dimensions and distributed by hand to deep well blocks in house and stored at 4°C.

Mini-prep kits were purchaced from Macherey-Nagel and Ni-NTA agarose from QIAGEN.

NuPAGE® 4-20% Bis-Tris Expedeon gels and NOVEX standard marker (Invitrogen) were both purchased from Expedeon and Instant Blue by Expedeon was used for all Coomassie staining.

Chloramphenicol and ampicillin were purchased from Applichem, Kanamycin and Isopropyl-β-D-Thiogalactopyranoside (IPTG) from Melford.

^{15}N-ammonium chloride, ^{13}C-glucose and deuterium oxide were purchased from Sigma-Aldrich.

Active kinases, Nek2, Erk1/2 and Cdc2 were purchased from Millipore.

2.2. Bioinformatics

Amino acid boundaries of the KH (150-260), NKKH (150-283) and STAR (97-283) domains of Sam68 were determined according to secondary structure predictions using RONN (Yang et al. 2005) and alignment of conserved sequences with other STAR proteins using ClustalW (Higgins, Sharp 1988).

2.3. Generation of protein constructs

2.3.1. Primer design

Primers were designed based on construct boundaries to be approximately 21 nucleotides in length complimentary to the DNA template sequence, with an additional homology region specific to the particular plasmid vector of use.

2.3.2. Cloning

All PCR reactions and cloning were carried out by the Protex Cloning Service at the University of Leicester using BD In-Fusion™ Universal PCR cloning system (BD Biosciences Clontech). The gene of interest was amplified by PCR with a homology tag specific to the plasmid vector of choice and fused to the linearized vector by incubation with BD in-fusion enzyme. The recombinant plasmid was then transformed into DH5α, which were grown on sucrose containing agar plates. The pLEICS vectors 1 and 3 contain a *sacB* gene flanked by restriction endonuclease recognition sites. This gene encodes for levansucrase, which breaks down sucrose into fructose and glucose, leading to the production of levan. Cells expressing this enzyme do not survive in the presence of sucrose; therefore the *sacB* gene is used as a negative selectable marker. (Jager et al. 1992).

2.3.3. Site-directed mutagenesis

Site directed mutagenesis was carried out using overlap extension PCR with primers of 50 nucleotides in length that contain the site of mutation centrally. Two PCR reactions were carried out initially, the first with the 5' primer for the construct of interest and the reverse mutagenesis primer, and the second with the 3' construct primer and forward mutagenesis primer. The products of this PCR reaction were purified and used as template material for a second round of PCR using the 5' and 3' construct primers. This results in annealing of the first set of PCR products, with incorporation of the mutation of interest, for cloning into a chosen vector as described previously.

2.3.4. Calculation of DNA Concentration

The concentration of plasmid DNA was determined by UV spectrophotometry at wavelength A_{260}, using Jenway 7315 spectrophotometer.

2.3.5. Sequencing

DNA sequencing was performed by PNACL service at Leicester University to confirm that the recombinant plasmids contain the DNA fragment or mutation of interest. Cycle sequencing reactions were carried out on templates and primers, cleaned using DyeEx columns and analysed using an Applied Biosystems 3730 Sequencer.

2.3.6. Transformation into DH5α

1µl of desired plasmid DNA was added at a concentration of 50ng/µl to 25µl *E.Coli* DH5α competent cells and incubated on ice for 30 minutes. 500µl 2TY media was added and samples were incubated for a further 30 minutes at 37°C in a shaking incubator at 200rpm. The cells were then separated from the media by centrifugation (Eppendorf 5417R) at 13000g, for 5 minutes, at room temperature and 400µl of the supernatant was removed. The cell pellet was resuspended in the remaining media and spread on to pre-warmed agar plate containing 2TY media and Ampicillin or Kanamycin using a sterile glass spreader. The plates were allowed to settle for 10minutes at room temperature before being incubated overnight at 37°C.

2.3.7. Plasmid purification

A single colony of DH5α containing the recombinant plasmid of choice was transferred into 50ml 2TY media and Ampicillin or Kanamycin (100µg/ml final concentration).

Cultures were incubated overnight at 37°C in a shaking incubator at 120rpm and the cells separated from the media the following day by centrifugation (Eppendorf 5810R) for 10 minutes at 20°C, 4000rpm. DNA was extracted using a Mini-prep kit (Macherey-Nagel) following set protocol.

2.4. Protein expression and purification

2.4.1. Preparation of Rosetta BL21 DE3 competent cells

Rosetta BL21 DE3 were plated out onto 2TY agar plates containing chloramphenicol and a single colony was transferred into 100ml 2TY with chloramphenicol (Appendix 7.4). Cultures were grown at 37° until reaching an optical density of 0.3-0.4. Cells were harvested and gently resuspended in 10ml ice cold, sterile 0.1M $MgCl_2$. After incubation on ice for 30 minutes, cells were harvested and gently resuspended in 10ml of ice cold 1M $CaCl_2$ and 14% glycerol. The bacteria were dispensed into 100µl aliquots and stored at -80°C.

2.4.2. Unlabelled Protein expression in Rosetta

Single bacterial colonies were transferred to autoclaved flasks containing 50ml 2TY media and Ampicillin/Kanamycin and Chloramphenicol (100mg/ml final concentration). After overnight incubation at 37°C in a shaking incubator at 200rpm, the cultures were transferred into 1L 2TY media and grown until an optical density of 0.4-0.6 was reached. Following a further hour of growth at 20°C, protein expression was induced by addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG) at 400µM final concentration. Constructs were expressed overnight at 20°C in a shaking incubator at 120rpm before harvesting by centrifugation for 20minutes at 4000rpm at 4°C. Following the removal of media, the cell pellet was transferred to a 50ml Falcon tube and the pellet was either purified immediately or stored at -20°C until required.

2.4.3. Labelled Protein expression in Rosetta

Single bacterial colonies were transferred into 10ml of 2TY media with Ampicillin/Kanamycin and Chloramphenicol (100mg/ml final concentration) in a sterile 50ml Falcon tube and incubated for 6-7 hours at 37°C in a shaking incubator at 200rpm. Cultures were then transferred to appropriately labelled 50ml M9 minimal media (Appendix 7.4) in a sterile flask with Ampicillin/Kanamycin and Chloramphenicol

(100mg/ml) and incubated overnight at 37°C. 50ml overnight cultures were then transferred to 1L M9 minimal media and grown to an optical density of 0.5-0.7 before being transferred to 20°C for one hour followed by induction of protein expression using IPTG (400µM final concentration) and incubation at 20°C overnight. Cells were harvested by centrifugation for 20minutes at 4000rpm at 4°C and the pellets transferred to a 50ml Falcon tube. The protein was then either purified immediately or the cell pellet stored at -20°C until required.

2.4.4. Protein Purification

Frozen or fresh cell pellets were resuspended in 10ml Lysis buffer containing 100µl Triton x100. Cells were subjected to sonication at 10amps for 6 series of 30 seconds on and 30 seconds off, on ice. The lysate was clarified by high-speed centrifugation at 18000rpm for 30 minutes at 4°C and the soluble fraction retained for purification by affinity chromatography. 5ml of packed Ni-NTA agarose (QIAGEN) was added to an extract clean column (GRACE) and storage buffer washed through by gravity flow. The resin was washed with 20ml of 50% ETOH, followed by 20ml of dH_2O and 10ml of lysis buffer before addition of the soluble fraction. This fraction was allowed to flow through the resin by gravity, allowing His-tagged constructs to bind the resin. The resin was then washed in a series of buffers with increasing concentrations of imidazole (Appendix 7.3). Constructs were most commonly eluted in the fractions containing 100mM and 250mM Imidazole, as verified by SDS-PAGE (section 2.4.5). TEV protease was then added for cleavage of the affinity tag. The purified fractions were dialysed with TEV overnight in 2L of dialysis buffer (Appendix 7.3) at room temperature with mixing in 10kDa upper weight limit dialysis tubing (Spectrrum Laboratories Inc) and for a further hour in fresh buffer the following morning. Samples were collected from the dialysis tubing and any precipitate removed before concentration by centrifugation (Millipore 10kDa centricon) at 4000rpm of 5 minute spins with resuspension of material between each spin. 700µl samples were removed from the concentrator and subject to size exclusion chromatography on a Superdex 75 10/300 (GE Healthcare) using an AKTA FPLC system into the desired buffer. 300µl fractions were collected and the efficiency of TEV cleavage and separation of material was determined using SDS-PAGE and the concentration of each fraction of interest was determined by optical density at 280nm (section 2.4.6) before fractions were pooled and further concentrated (Millipore 10kDa centricon) to desired concentration.

Because short ssRNA oligonucleotides are easily prone to degradation, 5 µl SUPERase IN RNase Inhibitor (Invitrogen) was added to the protein sample before gel filtration. RNAse activity was evaluated using Ambion RNAseAlert Lab Test kit according to manufacturer instructions. It is important to note that RNAse inhibitors should not be added to the final NMR sample because the storage buffer contains components with non-labile protons that interfere with the NMR measurements of the proteins and RNAs.

2.4.5. SDS-PAGE

Protein samples of 9µl were taken and added to 3µl 4x Novex NuPAGE SDS buffer (Life Technologies) and boiled for 5 minutes before loading. 8µl of Novex protein marker (Invitrogen) was used and all samples run on Run blue SDS pre-cast gels (Expedeon) at 120V for 1 hour in 1X running buffer (Expedeon). Proteins were stained using Expedeon instant blue staining for several hours.

2.4.6. Protein quantification

Protein concentration was quantified by measuring the absorbance at 280nm or using a Bradford Assay.

The concentration of each protein construct was derived from the OD value at wavelength A_{280} (Traycell Jenway 7315 spectrophotometer). 4µl of gel filtration buffer was used as a blank and then 4µl of sample used to measure the absorbance at OD_{280} and estimate the concentration using Beer's Law;

$$A = \varepsilon \cdot C \cdot L$$

With A corresponding to the absorbance value, ε the extinction coefficient, C the concentration in $mol.L^{-1}$ and L the path length of light in cm. Theoretical extinction coefficients were determined based on amino acid sequence using the Protparam tool of the ExPASy Bioinformatics Resource Portal (web.expasy.org/protparam/). The extinction coefficient of KH C238A is (4470), NKKH (7450) and STAR (9065).

Bradford Assay was carried out according to manufacturer's instructions, using 1ml solution of 20% Bio-rad reagent in dH_2O as a blank, before adding 1µl of protein. The sample was mixed well and incubated at room temperature for 5 minutes before measuring the absorbance at 595nm to give protein concentration in mg/ml using a previously calculated factor from a BSA standard curve available within the laboratory.

2.5. RNA Production

The RNA oligonucleotides were chemically synthesised at a 1 micromole scale (Dharmacon, Thermo Scientific), deprotected according to manufacturer instructions and lyophilised. RNAs were then resuspended in 100µl of dH_2O and pH was adjusted to 6.5-7. RNA concentration was measured by OD_{260} using the theoretical extinction coefficient provided by Dharmacon. Typical RNA stock concentrations ranged between 1 and 4mM.

2.6. Nuclear Magnetic Resonance

2.6.1. Experimental procedure for protein optimisation

NMR samples consisted of 330µl of proteins at concentrations of at least 200µM in different buffers with 20µl of D_2O and were placed in Shigemi tubes. NMR measurements were performed using Bruker AVIII-500 MHz, AVIII-600 MHz, AVIII-600 MHz (equipped with a cryoprobe) and Avance-800 MHz (equipped with a cryoprobe) spectrometers. Data were processed using Topspin (Bruker) and analysed with Sparky.

Optimisation of the buffer and temperature conditions as well as the protein constructs were evaluated using (^{1}H-^{15}N) HSQC experiments for visualizing the ^{15}N-labeled protein signals.

2.6.2. Experimental procedure for chemical shift perturbation experiments

Reference (^{1}H-^{15}N) HSQC experiments with watergate for solvent suppression were recorded of the free protein, then the RNA of interest was added directly to the sample tube at a ratio of 0.5:1 (RNA:protein) before recording another (^{1}H-^{15}N) HSQC and TOCSY experiment. This was repeated for ratios 1:1 and 1.5:1, and the final sample was lyophilized and resuspended in D_2O for recording a 2D NOESY with presaturation for solvent suppression.

The chemical shift changes in each dimension were combined using $\Delta\delta = \sqrt{(\Delta\delta H)^2 + (\frac{\Delta\delta N}{R})^2}$, where $\Delta\delta$ represents the combined chemical shift and $\Delta\delta H$ and $\Delta\delta N$ are the change in chemical shift in the proton and nitrogen dimensions, respectively. R is a scaling factor, 6.51.

2.6.3. Experimental procedure for kinase assay

Reference (^{1}H-^{15}N) HSQC experiments were recorded of the free protein, then 5mM $MgCl_2$ and 0.9mM ATP final concentration were added directly to the sample tube before another reference (^{1}H-^{15}N) HSQC was recorded. Finally 2.9µg of the appropriate kinase was added and a series of (^{1}H-^{15}N) HSQC experiments were recorded over an extended period of time.

2.6.4. Experimental procedure for backbone assignment

Samples of the NKKH and STAR domains were produced ^{15}N and ^{13}C labelled in several litres of M9 minimal medium containing 1g/L ^{15}N ammonium chloride and 2g/L ^{13}C glucose. The final concentration was at least 500µM and HNCA, HNCACB and HN(CO)CA experiments were recorded at 30°C. Additional ^{15}N samples of each construct were produced to record (^{15}N-^{1}H) NOESY experiments to aid assignment. Spectral analysis and backbone assignment was carried out using Sparky, and existing assignments of the NK domain (Meyer et al. 2010).

2.7. Crystallisation Experiments

All components used for crystallisation trials were prepared separately and were robotically dispensed onto 96 well crystallisation plates in 100nl drops. Prior to trials involving microseeding, the Cartesian robot was used and a stock solution of protein and RNA was prepared in various ratios before dispensing. For cross-microseeding experiments, each component was dispensed separately using a Mosquito.

2.7.1. Seed stock preparation

Crystal microseed stocks of T-STAR and Sam68 crystals were generated according to the Douglas instrument protocol (Shaw Stewart et al. 2011). Stocks were made in protein buffer.

2.7.2. Plate preparation

Commercially available, 96 condition screens (Hampton research) were used for crystallisation trials and 80µl of each condition was dispensed by hand into MRC 96 well sitting drop crystallisation plates. Protein, RNA and seed stocks were dispensed in equal volume with reservoir solution to form a 100nl drop.

2.7.3. Data collection

Crystals were harvested and cryofrozen in mother liquor containing cryoprotectant based on the mother liquor that had been pre-tested to minimise formation of ice crystals. Crystals were stored in liquid nitrogen for transportation to Diamond Light Source synchrotron for data collection.

2.8. Modelling

Homology modelling was carried out using the software Modeller (Eswar et al. 2007). This comparative modelling strategy incorporates fold assignment, target-template alignment and model generation and evaluation processes. Sequence alignments were generated between the protein of interest and protein of known structure using ClustalW.

2.9. Small angle X-ray scattering

2.9.1. Data collection

SAXS data were collected on the in-house equipment at Technische Universität München under the guidance of the laboratory of Professor Michael Sattler. Each sample was measured according to standard protocol within the group at three different concentrations, and buffer samples were recorded before and after each protein sample.

2.9.2. Data processing

SAXS data were processed using the ATSAS software suite (EMBL). The raw scattering measurements of the buffer samples were averaged and subtracted from those of the protein at each concentration. Scattering curves were produced using PRIMUS for calculation of the radius of gyration and distance distribution. At this stage the quality of the data was assessed and the range of data points to be considered adjusted to account for beam alignment. DAMMIF was then used to produce 5 SAXS envelope models per protein at each concentration, with and without symmetry that were then averaged to produce a final model envelope for each concentration.

2.10. Circular Dichroism

Circular Dichroism spectra were recorded in 1mm quartz cuvettes using a chirascan CD spectrometer (Applied Photophysics) at 20µM protein concentration in 400µl sample volume. Data was acquired at 25°C from wavelengths of 190nm to 280nm.

2.11. Fluorescence polarization

Fluorescence polarisation experiments were carried out in black 96-well plates (Life Sciences) with a 50µl sample volume per well. Proteins were serial diluted across the plate from 0 to 200µM and fluorescein labelled RNA added at 0.2µM final concentration across the whole plate. Plates were spun immediately after addition of RNA, and read using a Perkin Elmer Victor X5 plate reader. Plates were mixed on the in-built shaker before measurements at excitation wavelength of 531nm and emission at 595nm.

2.11.1. Data Analysis

The fluorescence polarization data were analysed using GraphPad Prism with the one-site specific binding algorithm that satisfies the equation y=Bmax*x/(Kd+x). Here, x is given by the protein concentration (µM) and y is the FP value (no units). Bmax describes the maximum specific binding of RNA and Kd represents the dissociation constant of the protein to the RNA.

2.12. *In vitro* kinase assay

2.5µg of purified protein was incubated with 200ng of active kinase of interest (Millipore) for 30 minutes at 30°C in 40µl of kinase buffer containing 1mCi [^{32}P]-γ-ATP (Appendix 7.5). 50µl sample buffer was added to stop the reaction. Samples were boiled and run on SDS-PAGE gel as described previously. Proteins were visualised by staining the gel with Expedeon instant blue for 2 hours and destaining with water. Gels were then dried onto blotting paper for 1 hour in a vacuum and exposed to X-ray film for 5 minutes before developing. The exposure time was adjusted as required.

2.13. Mammalian Cell Culture

Cells were removed from liquid nitrogen and fully thawed in a 37°C water bath before addition to warmed DMEM media (supplemented with 10% FBS and 1% Penstrep) in a 15ml falcon tube. The cells were then centrifuged for 5 minutes at 1000rpm and the

supernatant aspirated to remove all DMSO. After resuspension in fresh DMEM, cells were then added to plates and transferred to 37°C incubator with 5% CO_2.

Cells were incubated at 37°C in a humidified atmosphere of 5% CO_2 in either DMEM (MCF7, HBL100 and HeLa cell lines) or RPMI (PC3 and PTN2C2) media supplemented with 10% FBS and 1% penstrep (Sigma) in Petri dishes. At ~80% confluency, cells were diluted by washing in 1x PBS and incubating for 5 minutes in Trypsin at 37°C to detach cells from the plate. After further washing in 1x PBS, a subset of cells were added to fresh plates with fresh media and incubated for several days to allow re-attachment and growth back to 80% confluency.

2.13.1. Immunofluorescence

At 80% confluency, cells were stripped from plates and resuspended in fresh media for transfer to cover slips. After adherence to cover slips and growth to ~60% confluency, which took from 1 to 3 days, cells were washed in 1x PBS, fixed in ice cold methanol and stored at -20°C.

Cover slips were washed three times in 1x PBS, blocked in 1% BSA for 1 hour at room temperature and washed three more times in 1x PBS before addition of primary antibodies. Combinations of Nek2 (Rabbit), γ-tubulin (mouse and rabbit) and Sam68 (mouse) antibodies (Millipore) were added to the cover slips at a 1:1000 dilution in 0.3% BSA and incubated for 1 hour at room temperature. Following three further washes in 1x PBS, cover slips were incubated in secondary antibodies; either Goat anti-rabbit (Alexaflor 488) or Goat anti-mouse (alexaflor 594) and Hoescht dye in 0.3% BSA, for 1 hour in the dark. Finally, a final three washes in PBS were carried out before mounting the cover slips onto slides for viewing by a Nikon TE300 inverted fluorescence microscope.

Chapter 3. – Optimisation of Protein Expression and RNA binding

3.1. Introduction

The aim of this chapter is to determine the structure of Sam68 in complex with RNA using nuclear magnetic resonance (NMR). This will address the first objective; understanding structurally how Sam68 interacts with RNA and also the specificity of this recognition.

An overview of NMR studies of protein-RNA complexes is described first and discusses the necessity for optimisation of a suitable protein sample and RNA sequences. This will be followed by a description of the optimisation process of screening for an optimal protein construct and investigation of appropriate RNA sequences from the initial pool of RNAs designed.

3.2. Using NMR to study protein-ssRNA complexes

The two major techniques for solving the structure of protein-RNA complexes are X-ray crystallography and nuclear magnetic resonance (NMR). Although only around 10% of the total structures deposited in the protein database (pdb) last year were solved by NMR, 53% of the protein-RNA complexes were determined using this technique (Foot, Feracci & Dominguez, 2014). There are several reasons why NMR is such a useful tool in the structural investigation of such complexes (as reviewed in (Foot, Feracci & Dominguez 2014, Daubner, Cléry & Allain 2013, Cukier, Ramos 2011, Mackereth, Sattler 2012, Dominguez et al. 2011)). Firstly, this type of complex has some intrinsic properties that may interrupt the formation of crystals altogether. For example, most RBPs are composed of multiple RNA binding domains. These domains are globular and of approximately 100 amino acids in length, making them ideal targets for NMR, they are generally separated by flexible linkers that inhibit crystallisation.

Interactions between individual RBDs and RNA are relatively weak but high affinity binding is achieved when many RBDs bind a single RNA. Thus, many of these complexes are highly dynamic in nature, which can also contribute to difficulties in crystallisation. In addition, ssRNAs themselves are highly flexible and difficult to crystallise alone and with protein binding partners.

Aside from circumventing potential complications in studying these complexes by X-ray crystallography, NMR provides an extremely powerful method by which to rapidly screen different RNA targets of a particular RBP. Although there are several useful

techniques for identification of consensus binding sequences, such as cross-linking and immunoprecipitation (CLIP) (Ule et al. 2005, Ule et al. 2003) and systematic evolution of ligands by exponential enrichment (SELEX) (Tuerk, Gold 1990), the sequences generated are often degenerate and defining an optimal RNA sequence to investigate structurally is far from simple. NMR really lends itself to screening many different RNA sequences once a protein sample has been optimised.

3.2.1. Optimisation of protein expression

There are several different properties of an NMR protein sample that must be considered before embarking upon a structural project. The initial step in the optimisation process is to define the minimal region of the protein to express. Since NMR spectroscopy is limited by the size of the system, with a current upper molecular weight limit of approximately 50kDa for structure determination, it is often not feasible to investigate the full length of a protein. As previously discussed, many RNA binding proteins lend themselves to NMR spectroscopic analysis as they can be divided into distinct units, and/or globular RNA binding domains of suitable size for NMR studies. These can be defined by identification of conserved domains, multiple sequence alignment strategies, secondary structure predictions and disorder predictions. Even homologous domains and similar protein constructs have the potential to behave differently in terms of expression and purification, and therefore various strategies may have to be undertaken in order to produce an NMR sample that is stable, homogeneous and highly concentrated. The Sam68 constructs that are discussed in this chapter are defined in Figure 3.1.

To assess the quality of a protein sample for NMR experimentation and structure determination, (^{1}H-^{15}N) heteronuclear single quantum coherence (HSQC) spectra are often used. This experiment is a standard for determining the suitability of a protein sample for NMR studies. It is also used for investigating binding between a protein of interest and ligand, such as other proteins, nucleic acids and small molecules (Zuiderweg 2002). This two dimensional experiment transfers magnetisation between a proton and its attached nitrogen, which is made NMR visible by enrichment in ^{15}N isotope. This is accomplished by expressing recombinant proteins in bacteria that are grown in a minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source. The spectrum produced gives a crosspeak for each NH and NH_2 group at the frequency of each proton and nitrogen in each dimension. The position of each peak in the spectrum

is dependent on the local chemical environment of each particular residue, thus for folded proteins, each atom has a distinct frequency. It follows that for all amino acids (except for the N-terminal residue and prolines) there is a corresponding peak in the HSQC, given that they each contain a backbone amide group. Thus this spectrum is often referred to as the fingerprint of the protein and is useful in the optimisation process to determine the suitability of the sample for further experimentation by NMR. Furthermore, since the position of each peak in the spectrum is dependent on the chemical environment of that particular residue, if this is perturbed in any way, a shift in position will occur. In this way, residues that are affected by interactions with binding partners can be identified.

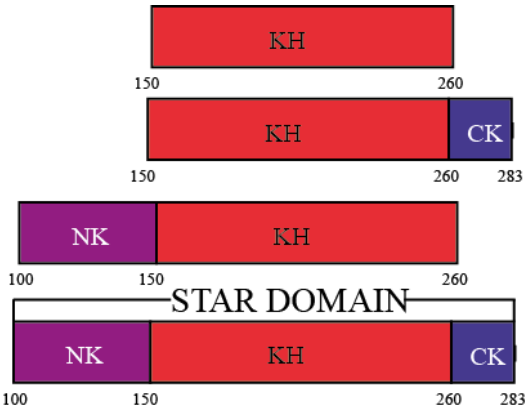


Figure 3.1: Sam68 constructs
Schematic representation of the domain boundaries of the KH, KHCK, NKKH and full STAR domain constructs used for cloning and NMR studies.

In order to obtain good quality spectra, there are several sample properties which must be fulfilled. Firstly, the construct, as discussed, must be less than 50kDa in size for NMR structure determination to be feasible. Following purification, the sample must be homogenous and free from impurities such as bacterial proteins. It is favourable for affinity tags used for purification to be cleaved from the construct since these are either highly flexible (His-tag) or very large (GST-, MBP-tags) and give additional NMR signals that interfere with those of the protein of interest. Finally, the sample must remain soluble to high concentrations (ideally >100μM).

In addition to the properties of the protein itself, the sample conditions also affect the quality of NMR spectra. The temperature, salt concentration, pH and type of buffer must all be considered during the optimisation process to produce an appropriate, stable sample for NMR studies. In general, NMR studies of proteins are conducted at 20-40°C, with a trade-off between better quality spectra at higher temperatures and better stability of protein samples at lower temperatures. It is not recommended to exceed 150-200mM

sodium or potassium chloride in the buffer used, as salt concentration is inversely proportional to NMR sensitivity. To prevent exchange of amide protons, the pH of the NMR buffer should be below pH 7.5 and take into consideration the theoretical isoelectric point (pI) of the construct. Finally, the buffer itself should not be protonated, as the concentration of the buffer is generally higher than that of the protein sample and therefore the buffer NMR signals would interfere. Buffers typically used for NMR include Tris-HCL, sodium phosphate and HEPES at concentrations between 10 and 50mM.

3.2.2. Selection of RNA oligonucleotides based on current literature

As described in section 1.5.3, STAR proteins gain functional specificity through recognition of particular RNA sequences (as reviewed in (Ryder, Massi 2010)). Identification of these consensus sequences has been undertaken for several STAR family proteins, including GLD-1 (Ryder, Williamson 2004), QKI (Galarneau, Richard 2005), SF1 (Berglund et al. 1997), Sam68, T-STAR (Lin, Taylor & Shalloway 1997) and HOW (Israeli, Nir & Volk 2007), using various techniques including CLIP, SELEX and SIA.

CLIP uses UV cross-linking to identify physiological RNA targets of proteins (Ule et al. 2005, Ule et al. 2003). This is done by irradiating sample material such as cell lysate, which results in the formation of covalent bonds between proteins and RNA. The protein-RNA complex can then be separated by immunoprecipitation and the bound RNA sequenced. The RNA sequences identified can then be used to derive a consensus binding sequence. SELEX and SIA are both *in vitro* techniques that identify non-physiological RNA binding sequences (Aquino-Jarquin, Toscano-Garibay 2011). SELEX comprises a series of selection rounds of interacting RNA from a randomised oligonucleotide library which can subsequently allow derivation of a consensus sequence. SIA is an NMR based method which uses (^{1}H-^{15}N) HSQC spectra to screen short synthetic randomised RNA sequences (Beuth et al. 2007). This technique is particularly useful when no information is available regarding RNA consensus sequences and for determining NMR spectral quality.

In all cases, the STAR protein family has been shown to preferentially bind AU-rich RNA (reviewed in (Feracci, Foot & Dominguez 2014)). There are however, notable differences in the consensus RNA sequences identified for different STAR proteins,

giving rise to functional diversity within the family (Table 1.2). For example, Gld-1 and QKI have nucleotide recognition sequences of 5'-UACU(C/A)A-3 (Ryder et al. 2003) and 5'-A(C/A)UAA-3' (Ryder, Williamson 2004), respectively, whereas Sam68 preferentially binds to RNA containing the 4mer sequence 5'-UAAA-3', as determined by SELEX (Lin, Taylor & Shalloway 1997) as well as poly(U) (Itoh et al. 2002) and T-STAR to UAA as determined by CLIP (personal communication with Dr Sushma Grellscheid and Professor David Elliot). Several structures of STAR proteins have been solved in complex with RNA. The STAR domains of QK1 and Gld-1 were solved by X-ray crystallography in complex with UUCA**CUAAC**AA and **CUAAC**AA, respectively (Teplova et al. 2013), whereas the KHCK domain of SF1 (which lacks the NK dimerisation domain) was solved by NMR with UAUA**CUAAC**AA (Liu et al. 2001). These RNA sequences all contain a CUAAC motif and in all cases the CK domain makes contacts with the RNA in addition to the KH domain. Given that Sam68 appears to bind an RNA sequence different to the other STAR proteins, and indeed a 4 nucleotide element rather than five or six nucleotides, suggests that Sam68 may interact with RNA in a different manner than other members of the family.

Figure 3.2: Sam68 RNA consensus sequences
Weblogo showing the consensus RNA sequence recognised by Sam68, as determined by SELEX (Lin, Taylor & Shalloway 1997).

Figure 3.2 shows a weblogo representation of the consensus RNA sequence for Sam68 derived from the SELEX data (Lin, Taylor & Shalloway 1997). UV cross-linking of endogenous Sam68 to physiological mRNA targets in NIH3T3 cells identified 23 RNAs containing similar motifs to those found by SELEX (Tremblay, Richard 2005). There are also several accounts in the literature of Sam68 interacting with poly(U) RNA (Taylor, Shalloway 1994) (Itoh et al. 2002). Together, these data were used to design a subset of short RNA to test by NMR (Table 3.1). The minimal binding sequence of 4 nucleotides is relatively short compared to the KH domain of other STAR proteins and therefore NMR studies were mainly focussed on 6mer RNA including the UAAA in order to ensure stable interactions.

UAAAAU	UAAAUA	UAAAUU	AAAUAA
UUUUUU	AAAUUU	UUUAAA	AUUAAA

Table 3.1: RNA oligonucleotides for NMR studies. Short RNA sequences designed based on CLIP and SELEX data for Sam68 and T-STAR used for RNA binding analysis by NMR.

3.2.3. Chemical Shift Perturbation Experiments

Once a suitable NMR sample has been obtained, this can be used in chemical shift perturbation experiments to screen the selected RNA. These assays make use of the ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) spectra, which are recorded first on the free protein as a reference and then again upon addition of increasing amounts of a ligand, in this case short RNA sequences. Since the chemical shift of each atom is dependent on the local chemical environment of the backbone amide of each amino acid, this peak position will be changed if that amino acid is affected by the addition of RNA. This results in a shift in the position of these affected peaks and allows identification of those residues that are either directly binding to the RNA or undergoing a conformational change upon binding. These simple experiments can also provide information regarding the stoichiometry of the complex and allow estimation of the dissociation constant (Kd). The stoichiometry and Kd can be determined by analysing how the peaks change as RNA is added, this is dependent on the exchange regime between the free and bound states of the protein in solution. The binding equilibrium between these two states is dependent on two main parameters; the rate of exchange of complex formation (K_{ex}) and the difference in resonance frequency of the nucleus at its free (v_A) and bound (v_B) position. The relationship between these properties on complex formation gives rise to three possible exchange regimes (Dominguez et al. 2011, Zuiderweg 2002).

Firstly, if a complex is in slow exchange on the NMR timescale, as RNA is added, two peaks become visible for residues whose N-H group is affected by the presence of RNA. One peak is located at the position of the free state and a new peak appears at the position of the bound state. As more RNA is added, the intensity of the peak at the free position will decrease and the intensity of the peak at the bound position will increase. This exchange regime is indicative of a high binding affinity, with a dissociation constant below 200nM and occurs when K_{ex} is much smaller than $2\pi(v_A-v_B)$.

If the exchange between the states is faster, only one peak will be visible in each spectrum, representing the average position between the two states. This occurs when K_{ex} is much larger than $2\pi(v_A-v_B)$ and as RNA is added this peak will gradually shift

from its free position to its bound position. This often indicates low affinity protein-RNA complex formation (dissociation constant greater than 20μM).

Finally, intermediate exchange has been observed for complexes with a dissociation constant between 400μM and 2μM where K_{ex} is similar to $2\pi(v_A-v_B)$. In this case the crosspeaks tend to disappear upon addition of RNA due to line broadening and some peaks will reappear once the protein becomes fully bound and there is an excess of RNA. This is not always the case and therefore this exchange regime is not ideal for studying protein-RNA complexes by NMR, although it is possible (Ramos et al. 2000). The exchange regime can be altered by optimising buffer conditions, temperature and pH of the sample to obtain a more suitable spectrum of the complex.

These relatively straightforward experiments are extremely powerful and can provide a wealth of information regarding complex formation in a variety of systems. The aim of this chapter is to understand the specificity of the interaction between Sam68 and RNA by solving the structure of this complex by NMR. In order to do so, a suitable protein sample must first be optimised for backbone assignment and chemical shift perturbation experiments. These will identify both the residues of Sam68 that bind RNA and a suitable RNA target for further NMR studies and structure determination. The screening of RNA will also provide an insight into the specificity of recognition by Sam68.

3.2.4. Triple Resonance Experiments and Assignment

Once a suitable protein-ssRNA complex has been determined, the process of structure determination begins by assigning each residue to their corresponding cross-peak in the (^{1}H-^{15}N) HSQC. Furthermore, by ascertaining the position of each amino acid, it is possible to identify those that are affected by RNA binding from chemical shift perturbation experiments.

There are several NMR experiments that are required to make these assignments, all of which extend the dimension of recording to include detection of carbon atoms, making it necessary to label the sample with both ^{15}N and ^{13}C isotopes. In general there are at least two triple resonance experiments required to gain enough information for assignment. These include a selection of HNCACB, HN(CO)CA, HN(CO)CACB and HNCA, amongst others (Sattler, Schleucher & Griesinger 1999). The names of these experiments refer to the atoms for which chemical shifts are recorded. For example, the HNCACB experiment provides the chemical shift information for the backbone amide for a particular residue (i), the Cα and Cβ of (i) and also the Cα and Cβ of (i-1).

Whereas the HN(CO)CACB gives the chemical shift information of the backbone amide of (i) but only the Cα and Cβ of the previous residue (i-1). It is therefore possible, by comparing these two spectra, to identify which carbon peaks belong to the residue in question, or to the previous in the sequence.

Since a subset of residues have characteristic and distinct carbon chemical shifts it is possible to identify the type of residue for each NH cross-peak. In this way, if chains of residues can be linked together using the HNCACB and HN(CO)CACB, once a type of residue has also been identified, these chains can be linked to the protein sequence.

Due to the insensitivity of these triple resonance experiments, they must be recorded for several days and the sample must be at least 500μM to optimise the signal:noise ratio.

3.3. Results

3.3.1. Optimisation of the KH and KHCK for NMR

3.3.1.1. Initial Expression and Purification

Initially, the KH domain of Sam68 (amino acids 151-260) was optimised because it is potentially the smallest domain required for RNA binding. The construct was based on the existing structure of SF1 with RNA, which was solved by NMR (Liu et al. 2001). SF1 is unique amongst the family as it lacks the NK dimerisation domain and is able to interact with RNA via its KH and CK domains only. Furthermore, chemical shift perturbation experiments conducted previously within the group on T-STAR indicated that the KH domain is sufficient for RNA binding (Foot, Feracci & Dominguez 2014). The particular domain boundaries were chosen based on disorder prediction at pH6, secondary structure prediction, multiple sequence alignments and identification of conserved domains through comparison to other STAR and KH domain-containing proteins (Figures 3.3 and 3.2).

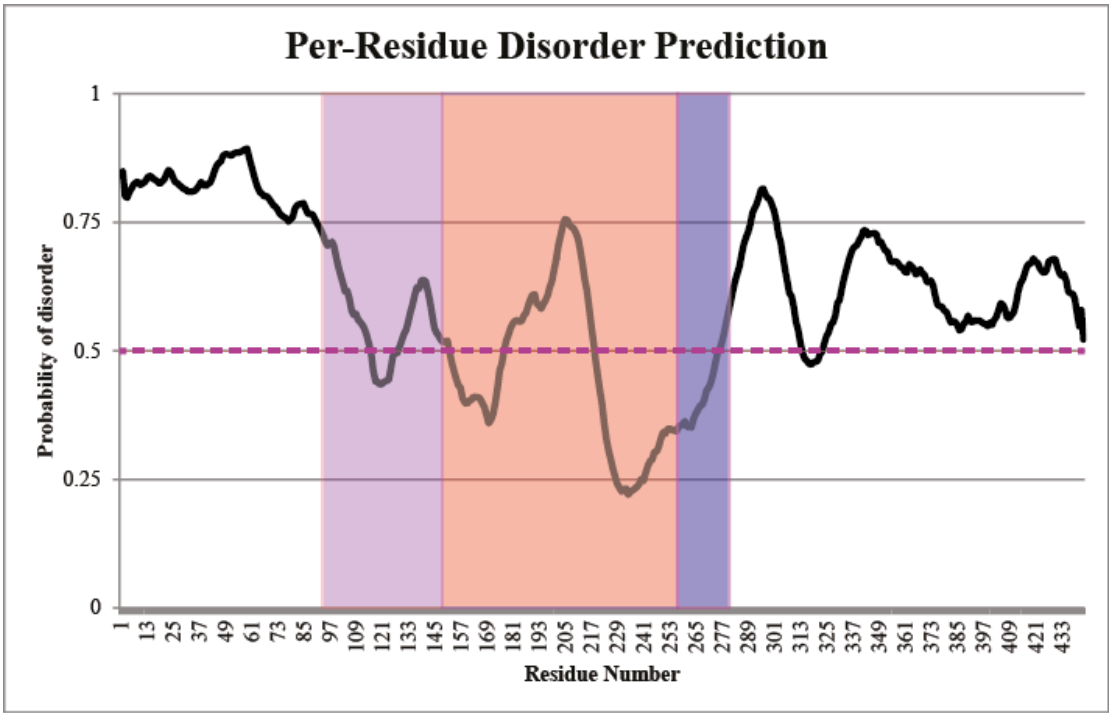


Figure 3.3: Disorder prediction of Sam68

Protein constructs were designed based on RONN disorder prediction with a score of more than 0.5 indicating predicted disordered residues. The NK, KH and CK domains are coloured pink, orange and blue, respectively.

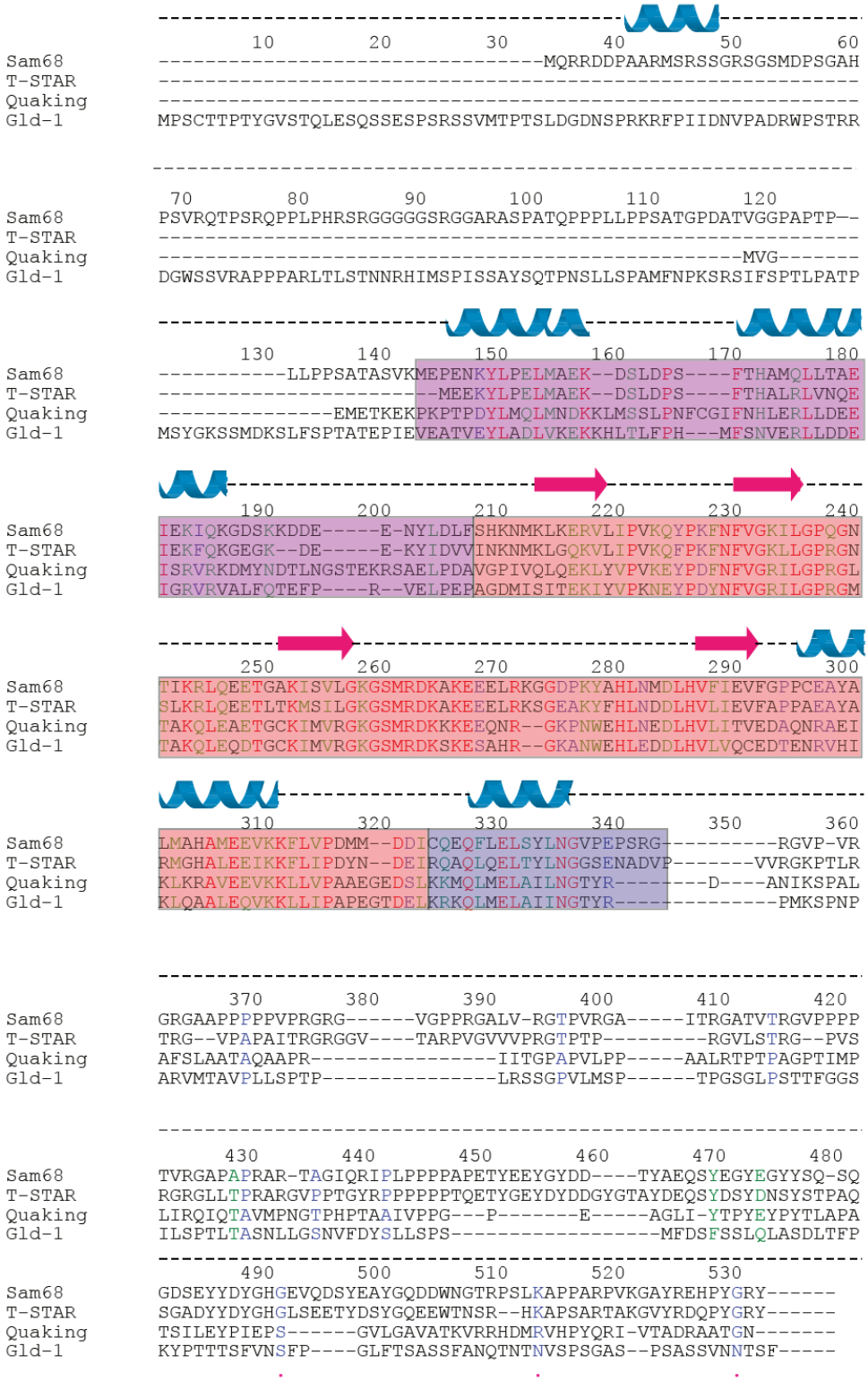


Figure 3.4: Domain conservation between STAR proteins
Protein constructs were selected based on multiple sequence alignment between Sam68 and T-STAR, Gld-1 and QKI. The NK domain is highlighted in pink, the KH in orange and the CK in blue (Higgins, Sharp 1988). Jpred secondary structure predictions are depicted above the sequences (α-helices in blue and β-strands in pink) (Cole, Barber & Barton 2008).

Initial expression and purification was carried out based on protocols previously developed in the laboratory for the homologous protein T-STAR, section 2.4.4 (Figure 3.5). Following affinity chromatography, the imidazole was removed by dialysis in NMR buffer (20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH6 or pH7 at room temperature) before cleavage of the His tag with TEV protease for four hours at room temperature. The final yield from 1L culture was 0.2mM for both 320µl samples at pH6 and pH7.

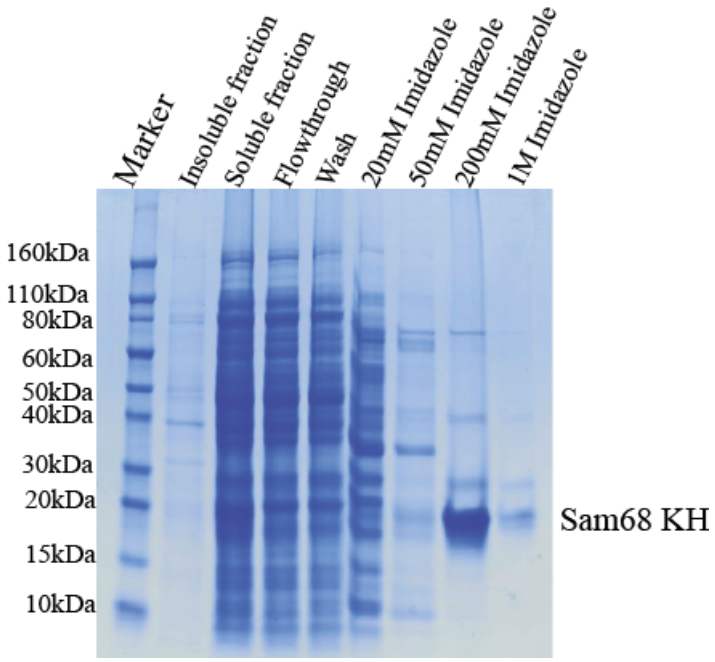


Figure 3.5: Affinity chromatography purification of Sam68
Sam68 KH domain containing an N-terminal poly-histidine affinity tag was purified by Nickel affinity chromatography from 1L *E.Coli*, grown in ^{15}N minimal media. SDS-PAGE was used to analyse the efficiency of the purification. The contents of each lane are indicated above the gel, and represent the protein content of the insoluble and soluble fractions following lysate clarification, and subsequent elutions of the Ni-NTA column with increasing concentrations of imidazole. The KH domain is eluted at 200mM Imidazole, as indicated, and the first lane contains protein marker, of molecular weights as shown.

During dialysis and cleavage of the affinity tag, the sample at pH6 precipitated, resulting in a ^{1}H-^{15}N HSQC spectrum at 20°C showing most amide crosspeaks concentrated in the centre of the spectrum (7.5-8.5ppm in the proton dimension and 115-125ppm in the nitrogen dimension) (Figure 3.6A). This suggests that in these conditions Sam68 KH may not be correctly folded. The sample dialysed at pH7 showed less precipitation and the ^{1}H-^{15}N HSQC was somewhat more dispersed and contained more peaks at a higher intensity (Figure 3.6B), indicating that the KH domain is more stable at pH7.

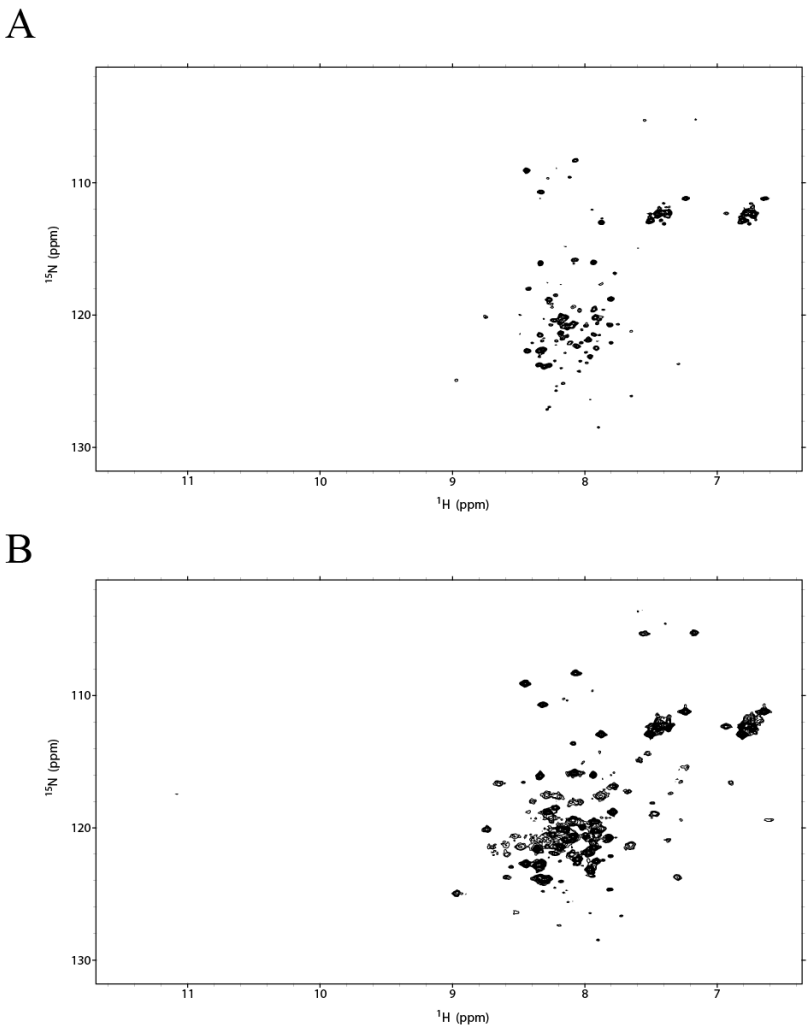


Figure 3.6: **(^{1}H-^{15}N) HSQC analysis of Sam68 KH in different pH conditions**
(^{1}H-^{15}N) HSQC experiment of the KH domain in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol was recorded at pH6 (A) and pH7 (B). Both spectra contain a fewer number of peaks than expected for this domain, but most at pH6 are concentrated in the centre of the spectrum indicating incorrect folding of the KH domain in these conditions. At pH7 there are more peaks, of greater intensity that are more dispersed, suggesting that at this pH the KH domain is folded.

3.3.1.2. TEV cleavage during dialysis

Considering that a significant amount of precipitation was observed during dialysis and TEV cleavage at room temperature, these purification steps were then performed together at 4°C. To allow for the reduced activity of TEV protease, the sample was dialysed overnight and resulted in less precipitation and a greater final protein yield.

The HSQC spectra produced a better dispersion of crosspeaks (Figure 3.7) although there is still some interference of high intensity peaks around the centre of the spectrum due to the presence of the histidine tag.

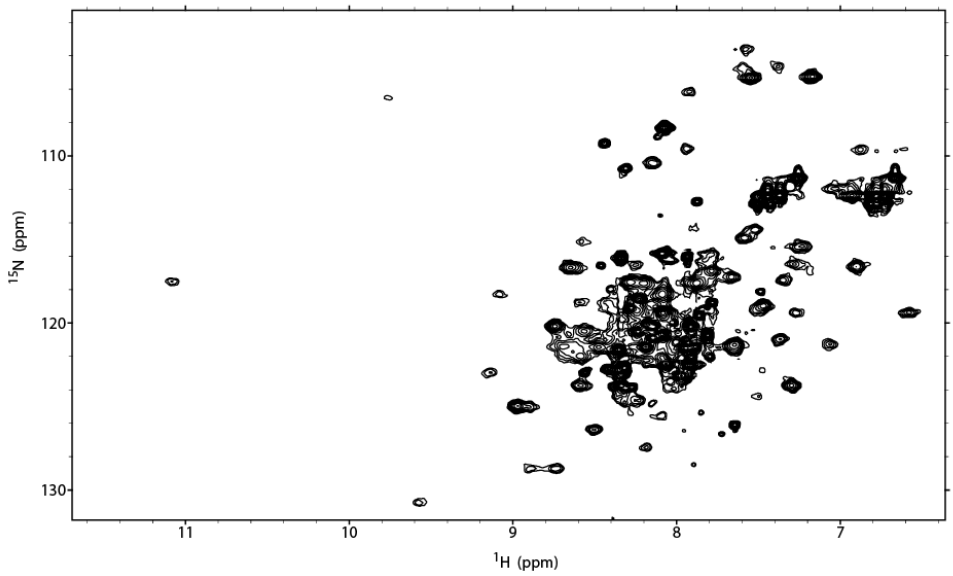


Figure 3.7: **(^{1}H-^{15}N) HSQC analysis of Sam68 KH following TEV cleavage**
The poly-histidine affinity tag was cleaved with TEV protease and the (^{1}H-^{15}N) HSQC spectrum of Sam68 KH in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7 is of better quality than the spectra with tagged-KH domain.

3.3.1.3. Gel Filtration

The presence of intense peaks in the centre of the spectrum was further resolved following size exclusion chromatography. After dialysis, samples were concentrated at 4°C using a 10kDa centricon to a final volume of 700µl before size exclusion chromatography on a Superdex 75 10/300 column in NMR buffer (Appendix 7.3).This allowed separation of the poly-histidine tag, any remaining uncleaved protein and any remaining bacterial proteins from the cleaved protein of interest that were not successfully removed during affinity chromatography (Figure 3.8A). Removal of these impurities resulted in a much better distribution of crosspeaks throughout the HSQC, with 93 peaks observed out of a total 103 expected in the KH domain (discounting prolines and the N-terminal residue) (Figure 3.8B). It was clear that despite improvements in spectral quality, further optimisation of conditions was necessary to obtain a sample of sufficient quality for structure determination. We investigated whether increasing the temperature could improve the quality of the spectra. However, increasing the temperature from 20°C to 25°C caused heavy precipitation and loss of the protein sample.

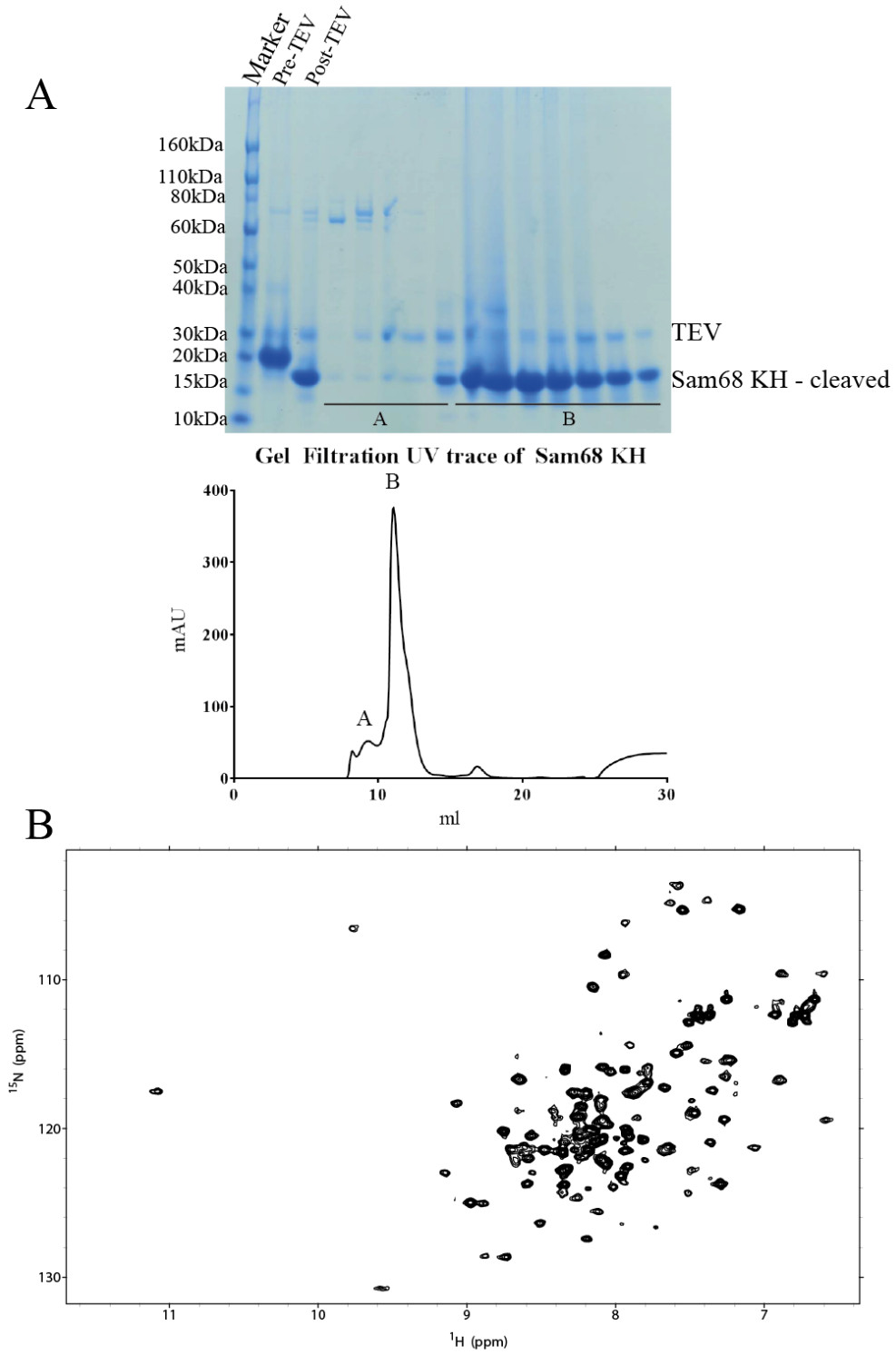


Figure 3.8: Size exclusion chromatography of Sam68 KH
Further purification of the KH domain to separate the tag from the cleaved protein was achieved by size exclusion chromatography (A). The UV trace in the lower panel was used to identify fraction eluates containing Sam68 and successful cleavage, confirmed by SDS PAGE analysis. The KH domain eluted in peak B, and some TEV protease remained in these fractions.(B) The (^{1}H-^{15}N) HSQC showed a much better distinction of peaks than for the uncleaved construct.

3.3.1.4. Optimising the NMR buffer

It was decided to screen various different buffer conditions to try and produce an NMR sample of sufficient quality for structure determination. The screen comprised a range of NaCl concentrations from 0 to 200mM and pH between 5.5 and 7.5. These conditions were tested at 20, 25 and 30°C and the quality of the HSQC was determined based on the number and line width of peaks.

To improve the solubility of the sample at higher temperatures, the ionic strength of the dialysis and NMR buffers was increased. The solubility was not improved by increasing the NaCl concentration to 200mM during purification at pH7 (Figure 3.9A) and protein was lost following gel filtration (Figure 3.9 B). Therefore, with reasonable spectra quality at lower salt concentration, a $^{15}N^{13}C$ sample was produced in buffer of 20mM Na_2HPO_4 pH 7, 100mM NaCl and 0.1% β-mercaptoethanol for NMR. However, the yield was extremely low and the HSQC was not sufficient to warrant recording of triple resonance experiments.

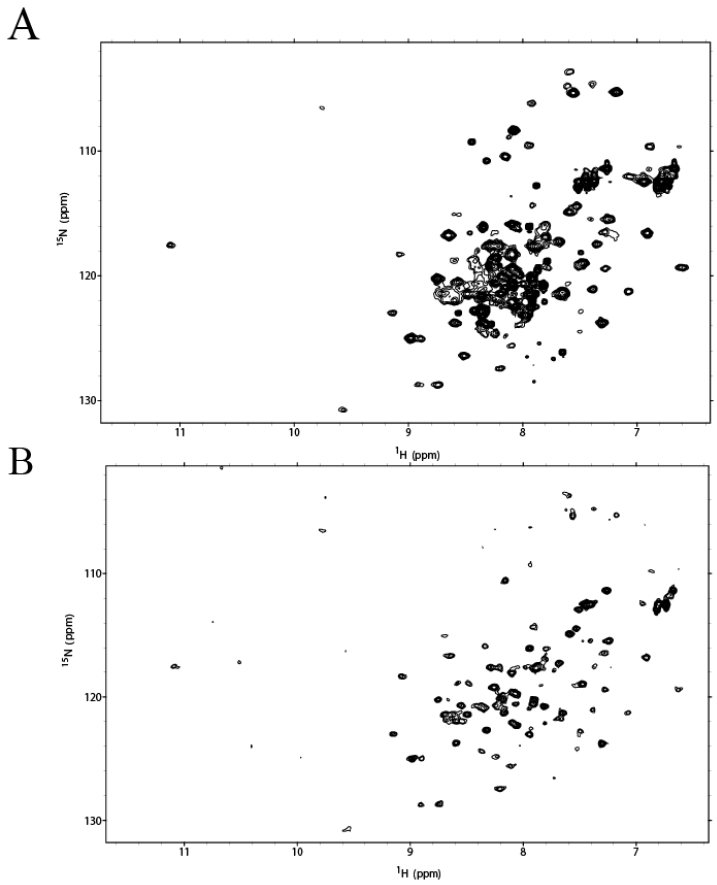


Figure 3.9: **(^{1}H-^{15}N) HSQC spectrum analysis of Sam68 KH in different salt concentrations** The ionic strength of the NMR buffer was increased to 200mM NaCl and a (^{1}H-^{15}N) HSQC spectrum recorded before gel filtration (A). This did not improve sample solubility or spectral quality. Size exclusion chromatography of this sample resulted in a loss of concentration due to precipitation and a poor quality spectrum (B).

To increase yield, Sam68 KH domain was expressed and purified from a total of 5L of ^{15}N minimal media. This allowed purification of several samples of 200µM each, in NMR buffer of 20mM Na_2HPO_4 pH 7, 0.1% β-mercaptoethanol with 100mM NaCl. The resulting HSQC exhibited an increase in quality, with better crosspeak dispersion, increase in the number of visible peaks and increase in peak intensity (Figure 3.10A). The salt concentration of this sample was then halved by dilution to 50mM NaCl in a salt-free buffer. After concentration back to 320µl, another good quality spectrum was recorded with an additional peak becoming visible at around 12ppm (Figure 3.10B).

To assign the backbone of Sam68 KH, triple resonance experiments were acquired using a (^{13}C-^{15}N) sample purified from 5L of culture. Unfortunately the sample was not stable for the duration of the experiments due to precipitation, likely due to aggregation, causing a loss in NMR signal due to a decrease in protein concentration and loss of homogenous magnetisation across the sample. Ultimately backbone assignment of this construct was not possible.

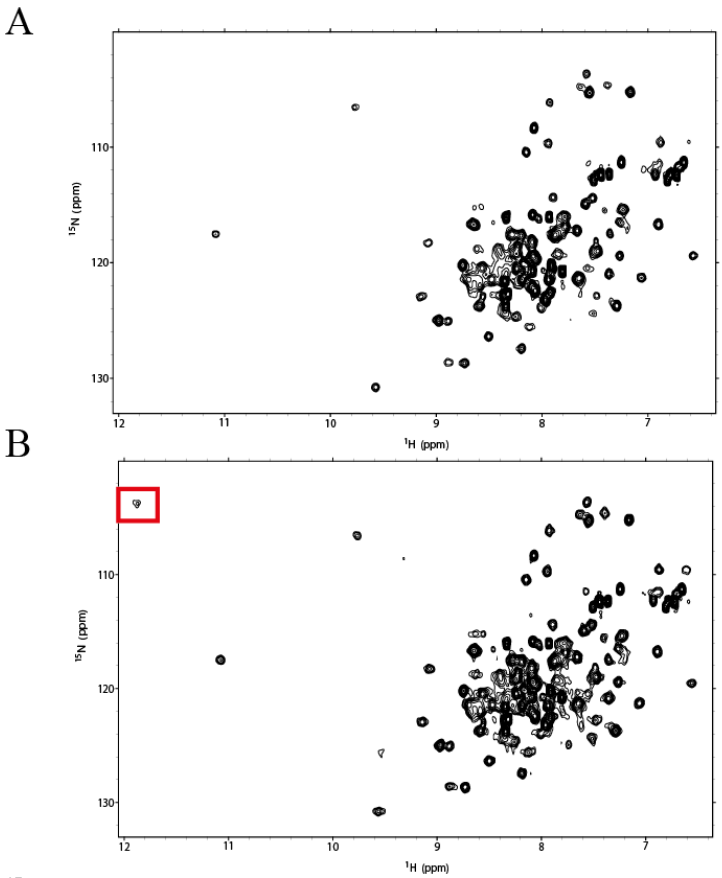


Figure 3.10: **(^{1}H-^{15}N) HSQC spectrum analysis of Sam68 KH from a 5L culture**
To increase protein yield and concentration of the NMR sample, the KH domain was purified from a 5L culture and purified in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7. The (^{1}H-^{15}N) HSQC spectrum contains a greater number of peaks of higher intensity, with an overall improvement in spectral quality (A). The salt concentration of this sample was halved to 50mM NaCl, resulting in one extra peak in the (^{1}H-^{15}N) HSQC (B).

3.3.2. Titration experiments with KH domain

Although the KH domain of Sam68 is unstable it still produced spectra of sufficient quality to perform chemical shift titration experiments to identify RNA sequences that interact with this domain. SELEX data for Sam68 identified a key consensus sequence of UAAA (Figure 3.2). As discussed previously, to ensure optimal binding, sequences of 6 nucleotides were designed to be tested by NMR, rather than a minimal 4mer sequence. Furthermore, the weblogo derived from SELEX data shows that Sam68 appears to preferentially bind adenine and uracil nucleobases over cytosine and guanine. Therefore the majority of the RNA tested flank the UAAA motif with U/As. A variety of RNAs were designed for these preliminary experiments that abide by the SELEX data, CLIP data and physiological RNA targets and a more systematic investigation of RNA binding is described in section 4.5.

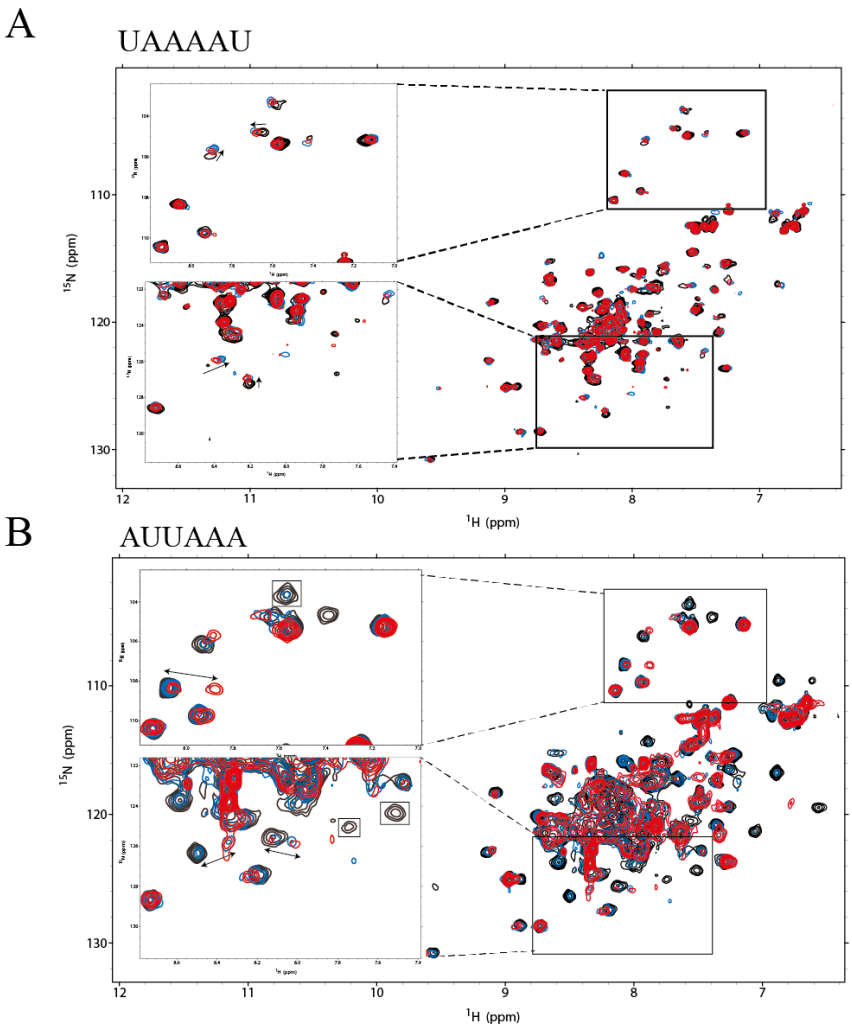


Figure 3.11: **Effect of RNA sequences on Sam68 KH-RNA complex formation**
Chemical shift perturbation experiments of the KH domain of Sam68 with UAAAAU (A) and AUUAAA (B), shown by overlaying the HSQC spectrum of Sam68 KH free (black), and in complex with RNA at protein:RNA molar ratios of 1:0.5 (blue) and 1:1 (red).

Three such sequences, UAAAAU (Figure 3.11A), UAAAUA and UAAAUU, all interacted with the KH domain in the same exchange regime, suggesting that the two nucleotides 3' of the UAAA motif do not affect the interaction. As shown in the full panel and two zoomed panels (Figure 3.11A), several cross peaks undergo chemical shift perturbations upon addition of RNA and can be followed throughout the titration. As previously described in section 3.2.3, this is an indication of fast exchange regime, and suggests that these RNA are binding to the KH domain with relatively low affinity in the µM range.

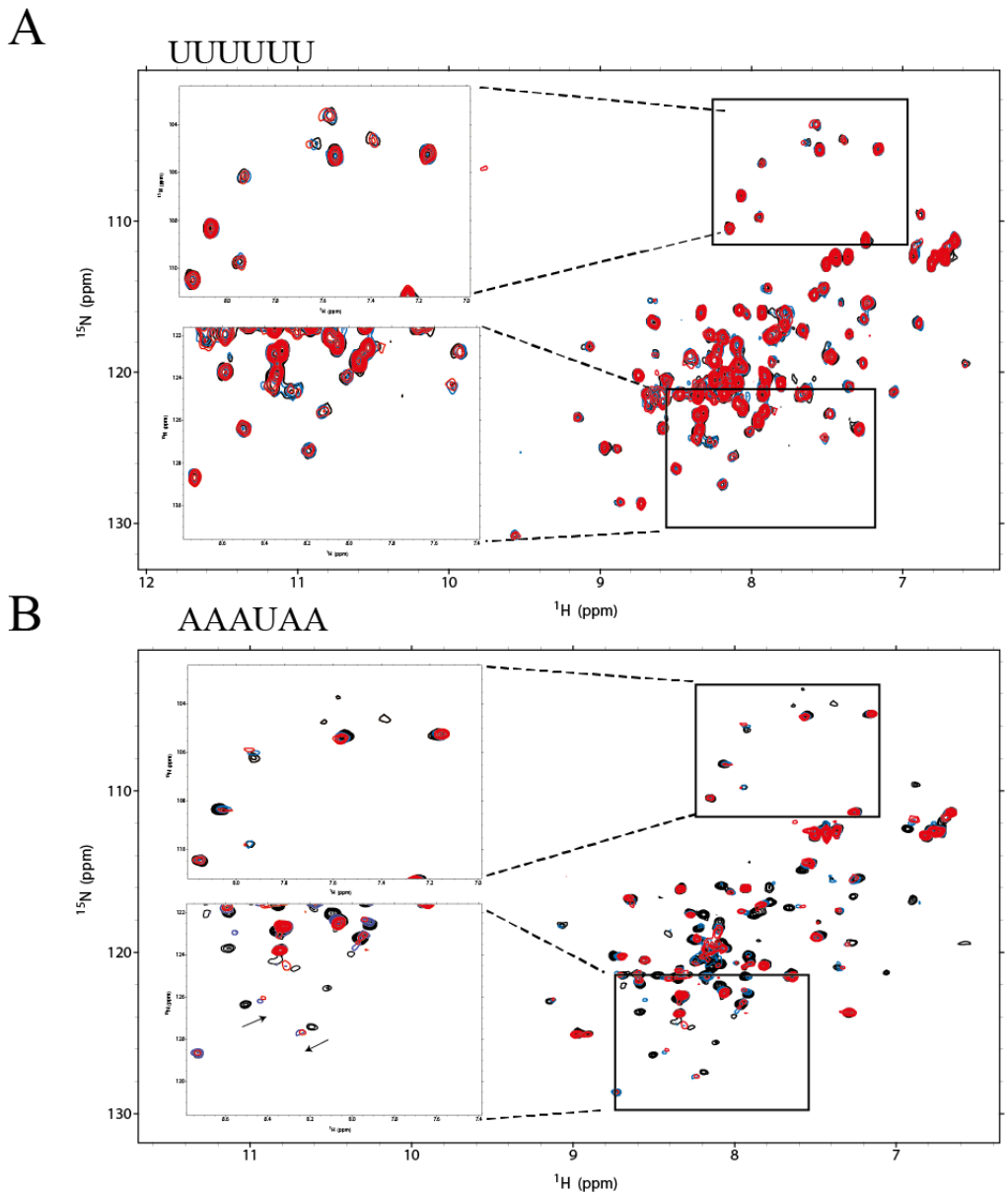


Figure 3.12: **Effect of RNA sequences on Sam68 KH-RNA complex formation**
Chemical shift perturbation experiments of the KH domain of Sam68 with UUUUUU (A) and AAAUAA (B), shown by overlaying the HSQC spectrum of Sam68 KH free (black), and in complex with RNA at protein:RNA molar ratios of 1:0.5 (blue) and 1:1 (red).

To determine the importance of nucleotides preceding the UAAA motif, the KH domain titrations were then performed with AUUAAA (Figure 3.11B). This resulted in the same peaks being affected by the addition of RNA. However, most peaks disappeared upon interaction rather than remaining visible throughout the experiment. Several crosspeaks did not reappear upon further addition of RNA (highlighted in black boxes), whereas others decreased in intensity between consecutive spectra. These peaks were often close in space to new resonances that appeared upon addition of RNA and increased in intensity as the ratio of RNA:protein was increased. This type of chemical shift perturbation is representative of the slow/intermediate exchange regime and indicates that this RNA binds with higher affinity than those in fast exchange. Surprisingly, it was not possible to reproduce this exchange regime with the KH domain and AUUAAA. Instead, this complex appeared to be in fast/intermediate exchange, indicating a lower binding affinity. The slow exchange regime observed in this experiment may be due to slight changes in temperature, salt concentration or magnetic field strength and is unlikely to represent the physiological RNA binding affinity, particularly as it does not confer with later estimates of the Kd (section 4.5) and was not observed for any other chemical shift perturbation experiments with different constructs and RNA.

Two further sequences were tested with this construct of Sam68. The first, UUUUUU has been shown *in vitro*, by differential display and cDNA representational analysis (cDNA-RDA), and *in vivo* by reverse transcription polymerase chain reaction (RT-PCR) from co-immunoprecipitation (co-IP) of Sam68 with nuclear RNA targets to have high affinity for Sam68 (Itoh et al. 2002). However, no chemical shift differences were observed between the reference spectrum and subsequent spectra after the addition of RNA (Figure 3.12A), suggesting that poly(U) RNA does not bind the KH domain of Sam68.

Finally, an RNA sequence was tested based on CLIP data collected for T-STAR (personal communication with Dr Sushma Grellscheid and Professor David Elliot). Interestingly, both SELEX and CLIP experiments focussed on the highly homologous protein T-STAR identified a consensus RNA binding sequence of UAA, potentially giving rise to subtle differences in the function of these two proteins. In particular, AAAUAA has been shown by NMR to bind T-STAR with strong affinity (Foot, Feracci & Dominguez 2014) and was therefore also tested with Sam68 (Figure 3.12B). Once again the same crosspeaks are affected by RNA binding and are in fast exchange. This was also observed for titrations of this construct with AAAUUU and UUUAAA.

These chemical shift perturbation experiments delivered preliminary data regarding the interaction of the KH domain of Sam68 with different 6mer RNAs. Each of these sequences, with the exception of poly(U), which did not interact, and AUUAAA, which was in slow exchange, bound this domain with similar affinity regardless of the nucleotides surrounding the UAAA motif. It was also noted that the same cross peaks underwent chemical shift perturbations in all cases, suggesting that the KH bound each RNA in similar manner. Given the instability of this domain and the challenge of obtaining an NMR backbone assignment, other constructs were optimised in order to further understand the specificity of RNA binding by Sam68.

3.3.2.2. Sam68 KHCK

Since the existing structures of STAR proteins Gld-1, QK1 and SF1, with RNA show that the CK domain also makes contacts with the RNA (Teplova et al. 2013, Liu et al. 2001), we extended the KH to include residues 260-283 of the CK domain. We postulated that addition of these residues would improve the stability of the KH domain within this construct. The KHCK domain was expressed and purified in a similar way to the KH domain alone, however, expression levels were low and some precipitation was observed throughout purification and gel filtration. The HSQC showed a relatively good dispersion of peaks (Figure 3.13); however the peak intensities were low due to the low concentration of sample. It was not possible to achieve higher sample concentration due to the tendency of this protein to precipitate, as was the case for the KH domain alone. This suggests that the addition of the CK domain does not improve the stability of the KH domain and does not influence its conformation.

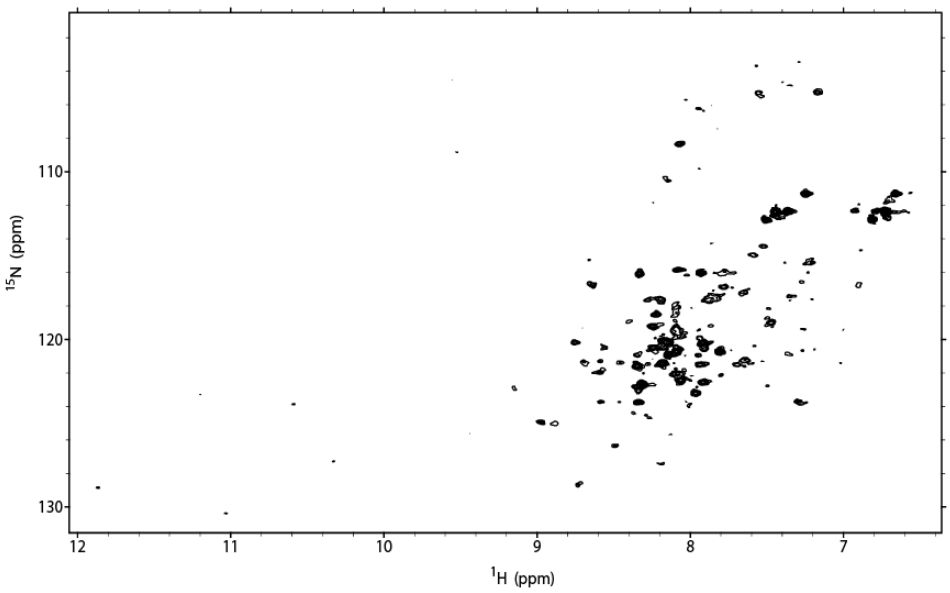


Figure 3.13: **Optimisation of the KHCK domain of Sam68**
(^{1}H-^{15}N) HSQC spectrum of the KHCK domain of Sam68, in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7. The spectrum contains few peaks, at low intensity as this construct was not stable and could not be further concentrated.

3.3.3. Optimisation of the KH C238A mutant

The stability issues of Sam68 KH were surprising because the equivalent construct of T-STAR is highly soluble for several days, at higher temperatures/concentrations and these two STAR proteins are 77% identical within the KH domain (Figure 3.4). The sequence alignment of Sam68 and T-STAR was carefully analysed and one notable difference was observed, which is that Sam68, but not T-STAR, contains one cysteine residue that could potentially form intermolecular disulphide bridges. To investigate this possibility a mutant was designed that replaced C238 with the equivalent residue in T-STAR, alanine. It should be noted that β-mercaptoethanol was included in in all dialysis and NMR buffers as a reducing agent.

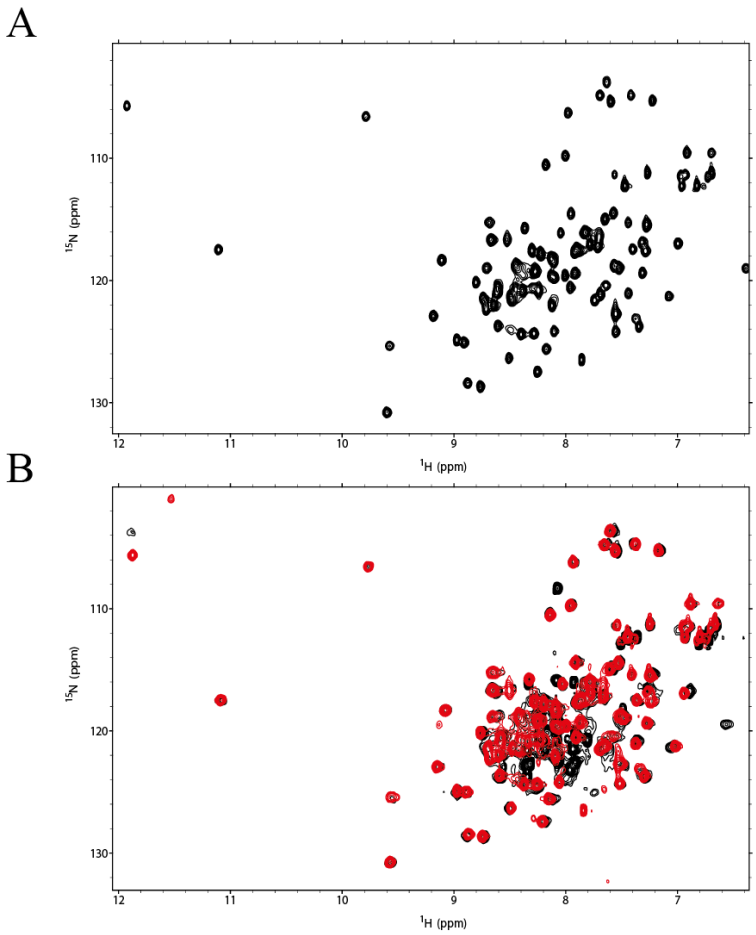


Figure 3.14: Optimisation of Sam68 KH C238A
(^{1}H-^{15}N) HSQC spectrum of the C238A mutant of Sam68 KH in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7 shows a huge improvement in sample solubility and spectral quality (A). Overlay of (^{1}H-^{15}N) HSQC spectra of the KH WT (black) and KH C238A (red) show that the majority of peaks have the same chemical shift in each domain (B).

The same expression and purification protocols as the wild-type (WT) KH domain were used. The HSQC spectrum of the mutant was significantly improved in comparison to the WT KH domain. It was also stable up to 25°C (Figure 3.14A) and overlaid well with the KH wild-type HSQC (Figure 3.14B), indicating that the fold of the KH is not affected by this mutation. This C238A construct was therefore (^{13}C-^{15}N) labelled for triple resonance experiments. Unfortunately, once again this KH domain construct did not remain stable for several days and the triple resonance spectra were of insufficient quality for backbone assignment (Figure 3.15), despite the continued use of β-mercaptoethanol. It is possible that sample stability may be improved by alteration of the reducing agent to TCEP or DTT.

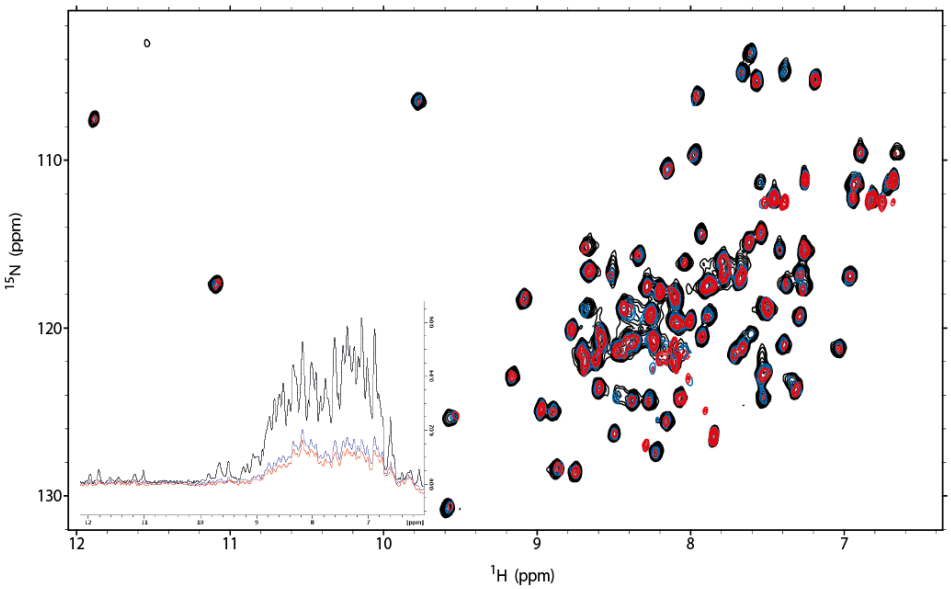


Figure 3.15: Optimisation of (^{13}C-^{15}N)-labelled Sam68 KH C238A
Overlay of (^{1}H-^{15}N) HSQC spectra representing ^{13}C^{15}N Sam68 KH C238A before recording of triple resonance experiments (black), after recording of a HNCA experiment (blue) and after a HNCACB (red) show a loss of peaks over time, due to loss of sample concentration as seen in the 1D spectra shown in the panel.

3.3.4. Titrations with C238A KH domain

Having improved upon the stability of the KH domain alone by mutating cysteine 238 to an alanine residue, chemical shift perturbation experiments were carried out on a subset of the 6mer RNAs to confirm that the mutant KH domain interacts with RNA in a similar manner to the WT KH domain. Analysis of titrations with AUUAAA (Figure 3.16A) and UAAAAU (Figure 3.16B) showed the same peaks shifting as in the WT KH domain. In all cases, some peaks displayed a fast exchange regime, whilst others displayed an intermediate exchange regime. Furthermore, the peaks disappearing upon addition of RNA, i.e. those in intermediate exchange, were the same as those that disappeared in experiments with the WT. So we can conclude that this mutant binds to these RNA in a similar manner and with similar affinity as the wild-type. It should be noted that the slow exchange regime observed between the WT and AUUAAA was not reproducible in these conditions.

Given that it was not possible to assign this domain even with the C238A mutation; it was decided to test other constructs to further investigate the mechanism of KH binding to RNA by NMR.

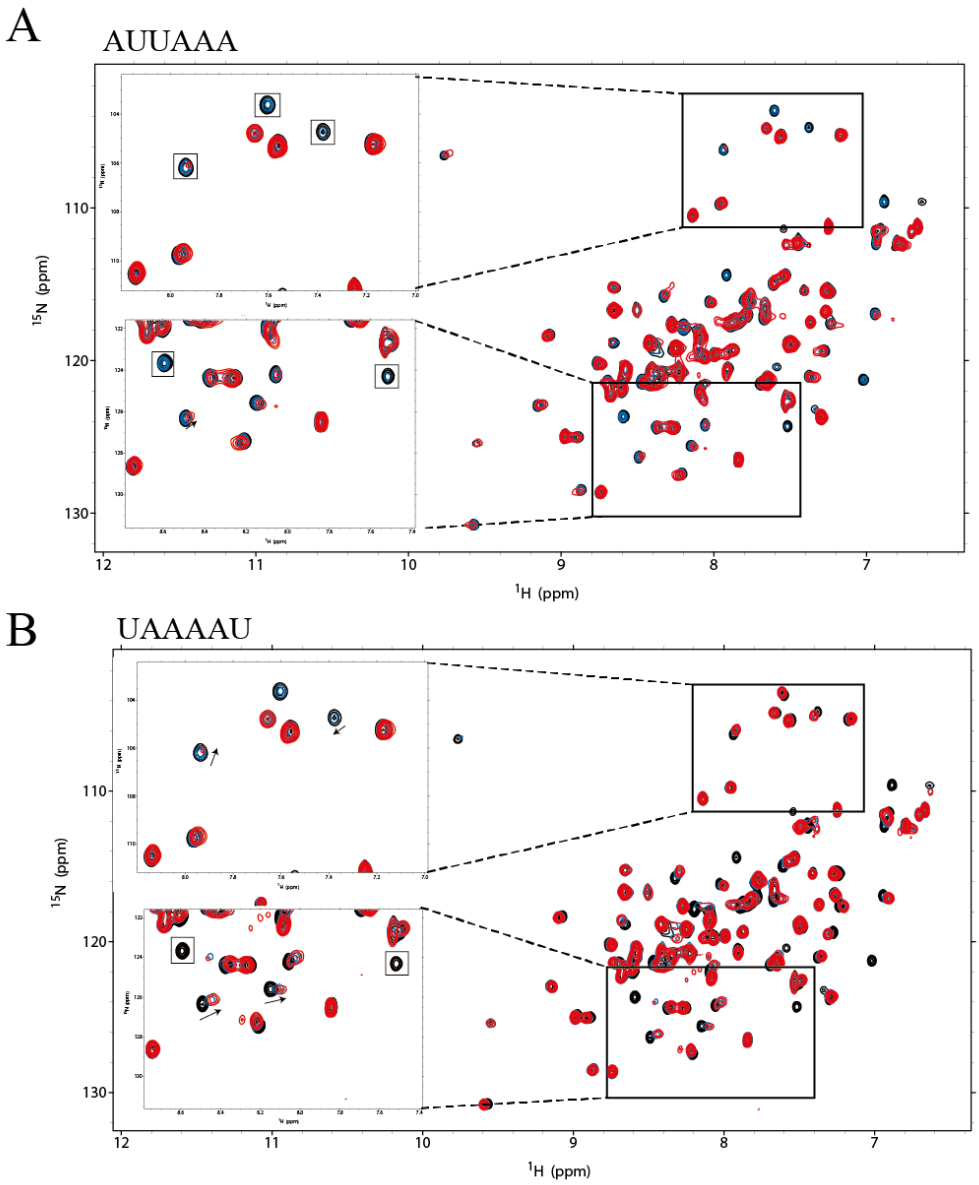


Figure 3.16: **Effect of RNA sequences on Sam68KH C238A-RNA complex formation** Chemical shift perturbation experiments of Sam68 KH C238A with AUUAAA (A) and UAAAAU (B). The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red). Both RNA bind to this domain in the fast to intermediate exchange regime.

3.3.5. Optimisation of the NKKH domain

A construct was designed to include the NK domain (amino acids 97-260) (Figure 3.1). It was anticipated that this construct would be more stable since the NK dimerisation domain structure has already been solved by NMR (Meyer et al. 2010), suggesting that this domain remains soluble at high concentrations.

The same expression protocols were applied and the protein yield was greater than the KH wild-type, cysteine mutant or KHCK domain. The final NMR buffer conditions were altered to 10mM Tris-HCl pH7, 100mM NaCl, 0.1% β-mercaptoethanol following

optimisation of an *in vitro* kinase assay using this domain (Chapter 5). NMR samples of this construct in these conditions could be concentrated to above 500µM to give good quality (^{1}H-^{15}N) HSQC spectra showing 135 out of 150 expected peaks that were well dispersed, suggesting that the domain is properly folded (Figure 3.17A). Furthermore, the NKKH domain remained stable for several weeks with no change in spectral quality at 30°C.

Interestingly, the NKKH spectrum did not overlay well with the spectrum of the KH domain (Figure 3.17B), suggesting that the KH domain within the NKKH dimer could be in a different conformation to the isolated KH domain, possibly explaining the poor stability of the isolated KH. This sample was therefore suitable for backbone assignment and for investigating RNA interactions.

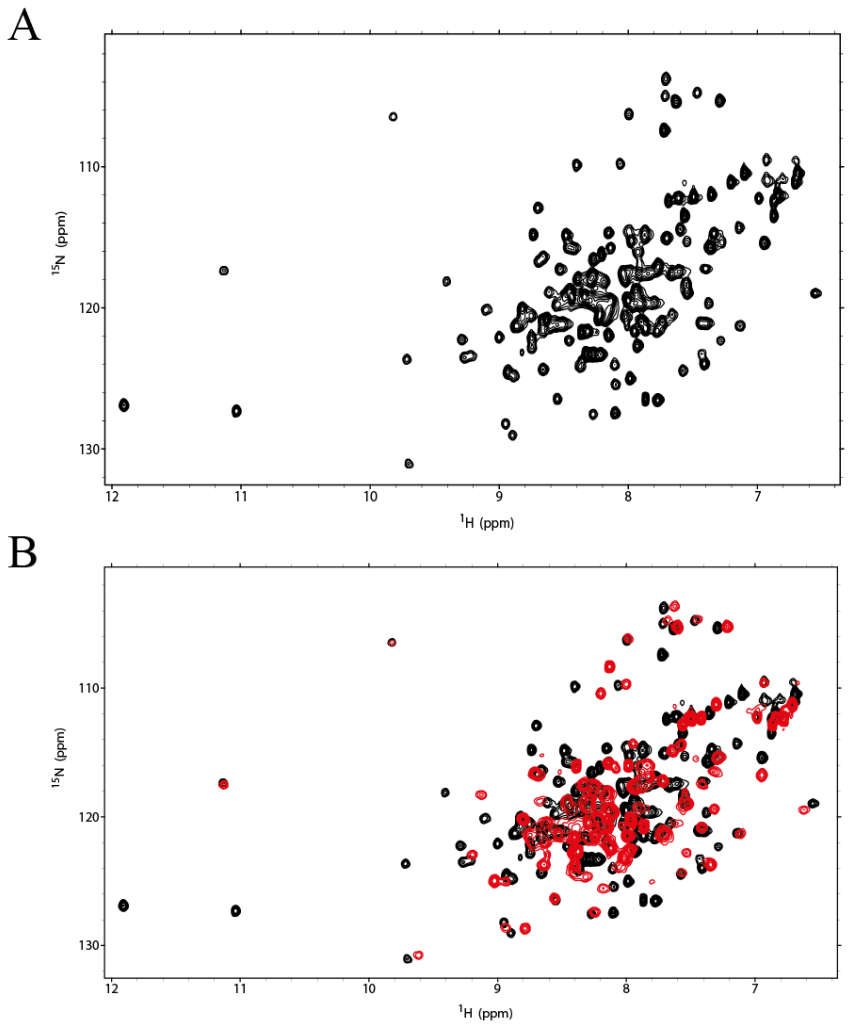


Figure 3.17: Optimisation of Sam68 NKKH
(^{1}H-15N) HSQC spectrum of Sam68 NKKH in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7 (A). This domain could be concentrated to a higher concentration, was more stable and gave a better quality (^{1}H-^{15}N) HSQC spectrum. Overlay of (^{1}H-^{15}N) HSQC spectra of the NKKH (black) and KH (red) domains show that most KH peaks experience a significant difference in chemical shift when part of the NKKH domain (B).

3.3.6. Assignment of the NKKH domain

Due to their instability at high concentrations, the KH (WT), KH (C238A) and KHCK domain constructs were not suitable for NMR assignment. The NKKH construct however, was highly stable for several days and protein concentrations greater than 500µM could be achieved without precipitation. This stability was retained with ^{13}C-^{15}N labelling and therefore good quality HNCA and HNCACB experiments were recorded.

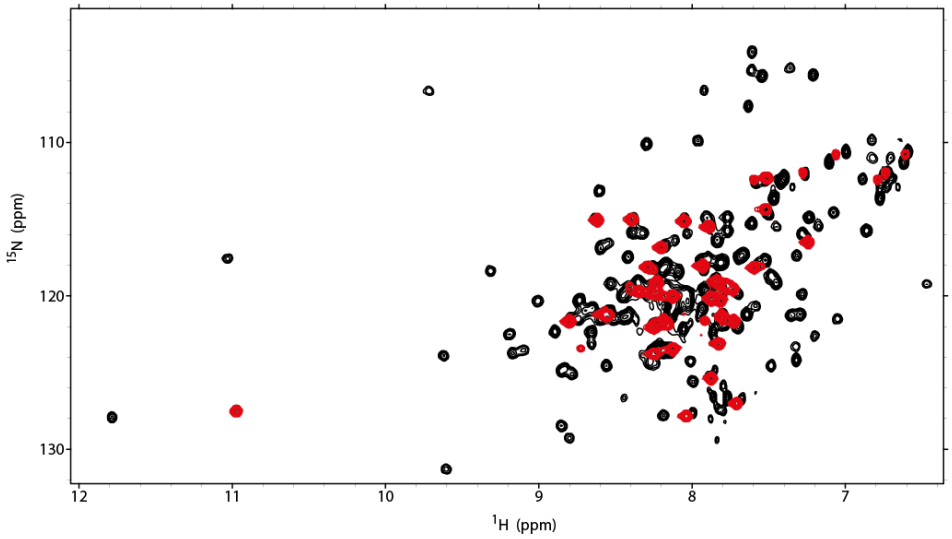


Figure 3.18: Chemical environment of the NK domain within the NKKH of Sam68
Overlay of (^{1}H-^{15}N) HSQC of the NKKH domain (black) and NK domain (red (Meyer et al. 2010b)) show that peaks of the NK domain experience the same chemical shift with addition of the KH domain, suggesting that the fold does not change.

As mentioned, the NK domain had previously been assigned (Meyer et al. 2010b). Since both ^{1}H-^{15}N HSQC spectra of the NK and our NKKH domain overlaid very well (Figure 3.18), these assignments provided a useful starting point and a means of confirmation of our own assignments as well as confirming that the isolated NK is in the same conformation as in the NKKH. In total, it was possible to assign 82% of the NKKH domain (Figures 3.19A and B and Table 7.2).

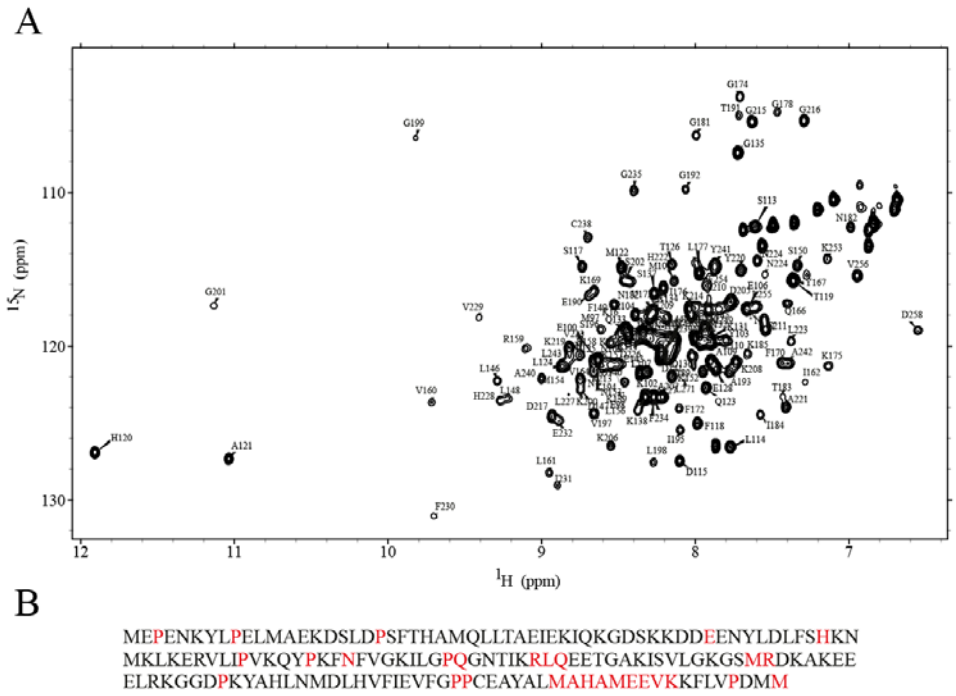


Figure 3.19: Assignment of Sam68 NKKH
(^{1}H-15N) HSQC spectrum of Sam68 NKKH with backbone chemical shift assignments (A), 82% of the NKKH domain could be assigned (red - unassigned). (B).

3.3.7. Titrations with NKKH

Inclusion of the NK domain will result in a dimer being formed in solution. It was therefore necessary to determine whether the KH domain within this structure interacts with RNA in the same manner as the KH domain alone. Analysis of a chemical shift perturbation experiment with the sequence AUUAAA, which bound with high affinity to the isolated KH WT, demonstrated that the NKKH construct also binds to this RNA (Figure 3.20A). Furthermore, the cross peaks that shift in the NKKH spectrum are the same that shift in the KH spectrum, suggesting that the KH domain binds to RNA in a similar manner in the dimeric structure. The peaks that shift are in intermediate to fast exchange, suggesting that the affinity of the interaction is similar to the KH WT and C238A mutant.

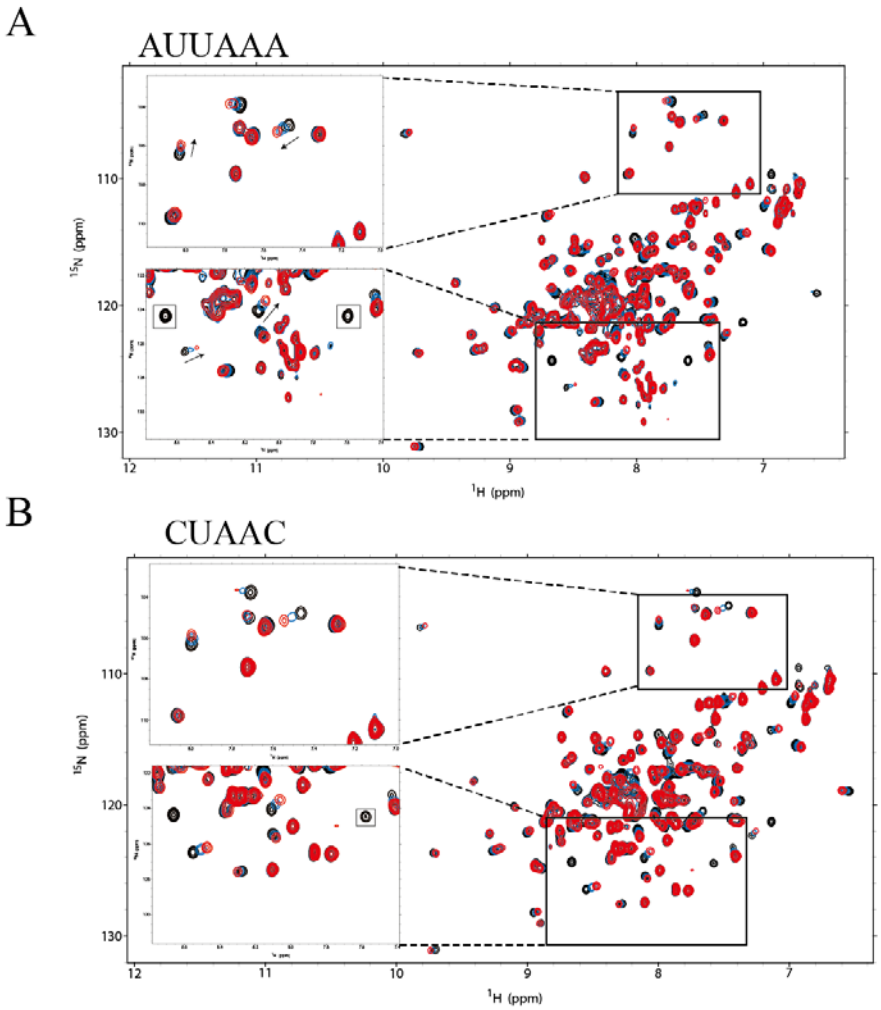


Figure 3.20: **Effect of RNA sequence and RNA length on Sam68 NKKH-RNA complex formation**
Chemical shift perturbation experiments of the NKKH domain of Sam68 with AUUAAA (A) and CUAAC (B). The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red).

Having confirmed that this construct interacts with RNA in the same manner, several new RNAs were then tested. As mentioned previously, the full STAR domain structure has been solved for QK1 and Gld-1 in complex with RNA containing the sequence CUAAC. Given that this NKKH construct is expected to form a dimer in a similar conformation to these two homologous proteins, CUAAC was tested by NMR (Figure 3.20B). Again, the affected cross-peaks either shifted a short distance or disappeared altogether, suggesting this sequence binds to the KH domain within this construct. This was also observed for the SELEX consensus sequence UAAA, although all shifting crosspeaks were in fast exchange and exhibited smaller perturbations (Figure 3.21A). These data suggest that each KH domain within the construct can bind short RNAs individually.

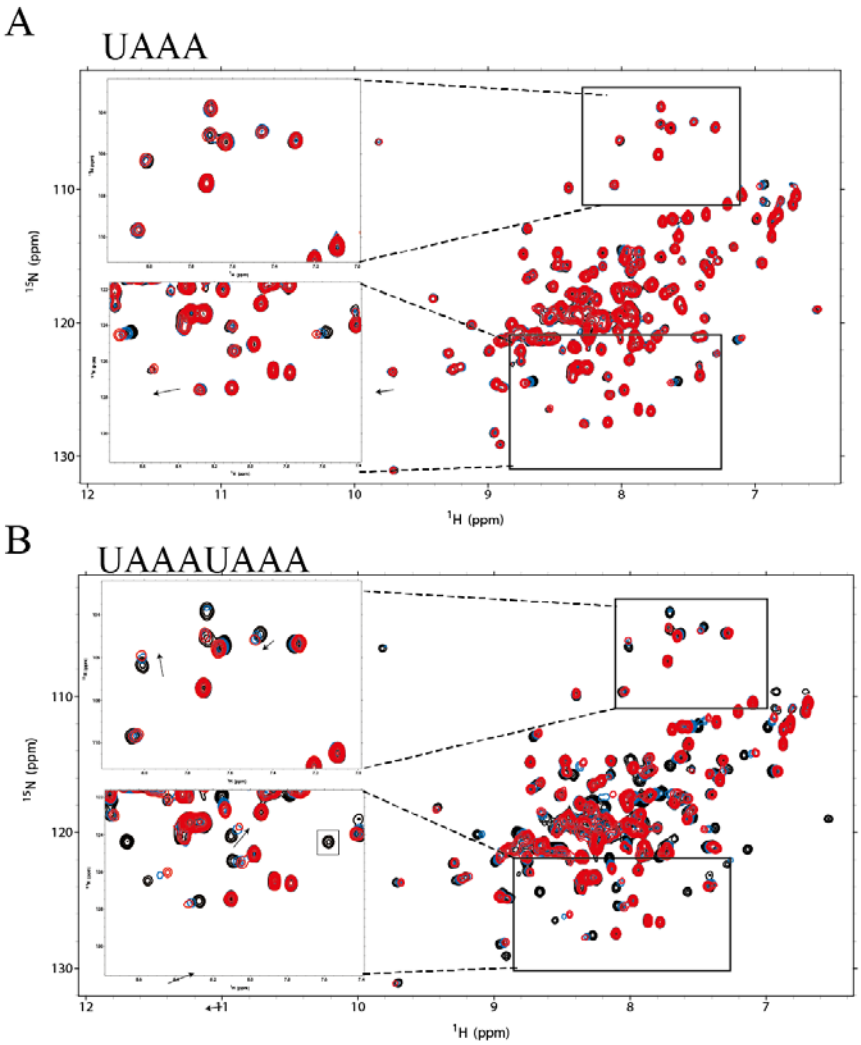


Figure 3.21: **Effect of RNA sequence and RNA length on Sam68 NKKH-RNA complex formation**
Chemical shift perturbation experiments of the NKKH domain of Sam68 with UAAA (A) and UAAAUAAA (B). The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red).

Assuming that this NKKH domain is forming a dimer, it may be possible that its affinity would be greater for longer RNAs with multiple potential binding sites. This is reflected in the SELEX data for Sam68 which shows a bipartite sequence, separated by a non-specific linker region of average length 16 nucleotides (Galarneau, Richard 2009). It is unknown whether Sam68 dimerises in the same conformation as other members of the STAR family and therefore the length of the non-specific linker region between the two binding sites is not known. Furthermore, the structures of several other STAR proteins have been solved in complex with RNA by X-ray crystallography and in each case the CK was found to make contacts with the RNA as well as the KH domain. Therefore to investigate bipartite RNA sequence recognition of this domain, longer RNAs were designed to test protein binding by NMR with the NKKH domain.

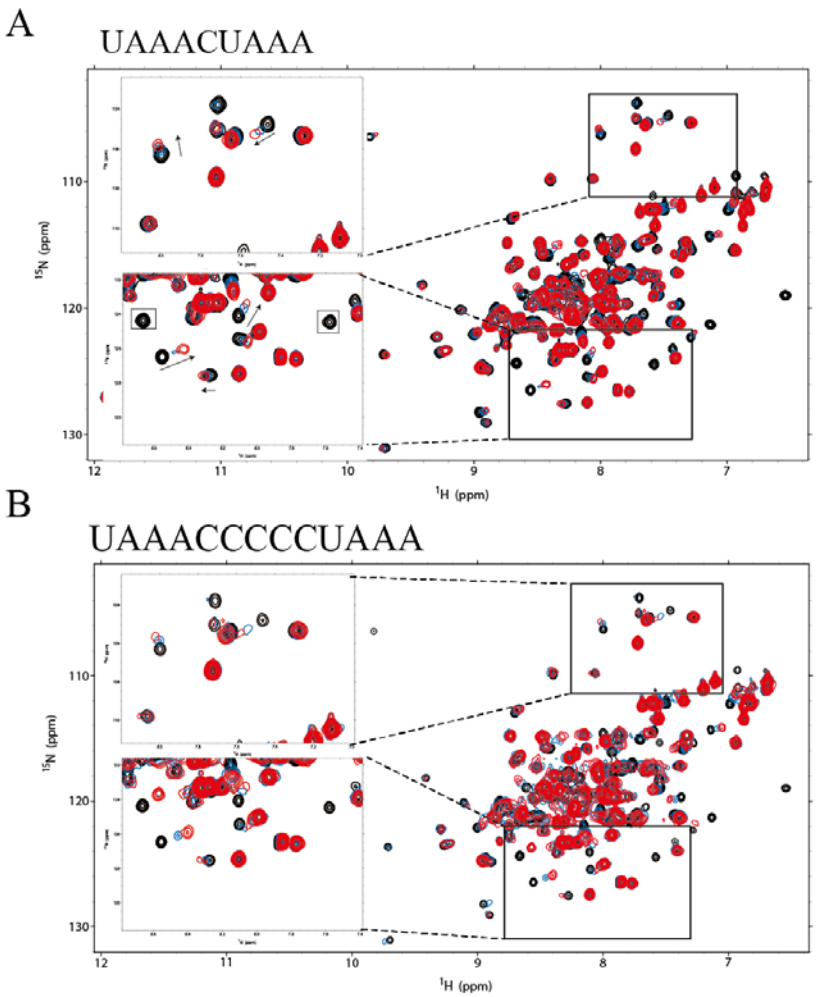


Figure 3.22: **Effect of RNA sequence and RNA length on Sam68 NKKH-RNA complex formation**
Chemical shift perturbation experiments of the NKKH domain of Sam68 with UAAACUAAA (A) and UAAACCCCCUAAA (B). The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red).

Firstly, the UAAA motif identified by SELEX was duplicated (UAAAUAAA) and titrated into the NKKH domain (Figure 3.21B). This RNA induced the disappearance of some crosspeaks and others to shift in fast exchange. This indicates that UAAAUAAA binds to Sam68 NKKH with higher affinity than UAAA, which may be due to the presence of more potential binding sites within the 8mer RNA, causing multiple register binding to the KH domain.

To simulate the non-specific linker region, a cytosine was introduced between the sequences to give UAAACUAAA (Figure 3.22A). Once more the residues affected demonstrate a mix of fast and intermediate exchange as with UAAAUAAA. Finally, this linker region was then expanded to five cytosines (UAAACCCCCUAAA) and the same peaks were seen to be in intermediate exchange (Figure 3.22B). Therefore it can

be concluded that the NKKH domain interacts with various length RNAs of similar sequences with similar binding affinity to that observed with the isolated KH domain. Since this domain was able to be assigned, it was possible to identify which residues the affected crosspeaks correspond to. This can be clearly assessed by plotting the combined chemical shift difference against the residue number (Figures 3.23 A-E). In these plots, peaks that disappear upon addition of RNA are denoted in red, and those displaying a combined chemical shift perturbation above 0.1ppm were considered to be affected by the addition of RNA. It is immediately apparent from these graphs that there are few changes to peak position upon addition of the 4mer UAAA sequence, suggesting that this sequence does not interact with the NKKH domain with high affinity. As the length of RNA increased to 6mer, with AUUAAA, more peaks are affected, particularly in the KH domain (purple shading). The combined chemical shift difference is higher; six peaks disappear altogether in the KH domain and one in the NK-KH flexible linker region. These indicate a subset of peaks that are affected by the addition of this RNA and likely correspond to the residues involved in RNA recognition. As the length of RNA is increased to 8, 9 and 13 nucleotides, more peaks disappear. These peaks only correspond to residues of the KH domain that shifted in the presence of other, shorter RNAs. More peaks in this domain experience a combined chemical shift difference above 0.15ppm, indicating an increase in the number of peaks in intermediate exchange and that those in fast exchange are moving further in space. This suggests that each KH domain within the dimer binds the longer RNA with higher affinity than that of the shorter RNA possibly due to these RNA binding in different registers.

In order to map the minimal protein interaction surface, the particular residues affected by the addition of RNA must be identified for mutational analysis of RNA binding (section 4.5.4). To determine length of the minimal RNA that interacts with this entire surface, various length RNAs must be tested for binding.

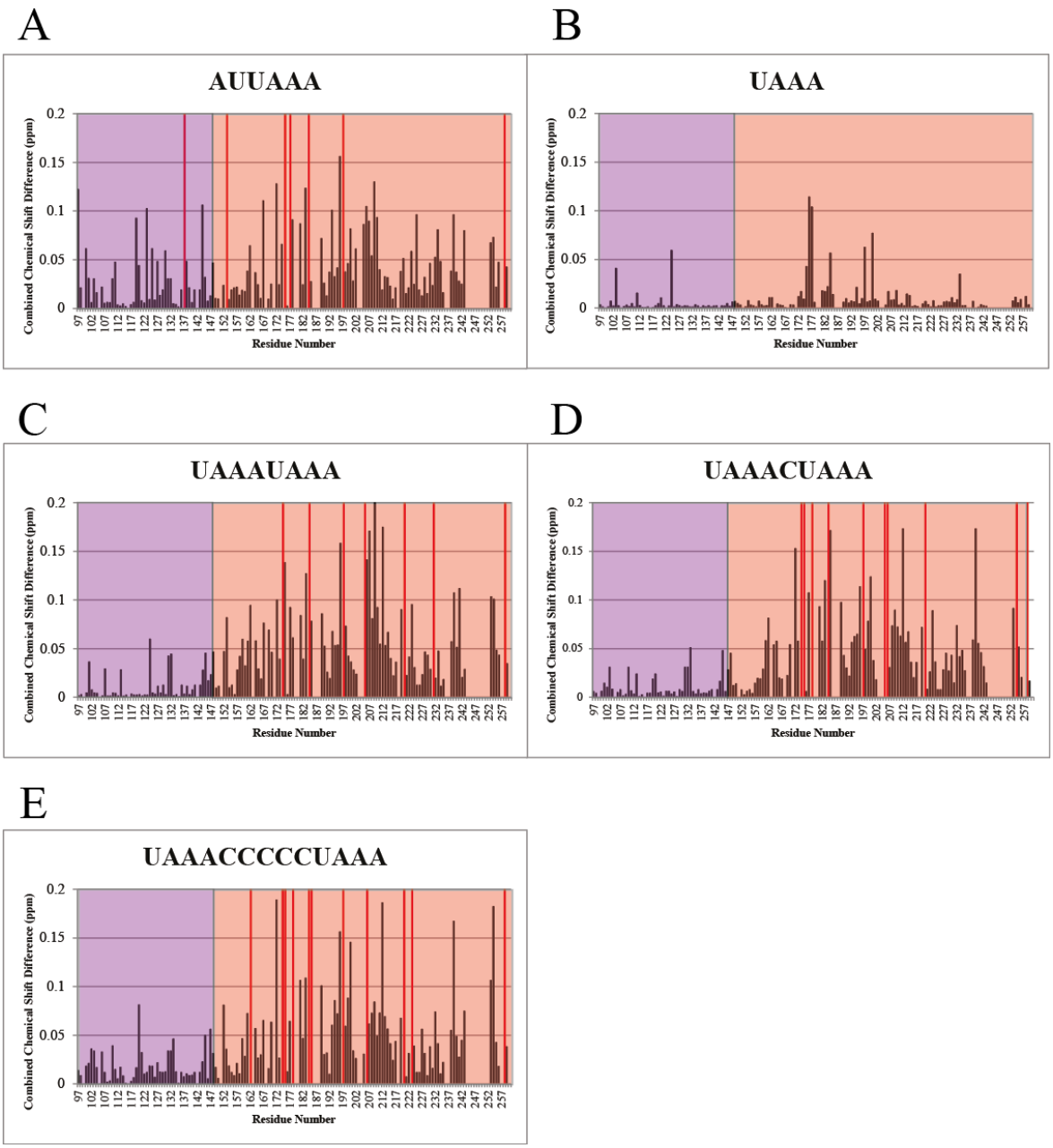


Figure 3.23: **Analysis of combined chemical shift differences**
Graphical representation of the combined chemical shift difference (ppm) against residue number indicate the peaks of the NKKH domain most affected by the addition of AUUUAA (A), UAAA (B), UAAAUAAA (C), UAAACUAAA (D) and UAAACCCCCUAAA (E). Bars in red represent residues whose peaks disappeared upon addition of RNA. The NK domain is highlighted in purple, and the KH domain in pink. The chemical shift changes in each dimension were combined using $\Delta\delta = \sqrt{(\Delta\delta H)^2 + (\frac{\Delta\delta N}{R})^2}$, where $\Delta\delta$ represents the combined chemical shift and $\Delta\delta H$ and $\Delta\delta N$ are the change in chemical shift in the proton and nitrogen dimensions, respectively. R is a scaling factor, 6.51.

3.3.8. Optimisation of the full STAR domain

The construct boundaries were then extended to include the NK and CK domains (amino acids 97-283). It was anticipated that this full STAR domain construct would be as stable as the NKKH domain, since the STAR domain is evolutionarily conserved throughout the family and the structure of both GLD-1 and QKI in complex with RNA have been solved by crystallography (Teplova et al. 2013).

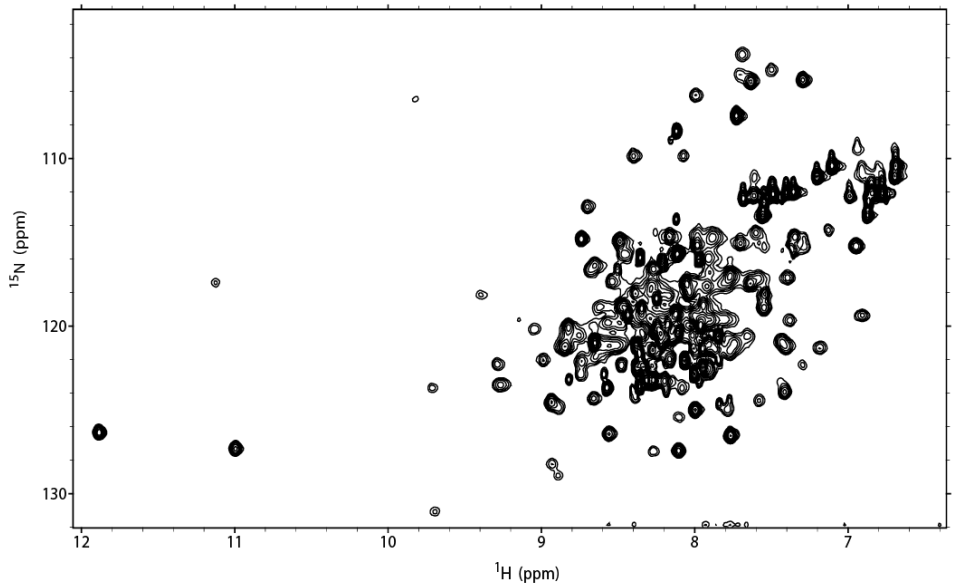


Figure 3.24: Optimisation of Sam68 STAR domain
(^{1}H-^{15}N) HSQC spectrum of Sam68 STAR in 20mM Na_2HPO_4, 100mM NaCl, 0.1% β-mercaptoethanol, pH7.

The same expression and purification protocols were used as described previously. The final yield was much greater than either the KH WT or KH C238A. In addition to an increase in expression level, this construct was remained stable at a high concentration, for several days and at a higher temperature of 30°C in 10mM Tris-HCL, 100mM NaCl and 0.1% β-mercaptoethanol (Figure 3.24). The (^{1}H-^{15}N) HSQC spectrum showed approximately 150 well dispersed peaks (of 186 amino acids) suggesting that this domain is properly folded. Interestingly, the STAR domain spectrum did not overlay well with the spectrum of the KH domain (Figure 3.25A), suggesting that the KH domain within the full STAR domain could be in a different conformation to the isolated KH domain and presumably explaining the poor stability of the isolated KH. The (^{1}H-^{15}N) HSQC spectrum overlays almost exactly with that of the NKKH domain, suggesting that the NK and KH domains have the same conformation within the NKKH domain, and that the CK domain undergoes degradation in the STAR construct (Figure 3.25B). The presence of intense peaks in the centre of the spectrum, suggests a flexible

linker either between the NK and KH domains or in the CK domain, as observed previously for QKI (Maguire et al. 2005). Despite this, chemical shift perturbation experiments with RNA were possible, and the stability of the sample allowed us to measure several triple-resonance.

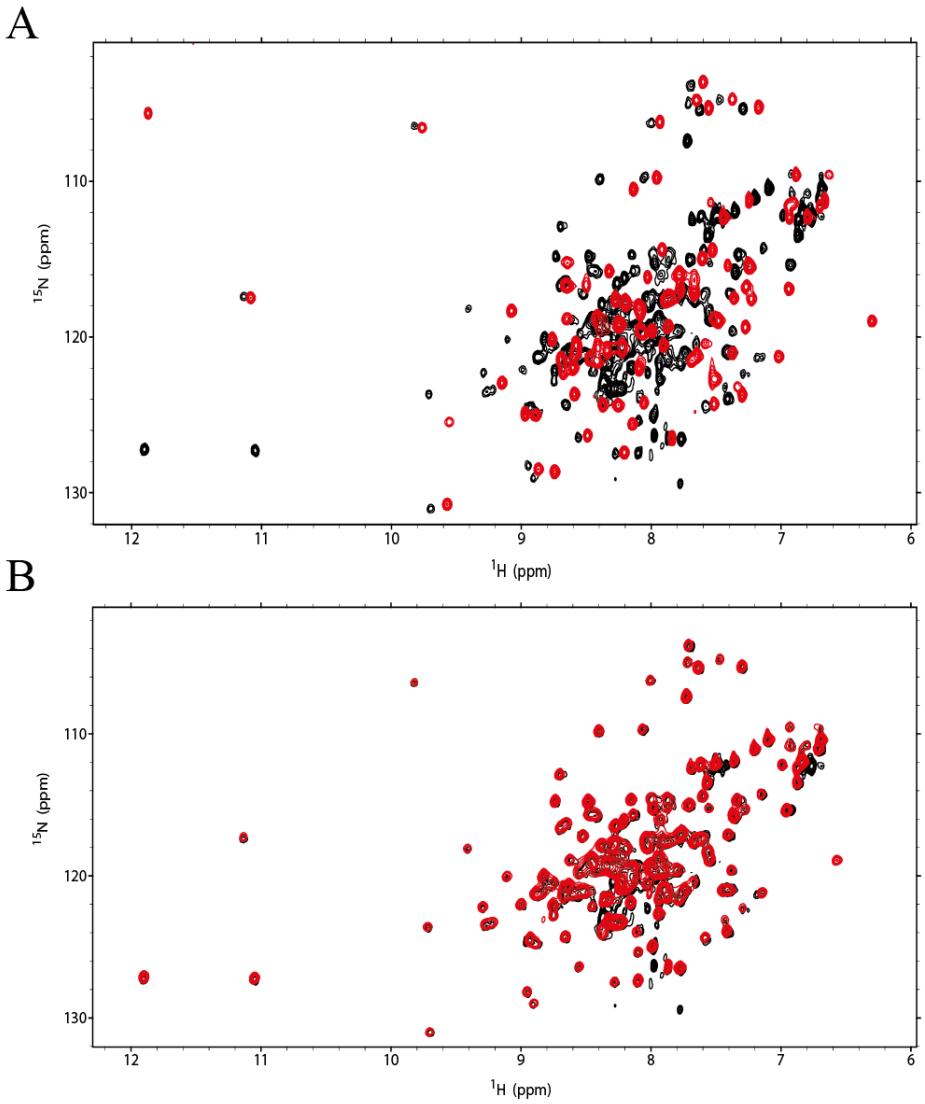


Figure 3.25: Comparison of (^{1}H-^{15}N) HSQC spectra of Sam68 STAR, NKKH and KH domains
Overlay of (^{1}H-^{15}N) HSQC spectra of the STAR (black) and KH (red) domains show that most KH peaks experience a different chemical shift as part of the STAR domain (A). (^{1}H-^{15}N) HSQC of the STAR domain (black) and NKKH domain (red) overlay almost exactly, with a few additional peak of the CK visible in the centre of the spectrum (B).

Given that the spectra of the STAR (black) and NKKH (red) domains overlay almost exactly (Figure 3.25B), with only a few additional peaks of the CK domain visible in the centre of the STAR domain spectrum, the NKKH assignments could be used to assign several peaks of the STAR domain (Figure 3.26A). It was possible to assign 77% of the full STAR domain (Figure 3.26B) and therefore this construct was suitable for

chemical shift perturbation experiments. As the KHCK domain was not stable enough for such experiments, the STAR domain construct is essential for investigation of the contribution of the CK domain in RNA binding by NMR.

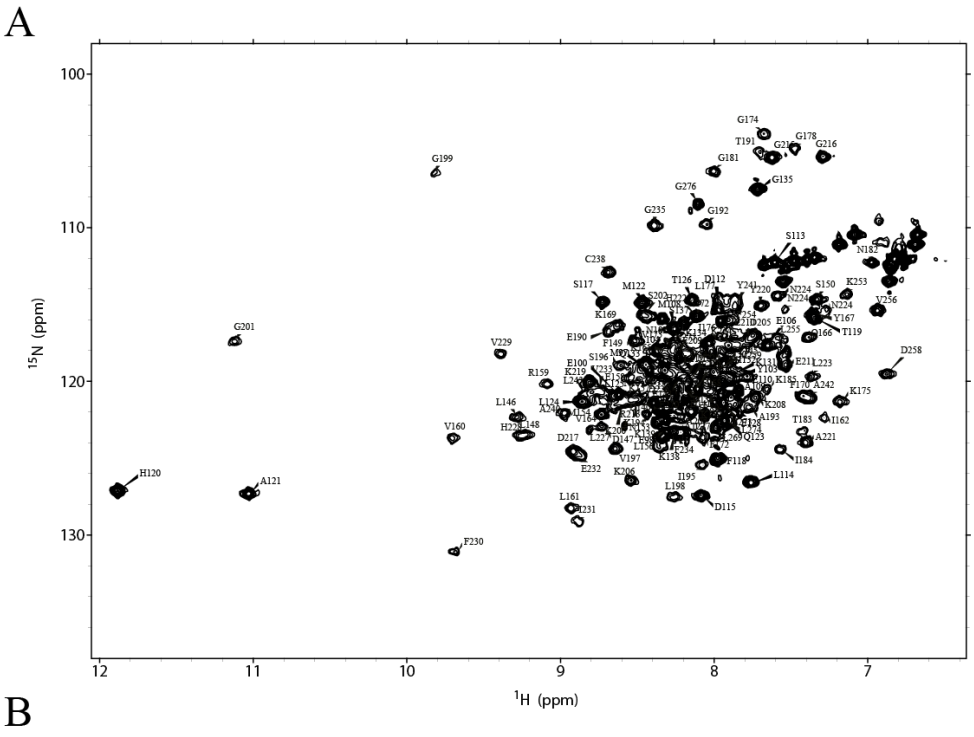


Figure 3.26: Assignment of Sam68 STAR domain
The backbone assignment of the NKKH domain could be transferred to the STAR domain (A), several additional peaks of the CK domain were assigned from triple resonance experiments of Sam68 STAR resulting in 77% amino acids of the STAR domain being assigned (B).

3.3.9. Titrations with the STAR domain

A subset of the RNAs tested with the KH domain were then used in chemical shift perturbation experiments with the STAR domain to determine whether dimerisation via the NK domain and presence of the CK domain affect RNA binding. The first RNA tested with the STAR domain was AUUAAA (Figure 3.27A). The full STAR domain binds this RNA and the exchange regime is fast to intermediate, as was observed for the KH domain. One other 6mer RNA was tested, UAAAUA, which also appeared to be in fast/intermediate exchange (Figure 3.27B), with again, the same crosspeaks undergoing chemical shift perturbations as in other constructs. These data suggest that the KH domain within the full STAR domain construct also interacts with RNA in the same way as the isolated KH domain and with similar affinity.

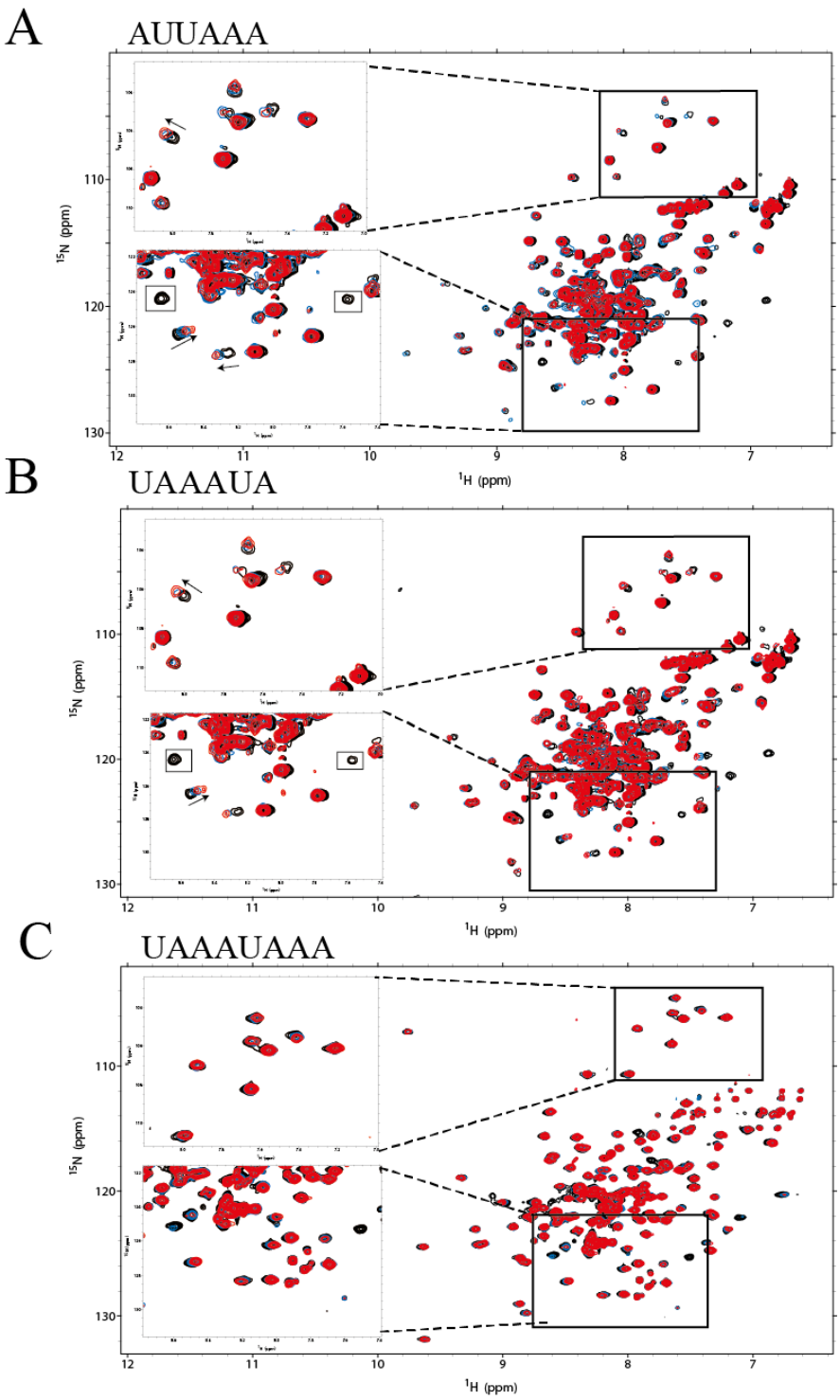


Figure 3.27: Effect of RNA sequence and RNA length on Sam68 STAR-RNA complex formation
Chemical shift perturbation experiments of the STAR domain of Sam68 with AUUAAA (A), UAAAUA (B) and UAAAUAAA (C). The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red).

Longer RNAs were then tested, to investigate the possible involvement of the CK domain in RNA binding and interaction with bipartite RNA sequences. The 8mer UAAAUAAA repeating sequence showed few changes upon addition of RNA, with

several peaks disappearing and others unaffected in the two zoomed regions of the spectra (Figure 3.27C). Several additional peaks disappeared upon addition of UAAACUAAA to the STAR domain, suggesting that it binds with higher affinity than the sequence without the central cytosine nucleotide (Figure 3.28A); however neither of these sequences bind with significantly stronger affinity than the shorter RNA.

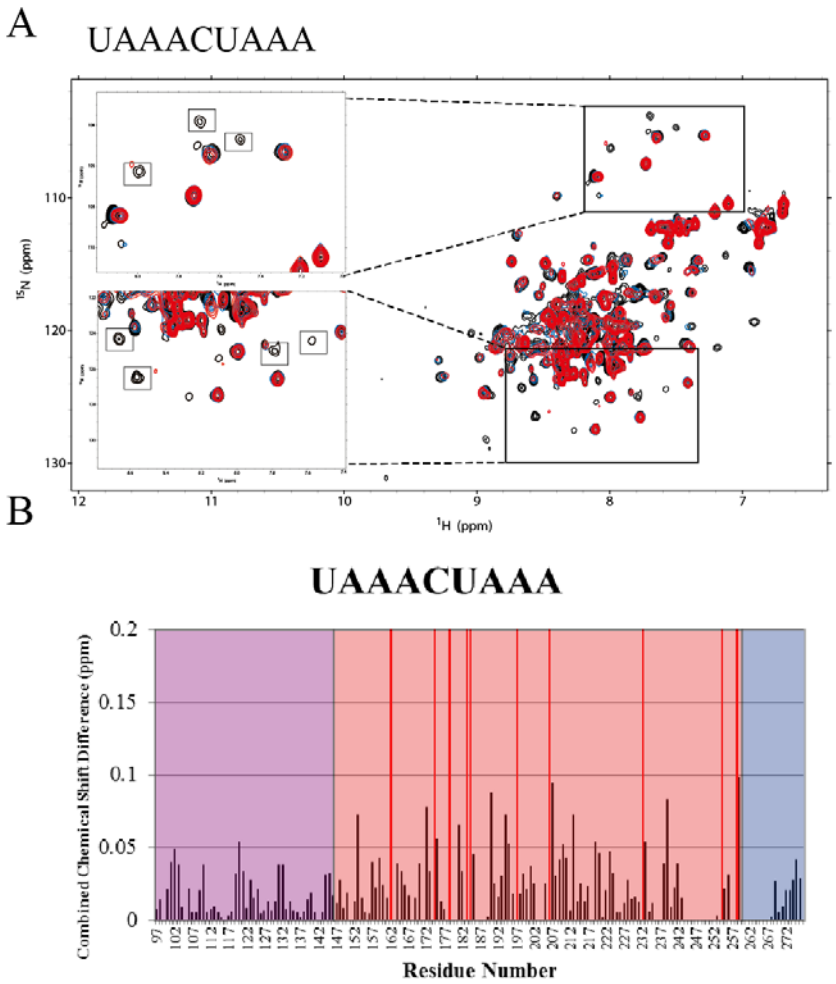


Figure 3.28: Effect of RNA sequence and RNA length on Sam68 STAR-RNA complex formation
Chemical shift perturbation experiments of the STAR domain of Sam68 with UAAACUAAA (A) The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red). Combined chemical shift difference was plotted against residue number for the STAR domain with UAAACUAAA and each subdomain highlighted, peaks that disappear are shown in red (B).

With 77% backbone assignment of the STAR domain, a plot combined of chemical shift difference against residue number could be generated to identify which residues are involved in RNA binding and whether the CK domain is involved in binding (Figure 3.28B). Similar to the NKKH domain, no peaks corresponding to the NK domain are significantly affected by the addition of these RNAs, and there are many residues in the KH domain that disappear or have a combined chemical shift difference of more than

0.15ppm. Interestingly, no peaks in the CK domain are shifting significantly or disappearing, suggesting that the CK domain is not involved in binding to these RNA (Figure 3.29).

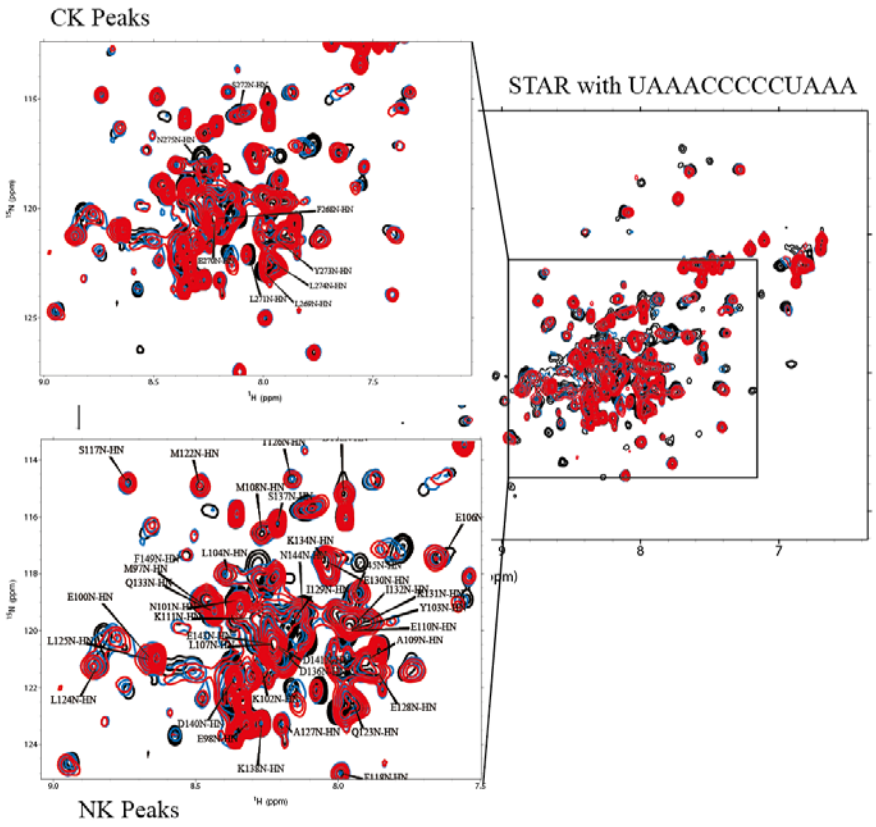


Figure 3.29: **Analysis of Sam68 STAR-RNA complex formation**
Analysis of titration of UAAACCCCCUAAA into the STAR domain with backbone assignments showed that the peaks corresponding to residues of the NK (lower panel) and CK (upper panel) domains are not affected by the addition of RNA. The overlaid spectra show free Sam68 KH (Black), 0.5:1 RNA:protein (blue) and 1:1 RNA:protein (red).

3.4. Discussion

The aim of this chapter was to determine the structure of Sam68 in complex with RNA using NMR, in order to understand how Sam68 interacts with RNA and the specificity of this recognition.

It was possible to produce samples of the KH, NKKH and STAR domains of sufficient quality for chemical shift perturbation experiments with various RNAs, and to assign 82% and 77% of the NKKH and STAR domains respectively. The optimisation process was particularly challenging for the KH domain of Sam68, which was surprising given the stability of the same construct of T-STAR, which is 77% identical to Sam68 KH domain. We postulated that the tendency for Sam68 to aggregate was due to the presence of a cysteine residue that is lacking in T-STAR. Mutation of C238 to alanine

did improve spectral quality but the sample was still not stable enough to record good quality triple resonance data required for structural determination. Recently, the KHCK domain of Gld-1 was solved by NMR in complex with a segment of the *tra-2* gene pre-mRNA. Mutations in the CK domain of this construct highlighted the need for interactions between the KH and CK domain for the correct protein function (Daubner et al. 2014). Given that Sam68 STAR is 55.4% homologous to the STAR domain of Gld-1 it is possible that the KH domain of Sam68 requires the presence of the CK domain to adopt a stable 3D conformation. The KHCK construct of Sam68 however was less stable than the KH alone and it was not possible to investigate RNA binding of this construct by NMR. This suggests that the KH domain of Sam68 may have an alternative conformation to other STAR proteins.

Inclusion of the NK domain to produce the NKKH construct had a significantly positive affect on protein expression and stability. The NK dimerisation domain has been structurally characterised amongst the STAR family and forms a helix-turn-helix motif in each of Sam68 (Meyer et al. 2010), Gld-1 and QKI (Teplova et al. 2013). Comparison of the NK spectra form the Sattler group and our own of the NKKH and STAR domains showed that the same residues are present at the same position in the spectrum, suggesting that the NK domain within the longer constructs has the same conformation as the NK alone. The improvement in stability of the samples, and quality of the (^{1}H-^{15}N) HSQC strongly suggest that the NK domain and dimerisation is required for stabilisation and correct folding of the KH domain of Sam68. The fact that the KH domain spectrum does not overlay well with the NKKH and STAR domains (Figures 3.17B and 3.25A) also supports that this domain is more stable and may have an alternative conformation within the dimer. This is significantly different to SF1 which is unique amongst the STAR family and lacks an NK domain altogether (Liu et al. 2001), although it has been shown to stabilise the overall fold of the STAR domain for Gld-1 and QKI (Teplova et al. 2013).

Chemical shift titration experiments were used to investigate binding of the KH, NKKH and STAR domain samples to RNA based on SELEX data for Sam68. Almost all RNA induced shifts in the fast to intermediate exchange for all three constructs. This indicates that the KH domain interacts with RNA with the same affinity as a single domain alone, or as part of the NKKH and STAR dimers. This is different to other STAR proteins, such as Gld-1 and QKI, whose KHCK domains have an increased affinity for RNA with the addition of the NK domain and dimerisation (Chen et al. 1997, Teplova et al. 2013).

Furthermore, Sam68 has similar affinity for a range of 6mer A-U rich RNA. Based on the exchange regime the Kd is likely to be in the low micromolar range. This is similar to the affinity of other KH domains for single stranded nucleic acids, such as hnRNPK KH3, which binds ssDNA with a Kd of 1μM (Braddock et al. 2002) and SF1 for ssDNA with a Kd of 3μM (Liu et al. 2001). To understand the specificity of Sam68 RNA binding it is necessary to estimate the binding affinity with better accuracy to distinguish the preference of Sam68 to similar AU-rich RNA and to distinguish an optimal binding partner for structure determination. This will be addressed in the following chapter.

Several other RNA sequences were tested for Sam68 binding by chemical shift perturbation experiments, including UUUUUU. Sam68 has been previously shown to bind this RNA via its RG-box motifs (Rho et al. 2007), and showed no binding to the KH domain of Sam68. This suggests that two distinct regions on Sam68 could be responsible for recognising and binding specific RNA sequences. The majority of KH domain containing proteins have multiple KH domains that all contribute to RNA binding specificity and affinity. Therefore although each individual KH domain may bind RNA with binding affinity in the low micromolar range, two KH domains can coordinate to increase affinity. STAR proteins are unique in that they have only one maxi KH domain and it has been suggested that they are able to increase binding affinity through dimerisation of the NK domain and coordination of two KH domains within this dimer. However, we have not observed a binding affinity as tight as that of multiple KH domains of other RBPs. Therefore, it may be that in addition to the two KH domains, the RG boxes at the N- and C- terminus also contribute to RNA binding and affinity. Analysis of longer constructs of Sam68 to include these regions, with longer RNA target sequences would reveal the presence of coordinated binding between these two RBDs.

The NKKH and STAR domains were tested with UAAAUAAA, UAAACUAAA and UAAACCCCCUAAA and in all cases the crosspeaks undergoing a chemical shift were in intermediate exchange, suggesting that these constructs have the same affinity for longer RNA sequences up to 13mer. This suggests that the CK domain within the STAR construct may not contribute to RNA binding and increase binding affinity of the KH domain. This is supported by analysis of the backbone assignment of the NKKH and STAR domains. In all cases, the same subset of crosspeaks were affected by the

addition of RNA and those corresponding to residues of the NK and CK domains did not undergo a chemical shift perturbation upon RNA binding.

This data has provided an insight into the contribution of each subdomain within the STAR domain to RNA binding, and that the affinity of the KH is similar amongst different length AU-rich RNA between 4 and 13mer. However, the fast to intermediate exchange regime of this complex is not ideal for further NMR studies and structure determination of the complex.

It is possible to determine the structure of a complex in intermediate exchange (Ramos et al. 2000), particularly if cross peaks reappear upon addition of excess RNA, which did not occur for any of the constructs and RNA tested so far. It is also possible to alter the exchange regime by adjusting the buffer composition, temperature and RNA target. However, due to time constraints, it was decided to attempt alternative structural and biophysical strategies to investigate the recognition of RNA by Sam68.

Chapter 4. Structural Investigation of Sam68-ssRNA complexes

4.1. Introduction

The aim of this chapter is to investigate the structure of Sam68 in complex with RNA in order to understand the mechanism and specificity of this interaction. Having produced a stable sample of the NKKH and STAR domains of Sam68 for RNA interaction studies by NMR, described in the previous chapter, we then had a well-characterised complex with which to begin further experiments.

This chapter will give an introduction to each structural and biophysical technique as they are discussed. These will outline the results obtained from crystallisation experiments, small angle x-ray scattering (SAXS) and homology modelling to determine an accurate structural model of Sam68 STAR structure free and in complex with RNA. These data, along with chemical shift perturbation data from the previous chapter will be used to determine the mechanism of RNA recognition by Sam68 using mutational analysis and fluorescence polarisation. This technique will also be used to investigate the specificity of RNA binding by Sam68.

4.2. Crystallisation experiments of Sam68

To produce good quality protein crystals, the protein molecules must aggregate in a well-ordered fashion and be able to diffract X-rays. This is dependent on many factors, such as the ionic strength, solvent concentration of the buffer or the pH of the sample. The solubility of the protein is dependent on its interaction with compounds within the solution and crystallisation occurs upon supersaturation of this solution. Therefore, in order to find the optimal conditions for a particular protein to form a crystal, many variations of precipitants, additives and temperatures must be tested, through the use of commercially available 96-well screens, such as PACT, JCSG+, Stura and Macrosol, PROPLEX, Morpheus (Molecular Dimensions). Each screen provides 96 conditions covering a range of pH, polyethylene glycol (PEGs), salt concentrations and additives, etc. There are several techniques for setting up crystallisation trials, such as the sitting drop vapour diffusion. Initially, the protein and precipitant concentrations are too low for crystallisation of the protein and the system is under saturated. As the drop and the reservoir equilibrate, the protein and precipitant concentrations increase slowly as water vaporizes from the protein droplet to the reservoir, and if the conditions are suitable, then crystal formation will occur. Often crystals of optimum quality are not produced

directly from such screens; however, they provide information regarding the type and concentration of precipitant which the protein favours. Based on these "hit" conditions, additional optimised screens can be designed around the original conditions to improve crystal growth, size and quality.

In all cases, the proteins or protein-RNA complexes at different molar ratios were dispensed into a droplet of 100nl of buffer and precipitant from a particular screen which was placed in a closed microwell next to a reservoir containing the same buffer and precipitants at higher concentrations. The crystallisation experiments undertaken for Sam68 KH, NKKH and STAR domains are summarised in Table 4.1.

4.2.1. Results

4.2.1.1. KH domain of Sam68

Crystallisation trials of the KH domain of Sam68 were undertaken during the early stages of optimisation of a suitable NMR sample. Since this construct was not soluble in 20mM Na_2HPO_4, 100mM NaCl, β-mercaptoethanol, pH7, it was not surprising that it did not readily crystallise in these conditions. One hit was obtained in the conditions shown in Table 4.1, but these did not diffract and were not reproducible.

Once the KH domain was successfully purified in a Tris based buffer, crystallisation trials were resumed using the following screens; Morpheus, Stura and Macrosol, PACT, Proplex and JCSG+ (Molecular Dimensions). The Cartesian robot was used to set up crystallisation experiments of the KH domain at 6mg/ml with and without AUUAAA, which had been identified as binding the KH domain with strong affinity (Figure 3.11B). No hits were obtained and therefore an alternative strategy was used to dispense the RNA separately from the protein, using an Oryx robot, rather than pre-mixing the two components together prior to distribution on the plate. The sample could only be concentrated to 3mg/ml and RNA was used at a 1:1 and 1:2 molar ratio of protein:RNA. A mixture of fine and granular precipitate was observed with no hits from any screen. In addition to the commercial screens, several plates were set up using conditions optimised for T-STAR KH crystallisation, and again no hits were obtained. The C238A mutant of the KH domain which was more stable than the WT as observed by NMR was then used in crystallisation trials with and without RNA using the Cartesian robot for distribution. This mutant could only be concentrated to 3mg/ml and the commercial screens Morpheus, MIDAS, Stura and Macrosol and PACT at either 4°C or 20°C did not yield any hits.

4.2.1.2. NKKH and STAR domains

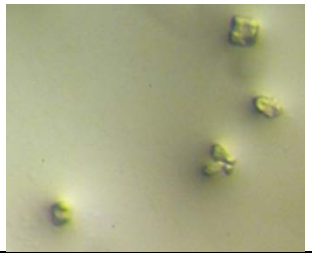
The STAR and NKKH domain samples produced for NMR studies were significantly more stable at higher concentrations than the KH domain constructs, and crystallisation experiments of these domains proved more successful. The strategy for crystallisation is outlined in Table 4.1, and describes the initial screens used to try and crystallise the NKKH and STAR domains free and in complex with various AU-rich RNA targets identified as Sam68 targets using SELEX and NMR data. These crystallisation trials were set up at a variety of protein concentrations, protein:RNA ratios, temperatures, screens and microseeds.

Microseeding proved to be a successful strategy for crystallisation of the STAR and NKKH domains. This technique involves adding a solution of crushed crystals to the drop, usually in a ratio of 3:1:2, protein:seed stock:reservoir solution, dispensed in this case using the Oryx robot. The presence of crystal seeds in a protein solution allows separation of crystal growth and nucleation, which is the initial stage of crystal formation and is usually achieved through the use of precipitants. Nucleation requires slow supersaturation of the protein sample in order to form nuclei from which crystals can grow, and this is more easily achieved when an existing protein surface, such as a crystal seed, is available since less energy is required than to form nuclei independently (Bergfors 2003). It has also been shown that cross-microseeding, using a crystal seed stock from a homologous protein, can be used to promote or improve crystal growth. Since the homologous protein, T-STAR can readily be crystallised, T-STAR STAR domain crystals were used to produce a microseed stock for use with the STAR and NKKH domains of Sam68. Both PACT and Morpheus screens were set up with 15mg/ml of Sam68 STAR or NKKH domains, both free or with AUUAAA and UAAACCCCCUAAA at a 1:1 molar ratio. The Mosquito robot was used to dispense the protein, T-STAR seed stock, RNA and reservoir solution. Plates were then incubated at 4°C or room temperature. No hits were obtained for the free protein or with UAAACCCCCUAAA at 4°C or room temperature for both the STAR and NKKH domains. However with AUUAAA several hits for each construct were obtained with the PACT and Morpheus screens at 4°C, forming needle shaped crystals growing from a single point (Table 4.1). Optimisation of crystallisation was carried out based on the hits obtained from the PACT screen because this screen generated more hits and bigger crystals than the Morpheus screen.

The best condition form the PACT screen was 0.2M NaCl, 0.1M HEPES (pH 7) and 20% PEG6000. Therefore an optimised plate (PACT Optimised 1) was designed to screen NaCl concentration from 0 to 0.4M, PEG 6000 from 16-26% and pH from 6.8 to 7.4 (Figure 7.3).

This optimised screen was dispensed with 22mg/ml Sam68 STAR and 21mg/ml NKKH in complex with AUUAAA at 1:1 molar ratio with microseed stock from the previous crystals of Sam68 STAR and NKKH crystals, respectively. The plates were kept at 4°C or room temperature and produced slightly bigger needle-shaped crystals growing out from a single point at 4°C. Only granular precipitant was observed at room temperature. The optimum conditions could be narrowed to 0.2/0.3M NaCl, 22-26% PEG6000 with constant 0.1M HEPES pH7. Therefore these conditions were optimised further by screening NaCl concentration from 0.2M to 0.5M, PEG6000 from 22 to 26% and a wider range of pH from 5 to 8 (PACT Optimised 2, Figure 7.4). Once again, the STAR and NKKH domains were dispensed at 25mg/ml and 23mg/ml, respectively, with a microseed stock from the first optimised screen and AUUAAA and UAAACUAAA at 1:1 molar ratio and incubated at 4°C only. There were no hits with the long RNA for both constructs; however with AUUAAA, larger needles or plate crystals were obtained from a single point. These STAR and NKKH crystals were cryofrozen in mother liquor containing 20% MPD and analysed at Diamond Light Source, beamline I24.

The plate crystals diffracted to 3Å and were of protein (Figure 4.1). They were however anisotropic, giving a smeary diffraction pattern in one direction and no diffraction at all in another (Figure 4.1). One data set was collected, and analysed. The space group was P2 and the solvent content was calculated to be 60% (using MATTPROB). Unit cell dimensions could be determined; a – 110.8, b – 49.4, c – 135.6, α – 90, β – 111 and γ – 90. However, due to the anisotropic nature of the crystal, the data set could not be processed further.

Construct	Sample conditions	Crystallisation Trials	Hit conditions	Outcome
KH	6mg/ml In phosphate No RNA	Non-divalent cation screen	0.8M KH_2PO_4 0.8M NaCl 100mM HEPES (NaOH) pH7 20°C 	No diffraction
	3mg/ml AUUAAA	Morpheus Stura and Macrosol PACT Proplex JCSG+	n/a	n/a
KH C238A	3mg/ml Free AUUAAA 1:1 and 2:1 (RNA:protein)	Morpheus MIDAS Stura and Macrosol PACT	n/a	n/a
STAR	12mg/ml Free AUUAAA 1:1	Morpheus	3 hits: (Ligand stock – Halogens/ethylene glycol/amino acid) 0.1M Imidazole MES pH6.5 30% PEG MME550, PEG20K 4°C Free STAR domain 	Optimised conditions (Figure 7.2) No further hits No diffraction
		Macrosol JCSG+ PACT (4°C and room temperature)	n/a	n/a
	10mg/ml Sam68 STAR microseeds from Morpheus hits	Morpheus Morpheus optimised (4°C and room temperature)	n/a	n/a

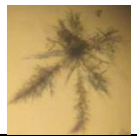
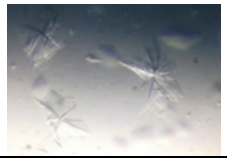
Construct	Sample conditions	Crystallisation Trials	Hit conditions	Outcome
STAR and NKKH	15mg/ml STAR and NKKH domains Free AUUAAA UAAACCCCCUAAA (1:1) T-STAR microseeds	PACT Morpheus (4°C or room temperature)	PACT common hit condition with AUUAAA: 0.1M HEPES (pH 7) 0.2M NaCl, 20% PEG6000 4°C 	Used as seed stock for further trials
	22mg/ml STAR 21mg/ml NKKH AUUAAA (1:1) Sam68 NKKH/STAR microseeds	PACT optimised 1	Several hits at 4°C, common condition: 0.1M HEPES 0.2/0.3M NaCl 22-26% PEG6000 (similar needles as above)	Used as seed stock for further trials
	25mg/ml STAR 23mg/ml NKKH AUUAAA UAAACUAAA (1:1) Sam68/NKKH microseeds from PACT Optimised 1	PACT optimised 2	Several hits for both domains at 4°C with AUUAAA Larger needles and plates of the STAR domain Common condition (plates): 0.1M HEPES (pH6) 0.2M NaCl 24% PEG6000 4°C 	Diffracted to 3Å and indicated protein. One data set collected but anisotropic and unable to fully process. Used as seed stock for further trials.
	23.5mg/ml STAR domain With (1:1) AUAAU, AAUAUU, AUUAAU, UAAAUAAA, UAAAUAAUU, UAAAUUAAU, UUUAAAUAA and UAAACCCCCUAAA	PACT optimised 2 (4°C)	n/a	n/a
	5, 10, 15, 20mg/ml STAR domain with AUUAAA (1:1) STAR plate crystal microseeds	PACT optimised 3 (4°C)	n/a	n/a
	25mg/ml STAR AUUAAA (1:1)	PACT optimised 4 (4°C)	n/a	n/a
	7mg/ml 14mg/ml NKKH (DTT)	PACT PACT optimised 1, 2, 3 and 4 JCSG+ Proplex (4°C)	n/a	n/a

Table 4.1: Sam68 crystallisation trials. Crystallisation experiments of Sam68 KH, NKKH and STAR domains with various RNA using various screens and microseed stocks.

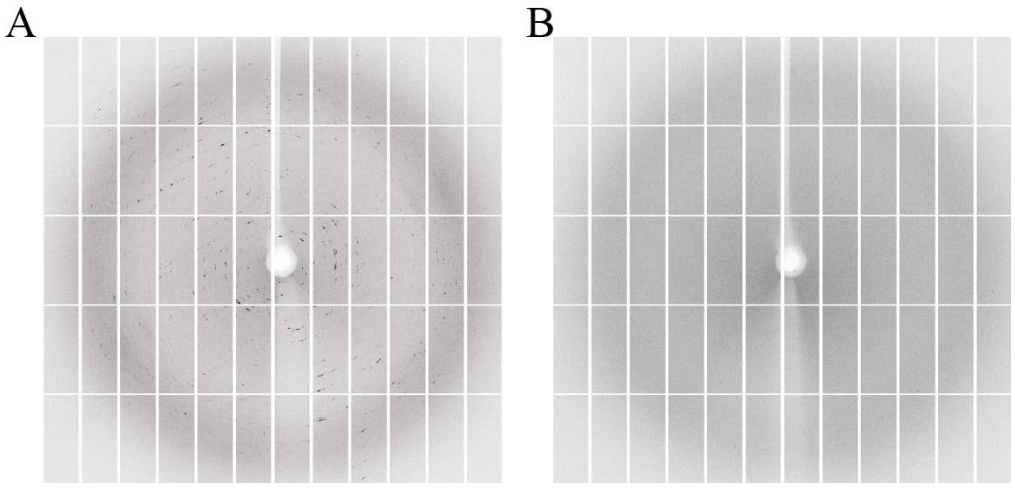


Figure 4.1: Diffraction of Sam68 STAR protein crystals
Sam68 STAR crystals diffracted in one direction to 3Å (A) but were anisotropic and no diffraction was detected in other directions (B).

4.2.1.3. Improvement of crystals

In order to improve crystal packing and obtain crystals that diffract isotropically, a range of different length RNAs were screened using the same optimised screen and the Sam68 needle crystals as a microseed stock. The STAR domain was dispensed at 23.5mg/ml with AUAAU, AAUAUU, AUUAAU, UAAAUAAA, UAAAUAAUU, UAAAUUAAU, UUUAAAUAA and UAAACCCCCUAAA and the plates incubated at 4°C.

No hits were obtained in these conditions and therefore the PEG molecular weight was varied in order to improve the crystal quality. PEG 3350, PEG4000, PEG6000 and PEG8000 were screened at 22, 24 and 26%, with NaCl from 0.1M to 0.5M and a constant buffer of 0.1M HEPES at pH6 (PACT Optimised 3, Figure 7.5). The STAR domain bound to AUUAAA crystallisation trials were performed with the protein:RNA microseed stock from the previous crystals, at four different concentrations to determine the effect of protein concentration on crystal growth as well. Once again no hits were obtained for the STAR domain at 5, 10, 15 or 20mg/ml.

In addition, glycerol was added to the second optimised screen from PACT, from 2.5% to 10% in order to improve crystal formation (PACT Optimised 4, Figure 7.6). Glycerol is a common additive that can be used to improve protein crystals by stabilising protein structure and crystal packing (Vagenende, Yap & Trout 2009). Sam68 STAR domain was dispensed across this screen at 25mg/ml with AUUAAA at a 1:1 molar ratio; however no hits were obtained in these conditions.

Finally, to investigate the potential for β-mercaptoethanol to disrupt crystal formation, dithiotreitol (DTT) was used as an alternative reducing agent throughout the purification

process of the NKKH domain. Crystallisation trials were set up with PACT, PACT optimised 1, 2, 3, PACT optimised 2 with glycerol, JCSG+ and Proplex, with the NKKH domain at 7mg/ml and 14mg/ml. No seed stock was used to determine the effect of reducing agent on crystallisation without this element, however, at 4°C, no hits were obtained.

4.3. Homology Modelling of Sam68 Based on Gld-1 STAR Structure

Having been unsuccessful in solving the structure of Sam68 STAR or NKKH domains by NMR or X-ray crystallography, when the crystal structures of Gld-1 and QKI became available in 2013, modelling techniques were then used to obtain a structural model of Sam68 based on these existing structures.

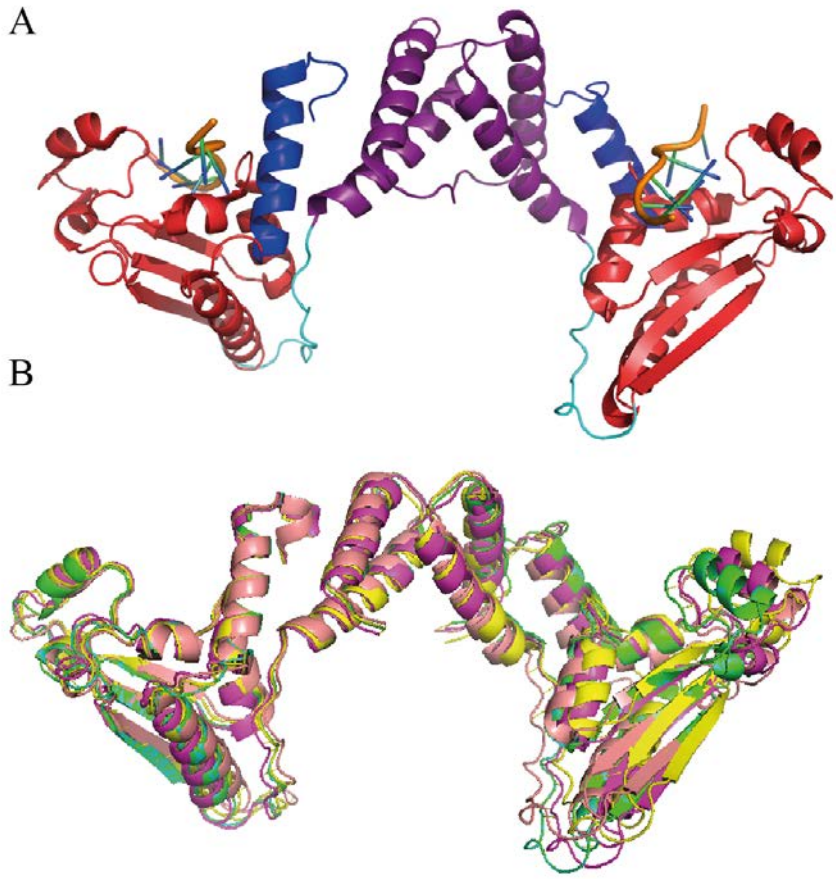


Figure 4.2: Homology modelling of Sam68 STAR based on Gld-1 STAR
The structure of the STAR domain of Gld-1 in complex with RNA, with NK (purple), NK-KH linker (cyan), KH (red) and CK (blue) (Teplova et al. 2013) (A), was used to generate five structural models of the STAR domain of Sam68 using the software modeller (Eswar et al. 2007)(B).

Sam68 STAR is 55.4% homologous (35.6% identical) to Gld-1 STAR domain. A structural model of Sam68 was therefore generated based on the structure of Gld-1

using the software Modeller (Eswar et al. 2007), which utilises sequence alignments of the protein of interest with that of one or more proteins of known structure. This method is highly reliable for proteins having more than 50% sequence identity (Eswar et al. 2007), and therefore the structure of Gld-1 STAR domain is a suitable template for estimating the structure of the STAR domain of Sam68 (Figure 4.2A). The alignment file that was generated for the STAR domains of Gld-1 and Sam68 using ClustalW and the input for Modeller can be found in Figure 7.8. Five models of the STAR domain of Sam68 were generated based on the crystal structure of Gld-1 using modeller by Dr Cyril Dominguez (Eswar et al. 2007) (Figure 4.2B). These models have the same arrangement of the three subdomains as Gld-1 and procheck was used to make further verifications in order to select the best model of the five models. The Ramachandran plot for the 5 models demonstrated 94.6% of residues in most favoured regions, 3.9% in additional allowed regions, 0.7% in generously allowed regions and 0.8% in disallowed regions. Furthermore, the wwPDB validation report for a selected model of the STAR domain gave no issues in terms of atomic clashes, peptide linkage, covalent geometry, chirality error or Phi/Psi torsion angles.

4.4. SAXS experiments for Sam68

To further investigate the structure of Sam68 STAR domain and validate the models generated based on the Gld-1 crystal structures, small angle X-ray scattering (SAXS) was used. This technique has undergone great progress in the development of instrumentation and data processing over the last decade and is now a major contender in structural investigation within the fields of physics, materials science and biology, particularly for the study of large proteins and macromolecular assemblies (as reviewed in (Blanchet, Svergun 2013).

During a SAXS experiment, as with X-ray crystallography, a protein sample is targeted with a monochromatic focused X-ray beam. In a SAXS sample, molecules are moving freely in solution and are therefore randomly orientated, unlike the regular positioning of molecules in a lattice for X-ray crystallography. This means that rather than yielding a group of diffraction peaks of specific intensities, from which a 3D electron density map can be calculated, no peaks are observed and no information regarding orientation is available. Instead, the scattered photons are radially averaged and those recorded from a buffer blank subtracted to obtain a scattering curve or pattern, which represents

the scattered intensity as a function of scattering angle (as reviewed in (Blanchet, Svergun 2013)(Jacques, Trewhella 2010)).

This scattering curve is then subjected to Guinier analysis (Guinier 1939), to determine the radius of gyration and forward scattering. SAXS data can be analysed in this way using the ATSAS program suite (Petoukhov et al. 2012) and gives an indication of the sample quality as well as the protein size, compactness, interactions, and oligomeric state. Linearity in the Guinier plot indicates good sample quality; however it is necessary to compare Guinier analysis from several samples at different concentrations to determine any concentration dependent behaviour as intermolecular interactions or aggregation can prevent accurate determination of the radius of gyration and further data analysis. In addition, an estimation of the molecular weight can be determined at this stage of analysis. A distance distribution curve is then generated using Fourier transformation of the scattering curve. This provides information on distances within the molecule, allowing more accurate calculation of the radius of gyration and indication of any aggregation within the sample and the suitability of the data for further manipulation. It is this distance distribution curve that can be used for construction of an envelope representing the space occupied by the protein molecules in solution using various programs within the ATSAS suite. Generally several different envelopes are calculated and averaged for each sample concentration.

A stable, soluble, homogeneous and contaminant free sample is required. The X-ray source available dictates the concentration of protein required and data collection times can vary from several seconds to several hours.

4.4.1. Results

Sam68 KH (C238A), NKKH and STAR domain constructs were expressed and purified at the University of Leicester, then transported at -20°C to TUM (Munich, Germany).

SAXS measurements and data processing were performed by Dr Ralf Stehle in collaboration with the laboratory of Professor Michael Sattler. Samples were tested by Circular Dichroism before and after freezing to determine whether freezing would affect the structure of the proteins. Figure 4.3 shows that each construct can be stored at -20°C and thawed with no overall effect on the protein quality. Sample quality of the NKKH domain was also assessed by size exclusion chromatography-multi-angle light scattering (SEC-MALS). As the name suggests, SEC-MALS first involves separation of the protein sample according to size using a Superdex 75 column. The flowthrough from

the column is then subjected to multi angle light scattering, during which laser beams are directed at the sample flow from different angles. Detectors measure the light scattering intensity as a function of the scattering angle, which is used to calculate the molecular weight of the protein in solution (Sahin, Roberts 2012). The molecular weight of the NKKH domain was calculated to be 38.57kDa (Figure 4.4), consistent with a dimer of theoretical molecular weight 37.5kDa. In addition, a portion of the sample appears to be tetrameric. These data suggest that at least two KH-RNA interaction surfaces are present within the dimer and that determination of the structure using SAXS and homology modelling will be useful in understanding the contribution of both to RNA recognition.

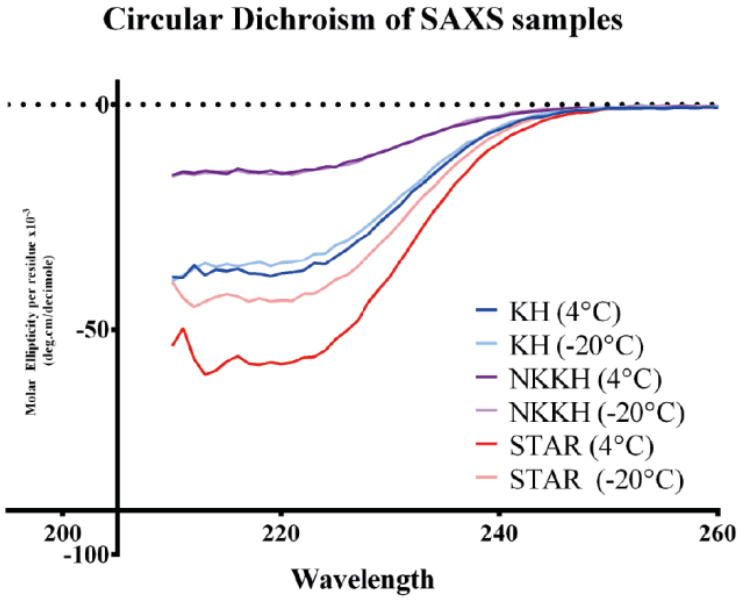


Figure 4.3: Preparation of samples for SAXS
CD was used to determine the effect of freezing Sam68 KH domain (light and dark blue), NKKH domain (light and dark purple) and STAR domain (light and dark red). The storage temperature of each sample is indicated in parentheses and all CD experiments were conducted at room temperature.

Each sample was transported at 3-5mg/ml and later concentrated to different concentrations for measurement so that any concentration dependent effects on the scattering (so called structure factor) could be determined. For each construct, a buffer measurement was recorded before and after the protein samples and these were merged and subtracted from the protein sample measurements.

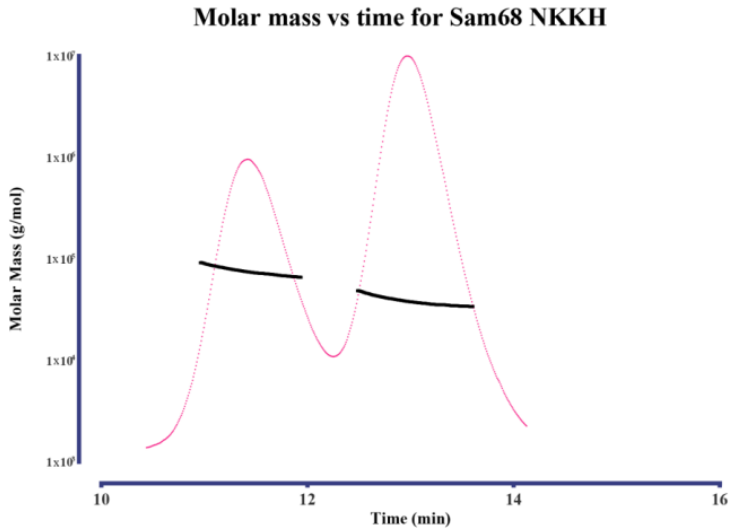


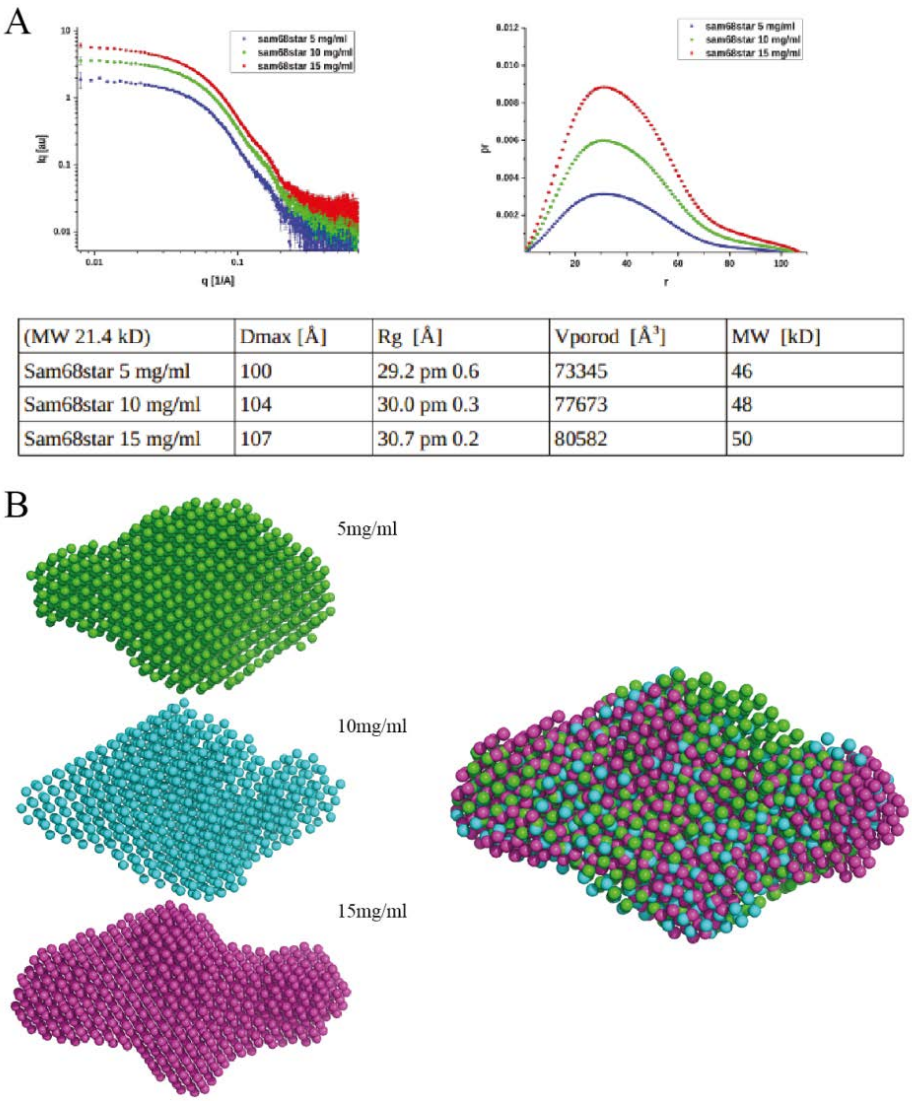
Figure 4.4: SEC-MALS analysis of Sam68 NKKH
Sam68 NKKH was analysed by SEC-MALS at 5mg/ml. The light scattering is plotted against time (pink), showing two eluted peaks representing the tetramer and dimer populations, respectively. Molar mass is plotted against time of elution (black) to calculate the molecular weight of each population, 78kDa and 39kDa, respectively.

4.4.2. Sam68 KH C238A

The tendency of the WT KH domain to aggregate and precipitate (section 3.3.1), makes it unsuitable for SAXS measurements. However, the C238A mutant is more stable and remains in solution at high concentrations for several days (section 3.3.3). This construct was therefore used at 5, 10 and 13mg/ml. Measurements were recorded for 6 hours and throughout this time the sample remained clear and showed no optically visible precipitation. The SAXS curves also did not change, having the same shape independent of protein concentration. However, the Guinier plot shows a constant curvature which is typical for aggregated particles and therefore it was not possible to calculate an Rg and no further data treatment was possible.

4.4.3. Sam68 STAR domain

The STAR domain of Sam68 was tested at 5, 10 and 15mg/ml and showed a concentration independent scattering behaviour. The molecular weight was calculated to be 48kDa, compared to the theoretical 42kDa of the dimer, suggesting that there was some aggregation. This was also seen in the slight increase in Dmax with the highest concentration, although there was no indication of aggregation at the lower concentrations from the Guinier plot (Figure 4.5). The data were of sufficient quality to further process using the ATSAS suite, generating a compact, globular envelope with some extensions in the 5 and 10mg/ml samples that are likely due to alternative conformations and flexibility of the construct rather than agglomeration (Figure 4.5).

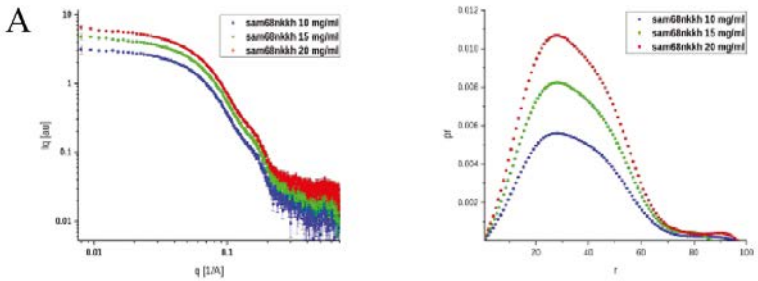


(MW 21.4 kD)	Dmax [Å]	Rg [Å]	Vporod [Å^3]	MW [kD]
Sam68star 5 mg/ml	100	29.2 pm 0.6	73345	46
Sam68star 10 mg/ml	104	30.0 pm 0.3	77673	48
Sam68star 15 mg/ml	107	30.7 pm 0.2	80582	50

Figure 4.5: SAXS of Sam68 STAR
The scattering curves and distance distributions were analysed for three sample of the STAR domain at concentrations of 5mg/ml, 10mg/ml and 15mg/ml. These gave an average molecular weight of 48kDa, confirming that the STAR domain is a dimer (A). The data were processed using the ATSAS suite to generate envelopes at each concentration, shown separately (left) and aligned (right) (B).

4.4.4. Sam68 NKKH domain

The NKKH sample is the most stable construct of Sam68, which is reflected in the CD data recorded before and after freezing (Figure 4.6) and our NMR analysis (section 3.3.5). Accordingly, the scattering curves did not show any concentration dependence and despite a hint of aggregation at the highest concentration, the molecular weight was calculated to be approximately 39kDa which corresponds to the theoretical molecular weight of this dimer (Figure 4.6). The envelopes calculated are similar to that of the STAR domain.



(MW 18.8 kD)	Dmax [Å]	Rg [Å]	Vporod [Å³]	MW [kD]
Sam68nkkh 10 mg/ml	95.75	27.357 pm 0.4	61937	38.7
Sam68nkkh 15 mg/ml	86.28	26.963 pm 1.4	62190	38.9
Sam68nkkh 20 mg/ml	96.53	27.579 pm 0.9	61847	38.7

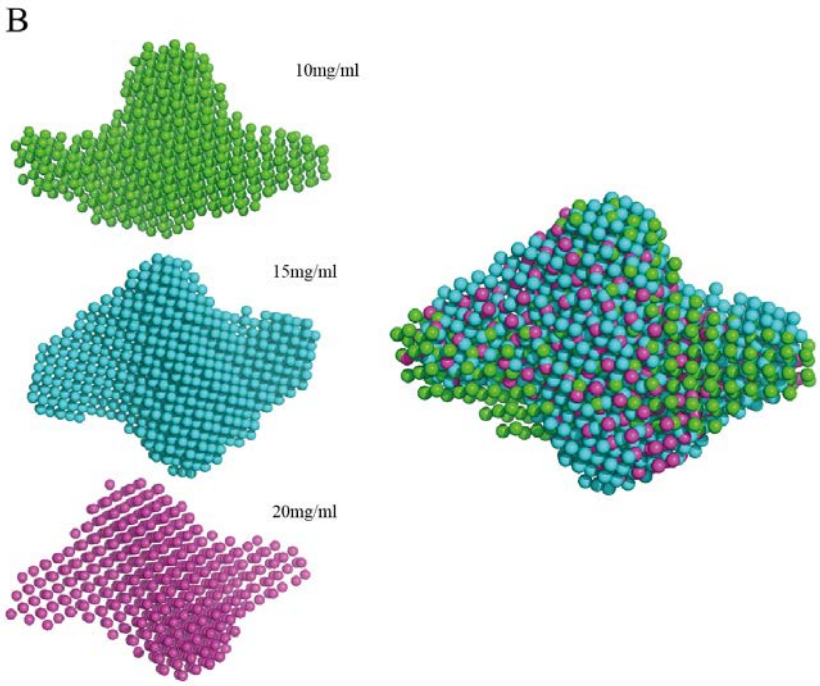


Figure 4.6: SAXS of Sam68 NKKH
The scattering curves and distance distributions were analysed for three sample of the STAR domain at concentrations of 10mg/ml, 15mg/ml and 20mg/ml. These gave an average molecular weight of 38.8kDa, confirming that the NKKH domain is a dimer (A). The data were processed using the ATSAS suite to generate envelopes at each concentration, shown separately (left) and aligned (right) (B).

4.4.5. Fitting the Sam68 STAR model into SAXS data

We compared the model of Sam68 STAR based on the Gld-1 STAR structure with the SAXS envelopes of the STAR domain of Sam68. As illustrated in (Figure 4.7), the structural model does not agree with the SAXS envelope. Importantly, the SAXS envelope suggests a compact conformation of the dimer, which is different to the model. This is supported by the high χ^2 value of 5.4, between the back calculated curve of the structural model with the real scattering curve of the STAR domain, calculated using CRYSOL. This suggests that Sam68 STAR has a different conformation than Gld-1.

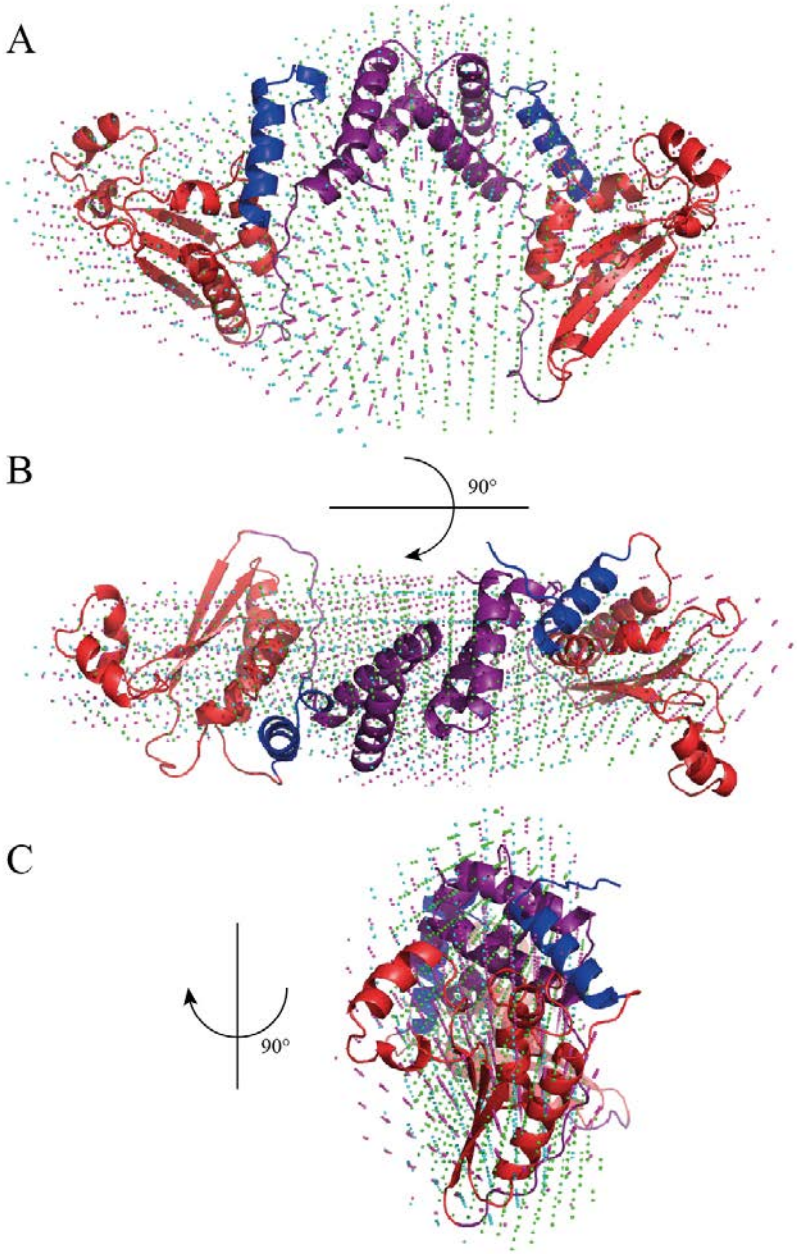


Figure 4.7: Fitting Sam68 STAR structural model within the STAR SAXS envelope
The structural model of Sam68 STAR based on the Gld-1 structure, with NK (purple), KH (red) and CK (blue) was fitted manually into the three SAXS envelopes calculated for this domain and is shown in various orientations (A, B and C).

The structure of Gld-1 STAR dimer reveals that contacts between the NK and the CK domains are necessary for the stability of the structure. Indeed, there are no direct contacts between the KH and the NK domains. The CK domain is sandwiched between them and has direct contact with both domains, stabilising the overall fold of the dimer. If that were true for the STAR domain of Sam68, then differences should be visible in the (^{1}H-^{15}N) HSQC spectra of the NKKH and STAR domains. In particular, one would expect to see a chemical shift perturbation of peaks within the NK domain that are

interacting with the CK domain. However, the spectra of Sam68 STAR and NKKH overlay perfectly, with only a few additional peaks in the central region of the STAR spectrum that likely correspond to peaks of the CK domain (Figure 3.25). This suggests that the presence of the CK domain does not affect the chemical environment of the NK and KH peaks and that the contacts observed in the Gld-1 structure are not present in the STAR domain of Sam68.

A further characteristic of the STAR domain of Gld-1 is the presence of contacts between helix 3 of the KH domain and the linker connecting the NK and KH domains (Figure 4.8A). Sequence alignment of this linker region shows that this region is not conserved in Sam68 and the residues of the Gld-1 linker that form hydrogen bonds with the KH domain helix are highlighted in red. These are mostly positive or uncharged amino acids, compared to the negatively charged substitutions in Sam68 and therefore it is unlikely that the same contacts can form in the Sam68 STAR structure (Figure 4.8B).

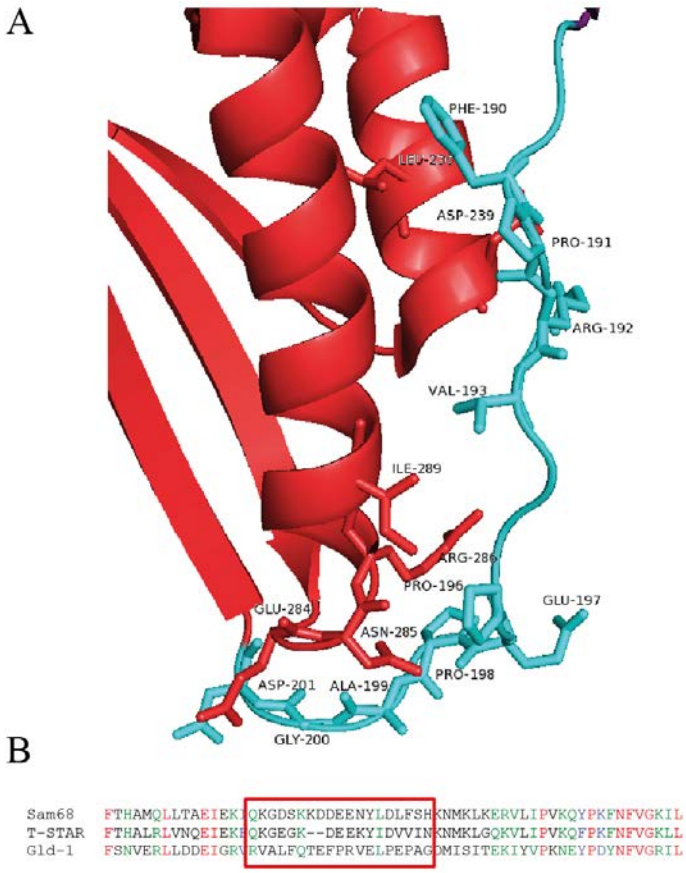


Figure 4.8: Lack of conservation of NK-KH linker between Sam68 and Gld-1
The residues of the NKKH linker (cyan) that contact the KH helix (red) are highlighted in the Gld-1 STAR domain structure (A). Sequence alignment of the linker region is shown between Sam68, T-STAR and Gld-1, showing a lack of residue conservation in this region.

4.4.6. Improved structural model of Sam68 NKKH based on SAXS data and T-STAR structure

Following collection of SAXS data, the crystal structures of T-STAR KH free and in complex with AAAUAA; KHCK in complex with AAUAAU; NKKH in complex with UAAU and the full STAR domain in complex with AUUAAA were solved within the laboratory by Dr Mikael Feracci (unpublished).

These structures revealed that the NK and KH domains form a compact dimer, with the NK domain adopting a helix-turn-helix motif forming the homodimerisation interface very similar to that of the isolated Sam68 NK structure (Meyer et al. 2010). The KH domain has a classical type I KH conformation, similar to those of other STAR proteins (Teplova et al. 2013) (Liu et al. 2001) and was shown to bind a single RNA molecule. In addition, a unique dimerisation interface was revealed between helix 3 of each KH domain in the dimer and the C-terminal of α-helix 3 of the KH domain and the C-terminal half of the NK-KH linker. This interface was found to be almost twice as large as the NK dimerisation interface of the NK domains and to be stabilised by a series of hydrophobic interactions in addition to an intermolecular hydrogen bond between Y141 in α-helix 3 and Q58 in β-strand 1. This linker region between the two subdomains was found to be flexible towards the N-terminus (residues 35-42), due to the lack of electron density observed, and more structured in the C-terminal half (residues 43-53). This more structured section of the linker was found to form a network of intermolecular hydrogen bonds with the KH domain of the other monomer. Interestingly, no electron density was observed for the CK domain, suggesting that this subdomain does not adopt a fixed conformation. Furthermore, no global structural changes to the KH domain were observed with addition of RNA, the NK or the CK domains.

This gives a strikingly different arrangement of the three subdomains compared with the crystal structures of Gld-1 and QKI. Of particular interest, in addition to the dimerisation of the KH domains within the STAR homodimer is the lack of contact between the NK-KH and KH-CK domains that is necessary for the stable fold of Gld-1 and QKI. This suggests that the Sam68 subfamily may have a distinct structure as well as distinct RNA sequence recognition to other STAR proteins.

The KH domains of Sam68 and T-STAR are 77% identical, and the majority of residues located within the dimerisation interface of T-STAR are conserved in Sam68, but not in QKI and Gld-1 (Figure 4.9C). This makes T-STAR a good candidate from which to

generate a structural model of Sam68, despite the obvious difference in crystallisation potential of these two proteins.

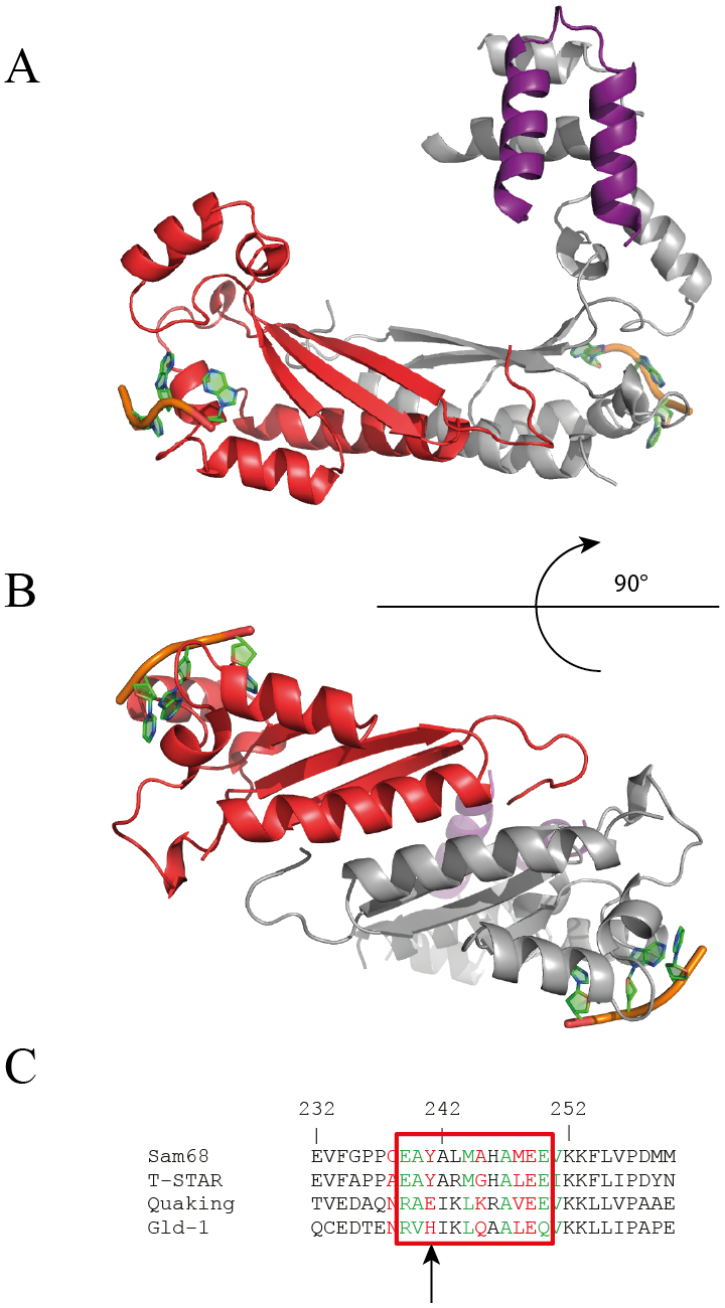


Figure 4.9: Crystal structure of T-STAR NKKH
Crystal structure of the NKKH domain of T-STAR with RNA (orange) shows one monomer in grey and the other in red (KH) and purple (NK), the electron density of the NK-KH linker was not visible (A). The KH domains of T-STAR dimerise via helix 3 (red and grey) (B). Multiple sequence alignment of the KH dimerisation interface between Sam68, T-STAR, QKI and Gld-1, highlights Y241 as a potential key residue for dimerisation of T-STAR and Sam68 only (C).

The properties of the NK-KH linker are another conserved feature between T-STAR and Sam68 that differs from other STAR proteins. This region is predicted to be inherently disordered in Sam68 (Figure 3.3) and the flexibility of the N-terminal portion compared to the C-terminal portion of the linker was confirmed using heteronuclear-NOE experiments. These provide an insight into the motion of the protein backbone NH bonds. A value greater than 0.7 indicates that the backbone of the residue is in a structured or rigid region, between 0.5 and 0.7 indicates a more flexible loop region and less than 0.5 represents a very flexible region of the protein (Kay, Torchia & Bax 1989).

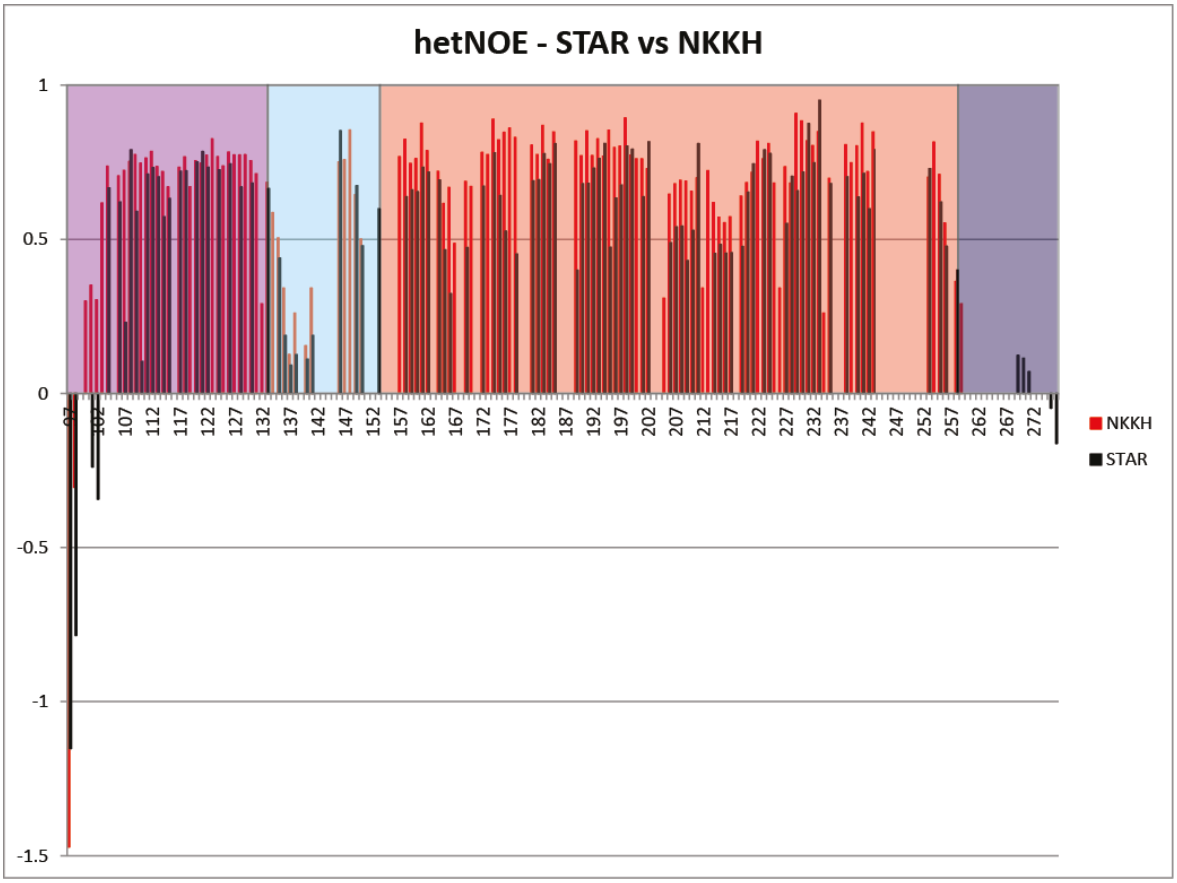


Figure 4.10: Heteronuclear NOE of Sam68 STAR and NKKH
Heteronuclear NOE data of the STAR (black) and NKKH (red) domains of Sam68, with NK shaded purple, the NK-KH linker in cyan, the KH in orange and the CK in blue. A value below 0.5 indicates that the residue is flexible.

Figure 4.10 shows the Heteronuclear NOE values for the NKKH and STAR domains of Sam68, demonstrating that the linker region is flexible in the same way as T-STAR.

The homology model of the NKKH domain of Sam68 based on T-STAR fits well within the SAXS envelope calculated from data acquired from Sam68 NKKH at 15mg/ml (Figure 4.11A), giving a χ^2 of 5.6 using CRYSOL. The extension of the envelope that is not occupied by the model is likely due to protein aggregation, or protein dynamics arising from flexibility between the NK and KH domains. Envelopes

were also calculated based on SAXS data from the NKKH domain with UAAAUAAA, these envelopes do not have this extension, fitting well with the NKKH structural model (Figure 4.11B), giving a χ^2 of 6.1 using CRYSOL. This more compact envelope is possibly a result of a decrease in protein flexibility upon RNA binding or improvement in protein stability as a result of RNA binding or decrease in sample concentration, reducing the aggregation of the sample.

This is still not optimal and closer inspection of the new Sam68 NKKH model within the envelope suggested that the KH dimer itself fits well but the orientation of the NK and KH domains do not. Given the flexibility of the NK-KH linker, the orientation of the KH domain relative to the NK domain can vary.

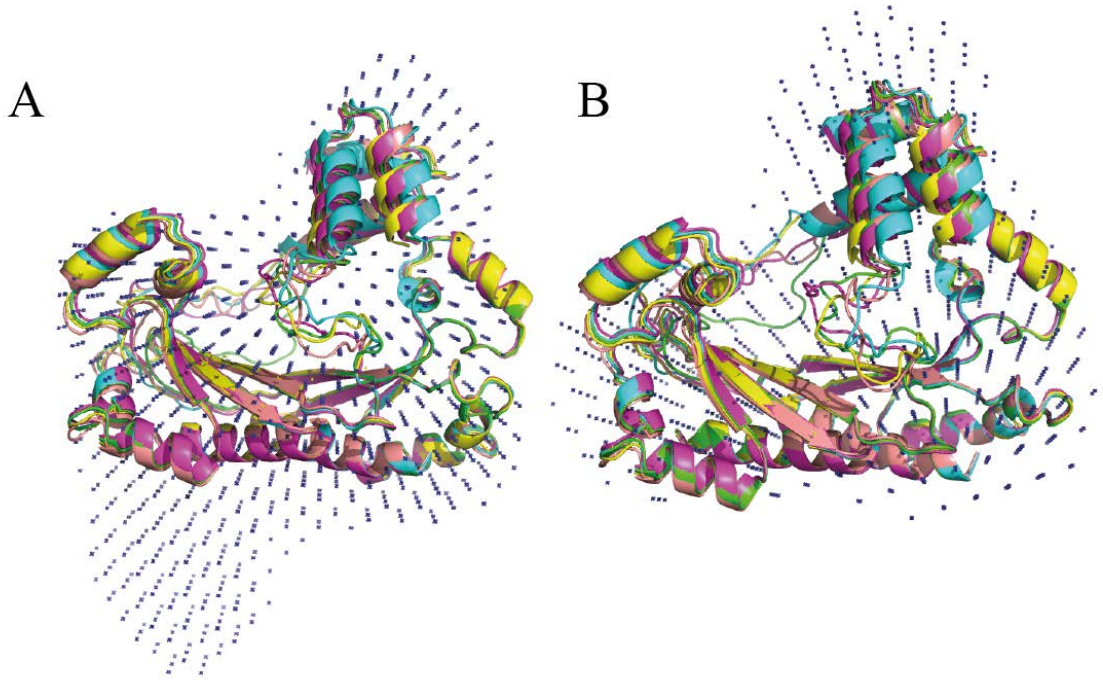


Figure 4.11: Structural models of Sam68 NKKH based on T-STAR NKKH
Overlay of five models of Sam68 NKKH based on the T-STAR NKKH crystal structure within the Sam68 NKKH SAXS envelopes of the free protein at 15mg/ml (A) and with UAAAUAAA at 10mg/ml (B).

Therefore, a structural model of Sam68 KH dimer and the NMR structure of the NK domain were independently fitted into the SAXS envelope of the NKKH domain. This orientation was used to produce another computational model of Sam68 NKKH (Figure 4.12). Fitting this structural model to the SAXS envelope of the free NKKH domain at 15mg/ml gave a χ^2 value of 3.2, calculated using CRYSOL (Figure 4.12A). This shows that this structural model fits better than the Gld-1 STAR model and the initial Sam68

NKKH model based on T-STAR NKKH, also fitting well to the SAXS envelope of Sam68 NKKH with UAAAUAAA (Figure 4.12B), with a χ^2 value of 3.2.

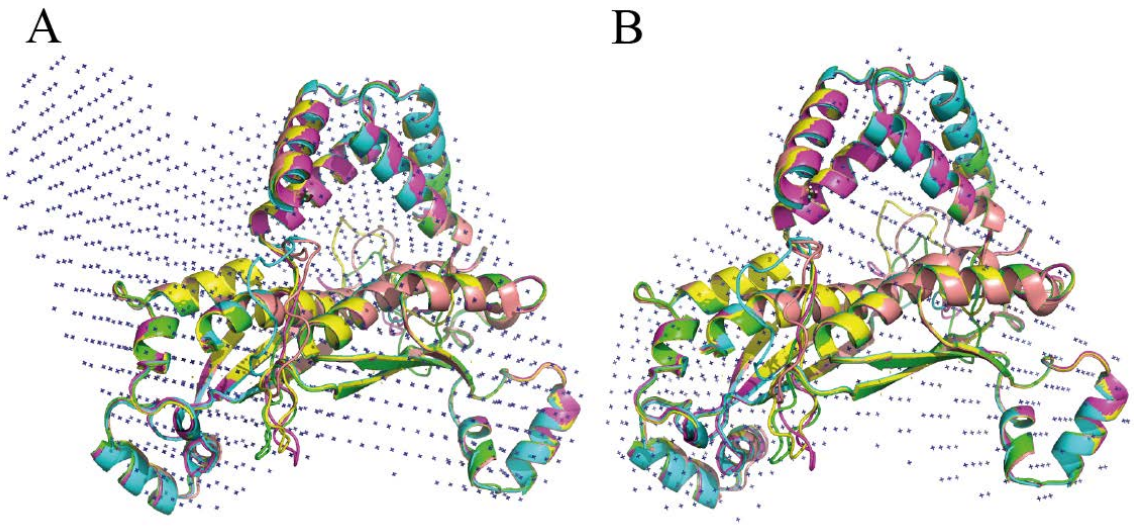


Figure 4.12: Structural models of Sam68 NKKH based on T-STAR KH and Sam68 NK
Overlay of five models of Sam68 NKKH based on physical manipulation of separate NK (Meyer et al. 2010) and T-STAR KH domains within the NKKH SAXS envelopes of the free protein at 15mg/ml (A) and with UAAAUAAA at 10mg/ml (B).

The structure of the KH domain of T-STAR with AUAAU was also solved by X-ray crystallography, making it possible to add this RNA to the NKKH model of Sam68 (Figure 4.13A and B). This structural model demonstrates the binding of one AUAAU to one KH domain within the dimer; another is able to bind to the other KH domain at the other end of the dimer. Figure 4.13 C and D shows the structural model of a single KH domain within the Sam68 NKKH dimer, with AUAAU. This clearly highlights the hydrophobic groove in which the RNA is binding. Further analysis of this region will be discussed in section 4.5.

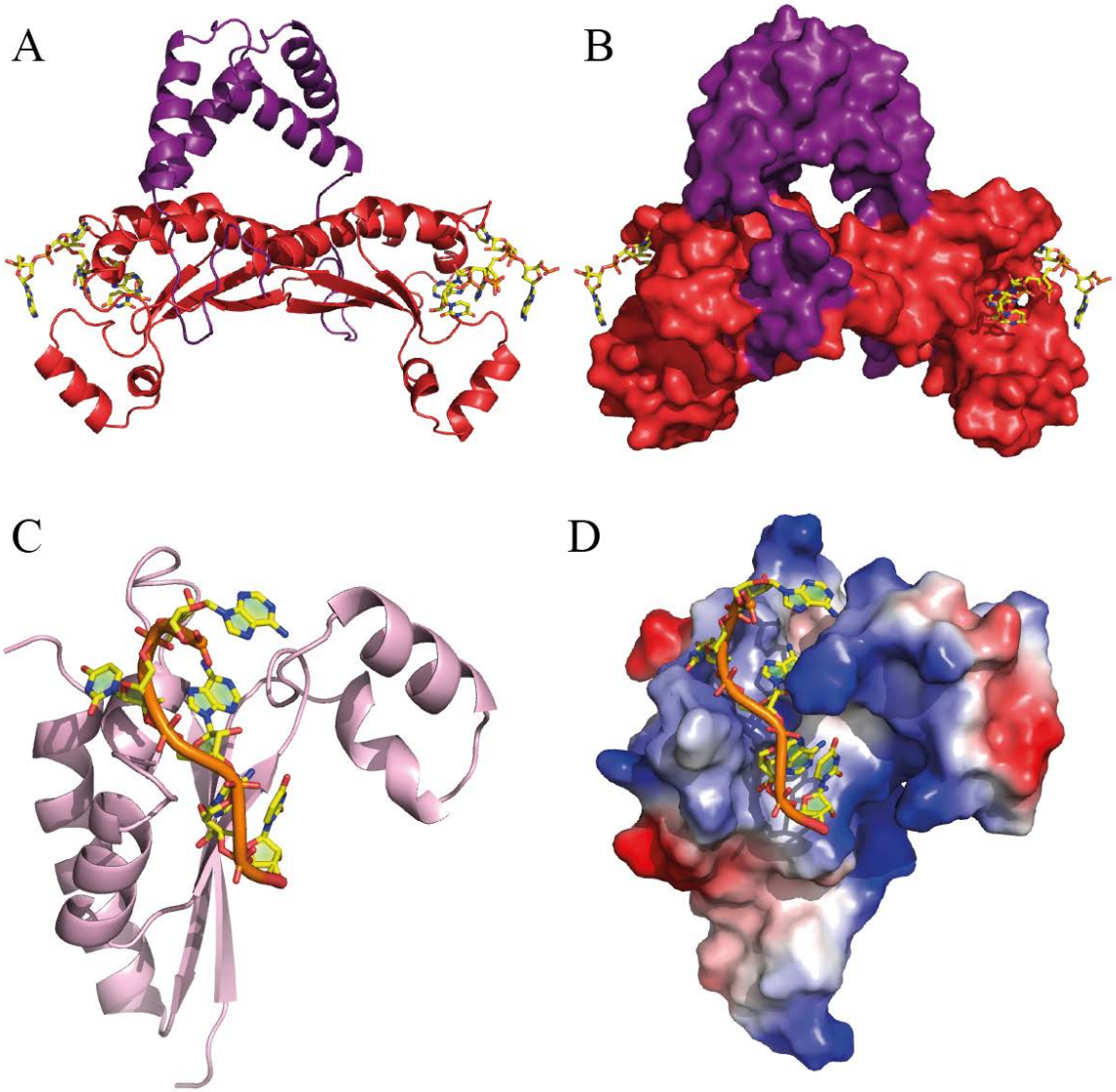


Figure 4.13: Structural models of Sam68 with AUAAU
Cartoon (A) and surface (B) structural models of the NKKH domain of Sam68 with AUAAU based on the T-STAR KH AUAAU complex X-ray structure. The NK-linker is coloured purple, and the KH domains in red. The hydrophobic RNA binding pocket is clearly visible in the KH domain (pink) (C) and electrostatic surface map, with blue indicating a positive, red a negative and white a neutral charge (D).

4.4.7. NMR studies of KH mutants to interrupt dimerisation

The homology model of Sam68 NKKH based on the T-STAR structure and the SAXS data suggest that each KH domain within the dimer is able to dimerise via helix 3 (Figure 4.14). Sam68 has been shown to self-associate *in vivo* by immunoprecipitation experiments; however c-myc tagged Sam68 KH domain alone was unable to pull untagged Sam68 out from cell lysates (Chen et al. 1997). Therefore it was essential to verify the presence of this self-association using several alternative techniques. In order

to validate this dimerisation, a mutation was introduced at the proposed interface of this interaction at position 241, replacing the tyrosine with a glutamic acid, since this is the residue at this position in the QKI structure (Figure 4.9). NMR was used in order to investigate the structural characteristics of this mutant of the STAR, NKKH and KH domains of Sam68.

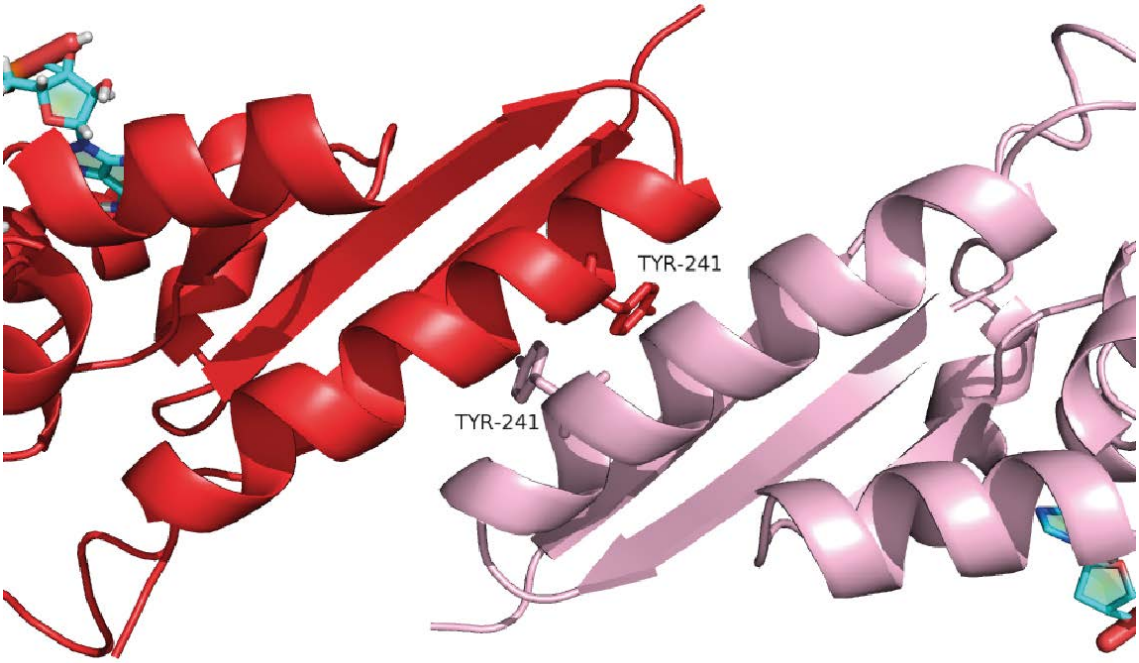


Figure 4.14: KH Dimerisation interface of Sam68
The KH domains (red and pink) within the NKKH domain of Sam68 may dimerise via helix 3, and it is possible that Y241 is essential for this dimerisation.

The Y241E mutants of the STAR, NKKH and KH domains of Sam68 were expressed and purified in the same manner as the WT proteins. This resulted in a significant decrease in the expression levels of each construct and in the case of the STAR domain, loss of the protein following gel filtration (Figure 4.15A). An NMR sample of the NKKH and KH domains could be produced, however the final sample concentration was much lower than for both wild-type constructs, but still sufficient to record ^{1}H-^{15}N HSQC experiments. Interestingly, the KH mutant gave a good quality spectrum, with well-dispersed peaks that overlay well with the WT KH domain (Figure 4.15B), suggesting that this construct is stable, folded and soluble. This is surprising given the instability of the WT KH domain and indicates that perhaps the WT is in exchange between the monomer/dimer/oligomer forms which is consistent with the SAXS data of the KH domain. The NKKH mutant, on the other hand, produced a poor quality spectrum, with fewer peaks, concentrated mostly in the centre of the spectrum (Figure 4.16A).

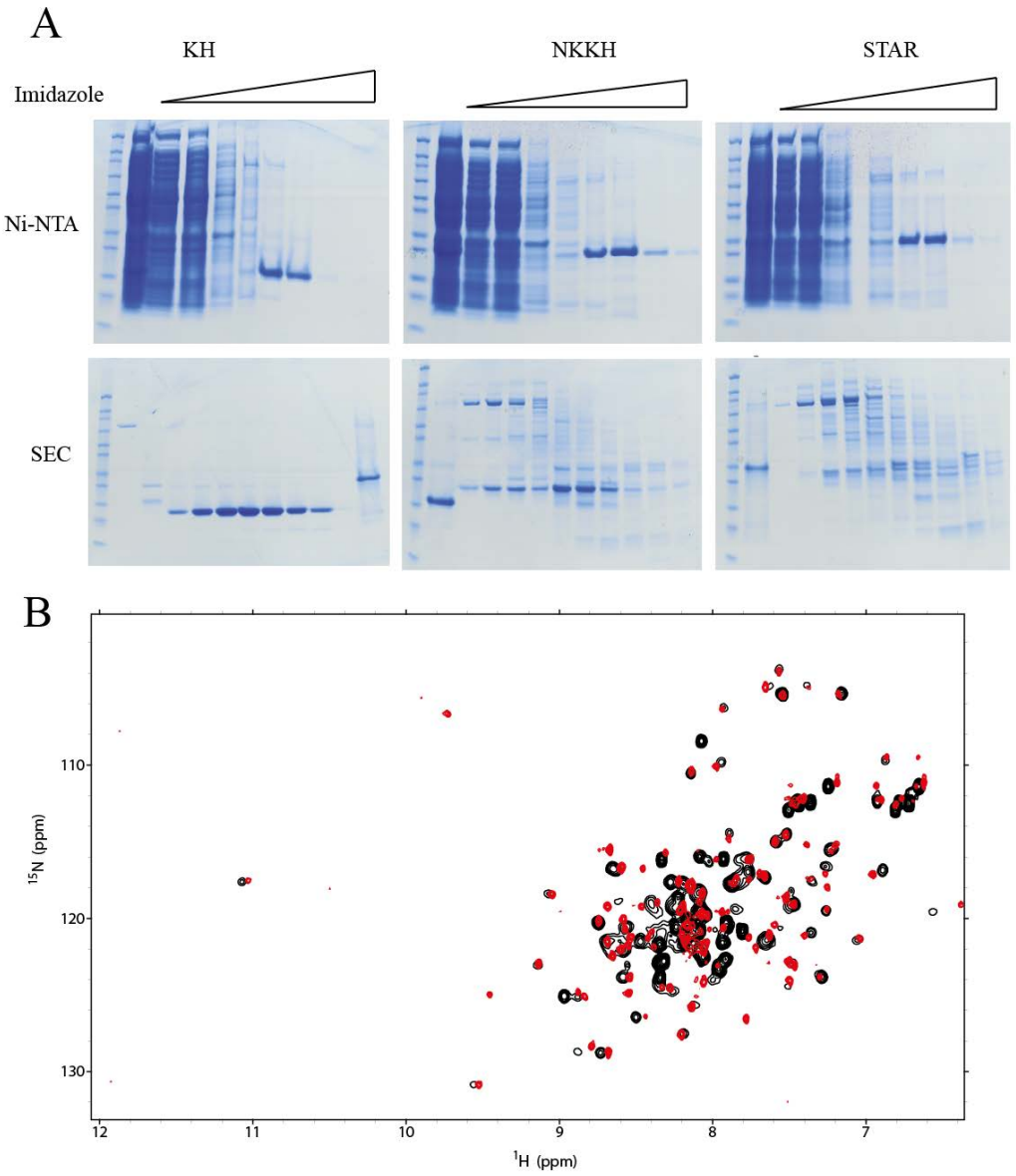


Figure 4.15: The effect of Y241E on the KH, NKKH and STAR domains of Sam68
SDS-PAGE analysis of affinity chromatography and size exclusion chromatography of Y241E mutant of the KH, NKKH and STAR domains of Sam68 (A). Overlay of (^{1}H-^{15}N) HSQC spectra of Sam68 KH WT (black) and Y241E mutant (red) (B).

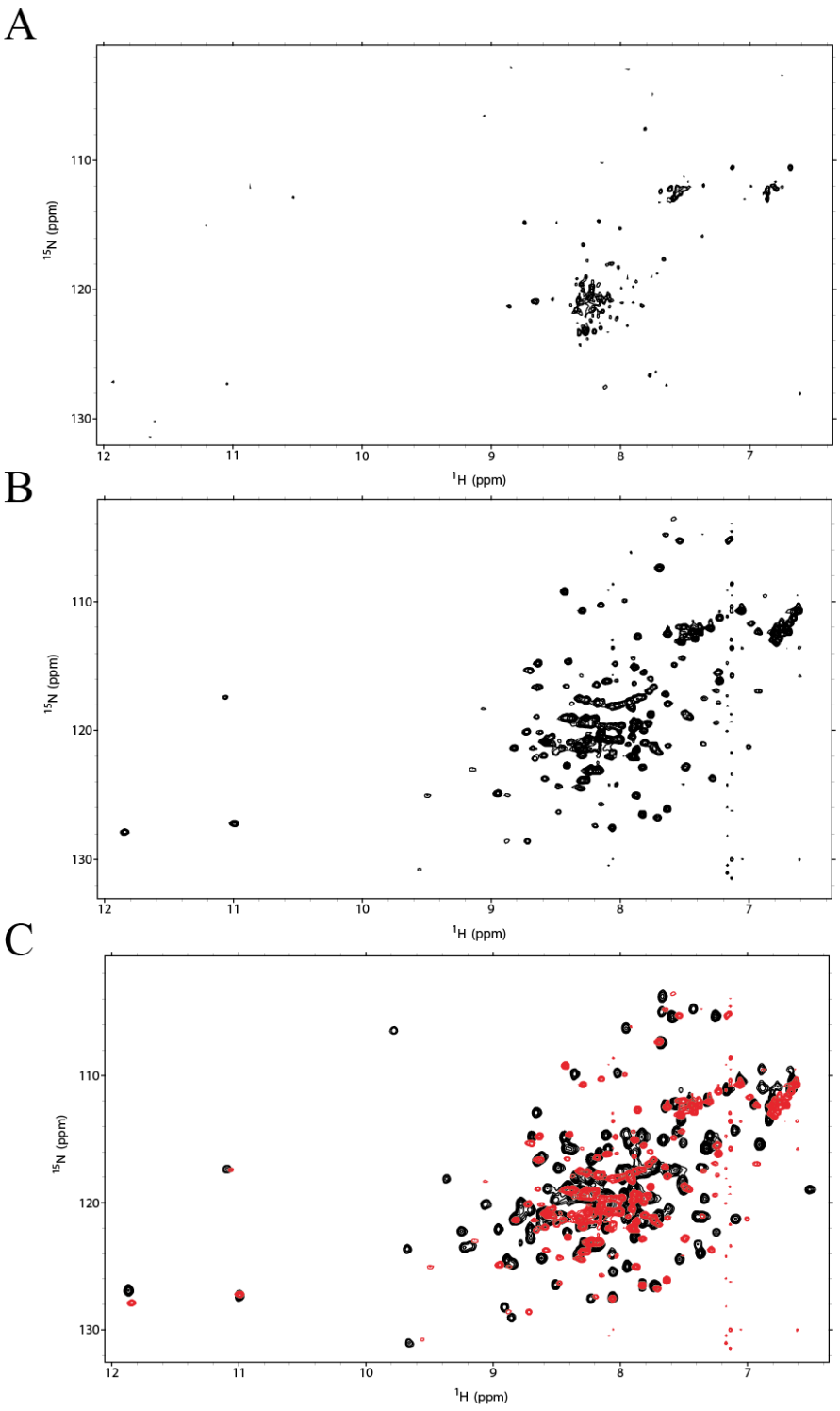


Figure 4.16: NMR analysis of Sam68 NKKH Y241E
(^{1}H-^{15}N) HSQC spectrum of Sam68 NKKH Y241E after gel filtration was of poor quality (A) compared to the sample without gel filtration (B). Overlay of (^{1}H-^{15}N) HSQC spectra of Sam68 NKKH WT (black) and Y241E mutant (red) shows that the NK peaks are not affected by the mutation but the KH peaks are.

The NKKH Y241E mutant was re-expressed and after Ni-NTA purification, the protein was dialysed straight into NMR buffer, without gel filtration and separation of the affinity tag and TEV protease. Although the homogeneity of this sample was

compromised due to the truncation of the purification protocol, it allowed production of a more concentrated sample that was stable enough to record a much better quality (^{1}H-^{15}N) HSQC in just 1 hour at 20°C (Figure 4.16B). Comparison with the WT NKKH spectrum highlights several peaks that are affected by the Y241E mutation (Figure 4.16C). The crosspeaks that correspond to the residues of the NK domain are not significantly affected, and show small chemical shift perturbations as the mutant spectrum was recorded at 20°C rather than 30°C. Many of the KH domain peaks, however, disappear or shift enough such that their position in the mutant spectrum cannot be determined. Figure 4.17A shows the structural model of the NKKH domain of Sam68, with the residues affected by the Y241E mutation highlighted in yellow. The NK domain is not affected by this mutation, there are several residues within the flexible linker but most are located in the KH domain (Figure 4.17B).

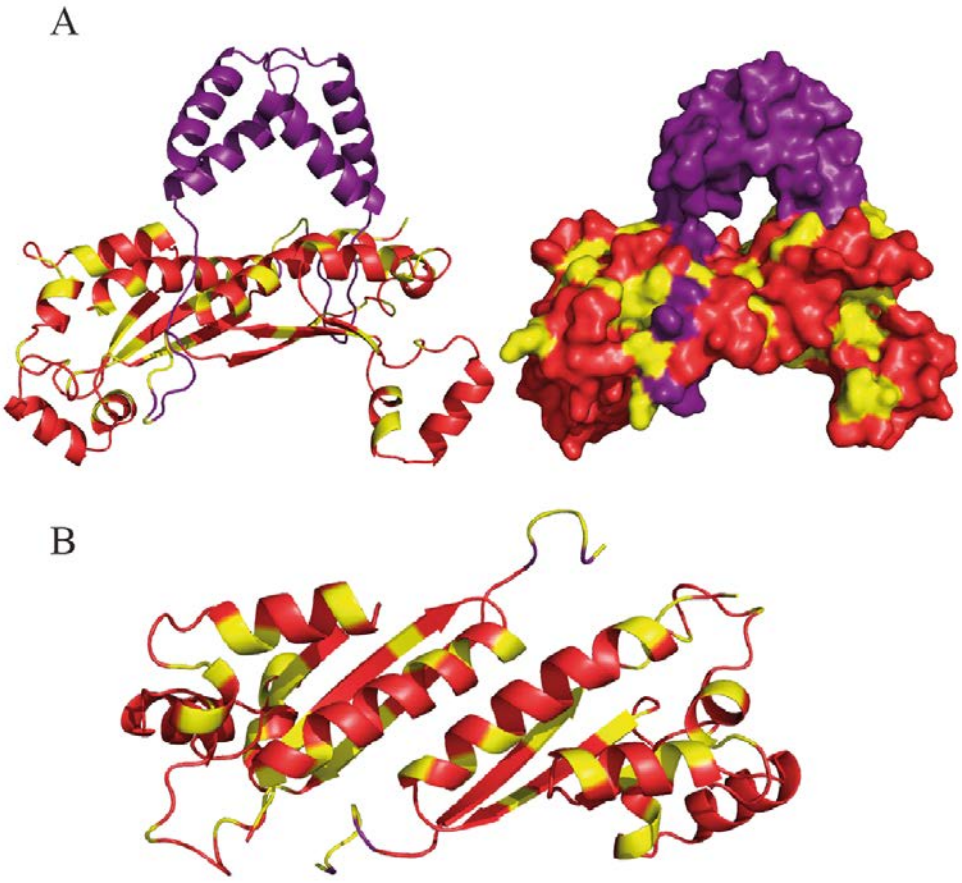


Figure 4.17: The effect of Y241E on the chemical shift of residues within Sam68 NKKH
Plotting the residues affected by the Y241E mutation on the Sam68 NKKH structural model (yellow), determined from Figure 4.17C show that the NK (purple) is unaffected (A). Residues throughout the KH domain (red) are affected, including those near the dimerisation interface (B).

Interestingly, the NKKH Y241E HSQC overlays better with the KH WT than the NKKH WT (Figure 4.18). These data suggest that the Y241E mutation does not disrupt dimerisation via the NK domain, but alters the conformation of the KH domain significantly.

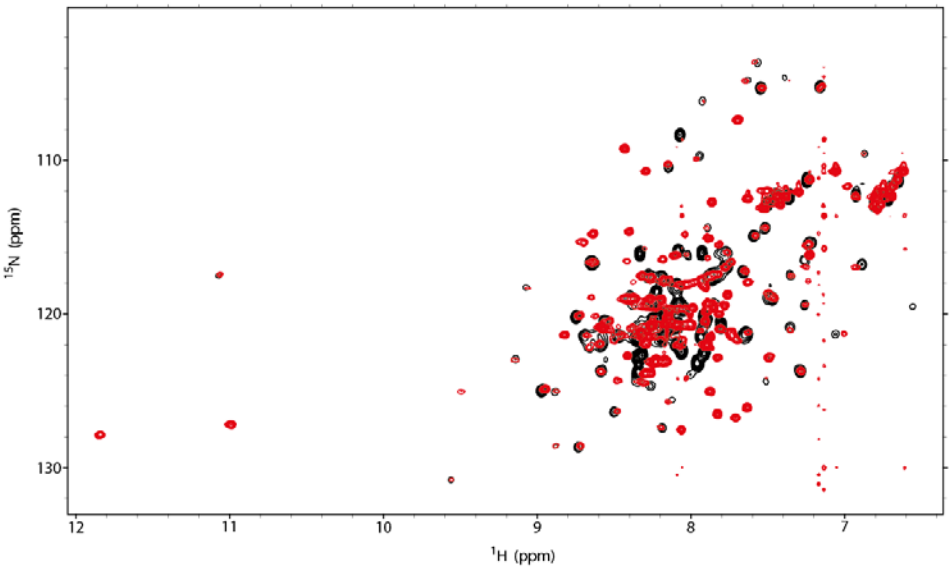


Figure 4.18: Comparison of NKKH Y241E and KH WT (^{1}H-^{15}N) HSQC spectra
Overlay of (^{1}H-^{15}N) HSQC spectra of Sam68 KH WT (black) and NKKH Y241E mutant (red) shows that many peaks of the NKKH Y241E that show a chemical shift perturbation compared to the NKKH WT spectrum have moved to corresponding positions in the KH WT spectrum.

4.5. Biophysical characterisation of Sam68-ssRNA interaction by Fluorescence Polarization

Fluorescence polarisation requires the ligand to be fluorescently labelled, and relatively small. For these experiments the RNA was labelled with a Fluorescein tag at the 5' end of the RNA. This fluorescent molecule absorbs and subsequently emits photons through excitation and de-excitation, which involves redistribution of electrons in the molecule. Therefore, excitation can only occur if the electric field of the applied light has a specific polarization relative to the fluorophore and it follows that the emitted light will also be of a particular orientation. In our case, the fluorescent tag is attached to a small RNA, which is tumbling rapidly in solution. As the molecule is changing orientation, most electrons are excited during exposure to polarized light and subsequently freely change their orientation before emitting photons. Therefore the overall light emitted is largely depolarised and this difference in polarization between incident and emitted light is dependent on the fluorescence lifetime compared to the rotational lifetime of the fluorophore. It follows that if this rotational lifetime is increased, i.e. if the tumbling

speed decreases, then the degree of decorrelation in polarization will decrease. This is observed when the fluorescently labelled RNA binds to the protein of interest, which is comparatively much larger and therefore forms a complex that tumbles much slower in solution than the free RNA. The difference between the polarization state of the light that is emitted from the free RNA and the complex with increasing concentrations of protein can be used to determine the constant of dissociation between the two partners.

It is routine to maintain the fluorescently labelled ligand at a concentration that is lower than that of the estimated dissociation constant. In our case, the Kd was anticipated to be in the low micro-Molar range based on studies of interactions of other KH domain-containing proteins, such as SF1, and our NMR data (Liu et al. 2001). The experiments were conducted in 50µl and the RNA concentration was kept constant at 0.2µM and the protein concentration was varied from 200µM to 0µM by 2-fold serial dilution across the 96-well plate.

4.5.1. Results

Having obtained a good quality model of the NKKH domain of Sam68, we then sought to characterise a consensus RNA binding motif of this domain and identify the residues required for interaction.

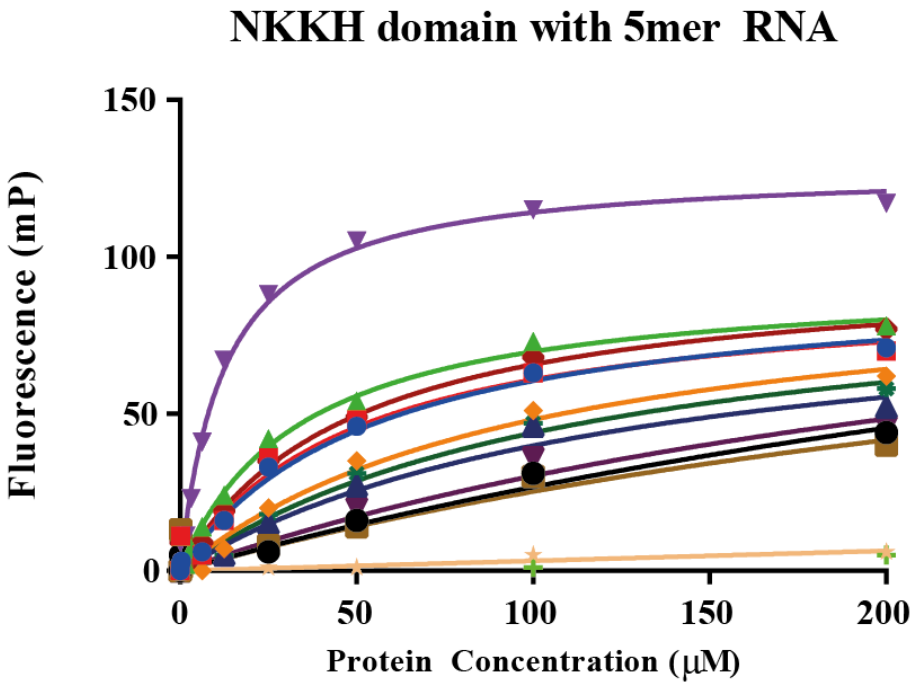
In order to do so, it was necessary to quantify the strength of the interaction between the two components. There are many different techniques that can be used to estimate binding affinity yielding binding parameters such as the dissociation constant (Kd), stoichiometry, enthalpy and entropy. These include isothermal titration calorimetry (ITC), fluorescence polarisation (FP), biolayer interferometry (BLI) and surface plasmon resonance (SPR). Based on the availability of such systems within the department, BLI, ITC and FP were tested. We found that FP gave consistent and reliable results. ITC, however, required a very large amount of protein and RNA and was extremely inconsistent and a Kd could not be estimated due to irreproducibility of a stable baseline, despite attempting to optimise buffer conditions. BLI was more successful, and a Kd similar to that estimated using FP was obtained, however there were also several issues with reproducibility. Therefore, for Sam68-ssRNA complexes, we concluded that FP was the most reliable technique, can be performed in 96 well plates allowing testing of many RNAs and protein constructs and requires a relatively small amount of protein and RNA.

Given that chemical shift perturbation experiments, described in Chapter 3, showed that the majority of RNA of differing lengths interacted with similar binding affinity to the KH domain, several of these sequences were fluorescently labelled for FP, along with several new A/U-rich RNA. These RNA sequences ranged in length from 4 to 13 nucleotides, and were synthesized with either a 5' or 3' fluorescein tag, with and without a 3-cytosine spacer between the label and the target sequence. RNAs with a 3' fluorescein tag with a CCC linker produced very low fluorescence polarization, suggesting that the fluorescein tag remains highly flexible upon complex formation or may be interfering with the interaction between the RNA and protein. Therefore only 5'-tagged RNAs with a CCC linker were used in the following fluorescence polarization experiments. Based on RNA binding experiments by NMR, the KH domain interacts with RNA with similar affinity on its own as well as within the NKKH and STAR domains. Since RNA binding is not affected by the presence of the NK or CK domains, but the NKKH domain is more stable than other constructs, this domain was used in FP, with just a few experiments carried out on the C238A and STAR constructs for confirmation of similar binding affinity. The same expression and purification protocols were used as for the production of NMR samples as well as the final buffer used in the assay (10mM Tris-HCl, 100mM NaCl, pH7 and 0.1% β-mercaptoethanol).

Protein concentration was estimated using the Bradford Assay to ensure the most consistent estimation of protein concentration for calculation of precise dissociation constants (section 2.4.6).

4.5.2. Specific RNA sequence recognised by Sam68 NKKH

In general, KH domains have been shown to directly bind 4 nucleotide sequences (Valverde, Edwards & Regan 2008) but the structure of T-STAR in complex with AUAAA suggested that only 3 nucleotides are specifically recognised by the KH domain. Therefore to investigate the importance of the nucleotide at each position within the AUAAA motif, a series of 5mer RNAs based on existing SELEX data were tested (Figure 4.19).



	Kd	Bmax	R square
AUAAA	54.89	93.75	0.967
UUAAA	49.51	91.05	0.9478
CUAAA	34.05	93.63	0.9824
AAAAA	12.52	128.5	0.9939
ACAAA	100.1	96.37	0.968
AUUAA	474.6	153.5	0.8616
AUCAA	368.5	118.4	0.8258
AUAUA	126.9	90.47	0.9039
AUACA	301.3	121.4	0.8169
AUAAU	48.53	97.64	0.9875
AUAAC	115.5	94.8	0.9543
CCCCC	-	-	n/a
UUUUU	-	-	n/a

Figure 4.19: Fluorescence polarisation of Sam68 NKKH with 5mer RNA
Fluorescence plotted against NKKH concentration and calculated dissociation constants, Bmax and r^2 with 5mer RNA. Systematic alterations of the nucleic acid at each position were made to determine a consensus RNA binding sequence.

The AUAAA motif itself bound to the NKKH domain with a Kd of 55μM. Changing the first nucleotide to a U or a C did not alter the affinity significantly, with Kds of 34 and 49μM, respectively, suggesting that this nucleotide is not important for specificity.

Sam68 had higher affinity for poly(A), suggesting that the second position can either be A or U, but prefers A, with a Kd of 12µM. However, a C at this position decreases the affinity to 100µM, indicating that the nucleotide at this position has an impact on specificity. Changing the third and fourth nucleotides from A to U or C decreased the binding affinity significantly, suggesting only an adenosine can be accommodated at these positions. Finally, changing the fifth position to U did not alter the affinity, with a Kd of 50µM, and there was some loss of affinity for the sequence ending in C, to 115µM. All together, these data suggest that the 5mer RNA consensus sequence for the NKKH domain of Sam68 is (A/U/C)-(**A**/U)-A-A-(A/U).

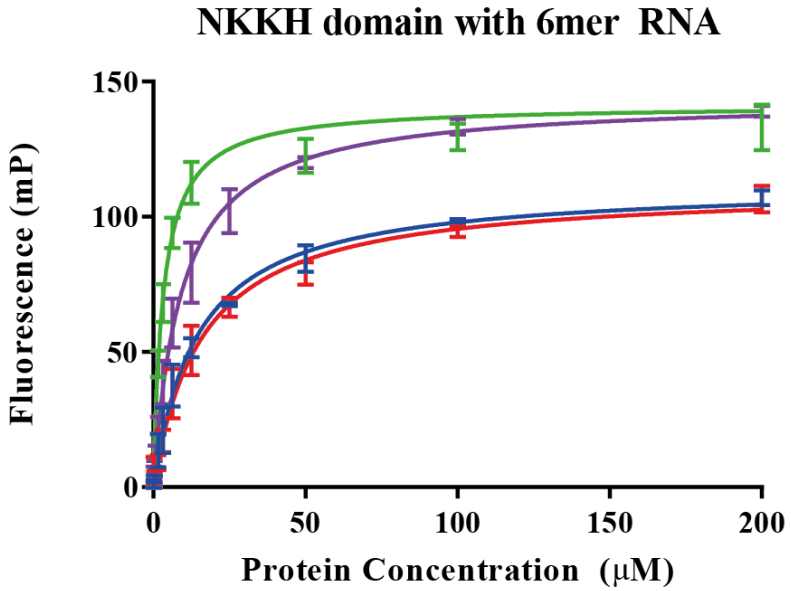


Figure 4.20: Fluorescence polarisation of Sam68 NKKH with 6mer RNA
Fluorescence plotted against NKKH concentration and calculated dissociation constants, Bmax and r^2 with several 6mer RNA used in NMR chemical shift titration experiments.

Interestingly, increasing the length of sequence to 6 nucleotides and adding A and U nucleotides either side of the UAA motif, further increased Sam68 binding affinity (Figure 4.20). For example, AAAUAA and AAUAAU have dissociation constants of 3.776µM and 9.223µM, respectively, indicating higher affinity than any of the 5mer RNAs despite lacking the full UAAA motif. In addition AUUAAA and AUUAAU have Kds of 18.59µM and 22.22µM, respectively.

4.5.3. The CK of Sam68 does not contribute to the affinity for RNA

FP experiments were also used to verify that the KH domain within the full STAR domain construct also interacts with RNA with similar affinity as it does within the NKKH domain (as seen in NMR experiments). Furthermore, FP experiments were important to confirm the lack of involvement of the CK domain in RNA recognition. To accomplish this, UAAAUAAA was used since it is of sufficient length to accommodate CK binding as suggested by Gld-1, QKI and SF1 structures. The Kd of the NKKH with this 8mer is 3.97µM, similar to the 6mer AAAUAA. The Kd of the STAR domain to this 8mer RNA was 10.4µM, strongly suggesting that the CK is indeed not involved in RNA binding (Figure 4.21).

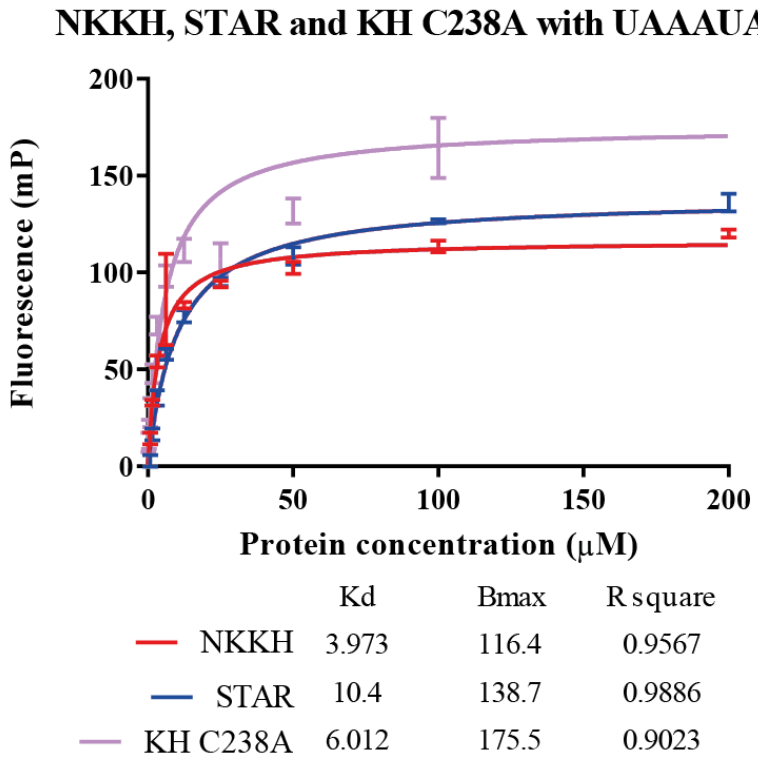


Figure 4.21: Fluorescence polarisation of Sam68 KH C238A, NKKH and STAR with UAAAUAAA
Fluorescence plotted against protein concentration and calculated dissociation constants, Bmax and r^2 with UAAAUAAA.

In addition, the KH C238A mutant was tested with UAAAUAAA for further comparison of binding affinity without the NK dimerisation domain. The Kd was estimated to be 6.012µM, similar to the Kd estimated for the NKKH and STAR domain constructs (Figure 4.21). This confirms that the KH domain is interacting with RNA with similar affinity in isolation or within larger domain constructs and that the NK and CK do not influence the affinity of the KH for the RNA.

This was also the case for several longer RNAs that were tested with Sam68 NKKH and STAR by FP (Figure 4.22) (full RNA sequences are listed in Appendix 7.9). Sequences G7.1 and G8.1 had the highest Kds for Sam68 of all RNAs identified by SELEX (Lin, Taylor & Shalloway 1997), SRE was found to be the most commonly bound to Sam68 of a pool of RNAs from SELEX data (Galarneau, Richard 2009) and N2L represents the Sam68 target region of the *Nrxn2* gene (Ehrmann et al. 2013). These sequences range from 29 to 52 nucleotides in length, and the affinity was found to be higher for Sam68 NKKH than for the full STAR domain, which was observed consistently throughout the FP experiments. This suggests that the CK domain does not contribute to RNA binding, and the reduction in affinity of the STAR domain compared to the NKKH domain may be attributed to some degradation of the STAR domain. There are significant differences in binding affinity for each of the physiological RNAs; however since FP is size limited with respect to the fluorescently labelled ligand, alternative techniques such as ITC or BLI would be ideal to confirm these patterns of binding.

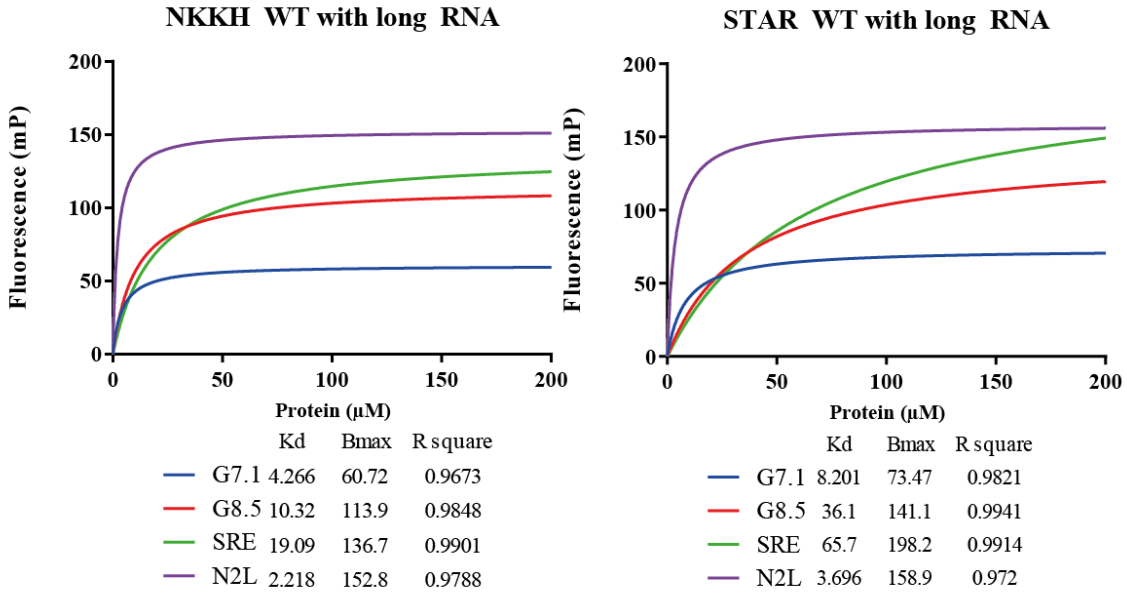


Figure 4.22: Fluorescence polarisation of Sam68 NKKH and STAR with longer RNA
Fluorescence plotted against protein concentration and calculated dissociation constants, Bmax and r^2 with different long RNAs previously shown to bind Sam68; G7.1, G8.5 (Lin, Taylor & Shalloway 1997), SRE-4 (Galarneau, Richard 2009) and N2L (Ehrmann et al. 2013).

4.5.4. Specific amino acids of Sam68 KH that are necessary for RNA binding

Having established fluorescence polarisation as a suitable technique for estimating the Kd of interaction between Sam68 and short RNAs, this was then used to investigate which residues in the KH domain are essential for RNA binding. Analysis of chemical

shift perturbation experiments identified several residues of the KH domain that are affected by the addition of RNA (Chapter 3). Residues that underwent a chemical shift perturbation greater than 0.8ppm (N171, F172, K175, L177, I184, K185, V197, S202, K206, E209, E210 and M258) were plotted on the structural model of Sam68.

This highlighted two putative binding pockets at opposite ends of the NKKH dimer (Figure 4.23). Figure 4.24 shows a single KH domain from the NKKH structural model in pink, and AUAAU modelled in from the T-STAR structure. The residues identified from NMR are highlighted in blue, and the conserved GXXG motif in green, highlighting a hydrophobic groove within the KH domain to accommodate RNA (Figure 4.24).

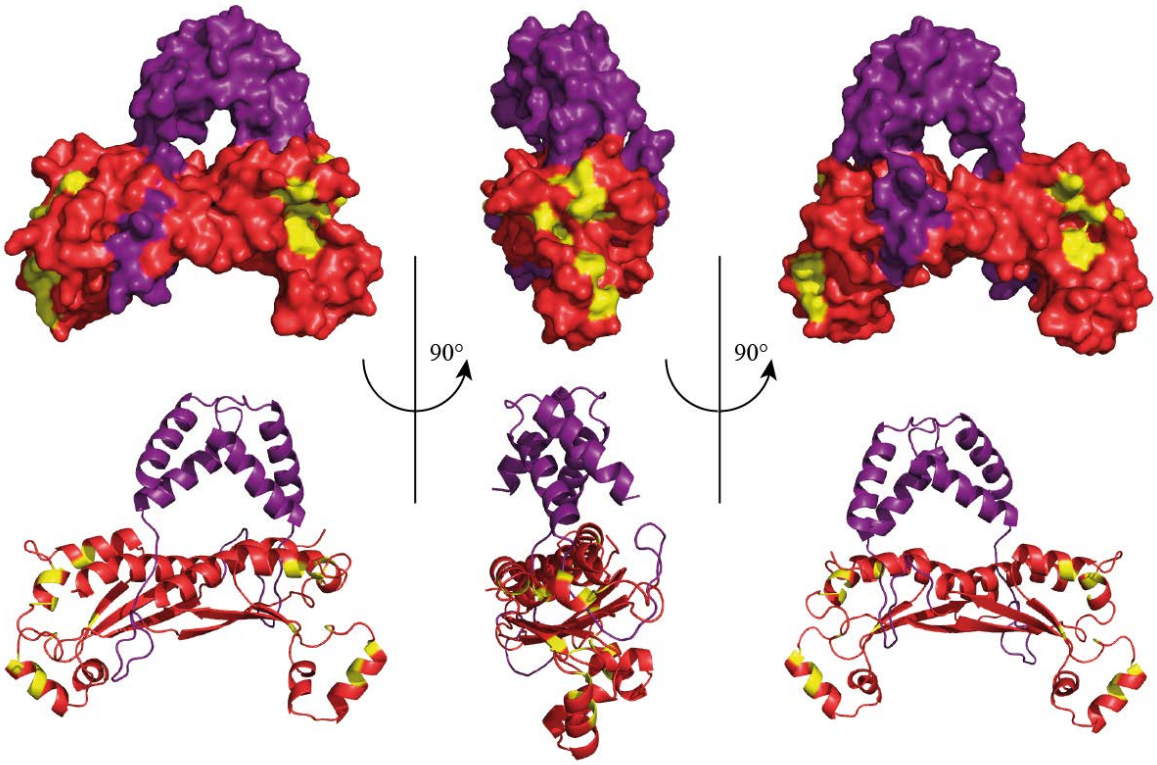


Figure 4.23: RNA binding residues of Sam68 NKKH domain
The residues involved in RNA binding as determined by chemical shift perturbation experiments plotted on the structural model of Sam68 NKKH (yellow) as shown in various orientations. These residues are located at either end of the dimer, within the KH domain (red). The peaks of the NK domain (purple) were not affected by the addition of RNA.

To investigate the contribution of each of these amino acids in RNA recognition, mutants of the NKKH domain were generated, replacing each residue with an alanine, for fluorescence polarisation assay with UAAAUAAA. The mutants M258A, V197A and I184A could not be purified in sufficient quantity for FP, suggesting that these residues may be important for correct folding of the NKKH domain. The remaining mutants were tested in two batches, along with a sample of the NKKH WT (Figure

4.25A). Since the binding of the WT to UAAAUAAA was different for each purification, for comparison, the percentage change in Kd between each WT and mutant was calculated (Figure 4.25B).

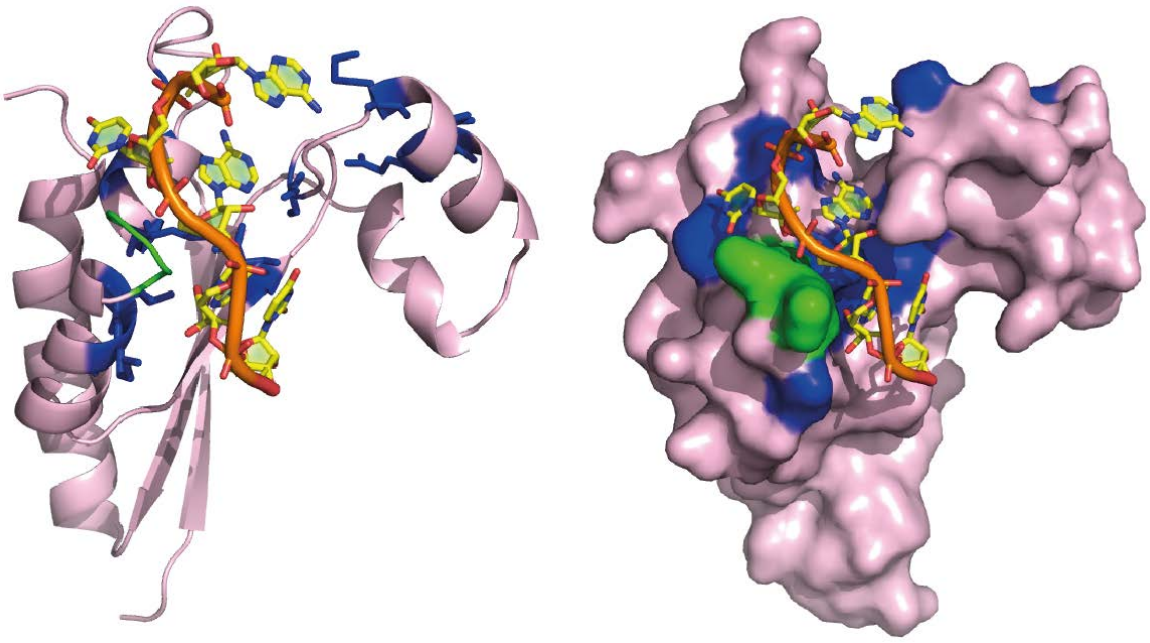
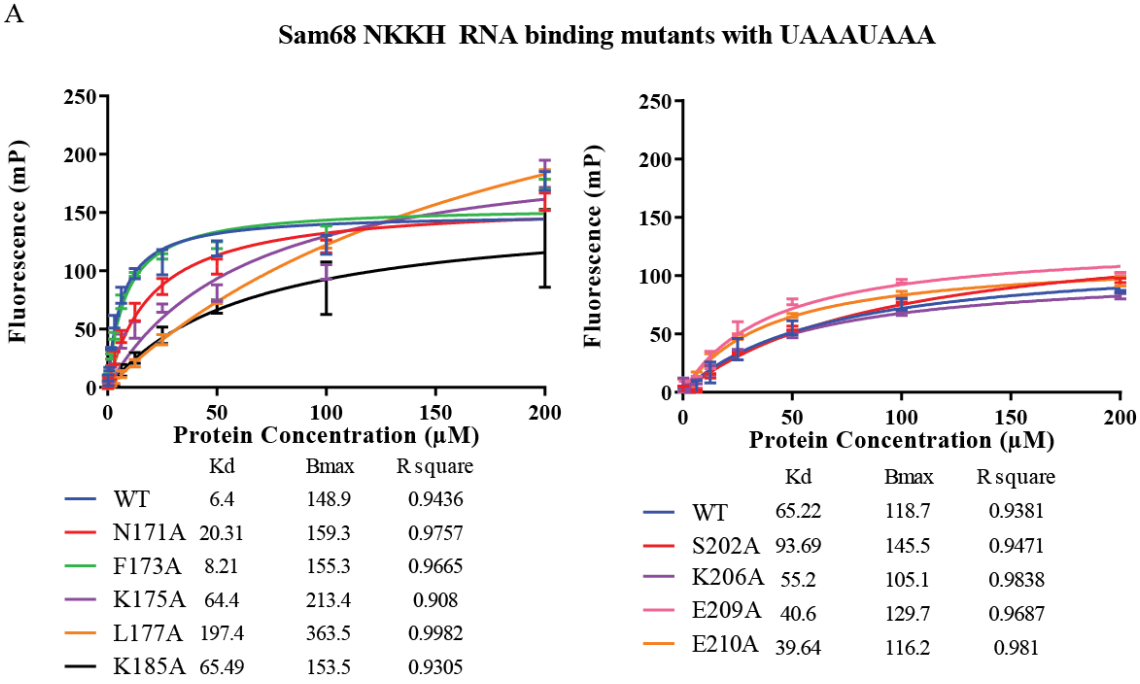


Figure 4.24: RNA binding pocket of Sam68 KH
RNA binding residues from NMR studies plotted on the structural model of the KH domain of Sam68 (blue), with RNA. The conserved GXXG loop is highlighted in green.

This graph shows that mutation of L177 caused the greatest change in RNA binding, from 6.4µM to 197.4µM, followed by K175A and K185A which both had a similar negative affect on RNA binding, with Kds of 64.4µM and 65.49µM, respectively. This suggests that these residues are crucial for RNA recognition. N-terminal to these residues, N171A and F172A both decreased NKKH RNA binding affinity to UAAAUAAA with Kds of 20.31µM and 8.21µM, respectively. Therefore F172 is less critical for RNA recognition, which is also the case for S202, K206, E209 and E210. Although the RNA binding affinity appeared to increase for K206A, E209A and E210A, this was not significant.

A



B

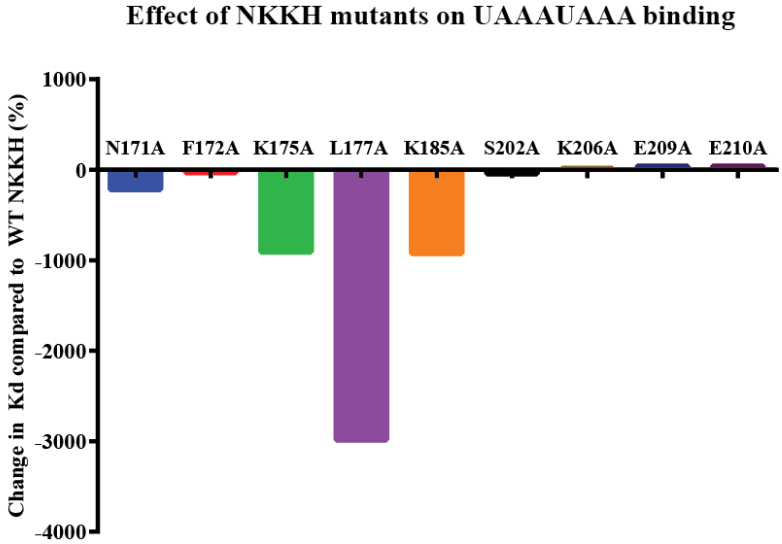


Figure 4.25: Fluorescence polarisation of Sam68 NKKH RNA binding mutants with UAAAUAAA

Fluorescence plotted against protein concentration and calculated dissociation constants, Bmax and r^2 of two sets of RNA binding mutants with UAAAUAAA (A). Percentage change in Kd was calculated using; ((x-y)/x)*100, where x is the Kd of the WT for each set of mutants, and y is the Kd of the mutant for the same RNA.

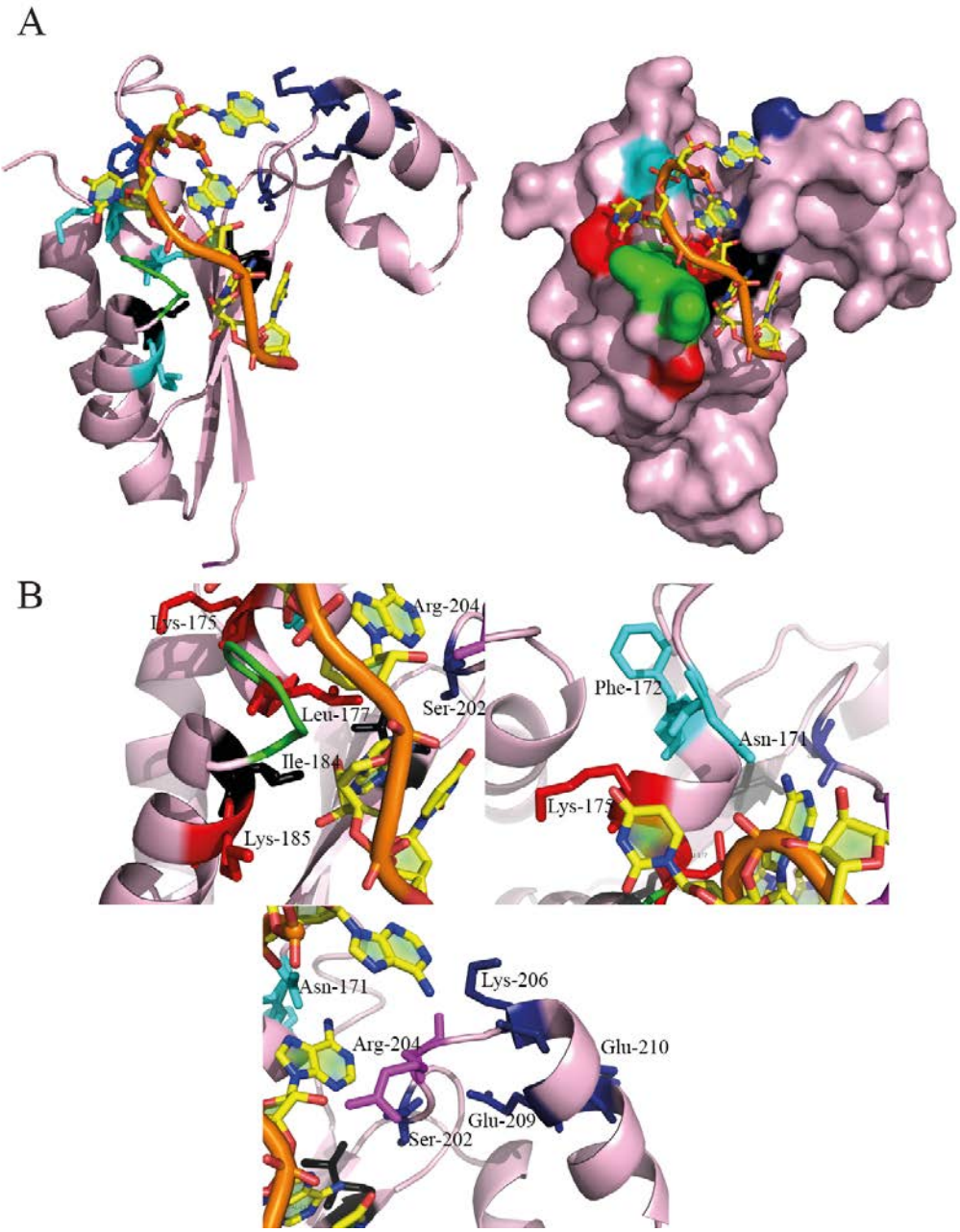


Figure 4.26: Effect of RNA binding mutant on Sam68 KH
RNA binding mutants highlighted on the structural model of Sam68 KH with AUAAU cartoon and surface representations. The non-expressing mutants are highlighted in black, the GXXG loop in green, the mutants with the most effect on RNA binding (K175A, L177A and K185A) in red, N171A and F172A in cyan and S202A, K206A, E209A and E210A in dark blue (A). Zoomed images of the three groups of RNA binding mutants, including R204 (purple) (B).

Identifying the position of these residues on the structural model of the NKKH domain highlights the hydrophobic binding pocket (Figure 4.26A). The GXXG motif is highlighted in green, and V197 and I184, the two mutants that did not express are in black. The residues that were most affected in RNA binding by mutation to alanine are highlighted in red and surround the GXXG loop. This is also shown in Figure 4.26B, which shows each labelled residue within the KH domain with and without AUAAU,

with the domain rotated 90° to the previous figure. These images show that this region is distinct from N171 and F172 and even more so for the third group of S202, K206, E209 and E210. These residues are likely to shift as a result of a conformational change of the protein through RNA binding to R204 (Figure 4.26B). In the T-STAR structure, the corresponding residue, R104, interacts directly with the RNA backbone and causes a conformational change in this region upon RNA binding. Therefore it is likely to be the same case for Sam68. This residue was not identified by NMR as it could not be assigned within the spectra of the NKKH or STAR domains.

4.5.5. Confirming the Sam68 NKKH structural model by Fluorescence Polarization

The chemical shift perturbation experiments described in Chapter 3 were undertaken with short RNAs to investigate the binding of a single KH domain within the STAR dimer. Physiologically, the dimer will bind one long section of RNA at two distinct sites that are separated by a non-specific linker, as illustrated in Figure 4.27A and B. The orientation of the dimer suggests that an RNA linker of more than 15 nucleotides should be present between the two UAAA binding sites. This is also reflected in the T-STAR SELEX data, which predicts an average linker region of at least 16 nucleotides between the two protein binding sites on the RNA (Figure 4.27C). If the STAR domain of Sam68 forms a dimer similar to that of Gld-1 then the distance between the two RNA binding sites can accommodate approximately 10 nucleotides indicating that the RNA binding sites on the KH domain of Sam68 within the dimer are likely to be in an alternative position relative to each other. To test this, RNAs were designed with a 5'-Fluorescein tag with two UAAA Sam68 binding sites, separated by different length stretches of cytosines, (1-30) which do not bind Sam68 (as demonstrated in the 4 and 5mer FP experiments).

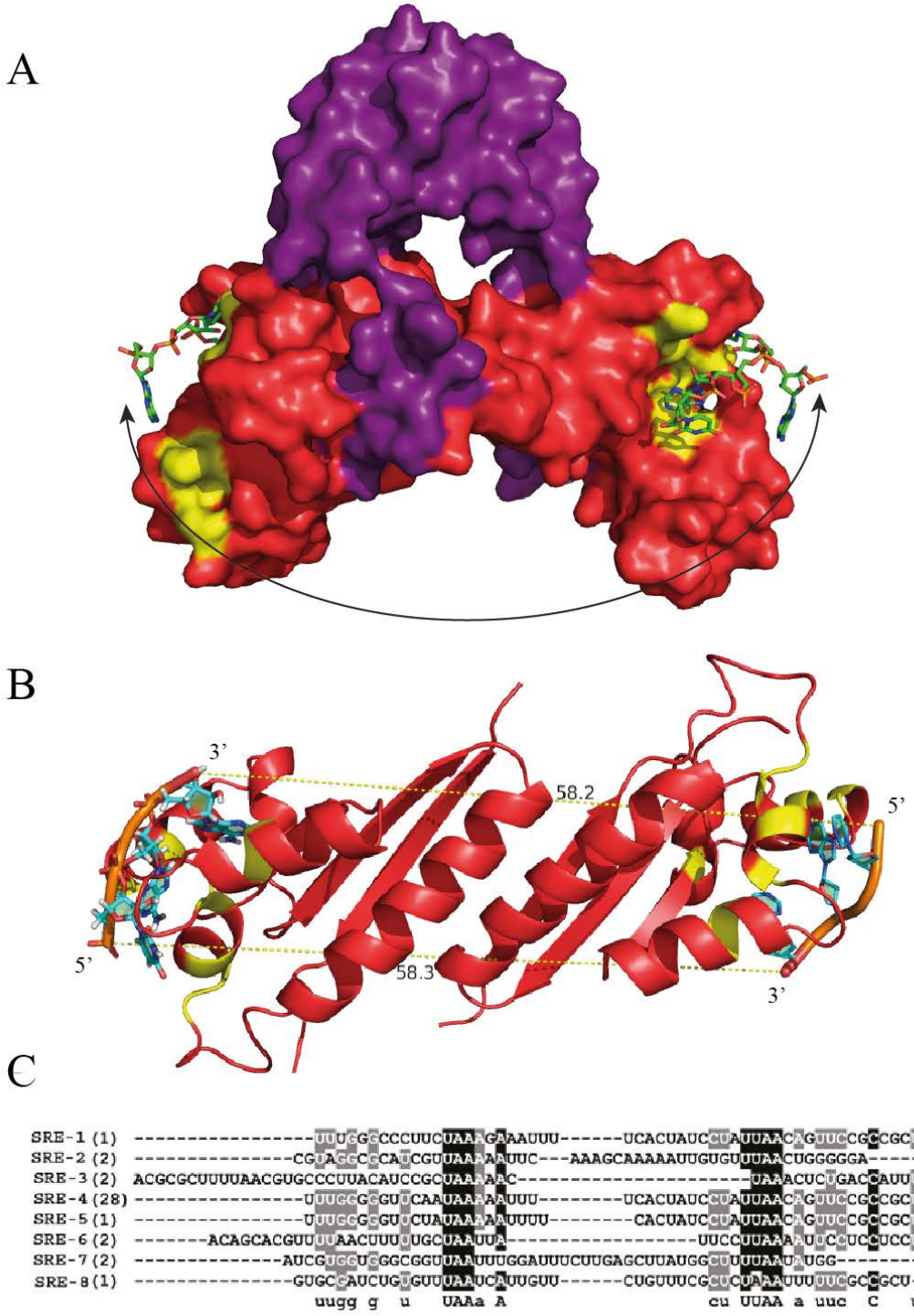


Figure 4.27: Distance between two RNA binding sites in Sam68 NKKH
Physiologically, Sam68 recognises two RNA motifs at either end of the dimer, indicated in yellow (A). The distance between these two sites is approximately 58Å (B), suggesting that at least 16 non-specific nucleotides are necessary to link these two sequences, which is predicted from SELEX experiments (Galarneau, Richard 2009) (C).

The UAAA-5C-UAAA RNA bound the NKKH domain with a similar affinity than the single site UAAACCC sequence with Kds of 49 and 42µM, respectively, and with

similar strength binding as the longer UAAA-10C-UAAA and UAAA-15C-UAAA RNA (Figure 4.28A). The Kd then decreased, indicating higher affinity binding for the UAAA-20C-UAAA, and particularly the UAAA-25C-UAAA and UAAA-30C-UAAA RNAs, suggesting that a spacing of at least 20 nucleotides is required for additive binding between the two KH domains of the NKKH construct. This is consistent with the structural model of Sam68 NKKH.

The STAR WT was also tested with these RNA and although an accurate Kd with UAAACCC could not be determined, the same pattern of binding was observed (Figure 4.28B). This defines a RNA consensus sequence of (A/U/C)-(**A**/U)-A-A-(A/U)-$N_{>15}$-(A/U/C)-(**A**/U)-A-A-(A/U).

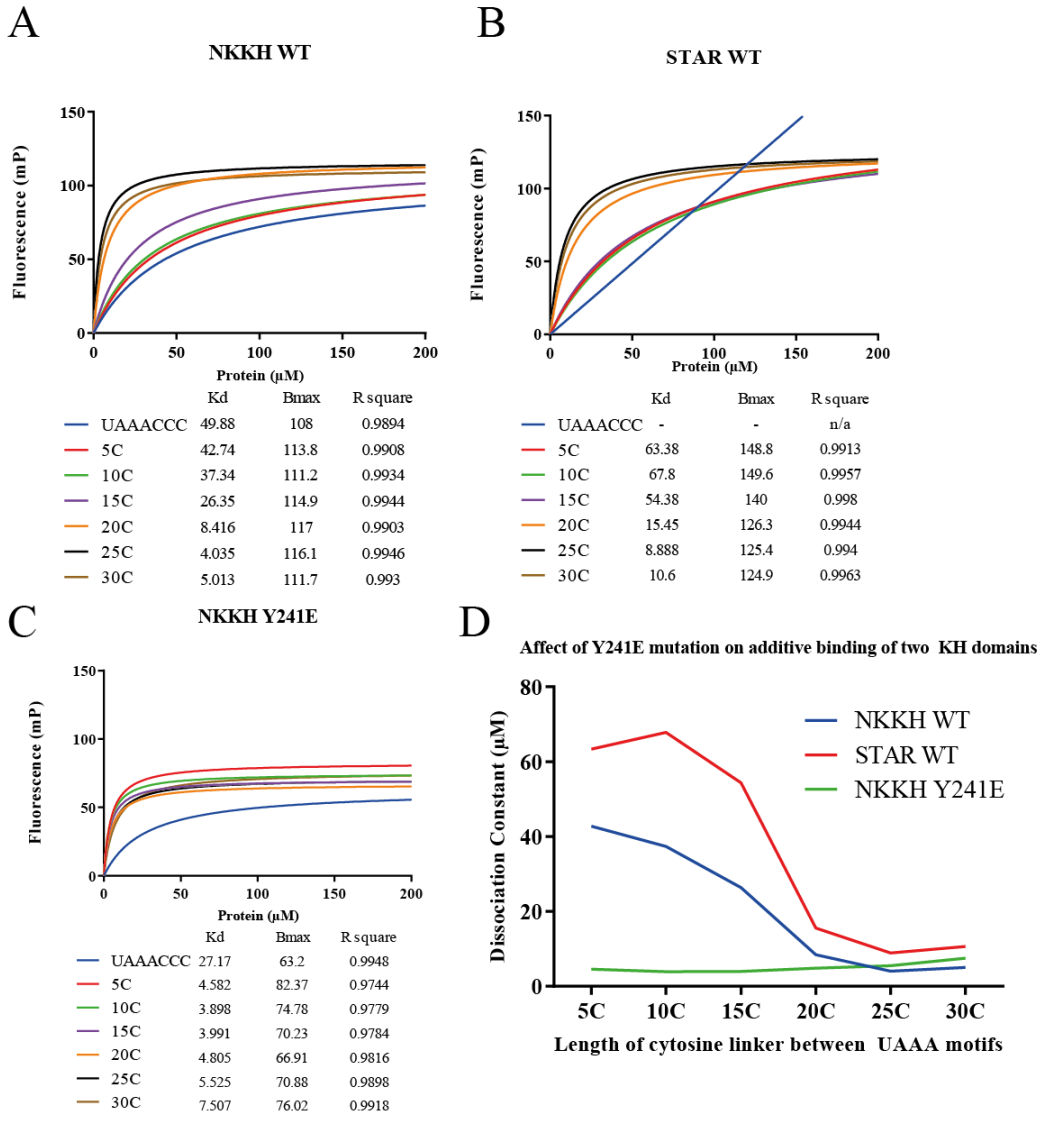


Figure 4.28: Fluorescence polarisation of Sam68 constructs with different length RNA
Fluorescence plotted against protein concentration and calculated dissociation constants, Bmax and r^2 of NKKH WT (A), STAR WT (B) and NKKH Y241E (C) with RNA of UAAAxUAAA where x corresponds to 5, 10, 15, 20, 25 and 30 cytosines. The Kd values for each construct are plotted against cytosine linker length (D).

4.5.5.2. Interaction of Sam68 NKKH Y241E with AU-rich RNA of different lengths

Having demonstrated that the RNA binding interface of Sam68 KH domain within the NKKH computational model based on T-STAR differs to that of the Gld-1 model, it was also possible to investigate the dimerisation of the KH domain using FP. We have shown that the Y241E mutant disrupts the stability of the NKKH and STAR domains of Sam68 by NMR, strongly suggesting that this mutation affects the conformation of the dimer. If that is the case, then the affinity of this mutant for RNA of different lengths should also be different to that of the WT NKKH domain.

To investigate the effect of this mutation on RNA binding, unlabelled NKKH Y241E was tested with each RNA containing different length cytosine linkers as for the wild-type NKKH (Figure 4.28C). Interestingly, the affinity increased significantly with two UAAA sites separated by 5 Cs compared to a single UAAA motif, unlike the WT. This may be due to a loss of KH dimerisation and an alternative arrangement of the KH domains, allowing binding of two RNA sites that are close together. Interestingly, the Kd for 5C, 10C, 15C, 20C, 25C and 30C were very similar. This demonstrates that the Y241E mutation has affected the RNA binding properties of the two KH domains within the NKKH dimer (Figure 4.28D). Additionally, the Kd of the Y241E mutant for each RNA was lower than those of the WT, suggesting that this construct has a stronger affinity for RNA, although this may be due to the experiments being carried out at different times.

4.6. Functional effect of Y241E mutation

In order to determine whether mutation of tyrosine 241 to glutamic acid and disruption of the KH dimer has a functional effect on Sam68, *in vitro* splicing assays were conducted using a *Nrxn3* minigene. Experiments were performed by Oksana Gonchar and Marina Danilenko in collaboration with Professor Ian Eperon and Professor David Elliot. Splicing of this neurexin gene has been shown to be modulated by Sam68 and T-STAR (Ehrmann et al. 2013). As expected, in contrast to wild-type Sam68, Sam68 Y241E mutant has no effect on the splicing of the neurexin3 minigene (Figure 4.29). This suggests that mutation of Y241 inhibits Sam68 splicing function through prevention of KH dimerisation (Figure 4.29).

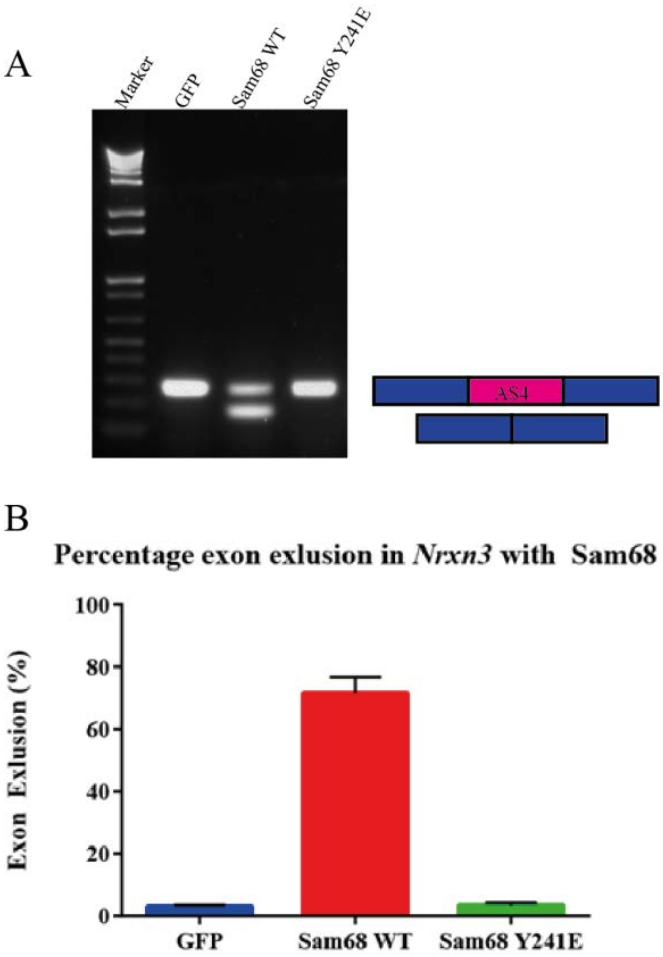


Figure 4.29: Splicing of Neurexin3 by Sam68
Splicing gel showing induction of exon AS4 exclusion in *Nrxn3* RNA with Sam68 WT but not GFP alone or Sam68 Y214E mutant (A). Percentage exon inclusion for each construct shows that the Y241E mutation inhibits Sam68 splicing function (B). (Experiment performed by Oksana Gonchar and Marina Danilenko in collaboration with Professor Ian Eperon and Professor David Elliot).

4.7. Discussion

The aim of this chapter was to investigate the mechanism and specificity of RNA binding by Sam68, by determining the structure of this STAR protein in complex with RNA.

The STAR and NKKH complexes with RNA were found to be unsuitable for structural studies by NMR; however the sample optimisation strategy outlined in Chapter 3 provided an excellent point at which to begin crystallisation trials. Hits were obtained of the NKKH and STAR domains of Sam68 in complex with RNA, and optimisation resulted in protein crystals that diffracted to 3Å. These were anisotropic and the data set could not be processed, so several strategies were applied in order to further optimise crystal growth. These have not yet been successful; however it is possible that further

trials may yield better quality crystals. For example, the use of fresh microseed crystals, for immediate use rather than following storage at -80°C is likely to improve crystal growth and increase the number of crystals obtained. In order to improve crystal packing, applying dehydration techniques to existing crystals is likely to improve the diffraction potential and reduce anisotropy (as reviewed in (Krausse et al. 2012)).

The likelihood that Sam68 is able to be effectively crystallised is increased due to the fact that a homologous STAR protein, T-STAR, is readily crystallisable. As well as being useful as a good quality source of microseed stock for Sam68 crystallisation trials, the crystal structures of T-STAR NKKH and KH also provided a template for structural modelling of Sam68.

Several structural models of Sam68 were generated using the software Modeller. These were based on the structures of Gld-1 and T-STAR crystal structures in complex with RNA. Verification of these structures was investigated using SAXS data obtained for the STAR and NKKH domains of Sam68. Fitting the structural models within SAXS envelopes of these domains can be assessed by eye and quantified using the ATSAS suite. It was clear that the Gld-1 model of Sam68 did not fit the envelope as well as the T-STAR model. There are several other indications that Sam68 has a different conformation than other STAR subfamily members. In the Gld-1 STAR domain structure, the position of each subdomain is stabilised by contacts between the NK and CK domains and contacts between the NK-KH flexible linker and KH domain. These contacts are unlikely to occur in the same way in Sam68, due to differences in amino acids at the interface of these interactions, and from our NMR data showing that the NK peaks within the spectra of the NKKH and STAR domains are in the same position. Additionally, in terms of RNA binding, the CK is involved in recognition of RNA by Gld-1, however that does not seem to be the case for Sam68. Additionally, the distance between the two RNA binding sites in the Gld-1 dimer suggests a 10 nucleotide link between two protein binding sites on the RNA, which is different to the SELEX data for Sam68 and our fluorescence polarisation data.

Therefore the model of the NKKH domain based on T-STAR is more suitable, and is supported by CRYSOL analysis of the model and SAXS envelope. Since the NK-KH linker has been shown by Heteronuclear NOE to be flexible, it is possible that the orientation of these two subdomains is not fixed. Therefore an improved model was generated based on manipulation of the NK solution structure and T-STAR KH model of Sam68 within the NKKH SAXS envelope.

There are several methods by which the relative orientation of the NK and KH domains within the dimer can be determined. Firstly, the SAXS data obtained for the NKKH and STAR domains can be analysed using ensemble optimisation method (EOM) to quantitatively characterise flexible proteins in solution. This method accounts for several conformations of different domains that contribute to the scattering pattern to determine an ensemble of optimal conformers (Bernado, P. et al. 2007). In addition, orientational information between two protein domains can be acquired by measuring residual dipolar couplings (RDCs). This data is recorded using specific NMR experiments on a protein sample in the presence of an alignment medium, such as Pf1 phages, and gives long range angular information between atoms of different domains relative to the external magnetic field (Fischer et al 1999). This method is particularly powerful for refinement of global and local structure and would provide strong evidence for the quality of the Sam68 homology model.

An interesting feature of the T-STAR NKKH structure is that the KH itself forms a dimer. There are several KH domain containing proteins that have been reported to have self-association properties within this domain (as reviewed in (Valverde, Edwards & Regan 2008)). However, in several cases altering the protein construct and conditions resulted in alterations in crystal packing and a monomeric KH domain suggesting this observed dimerisation is an artefact of crystallisation. Several biochemical studies have been undertaken on KH domain containing proteins to determine whether this domain dimerises in solution. It was estimated that the KH3 domain of Nova-2, which dimerises via the KH domain in crystals (Lewis et al. 1999), self-associates with a Kd of approximately 300μM based on limited equilibrium ultracentrifugation experiments. However, in these experiments only 10-20% dimer is present (Ramos et al. 2002).

Therefore to verify that the KH undergoes dimerisation within the NKKH domain of Sam68, an Y241E mutant of the KH, NKKH and STAR domains were generated. This residue lies at the interface and is likely to be crucial for dimerisation. The stability of KH Y241E was similar to the WT, however protein expression and solubility of the NKKH and STAR mutants was reduced. This already suggests that this mutation has had a significant effect on the structure of these domains. A sample of NKKH Y241E could be produced of sufficient quality for NMR, and showed that the NK domain is not affected by this mutation. Many peaks corresponding to residues of the KH domain disappeared or shifted, in some cases to positions of peaks observed in the KH WT spectrum. SAXS data showed that isolated Sam68 WT KH domain forms oligomers,

which would explain the tendency for this construct to aggregate in solution. The addition of the NK domain may facilitate the stable dimerisation of the KH domain.

With a good quality structural model of the NKKH domain of Sam68, it was then possible to investigate the mechanism and specificity of RNA binding. NMR studies described in Chapter 3 identified that the KH, NKKH and STAR domains of Sam68 bind to AU-rich RNA of 4-13mer with similar binding affinity and only residues of the KH domain are affected. Therefore FP was used to estimate the binding affinity with greater accuracy to determine the intricacies of RNA recognition of Sam68. Having determined that the NK domain is required for correct dimerisation of the KH domains, and that the CK domain is not involved in RNA binding, the NKKH construct was used to obtain a 5mer consensus RNA binding sequence of (A/U/C)-(**A**/U)-A-A-(A/U). This sequence satisfies the binding motif identified by SELEX experiments (Figure 3.2) and is similar but distinct from the consensus sequences of other STAR proteins, such as Gld-1, QKI and SF1.

Longer RNAs were also tested to determine the cooperativity of binding between the two KH domains of the dimer. It was found that for the NKKH and STAR domains, a non-specific linker of greater than 15 nucleotides between UAAA motifs is required for additive binding of these domains. This was not observed for the Y241E NKKH mutant, showing that the two RNA binding domains are disrupted by this mutation. Interestingly, the affinity of the NKKH Y241E sample increased significantly between UAAACCC and UAAACCCCCUAAA with Kds of 27.17μM and 4.582μM, respectively. There was not a significant difference between these two RNAs for the WT construct. The fact that the Y241E mutant binds UAAACCCCCUAAA with higher affinity, and that this is consistent for the RNAs with longer cytosine linkers suggests that both KH domains within the dimer may bind to a single RNA. This is different to the WT constructs, for which an increase in binding affinity is not observed until the cytosine linker is increased to 20 nucleotides in length. This suggests that the KH domains within the NKKH Y241E sample are more flexible and the RNA binding pockets can be orientated closer together to accommodate this short bipartite RNA.

These data supports the notion that this mutation has had a significant effect on the structure of the NKKH domain, by interrupting the dimerisation of the KH domain. It also seems that the NK domain is required to facilitate the dimerisation of the KH domain, and for stabilisation of this KH dimer, but that the CK domain has no effect on

the structure. It has also been useful in understanding the specific function of Sam68 in alternative splicing.

These data strongly indicate that the dimerisation of the KH domain exists in solution, and is not an artefact of crystallisation of the STAR domain of T-STAR. In order to directly confirm this dimerisation, SEC-MALS could be recorded for the KH and KH Y241E constructs to determine an accurate molecular weight of each. Since the KH WT is not very stable, extension of this domain to include the C-terminal half of the NK-KH linker could improve stability and reduce the risk of the sample precipitating on the column. Further evidence for dimerisation could be obtained using paramagnetic relaxation enhancement (PREs) (Battiste and Wagner 2000). This NMR technique measures long range distance information using paramagnetic probes introduced to specific sites on a protein. This is particularly useful for studies of heterodimers, however in this situation measurements of the NKKH homodimer would be problematic. Therefore solvent PREs could be used, in which a soluble paramagnetic agent is titrated into the sample and associated with the solvent accessible regions of the protein (Madl et al. 2011). This results in line broadening and loss of peaks corresponding to the solvent exposed residues, revealing those at the dimer interface. These measurements provide translational information complimentary to the orientational data provided by RDCs, and a combination of these techniques could offer direct evidence for the dimerisation of the KH domain and relative orientation of the NK and KH domains.

There are many known gene targets of Sam68 (Figure 1.5), several of which have been identified to be enriched in Sam68 binding motifs, situated approximately 200 nucleotides upstream and downstream of target exons (Chawla et al. 2009). In addition, the influence of Sam68 on splicing outcome has been described for several different pre-mRNAs. For example, Sam68 induces skipping of exon 8 from the sgce gene and two Sam68 binding sites have been identified, upstream and downstream of this exon. Our data suggests that the Sam68 STAR domain dimer requires a distance of at least 20 nucleotides between each RNA consensus motif, but that this distance can be increased, such that Sam68 brings together these two sites in close proximity to facilitate “looping out” of the alternative exon. This may also be the case for exon inclusion, for instance in the case of SRSF1, Sam68 is known to promote the inclusion of exon 5 (Valacca et al. 2010). Once more, two Sam68 binding sites have been identified, one close to the exon 5 3’ splice site and the other close to the exon 4 5’ splice site. These regions are

separated by approximately 600 nucleotides, and therefore the two KH domains within the Sam68 dimer may each recognise one site and bring the two exons together for inclusion (Valacca et al. 2010).

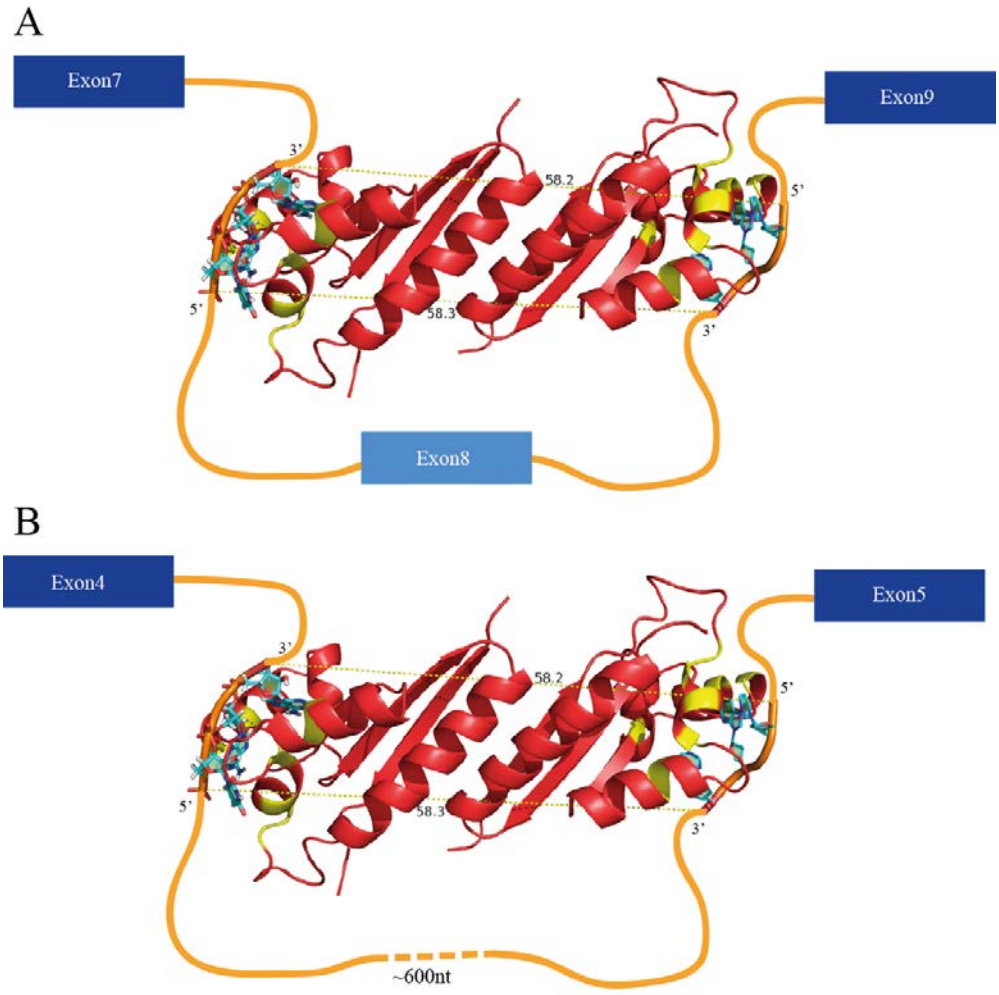


Figure 4.30: Physiological targets of Sam68
Structural models of Sam68 KH interaction with Sgce (A) and SRSF1 (B) pre-mRNAs suggesting that Sam68 functions to loop out regions of exonic and/or intronic RNA.

To summarise, we propose that Sam68 functions in alternative splicing by bringing two distant UAAA motifs together in close proximity and that depending on the location of these sites with respect to splice sites and splice regulatory regions, exon inclusion or exclusion can be induced.

Sam68 alternative splicing function is also well established to be regulated by various post-translational modifications. Having established the mechanisms by which Sam68 recognises RNA, the next chapter will investigate the effect of serine and threonine phosphorylation on the function of this STAR protein.

Chapter 5. Effect of serine/threonine phosphorylation on Sam68 RNA recognition

5.1. Introduction

Particular attention has been drawn to STAR proteins due to their potential function as a regulatory link between signal transduction pathways and alternative splicing. As mentioned in Chapter 1, Sam68 is subject to a variety of different post-translational modifications that alter its intracellular localisation and function. In Chapters 3 and 4 of this thesis, the emphasis has been on understanding how this STAR protein recognises its RNA targets; we will now shift the focus onto investigating how Sam68 RNA binding is affected by post-translational modifications.

This chapter will discuss the techniques used to determine the domains of Sam68 phosphorylated by Cdc2, Erk1 and Nek2, and the particular residues of the STAR domain found to be phosphorylated by Nek2. The implications of these modifications on RNA binding will be investigated using the data acquired in the previous two chapters by fluorescence polarisation.

5.2. Results

Serine/threonine phosphorylation of Sam68 has been shown to affect its function in splicing. Indeed, this post-translational modification appears to enhance the activity of Sam68 and it is likely that this subsequently impacts the development of cancer (Matter, Herrlich & Konig 2002). However, the mechanism by which these phosphorylation events affect Sam68 function is not yet known. It is possible that this effect on alternative splicing function is achieved through alterations in Sam68 RNA recognition and/or through changes in protein-protein interactions.

It has not yet been investigated which serine or threonine residues of Sam68 are phosphorylated by Cdc2 and Nek2 to induce these functional changes. Phosphorylation by ERK1 and the subsequent effects on murine Sam68 splicing function have been investigated more thoroughly. Matter et al identified eight potential ERK1 target sequences based on kinase consensus sequences, five serine/threonine residues in the N-terminal region and three at the C-terminus of the protein (Matter, Herrlich & Konig 2002). Replacement of all eight residues with alanine inhibited Sam68 mediation of exon inclusion following phorbol-ester stimulation of the MAPK pathway. Systematic

mutation of different groups of residues identified S58, T71 and T84 as the predominant ERK1 phosphorylation sites, however residual splicing was still observed using this assay which suggests that either endogenous WT Sam68 was exerting some effect, or that additional ERK1 targets exist. In any case, residues within the STAR domain of Sam68 were not investigated, and additionally RNA/protein interactions were not specifically explored, rather the overall outcome of splicing through this particular pathway. Given that mutated residues lie outside of the RNA binding domain, it is likely that the changes in Sam68 function are through alterations in protein-protein interactions.

There are twelve serine and threonine residues within the STAR domain of Sam68; S113, S117 T119, T126, S137 and S150 within the NK domain, T183, T191, S196 and S202 within the KH domain and S272 and S181 within the CK domain (Figure 5.1). We therefore postulated that phosphorylation at these sites by ERK1, Cdc2 and/or Nek2 might affect RNA interaction.

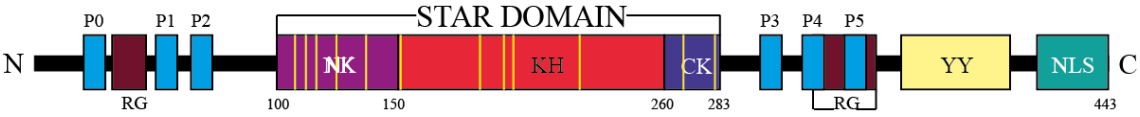


Figure 5.1: Serine/threonine residues if Sam68 STAR domain
Domain structure of full length Sam68, with serine and threonine residue positions within the STAR domain highlighted in yellow.

A consensus sequence has been identified for each of these kinases, as follows; L-X-X-(S/T)-ϕ for Nek2 (where X represents any residue and ϕ a hydrophobic residue) (Hardy et al. 2014), (K/R)-(S/T)-P-X-(K/R) for Cdc2 (Nigg 1991) and P-X-X-(S/T)-P for ERK1 (Clark-Lewis, Sanghera & Pelech 1991). There are two sites within Sam68 that conform to these target sequences; S117 for Nek2 and T84 for Erk1. These consensus sequences are not strict, however, and do not discount other residues from being potential kinase targets. Therefore initially the domains that undergo phosphorylation by these kinases were identified using *in vitro* kinase assays.

5.2.2. ^{32}P-ATP radiolabelled kinase assays

In order to identify whether the STAR domain of Sam68 is targeted by these kinases, a ^{32}P radiolabelled ATP assay was used (section 2.12) (Hastie, McLauchlan & Cohen 2006). This protocol involves incubating the purified protein construct of interest with the selected kinase in a reaction buffer that contains ^{32}P-ATP. It follows that if residues

within the construct are phosphorylated, ^{32}P will become incorporated into them and a band will be visible on X-ray film that is exposed to these samples on an SDS-PAGE gel. This assay was used to determine the phosphorylation of the STAR, KH, N- and C-terminal constructs of Sam68 with Cdc2, ERK1 and Nek2. The N- and C-terminal domains of Sam68 did not express well and could not be purified in sufficient quantity for these experiments.

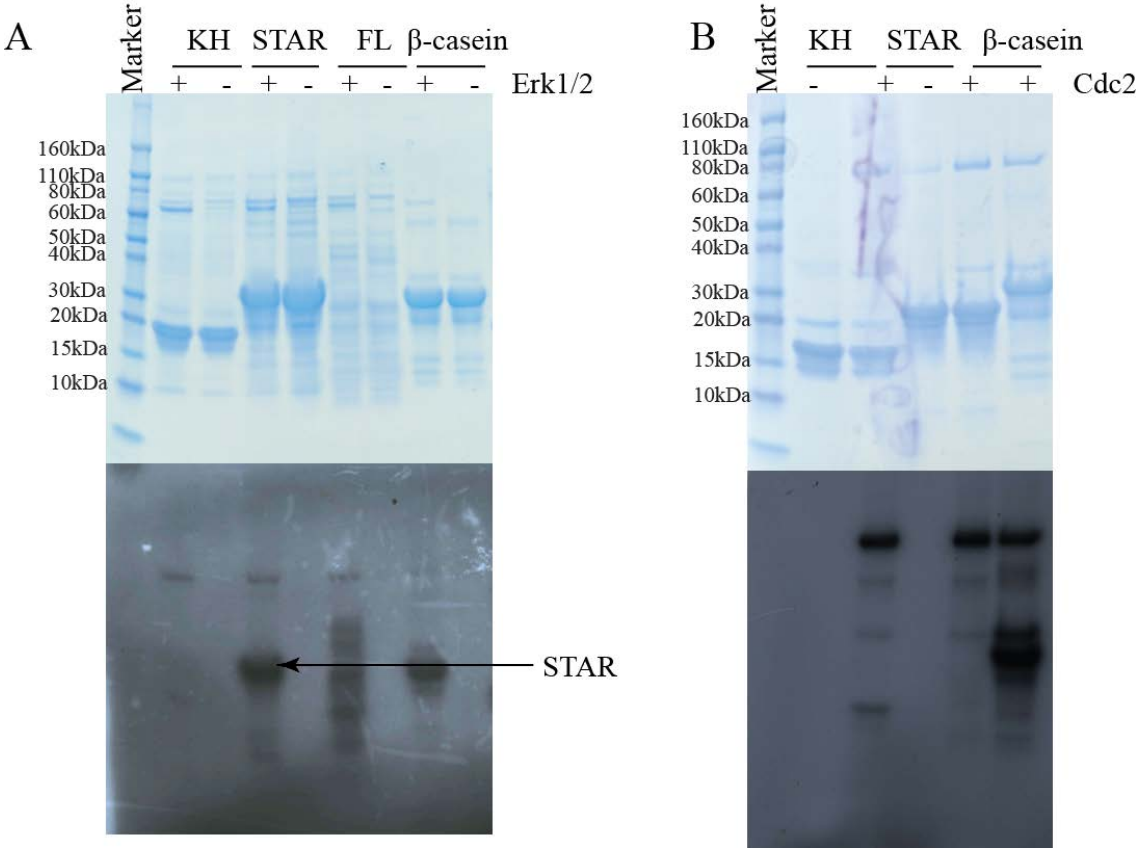


Figure 5.2: Phosphorylation of Sam68 by Erk1 and Cdc2
SDS-PAGE gel (upper panel) and x-ray exposure images (lower panel) of ^{32}P ATP radiolabelled kinase assay of Sam68 KH, STAR and full length with Erk1 (A) and Sam68 KH and STAR with Cdc2 (B). Positive controls were conducted using β-casein as a substrate and negative controls lacking active kinase were run as indicated.

Our results show that the STAR but not the KH domain of Sam68 is phosphorylated by ERK1 but both the STAR and KH domains are phosphorylated by Nek2 (Figures 5.2A and 5.3). Results with Cdc2 were inconclusive, with a small band visible for the KH domain and not for the STAR domain (Figure 5.2B). Therefore, as the only kinase to target the KH RNA binding domain, Nek2 was focussed on in subsequent experiments.

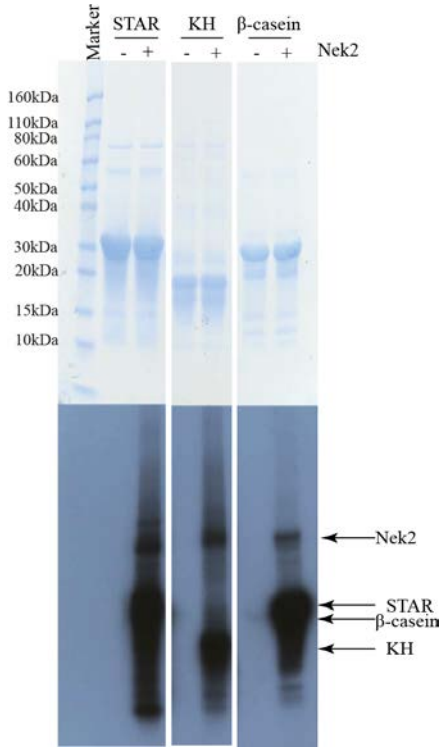


Figure 5.3: Phosphorylation of Sam68 by Nek2
SDS-PAGE gel (upper panel) and x-ray exposure images (lower panel) of ^{32}P ATP radiolabelled kinase assay of Sam68 KH and STAR domains with Nek2. Positive controls were conducted using β-casein as a substrate and negative controls lacking active kinase were run as indicated.

5.2.3. Co-localisation of Sam68 and Nek2 in cancer and non-cancer derived cell lines

Nek2 is well documented as a modulator of centrosomal activity, but has also been proposed to be localised in the nucleus of cancer cell lines and to associate with various splicing factors including Sam68 (Naro et al. 2013). We therefore investigated the localisation of these proteins in several cancer and non-cancer derived cell lines by immunofluorescence microscopy in order to determine whether these two proteins localise in the same cellular compartment. Experiments were carried out under the guidance of Dr Laura O'Regan and Professor Andrew Fry.

Since Nek2 is primarily a centrosomal protein, control experiments were carried out on MCF7 (breast cancer), HBL100 (breast non-cancer), PC3 (prostate cancer), PNCT (prostate non-cancer) and HeLa (cervical cancer) cell lines to determine the localisation of Nek2 with respect to the nuclei (by staining with Hoescht dye to highlight the DNA) and γ-Tubulin, a centrosomal marker (Figure 5.4A). These images demonstrate that these two proteins are concentrated in one or two "bright spots" per cell which correspond to the centrosomes.

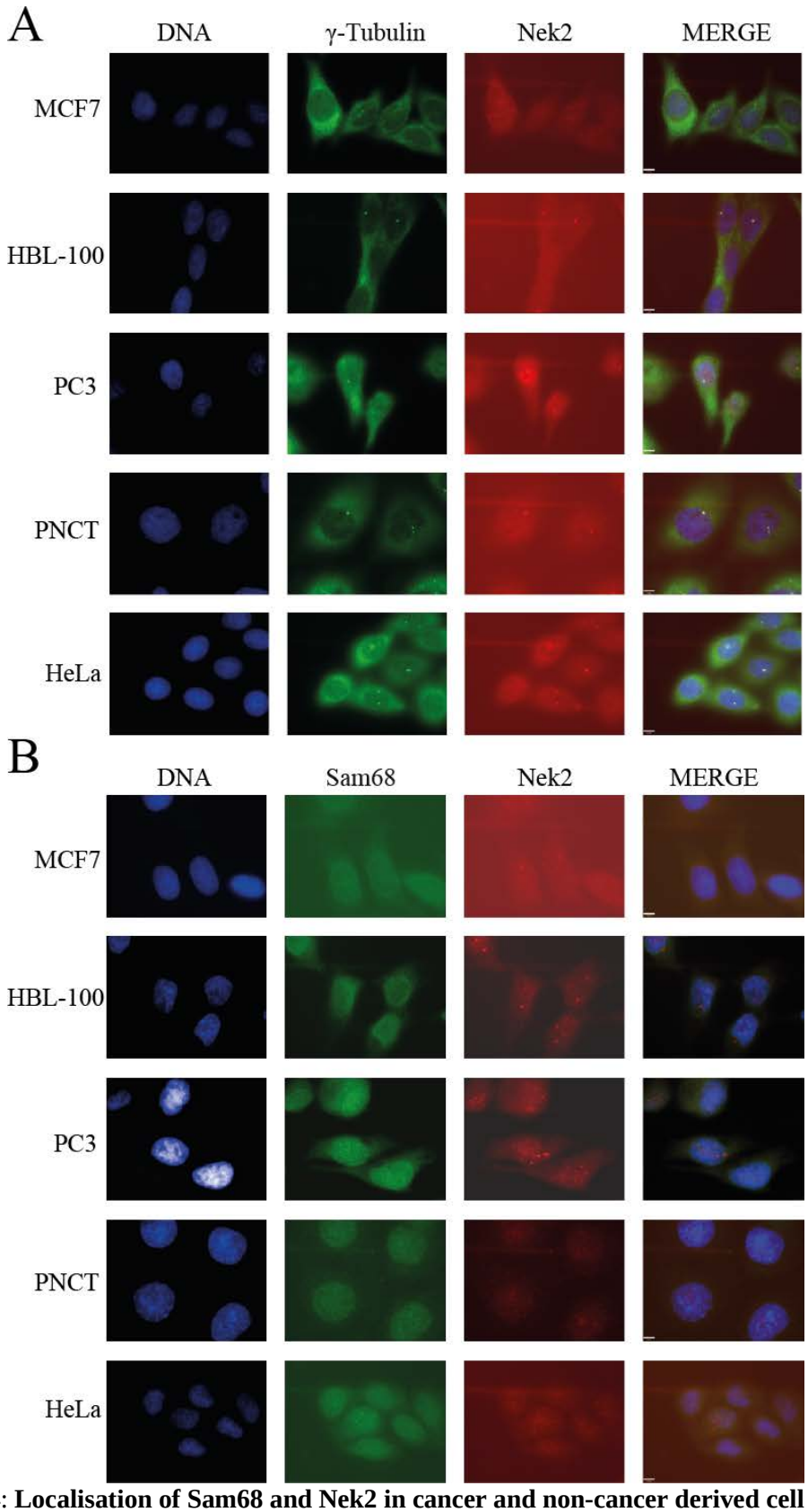


Figure 5.4: **Localisation of Sam68 and Nek2 in cancer and non-cancer derived cell lines** Immunofluorescence microscopy of five different cell lines to show the localisation of DNA, γ-tubulin (centrosomes) and Nek2 (A), and DNA, Sam68 and Nek2 (B).

They are each diffused within the cytoplasm and nucleus, and in speckles within the nucleus. This confirms that Nek2 is present at the centrosomes as expected but is also located in other cellular compartments, most importantly the nucleus.

Immunofluorescence microscopy was then carried out on the same cell lines with Sam68 and Nek2 antibodies in addition to Hoescht dye to identify the localisation of the nuclei (Figure 5.4B). We observed that both Nek2 and Sam68 are diffused in the nucleus and cytoplasm, but more concentrated in the nucleus. They both seem to concentrate within particular spots in the nucleus, although these do not appear to be the same speckles for each protein. These data suggest that in these cell lines, Nek2 and Sam68 are present in the same cellular compartments and it is therefore possible for Sam68 to interact with and be phosphorylated by Nek2.

5.2.4. Mass Spectrometry

Having demonstrated that Nek2 phosphorylates the KH and STAR domains of Sam68 and that this interaction is physiologically possible, we then set about locating the particular residues within these domains that are targeted by Nek2.

The primary method used was mass spectrometry analysis of cold samples that were produced in tandem to the ^{32}P-labelled samples during the *in vitro* kinase assays. Mass spectrometry was carried out on the STAR domain of Sam68 using the Protein Nucleic Acid Chemistry Laboratory (PNACL, University of Leicester) and revealed two sites of phosphorylation within the KH domain of Sam68 that were targeted by Nek2: T183 and S196.

To determine whether these are the only two sites of phosphorylation, a double mutant of the STAR domain and KH domain were expressed, replacing both of these residues with an alanine. The KH double mutant did not express (Figure 5.5); however the STAR double mutant construct could be expressed and purified in sufficient quantity. Phosphorylation assays showed that this double mutant is still phosphorylated by Nek2, suggesting the presence of other sites within the STAR domain that are targeted by this kinase (Figure 5.5).

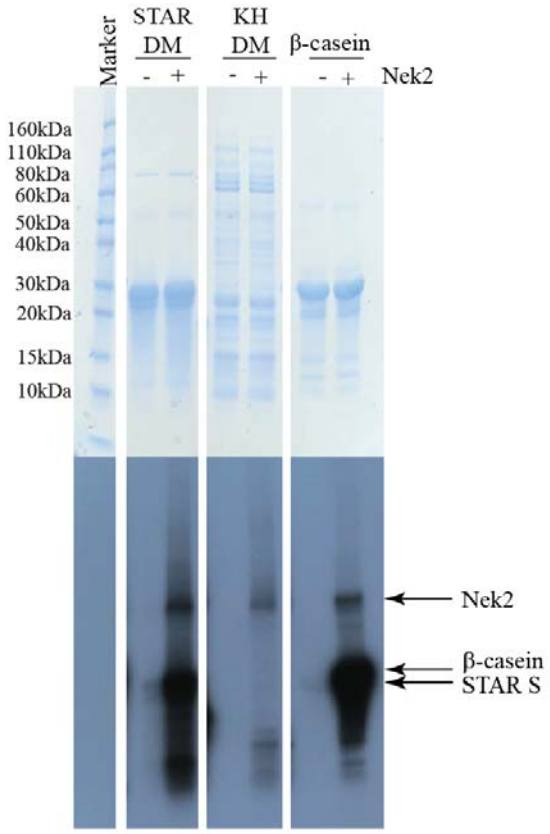


Figure 5.5: Phosphorylation of Sam68 serine/threonine mutants by Nek2
SDS-PAGE gel (upper panel) and x-ray exposure images (lower panel) of ^{32}P ATP radiolabelled kinase assay of Sam68 KH and STAR double mutants (T183A and S196A) with Nek2. Positive controls were conducted using β-casein as a substrate and negative controls lacking active kinase were run as indicated.

5.2.5. Kinase Assay by NMR spectroscopy

Although quantification of relative phosphorylation levels of each construct is possible using scintillation counting, it was decided to use NMR as an alternative technique to identify further sites of phosphorylation on Sam68. It has been shown recently that NMR can be used to follow the phosphorylation of proteins by serine/threonine kinases, as reviewed in (Theillet et al. 2013). These experiments are similar in principle to the chemical shift perturbation experiments used to investigate protein-RNA interactions (Chapter 3). Given that the chemical shift of a particular amino acid in a (^{1}H-^{15}N) HSQC spectrum is dependent on the local chemical environment of its own backbone amide, if this is altered significantly then a change in chemical shift will occur. As well as during interaction with a ligand, this shift also occurs upon covalent attachment of a phosphate group to the side chain of a serine residue for example, since the local chemical environment of atomic nuclei in the vicinity is altered. This therefore gives

information at amino acid resolution and theoretically one is able to distinguish between multiple phosphorylation events on closely spaced sites.

As with the chemical shift perturbation experiments described previously, a (^{1}H-^{15}N) HSQC experiment is recorded of the ^{15}N-labelled protein of interest. This is followed by addition of ATP and $MgCl_2$ and recording of a reference spectrum before addition of the kinase of interest. HSQC experiments are then recorded over time. This selective isotope labelling strategy allows detection of the substrate only, and continuous monitoring of affected residues in a time-resolved manner since NMR is a non-destructive technique. Furthermore, since both phosphorylated and unmodified species of each residue are visualised simultaneously, phosphorylation events can be quantified over time in a site specific manner by analysis of changes in crosspeak intensities. There are therefore several inherent properties of NMR spectroscopy that means this technique really lends itself to investigation of post-translational modifications.

The advantages of this technique compared to the ^{32}P labelled assay is that it is possible to acquire information at the residue level in a time-dependent manner, allowing calculation of kinetic parameters. Furthermore, once optimised, it is possible that following a kinase assay by NMR, the same sample may be used to investigate RNA binding using the chemical shift perturbation experiments described previously.

5.2.5.1. Optimisation of NMR kinase assay

Since most of the backbone assignment of the STAR domain is available (Chapter 3), in theory it is possible to identify particular residues that are affected by incubation with Nek2 using NMR spectroscopy.

Initially, these experiments were carried out during the optimisation process of the STAR domain of Sam68 (Chapter 3) and therefore were recorded on proteins purified in phosphate buffer and at 20°C. In such conditions, no chemical shift differences were observed following incubation with Nek2 for several days. Since these conditions vary slightly from those used in the *in vitro* ^{32}P-ATP assay, Nek2 activity was tested using the buffer used for NMR experiments at 20°C (Figure 5.6). The phosphorylation of the STAR and KH domains of Sam68 by Nek2 was tested in Tris and phosphate buffer, and showed that Nek2 activity is decreased in phosphate buffer. Furthermore, the level of phosphorylation of the KH and STAR domains in Tris buffer was reduced at 25°C, suggesting that Nek2 activity is lower at this temperature. These data indicated the necessity to optimise an NMR sample in Tris buffer that was stable up to at least 25°C.

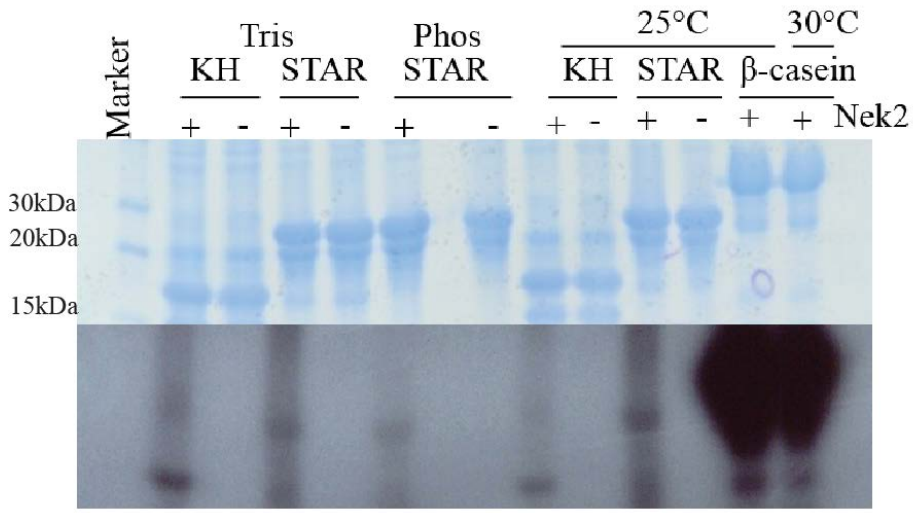


Figure 5.6: **Optimisation of NMR kinase assay sample**
SDS-PAGE gel (upper panel) and x-ray exposure images (lower panel) of ^{32}P radiolabelled kinase assay of Sam68 KH and STAR domains in Phosphate and Tris buffer, and at 25°C and 30°C. Positive controls were conducted using β-casein as a substrate and negative controls lacking active kinase were run as indicated.

It has also previously been established that members of the Nek kinase family are more active in reactions containing $MnCl_2$ rather than $MgCl_2$, (personal communication with Professor Andrew Fry). The levels of phosphorylation of the STAR domain were determined in the presence of each (Figure 5.7), demonstrating that indeed, Nek2 appears to phosphorylate the STAR domain less efficiently in $MgCl_2$ compared to $MnCl_2$. We therefore further optimised the STAR domain sample to accommodate these conditions and observed that it is very stable in Tris buffer at pH7 at 30°C. However, given that manganese is paramagnetic it induces a line broadening effect and renders the sample invisible by NMR. Therefore, since Nek2 is still active in a reaction mix containing $MgCl_2$, it was deemed sufficient to maintain this in the NMR sample.

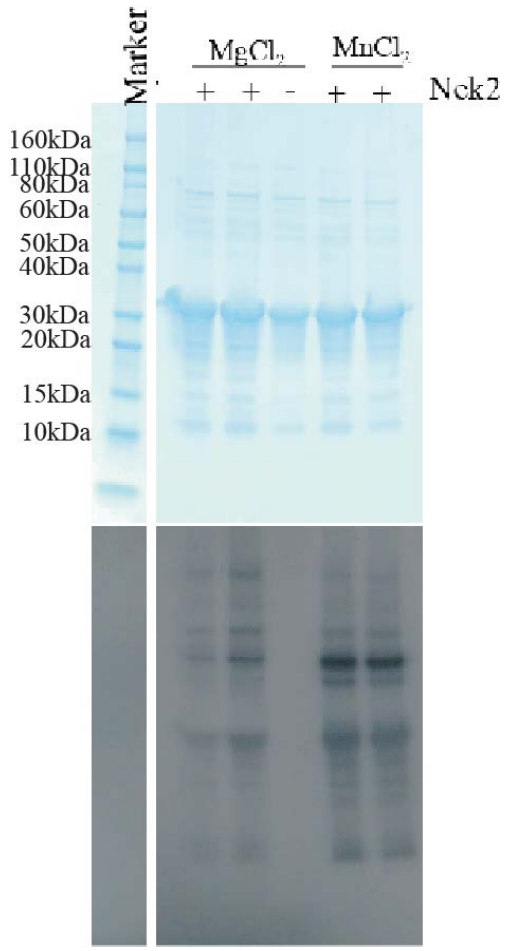


Figure 5.7: The effect of $MgCl_2$ vs $MnCl_2$ on Nek2 kinase activity
SDS-PAGE gel (upper panel) and x-ray exposure images (lower panel) of ^{32}P radiolabelled kinase assay of Sam68 STAR domains with Nek2 in the presence of $MgCl_2$ or $MnCl_2$. Negative controls lacking active kinase were run as indicated.

As described in Chapter 3, it was possible to obtain a sample of the STAR domain in Tris buffer that was stable up to 30°C, resulting in a suitable quality HSQC for phosphorylation assays with Nek2 (Figure 5.8). This spectrum demonstrates that the addition of $MgCl_2$ and ATP (red) did not cause any chemical shift perturbations and therefore 2.9µg of Nek2 was added to the sample and HSQC spectra were measured overnight. Comparison of these spectra over time showed changes in the position of several crosspeaks, indicating that phosphorylation had occurred (Figure 5.9).

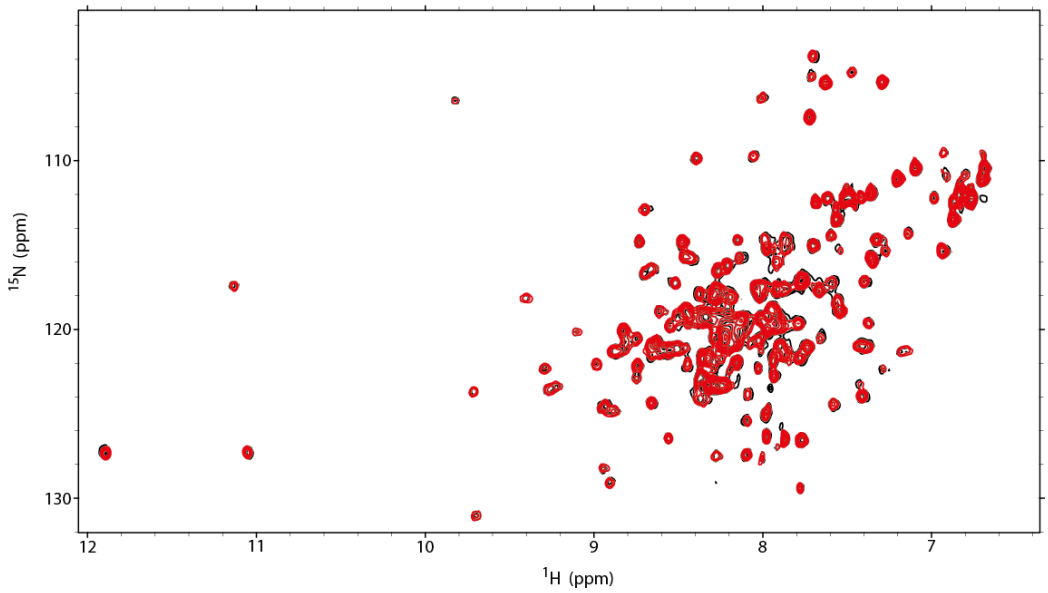


Figure 5.8: Sam68 STAR domain sample optimisation for NMR kinase assay
Overlay of (^{1}H-^{15}N) HSQC spectra of Sam68 STAR domain free (black) and with a final concentration of 5mM $MgCl_2$ and 0.9mM ATP (red).

5.2.5.2. Identification of phosphorylated residues by NMR

Three residues were clearly undergoing chemical shift perturbations over 44 hours, H120, A121 and R159 (Figure 5.9A and B). It is unlikely that any of these are being directly phosphorylated by Nek2, however it is possible that they are affected by a local phosphorylation event or conformational change in protein structure as a result of phosphorylation, as H120 and A121 are located at the NK dimerisation interface (FIGURE 5.9D). Interestingly, H120 and A121 are adjacent in sequence to T119, which also experiences a chemical shift perturbation (FIGURE 5.9C), suggesting that it may be a phospho-target of Nek2. The quality of the final few spectra was poor, with loss of peak intensity, due to precipitation of the sample and loss of peak signals. This was a result of a pH change from 7 to 5. This is a consequence of the phosphorylation reaction as ATP undergoes hydrolysis to form ADP and Pi. This reaction uses a hydroxyl group on the side chain of the target serine or threonine, leaving the oxygen to form the new phosphate group from ATP, and releasing a proton which accumulates over time as more protein molecules are phosphorylated, altering the pH of the sample. Therefore it is unclear whether the chemical shift perturbations of H120, A121 and R159 are caused directly by a phosphorylation event, a change in structural conformation due to phosphorylation or a change in pH of the sample, particularly as the imidazole ring of histidine residues has a tendency to become protonated and therefore is sensitive to pH changes with regard to NMR chemical shift.

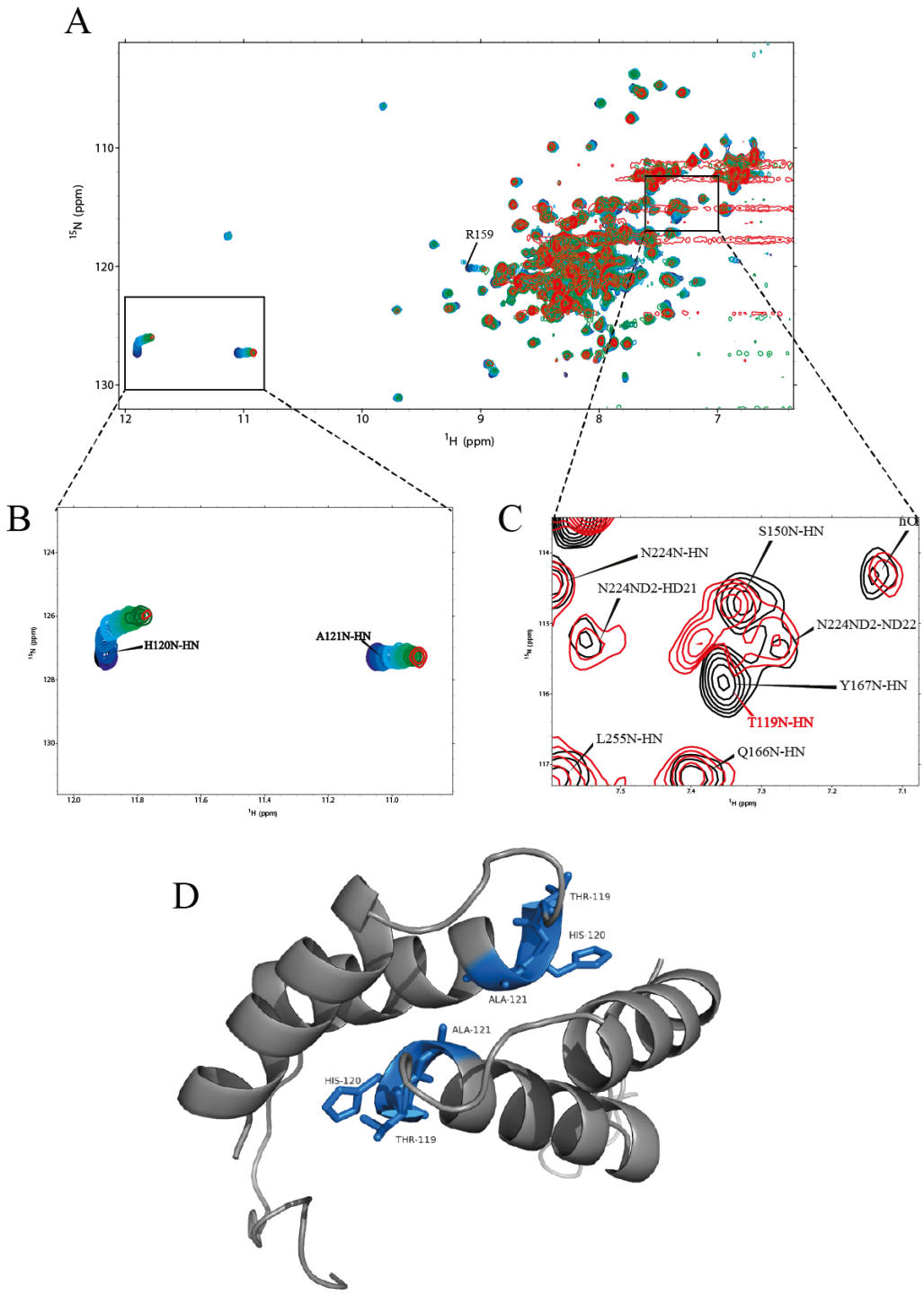


Figure 5.9: **NMR kinase assay of Sam68 STAR domain with Nek2**
Overlay of (^{1}H-^{15}N) HSQC of the STAR domain of Sam68 (black) with Nek2 every 4 hours over a total of 44 hours (shades of blue to red) (A) showed a chemical shift perturbation of the crosspeaks corresponding to H120 and A121 (B). T119 also changed position after 44 hours, suggesting it may be phosphorylated by Nek2 (C). Location of T119, H120 and A121 are highlighted in blue on the structure of the NK domain of Sam68 (D).

To investigate this further, a triple mutant of the STAR domain was produced, replacing T119 in addition to T183 and S196 as identified by mass spectrometry, with alanine residues. This triple mutant construct expressed well and could be purified to produce a

good quality (^{1}H-^{15}N) HSQC spectrum in the same conditions as the WT (Figure 5.10). Once again, the only clear crosspeaks to change over time were H120, A121 and R159. Once more, the pH dropped from 7 to 5 and a reduction in spectral quality was observed, indicating that some phosphorylation is still occurring. However, these changes took approximately 16 hours as opposed to 4 hours for the WT to occur, and the chemical shift difference was not as significant as for the WT. This suggests that at least some of the phosphorylation events have been inhibited as a result of these mutations.

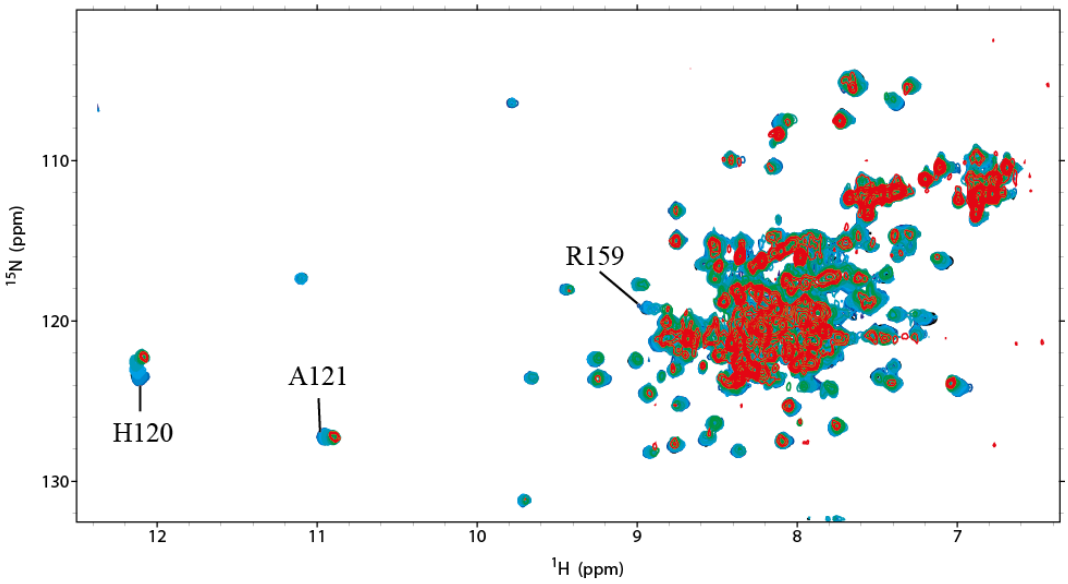


Figure 5.10: **NMR kinase assay of Sam68 STAR domain triple mutant with Nek2**
Overlay of (^{1}H-^{15}N) HSQC of the STAR domain triple mutant of Sam68 (T183A, S196A and T119A) (black) with Nek2 every 4 hours over a total of 44 hours (shades of blue to red).

Given that the HSQC of the STAR domain is relatively overlapped in the centre due to high intensity peaks of the CK domain, we performed a kinase assay on the NKKH sample with Nek2. During the experiment, the sample pH was measured and readjusted to 7 every 24 hours. After 5 days, at pH7 the peaks of H120 and A121 remained in their original position and the spectral quality was still sufficient for analysis (Figure 5.11). The most significant change observed was the disappearance of the peak corresponding to S137 after this time, which was unclear from the STAR domain spectra.

In order to further optimise these experiments and eradicate the need to constantly monitor the pH of the sample, the STAR and NKKH WTs were produced in 10mM HEPES buffer, with 100mM NaCl and β-mercaptoethanol as before at pH7, since this is a stronger buffer at this pH. However, despite producing a good quality HSQC, no changes were observed over time with Nek2 for either construct of Sam68. This is

surprising, since the main component of the ^{32}P kinase assay is HEPES buffer (Appendix 7.5), in which Nek2 has been shown to be active and phosphorylate Sam68. Therefore it may be that the proportion of Sam68 that is phosphorylated is not high enough to induce visible changes in the spectrum.

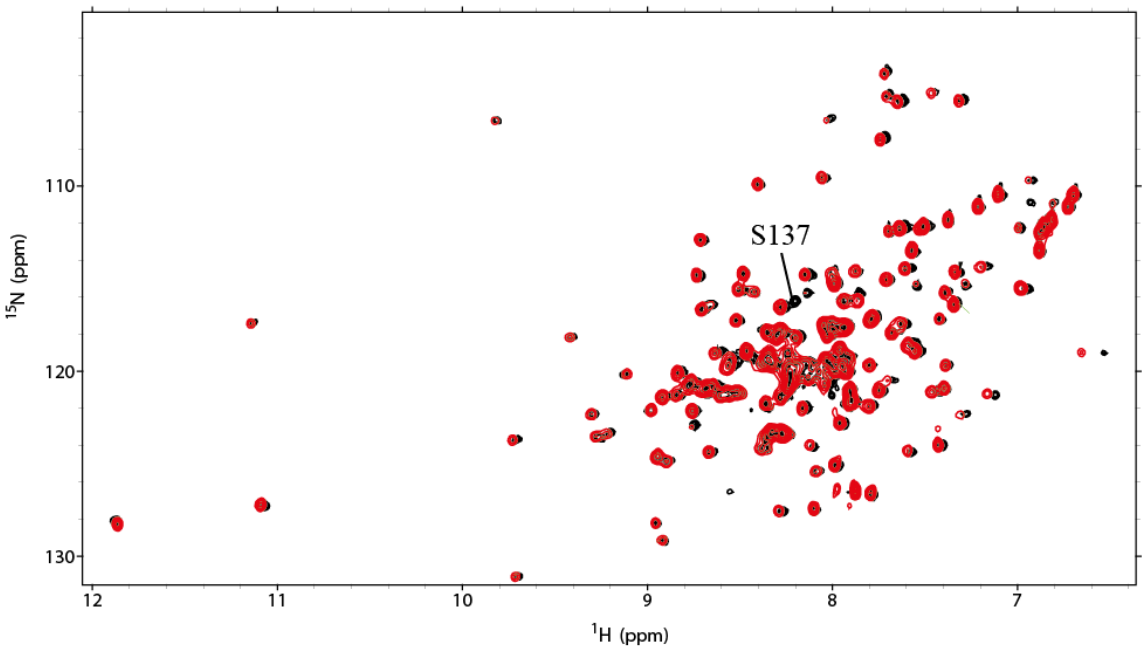


Figure 5.11: NMR kinase assay of Sam68 NKKH domain with Nek2
Overlay of (^{1}H-^{15}N) HSQC of the NKKH domain of Sam68 (black) with Nek2 after 5 days (red) with pH adjustment.

5.2.5.3. NMR kinase assay in nuclear extract

It is also possible to use NMR to investigate post-translational modifications and their effect on protein-structure and function in cells such as *E.coli*, Mammalian cells and Xenopus oocytes (as reviewed in (Lippens, Landrieu & Hanoulle 2008)). It has been shown that these events can also be observed in cellular extracts (Theillet et al. 2013). Since the phosphorylation of Sam68 by Nek2 most likely occurs in the nucleus, we investigated this post-translational event in nuclear extracts. In this case a reference (^{1}H-^{15}N) HSQC spectrum of the bacterially expressed, labelled protein was recorded, and then nuclear extract was added before recording a series of HSQC experiments. The nuclear extract should contain all components required for phosphorylation, provided that the kinase is present. Having determined that Nek2 and Sam68 are both present in the nucleus of HeLa cells (Figure 5.4), commercially available HeLa cell nuclear extracts were used in preliminary NMR experiments with Sam68 STAR and NKKH domains. These extracts were found to be splicing efficient within the laboratory with Sam68.

Initially the STAR domain of Sam68 was investigated in HEPES buffer, having shown that phosphorylation events in Tris buffer result in a pH drop that affects the chemical shift of several residues and spectral quality. A reference spectrum was recorded, followed by the addition of 20% nuclear extract, which is the standard protocol for *in vitro* splicing assays. The addition of nuclear extract did not alter the spectrum, and indeed after incubation at 30°C for 10 days, no chemical shift perturbations were observed (Figure 5.12A). This suggests that no phosphorylation has occurred. This is surprising given that the kinase reaction mix used for the ^{32}P radiolabelled kinase assays contains HEPEs buffer, suggesting that Nek2 should be active in these conditions, however the kinase composition of the nuclear extract is not known, and therefore it is possible that Nek2 is not present.

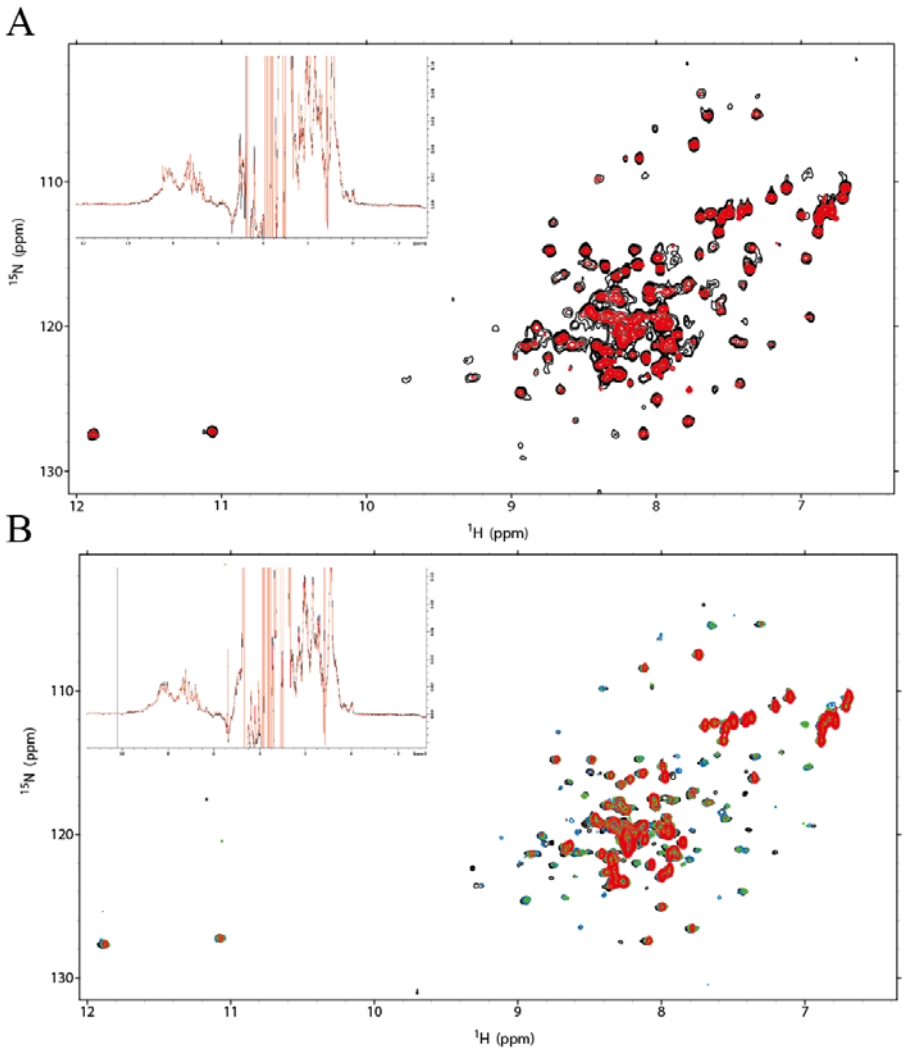


Figure 5.12: **NMR kinase assay of the STAR domain in HEPES buffer with nuclear extract** Overlay of (^{1}H-^{15}N) HSQC of the STAR domain of Sam68 in HEPES buffer before (black) and after 10 days in the presence of 20% nuclear extract (red) (A) and of the free STAR domain in HEPEs buffer (black), and in the presence of 20% nuclear extract, ATP, MgCl2 and Nek2 over 20 hours (blue to red) (B).

Having determined that HSQC spectra of Sam68 STAR domain can be measured in nuclear extracts, a new sample of the STAR domain in HEPES was prepared and ATP and $MgCl_2$ added along with 20% nuclear extract (Figure 1.12B). However, after 20 hours no changes were observed other than a decrease in intensity of some cross-peaks, indicating that endogenous kinases present in the nuclear extract are not sufficient to phosphorylate Sam68 STAR domain. This was not attributed to a loss in sample concentration by analysis of 1D spectra recorded before and after 20 hours.

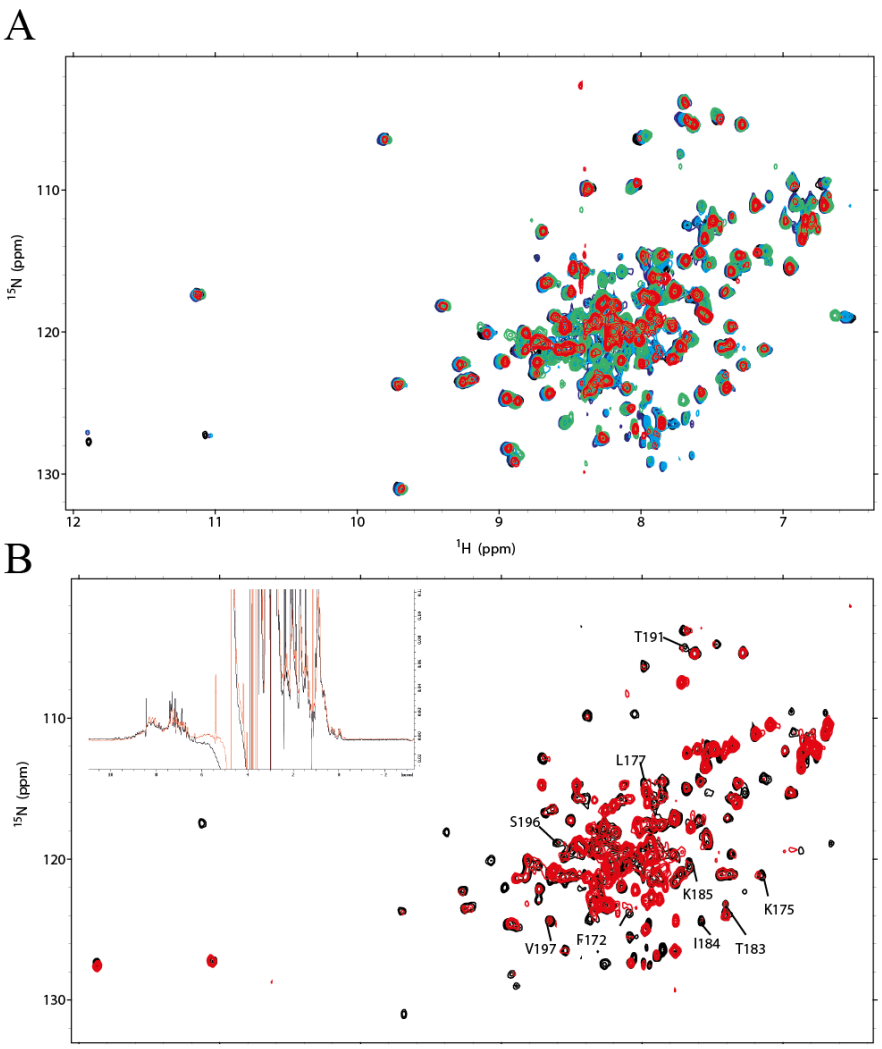


Figure 5.13: **NMR kinase assay of the STAR domain in Tris buffer with nuclear extract** Overlay of (^{1}H-^{15}N) HSQC of the NKKH domain of Sam68 in Tris buffer (black) and after several days with nuclear extract, ATP, MgCl2 and Nek2 (blue to red) (A), and of the STAR domain in Tris buffer with nuclear extract (black) and after addition of Nek2 (red) (B).

Having observed changes in spectra due to phosphorylation in Tris buffer, this experiment was repeated with the NKKH domain in Tris buffer at pH7 in nuclear extract (Figure 5.13A).

Spectra were recorded for several days. As a decrease in pH was observed over time, the pH was readjusted to 7 several times. This suggests that in Tris buffer with nuclear extract and supplemented with ATP and $MgCl_2$, some phosphorylation event is occurring that is altering the pH of the buffer. Given that the nuclear extract is likely to contain many enzymes in addition to serine/threonine kinases it is possible that this is due to other post-translational modification events.

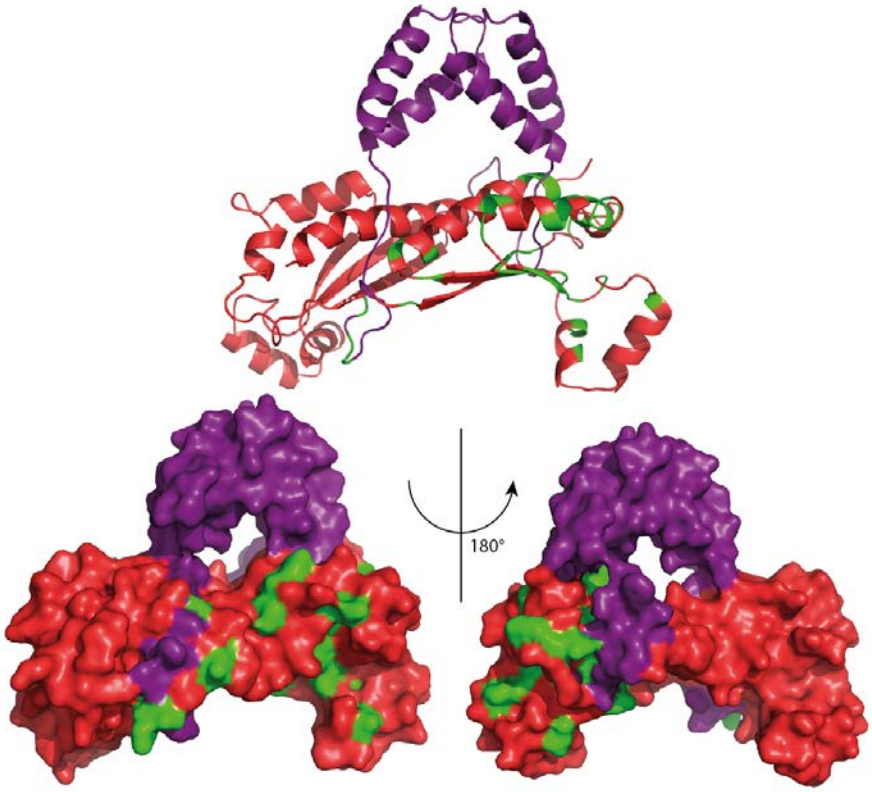


Figure 5.14: Analysis of NMR kinase assay in nuclear extract
Structural model of Sam68 NKKH (NK purple, KH red), in cartoon and surface representation at two different angles. Residues affected by incubation with nuclear extract and Nek2 are highlighted in green.

To determine if elements of the nuclear extract affect phosphorylation of Sam68 by Nek2, both nuclear extract and Nek2 were added to the STAR domain in Tris pH7 (Figure 5.13B). This resulted in the immediate loss of many peaks, without a loss in concentration of the protein as observed from the 1D spectrum. This suggests that the chemical environment of the residues corresponding to the disappearing peaks has been affected by the presence of Nek2 in nuclear extract, possibly by serine/threonine phosphorylation by this kinase or indeed other post-translational modifications. The affected crosspeaks correspond to many types of residue, not just serine or threonines, and it is therefore unclear exactly what has caused these perturbations. Of the five residues identified as potential targets of Nek2, the cross peaks corresponding to T183

and S196 both lost intensity, in addition to T191 suggesting that they may have been phosphorylated. Plotting all affected residues on the structural model of Sam68 NKKH it is clear that only the KH domain is affected by the presence of Nek2 in nuclear extract (Figure 5.14). Interestingly, many of the peaks that lose intensity, or disappear altogether, are those that correspond to residues involved in RNA binding (F172, K175, L177, I184, K185 and V197) and most others are in close proximity. This suggests that there are post-translational modification events occurring in this region that have the potential to alter RNA binding through changing the chemical properties of residues involved in RNA recognition and/or conformational changes to the structure of this region.

5.2.6. RNA binding of S/T mutants

Having identified residues T119, S137, T183, T191 and S196 as being potential targets of Nek2, the impact of phosphorylation of these sites on RNA binding was then investigated. Location of these residues on the model of Sam68 NKKH along with the RNA binding sites shows that several of these residues are in close proximity to the RNA, suggesting that phosphorylation at these sites might influence RNA recognition (Figure 5.15). In addition, S202 was highlighted as a potential site of phosphorylation by Nek2 using KinasePhos (Huang et al. 2005). Since this residue was also identified as being directly involved in RNA binding, this residue was also tested as a potential site of phosphorylation.

Fluorescence polarization experiments had already been optimised for the NKKH domain of Sam68 (section 4.5) to quantify RNA binding and highlight the residues that are essential for this recognition. Therefore this method was also used to determine the effect of phosphorylation on RNA binding. Each of the serine and threonine residues identified by NMR and the ^{32}P assay were mutated into a glutamic acid, as a mimic of phosphorylation.

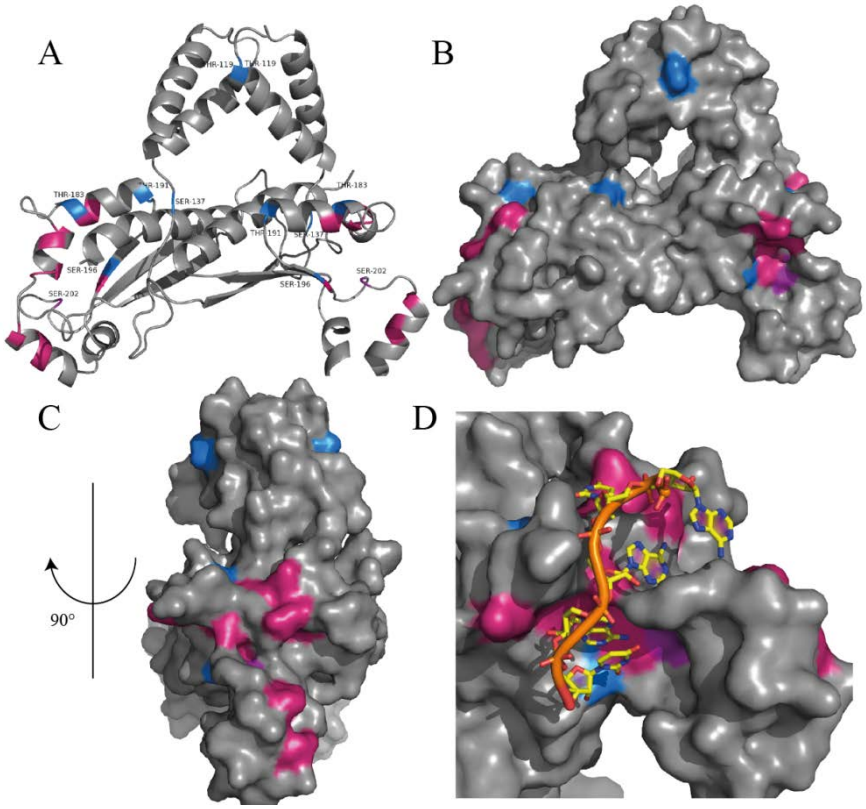


Figure 5.15: Sites of serine and threonine phosphorylation on Sam68 NKKH
Structural model of Sam68 NKKH with RNA binding sites highlighted in pink, serine and threonine target sites of Nek2 highlighted in blue, and S202 in purple. The structural model and labelled S/T residues are shown in cartoon (A) and surface representation (B), with focus on the RNA binding site at 90° to A/B (C), and a zoom of the RNA binding pocket with AUAAU (D).

T191E and S137E did not express well, which may indicate that these mutations affect the conformation of the protein, causing aggregation and poor expression. This is interesting, as S137 is located in the flexible linker between the NK and KH domains, suggesting that this region may be important for the stability of the dimer. T191 is located at the C-terminal end of helix 2 within the KH domain and may be essential for proper folding of this domain. Serine 202 was identified as being involved in RNA binding, and the alanine mutant was expressed, purified and tested by fluorescence polarisation (Figure 4.26A). The phosphomimetic mutant S202E however, was lost during gel filtration, suggesting that this amino acid substitution may also affect the stability of the NKKH domain.

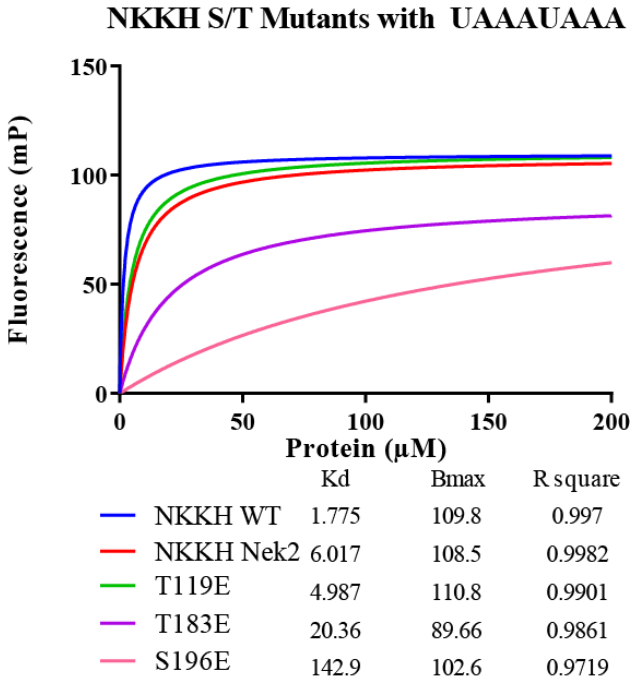


Figure 5.16: Fluorescence polarisation of Sam68 NKKH serine and threonine mutants with UAAAUAAA

Fluorescence plotted against NKKH mutant concentrations and calculated dissociation constants, Bmax and r^2 with 5mer RNA. NKKH WT, T119E, T1183E and S196E refer to each protein construct and NKKH Nek2 refers to WT sample after 5 days incubation with Nek2 at 30°C.

The remaining three mutants, T119E, T183E and S196E, could be expressed and purified to produce a sample for FP. Each construct, along with the WT, was titrated across the plate from 200 to 0µM, with 0.2µM UAAAUAAA so as to compare with the RNA binding mutants (Figure 5.16). Additionally, the NKKH WT was incubated with Nek2, ATP and $MgCl_2$ for five days at 30°C, and the pH adjusted back to pH 7 each day. T119E showed a small decrease in RNA binding affinity with a Kd of 4.987µM compared to the WT with a Kd of 1.775µM. T119 is located in the NK domain, far from the RNA binding sites and not at the NK dimer interface and therefore unlikely to have a significant effect on RNA binding or protein folding. The WT sample incubated with Nek2 for five days did not have a significant effect on RNA binding, despite a drop in pH from 7 to 5 each day, indicating that some phosphorylation event is occurring.

T183E and S196E mutations had a more significant effect on RNA binding, with a reduction in Kd to 20.36µM and 142.9µM, respectively. T183 is located between two RNA binding sites, K185 and I184, and in proximity to L177 and K175 (Figure 5.17A). It is possible that the addition of a negative charge to this region affects the structure of this binding pocket and inhibits RNA binding. S196 is located within a β-strand, adjacent to V197, which has been identified as a potential RNA binding site and

therefore phosphorylation of this residue may affect the interaction between V197 and the RNA (Figure 5.17B).

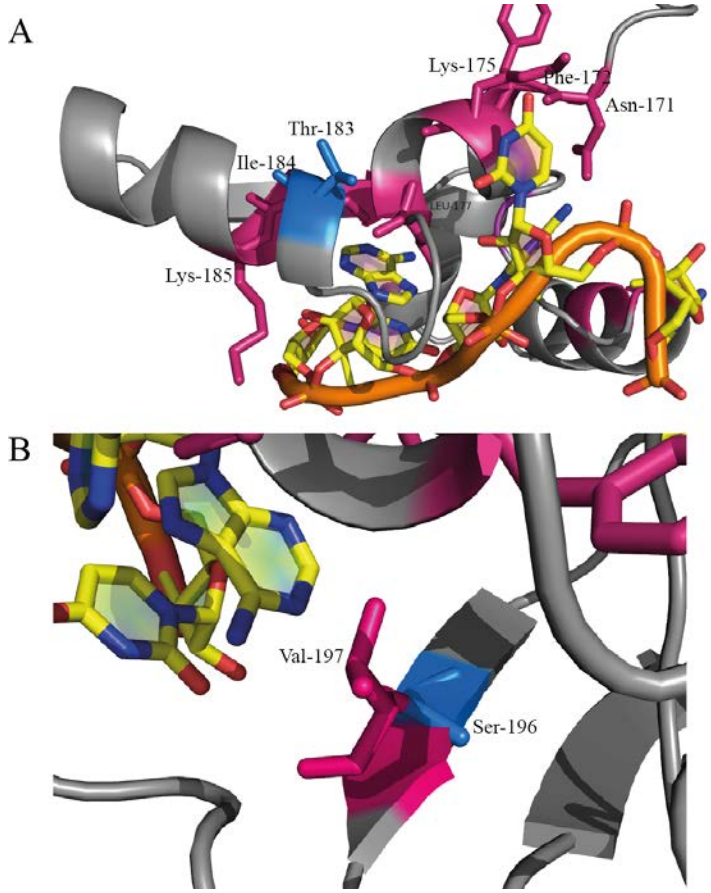


Figure 5.17: Effect of serine/threonine phosphorylation on Sam68 RNA binding
Structural model of Sam68 KH domain with AUAAU. Residues involved in RNA binding are highlighted in pink, potential sites of phosphorylation by Nek2 are highlighted in blue.

5.3. Discussion

Sam68 function is tightly regulated by various post-translational modifications; however these mechanisms are not yet well understood. Thus far, it has been established that amongst these modifications, serine/threonine phosphorylation of Sam68 enhances its splicing activity (Matter, Herrlich & Konig 2002). The mechanisms of action are unclear and could be attributed to alterations in RNA binding and recognition or protein-protein interactions. The previous two chapters have been focussed on understanding structurally how Sam68 specifically recognises RNA. The aim of this chapter was to investigate how serine and threonine phosphorylation affects Sam68 RNA binding, to determine if this type of PTM enhances splicing activity through RNA binding.

Sam68 has been reported to be phosphorylated by three serine/threonine kinases; Cdc2 (Resnick et al. 1997), Erk1 (Matter, Herrlich & Konig 2002) and Nek2 (personal

communication with Professor Claudio Sette). To determine if these kinases target the STAR domain of Sam68, ^{32}P radiolabelled kinase assays were used on the KH and STAR domain constructs. Nek2 was found to phosphorylate both samples and Erk1 only the STAR domain. The Cdc2 experiment was inconclusive and must be repeated, along with Erk1 to be confident that this kinase does not target the KH domain of Sam68.

Further investigation was focussed on the phosphorylation of Sam68 by Nek2. These proteins have been shown to colocalise in splicing speckles (Naro et al. 2013), and Figure 5.4 demonstrates that these proteins colocalise in the nucleus of various cancer and non-cancer derived cell lines. Mass spectrometry was carried out on the ^{32}P assay samples and identified T183 and S196 as potential sites of phosphorylation. NMR kinase assays further identified T119 and S137. In addition, S202 was investigated due to its known involvement in RNA binding. This gave a total of five potential targets of a total twelve serine and threonine residues within the STAR domain. Interestingly, S117, which fits the consensus sequence of Nek2, was not identified in these experiments. It is possible that with further optimisation of the NMR kinase assay, particularly in terms of buffer conditions and maintenance of pH, that further sites would be identified. The same is true for development of such an assay with incorporation of nuclear extract in order to obtain the best physiologically relevant sites of phosphorylation.

Plotting these residues on the model of Sam68 shows the position of T119 close to the NK dimerisation interface. Mutation of this residue did not result in significant changes to the stability of the NKKH domain and is therefore unlikely to affect the dimerisation. The Kd of interaction between NKKH T119E and UAAAUAAA was increased from 1.775μM for NKKH WT to 4.987μM, suggesting that this phospho-mimic has a slightly negative effect on RNA binding, which is surprising, given its distance from the RNA binding site. S137 is located in the flexible linker between the NK and KH domains and therefore its precise position cannot be accurately determined with respect to the RNA binding pocket. S137E, along with S202E, did not express well in *E.coli* and could not be assessed by FP. T183 is directly adjacent to I184 and K185, and sits directly between these residues and the GXXG loop and RNA binding residues of K175 and L177. This suggests that phosphorylation of this residue is likely to affect RNA binding and mutation of this residue to glutamic acid resulted in a decrease in the Kd to 20.36μM. Finally S196E had a Kd of 142.9μM for UAAAUAAA, which suggests it binds with much less affinity than the WT. S196 is adjacent to V197, which was identified as being

involved in RNA binding by NMR studies and therefore may be affected by phosphorylation of S196 (Chapters 3 and 4).

It is notable that in all cases, the affinity of the NKKH domain for UAAAUAAA was decreased upon phosphomimetic mutation of potential Nek2 target sites. It had been hypothesised that serine/threonine phosphorylation of Sam68 enhanced splicing activity through alterations in RNA binding. Initially, it had been anticipated that an enhancement in activity may have resulted from an increase in RNA binding affinity. This does not appear to be the case, under these conditions.

Mutation of these residues to glutamic acid may not be an accurate mimic of phosphorylation and may alter protein structure which in turn would affect RNA binding. Mutation of T119, T183 and S196 to alanines resulted in a stable sample for NMR, with a spectrum indicative of a well-folded protein. However, it would be important to determine the effect of glutamic acid mutation on protein structure either by circular dichroism or NMR. Adjusting the phosphomimetic mutation to an aspartic acid may also result in better protein stability. Further optimisation of Sam68 samples incubated with Nek2 rather than phosphomimetic mutants for fluorescence polarisation could circumvent the necessity for mutational analysis. Use of a longer RNA, such as UAAA(20C)UAAA which binds the NKKH domain with higher affinity and cooperation of both KH domains within the dimer would also be important. Protein kinases are directed by consensus sequences as well as the structure of their targets. More accessible regions, i.e. those that are disordered, flexible or exposed, are more likely to be phosphorylated, hence most NMR studies of phosphorylation have been undertaken on peptides or disordered proteins. This may explain the challenges encountered monitoring phosphorylation by NMR of a folded domain, and why those sites identified are within the flexible NK-KH linker and those within the exposed RNA binding region. The flexibility and accessibility will be affected by the temperature of the sample, and therefore by increasing the temperature of the NMR kinase assay to 35°C may yield further information.

Identification of additional phosphorylation sites of Nek2 and other serine threonine kinases and the effect on RNA binding may be possible by optimisation of NMR and FP in the presence of nuclear extract, or by systematic mutation of each serine and threonine residue within the STAR domain.

There are several possible approaches to determine the potential physiological relevance of each of the phosphorylation sites. Firstly, quantification of ^{32}P radiolabelled assays of

Sam68 mutants with Nek2 using scintillation counting would determine the relative importance of each residue as a Nek2 target. In addition, expression of WT Sam68 and such mutants in mammalian cells, followed by western blotting with phospho-specific antibodies would highlight whether these sites are phosphorylated *in cellulo*. Identification of the kinases responsible would be possible using immunoprecipitation and western blotting or mass-spectrometry.

Regulation of Sam68 alternative splicing function through PTMs may arise from changes in RNA sequence recognition and therefore a range of RNAs should be tested with phosphorylated Sam68. It is also likely that this enhancement in splicing activity arises through alterations in protein-protein interactions. Sam68 is known to interact with various members of the spliceosomal machinery (Bedford et al. 2000) and splicing factors (Venables et al. 1999) and it is likely that these interactions are mediated through post-translational modifications to affect alternative splicing.

Therefore these data provide a good starting point for further investigation of the effect of post-translational modifications on Sam68 function, using structural and biophysical techniques.

Chapter 6. General Discussion

Sam68 is a member of the STAR family of proteins, characterised by their RNA binding properties and involvement in signal transduction pathways. STAR proteins are one of several RNA binding protein families that function as a mediator between cell signalling and RNA processing events such as alternative splicing. The incredible diversity of protein products produced from a limited number of genes is reflected in the complexity of alternative splicing and its highly cell cycle and tissue specific nature. It is important to understand the intricacies of this process in order to prevent and treat the many diseases that arise from aberrant alternative splicing. The development of as many as 50% of all human genetic diseases have been at least in part attributed to changes in alternative splicing and incorrect isoform expression (Tazi, Bakkour & Stamm 2009). Splice site selection is regulated by a number of splicing factors, ribonucleoproteins and RBPs, including Sam68 and other members of the STAR family.

The role of Sam68 in RNA processing is in itself incredibly complex (as discussed in the introduction to this thesis) and in order to get a clear picture of the role of this STAR protein in alternative splicing, it is necessary to understand the mechanism and specificity of Sam68 RNA recognition. Structural studies of several other STAR proteins, including Gld-1, QKI and SF1, have revealed the contribution of the KH and CK domains to RNA recognition (Teplova et al. 2013)(Daubner et al. 2014, Liu et al. 2001). These data also revealed the specificity of RNA targets, in accordance with specific consensus RNA sequences determined by a range of techniques including SELEX, CLIP and site-directed mutagenesis (Volk, Artzt 2010). Thus far, however, the structures of the other STAR protein subfamily, comprised of Sam68, T-STAR and SLM2, have not been solved.

This project therefore set out to determine the structure of Sam68 with RNA, both identifying the subdomains and particular residues involved in RNA binding and to characterise the specificity of RNA recognition and define a consensus RNA target sequence. NMR, X-ray crystallography, SAXS, computational modelling, site-directed mutagenesis and FP were used to produce and verify a novel structural model of Sam68 STAR domain, highlighting the complimentary nature of these structural and biophysical techniques.

The structural model, based on the crystal structure of T-STAR (unpublished), revealed a dimerisation interface between helix 3 of the two KH domains of the STAR domain.

This demonstrates that the Sam68 subfamily of STAR proteins adopts a significantly different arrangement than those of Gld-1, QKI and SF1. Other KH domain containing proteins have been reported to self-associate, however this remains contentious and has often been attributed to crystal packing rather than a physiological arrangement. Several strategies were used to verify the structural model of Sam68 STAR.

SAXS envelopes were calculated for the NKKH and STAR domains of Sam68. This demonstrated that the structural model of Sam68 NKKH based on T-STAR had a better fit to the NKKH envelope than the Sam68 STAR structural model based on Gld-1 with the STAR domain envelope. This suggests that Sam68 NKKH forms a more compact arrangement with flexibility between the NK and KH domains than Gld-1.

This flexibility is possible through a 20 amino acid linker between the two subdomains. In the Gld-1 structure this linker forms contacts with the KH domain, resulting in a specific arrangement of the NK and CK domains. Sequence alignment of Sam68 and Gld-1 in this linker region revealed a lack of residue conservation, suggesting that these contacts do not form in the same arrangement in Sam68. The dynamic and flexible nature of this linker is also observed in heteronuclear NOE experiments of the STAR and NKKH domains. Furthermore, the three-dimensional conformation of Gld-1 STAR domain is stabilised by sandwiching of the CK domain between the NK and KH domains. NMR studies of Sam68 show that there are not direct contacts between the CK and KH domains or the CK and NK domains as in Gld-1. Instead, the NK domain was revealed to be essential for correct folding of the KH domain of Sam68. NMR studies of the KH domain of Sam68 showed that this domain alone is unstable and prone to aggregation. Inclusion of the NK domain had a significant effect on protein stability, suggesting that the NK facilitates bringing together and stabilising the dimerisation of the KH domains within the NKKH dimer.

Tyrosine 241 was identified as one of several residues likely to be at the interface of interaction between the two KH helices in the Sam68 structural model. This residue is a glutamic acid in Gld-1, and therefore site-directed mutagenesis was used to replace Y241 with the equivalent Gld-1 residue and to investigate the subsequent effect on Sam68 structure. Sam68 KH Y241E was more stable than the WT, suggesting that the dimerisation may lead to aggregation of this construct. The STAR and NKKH Y241E samples were less stable than the WT constructs, and the HSQC of NKKH Y241E revealed that the NK domain is not affected by the mutation, whereas many of the KH peaks disappeared or moved to the equivalent position in the KH WT spectrum.

We can therefore conclude that the NK domain stabilises dimerisation of the KH domain and that this is required for correct folding of the NKKH and STAR domains of Sam68.

RNA binding studies also support the STAR structural model of Sam68. Existing SELEX data for Sam68 suggested an RNA consensus sequence of UAAA, which is a different, and shorter binding sequence than that of Gld-1 and QKI. NMR and FP were used to define a consensus RNA binding sequence of (A/U/C)-(**A**/U)-A-A-(A/U), agreeing with SELEX data. Therefore the Sam68 subfamily appears to bind just 4 nucleotides, rather than 6 as for the Gld-1 subfamily. This is supported by NMR and FP studies of Sam68 that show a lack of interaction between the CK domain and RNA, that is characteristic of Gld-1, QKI and SF1.

These data also revealed the KH residues of Sam68 that are involved in RNA binding. Mutational analysis confirmed the contribution of each of these to RNA binding, and supported their position in the Sam68 KH structural model with AUAAU. The two RNA binding pockets within the NKKH dimer are found to be approximately 50Å apart, suggesting that Sam68 binds a single RNA containing two RNA binding sequences separated by at least 15 nucleotides. To test this, FP was conducted with Sam68 NKKH, STAR and NKKH Y241E with a series of RNAs comprising two UAAA motifs separated by a non-specific cytosine sequence of 5, 10, 15, 20, 25 and 30 nucleotides. This revealed that binding affinity of Sam68 NKKH and STAR for the RNA increased significantly for a cytosine linker of 15 nucleotides or more, confirming our structural model prediction. Interestingly there was no trend in binding for Sam68 NKKH Y241E and the same affinity was seen for 5-20 nucleotide linkers. This affinity was significantly higher than the UAAA motif alone, suggesting that the KH domains may be flexible within the NKKH Y241E construct, and can orientate themselves in such a way as to accommodate two UAAA motifs closer together than the WT construct. This bipartite RNA recognition sequence was suggested from SELEX data for Sam68, and several of these identified sequences were also tested with Sam68 NKKH and STAR domains. This also concluded that the CK is not required for RNA recognition and that this distance between RNA binding sites would not be accommodated for the Gld-1 structural model of Sam68.

We therefore concluded that the consensus RNA binding sequence of Sam68 is (A/U/C)-(**A**/U)-A-A-(A/U)-$N_{>15}$-(A/U/C)-(**A**/U)-A-A-(A/U).

Of the many known pre-mRNA targets of Sam68, several have been identified to contain AU-rich sequences that may serve as Sam68 recognition sites (Chawla et al. 2009). Both SRSF1 and Sgce contain two AU-rich motifs, separated by a non-Sam68 specific region of mRNA. In the case of SRSF1, Sam68 promotes inclusion of exon 5 and may facilitate looping out of a ~600 nucleotide stretch of intronic RNA to bring exon 4 and 5 closer together for inclusion in mature mRNA (Valacca et al. 2010). This looping out functionality may also serve to promote exon exclusion, as for the alternative splicing of the Sgce gene. Sam68 has been shown to promote exclusion of exon 8 by binding to sites close to the 3' end of exon 4 and the 5' end of exon 5 (Valacca et al. 2010) (Figure 4.30).

We conclude that Sam68 could function in alternative splicing by bringing two distant RNA sites close together to loop out regions of RNA and dictate the inclusion and/or exclusion of particular intronic and/or exonic RNA in the final transcript.

It has been shown that Sam68 function is regulated by post-translational modifications through signal transduction pathways. In particular, serine and threonine phosphorylation has been reported to enhance Sam68 splicing activity (Matter, Herrlich & Konig 2002). We postulated that this occurred through alterations in RNA binding of Sam68 and determined that the KH and STAR domains were phosphorylated by Nek2. After confirming that these proteins colocalise in the nucleus, ^{32}P-radiolabelled kinase assays, mass spectrometry and NMR were used to find particular sites of phosphorylation within the RNA binding domain of Sam68. Residues T119, S137, T183, T191, S196 and S202 were all identified as being potential phosphorylation sites of Nek2 and several of these were found to be in close proximity to the RNA binding sites. Phosphomimetic mutation of these residues and FP analysis with UAAAUAAA showed that some of these mutations had a negative effect on RNA binding, particularly for T183 which is close in space to RNA binding residues K175, L177 and K185 and S196 which is adjacent to V197 which interacts with RNA. This was not the anticipated outcome, since S/T phosphorylation was previously shown to enhance Sam68 splicing function. It may be that the mutation of these residues to glutamic acids, which is the common substitution for mimicking phosphorylation, does not accurately simulate this post-translational modification. Therefore development of a stable, phosphorylated Sam68 sample through incubation with active Nek2 kinase for FP would be preferred. In addition, developing the NMR kinase assay and FP in the presence of nuclear extract and with longer, bipartite RNA sequences may yield a more physiological outlook on

RNA binding in the presence of serine threonine kinases. This would be the case for investigation of kinases other than Nek2 that are known to phosphorylate Sam68 such as, Cdc2 and Erk1 (Matter, Herrlich & Konig 2002, Resnick et al. 1997). It would also be useful for investigating other regions of Sam68 outside the STAR domain that may be phosphorylated.

It is possible that this post-translational modification of Sam68 does reduce its affinity for RNA, in which case the alteration in splicing function may be brought about via changes in the specificity of RNA binding and/or by changes in protein-protein interactions with other members of the spliceosomal machinery.

The structural model of Sam68 and biophysical studies of this protein-RNA complex have highlighted a novel protein conformation and function of this STAR protein, compared to other members of the family. This provides an attractive avenue for drug discovery to interrupt KH dimerisation and RNA binding. In that view we were successful in application to a high throughput binding assay development program using FP through the Scottish Universities Life Sciences Alliance (SULSA) Assay Development Fund. The assay development was successful and we will apply for screening of 500,000 compounds through the European Lead Factory (ELF), an international consortium providing a novel platform for discovery of new drug lead molecules.

Chapter 7. Appendices

7.1. Plasmid vectors

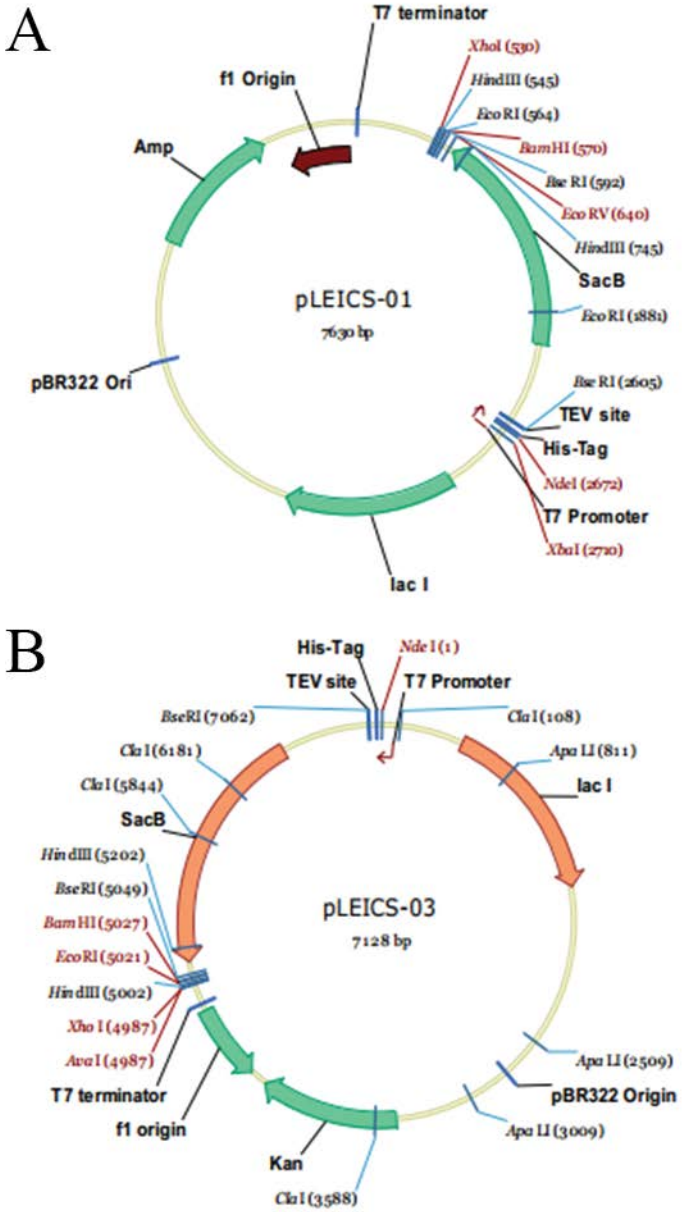


Figure 7.1: Plasmid vectors used for molecular cloning
Bacterial plasmid vectors pLEICS-01 (A) and pLEICS-03 (B) were used to create recombinant proteins of mutant and WT constructs, respectively.

7.2. Oligonucleotide primers

Protein Construct	Oligonucleotide Primers
KH	5’ - TACTTCCAATCCATGTCTCATAAGAACATGAAACTG 3’ - TACTTCCAATCCATGTTACATCATATCCGGTACTAGAAA
KHCK	5’ - TACTTCCAATCCATGTCTCATAAGAACATGAAACTG 3’ - TATCCACCTTTACTGTCAACCACGAGAGGGTTCAGGTAC
NKKH	5’ - TACTTCCAATCCATGATGGAGCCAGAGAACAAGTAC 3’ - TACTTCCAATCCATGTTACATCATATCCGGTACTAGAAA
STAR	5’ - TACTTCCAATCCATGATGGAGCCAGAGAACAAGTAC 3’ - TATCCACCTTTACTGTCAACCACGAGAGGGTTCAGGTAC
N-terminus	5’ – TACTTCCAATCCATGCAGCGCCGGGACGACCCC 3’ - TATCCACCTTTACTGTCATTACTTGACCGAGGCTGTGGC
C-terminus	5’ – TACTTCCAATCCATGCGTGGGGTGCCAGTGAGA 3’ - TATCCACCTTTACTGTCATTAATAACGTCCATATGGGTG
Full-length	5’ - ATGCAGCGCCGGGACGAC 3’ - TTAATAACGTCCATATGGGTG
C238A	5’ - TTCATTGAAGTCTTTGGACCCCCAGCTGAGGCTTATGCTCTTATGGCCCAT 3’ -TAATGGGCCATAAGAGCATAAGCCTCAGCTGGGGGTCCAAAGACTTCAATGAA
Y241E	5’ - GTCTTTGGACCCCCATGTGAGGCTGAAGCTCTTATGGCCCATGCCATGGAG 3’ - CTCCATGGCATGGGCCATAAGAGCTTCAGCCTCACATGGGGGTCCAAAGAC
N171A	5’ - CCTGTCAAGCAGTATCCCAAGTTCGCTTTTGTGGGGAAGATTCTT 3’ - TGGTCCAAGAATCTTCCCCACAAAAGCGAACTTGGGATACTGCTT
F172A	5’ – GTCAAGCAGTATCCCAAGTTCAATGCTGTGGGGAAGATTCTTGGA 3’ - TTGTGGTCCAAGAATCTTCCCCACAGCATTGAACTTGGGATACTG
K175A	5’ – TATCCCAAGTTCAATTTTGTGGGGGCGATTCTTGGACCACAAGGG 3’ - TGTATTCCCTTGTGGTCCAAGAATCGCCCCCACAAAATTGAACTT
L177A	5’ - AAGTTCAATTTTGTGGGGAAGATTGCTGGACCACAAGGGAATACA 3’ - TTTGATTGTATTCCCTTGTGGTCCAGCAATCTTCCCCACAAAATT
I184A	5’ – ATTCTTGGACCACAAGGGAATACAGCCAAAAGACTGCAGGAAGAG 3’ - ACCAGTCTCTTCCTGCAGTCTTTTGGCTGTATTCCCTTGTGGTCC
K185A	5’ – CTTGGACCACAAGGGAATACAATCGCAAGACTGCAGGAAGAGACT 3’ - TGCACCAGTCTCTTCCTGCAGTCTTGCGATTGTATTCCCTTGTGG
V197A	5’ - GAAGAGACTGGTGCAAAGATCTCTGCATTGGGAAAGGGCTCAATG 3’ - GTCTCTCATTGAGCCCTTTCCCAATGCAGAGATCTTTGCACCAGT
S202A	5’ – AAGATCTCTGTATTGGGAAAGGGCGCAATGAGAGACAAAGCCAAG 3’ - TTCCTCCTTGGCTTTGTCTCTCATTGCGCCCTTTCCCAATACAGA
K206A	5’ – TTGGGAAAGGGCTCAATGAGAGACGCAGCCAAGGAGGAAGAGCTG 3’ - TTTGCGCAGCTCTTCCTCCTTGGCTGCGTCTCTCATTGAGCCCTT
E209A	5’ - GGCTCAATGAGAGACAAAGCCAAGGCGGAAGAGCTGCGCAAAGGT 3’ - GTCTCCACCTTTGCGCAGCTCTTCCGCCTTGGCTTTGTCTCTCAT
E210A	5’ – TCAATGAGAGACAAAGCCAAGGAGGAAGAGCTGCGCAAAGGTGGA 3’ - GGGGTCTCCACCTTTGCGCAGCTCTTCCTCCTTGGCTTTGTCTCT
M258A	3’ - TATCCACCTTTACTGTCACATCGCATCCGGTACTAGAAATTTCTT

S113A	5’ - CCCGAACTCATGGCCGAGAAGGACGCGCTCGACCCGTCCTTCACTCACGCC 3’ - GGCGTGAGTGAAGGACGGGTCGAGCGCGTCCTTCTCGGCCATGAGTTCGGG
S113E	5’ - CCCGAACTCATGGCCGAGAAGGACGAGCTCGACCCGTCCTTCACTCACGCC 3’ - GGCGTGAGTGAAGGACGGGTCGAGCTCGTCCTTCTCGGCCATGAGTTCGGG
S117A	5’ - GCCGAGAAGGACTCGCTCGACCCGGCCTTCACTCACGCCATGCAGCTGCTG 3’ - CAGCAGCTGCATGGCGTGAGTGAAGGCCGGGTCGAGCGAGTCCTTCTCGGC
T117E	5’ - GCCGAGAAGGACTCGCTCGACCCGGAATTCACTCACGCCATGCAGCTGCTG 3’ - CAGCAGCTGCATGGCGTGAGTGAATTCCGGGTCGAGCGAGTCCTTCTCGGC
T119A	5’ - AAGGACTCGCTCGACCCGTCCTTCGCTCACGCCATGCAGCTGCTGACGGCA 3’ - TTATGCCGTCAGCAGCTGCATGGCGTGAGCGAAGGACGGGTCGAGCGAGTCCTT
T119E	5’ - AAGGACTCGCTCGACCCGTCCTTCGAACACGCCATGCAGCTGCTGACGGCA 3’ - TGCCGTCAGCAGCTGCATGGCGTGTTCGAAGGACGGGTCGAGCGAGTCCTT
T126A	5’ - ACTCACGCCATGCAGCTGCTGGCGGCAGAAATTGAGAAGATTCAGAAA 3’ - TTTCTGAATCTTCTCAATTTCTGCCGCCAGCAGCTGCATGGCGTGAGT
T126E	5’ - ACTCACGCCATGCAGCTGCTGGAGGCAGAAATTGAGAAGATTCAGAAA 3’ - TTTCTGAATCTTCTCAATTTCTGCCTCCAGCAGCTGCATGGCGTGAGT
S137A	5’ - ATTGAGAAGATTCAGAAAGGAGACGCAAAAAAGGATGATGAGGAGAATTAC 3’ - GTAATTCTCCTCATCATCCTTTTTTGCGTCTCCTTTCTGAATCTTCTCAAT
S137E	5’ - ATTGAGAAGATTCAGAAAGGAGACGAAAAAAAGGATGATGAGGAGAATTAC 3’ - GTAATTCTCCTCATCATCCTTTTTTTCGTCTCCTTTCTGAATCTTCTCAAT
S150A	5’ - GAGGAGAATTACTTGGATTTATTTGCTCATAAGAACATGAAACTGAAAGAG 3’ - CTCTTTCAGTTTCATGTTCTTATGAGCAAATAAATCCAAGTAATTCTCCTC
S150E	5’ - GAGGAGAATTACTTGGATTTATTTGAACATAAGAACATGAAACTGAAAGAG 3’ - CTCTTTCAGTTTCATGTTCTTATGTTCAAATAAATCCAAGTAATTCTCCTC
T183A	5’ - AAGATTCTTGGACCACAAGGGAATGCAATCAAAAGACTGCAGGAAGAGACT 3’ - AGTCTCTTCCTGCAGTCTTTTGATTGCATTCCCTTGTGGTCCAAGAATCTT
183E	5’ - AAGATTCTTGGACCACAAGGGAATGAAATCAAAAGACTGCAGGAAGAGACT 3’ - AGTCTCTTCCTGCAGTCTTTTGATTTCATTCCCTTGTGGTCCAAGAATCTT
T191A	5’ - ACAATCAAAAGACTGCAGGAAGAGGCTGGTGCAAAGATCTCTGTATTGGGA 3’ - TCCCAATACAGAGATCTTTGCACCAGCCTCTTCCTGCAGTCTTTTGATTGT
T191E	5’ - ACAATCAAAAGACTGCAGGAAGAGGAAGGTGCAAAGATCTCTGTATTGGGA 3’ - TCCCAATACAGAGATCTTTGCACCTTCCTCTTCCTGCAGTCTTTTGATTGT
S196A	5’ - CAGGAAGAGACTGGTGCAAAGATCGCTGTATTGGGAAAGGGCTCAATGAGA 3’ - TCTCATTGAGCCCTTTCCCAATACAGCGATCTTTGCACCAGTCTCTTCCTG
S196E	5’ - CAGGAAGAGACTGGTGCAAAGATCGAGGTATTGGGAAAGGGCTCAATGAGA 3’ - TCTCATTGAGCCCTTTCCCAATACCTCGATCTTTGCACCAGTCTCTTCCTG
S202A	5’ - AAGATCTCTGTATTGGGAAAGGGCGCAATGAGAGACAAAGCCAAGGAGGAA 3’ - TTCCTCCTTGGCTTTGTCTCTCATTGCGCCCTTTCCCAATACAGAGATCTT
S202E	5’ - AAGATCTCTGTATTGGGAAAGGGCGAAATGAGAGACAAAGCCAAGGAGGAA 3’ - TTCCTCCTTGGCTTTGTCTCTCATTTCGCCCTTTCCCAATACAGAGATCTT

Table 7.1: Sam68 oligonucleotide primers.

7.3. Purification Buffers

Lysis Buffer	Elution 1
50mM Na_2HPO_4	50mM Na_2HPO_4
1M NaCl	1M NaCl
10mM Imidazole	20mM Imidazole

Elution2	Elution 3
50mM Na_2HPO_4	50mM Na_2HPO_4
1M NaCl	1M NaCl
50mM Imidazole	100mM Imidazole

Elution4	Elution 5
50mM Na_2HPO_4	50mM Na_2HPO_4
1M NaCl	1M NaCl
250mM Imidazole	500mM Imidazole

Dialysis Buffer	NMR/X-ray/FP/Kinase Assay Buffer
20mM Na_2HPO_4	10mM Bis-Tris HCl pH7
100mM NaCl	100mM NaCl
0.1% β-Mercaptoethanol	0.1% β-Mercaptoethanol

7.4. Bacterial growth media (1 Litre)

M9 Salts	
Na_2HPO_4	34.1g
KH_2PO_4	15g
NaCl	2.5g

2TY	
Tryptone	16g
Yeast extract	10g
NaCl	5g

M9 minimal medium	
dH_2O	700ml
5xM9 Salts	200ml
1M $MgSO_4$	2ml
1M $CaCl_2$	0.1ml
NH_4Cl	1g in 50ml dH_2O
Glucose	4g in 50ml dH_2O
^{13}C Glucose	1g in 50ml dH_2O

7.5. ^{32}P Kinase assay

Buffer mastermix (10 reactions)	Reaction mix (Per reaction)
50mM HEPES.KOH pH 7.4	40µl Kinase buffer
5mM $MnCl_2$	100ng kinase
5mM β-glycerophosphate	5µl substrate protein (at ~5mg/ml)
5mM NaF	
4nM ATP	
1mM DTT	
10mCi[^{32}P]-γ-ATP	
dH2O (up to 500µl total volume for 10 reactions)	

7.6. Table of Assignments

Residue	Number	N	H
Met	97	119.282	8.419
Glu	98	123.342	8.304
Glu	100	120.875	8.62
Asn	101	119.283	8.321
Lys	102	121.73	8.31
Tyr	103	120.011	7.946
Leu	104	118.002	8.367
Glu	106	117.688	7.652
Leu	107	120.724	8.187
Met	108	116.578	8.257
Ala	109	121.06	7.889
Glu	110	119.872	7.88
Lys	111	119.54	8.289
Asp	112	115.409	7.969
Ser	113	112.258	7.605
Leu	114	126.609	7.761
Asp	115	127.441	8.085
Ser	117	114.806	8.729
Phe	118	124.991	7.976
Thr	119	115.863	7.349
His	120	127.125	11.884
Ala	121	127.312	11.029
Met	122	114.874	8.472
Gln	123	122.56	7.933
Leu	124	121.329	8.86
Leu	125	120.972	8.654
Thr	126	114.732	8.145
Ala	127	123.318	8.217
Glu	128	121.532	7.864
Ile	129	119.728	8.19
Glu	130	118.067	8.017
Lys	131	119.709	7.915
Ile	132	119.447	7.949
Gln	133	118.939	8.443
Lys	134	117.519	8.033
Gly	135	107.446	7.718

Residue	Number	N	H
Asp	136	120.393	8.181
Ser	137	116.232	8.196
Lys	138	123.318	8.26
Lys	139	123.279	8.309
Asp	140	121.764	8.348
Asp	141	120.309	8.203
Glu	143	120.456	8.231
Asn	144	119.399	8.107
Tyr	145	118.858	7.937
Leu	146	122.332	9.283
Asp	147	122.913	8.73
Leu	148	123.472	9.221
Phe	149	117.359	8.517
Ser	150	114.677	7.331
Lys	152	121.589	7.94
Asn	153	122.18	8.437
Met	154	121.288	8.833
Lys	155	120.007	8.215
Leu	156	123.642	8.343
Lys	157	119.458	8.378
Glu	158	119.8	8.541
Arg	159	120.189	9.085
Val	160	123.667	9.71
Leu	161	128.26	8.932
Ile	162	122.382	7.283
Val	164	122.154	8.734
Lys	165	118.027	8.279
Gln	166	117.137	7.383
Tyr	167	115.61	7.354
Lys	169	116.317	8.629
Phe	170	120.965	7.418
Phe	172	123.67	8.072
Val	173	117.904	8.263
Gly	174	103.892	7.676
Lys	175	121.351	7.181
Ile	176	117.477	7.977

Residue	Number	N	H
Leu	177	114.544	7.974
Gly	178	104.896	7.478
Gly	181	106.309	8.005
Asn	182	117.623	8.272
Thr	183	123.212	7.412
Ile	184	124.453	7.572
Lys	185	120.498	7.654
Glu	189	120.663	8.02
Glu	190	116.774	8.687
Thr	171	105.069	7.7
Gly	182	109.796	8.045
Ala	193	121.702	7.775
Lys	194	121.193	8.507
Ile	195	125.457	8.073
Ser	196	118.974	8.605
Val	197	124.38	8.644
Leu	198	127.555	8.259
Gly	199	106.469	9.825
Lys	200	121.409	8.6
Gly	201	117.415	11.121
Ser	202	115.721	8.449
Asp	205	117.006	7.747
Lys	206	126.456	8.54
Ala	207	121.989	8.148
Lys	208	121.073	7.732
Glu	209	118.218	8.18
Glu	210	117.244	7.773
Glu	211	118.93	7.535
Leu	212	119.765	8.164
Arg	213	121.2	8.542
Lys	214	117.665	7.907
Gly	215	105.434	7.624
Gly	216	105.361	7.29
Asp	217	124.562	8.916
Lys	219	120.083	8.812
Tyr	220	115.103	7.701

Residue	Number	N	H
Ala	221	123.994	7.406
His	222	115.772	8.113
Leu	223	119.721	7.371
Asn	224	114.431	7.586
Met	255	120.06	8.167
Asp	266	120.159	8.223
Leu	277	122.947	8.73
His	228	123.547	9.259
Val	229	118.256	9.386
Phe	230	131.128	9.684
Ile	231	129.123	8.886
Glu	232	124.849	8.865
Val	233	120.652	8.746
Phe	234	123.469	8.238
Gly	235	109.853	8.391
Cys	238	112.911	8.69
Glu	239	120.491	8.107
Ala	240	122.044	8.973
Tyr	241	115.092	7.863
Ala	242	121.045	7.38
Leu	243	120.703	8.795
Lys	253	114.342	7.137
Phe	254	116.106	7.905
Leu	255	117.197	7.582
Val	256	115.401	6.938
Asp	258	119.516	6.876
Met	259	119.384	8.006
Phe	268	120.541	8.155
Leu	269	123.051	7.991
Glu	270	121.263	8.226
Leu	271	122.224	8.065
Ser	272	115.772	8.113
Tyr	273	121.397	7.939
Leu	274	122.56	7.933
Asn	275	118.371	8.232
Glu	276	108.418	8.106

Table 7.2: Table of Sam68 STAR backbone assignments

7.7. Crystallisation Trials

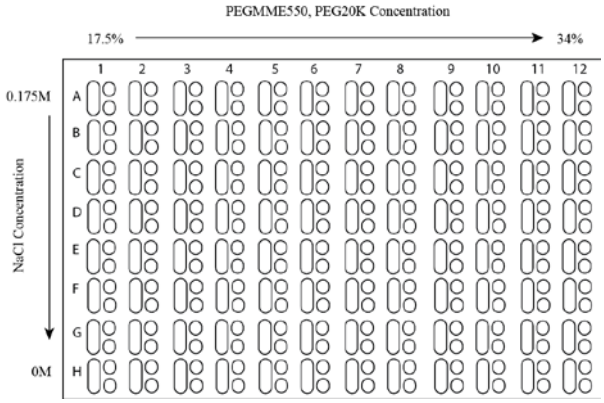


Figure 7.2: Morpheus optimised crystallisation screen
Microwell plate image showing the conditions screened from hits obtained from the Morpheus screen (Hampton Research).

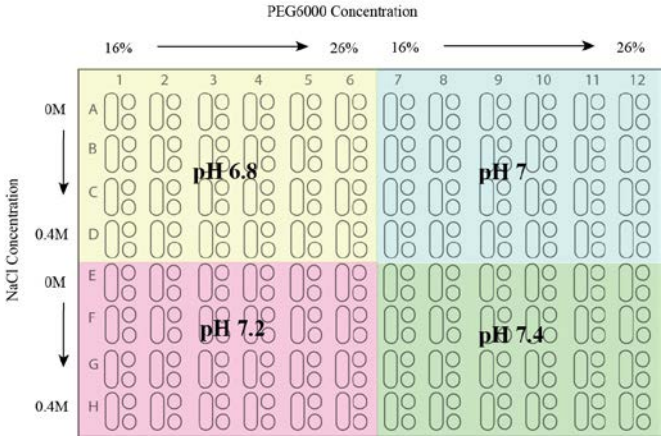


Figure 7.3: PACT Optimised 1 crystallisation screen
Microwell plate image showing the conditions screened from hits obtained from the PACT screen (Hampton Research).

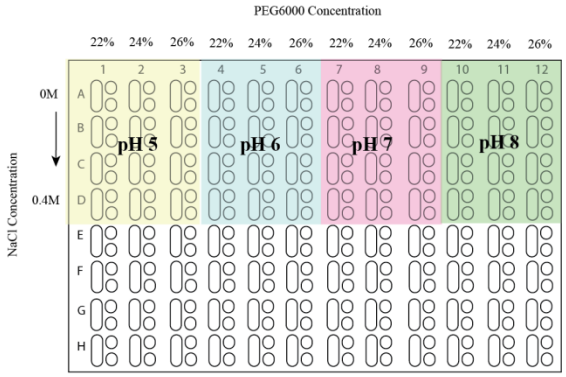


Figure 7.4: PACT Optimised 2 crystallisation screen
Microwell plate image showing the conditions screened from hits obtained from the PACT optimised 1 screen.

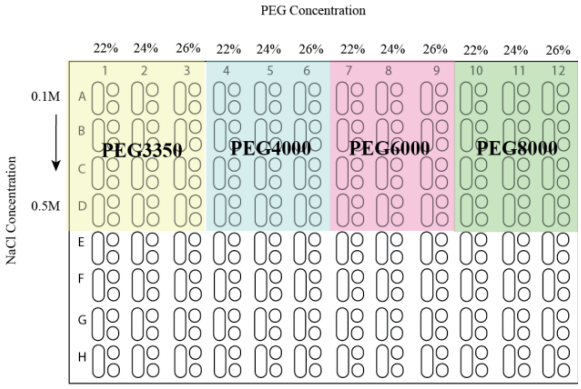


Figure 7.5: PACT Optimised 3 crystallisation screen
Microwell plate image showing the conditions screened from hits obtained from the PACT optimised 2screen.

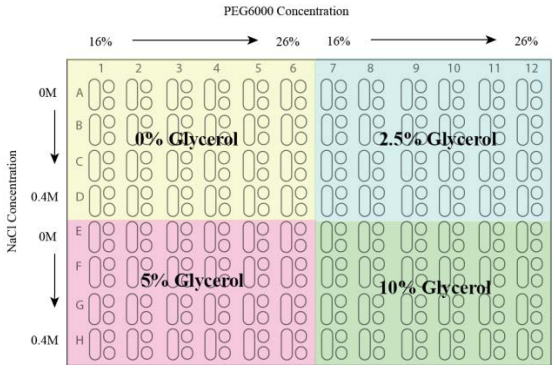


Figure 7.6: PACT Optimised 4 crystallisation screen
Microwell plate image showing the conditions screened from hits obtained from the PACT optimised 3 screen.

7.8. Modeller input data

A

```
>P1;Gld-1STAR
StructureX:GLD1-RNA-4JVY:144  :A:331  :B::::
EATVEYLADLVKEKKHLTLFPHMFSNVERLLDDEIGRVRVALFQTEFPRVELPEPAGDMISITEKIYVPKNEYPDYNFVGRILGPRGMTAKQLEQDTGCKIMVRGKGSMRDKSKESAHR--
GKANWEHLEDDLHVLVQCEDTENRVHIKLQAALEQVKKLLIPAPEGTDELKRKQLMELAIINGTYRPMKS/
SHEATVEYLADLVKEKKHLTLFPHMFSNVERLLDDEIGRVRVALFQTEFPRVELPEPAGDMISITEKIYVPKNEYPDYNFVGRILGPRGMTAKQLEQDTGCKIMVRGKGSMRDKSKESAHR--
GKANWEHLEDDLHVLVQCEDTENRVHIKLQAALEQVKKLLIPAPEGTDELKRKQLMELAIINGTYRPMK*

>P1;Sam68STAR
sequence:Sam68KH:98   :A :282  :B ::::
EPENKYLPELMAEKDSLDPS---
FTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVP--
DMMDDICQEQFLELSYLNGVPEPSRG/
--EPENKYLPELMAEKDSLDPS---
FTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVP--
DMMDDICQEQFLELSYLNGVPEPSR*
```

B

```
# Homology modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

a = automodel(env,
              alnfile  = 'alignment.ali',     # alignment filename
              knowns   = 'Gld-1STAR',              # codes of the templates
              sequence = 'Sam68STAR')              # code of the target
a.starting_model= 2                 # index of the first model
a.ending_model  = 5                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modeling
```

Figure 7.7: Homology modelling of Sam68 using Modeller
Sequence alignment file input for Modeller (A) and input command including Gld-1 pdb file (B).

7.9. Long RNA sequences

G8.5

CUGGGUGACACACUAGCUAUAGCAUUAAAAGACCGAGCAAGU

G7.1

UCCGGAUUGGCCUAAAUAGAUGCGCGAUAAUAAUAGAGUA

SRE-4

UUUGGGGGUUCAAUAAAAAUUUUCACUAUCCUAUUAACAGUUCCGCCGC
UCC

Nrx2

CCCAAUUAACUAACUAACUAACUUUAAAA

Chapter 8. Bibliography

Andreotti, A.M., Bunnell, S.C., Feng, S., Berg, L.J. & Schreiber, S.L. 1997, "Regulatory intramolecular association in a tyrosine kinase of the Tec family", *Nature*, vol. 385, pp. 93-97.

Änkö, M. 2014, "Regulation of gene expression programmes by serine–arginine rich splicing factors", *Seminars in cell & developmental biology*, vol. 32, no. 0, pp. 11-21.

Aquino-Jarquin, G. & Toscano-Garibay, J.D. 2011, "RNA Aptamer Evolution: Two Decades of SELEction", *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 9155-9171.

Babic, I., Cherry, E. & Fujita, D.J. 2006, "SUMO modification of Sam68 enhances its ability to repress cyclin D1 epxression and inhibits its ability to induce apoptosis", *Oncogene*, vol. 25, pp. 4955-4964.

Babic, I., Jakymiw, A. & Fujita, D.J. 2004, "The RNA binding protein Sam68 is acetylated in tumor cell lines, and its acetylation correlates with enhanced RNA binding activity", *Oncogene*, vol. 23, pp. 3781–3789.

Baltz, A., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C. & Landthaler, M. 2012, "The mRNA-Bound Proteome and Its Global

Occupancy Profile on Protein-Coding Transcripts", *Molecular cell,* vol. 46, no. 5, pp. 674-690.

Barlat, I., Maurier, F., Duchesne, M., Guitard, E., Tocque, B. & Schweighoffer, F. 1997, *A Role for Sam68 in Cell Cycle Progression Antagonized by a Spliced Variant within the KH Domain*.

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,* vol. 39, no. 18, pp. 5355-5365.

Bedford, M.T., Frankel, A., Yaffe, M.B., Clarke, S., Leder, P. & Richard, S. 2000, "Arginine Methylation Inhibits the Binding of Proline-rich Ligands to Src Homology 3, but Not WW, Domains", *Journal of Biological Chemistry,* vol. 275, no. 21, pp. 16030-16036.

Berget, S.M., Moore, C. & Sharp, P.A. 1977, "Spliced segments at the 5' terminus of adenovirus 2 late mRNA", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 74, no. 8, pp. 3171–3175.

Bergfors, T. 2003, "Seeds to crystals", *Journal of structural biology,* vol. 142, no. 1, pp. 66-76.

Berglund, J.A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. 1997, "The Splicing Factor BBP Interacts Specifically with the Pre-mRNA Branchpoint Sequence UACUAAC", *Cell,* vol. 89, no. 5, pp. 781-787.

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., Svergun, D.I., 2007 "Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering". *Journal of the American. Chemistry Society,* vol. 129, no. 17, pp. 5656-5664.

Beuck, C., Qu, S., Fagg, W.S., Ares Jr., M. & Williamson, J.R. 2012, "Structural Analysis of the Quaking Homodimerization Interface", *Journal of Molecular Biology,* vol. 423, no. 5, pp. 766-781.

Beuck, C., Szymczyna, B.R., Kerkow, D.E., Carmel, A.B., Columbus, L., Stanfield, R.L. & Williamson, J.R. 2010, "Structure of the GLD-1 Homodimerization Domain: Insights into STAR Protein-Mediated Translational Regulation", *Structure,* vol. 18, no. 3, pp. 377-389.

Beuth, B., García-Mayoral, M.R., Taylor, I.A. & Ramos, A. 2007, "Scaffold-Independent Analysis of RNA−Protein Interactions: The Nova-1 KH3−RNA Complex", *Journal of the American Chemical Society*, vol. 129, no. 33, pp. 10205-10210.

Bianchi, E., Barbagallo, F., Valeri, C., Geremia, R., Salustri, A., De Felici, M. & Sette, C. 2010, "Ablation of the Sam68 gene impairs female fertility and gonadotropin-dependent follicle development", *Human molecular genetics*, vol. 19, no. 24, pp. 4886-4894.

Black, D.L. 2003, "Mechanisms of alternative pre-messenger RNA splicing", *Annual Review of Biochemistry*, vol. 72, pp. 291-336.

Blanchard, J., Brunel, C. & Jeanteur, P. 1977, "Phosphorylation in vivo and in vitro of Proteins from HeLa Cells: Heterogeneous Nuclear Ribonucleoprotein Particles", *Biochemical Society Transactions*, vol. 5, no. 3, pp. 670-671.

Blanchet, C.E. & Svergun, D.I. 2013, "Small-Angle X-Ray Scattering on Biological Macromolecules and Nanocomposites in Solution", *Annual Review of Physical Chemistry*, vol. 64, pp. 37-54.

Boise, L.H., González-García, M., Postema, C.E., Ding, L., Lindsten, T., Turka, L.A., Mao, X., Nuñez, G. & Thompson, C.B. 1993, "bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death", *Cell*, vol. 74, no. 4, pp. 597-608.

Boucard, A.A., Chubykin, A.A., Comoletti, D., Taylor, P. & Südhof, T.C. 2005, "A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to a- and ß-neurexins", *Neuron*, vol. 48, no. 2, pp. 229-236.

Braddock, D.T., Baber, J.L., Levens, D. & Marius Clore, G. 2002, "Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: solution structure of a complex between hnRNP K KH3 and single-stranded DNA", *The EMBO Journal*, vol. 21, pp. 3476-3485.

Burd, C.G. & Dreyfuss, G. 1994, "Conserved structures and diversity of functions of RNA-binding proteins", *Science*, vol. 265, no. 5172, pp. 615-621.

Busa, R., Paronetto, M.P., Farini, D., Pierantozzi, E., Botti, F., Angelini, D.F., Attisani, F., Vespasiani, G. & Sette, C. 2007, "The RNA-binding protein Sam68 contributes to proliferation and survival of human prostate cancer cells", *Oncogene*, vol. 26, pp. 4372–4382.

Busà, R., Geremia, R. & Sette, C. 2010, "Genotoxic stress causes the accumulation of the splicing regulator Sam68 in nuclear foci of transcriptionally active chromatin", *Nucleic acids research,* vol. 38, no. 9, pp. 3005-3018.

Cappellari, M., Bielli, P., Paronetto, M.P., Ciccosanti, F., Fimia, G.M., Saarikettu, J., Silvennoinen, O. & Sette, C. 2014, "The transcriptional co-activator SND1 is a novel regulator of alternative splicing in prostate cancer cells", *Oncogene,* vol. 33, pp. 3794–3802.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B., Strein, C., Davey, N., Humphreys, D., Preiss, T., Steinmetz, L., Krijgsveld, J. & Hentze, M. 2012, "Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins", *Cell,* vol. 149, no. 6, pp. 1393-1406.

Chawla, G., Lin, C., Han, A., Shiue, L., Ares, M. & Black, D.L. 2009, "Sam68 Regulates a Set of Alternatively Spliced Exons during Neurogenesis", *Molecular and cellular biology,* vol. 29, no. 1, pp. 201-213.

Chen, T., Damaj, B.B., Herrera, C., Lasko, P. & Richard, S. 1997, "Self-association of the single-KH-domain family members Sam68, GRP33, GLD-1, and Qk1: role of the KH domain", *Molecular and Cellular Biology,* vol. 17, no. 10, pp. 5707-18.

Chen, Z., Cai, L., Zhu, J., Chen, M., Chen, J., Li, Z., Liu, X., Wang, S., Bie, P., Jiang, P., Dong, J. & Li, X. 2011, "Fyn requires HnRNPA2B1 and Sam68 to synergistically regulate apoptosis in pancreatic cancer", *Carcinogenesis,* vol. 32, no. 10, pp. 1419-1426.

Cheung, N., Chan, L.C., Thompson, A., Cleary, M.L. & So, C.W.E. 2007, "Protein arginine-methyltransferase-dependent oncogenesis", *Nature Cell Biology*, vol. 9, pp. 1208 - 1215.

Chunyun, H., Youyu, S., Jack, J. & Lianjun, C. 2014, "Identification of melanoma biomarkers based on network modules by integrating the human signaling network with microarrays", *Journal of Cancer Research and Therapeutics,* vol. 10, no. 7, pp. 114-124.

Clark-Lewis, I., Sanghera, J.S. & Pelech, S.L. 1991, "Definition of a consensus sequence for peptide substrate recognition by p44mpk, the meiosis-activated myelin basic protein kinase.", *Journal of Biological Chemistry,* vol. 266, no. 23, pp. 15180-15184.

Cole, C., Barber, J.D. & Barton, G.J. 2008, "The Jpred 3 secondary structure prediction server", *Nucleic acids research,* vol. 36, no. suppl 2, pp. W197-W201.

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J.L., Bell, J.C. & Duncan, P.I. 1996, "The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution", *The EMBO Journal*, vol. 15, no. 2, pp. 265–275.

Côté, J., Boisvert, F., Boulanger, M., Bedford, M.T. & Richard, S. 2003, *Sam68 RNA Binding Protein Is an In Vivo Substrate for Protein Arginine N-Methyltransferase 1*.

Courtneidge, S.A. & Fumagalli, S. 1994, "A mitotic function for Src?", *Trends in Cell Biology*, vol. 4, no. 10, pp. 345–347.

Cukier, C.D. & Ramos, A. 2011, "Modular protein-RNA interactions regulating mRNA metabolism: a role for NMR", *European Biophysics Journal*, vol. 40, no. 12, pp. 1317–1325.

Cukier, C.D. & Ramos, A. 2010, "Creating a Twin STAR", *Structure,* vol. 18, no. 3, pp. 279-280.

Daubner, G.M., Brümmer, A., Tocchini, C., Gerhardy, S., Ciosk, R., Zavolan, M. & Allain, F.H.-. 2014, "Structural and functional implications of the QUA2 domain on RNA recognition by GLD-1", *Nucleic acids research,* vol. 42, no. 12, pp. 8092-8105.

Daubner, G.M., Cléry, A. & Allain, F.H. 2013, "RRM–RNA recognition: NMR or crystallography…and new findings", *Current opinion in structural biology,* vol. 23, no. 1, pp. 100-108.

Dejgaard, K. & Leffers, H. 1996, "Characterisation of the Nucleic-Acid-Binding Activity of KH Domains Different Properties of Different Domains", *European Journal of Biochemistry*, vol. 241, no. 2, pp. 425-431.

Derry, J.J., Richard, S., Valderrama Carvajal, H., Ye, X., Vasioukhin, V., Cochrane, A.W., Chen, T. & Tyner, A.L. 2000, "Sik (BRK) Phosphorylates Sam68 in the Nucleus and Negatively Regulates Its RNA Binding Ability", *Molecular and cellular biology,* vol. 20, no. 16, pp. 6114-6126.

Di Fruscio, M., Chen, T. & Richard, S. 1999, "Characterization of Sam68-like mammalian proteins SLM-1 and SLM-2: SLM-1 is a Src substrate during mitosis", *Proceedings of the National Academy of Sciences,* vol. 96, no. 6, pp. 2710-2715.

Dodson, R.E. & Shapiro, D.J. 1997, "Vigilin, a Ubiquitous Protein with 14 K Homology Domains, Is the Estrogen-inducible Vitellogenin mRNA 3'-Untranslated Region-binding Protein", *Journal of Biological Chemistry*, vol. 272, no. 19, pp. 12249-12252.

Dominguez, C., Schubert, M., Duss, O., Ravindranathan, S. & Allain, F.H.-. 2011, "Structure determination and dynamics of protein–RNA complexes by NMR spectroscopy", *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 58, no. 1–2, pp. 1-61.

Dormann, D., Madl, T., Valori, C.F., Bentmann, E., Tahirovic, S., Abou-Ajram, C., Kremmer, E., Ansorge, O., Mackenzie, I.R.A., Neumann, M. & Haass, C. 2012, "Arginine methylation next to the PY-NLS modulates Transportin binding and nuclear import of FUS", *The EMBO Journal*, vol. 31, pp. 4258-4275.

Dreyfuss, G., Matunis, M.J., Pinol-Roma, S. & Burd, C.G. 1993, "hnRNP Proteins and the Biogenesis of mRNA", *Annual Review of Biochemistry*, vol. 62, no. 1, pp. 289-321.

Edmond, V., Moysan, E., Khochbin, S., Matthias, P., Bramilla, C., Brambilla, E., Gazzeri, S. & Eymin, B. 2011, "Acetylation and phosphorylation of SRSF2 control cell fate decision in response to cisplatin", *The EMBO Journal*, vol. 30, no. 3, pp. 451-628.

Ehrmann, I., Dalgliesh, C., Liu, Y., Danilenko, M., Crosier, M., Overman, L., Arthur, H.M., Linsday, S., Clowry, G., Venables, J.P., Fort, P. & Elliott, D.J. 2013, "T-STAR Controls Regional Splicing Patterns of Neurexin Pre-mRNAs in the Brain", *PLOS Genetics*, .

Ehrmann, I. & Elliott, D.J. 2010, "Expression and functions of the star proteins Sam68 and T-STAR in mammalian spermatogenesis", *Advances in Experimental Medicine and Biology*, vol. 693, pp. 67-81.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U. & Sali, A. 2007, "Comparative Protein Structure Modeling Using Modeller", *Current Protocols in Protein Science*, vol. 2, no. 2.9.

Feracci, M., Foot, J.N. & Dominguez, C. 2014, "Structural investigations of the RNA-binding properties of STAR proteins", *Biochemical Society Transactions*, vol. 42, no. 4, pp. 1141–1146.

Fischer, M. W. F., Losonczi, J. A., Weaver, J. L., Prestegard, J. H., 1999, "Domain Orientation and Dynamics in Multidomain Proteins from Residual Dipolar Couplings", *Biochemistry*, vol. 38, pp. 9013-9022.

Foot, J.N., Feracci, M. & Dominguez, C. 2014, "Screening protein – Single stranded RNA complexes by NMR spectroscopy for structure determination", *Methods,* vol. 65, no. 3, pp. 288-301.

Fry, A.M. 2002, "The Nek2 protein kinase: a novel regulator of centrosome structure", *Oncogene,* vol. 21, no. 40, pp. 6184-6194.

Fry, A.M., O'Regan, L., Sabir, S.R. & Bayliss, R. 2012, "Cell cycle regulation by the NEK family of protein kinases", *Journal of cell science,* vol. 125, no. 19, pp. 4423-4433.

Fusaki, N., Iwamatsu, A., Iwashima, M. & Fujisawa, J. 1997, "Interaction between Sam68 and Src Family Tyrosine Kinases, Fyn and Lck, in T Cell Receptor Signaling", *Journal of Biological Chemistry,* vol. 272, no. 10, pp. 6214-6219.

Galarneau, A. & Richard, S. 2009, "The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs", *BMC Molecular Biology,* vol. 10.

Galarneau, A. & Richard, S. 2005, "Target RNA motif and target mRNAs of the Quaking STAR protein", *Nature Structural & Molecular Biology,* vol. 12, pp. 691 - 698.

Galarneau, A. & Richard, S. 2009, "The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs", *BMC Molecular Biology,* vol. 10, no. 1, pp. 47.

García-Mayoral, M.F., Hollingworth, D., Masino, L., Díaz-Moreno, I., Kelly, G., Gherzi, R., Chou, C., Chen, C. & Ramos, A. 2007, "The Structure of the C-Terminal KH Domains of KSRP Reveals a Noncanonical Motif Important for mRNA Degradation", *Structure,* vol. 15, no. 4, pp. 485-498.

Ghosh, G. & Adams, J.A. 2011, "Phosphorylation mechanism and structure of serine-arginine protein kinases", *FEBS Journal,* vol. 278, no. 4, pp. 587-597.

Gui, J., Lane, W.S. & Fu, X. 1994, "A serine kinase regulates intracellular localization of splicing factors in the cell cycle", *Nature,* vol. 369, pp. 678 - 682.

Guinier, A. 1939, "La diffraction des rayons X aux très petits angles; application a l'étude de phénomènes ultramicroscopiques", *Annales de Physique Paris,* vol. 12.

Hardy, T., Lee, M., Hames, R.S., Prosser, S.L., Cheary, D., Samant, M.D., Schultz, F., Baxter, J.E., Rhee, K. & Fry, A.M. 2014, "Multisite phosphorylation of C-Nap1 releases

it from Cep135 to trigger centrosome disjunction", *Journal of Cell Science,* vol. 127, no. 11, pp. 2493-2506.

Hartmann, A.M., Nayler, O., Schwaiger, F.W., Obermeier, A. & Stamm, S. 1999, "The Interaction and Colocalization of Sam68 with the Splicing-associated Factor YT521-B in Nuclear Dots Is Regulated by the Src Family Kinase p59fyn", *Molecular biology of the cell,* vol. 10, no. 11, pp. 3909-3926.

Hastie, C.J., McLauchlan, H.J. & Cohen, P. 2006, "Assay of protein kinases using radiolabeled ATP: a protocol", *Nature Protocols,* vol. 1, pp. 968-971.

Hayward, D.G. & Fry, A.M. 2006, "Nek2 kinase in chromosome instability and cancer", *Cancer Letters,* vol. 237, pp. 155-166.

He, J.J., Henao-Mejia, J. & Liu, Y. 2009, "Sam68 functions in nuclear export and translation of HIV-1 RNA", *RNA Biology,* vol. 6, no. 4, pp. 384-386.

Higgins, D.G. & Sharp, P.M. 1988, "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.", *Gene,* vol. 73, no. 1, pp. 237-44.

Hinterberger, M., Pettersson, I. & Steitz, J.A. 1983, "Isolation of small nuclear ribonucleoproteins containing U1, U2, U4, U5, and U6 RNAs.", *Journal of Biological Chemistry,* vol. 258, no. 4, pp. 2604-2613.

Hong, W., Resnick, R.J., Rakowski, C., Shalloway, D., Taylor, S.J. & Blobel, G.A. 2002, "Physical and Functional Interaction Between the Transcriptional Cofactor CBP and the KH Domain Protein Sam68", *Molecular Cancer Research,* vol. 1, no. 1, pp. 48-55.

Huang, H., Lee, T., Tzeng, S. & Horng, J. 2005, "KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites", *Nucleic acids research,* vol. 33, no. suppl 2, pp. W226-W229.

Huot, M., Brown, C.M., Lamarche-Vane, N. & Richard, S. 2009, "An Adaptor Role for Cytoplasmic Sam68 in Modulating Src Activity during Cell Polarization", *Molecular and cellular biology,* vol. 29, no. 7, pp. 1933-1943.

Huot, M. & Richard, S. 2012, "Stay lean without dieting Lose Sam68", *Adipocyte,* vol. 1, no. 4, pp. 246–249.

Huot, M., Vogel, G., Zabarauskas, A., Ngo, C., Coulombe-Huntington, J., Majewski, J. & Richard, S. 2012a, "The Sam68 STAR RNA-Binding Protein Regulates mTOR Alternative Splicing during Adipogenesis", *Molecular cell,* vol. 46, no. 2, pp. 187-199.

Huot, M., Vogel, G., Zabarauskas, A., Ngo, C., Coulombe-Huntington, J., Majewski, J. & Richard, S. 2012b, "The Sam68 STAR RNA-Binding Protein Regulates mTOR Alternative Splicing during Adipogenesis", *Molecular cell,* vol. 46, no. 2, pp. 187-199.

Iijima, T., Wu, K., Witte, H., Hanno-Iijima, Y., Glatter, T., Richard, S. & Scheiffele, P. 2011a, "SAM68 Regulates Neuronal Activity-Dependent Alternative Splicing of Neurexin-1", *Cell,* vol. 147, no. 7, pp. 1601-1614.

International Human Genome Sequencing Consortium 2001, "Initial sequencing and analysis of the human genome", *Nature,* vol. 409, no. 6822, pp. 860-921.

Ishidate, T., Yoshihara, S., Kawasaki, Y., Roy, B.C., Toyoshima, K. & Akiyama, T. 1997, "Identification of a novel nuclear localization signal in Sam68", *FEBS letters,* vol. 409, no. 2, pp. 237-241.

Israeli, D., Nir, R. & Volk, T. 2007, "Dissection of the target specificity of the RNA-binding protein HOW reveals dpp mRNA as a novel HOW target", *Development,* vol. 134, no. 11, pp. 2107-2114.

Itoh, M., Haga, I., Li, Q. & Fujisawa, J. 2002, "Identification of cellular mRNA targets for RNA-binding protein Sam68", *Nucleic acids research,* vol. 30, no. 24, pp. 5452-5464.

Jacques, D.A. & Trewhella, J. 2010, "Small-angle scattering for structural biology?Expanding the frontier while avoiding the pitfalls", *Protein Science,* vol. 19, no. 4, pp. 642-657.

Jager, W., Schaffer, A., Puhler, G., Labes, G. & Wohlleben, W. 1992, "Expression of the Bacillus subtilis sacB Gene Leads to Sucrose Sensitivity in the Gram-Positive Bacterium Corynebactenium glutamicum but Not in Streptomyces lividans", *Journal of Bacteriology,* vol. 174, no. 16, pp. 5462-5465.

Javier Lopez, A. 1998, "ALTERNATIVE SPLICING OF PRE-mRNA: Developmental Consequences and Mechanisms of Regulation", *Annual Review of Genetics,* vol. 32, pp. 279 -305.

Jean-Philippe, J., Paz, S. & Caputi, M. 2013, "hnRNP A1: The Swiss Army Knife of Gene Expression", *International Journal of Molecular Sciences,* vol. 14, no. 9, pp. 18999-19024.

Jones, A.R. & Schedl, T. 1995, "Mutations in gld-1, a female germ cell-specific tumor suppressor gene in Caenorhabditis elegans, affect a conserved domain also found in Src-associated protein Sam68.", *Genes & development,* vol. 9, no. 12, pp. 1491-1504.

Kamma, H., Portman, D.S. & Dreyfuss, G. 1995, "Cell Type-Specific Expression of hnRNP Proteins", *Experimental cell research,* vol. 221, no. 1, pp. 187-196.

Kannan, N. & Neuwald, A.F. 2004, "Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2?", *Protein Science,* vol. 13, no. 8, pp. 2059-2077.

Kay, L.E., Torchia, D.A. & Bax, A. 1989, "Backbone Dynamics of Proteins As Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease", *Biochemistry,* vol. 28, pp. 8972-8979.

Kim, E., Magen, A. & Ast, G. 2007, "Different levels of alternative splicing among eukaryotes", *Nucleic acids research,* vol. 35, no. 1, pp. 125-131.

Krainer, A.R., Conway, G.C. & Kozak, D. 1990, "The essential pre-mRNA splicing factor SF2 influences 5′ splice site selection by activating proximal sites", *Cell,* vol. 62, no. 1, pp. 35-42.

Krausse, I.R., Sica, F., Mattia, C.A. & Merlino, A. 2012, "Increasing the X-ray Diffraction Power of Protein Crystals by Dehydration: The Case of Bovine Serum Albumin and a Survey of Literature Data", *International Journal of Molecular Sciences,* vol. 13, no. 3, pp. 3782–3800.

Lamond, A.I. & Spector, D.L. 2003, "Nuclear speckles: a model for nuclear organelles", *Nature Reviews Molecular Cell Biology,* vol. 4, pp. 605-612.

Laroia, G. & Schneider, R.J. 2002, "Alternate exon insertion controls selective ubiquitination and degradation of different AUF1 protein isoforms", *Nucleic acids research,* vol. 30, no. 14, pp. 3052-3058.

Lawe, D.C., Hahn, C. & Wong, A.J. 1997, "The Nck SH2/SH3 adaptor protein is present in the nucleus and associates with the nuclear protein SAM68", *Oncogene,* vol. 14, no. 2, pp. 223-231.

Lazer, G., Pe'er, L., Schapira, V., Richard, S. & Katzav, S. 2007, "The association of Sam68 with Vav1 contributes to tumorigenesis", *Cellular signalling,* vol. 19, no. 12, pp. 2479-2486.

Lewis, H.A., Chen, H., Edo, C., Buckanovich, R.J., Yang, Y.Y., Musunuru, K., Zhong, R., Darnell, R.B. & Burley, S.K. 1999, "Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains", *Structure,* vol. 7, no. 2, pp. 191-203.

Lewis, H.A., Musunuru, K., Jensen, K.B., Edo, C., Chen, H., Darnell, R.B. & Burley, S.K. 2000, "Sequence-Specific RNA Binding by a Nova KH Domain: Implications for Paraneoplastic Disease and the Fragile X Syndrome", *Cell,* vol. 100, no. 3, pp. 323-332.

Li, Q.H., Haga, I., Shimizu, T., Itoh, M., Kurosaki, T. & Fujisawa, J. 2002, "Retardation of the G2-M phase progression on gene disruption of RNA binding protein Sam68 in the DT40 cell line", *Federation of European Biochemical Societies Letters,* vol. 525, pp. 145-150.

Li, T., Evdokimov, E., Shen, R.F., Chao, C.C., Tekle, E., Wang, T., Stadtman, E.R., Yang, D.C. & Chock, P.B. 2004, "Sumoylation of heterogeneous nuclear ribonucleoproteins, zinc finger proteins, and nuclear pore complex proteins: a proteomic analysis", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. 23, pp. 8551-6.

Li, J., Liu, Y., Kim, B.O. & He, J.J. 2002a, "Direct Participation of Sam68, the 68-Kilodalton Src-Associated Protein in Mitosis, in the CRM1-Mediated Rev Nuclear Export Pathway", *Journal of virology,* vol. 76, no. 16, pp. 8374-8382.

Li, J., Liu, Y., Park, I. & He, J.J. 2002b, "Expression of Exogenous Sam68, the 68-Kilodalton Src-Associated Protein in Mitosis, Is Able To Alleviate Impaired Rev Function in Astrocytes", *Journal of virology,* vol. 76, no. 9, pp. 4526-4535.

Li, Z., Yu, C.-., Zhong, Y., Liu, T.-., Huang, Q.-., Zhao, X.-., Huang, H., Tu, H., Jiang, S., Zhang, Y., Liu, J.-. & Song, L.-. 2012, "Sam68 expression and cytoplasmic localization is correlated with lymph node metastasis as well as prognosis in patients with early-stage cervical cancer", *Annals of Oncology,* vol. 23, no. 3, pp. 638-646.

Liao, W., Liu, J., Wang, Z., Cui, Y., Shi, L., Li, T., Zhao, X., Chen, X., Ding, Y. & Song, L. 2013, "High expression level and nuclear localization of Sam68 are associated with progression and poor prognosis in colorectal cancer", *BMC Gastroenterology,* vol. 13, no. 126.

Lin, Q., Taylor, S.J. & Shalloway, D. 1997, "Specificity and Determinants of Sam68 RNA Binding", *The Journal of Biological Chemistry,* vol. 272, no. 43, pp. 27274-27280.

Lippens, G., Landrieu, I. & Hanoulle, X. 2008, "Studying Posttranslational Modifications by In-Cell NMR", *Chemistry & biology*, vol. 15, no. 4, pp. 311-312.

Liu, Z., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Kramer, A. & Sattler, M. 2001, "Structural basis for recognition of the intron branch site RNA by splicing factor 1", *Science*, vol. 294, no. 5544, pp. 1098-102.

Liu, K., Li, L., Nisson, P.E., Gruber, C., Jessee, J. & Cohen, S.N. 2000, "Neoplastic transformation and tumorigenesis associated with Sam68 protein deficiency in cultured murine fibroblasts", *Journal of Biological Chemistry*, .

Locatelli, A., Lofgren, K.A., Daniel, A.R., Castro, N.E. & Lange, C.A. 2011, "Mechanisms of HGF/Met Signalling to Brk and Sam68 in Breast Cancer Progression", *H*, vol. 10.

Locatelli, A. & Lange, C.A. 2011, "Met Receptors Induce Sam68-dependent Cell Migration by Activation of Alternate Extracellular Signal-regulated Kinase Family Members", *Journal of Biological Chemistry*, vol. 286, no. 24, pp. 21062-21072.

Lock, P., Fumagalli, S., Polakis, P., McCormick, F. & Courtneidge, S.A. 1996, "The Human p62 cDNA Encodes Sam68 and Not the RasGAP-Associated p62 Protein", *Cell*, vol. 84, no. 1, pp. 23-24.

Lukong, K.E., Chang, K., Khandijan, E.W. & Richard, S. 2008, "RNA-binding proteins in human genetic disease", *Trends in Genetics*, vol. 24, no. 8, pp. 416-425.

Lukong, K.E. & Richard, S. 2003, "Sam68, the KH-domain containing superSTAR", *Biochimica et Biophyisica Acta*, , pp. 73-86.

Lukong, K.E. & Richard, S. 2008, "Motor coordination defects in mice deficient for the Sam68 RNA-binding protein", *Behavioural brain research*, vol. 189, no. 2, pp. 357-363.

Lukong, K.E., Larocque, D., Tyner, A.L. & Richard, S. 2005, "Tyrosine Phosphorylation of Sam68 by Breast Tumor Kinase Regulates Intranuclear Localization and Cell Cycle Progression", *Journal of Biological Chemistry*, vol. 280, no. 46, pp. 38639-38647.

Maa, M.C., Leu, T.H., Trandel, B.J., Chang, J.H. & Parsons, S.J. 1994, "A protein that is highly related to GTPase-activating protein-associated p62 complexes with phospholipase C gamma.", *Molecular and cellular biology*, vol. 14, no. 8, pp. 5466-5473.

Mackereth, C.D. & Sattler, M. 2012, "Dynamics in multi-domain protein recognition of RNA", *Current opinion in structural biology*, vol. 22, no. 3, pp. 287-296.

Madl, T., Guttler, T., Gorlich, D. and Sattler, M., 2011, "Structural Analysis of Large Protein Complexes Using Solvent Paramagnetic Relaxation Enhancements", *Angewandte Chemie*, vol. 50, no. 17, pp 3993-3997

Maguire, M.L., Guler-Gane, G., Nietlispach, D., Raine, A.R.C., Zorn, A.M., Standart, N. & Broadhurst, R.W. 2005, "Solution Structure and Backbone Dynamics of the KH-QUA2 Region of the Xenopus STAR/GSG Quaking Protein", *Journal of Molecular Biology*, vol. 348, no. 2, pp. 265-279.

Maris, C., Dominguez, C. & Allain, F.H.-. 2005, "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression", *FEBS Journal*, vol. 272, no. 9, pp. 2118-2131.

Maroni, P., Citterio, L., Piccoletti, R. & Bendinelli, P. 2009, "Sam68 and ERKs regulate leptin-induced expression of OB-Rb mRNA in C2C12 myotubes", *Molecular and cellular endocrinology*, vol. 309, no. 1–2, pp. 26-31.

Martín-Romero, C. & Sánchez-Margalet, V. 2001, "Human Leptin Activates PI3K and MAPK Pathways in Human Peripheral Blood Mononuclear Cells: Possible Role of Sam68", *Cellular immunology*, vol. 212, no. 2, pp. 83-91.

Matlin, A.J., Clark, F. & Smith, C.W.J. 2005, "Understanding alternative splicing: towards a cellular code", *Nature Reviews Molecular Cell Biology*, vol. 6, pp. 386-398.

Matter, N., Herrlich, P. & Konig, H. 2002, "Signal-dependent regulation of splicing via phosphorylation of Sam68", *Nature*, vol. 420, pp. 691-695.

Mclaren, M., Asai, K. & Cochrane, A. 2004, "A novel function for Sam68: Enhancement of HIV-1 RNA 3' end processing", *RNA*, vol. 10, no. 7, pp. 1119-1129.

Meyer, N.H., Tripsianes, K., Vincendeau, M., Madl, T., Kateb, F., Brack-Werner, R. & Sattler, M. 2010a, "Structural Basis for Homodimerization of the Src-associated during Mitosis, 68-kDa Protein (Sam68) Qua1 Domain", *Journal of Biological Chemistry*, vol. 285, no. 37, pp. 28893-28901.

Modem, S., Badri, K.R., Holland, T.C. & Reddy, T.R. 2005, "Sam68 is absolutely required for Rev function and HIV-1 production", *Nucleic Acids Research*, vol. 33, no. 3, pp. 873-879.

Modem, S., Chinnakannu, K., Bai, U., Reddy, G.P. & Reddy, T.R. 2011, "Hsp22 (HspB8/H11) knockdown induces sam68 expression and stimulates proliferation of glioblastoma cells", *Journal of cellular physiology,* vol. 226, no. 11, pp. 2747-2751.

Moore, M.J., Wang, Q., Kennedy, C.J. & Silver, P.A. 2010, "An alternative splicing network links cell-cycle control to apoptosis", *Cell,* vol. 142, no. 4, pp. 625-636.

Morishita, E., Murayama, K., Kato-Murayama, M., Ishizuka-Katsura, Y., Tomabechi, Y., Hayashi, T., Terada, T., Handa, N., Shirouzu, M., Akiyama, T. & Yokoyama, S. 2011, "Crystal Structures of the Armadillo Repeat Domain of Adenomatous Polyposis Coli and Its Complex with the Tyrosine-Rich Domain of Sam68", *Structure,* vol. 19, no. 10, pp. 1496-1508.

Musco, G., Stier, G., Joseph, C., Morelli, M.A.C., Nilges, M., Gibson, T.J. & Pastore, A. 1996, "Three-Dimensional Structure and Stability of the KH Domain: Molecular Insights into the Fragile X Syndrome", *Cell,* vol. 85, no. 2, pp. 237-245.

Najib, S., Martin-Romero, C., Gonzalez-Yanes, C. & Sanchez-Margalet, V. 2005a, "Role of Sam68 as an adapter protein in signal transduction", *Cellular and Molecular Life Sciences,* vol. 62, pp. 36-43.

Najib, S., Rodríguez-Baño, J., Ríos, M.J., Muniain, M.A., Goberna, R. & Sánchez-Margalet, V. 2005b, "Sam68 is tyrosine phosphorylated and recruited to signalling in peripheral blood mononuclear cells from HIV infected patients", *Clinical & Experimental Immunology,* vol. 141, no. 3, pp. 518-525.

Naro, C., Barbagallo, F., Chieffi, P., Bourgeois, C.F., Paronetto, M.P. & Sette, C. 2013, "The centrosomal kinase NEK2 is a novel splicing factor kinase involved in cell survival", *Nucleic acids research.*

Nicastro, G., Taylor, I.A., Ramos, A. 2015, "KH-RNA interactions: back in the groove", *Current Opinion in Structural Biology,* vol. 30, pp 63-70.

Nigg, E.A. 1991, "The substrates of the cdc2 kinase", *Seminars in Cell Biology,* vol. 2, no. 4, pp. 261-70.

Okazaki, K. & Niwa, O. 2000, "mRNAs Encoding Zinc Finger Protein Isoforms are Expressed by Alternative Splicing of an In-frame Intron in Fission Yeast", *DNA Research,* vol. 7, no. 1, pp. 27-30.

Ozawa, E., Mizuno, Y., Hagiwara, Y., Sasaoka, T. & Yoshida, M. 2005, "Molecular and cell biology of the sarcoglycan complex", *Muscle & nerve,* vol. 32, no. 5, pp. 563-576.

Paronetto, M.P., Cappellari, M. & Busa, R. 2010, "Alternative Splicing of the Cyclin D1 Proto-Oncogene Is Regulated by the RNA-Binding Protein Sam68", *Cancer Research,* vol. 70, pp. 229-239.

Paronetto, M.P., Achsel, T., Massiello, A., Chalfant, C.E. & Sette, C. 2007, "The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x", *The Journal of cell biology*, vol. 176, no. 7, pp. 929-939.

Paronetto, M.P., Messina, V., Barchi, M., Geremia, R., Richard, S. & Sette, C. 2011, "Sam68 marks the transcriptionally active stages of spermatogenesis and modulates alternative splicing in male germ cells", *Nucleic acids research,* vol. 39, no. 12, pp. 4961-4974.

Paronetto, M.P., Messina, V., Bianchi, E., Barchi, M., Vogel, G., Moretti, C., Palombi, F., Stefanini, M., Geremia, R., Richard, S. & Sette, C. 2009, "Sam68 regulates translation of target mRNAs in male germ cells, necessary for mouse spermatogenesis", *The Journal of cell biology,* vol. 185, no. 2, pp. 235-249.

Pedrotti, S., Bielli, P., Paronetto, M.P., Ciccosanti, F., Fimia, G.M., Stamm, S., Manley, J.L. & Sette, C. 2010, "The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy", *The EMBO Journal,* vol. 29, pp. 1235-1247.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V. & Svergun, D.I. 2012, "New developments in the ATSAS program package for small-angle scattering data analysis", *Journal of Applied Crystallography,* vol. 45, pp. 342–350.

Pillay, I., Nakano, H. & Sharma, S. 1996, "Radicicol inhibits tyrosine phosphorylation of the mitotic Src substrate Sam68 and retards subsequent exit from mitosis of Src-transformed cells", *Cell Growth Differentiation,* vol. 7, no. 11, pp. 1487-1499.

Pinol-Roma, S. & Dreyfuss, G. 1992, "Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm", *Nature 3 Serafm Piñol-Roma & Gideon Dreyfuss*,* vol. 355, pp. 730 - 732.

Rajan, P., Gaughan, L., Dalgliesh, C., El-Sherif, A., Robson, C., Leung, H. & Elliott, D. 2008, "The RNA-binding and adaptor protein Sam68 modulates signal-dependent splicing and transcriptional activity of the androgen receptor", *The Journal of pathology,* vol. 215, no. 1, pp. 67-77.

Rajpurohit, R., Paik, W.K. & Kim, S. 1992, "Enzymatic methylation of heterogeneous nuclear ribonucleoprotein in isolated liver nuclei", *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, vol. 1122, no. 2, pp. 183-188.

Ramos, A., Grünert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D. & Varani, G. 2000, "RNA recognition by a Staufen double-stranded RNA-binding domain", *EMBO J.*, vol. 19, no. 5, pp. 997-1009.

Ramos, A., Hollingworth, D., Major, S., Adinolfi, S., Kelly, G., Muskett, F. & Pastore, A. 2002, "Role of dimerization in KH/RNA complexes: The example of Nova KH3", *Biochemistry*, vol. 41, no. 13, pp. 4193-4201.

Reddy, T.R., Xu, W., Mau, J.K.L., Goodwin, C.D., Suhasini, M., Tang, H., Frimpong, K., Rose, D.W. & Wong-Staal, F. 1999, "Inhibition of HIV replication by dominant negative mutants of Sam68, a functional homolog of HIV-1 Rev", *Nature Medicine*, vol. 5, pp. 635 - 642.

Resnick, R.J., Taylor, S.J., Lin, Q. & Shalloway, D. 1997, "Phosphorylation of the Src substrate Sam68 by Cdc2 during mitosis", *Oncogene*, vol. 15, pp. 1247-1253.

Revil, T., Pelletier, J., Toutant, J., Cloutier, A. & Chabot, B. 2009, "Heterogeneous Nuclear Ribonucleoprotein K Represses the Production of Pro-apoptotic Bcl-xS Splice Isoform", *Journal of Biological Chemistry*, vol. 284, no. 32, pp. 21458-21467.

Rho, J., Choi, S., Jung, C. & Im, D. 2007, "Arginine methylation of Sam68 and SLM proteins negatively regulates their poly(U) RNA binding activity", *Archives of Biochemistry and Biophysics*, vol. 466, no. 1, pp. 49-57.

Richard, S., Torabi, N., Franco, G.V., Tremblay, G.A., Chen, T., Vogel, G., Morel, M., Cleroux, P., Forget-Richard, A., Komarova, S. & Tremblay, M. 2005, "Ablation of the Sam68 RNA binding protein protects mice from age-related bone loss", *PLOS Genetics*, vol. 1, no. 6.

Richard, S., Yu, D., Blumer, K.J., Hausladen, D., Olszowy, M.W., Connelly, P.A. & Shaw, A.S. 1995, "Association of p62, a multifunctional SH2- and SH3-domain-binding protein, with src family tyrosine kinases, Grb2, and phospholipase C gamma-1.", *Molecular and cellular biology*, vol. 15, no. 1, pp. 186-197.

Ryder, S.P., Frater, L.A., Abramov, D.L., Goodwin, E.B. & Williamson, J.R. 2003, "RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1", *Nature Structural & Molecular Biology*, vol. 11, pp. 20-28.

Ryder, S.P. & Massi, F. 2010, "Insights into the structural basis of RNA recognition by STAR domain proteins", *Advances in Experimental Medicine and Biology*, vol. 693, pp. 37-53.

Ryder, S.P. & Williamson, J.R. 2004, "Specificity of the STAR/GSG domain protein Qk1: Implications for the regulation of myelination", *RNA*, vol. 10, no. 9, pp. 1449-1458.

Sahin, E. & Roberts, C.J. 2012, "Size-Exclusion Chromatography with Multi-angle Light Scattering for Elucidating Protein Aggregation Mechanisms", *Methods in Molecular Biology*, vol. 899, pp. 403-423.

Sattler, M., Schleucher, J. & Griesinger, C. 1999, "Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients", *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 34, pp. 93-158.

Shaw Stewart, P.D., Kolek, S.A., Briggs, R.A., Chayen, N.E. & Baldock, P.F.M. 2011, "Random Microseeding: A Theoretical and Practical Exploration of Seed Stability and Seeding Techniques for Successful Protein Crystallization", *Crystal Growth and Design*, vol. 11, no. 8, pp. 3432-3441.

Shen, E.C., Henry, M.F., Weiss, V.H., Valentini, S.R., Silver, P.A. & Lee, M.S. 1998, *Arginine methylation facilitates the nuclear export of hnRNP proteins*.

Shin, C. & Manley, J.L. 2004, "Cell Signalling And The Control Of Pre-mRNA Splicing", *Nature Reviews Molecular Cell Biology*, vol. 5, pp. 727-738.

Singh, R.K. & Cooper, T.A. 2012, "Pre-mRNA splicing in disease and therapeutics", *Trends in Molecular Medicine*, vol. 18, no. 8, pp. 472-482.

Sinha, R., Allemand, E., Zhang, Z., Karni, R., Myers, M.P. & Krainer, A.R. 2010, "Arginine Methylation Controls the Subcellular Localization and Functions of the Oncoprotein Splicing Factor SF2/ASF", *Molecular and cellular biology*, vol. 30, no. 11, pp. 2762-2774.

Siomi, H., Choi, M., Siomi, M.C., Nussbaum, R.L. & Dreyfuss, G. 1994, "Essential role for KH domains in RNA binding: Impaired RNA binding by a mutation in the KH domain of FMR1 that causes fragile X syndrome", *Cell*, vol. 77, no. 1, pp. 33-39.

Siomi, H., Matunis, M.J., Michael, W.M. & Dreyfuss, G. 1993a, "The pre-mRNA binding K protein contains a novel evolutionary conserved motif", *Nucleic acids research*, vol. 21, no. 5, pp. 1193-1198.

Siomi, H., Matunis, M.J., Michael, W.M. & Dreyfuss, G. 1993b, "The pre-mRNA binding K protein contains a novel evolutionary conserved motif", *Nucleic acids research*, vol. 21, no. 5, pp. 1193-1198.

Smith, C.W.J. & Valcárcel, J. 2000, "Alternative pre-mRNA splicing: the logic of combinatorial control", *Trends in biochemical sciences*, vol. 25, no. 8, pp. 381-388.

Song, L., Wang, L., Li, Y., Xiong, H., Wu, J., Li, J. & Li, M. 2010, "Sam68 up-regulation correlates with, and its down-regulation inhibits, proliferation and tumourigenicity of breast cancer cells", *The Journal of pathology*, vol. 222, no. 3, pp. 227-237.

Soros, V.B., Carvajal, H.V., Richard, S. & Cochrane, A.W. 2001, *Inhibition of Human Immunodeficiency Virus Type 1 Rev Function by a Dominant-Negative Mutant of Sam68 through Sequestration of Unspliced RNA at Perinuclear Bundles*.

Taylor, S.J., Resnick, R.J. & Shalloway, D. 2004, "Sam68 exerts separable effects on cell cycle progression and apoptosis", *BMC Cell Biology*, vol. 5, no. 5.

Taylor, S.J. & Shalloway, D. 1994, "An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis", *Nature*, vol. 368, pp. 867 - 871.

Taylor, S.J., Anafi, M., Pawson, T. & Shalloway, D. 1995, "Functional Interaction between c-Src and Its Mitotic Target, Sam 68", *Journal of Biological Chemistry*, vol. 270, no. 17, pp. 10120-10124.

Tazi, J., Bakkour, N. & Stamm, S. 2009, "Alternative splicing and disease", *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1792, no. 1, pp. 14-26.

Teplova, M., Hafner, M., Teplov, D., Essig, K., Tuschl, T. & Patel, D.J. 2013, "Structure–function studies of STAR family Quaking proteins bound to their in vivo RNA target sites", *Genes & development*, vol. 27, no. 8, pp. 928-940.

Theillet, F., Rose, M.R., Liokatis, S., Binolfi, A., Thongwichian, R., Stuiver, M. & Selenko, P. 2013, "Site-specific NMR mapping and time-resolved monitoring of serine and threonine phosphorylation in reconstituted kinase reactions and mammalian cell extracts", *Nature Protocols*, vol. 8, pp. 1416–1432.

Thiery, J.P. 2002, "Epithelial–mesenchymal transitions in tumour progression", *Nature Reviews Cancer*, vol. 2, pp. 442-454.

Tremblay, G.A. & Richard, S. 2005, "mRNAs Associated with the Sam68 RNA Binding Protein", *RNA Biology*, vol. 2, no. 2.

Trüb, T., Frantz, J.D., Miyazaki, M., Band, H. & Shoelson, S.E. 1997, "The Role of a Lymphoid-restricted, Grb2-like SH3-SH2-SH3 Protein in T Cell Receptor Signaling", *Journal of Biological Chemistry*, vol. 272, no. 2, pp. 894-902.

Tuerk, C. & Gold, L. 1990, "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase", *Science*, vol. 249, no. 4968, pp. 505-510.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A. & Darnell, R.B. 2003, "CLIP Identifies Nova-Regulated RNA Networks in the Brain", *Science*, vol. 302, no. 5648, pp. 1212-1215.

Ule, J., Jensen, K., Mele, A. & Darnell, R.B. 2005, "CLIP: A method for identifying protein–RNA interaction sites in living cells", *Methods*, vol. 37, no. 4, pp. 376-386.

Ushkaryov, Y.A., Petrenko, A.G., Geppert, M. & Südhof, T.C. 1992, "Neurexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin.", *Science*, vol. 257, no. 5006, pp. 50-56.

Vagenende, V., Yap, M.G.S. & Trout, B.L. 2009, "Mechanisms of Protein Stabilization and Prevention of Protein Aggregation by Glycerol", *Biochemistry*, vol. 48, no. 46, pp. 11084-11096.

Valacca, C., Bonomi, S., Buratti, E., Pedrotti, S., Baralle, F.E., Sette, C., Ghigna, C. & Biamonti, G. 2010, "Sam68 regulates EMT through alternative splicing-activated nonsense-mediated mRNA decay of the SF2/ASF proto-oncogene.", *Journal of Cell Biology*, vol. 191, no. 1, pp. 87-99.

Valverde, R., Edwards, L. & Regan, L. 2008, "Structure and function of KH domains", *FEBS Journal*, vol. 275, no. 11, pp. 2712-2726.

Venables, J.P., Vernet, C., Chew, S.L., Elliott, D.J., Cowmeadow, R.B., Wu, J., Cooke, H.J., Artzt, K. & Eperon, I.C. 1999, "T-STAR/ÉTOILE: A Novel Relative of SAM68 That Interacts with an RNA-Binding Protein Implicated in Spermatogenesis", *Human molecular genetics*, vol. 8, no. 6, pp. 959-969.

Vernet, C. & Artzt, K. 1997, "STAR, a gene family involved in signal transduction and activation of RNA", *Trends in Genetics*, vol. 13, no. 12, pp. 479-484.

Volk, T. & Artzt, K.J. 2010, *Post-transcriptional regulation by STAR proteins: control of RNA metabolism in development and disease*, Springer Science+Business Media, LLC, New York; Austin, Tex.

Wang, B.B. & Brendel, V. 2006, "Genomewide comparative analysis of alternative splicing in plants", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 18, pp. 7175-7180.

Watson, J. & Crick, F. 1953, "A Structure for Deoxyribose Nucleic Acid", *Nature*, vol. 171, pp. 737-738.

Weg-Remers, S., Ponta, H., Herrlich, P. & Konig, H. 2001, "Regulation of alternative pre-mRNA splicing by the ERK MAP-kinase pathway", *The EMBO Journal*, vol. 20, pp. 4194-4203.

Wong, G., Müller, O., Clark, R., Conroy, L., Moran, M.F., Polakis, P. & McCormick, F. 1992, "Molecular cloning and nucleic acid binding properties of the GAP-associated tyrosine phosphoprotein p62", *Cell*, vol. 69, no. 3, pp. 551-558.

Wu, J.Y. & Maniatis, T. 1993, "Specific interactions between proteins implicated in splice site selection and regulated alternative splicing", *Cell*, vol. 75, no. 6, pp. 1061-1070.

Yang, Z.R., Thomson, R., McMeil, P. & Esnouf, R.M. 2005, "RONN: the bio-basis function neural network technique applied to the dectection of natively disordered regions in proteins", *Bioinformatics*, vol. 21, pp. 3369-3376.

Zhang, Z., Xu, Y., Sun, N., Zhang, M., Xie, J. & Jiang, Z. February 2014, "High Sam68 expression predicts poor prognosis in non-small cell lung cancer", *Clinical and Translational Oncology*, .

Zhang, Z., Li, J., Zheng, H., Yu, C., Chen, J., Liu, Z., Li, M., Zeng, M., Zhou, F. & Song, L. 2009, "Expression and Cytoplasmic Localization of SAM68 Is a Significant and Independent Prognostic Marker for Renal Cell Carcinoma", *Cancer Epidemiology Biomarkers & Prevention*, vol. 18, no. 10, pp. 2685-2693.

Zhou, Z. & Fu, X. 2013, "Regulation of splicing by SR proteins and SR protein-specific kinases", *Chromosoma*, vol. 122, no. 3, pp. 191-207.

Zuiderweg, E.R.P. 2002, "Mapping Protein−Protein Interactions in Solution by NMR Spectroscopy", *Biochemistry,* vol. 41, no. 1, pp. 1-7.

Methods 65 (2014) 288–301

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Screening protein – Single stranded RNA complexes by NMR spectroscopy for structure determination

Jaelle N. Foot [1], Mikael Feracci [1], Cyril Dominguez *

Department of Biochemistry, Henry Wellcome Laboratories of Structural Biology, University of Leicester, UK

ARTICLE INFO

Article history:
Available online 1 October 2013

Keywords:
NMR spectroscopy
Protein–RNA complex
STAR proteins
Sam68
T-STAR

ABSTRACT

In the past few years, RNA molecules have been revealed to be at the center of numerous biological processes. Long considered as passive molecules transferring genetic information from DNA to proteins, it is now well established that RNA molecules play important regulatory roles. Associated with that, the number of identified RNA binding proteins (RBPs) has increased considerably and mutations in RNA molecules or RBP have been shown to cause various diseases, such as cancers. It is therefore crucial to understand at the molecular level how these proteins specifically recognise their RNA targets in order to design new generation drug therapies targeting protein–RNA complexes. Nuclear magnetic resonance (NMR) is a particularly well-suited technique to study such protein–RNA complexes at the atomic level and can provide valuable information for new drug discovery programs. In this article, we describe the NMR strategy that we and other laboratories use for screening optimal conditions necessary for structural studies of protein-single stranded RNA complexes, using two proteins, Sam68 and T-STAR, as examples.

1. Introduction

While RNA molecules have long been considered as passive molecules that transfer information from genes to proteins, the last few years have seen the emergence of a massive but still poorly understood RNA world. For example, recent studies from the ENCODE project (http://encodeproject.org) suggested that, while only 1.5% of our genome corresponds to protein-coding sequences, between 20% and 80% of it is transcribed into RNA [1]. It is now clear that RNA molecules are highly abundant and play crucial roles in multiple cellular functions [2,3]. Associated with this, the number of RNA binding proteins (RBPs) identified has also increased significantly in the last decade. Recently, more than 800 human proteins have been identified that directly bind messenger RNAs (mRNAs) [4,5]. While half of these proteins contain well-known RNA binding domains (RBDs) such as the RNA recognition motif (RRM), the double-stranded RNA binding domain (dsRBD) or the hnRNP K homology (KH) domain, the other half were not previously predicted to be RNA binding proteins. Mutations found in RNA and RBPs have been shown to cause numerous diseases such as neurological disorders, genetic diseases and cancers [6–8]. It is therefore crucial to obtain structural information of these protein–RNA complexes in order to, first understand the specificity of recognition, and second to target these complexes for novel therapeutic strategies.

RNA molecules are single stranded in cells and a majority of RBPs recognise and bind short single-stranded RNA (ssRNA) motifs through specific contacts with the nucleic acid bases. The two major techniques to solve structures of macromolecular complexes such as protein-ssRNA complexes are X-ray crystallography and nuclear magnetic resonance (NMR). While overall, NMR contributes to around 10% of the structures deposited into the protein data bank (PDB), the contribution of NMR for structure determination of protein-ssRNA complexes is 53% (out of 62 protein-ssRNA complexes, 33 were solved by NMR and 29 by X-ray crystallography). This emphasises that NMR is a major technique for the investigation of such complexes. This fact can be explained by numerous intrinsic properties of protein-ssRNA complexes [9–11]. First, most known RBPs contain small globular RBDs, such as the RRM or the KH domains, that are around 100 amino acids in length and therefore suitable for NMR studies. Second, single-stranded RNA molecules are highly flexible which can interfere with the crystallisation process of such complexes. Third, many RBPs are modular proteins containing more than one RBD separated by flexible linkers. The presence of such flexible regions and the lack of well-defined relative orientation of the RBDs can also prevent formation of crystals. Additionally, while full-length RBPs often bind RNA with high affinity, they act through a modular interaction

* Corresponding author. Address: Department of Biochemistry, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 9HN, UK.
E-mail address: cyril.dominguez@leicester.ac.uk (C. Dominguez).
[1] These authors contributed equally to this work.

1046-2023
http://dx.doi.org/10.1016/j.ymeth.2013.09.018

approach where each RBD binds rather weakly to its RNA substrate (sometimes in the millimolar range) and the high affinity is provided by the presence of multiple RBDs within one RBP. Therefore RBD-ssRNA complexes are often dynamic and can prevent the formation of a well-defined crystal of the complex, while still being suitable for NMR studies. Finally, although new methodological developments have allowed a precise definition of the RNA sequence specifically recognised by RBPs or RBDs [12], these sequences are often degenerate and the identification of the optimal ssRNA sequence for structural studies is far from being straightforward. From that point of view NMR is very powerful because it allows the screening of multiple RNA sequences at an early stage of the structural investigation, as will be detailed in this article for two proteins, Sam68 and T-STAR, that belong to the STAR family of proteins [13,14]. Sam68 is the best-characterised member of this family and is involved in various post-transcriptional regulation events, such as alternative splicing and RNA export [15,16]. T-STAR, also known as SLM2, is closely related to Sam68 but its biological function is less well characterised [17,18]. T-STAR was recently identified as a specific neuronal splicing factor [19]. STAR proteins are characterised by the presence of a STAR domain necessary for RNA binding [13]. This domain can be subdivided into a central KH domain flanked by highly conserved regions, QUA1 and QUA2 (Fig. 1). Previous structural studies on other STAR proteins indicated that the KH-QUA2 region of the STAR domain is involved in RNA binding and the QUA1 region is involved in the homo-dimerisation of the protein [20–24]. While KH domains generally accommodate 4 nucleotides [25], the NMR structure of SF1 KH-QUA2 demonstrated that the QUA2 region adopts an α-helical conformation packed against the KH domain and interacts with three additional nucleotides [20]. This large RNA interaction involving the QUA2 region was recently confirmed by the X-ray structures of GLD-1 and Quaking STAR domains in complex with RNA [22]. In that case, the KH-QUA2 region accommodates five and six nucleotides, respectively. While SF1, Quaking and GLD-1 specifically bind similar RNA sequences containing a CUAAC motif, SELEX data indicate that Sam68 and T-STAR specifically recognise A/U rich RNA sequences [26,27]. Consistently, sequence alignment between STAR proteins suggested that the RNA binding mode of Sam68 and T-STAR might be different to other STAR proteins [28]. Additionally, because the QUA2 amino acids of Quaking and GLD-1 that interact with the RNA are not conserved in Sam68 and T-STAR, it has been proposed that the QUA2 region of Sam68 might not be involved in RNA binding [28]. This is supported by previous data showing that a construct of Sam68, QUA1-KH, lacking the QUA2 region is able to bind RNA as well as the full-length protein [26] and SELEX data indicating that Sam68 specifically binds four nucleotides as opposed to the six nucleotides bound by GLD-1, and QKI [27].

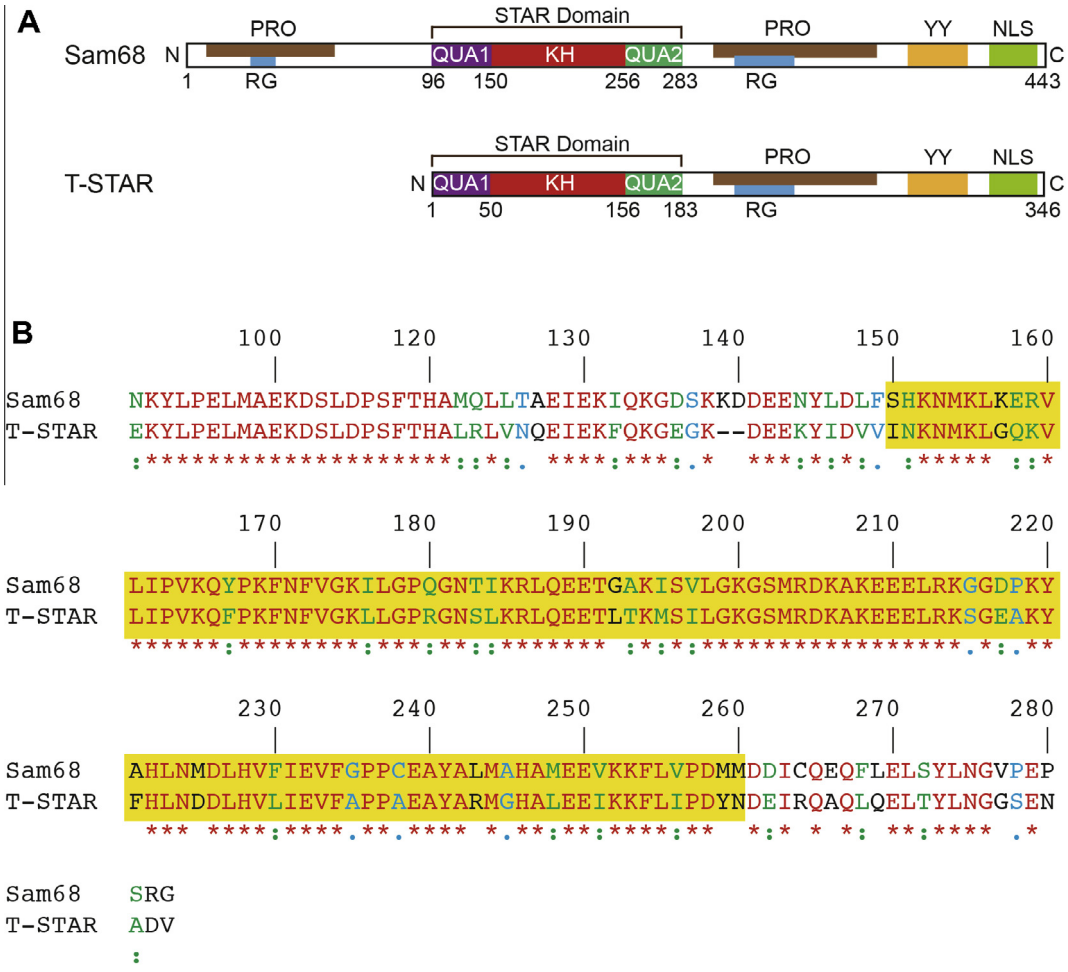


Fig. 1. Domain organisation of Sam68 and T-STAR. (A) Sam68 and T-STAR contain a STAR domain responsible for RNA binding that is composed of a central KH domain flanked by QUA1 and QUA2 regions. In addition, these proteins contain various motifs necessary for the cellular function of the protein. The nuclear localisation signal (NLS) allows the proteins to shuttle between the cytoplasm and the nucleus. The proline-rich (PRO) and tyrosine-rich regions (YY) are necessary for tyrosine phosphorylation of these proteins and the arginine–glycine rich regions (RG) are target sites for arginine methylation. (B) Sequence alignment of Sam68 and T-STAR STAR domains. The amino acids of these two proteins in this region are 69% identical (red), 16% highly homologous (green) and 5% homologous (blue). The alignment was performed using CLUSTALW. The KH domain is highlighted in yellow.

In this article, we will describe the strategy that we and other NMR laboratories commonly use to define optimal protein constructs and RNA sequences for structural studies, using Sam68 and T-STAR proteins as examples.

2. Materials and methods

2.1. RNA production

The RNA oligonucleotides used for NMR studies were chemically synthesised at a 1 micromole scale (Dharmacon, Thermo Scientific), deprotected according to manufacturer instructions, and lyophilised. RNAs were then resuspended in 100 µl of water and pH was adjusted to 6.5 or 7.0. RNA concentration was measured by OD_{260} using the theoretical extinction coefficient provided by Dharmacon. Typical RNA stock concentrations ranged between 1 and 4 mM.

2.2. Protein production

Sam68 STAR (amino acids 97-283), KH-QUA2 (150-283) and KH (150-260) domains and T-STAR STAR (1-183), KH-QUA2 (50-183) and KH (50-160) domains (Fig. 1) were cloned by the University of Leicester Protein Expression Facility (PROTEX, http://www2.le.ac.uk/departments/biochemistry/research-groups/protex) using the pLeics03 expression vector that contains an N-terminal poly-histidine tag followed by a tobacco-etch virus protease (TEV) cleavage site. All plasmid constructs were verified by DNA sequencing (PNACL, Leicester). Recombinant plasmids were transformed into Rosetta BL21 DE3 cells and expressed in 4 litres of 2TY medium or M9 minimal medium supplemented with $^{15}NH_4Cl$. At an optical density of 0.5, cultures were transferred to an incubator at 20 °C for 1 h and protein expression was induced with 400 µM IPTG for 16 h at 20 °C.

The proteins of interest were purified by affinity chromatography using Ni-NTA agarose (Qiagen) followed by TEV cleavage during overnight dialysis in phosphate buffer (20 mM sodium phosphate pH7, 100 mM sodium chloride, 10 mM β-mercaptoethanol) at 4 °C. Because short ssRNA oligonucleotides are easily prone to degradation, 5 µl SUPERase IN RNase Inhibitor (Invitrogen) was added to the protein sample that was further purified by size-exclusion chromatography on a Superdex 75 10/300 (GE Healthcare) into the desired buffer for NMR analysis (see Section 3). Selected fractions were pooled and concentrated (Millipore 10 kDa centricon) to approximately 0.2 mM for NMR studies or to approximately 10 mg ml^{-1} for crystallisation screenings. Protein concentrations were estimated by measuring the OD_{280} and using a theoretical extinction coefficient (web.expasy.org/protparam/) derived from the protein sequence. RNAse activity was evaluated using Ambion RNAseAlert Lab Test kit according to manufacturer instructions. It is important to note that RNAse inhibitors should not be added to the final NMR sample because the storage buffer contains components with non-labile protons that interfere with the NMR measurements of the proteins and RNAs.

2.3. NMR measurements

NMR samples consisted of 330 µl of proteins at concentrations of at least 200 µM in different buffers and 20 µl of D_2O. NMR measurements were performed using Bruker AVIII-500 MHz, AVIII-600 MHz, AVIII-600 MHz (equipped with a cryoprobe) and Avance-800 MHz (equipped with a cryoprobe) spectrometers. Data were processed using XWINNMR (Bruker) and analyzed with Sparky (http://www.cgl.ucsf.edu/home/sparky/).

Optimisation of the buffer and temperature conditions as well as the protein constructs and RNA sequences were evaluated using 2D ^{15}N–^{1}H HSQC experiments for visualizing the ^{15}N-labeled protein signals, and 2D-TOCSY and 2D-NOESY experiments to visualise the RNA signals and the presence of intermolecular NOEs. TOCSY and NOESY experiments were recorded in D_2O with mixing times of 50 and 150 ms, respectively.

2.4. X-ray crystallography

Six different crystallisation screens (Proplex, NR-LBD, Morpheus, PACT, JCSG+ and Stura & Macrosol) have been used with different protein and RNA concentrations (10 and 20 mg mL^{-1}) using the Douglas Instrument Oryx 4 robot. For optimisation, T-STAR KH crystals were grown by sitting drop vapour diffusion at 4 °C in 200 mM ammonium sulfate, 100 mM HEPES pH 7.5 and 20% PEG 3350. Crystals were flash-frozen in mother liquor containing 15% MPD as a cryoprotectant. The KH–AAAUAA complex was crystallised in 2 M lithium sulfate and 100 mM Tris pH7.0 at 4 °C using the sitting drop vapour diffusion at a protein concentration of 15 mg mL^{-1} (protein:RNA molar ratio of 1:1.5). Data were collected on single crystals at the diamond synchrotron beamline I04 and microfocus beamline I24 and processed using XDS [29].

3. Results and discussion

3.1. Defining the optimal protein construct for NMR studies

NMR spectroscopy is limited by the size of the system under study. The upper molecular weight limit for structure determination is currently approximately 50 kDa, which means that it is often not possible to study a full-length protein by NMR. Fortunately, most RBPs are composed of small structurally independent domains that are sufficient for RNA binding. It is therefore possible to subclone RBPs into distinct domains whose size is suitable for NMR studies. If little is known about this RBP, potential functional domains can be identified using multiple sequence alignment algorithms, secondary structure prediction and identification of conserved domains. For well-characterised RBDs, specific constructs can easily be designed. Different protein constructs, although highly homologous, can however behave differently and various expression and purification strategies may have to be attempted in order to obtain a highly concentrated, pure and soluble sample suitable for NMR studies. If protein yield is too low following overexpression in a bacterial host, it may be that the construct is toxic to the cells or prone to aggregation. In this case, adjustment of the domain boundaries, use of an alternative affinity tag or use of a solubility tag may result in a more stable sample [30,31]. Other options for protein production such as baculovirus or mammalian cells are still not commonly used for NMR studies because isotope labelling is either not possible or not financially viable [32]. The ^{15}N–^{1}H HSQC is the most commonly used NMR experiment to investigate the suitability of a protein construct for further NMR studies and for investigating the complex formation between a protein and its partner, such as protein, RNA, DNA or small molecules. Such experiments requires a ^{15}N-labeled protein that can be obtained by expressing the protein in *Escherichia coli* grown in a minimum medium in which the sole nitrogen source is provided by the addition of $^{15}NH_4Cl$. The ^{15}N–^{1}H HSQC is a two-dimensional NMR experiment that allows a magnetisation transfer between a proton and its attached NMR visible ^{15}N isotope. This results in a spectrum in which each NH and NH2 groups give a crosspeak at the specific frequency of the proton in one dimension and the nitrogen in the other dimension. These frequencies are dependent on the atom chemical environments and therefore, in folded

proteins, different atoms have different frequencies. Since every amino acid (except proline and the N-terminal amino acid) contains an amide group in its backbone, this spectrum is often referred as the NMR fingerprint of the protein and can be used to optimise the protein construct and the buffer conditions of the sample. Indeed, the quality of the ^{15}N–^{1}H HSQC depends on the folding and stability of the protein, that in turn is dependent on various parameters such as protein concentration, type of buffer, salt concentration, pH and temperature.

Various types of buffer are suitable for NMR studies and some parameters must be taken into consideration when optimizing buffer conditions. Ideally, the buffer used should not be protonated. Indeed, the concentration of buffer is generally higher than that of the protein and, if protonated, the buffer NMR signals will interfere with the protein signals. The most commonly used NMR buffer is sodium phosphate at a concentration ranging between 10 and 50 mM. Alternatively, other buffers such as Tris–HCl or HEPES can be used at similar concentrations. As these buffers contain protons, it is preferable to purchase them in a deuterated form. NMR sensitivity is inversely correlated with the ionic strength of the buffer. Typically, the ionic strength should be minimal and not exceed 150 or 200 mM sodium or potassium chloride. If protein solubility requires high ionic strength, a high concentration of salt can efficiently be replaced by low-conductivity salts such as 50 mM L-Arginine and L-Glutamate [33]. Additionally, the pH of the buffer must be neutral or slightly acidic (typical range between 5 and 7). Amide protons exchange with the solvent and high pH increases this exchange rate leading to a loss of amide signals in NMR experiments such as ^{15}N–^{1}H HSQC. Below pH 5, the acidic condition might induce protein unfolding and aggregation. It should also be noted that the pH should be different from the isoelectric point (pI) of the protein or protein domain by at least 0.5 to avoid problems of solubility. Finally, compounds that are protonated or that increase the viscosity of the buffer, such as glycerol, must be avoided.

The success of the structure determination depends on the sample conditions, and thus a screen of different conditions can be designed at an early stage of the NMR study to improve the spectra quality. Typically, in our laboratory, as in other laboratories, we initially measure a ^{15}N–^{1}H HSQC in standard buffer conditions (20 mM sodium phosphate pH 7, 50 mM sodium chloride, 20 °C) and estimate the quality of the spectrum. A screen of conditions is then applied where the pH is varied from 5.5 to 7.5 and the salt concentration from 0 to 200 mM NaCl. For each condition, a ^{15}N–^{1}H HSQC is measured at 20, 25 and 30 °C and the quality of the spectrum is estimated based on the number and the line width of the visible crosspeaks (Figs. 2 and 3).

3.1.1. Optimisation of T-STAR constructs

In order to obtain a suitable protein sample for NMR studies, we tested different constructs of T-STAR that are expected to be sufficient for RNA binding: the full STAR domain (amino acids 1–183), and shorter constructs containing the KH (50–160) and the KH-QUA2 (50–183) domains. In all cases, the protein constructs expressed very well in *E. coli*, were soluble and could be purified using Ni-NTA agarose followed by TEV cleavage and gel filtration. These protein constructs remained soluble in various buffers suitable for NMR studies and could be concentrated to a final protein concentration above 200 μM.

We started the project by investigating the full STAR domain of T-STAR and preparing a sample of this protein in a common NMR buffer containing 20 mM sodium phosphate pH7, 50 mM sodium chloride and measuring a ^{15}N–^{1}H HSQC experiment at 20 °C. Despite the fact that the protein is soluble in this buffer condition, the ^{15}N–^{1}H HSQC spectrum was poorly defined (Fig. 2A). With this protein construct containing 183 amino acids including 8 prolines, one would expect 174 amide crosspeaks in the spectrum. The number of crosspeaks that could be observed in the amide region of the spectrum was only 107. The spectrum shows that most peaks are located in the center of the spectrum and have a high intensity, which is typical for flexible regions of proteins. This clearly indicates that in these conditions, although the protein construct is soluble, the quality of the spectrum is not suitable for structural analysis. Attempts to optimise the sample conditions by varying the pH, the temperature and the salt concentration did not improve the quality of the ^{15}N–^{1}H HSQC spectrum (data not shown). We then expressed and purified a truncated version of the STAR domain, the KH-QUA2, which lacks the QUA1 dimerisation domain but is expected to be sufficient for RNA binding. This construct was also soluble at suitable NMR concentrations in various buffer conditions and the quality of the ^{15}N–^{1}H HSQC spectrum improved dramatically (Fig. 2B). Crosspeaks are well-dispersed in the proton dimension indicating that the KH-QUA2 construct is correctly folded. 101 out of 126 crosspeaks were observed. The central region of the spectrum still contains many intense peaks suggesting that some parts of the protein construct are flexible. This is consistent with previous structural studies on the STAR protein Quaking, showing that the QUA2 region is flexible in solution [34]. We thus tested another shorter protein construct of T-STAR, the isolated KH domain. As for the other constructs, the KH domain of T-STAR expressed very well in *E. coli* and remained soluble at concentrations above 200 μM. The ^{15}N–^{1}H HSQC spectrum of the KH domain was of excellent quality. Condition optimisations for this domain were performed and it appeared that the NMR spectra of this domain remained suitable for NMR studies under various buffer conditions, pH, salt concentration and temperature ranges. From our initial screen, we defined the optimal conditions as 20 mM sodium phosphate pH 6.3, 50 mM NaCl, 30 °C. A final optimisation of these conditions was performed replacing the sodium phosphate by TRIS–HCl or HEPES. We observed that changing the buffer to 10 mM TRIS–HCl pH 7 improved the stability of the sample and the quality of the spectrum, although the buffering capacity of TRIS–HCl is not effective at this pH. These conditions were subsequently used for our NMR studies (Fig. 2C). In these conditions, the ^{15}N–^{1}H HSQC spectrum of T-STAR KH displayed 100 crosspeaks out of the 103 expected. Furthermore, an overlay of the ^{15}N–^{1}H HSQC spectra of T-STAR KH and KH-QUA2 shows that the fold of the KH domain is similar in both constructs (overlap of crosspeaks) and that the QUA2 region is flexible since most additional crosspeaks of KH–QUA2 are located in the center of the spectrum and more intense than the crosspeaks corresponding to the KH domain (Fig. 2D).

3.1.2. Optimisation of Sam68 constructs

Sam68 and T-STAR are highly homologous proteins. The main difference between these proteins is the presence of a 100 amino acid N-terminal region of Sam68 that is not present in T-STAR (Fig. 1A). Considering the STAR domain of both proteins, sequence alignment indicates that 69% of the amino acids are identical and 16% display a strong similarity (Fig. 1B). When considering the KH domain only, the identity increases to 77% with a strong similarity of 14%. We therefore anticipated that the KH domain of Sam68 would behave similarly to the KH domain of T-STAR in solution and initiated an NMR study of the Sam68 KH construct using the optimal conditions defined for T-STAR (10 mM Tris pH 6.5, 50 mM NaCl, 30 °C). Sam68 KH expressed well and was soluble in *E. coli*, although with a lower yield than T-STAR KH. The affinity chromatography purification procedure was the same as for T-STAR. Dialysing the protein in the T-STAR NMR buffer, however, resulted in a large amount of precipitation and we could not recover soluble forms of Sam68 KH. Changing the buffer from 10 mM TRIS–HCl to 20 mM sodium phosphate and increasing the

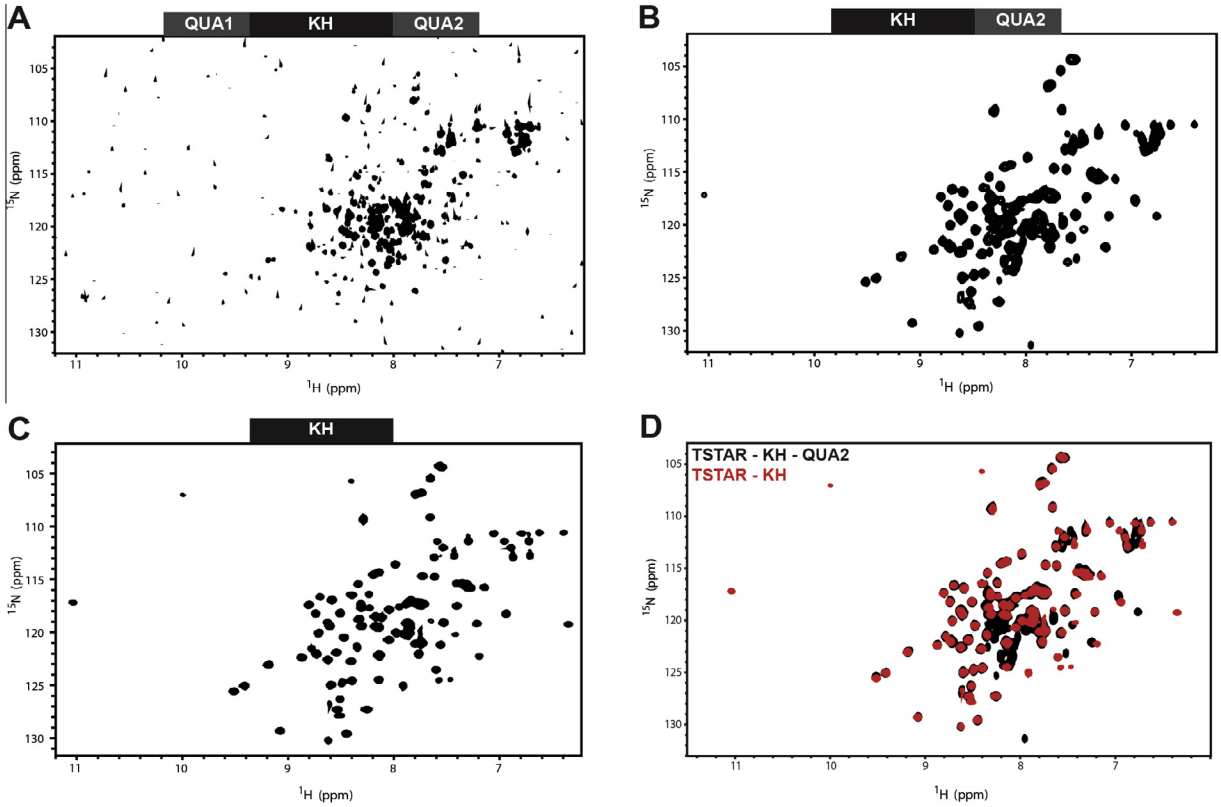


Fig. 2. Condition optimisation for NMR studies of T-STAR constructs. ^{15}N–^{1}H HSQC spectra of T-STAR constructs. (A) STAR domain in 20 mM sodium phosphate pH 6.5, 50 mM NaCl, 30 °C. (B) KH-QUA2 in 20 mM sodium phosphate pH 6.5, 50 mM NaCl, 30 °C. (C) KH in 10 mM Tris–HCl pH 6.5, 50 mM NaCl, 30 °C. (D) Overlay of the ^{15}N–^{1}H HSQC spectra of T-STAR KH-QUA2 (black) and KH (red).

salt concentration of the dialysis and gel filtration buffers to 100 mM NaCl allowed us to maintain the solubility of the protein. We could obtain a sample of ^{15}N labelled Sam68 KH at approximately 0.2 mM that was sufficient for measuring a ^{15}N–^{1}H HSQC experiment at 20 °C (Fig. 3A). All amide crosspeaks were very intense and located in the central region of the spectrum, indicating that in these conditions only the flexible regions of Sam68 KH were visible. We performed a screen of conditions as described above. In summary, increasing the pH to 7.0 led to the appearance of well-dispersed crosspeaks at 20 °C (Fig. 3B). Increasing the temperature to 25 °C resulted in sample precipitation (Fig. 3C). Finally, decreasing the NaCl concentration from 100 to 50 mM, improved the signal to noise ratio of the spectrum (Fig. 3D). These conditions could be used to investigate the binding of RNA to Sam68 KH domain (see Section 3.2.2.), although the protein could not be kept in its stable folded state for a long period of time. For this reason, we recently tested the expression and solubility of alternative constructs of Sam68. Initially, the KH–QUA2 domain was expressed and purified using the same protocol as for the KH domain. In the same sample conditions as Sam68 KH, we were unable to concentrate this construct adequately and the protein was unstable, even at 20 °C (Fig. 3E). The full STAR domain of Sam68 was then expressed and yielded larger amounts of protein than either the KH or KH–QUA2 constructs. We were able to concentrate this sample up to ~500 μM and it remained stable at 30 °C for a long period of time (several weeks at room temperature), making it highly suitable for NMR analysis. The ^{15}N–^{1}H HSQC spectrum shows that, in contrast to T-STAR STAR, the STAR domain of Sam68 is well folded and we observed 172 crosspeaks out of 174 expected (Fig. 3F). In addition, the spectrum of Sam68 STAR overlays well with the spectrum of isolated KH domain and of isolated QUA1 [21], suggesting that the QUA1 dimerisation domain and the KH domain of Sam68 are properly folded in our STAR construct.

3.2. Defining the optimal ssRNA sequence for NMR studies

As described in the previous section, the ^{15}N–^{1}H HSQC spectrum can be considered the fingerprint of the protein. Since the frequency of each nucleus depends on its chemical environment, NMR can be used to investigate the binding of partner molecules to a protein. The NMR chemical shift perturbation assay consists of adding increasing amounts of unlabeled partner to a ^{15}N labeled protein and measuring ^{15}N–^{1}H HSQC experiments for various partner-protein molar ratios [35,36]. If binding occurs, the amino acids at the interface with the partner will experience a different chemical environment and therefore their chemical shift will be different. This experiment provides precise information on the complex formation, such as an estimation of the dissociation constant, the stoichiometry of the complex and the amino acids involved in the interaction. In solution, the protein and the RNA are in equilibrium between their free and bound states and this equilibrium depends on the dissociation constant of the complex. During the NMR experiment, depending on the exchange rate of the complex formation, three different events can occur. In the slow exchange regime, the progressive addition of RNA leads to the presence of two crosspeaks for one perturbed N–H, one corresponding to the free and one to the bound form of the protein. The intensity of each crosspeak is directly proportional to the protein:RNA molar ratio. This exchange regime is reported for protein–RNA complexes with high affinities (dissociation constant below 200 nM). In the fast exchange regime, only one signal corresponding to an average of the free and bound state of the protein is visible. The addition of increasing amounts of RNA will gradually shift the signal from the free state towards the bound state of the protein. This exchange regime is generally reported for protein–RNA complexes with relatively low affinities (dissociation constant higher than 20 μM). Finally, in the intermediate exchange

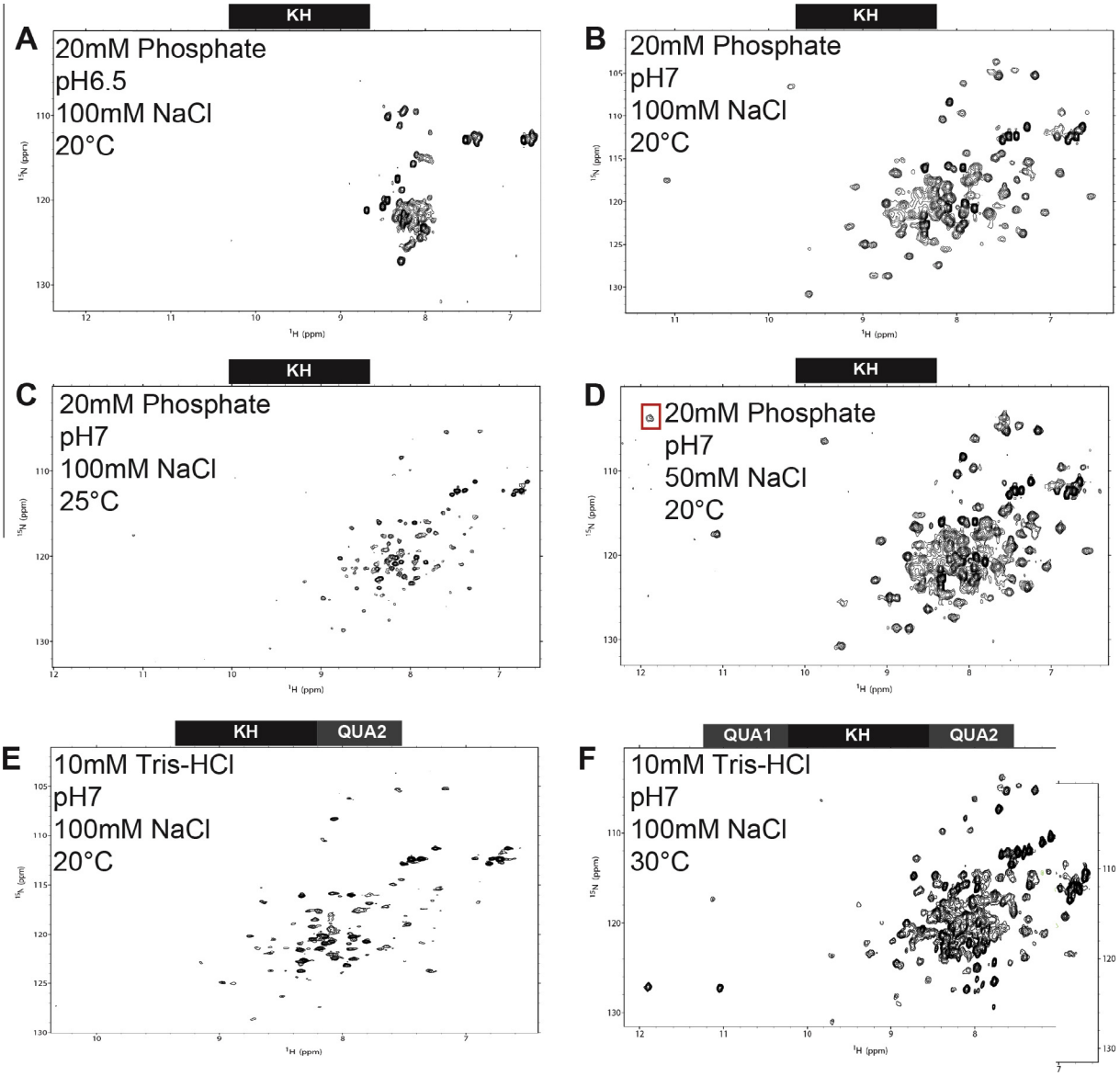


Fig. 3. Condition optimisation for NMR studies of Sam68 constructs. ^{15}N–^{1}H HSQC spectra of Sam68 constructs. (A) KH domain in 20 mM sodium phosphate pH 6.5, 100 mM NaCl, 20 °C. (B) KH in 20 mM sodium phosphate pH 7.0, 100 mM NaCl, 20 °C. (C) KH in 20 mM sodium phosphate pH 7, 100 mM NaCl, 25 °C. (D) KH in 20 mM sodium phosphate pH 7.0, 50 mM NaCl, 20 °C. (E) KH-QUA2 in 10 mM Tris–HCl pH 7.0, 100 mM NaCl, 20 °C. (F) STAR domain in 10 mM Tris–HCl pH 7.0, 100 mM NaCl, 30 °C.

regime, crosspeaks tend to disappear upon addition of RNA due to line broadening and reappear when the stoichiometry of the complex is reached. In many cases however, crosspeaks do not reappear, even in the presence of excess RNA. In that case, optimisation of the conditions (buffer, salt concentration, temperature) should be performed to obtain a suitable NMR spectrum of the protein–RNA complex. The intermediate exchange regime is reported for protein–RNA complexes with dissociation constants between 400 and 2 μM. NMR chemical shift perturbation experiment is very powerful and allows screening of different RNA sequences at an early stage of the structural work, permitting the identification of the optimal RNA sequence for structural investigation of protein–RNA complexes.

Chemical shift perturbation experiments performed using ^{15}N–^{1}H HSQC experiments, as detailed above, only provide information on the quality of the protein NMR signals. In addition, it is important to investigate the quality of the RNA signals, since the structure determination of the complex will rely on NMR derived restraints from both the protein and the RNA. As short single-stranded RNAs are mainly obtained by chemical synthesis, they can not be easily labeled isotopically. Observing solely the RNA resonances in the protein–RNA complex can therefore only be achieved by labeling proteins with ^{15}N and ^{13}C and using specific NMR experiments that cancel protein signals (reviewed in [37]). Since certain RNA chemical shifts are distinct from the protein ones, it is still possible to evaluate the quality of the RNA spectra using proton NMR experiments such as 2D DQF-COSY (Double Quantum Filtered Correlation Spectroscopy), 2D TOCSY (Total Correlation Spectroscopy) and 2D NOESY (Nuclear Overhauser Effect Spectroscopy) experiments (for more details, see [36]) without the need for producing ^{15}N/^{13}C-labeled protein samples. Homonuclear DQF-COSY and TOCSY are through-bond NMR experiments. Crosspeaks are observed between protons connected by two or three covalent bonds. For example, in RNA pyrimidines, the base contains two protons, H5 and H6, connected by three bonds through carbon atoms. Homonuclear NOESY is a through-space NMR experiment. Crosspeaks are observed between protons that are close in space (typically less than 5 Å). This experiment is

crucial in NMR structure determination for obtaining inter-proton (intra-protein, intra-RNA and intermolecular) distance restraints.

In order to design a pool of RNA targets for NMR screening, prior knowledge of the protein–RNA specificity is highly desirable. Several biochemical methods allow the identification of specific RNA sequences bound by RBPs, including footprinting, Systematic evolution of ligands by exponential enrichment (SELEX) or Cross-linking immunoprecipitation (CLIP) techniques. Footprinting experiments have been used for decades, using enzymes or chemicals that specifically cleave RNA molecules at certain positions, allowing the investigation of RNA structures and the identification of RNA sequences specifically bound by RBPs [38]. SELEX is an *in vitro* method consisting of a series of selection cycles of interacting RNA from a randomised oligonucleotide library. This generally allows for the identification of a consensus RNA sequence bound by the protein of interest [39]. CLIP experiments make use of the fact that UV irradiation of sample material, such as a cell lysate, causes covalent bond formation between RNA and proteins [40,41]. This technique allows the identification of natural RNA targets for the protein of interest and a consensus RNA sequence can be derived. Alternatively, when no specific sequence is known to bind an RBP, an NMR based method, called scaffold-independent analysis (SIA), has been developed using short synthetic randomised RNA sequences that are tested for binding to an RBP or RBD by NMR ^{15}N–^{1}H HSQC [42].

Each of these techniques provides useful preliminary information to define a pool of RNA sequences to screen for protein–ssRNA complex structure determination. It should be noted that consensus sequences derived from CLIP, SELEX or SIA are often degenerate and differ from natural sequences bound by RBDs. Nonetheless, the optimal RNA sequence for structure determination is not necessarily found naturally, nor has the highest affinity for the protein. This is due to the fact that a precise structure determination of a protein–RNA complex requires a single and stable conformation of the complex. For example, natural and/or high affinity RNA sequences often contain multiple, similar, and juxtaposed binding sites and are not suitable for structural work because the protein can bind these sequences in multiple registers leading to an inhomogeneity of the sample and a loss of NMR signal. It is therefore crucial to identify the optimal RNA sequence that has reasonably high affinity to obtain a stable complex, specificity to obtain a homogeneous complex, and is still similar to natural sequences to derive biologically relevant structural information.

Structure analysis of various KH domains in complex with DNA or RNA showed that the classical nucleotide binding pocket of KH domains accommodates 4 nucleotides and structures of these complexes were solved with DNAs or RNAs varying from 4 to 12 nucleotides in length [25]. In the case of Sam68, SELEX experiments defined three consensus RNA motifs with different binding affinities, UAAA having the highest affinity, followed by UUUA and AAAA [26,27]. Accordingly, these motifs have been identified in numerous pre-mRNAs bound by Sam68 [43–48]. However, other RNA sequences have been identified in other pre-mRNAs, such as AAAUU [49,50]. Interestingly, it has been recently reported that Sam68 bound a UAAUAAA motif present in the Neurexin pre-mRNA but not a truncated RNA containing only the UAAA motif [51]. Finally, Sam68 was also shown to bind poly(U) RNA sequences [52,53]. In the case of T-STAR, SELEX experiments identified A/U-rich sequences similar to the one bound by Sam68 [27]. Recently, a novel method, RNAcompete, defined the core binding site of T-STAR as UAA [54]. Similar AU-rich motifs have also been identified by CLIP experiments (S. Grellscheid, D. Elliot, personal communication). Finally, NMR-based SIA experiments with T-STAR KH suggested a preference for A-rich RNA sequences (K. Collin, A. Ramos, personal communication). The biological role of T-STAR is still unclear, and only one pre-mRNA target has been identified to date with a T-STAR binding site defined as 4 × (UUAA) [19]. Interestingly, Sam68 and T-STAR share 77% identity in their KH domains and correspondingly both proteins bind A/U rich RNA sequences. Yet, a comparison of the SELEX outputs suggest that these two proteins could specifically bind slightly different RNA sequences which could explain the fact that these proteins are not biologically equivalent (Fig. 4) [19]. Indeed, Sam68 seems to favour a UAAA motif surrounded preferentially by A (Fig. 4A), while T-STAR favours a UAA motif preferentially preceded by U and followed by A (Fig. 4B). Based on these consensus sequences, we have designed a series of 6mer A/U-rich RNAs (Table 1). For instance, sequences AAAUAA and AAUAAA resemble the Sam68 consensus sequence; UUUAAA resembles the T-STAR consensus sequences, while sequences like UAAAAA resemble both Sam68 and T-STAR consensus sequences. In addition, other sequences were derived based on pre-mRNA target sites such as AAAUUU and UAAAUU. Finally, we designed derivatives of these sequences, as well as 6mer polyA and polyU. Series of longer and shorter RNAs were also designed to reflect natural targets of Sam68 (UAAUAAAUU) or T-STAR (UAAUUAAA and AUUAAUUA) and to investigate whether the length of the optimal RNA sequence could improve the structural quality of the protein–RNA complex (Table 2).

3.2.1. Defining the optimal ssRNA sequence bound by T-STAR KH

As mentioned in Section 3.1.1, T-STAR KH and T-STAR KH–QUA2 constructs are highly soluble and stable, and the ^{15}N–^{1}H HSQC spectra of these domains were of excellent quality. In contrast, T-STAR STAR construct resulted in poor NMR spectra. As it has been shown for other STAR proteins that the KH–QUA2 region is sufficient for RNA binding [20] and that in the case of Sam68 (and by homology of T-STAR), the QUA2 region might not be involved in RNA binding [26,28], we tested the RNA binding ability of the constructs KH–QUA2 and KH. NMR chemical shift perturbation experiments were performed by measuring a ^{15}N–^{1}H HSQC experiment of a 0.2 mM sample of the free protein as reference. RNA was then gradually added to the protein sample at different molar ratios (protein:RNA ratio of 1:0.5 and 1:1). In all cases, the pH of the RNA stock solution was adjusted to correspond to the pH of the protein solution and RNAs were prepared at high concentration (up to 4 mM) to restrict the issue of RNA to be added to the protein and avoid a dilution of the protein that could affect the chemical shifts.

We initially tested the binding of T-STAR KH–QUA2 with some of our 6mer RNAs. With all tested RNA sequences, we observed changes of the protein ^{15}N–^{1}H HSQC spectrum, some peaks

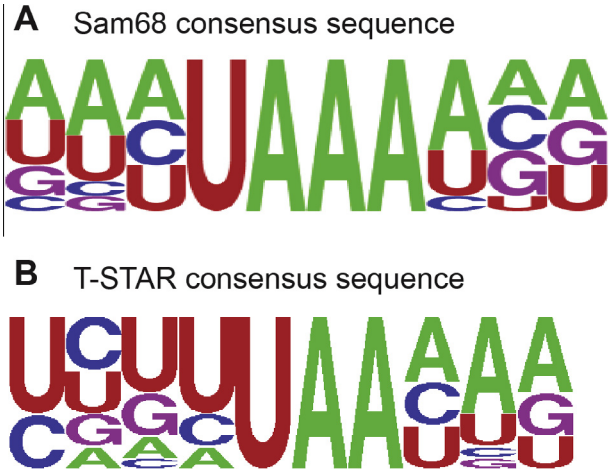


Fig. 4. Consensus RNA sequences derived from SELEX experiments. (A) Sam68 derived consensus RNA sequence. (B) T-STAR derived consensus RNA sequence. Figures were generated using WEBLOGO [57].

Table 1
List of 6mer RNAs used to study the T-STAR–RNA and Sam68–RNA complexes.

AAAUAA	AAUAAA	UUUAAA	UAAAAA	AAAUUU
UAAAUU	UAAAUA	UAAAAU	AAAUAU	AAUAUU
AUUAAA	AAUUUU	AUUUUU	AAAAAA	UUUUUU

Table 2
List of RNA with various lengths used to study the T-STAR–RNA and Sam68–RNA complexes.

UAAUAAAUU	UAAUUAAU	AUUAAUUA	UUUAAAUAA	AAAAAAUAA
UAAAUAAUU	UAAAAAUUUU	UAAAAUUUUU	UAAAUUUUUU	UAAAUAUUUU
AAAU	AAUA	AUAA	AAAUA	AAUAA

disappearing and others changing position, clearly indicating that these RNA sequences are able to bind T-STAR KH–QUA2 (Fig. 5A). A careful analysis of the chemical shift perturbations showed that all the peaks affected by the RNA addition corresponded to amino acids of the KH domain, while peaks of the QUA2 in the central region of the spectrum were not affected. This suggested that the KH domain of T-STAR could be sufficient for RNA binding. We therefore performed the same experiments with the T-STAR KH construct and indeed observed that the KH domain is sufficient for RNA binding and the chemical shift perturbation observed on the KH construct were similar to those observed on the KH–QUA2 construct. Further screening of RNA sequences was therefore performed on the KH construct of T-STAR.

All the RNA sequences tested showed a clear binding to the KH domain. Typical examples are displayed in Fig. 5B–D. In all cases, the same protein crosspeaks were affected, indicating that, whatever the A/U-rich RNA sequence, the same amino acids are involved in binding. However, the effect of RNA addition on the crosspeaks varied significantly with different RNA sequences (Fig. 5 B-D). For example, the AAAUAA RNA shows a clear fast exchange regime, with crosspeaks gradually shifting from their free to their bound position as a function of the protein: RNA molar ratio. This allows us to follow all the chemical shift perturbation and obtain a complete spectrum of the bound form of the protein (Fig. 5B). Other RNA sequences such as AAAUUU induce chemical shift perturbation in the protein crosspeaks but the intensity of the shift is weaker indicating that these RNAs have a lower affinity for the protein than AAAUAA (Fig. 5C). Other RNAs such as polyA induce perturbations similar to AAAUAA, but some peaks disappeared indicating a fast to intermediate exchange regime (Fig. 5D). Unfortunately, while this implies that these RNA sequences have a higher affinity for the protein, the peaks that disappear do not reappear in the spectrum even in excess of RNA, which is not optimum for acquisition of sufficient data for the structure determination of the protein–RNA complex. Taken together, the analysis of 6-mer RNA sequences showed that they all bound the T-STAR KH protein construct, but with different affinities, leading to different intermediate or fast exchange regimes. Our study showed that the RNA sequence AAAUAA was the optimal one because it induced the largest chemical shift perturbations of the protein crosspeaks and all crosspeaks were visible in the bound state.

We then investigated whether the length of the RNA sequence could influence the quality of the NMR spectra. Various derivatives of the AAAUAA sequence were synthesised (Table 2). This included shorter RNA sequences (5mers and 4mers) as well as longer sequences with extension in 5′, 3′ or both. Shorter versions of the RNA sequence were still sufficient for binding the protein but the chemical shift perturbations were smaller than with the 6mer sequence suggesting a lower affinity (data not shown). We then tested longer RNA sequences (9mers) with polyA or polyU extensions in the 5′ or 3′ of the AAAUAA central part. With these RNAs, the chemical shift perturbations have the same effect as the 6mer sequence on the KH domain. They affect the same area of the spectrum but instead of a clear chemical shift perturbation, crosspeaks disappeared and reappeared indicating an intermediate exchange regime and meaning a higher affinity of the protein for these RNAs compared to AAAUAA. However, not all crosspeaks of the protein reappeared when fully bound and these longer RNAs were therefore not suitable for structure determination (Fig. 6). In conclusion, we optimised both the composition and the length of the RNA sequences bound to T-STAR KH and concluded that the optimal sequence for structure determination was AAAUAA. Interestingly, this sequence could not be derived from the T-STAR specific RNA consensus sequence, but resemble the Sam68 consensus. Nevertheless, this sequence still contains the UAA core consensus sequence for T-STAR.

TOCSY and NOESY NMR experiments have also been used to investigate the NMR signal quality of the different RNA sequences in complex with T-STAR KH. As most RNA base protons are non-labile and have chemical shift values overlapping with the amide protein protons and with water, these experiments were recorded in 100% D_2O (see Section 2). In these conditions, the amide protons of the protein exchange with deuterium and the RNA crosspeaks can easily be analyzed. TOCSY spectra were used to identify the crosspeaks of the uridine H5–H6 bases. As expected, the TOCSY spectra of the AAAUAA and AAAUUU RNAs in complex with T-STAR KH displayed one and three crosspeaks, respectively, indicating that the uridine bases experience a single chemical environment when bound to the protein (Fig. 7A). In contrast, the TOCSY spectra of the longer sequences UAAAUAAUU and UAAAAUUUUU displayed two and one intense crosspeaks, instead of the four and six expected (Fig. 7A). This indicates that chemical exchange of these protons occur during binding and could be due to the RNA binding the protein in different registers. NOESY spectra provide useful information on the quality of the complex for NMR studies. When measured in 100% D_2O, the resonances in the 8 ppm frequency region correspond mainly to the RNA bases (in our case, adenine H8 and H2 and uridine H6). Crosspeaks from this region of the spectrum to the RNA sugar region (3–6.5 ppm) arise from RNA base protons in close proximity to RNA sugar protons (intra-RNA NOES) while crosspeaks to other regions of the spectrum (0–3 ppm) arise from RNA base protons in close proximity to protein protons (intermolecular NOES). As shown in Fig. 7B, the NOESY spectrum of AAAUAA in complex with T-STAR KH displays many NOE crosspeaks in the intra-RNA region, suggesting that the RNA adopts a well-defined conformation and is not disordered. Many NOES can also be observed in the intermolecular region, suggesting that the protein–RNA complex adopts a well-defined orientation and that intermolecular distances can be extracted, which are crucial for the structure determination of a protein–RNA complex by NMR. In contrast, the NOESY spectra of the other RNAs in complex with T-STAR KH displayed no or few intra-RNA and intermolecular NOES indicating that these RNA sequences are not suitable for structure determination of the protein–RNA complex. These NMR experiments confirmed our previous conclusion that the RNA sequence AAAUAA is the optimal sequence for the NMR structure determination of the T-STAR KH-RNA complex.

3.2.2. Defining the optimal ssRNA sequence bound by Sam68 KH

Given the high sequence homology between Sam68 and T-STAR KH, we performed chemical shift perturbation experiments on the KH domain of Sam68 with various AU-rich 6mer RNAs. The quality of the spectra was not as good as that of T-STAR KH, and the sample

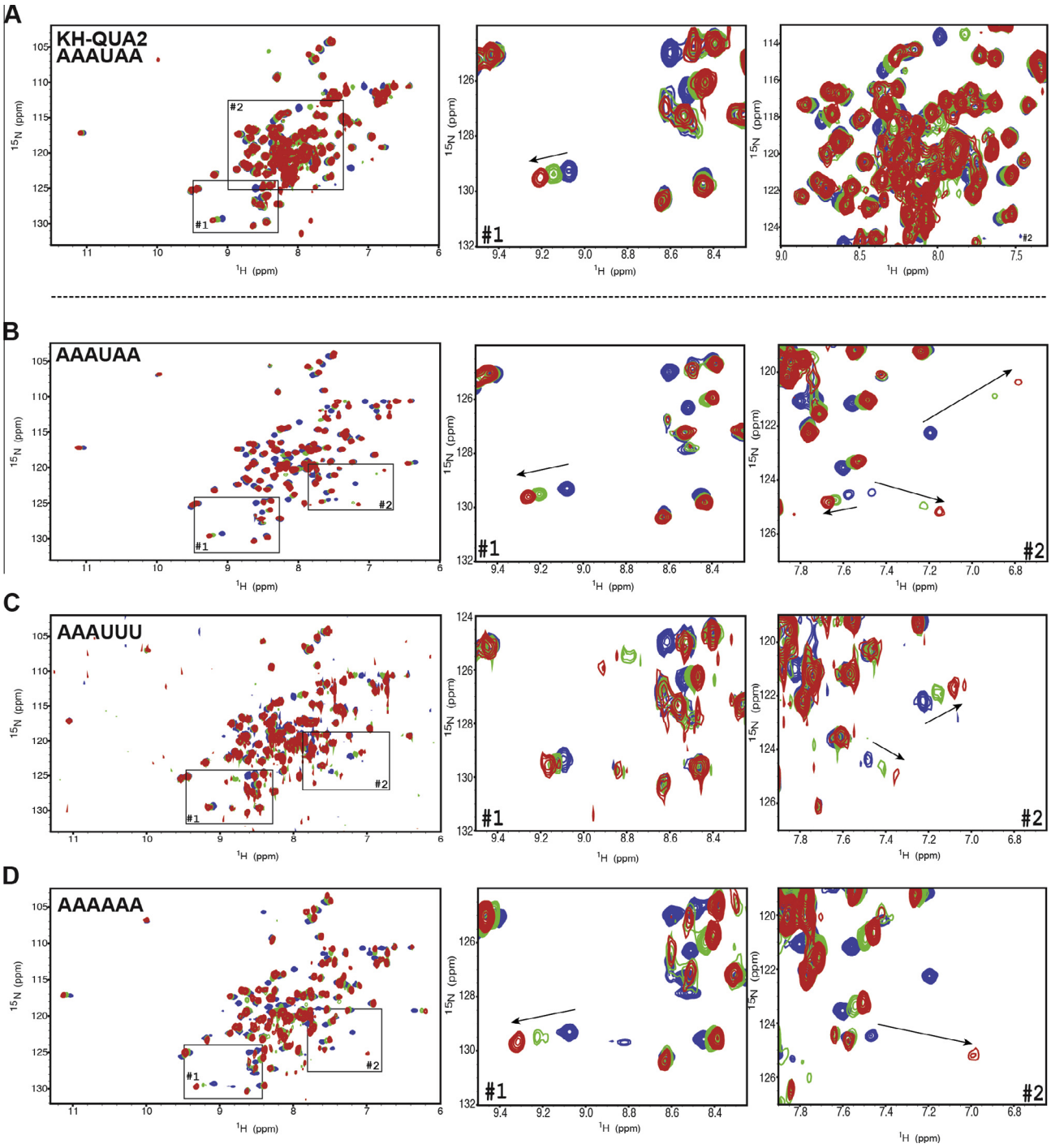


Fig. 5. Effect of RNA sequences on T-STAR–RNA complex formation. Chemical shift perturbation experiments of (A) T-STAR KH–QUA2 with AAAUAA, and T-STAR KH with (B) AAAUAA, (C) AAAUUU, and (D) AAAAAA. In all cases, an overlay of ^{15}N–^{1}H HSQC spectra is displayed for the free protein (blue), a protein:RNA molar ratio of 1:0.5 (green) and a protein:RNA molar ratio of 1:1 (red).

was not as stable (see Section 3.1.2). It was however sufficient to identify changes in the protein spectrum upon addition of increasing amounts of RNA. This suggested that, as for T-STAR, the KH domain of Sam68 is sufficient for RNA binding. We tested different 6mer RNAs designed according to SELEX and published biological data (Table 1). Interestingly, while Sam68 has previously been shown to bind poly(U) RNAs [52,53], the addition of UUUUUU RNA to Sam68 KH did not affect the NMR spectrum, indicating that, in our conditions, Sam68 KH does not bind poly(U) (Fig. 8A). All other tested RNAs affected the ^{15}N–^{1}H HSQC spectrum of Sam68 KH, indicating complex formation. Furthermore, the same peaks of Sam68 KH were affected by the addition of RNA, suggesting that the same residues are involved in binding. Different RNA sequences led to a combination of intermediate and fast exchange regimes, with many peaks disappearing and others shifting upon RNA addition. Surprisingly, crosspeaks in fast exchange shifted in different directions depending on the RNA sequence used (Fig. 8B–D), indicating that the chemical environment of these amino acids is different when bound to different RNA sequences. This suggests that although the same amino acids are affected by the

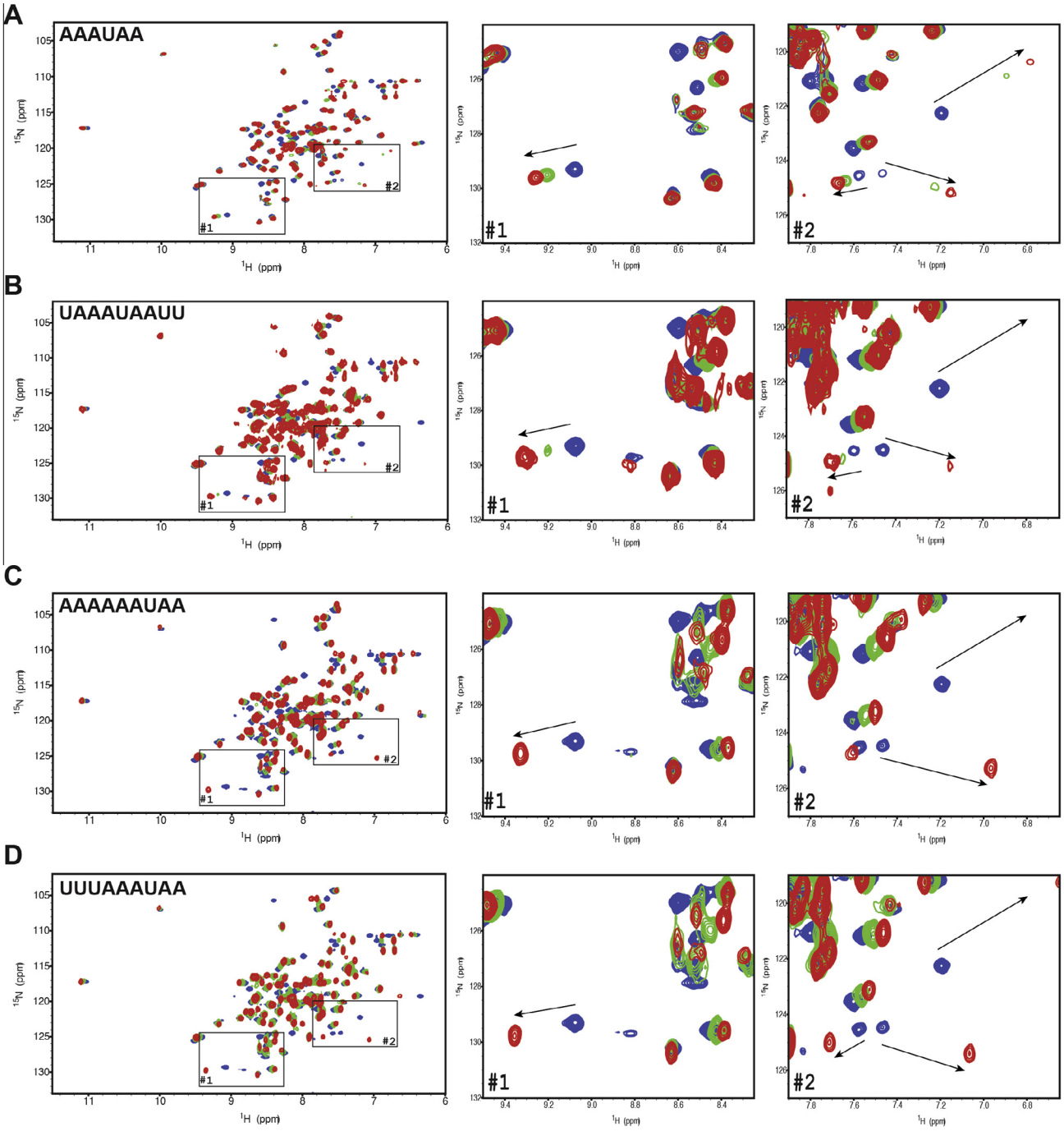


Fig. 6. Effect of RNA size on T-STAR–RNA complex formation. Chemical shift perturbation experiments of (A) T-STAR KH with AAAUAA (similar to Fig. 3B), (B) AAAAAAUAA, and (C) UUUAAAUAA. In all case, an overlay of ^{15}N–^{1}H HSQC spectra is displayed for the free protein (blue), a protein:RNA molar ratio of 1:0.5 (green) and a protein:RNA molar ratio of 1:1 (red).

various RNAs, the KH domain binds these RNAs in a slightly different way. An interesting RNA sequence is AUUAAA. The chemical shift perturbation experiment with this RNA was in the slow exchange regime indicating a strong affinity for the protein (Fig. 8E). In this case, most crosspeaks corresponding to the bound form of the protein were visible, making it a suitable candidate for further structural studies. However, since the sample was not stable, the quality of the ^{15}N–^{1}H HSQC spectrum remained poor and we could not measure additional NMR experiments such as NOESY.

Recently, we have produced samples of the STAR domain that are stable (Fig. 3F). These new samples are suitable for NMR structural studies and we will therefore investigate the binding of the different A/U-rich RNA sequences to the STAR domain of Sam68.

3.3. Using NMR data to optimise crystallisation conditions

X-ray crystallography is the primary method to determine the molecular structure of various biological molecules. This requires the molecules to aggregate in a well-ordered crystal. The principal factor for crystallisation is the buffer composition that, as for NMR, must be optimised. Because our NMR analysis showed that the KH

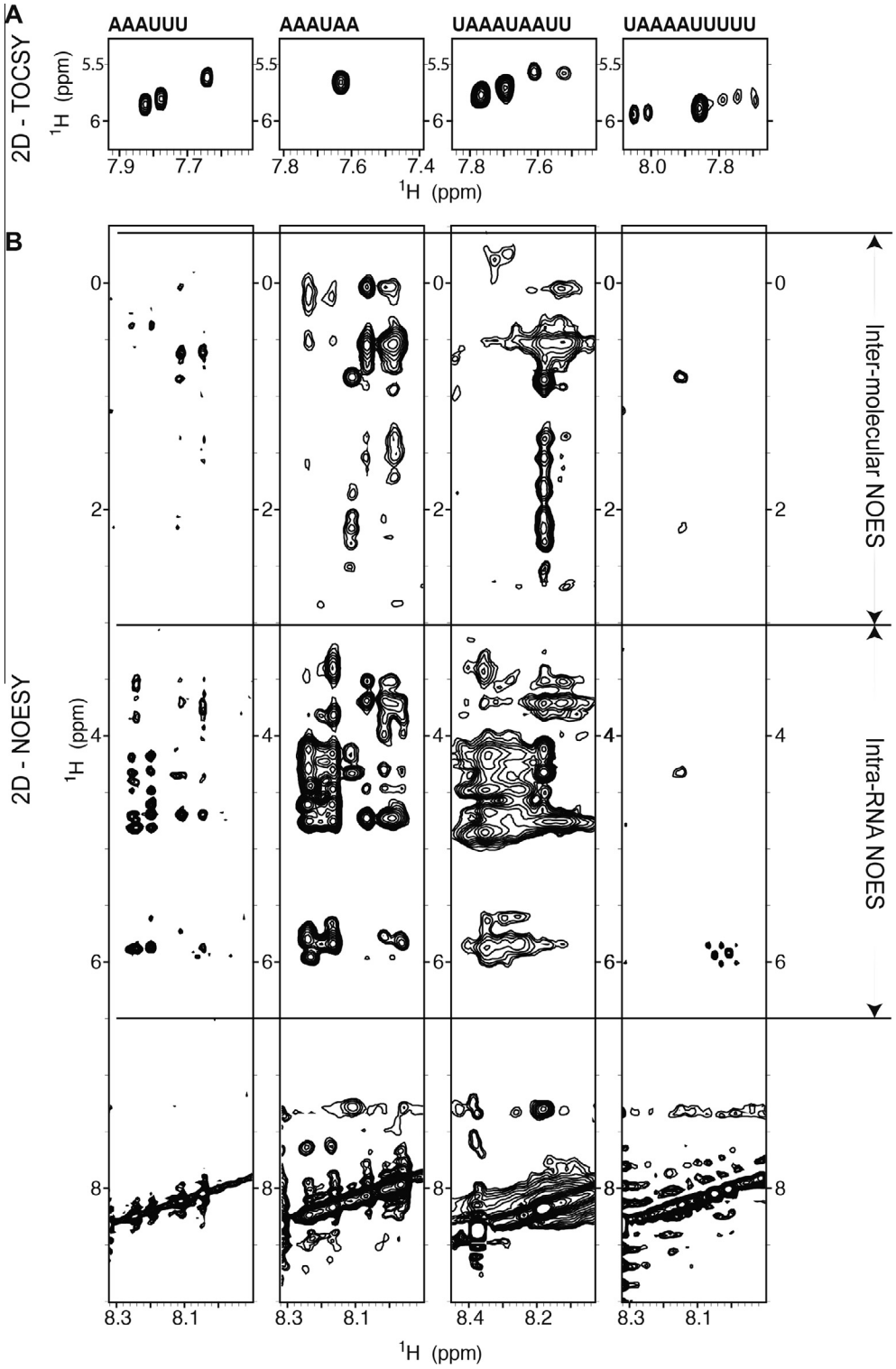


Fig. 7. Analyzing the NMR resonances of various RNAs in complex with T-STAR KH. (A) TOCSY and (B) NOESY spectra of different RNA sequences in complex with T-STAR KH measured in D_2O. The displayed section of the TOCSY spectra shows to the H5/H6 region of RNA pyrimidines and the section of the NOESY spectra shows the NOES between the RNA bases and either the RNA sugars (intra-RNA NOES) or the protein (intermolecular NOES).

domain of T-STAR was highly soluble and structured in solution (see Section 3.1.1), we set up crystallisation screens for this domain using six commercially available screens and protein concentrations ranging from 10 to 20 mg ml^{-1} in our optimised NMR buffer. We obtained various hits and optimised the conditions in order to obtain protein crystals of sufficient size. Our optimised crystals were rectangular and diffracted to a resolution of 1.6 Å (Fig. 9A). Interestingly, we observed that, in contrast to T-STAR KH, Sam68 KH does not behave well in solution (Section 3.1.2). Accordingly, crystallisation trials of Sam68 KH did not produce any crystal hits suggesting that NMR preliminary experiments on the solubility and stability of proteins (Sections 3.1.1 and 3.1.2) can provide useful information for crystallisation trials of proteins.

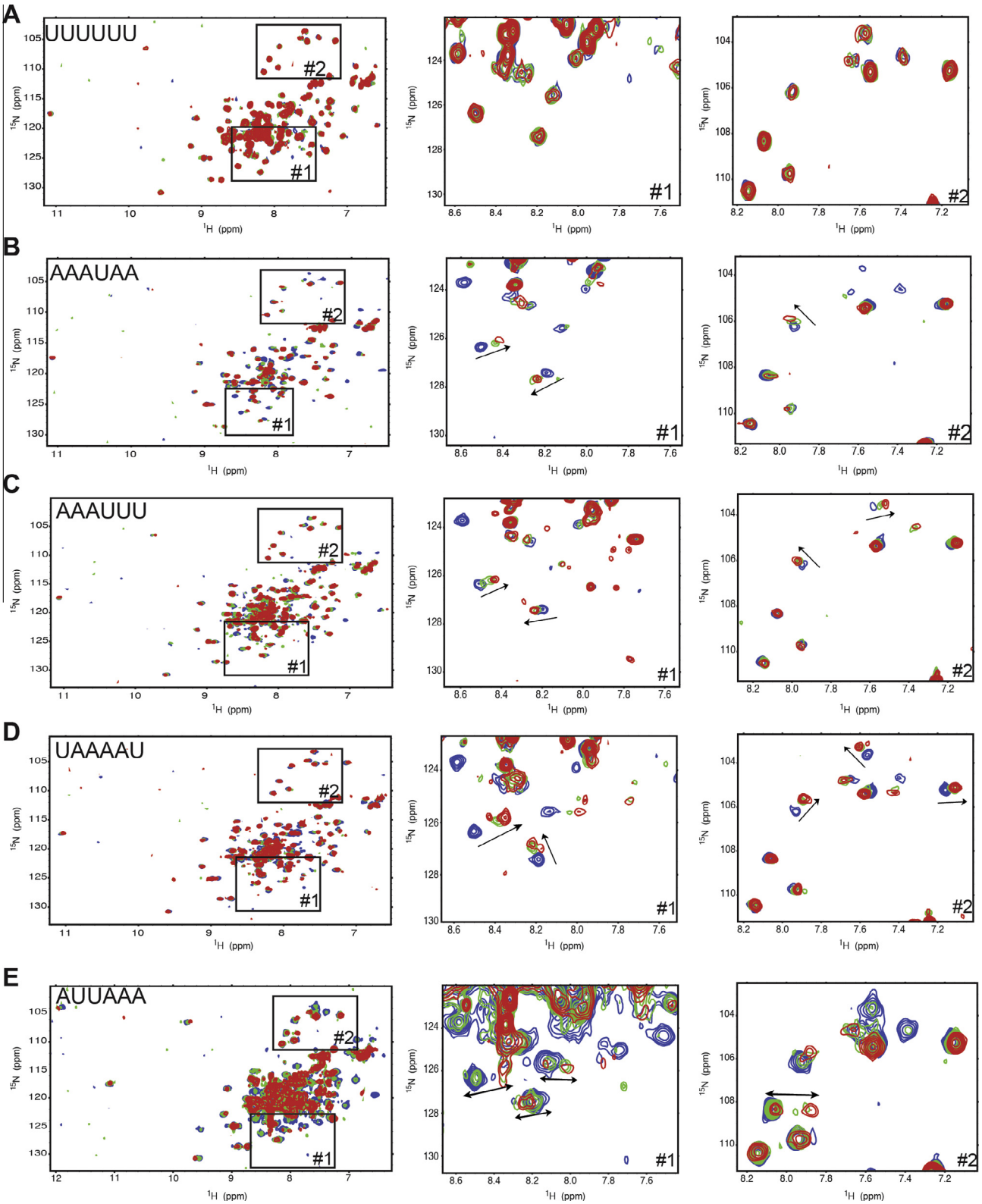


Fig. 8. Effect of RNA sequences on Sam68-RNA complex formation. Chemical shift perturbation experiments of Sam68 KH domain with (A) UUUUUU, (B) AAAUAA, (C) AAAUUU, (D) UAAAAU and (E) AUUAAA. In all case, an overlay of ^{15}N–^{1}H HSQC spectra is displayed of the free protein (blue), a protein:RNA molar ratio of 1:0.5 (green) and a protein:RNA molar ratio of 1:1 (red).

Using NMR chemical shift perturbation experiments, we have tested a large number of RNA sequences for binding T-STAR KH (Section 3.2.1) and concluded that the AAAUAA RNA sequence was the most suitable candidate for the structure determination of T-STAR KH in complex with RNA (Section 3.2.1). We therefore initiated a crystallisation trial of T-STAR with various 6-mer RNA sequences. Interestingly, only the complex of T-STAR with the AAAUAA RNA crystallised. In this case, crystals were hexagonal

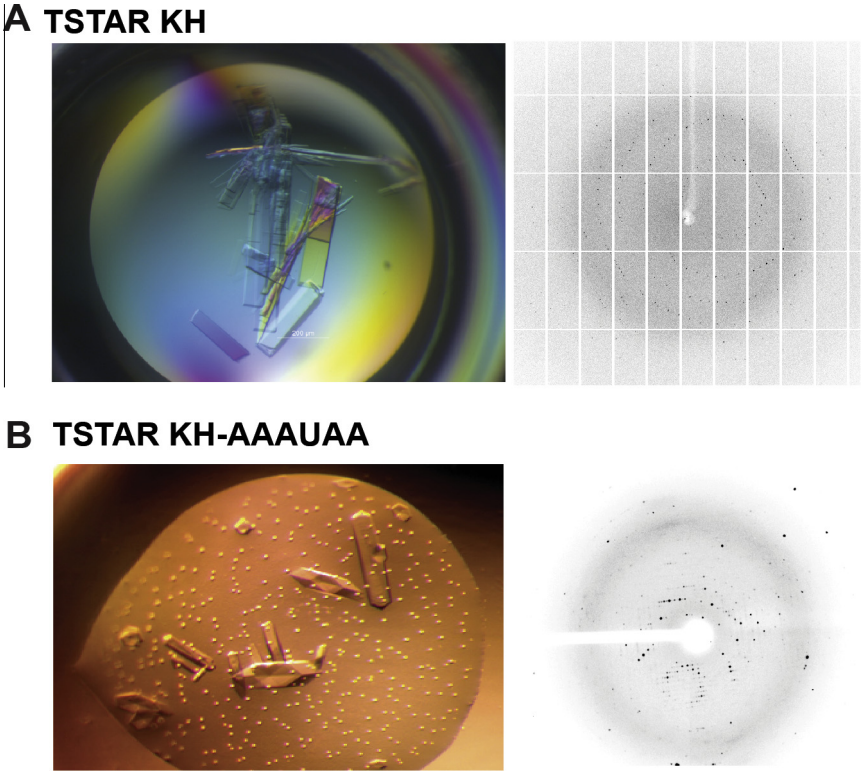


Fig. 9. Using NMR screening for X-ray crystallography. Crystals and diffraction pattern of (A) free T-STAR KH and (B) T-STAR KH–AAAUAA complex.

and diffracted to a resolution of 2.0 Å (Fig. 9B). Interestingly, the crystallogenesis condition and the space group are different than from the free KH suggesting that these crystals contain both protein and RNA. Furthermore, these data suggest that NMR chemical shift perturbation experiments of protein–RNA complexes can be used as a screening method to optimise the crystallisation procedure of such complexes.

4. Concluding remarks

Over the past few years, there has been an increasing interest in RNA biology and RNA binding proteins. Structural studies of protein–RNA complexes are therefore needed if we want to understand how proteins recognise specifically their RNA targets and to derive a general code for RNA recognition [55,56]. The intrinsic properties of such complexes, however, make them difficult to study structurally. In this article, we have shown how NMR can be used at an early stage of structural studies to first identify which protein constructs are suitable and, second to screen many RNA sequences in order to identify the optimal protein–RNA complex for structure determination.

Acknowledgments

The authors are grateful to S. Grellscheid and D. Elliot for providing T-STAR CLIP results, K. Collin and A. Ramos for SIA experiments, X. Yang (PROTEX) for the cloning facility, F. Muskett for NMR support, K. Sidhu for IT support and useful discussion, P. Watson, L. Fairall, P. Moody, J. Schwabe and the staff at beamlines I04 and I24 at the Diamond Light Source for assistance with X-ray crystallisation and data collection, and C. Weldon, O. Gonchar, I. Eperon and S. Jayne for useful discussion.

This work was supported by a Medical Research Council Career Development Award to C.D. (G1000526) and by a College of Medicine, Biological Sciences and Psychology, University of Leicester, studentship to J.F.

References

[1] S. Djebali, C.A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, et al., Nature 488 (2012) 101–108.
[2] L.R. Sabin, M.J. Delás, G.J. Hannon, Mol. Cell 49 (2013) 783–794.
[3] J.E. Wilusz, H. Sunwoo, D.L. Spector, Genes Dev. 23 (2009) 1494–1504.
[4] A.G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, et al., Mol. Cell 46 (2012) 674–690.
[5] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B.M. Beckmann, C. Strein, et al., Cell 149 (2012) 1393–1406.
[6] T.A. Cooper, L. Wan, G. Dreyfuss, Cell 136 (2009) 777–793.
[7] A. Castello, B. Fischer, M.W. Hentze, T. Preiss, Trends Genet. (2013) 1–10.
[8] K.E. Lukong, K.-W. Chang, E.W. Khandjian, S. Richard, Trends Genet. 24 (2008) 416–425.
[9] G.M. Daubner, A. Cléry, F.H.T. Allain, Curr. Opin. Struct. Biol. 23 (2013) 100–108.
[10] C.D. Cukier, A. Ramos, Eur. Biophys. J. 40 (2011) 1317–1325.
[11] C.D. Mackereth, M. Sattler, Curr. Opin. Struct. Biol. 22 (2012) 287–296.
[12] J. Konig, K. Zarnack, N.M. Luscombe, J. Ule, Nat. Rev. Genet. 13 (2012) 77–83.
[13] C. Vernet, Trends Genet. 13 (1997) 479–484.
[14] K. Artzt, J.I. Wu, Adv. Exp. Med. Biol. 693 (2010) 1–24.
[15] P. Bielli, R. Busa, M.P. Paronetto, C. Sette, Endocr. Relat. Cancer 18 (2011) R91–R102.
[16] J.J. He, J. Henao-Mejia, Y. Liu, RNA Biol. 6 (2009) 384–386.
[17] J.P. Venables, C. Vernet, S.L. Chew, D.J. Elliot, R.B. Cowmeadow, J. Wu, et al., Hum. Mol. Genet. 8 (1999) 959–969.
[18] M. Di Fruscio, T. Chen, S. Richard, Proc. Natl. Acad. Sci. USA 96 (1999) 2710–2715.
[19] I. Ehrmann, C. Dalgliesh, Y. Liu, M. Danilenko, M. Crosier, L. Overman, et al., PLoS Genet. 9 (2013) e1003474.
[20] Z. Liu, I. Luyten, M.J. Bottomley, A.C. Messias, S. Houngninou-Molango, R. Sprangers, et al., Science 294 (2001) 1098–1102.
[21] N.H. Meyer, K. Tripsianes, M. Vincendeau, T. Madl, F. Kateb, R. Brack-Werner, et al., J. Biol. Chem. 285 (2010) 28893–28901.
[22] M. Teplova, M. Hafner, D. Teplov, K. Essig, T. Tuschl, D.J. Patel, Genes Dev. 27 (2013) 928–940.
[23] C. Beuck, B.R. Szymczyna, D.E. Kerkow, A.B. Carmel, L. Columbus, R.L. Stanfield, et al., Structure 18 (2010) 377–389.
[24] C. Beuck, S. Qu, W.S. Fagg, M.A. Jr, J.R. Williamson, J. Mol. Biol. 423 (2012) 766–781.
[25] R. Valverde, L. Edwards, L. Regan, FEBS J. 275 (2008) 2712–2726.

[26] Q. Lin, S.J. Taylor, D. Shalloway, J. Biol. Chem. 272 (1997) 27274–27280.
[27] A. Galarneau, S. Richard, BMC Mol. Biol. 10 (2009) 47.
[28] S.P. Ryder, F. Massi, Adv. Exp. Med. Biol. 693 (2010) 37–53.
[29] W. Kabsch, Acta Crystallogr. D66 (2010) 125–132.
[30] D. Esposito, D.K. Chatterjee, Curr. Opin. Biotechnol. 17 (2006) 353–358.
[31] D.S. Waugh, Trends Biotechnol. 23 (2005) 316–320.
[32] H. Takahashi, I. Shimada, J. Biomol. NMR 46 (2009) 3–10.
[33] G.M. Hautbergue, A.P. Golovanov, J. Magn. Reson. 191 (2008) 335–339.
[34] M.L. Maguire, G. Guler-Gane, D. Nietlispach, A.R.C. Raine, A.M. Zorn, N. Standart, et al., J. Mol. Biol. 348 (2005) 265–279.
[35] E.R.P. Zuiderweg, Biochemistry 41 (2002) 1–7.
[36] C. Dominguez, M. Schubert, O. Duss, S. Ravindranathan, F.H.T. Allain, Prog. Nucl. Magn. Reson. Spectrosc. 58 (2011) 1–61.
[37] A.L. Breeze, Prog. Nucl. Magn. Reson. Spectrosc. 36 (2000) 323–372.
[38] D. Fourmy, S. Yoshizawa, WIREs RNA. 3 (2012) 557–566.
[39] G. Aquino-Jarquin, J.D. Toscano-Garibay, Int. J. Mol. Sci. 12 (2011) 9155–9171.
[40] J. Ule, K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, R.B. Darnell, Science 302 (2003) 1212–1215.
[41] J. Ule, K. Jensen, A. Mele, R.B. Darnell, Methods 37 (2005) 376–386.
[42] B. Beuth, M.F. García-Mayoral, I.A. Taylor, A. Ramos, J. Am. Chem. Soc. 129 (2007) 10205–10210.
[43] G.A. Tremblay, S. Richard, RNA Biol. 3 (2006) 90–93.
[44] M.P. Paronetto, T. Achsel, A. Massiello, C.E. Chalfant, C. Sette, J. Cell Biol. 176 (2007) 929–939.
[45] G. Chawla, C.-H. Lin, A. Han, L. Shiue, M. Ares, D.L. Black, Mol. Cell. Biol. 29 (2009) 201–213.
[46] M.P. Paronetto, M. Cappellari, R. Busà, S. Pedrotti, R. Vitali, C. Comstock, et al., Cancer Res. 70 (2010) 229–239.
[47] S. Pedrotti, P. Bielli, M.P. Paronetto, F. Ciccosanti, G.M. Fimia, S. Stamm, et al., EMBO J. 29 (2010) 1235–1247.
[48] M.-É. Huot, G. Vogel, A. Zabarauskas, C.T.-A. Ngo, J. Coulombe-Huntington, J. Majewski, et al., Mol. Cell 46 (2012) 187–199.
[49] N. Matter, P. Herrlich, H. König, Nature 420 (2002) 691–695.
[50] C. Valacca, S. Bonomi, E. Buratti, S. Pedrotti, F.E. Baralle, C. Sette, et al., J. Cell Biol. 191 (2010) 87–99.
[51] T. Iijima, K. Wu, H. Witte, Y. Hanno-Iijima, T. Glatter, S. Richard, et al., Cell 147 (2011) 1601–1614.
[52] S.J. Taylor, D. Shalloway, Nature 368 (1994) 867–871.
[53] M. Itoh, I. Haga, Q.-H. Li, J.-I. Fujisawa, Nucleic Acids Res. 30 (2002) 5452–5464.
[54] D. Ray, H. Kazan, E.T. Chan, L.P. Castillo, S. Chaudhry, S. Talukder, et al., Nat. Biotechnol. 27 (2009) 667–670.
[55] S.D. Auweter, F.C. Oberstrass, F.H.T. Allain, Nucleic Acids Res. 34 (2006) 4943–4959.
[56] A. Serganov, D.J. Patel, Curr. Opin. Struct. Biol. 18 (2008) 120–129.
[57] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, Genome Res. 14 (2004) 1188–1190.

Biochemical Society TRANSACTIONS
www.biochemsoctrans.org
Biochemical Society Transactions

Structural investigations of the RNA-binding properties of STAR proteins

Mikael Feracci*, Jaelle Foot* and Cyril Dominguez*[1]
*Department of Biochemistry, Henry Wellcome Laboratories of Structural Biology, University of Leicester, Lancaster Road, Leicester LE1 9HN, U.K.

Abstract
STAR (signal transduction and activation of RNA) proteins are a family of RNA-binding proteins that regulate post-transcriptional gene regulation events at various levels, such as pre-mRNA alternative splicing, RNA export, translation and stability. Most of these proteins are regulated by signalling pathways through post-translational modifications, such as phosphorylation and arginine methylation. These proteins share a highly conserved RNA-binding domain, denoted STAR domain. Structural investigations of this STAR domain in complex with RNA have highlighted how a subset of STAR proteins specifically recognizes its RNA targets. The present review focuses on the structural basis of RNA recognition by this family of proteins.

The STAR (signal transduction and activation of RNA) family of proteins

Throughout the 1990s, a novel family of proteins that functions in RNA metabolism and is regulated through signal transduction pathways, was discovered and named the STAR family [1]. First, Sam68 (Src-associated in mitosis of 68 kDa) was initially identified as a 62 kDa protein binding to Ras-GTP and was later found to be a downstream target of Src and Fyn [2,3]. Sam68 is now the best-characterized member of the STAR family, and two mammalian orthologues have also been identified, SLM-1 and SLM-2/T-STAR (Sam68-like mammalian proteins 1 and 2/testis STAR) [4,5]. Shortly after the discovery of Sam68, a novel subfamily of evolutionarily conserved proteins that share sequence homology with Sam68 was described. These proteins are *Caenorhabditis elegans* GLD-1 (defective in germline development protein 1) [6], mouse QKI (quaking) [7] and *Drosophila* HOW (held out of wings) [8,9] and play roles in post-transcriptional gene regulation and developmental processes including muscle development, oogenesis and spermatogenesis. Finally, another member, SF1 (splicing factor 1) was identified in 1996 [10].

Domain organization of STAR proteins

Together, the STAR proteins are related by an evolutionarily conserved domain of approximately 200 amino acids, which is responsible for RNA binding. This domain, denoted STAR domain, can be divided into three distinct subdomains: the central well-known hnRNP (heterogeneous nuclear ribonucleoprotein) KH (K homology) RNA-binding domain surrounded by two flanking regions, the QUA1 and QUA2, which are involved in dimerization and RNA binding respectively [1] (Figure 1). In addition to this STAR domain, most STAR proteins contain domains or regions important for their proper function (Figure 1). Sam68 possess a nuclear localization signal and a tyrosine-rich domain at the C-terminus, which is phosphorylated by tyrosine kinases such as Src and Brk [2,11]. Sam68 also contain six proline-rich motifs, three to the N-terminal and three to the C-terminal of the STAR domain, which bind SH3 and WW domains to facilitate tyrosine phosphorylation [2], and two RG-rich (arginine–glycine-rich) motifs that are target sites for arginine methylation by PRMT1 (protein methyltransferase 1) [12]. Finally, Sam68 is also subject to lysine acetylation [13], SUMOylation [14] and serine/threonine phosphorylation by kinases such as Cdc2 (cell division control protein 2 homologue) [15], ERK1/2 (extracellular-signal-regulated kinase 1/2) [16] and Nek2 [NIMA (never in mitosis kinase)-related kinase 2] [17]. The domain architectures of SLM1 and SLM2/T-STAR are very similar to that of Sam68, with a deletion of the first 100 amino acids that form the N-terminus of Sam68 [4,5]. In addition to its STAR domain, QKI also contains a nuclear localization signal, a tyrosine-rich region and a proline-rich region. QKI has also been shown to be tyrosine phosphorylated by Src and Fyn [18]. In contrast, GLD-1 lacks such domains but has been shown to be phosphorylated by the serine/threonine kinases cyclin E and CDK2 (cyclin-dependent kinase 2) [19]. SF1 is the most divergent member of the STAR family. Its STAR domain contains the KH and the QUA2 regions but lacks the QUA1 dimerization domain. In addition, an N-terminal nuclear localization signal, and C-terminal zinc-knuckle and proline-rich regions are present in SF1.

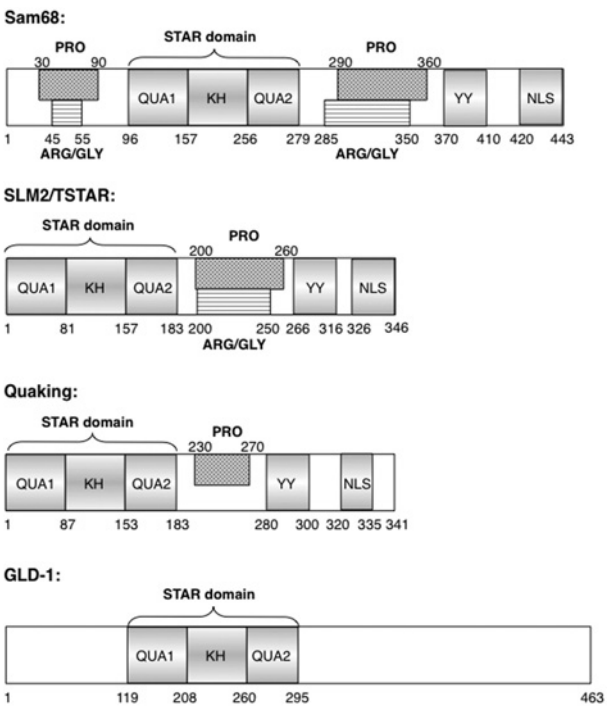
Function of STAR proteins in RNA metabolism

All identified STAR proteins have been shown to bind RNA and to be involved in post-transcriptional gene regulation. Sam68 affects various cellular processes such as

Key words: alternative splicing, branchpoint sequence, pre-mRNA, RNA binding, STAR proteins.
Abbreviations: FMR1, fragile X mental retardation 1; GLD-1, germline development protein 1; KH, K homology; NOVA, neuro-oncological ventral antigen; QKI, quaking; SELEX, systematic evolution of ligands by exponential enrichment; Sam68, Src-associated in mitosis of 68kDa; SF1, splicing factor 1; SLM, Sam68-like mammalian protein; STAR, signal transduction and activation of RNA.
[1]To whom correspondence should be addressed (email cyril.dominguez@le.ac.uk).

Figure 1 | Domain organization of STAR proteins

All STAR proteins possess a STAR domain responsible for RNA binding and composed of a central KH domain flanked by the QUA1 and QUA2 regions. In addition, other domains or motifs are present in STAR proteins, such as NLS (nuclear localization signals), tyrosine-rich (YY) regions, proline-rich (PRO) regions, and arginine/glycine-rich regions (ARG/GLY).



differentiation, cell cycle progression and apoptosis, through its direct involvement in alternative splicing [20]. Indeed, many pre-mRNA targets of Sam68 have been identified that encode proteins spanning a variety of cellular functions: mTOR (mammalian target of rapamycin) in adipogenesis [21], neurexin 1 in synapse function [22], cyclin D1 in cell cycle progression [23], SRSF1 (serine/arginine-rich splicing factor 1) (formerly ASF/SF2) in splicing [24], CD44 in cell migration [16] and Bcl-x (B-cell lymphoma X) in apoptosis [25]. In addition, Sam68 is involved in the nuclear export of the unspliced HIV RNA through binding the viral Rev protein and the viral RNA directly [26]. Whereas Sam68 is ubiquitously expressed in all tissues, SLM1 and SLM2/T-STAR exhibit preferential expression in the brain where they specifically act as alternative splicing regulators of the neurexin 1–3 genes [27,28]. QKI is involved in the regulation of alternative splicing, RNA export and mRNA stability of myelin-associated pre-mRNAs [29,30]. GLD-1 is a cytoplasmic protein that functions as a translational repressor [31,32]. SF1 is an essential SF that binds the pre-mRNA branchpoint sequence during the early stage of spliceosome formation [33].

Specific RNA recognition by STAR proteins

Several studies have investigated the RNA-binding specificity of STAR proteins. First, SELEX (systematic evolution of ligands by exponential enrichment) experiments have identified purine-rich RNA sequences specifically bound by Sam68 [34]. Most of these sequences contained a conserved UAAA motif. The consensus sequence has been later confirmed and extended to SLM2/TSTAR that was found, like Sam68, to bind specifically a U(A/U)AA motif [35]. The RNA-binding motifs specifically bound by QKI and GLD-1 have also been identified [36,37]. Both proteins recognize very similar motifs, with the consensus sequences being UACU(C/A)A for GLD-1 [36] and NA(A/C)UAA for QKI [37]. These motifs have been confirmed *in vitro* by SELEX [35] and *in vivo* by RIP (RNP immunoprecipitation)–ChIP microarray for GLD-1 [38] and PAR–CLIP (photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) for QKI [39]. The RNA motif recognized by QKI and GLD-1 is very similar to the highly conserved branchpoint sequence (UACUAAC) specifically bound by SF1 [33]. STAR proteins can therefore be divided into two subfamilies based on their RNA-binding specificities: the QKI/GLD-1/SF1 subfamily that recognizes a six-nucleotide UACUAA motif and the Sam68/SLM1/SLM2/TSTAR subfamily that recognizes specifically AU-rich RNAs containing the four-nucleotide UAAA motif.

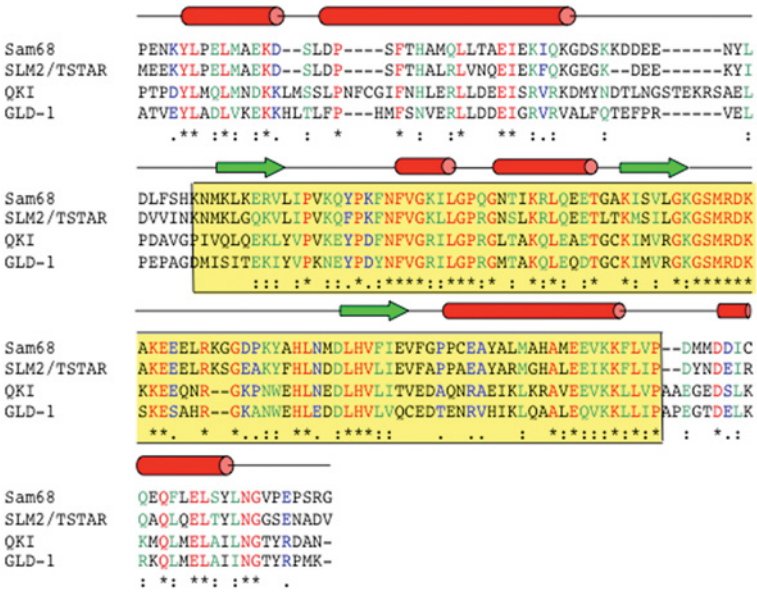
Structural investigations of RNA binding by STAR proteins

The RNA-binding STAR domain is very well conserved within the STAR family (Figure 2) and this domain is crucial for the proper function of these proteins in RNA metabolism. Structural studies of STAR proteins have mainly so far focused on the STAR domain free and in complex with RNA. These structures, as detailed below, explain the specificity of RNA recognition by certain members of the STAR family.

The first structure of a KH domain from the human FMR1 (fragile X mental retardation 1) protein was solved by NMR in 1996 [40]. This structure showed that the KH domain adopts a $\beta\alpha\alpha\beta\beta\alpha$ fold and demonstrated that a single-point mutation on this protein (I304N) leading to the phenotype of fragile X syndrome [41] was responsible for the unfolding of the KH domain. Later the structure of NOVA-2 (neuro-oncological ventral antigen 2) KH3 was solved in complex with an RNA hairpin [42]. This structure revealed how KH domains specifically bind RNA molecules and highlighted the importance of a highly conserved GXXG motif in RNA binding.

In 2001, the first structure of a KH domain from a STAR protein SF1 was solved by NMR in complex with the canonical branchpoint sequence, UAUACUAACAA [43] (Figure 3a). In contrast with the KH domains of FMR1 and NOVA, the KH domain of SF1 contains an additional flexible loop located between the second and third β-strands. Interestingly, this additional loop is conserved among the STAR family of proteins (Figure 2). Although this loop is not involved in RNA binding by SF1, deletion of this loop in Sam68 disrupted its RNA-binding ability [34].

Figure 2 | **Amino acid sequence alignment of the STAR domain of Sam68, SLM2/TSTAR, QKI and GLD-1**
The KH domain is highlighted in yellow. Secondary-structure elements are shown above the sequences.



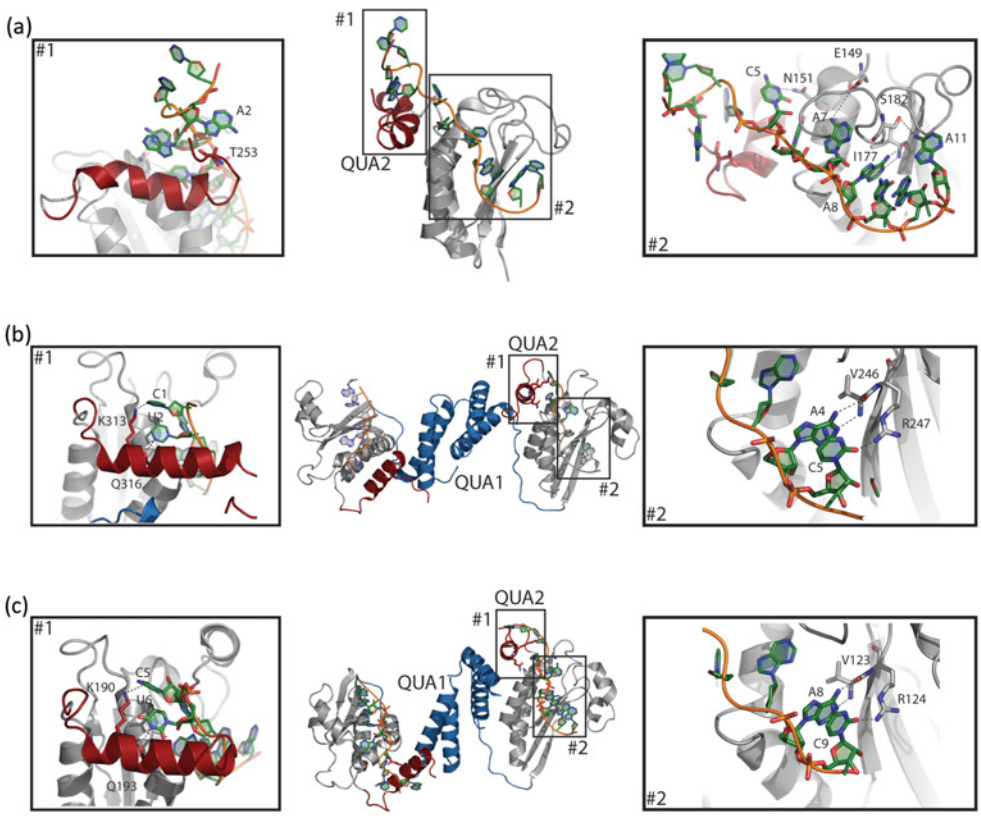
The RNA recognition by SF1 is driven by a mixture of hydrophobic, electrostatic and hydrogen bond contacts [43]. The KH domain of SF1 specifically recognizes the $U^{6}AAC^{9}$ sequence of the $U^{1}AUACUAACAA^{11}$ RNA. The GXXG motif (GPRG in the case of SF1) recognizes specifically the base of U^{6} and contacts the backbone of the RNA. Specific hydrogen bonds to the bases of the RNA are formed in SF1 between Glu^{149} and A^{7}, Ile^{177} and A^{8}, and Ser^{182} and A^{11} (Figure 3a). In addition, Asn^{151} contacts specifically the base of C^{5} via a hydrogen bond. Interestingly, this structure also revealed that the QUA2 region that is conserved among the STAR family (Figure 2) adopts a helical conformation, packs against the KH domain and is also involved in RNA binding (Figure 3a). Indeed, the QUA2 domain recognizes four nucleotides, $A^{2}UAC^{5}$, of the RNA. Specifically, Thr^{253} contacts A^{2} through specific hydrogen bonds, and Arg^{255}, Ala^{248}, Leu^{244} and Leu^{247} make hydrophobic contacts to A^{4} and C^{5}. This study also demonstrated that, although QUA2 is flexible in the absence of RNA, the α-helix is already pre-formed in solution and does not occur upon RNA binding [43]. This was confirmed later through an NMR dynamic study of the KH–QUA2 domain of the *Xenopus* QKI protein showing that the QUA2 region is helical but highly flexible in the absence of RNA and becomes rigid in the presence of RNA [44].

More recently, the structure of the QUA1 domains of Sam68, GLD-1 and QKI have been solved [45–47]. These structures showed that the QUA1 region of these proteins consists of a helix–turn–helix motif stabilized by a network of hydrophobic contacts. The dimer is formed by an almost perpendicular stacking of the two monomers and is stabilized by a network of hydrophobic contacts and hydrogen bonds involving mainly residues of the turn between the two helices. The structures demonstrated that the QUA1 domain alone is sufficient for STAR protein dimerization. In the case of Sam68, it has been proposed that the phosphorylation of one tyrosine residue (Tyr^{103}), localized in the QUA1 domain, could be involved in the dissociation of the homodimer, and mutation of this residue leads to a loss of splicing activity [45]. This suggests that the dimerization of the protein is compulsory for its function in alternative splicing and that tyrosine phosphorylation could regulate Sam68 function by disrupting its ability to dimerize.

In 2013, the structures of the full STAR domains of QKI and GLD-1 in complex with their target RNAs were solved by X-ray crystallography [48] (Figures 3b and 3c). These structures revealed that the RNA is only bound by the KH and the QUA2 regions, whereas the QUA1 region is involved in dimerization and protein–protein interaction with the QUA2 region. The RNA sequences bound by QKI and GLD-1 are very similar to the one bound by SF1 (CUAAC) and, consequently, the RNA recognition by GLD1 and QKI proteins involves the same conserved residues of KH and QUA2 as SF1. Specifically, the base of C^{1} is specifically recognized by Lys^{190} and Lys^{313} of QKI and GLD-1 respectively. The UAA motif that overlaps perfectly on all the structures contacts Gln^{193} (QKI) and Gln^{316} (GLD1) for U^{2}, and Asn^{97} (QKI) for A^{3}. The last adenine forms two hydrogen bonds with the main chain of the protein, a valine residue in both QKI and GLD-1. The last cytosine (C5) is recognized by an arginine residue (Arg^{124} and Arg^{247} for QKI and GLD-1 respectively). As for the structure of SF1, the QUA2 domain of QKI and GLD1 is strongly involved in the interaction with the RNA. These structures also show that the QUA1 domain interacts with the QUA2 region, forming a three-helix bundle. This

Figure 3 | Structures of members of the STAR family of proteins with their RNA

(**a**) SF1, (**b**) QKI and (**c**) GLD-1. The QUA1 domain is coloured blue, the KH domain is coloured grey, the QUA2 domain is coloured red, and the RNA is coloured orange and green. The structures of GLD-1 and QKI show the dimerization of the QUA1 domain. Specific contacts between the protein and the RNA are enlarged and labelled.



interaction induces a precise orientation of the KH-QUA2 domains in the dimer and positions the two RNA-binding surfaces at opposite ends of the dimer. This orientation suggests that one dimer can bind a single RNA molecule only if the two binding motifs (CUAAC) are separated by more than ten nucleotides. In addition to the QUA1–QUA2 interaction, the structure of the GLD-1 STAR domain shows that the QUA1–KH linker contacts the KH domain through hydrogen bonds and van der Walls contacts. It was proposed that these contacts facilitate the orientation of the QUA1, KH and QUA2 regions within the STAR domain. It should be noted, however, that this linker is clearly visible in the GLD1 structure, but not in the QKI structure, suggesting that the linker is more flexible in the case of QKI. This flexibility could be important for STAR proteins to bind two RNA motifs that are relatively close (fewer than ten nucleotides on the same RNA) [49]. Further structural studies will be needed to address this issue.

In contrast with SF1, QKI and GLD-1 that specifically bind a CUAAC RNA motif, Sam68, SLM1 and SLM2/TSTAR bind specifically AU-rich RNA sequences (see above). The molecular details of the specific AU-rich RNA recognition by these proteins remain unknown. We have therefore initiated a structural study of Sam68 and STAR in complex with AU-rich RNA [50]. Using NMR spectroscopy, we have identified the optimal protein constructs of Sam68 and SLM2/TSTAR, and, through an AU-rich RNA NMR screen, the optimal RNA sequence for the structure determination of these complexes.

Funding

This work was supported by a Medical Research Council Career Development Award to C.D. [grant number G1000526] and by a College of Medicine, Biological Sciences and Psychology, University of Leicester, studentship to J.F.

References

1 Vernet, C. and Artzt, K. (1997) STAR, a gene family involved in signal transduction and activation of RNA. Trends Genet. **13**, 479–484 CrossRef PubMed
2 Taylor, S.J. and Shalloway, D. (1994) An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis. Nature **368**, 867–871 CrossRef PubMed
3 Fumagalli, S., Totty, N.F., Hsuan, J.J. and Courtneidge, S.A. (1994) A target for Src in mitosis. Nature **368**, 871–874 CrossRef PubMed
4 Di Fruscio, M., Chen, T. and Richard, S. (1999) Characterization of Sam68-like mammalian proteins SLM-1 and SLM-2: SLM-1 is a Src substrate during mitosis. Proc. Natl. Acad. Sci. U.S.A. **96**, 2710–2715 CrossRef PubMed

5 Venables, J.P., Vernet, C., Chew, S.L., Elliot, D.J., Cowmeadow, R.B., Wu, J., Cooke, H.J., Artzt, K. and Eperon, I.C. (1999) T-STAR/ETOILE: a novel relative of SAM68 that interacts with an RNA-binding protein implicated in spermatogenesis. Hum. Mol. Genet. **8**, 959–969 CrossRef PubMed

6 Jones, A.R. and Schedl, T. (1995) Mutations in *gld-1*, a female germ cell-specific tumor suppressor gene in *Caenorhabditis elegans*, affect a conserved domain also found in Src-associated protein Sam68. Genes Dev. **9**, 1491–1504 CrossRef PubMed

7 Ebersole, T.A., Chen, Q., Justice, M.J. and Artzt, K. (1996) The *quaking* gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins. Nat. Genet. **12**, 260–265 CrossRef PubMed

8 Zaffran, S., Astier, M., Gratecos, D. and Semeriva, M. (1997) The *held out wings* (*HOW*) *Drosophila* gene encodes a putative RNA-binding protein involved in the control of muscular and cardiac activity. Development **124**, 2087–2098 PubMed

9 Baehrecke, E.H. (1997) *who* encodes a KH RNA binding protein that functions in muscle development. Development **124**, 1323–1332 PubMed

10 Arning, S., Gruter, P., Graeme, B. and Kramer, A. (1996) Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. RNA **2**, 794–810 PubMed

11 Derry, J.J., Richard, S., Valderrama Carvajal, H., Ye, X., Vasioukhin, V., Cochrane, A.W., Chen, T. and Tyner, A.L. (2000) Sik (BRK) phosphorylates Sam68 in the nucleus and negatively regulates its RNA binding ability. Mol. Cell. Biol. **20**, 6114–6126 CrossRef PubMed

12 Côté, J., Boisvert, F.-M., Boulanger, M.-C., Bedford, M.T. and Richard, S. (2003) Sam68 RNA binding protein is an *in vivo* substrate for protein arginine N-methyltransferase 1. Mol. Biol. Cell **14**, 274–287 CrossRef PubMed

13 Babic, I., Jakymiw, A. and Fujita, D.J. (2004) The RNA binding protein Sam68 is acetylated in tumor cell lines, and its acetylation correlates with enhanced RNA binding activity. Oncogene **23**, 3781–3789 CrossRef PubMed

14 Babic, I., Cherry, E. and Fujita, D.J. (2006) SUMO modification of Sam68 enhances its ability to repress cyclin D1 expression and inhibits its ability to induce apoptosis. Oncogene **25**, 4955–4964 CrossRef PubMed

15 Resnick, R.J., Taylor, S.J., Lin, Q. and Shalloway, D. (1997) Phosphorylation of the Src substrate Sam68 by Cdc2 during mitosis. Oncogene **15**, 1247–1253 CrossRef PubMed

16 Matter, N., Herrlich, P. and König, H. (2002) Signal-dependent regulation of splicing via phosphorylation of Sam68. Nature **420**, 691–695 CrossRef PubMed

17 Naro, C., Barbagallo, F., Chieffi, P., Bourgeois, C.F., Paronetto, M.P. and Sette, C. (2014) The centrosomal kinase NEK2 is a novel splicing factor kinase involved in cell survival. Nucleic Acids Res. **42**, 3218–3227 CrossRef PubMed

18 Zhang, Y., Lu, Z., Ku, L., Chen, Y., Wang, H. and Feng, Y. (2003) Tyrosine phosphorylation of QKI mediates developmental signals to regulate mRNA metabolism. EMBO J. **22**, 1801–1810 CrossRef PubMed

19 Jeong, J., Verheyden, J.M. and Kimble, J. (2011) Cyclin E and Cdk2 control GLD-1, the mitosis/meiosis decision, and germline stem cells in *Caenorhabditis elegans*. PLoS Genet. **7**, e1001348 CrossRef PubMed

20 Sánchez-Jiménez, F. and Sánchez-Margalet, V. (2013) Role of Sam68 in post-transcriptional gene regulation. Int. J. Mol. Sci. **14**, 23402–23419 CrossRef PubMed

21 Huot, M.-É., Vogel, G., Zabarauskas, A., Ngo, C.T.-A., Coulombe-Huntington, J., Majewski, J. and Richard, S. (2012) The Sam68 STAR RNA-binding protein regulates mTOR alternative splicing during adipogenesis. Mol. Cell **46**, 187–199 CrossRef PubMed

22 Iijima, T., Wu, K., Witte, H., Hanno-Iijima, Y., Glatter, T., Richard, S. and Scheiffele, P. (2011) SAM68 regulates neuronal activity-dependent alternative splicing of neurexin-1. Cell **147**, 1601–1614 CrossRef PubMed

23 Paronetto, M.P., Cappellari, M., Busà, R., Pedrotti, S., Vitali, R., Comstock, C., Hyslop, T., Knudsen, K.E. and Sette, C. (2010) Alternative splicing of the cyclin D1 proto-oncogene is regulated by the RNA-binding protein Sam68. Cancer Res. **70**, 229–239 CrossRef PubMed

24 Valacca, C., Bonomi, S., Buratti, E., Pedrotti, S., Baralle, F.E., Sette, C., Ghigna, C. and Biamonti, G. (2010) Sam68 regulates EMT through alternative splicing-activated nonsense-mediated mRNA decay of the SF2/ASF proto-oncogene. J. Cell Biol. **191**, 87–99 CrossRef PubMed

25 Paronetto, M.P., Achsel, T., Massiello, A., Chalfant, C.E. and Sette, C. (2007) The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. J. Cell Biol. **176**, 929–939 CrossRef PubMed

26 Reddy, T.R., Xu, W., Mau, J.K., Goodwin, C.D., Suhasini, M., Tang, H., Frimpong, K., Rose, D.W. and Wong-Staal, F. (1999) Inhibition of HIV replication by dominant negative mutants of Sam68, a functional homolog of HIV-1. Rev. Nat. Med. **5**, 635–642 CrossRef

27 Ehrmann, I., Dalgliesh, C., Liu, Y., Danilenko, M., Crosier, M., Overman, L., Arthur, H.M., Lindsay, S., Clowry, G.J., Venables, J.P. et al. (2013) The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. PLoS Genet. **9**, e1003474 CrossRef PubMed

28 Iijima, T., Iijima, Y., Witte, H. and Scheiffele, P. (2014) Neuronal cell type-specific alternative splicing is regulated by the KH domain protein SLM1. J. Cell Biol. **204**, 331–342 CrossRef PubMed

29 Wu, J.I., Reed, R.B., Grabowski, P.J. and Artzt, K. (2002) Function of quaking in myelination: regulation of alternative splicing. Proc. Natl. Acad. Sci. U.S.A. **99**, 4233–4238 CrossRef PubMed

30 Li, Z., Zhang, Y., Li, D. and Feng, Y. (2000) Destabilization and mislocalization of myelin basic protein mRNAs in *quaking* dysmyelination lacking the QKI RNA-binding proteins. J. Neurosci. **20**, 4944–4953 PubMed

31 Jan, E., Motzny, C.K., Graves, L.E. and Goodwin, E.B. (1999) The STAR protein, GLD-1, is a translational regulator of sexual identity in *Caenorhabditis elegans*. EMBO J. **18**, 258–269 CrossRef PubMed

32 Schumacher, B., Hanazawa, M., Lee, M.-H., Nayak, S., Volkmann, K., Hofmann, R., Hengartner, M., Schedl, T. and Gartner, A. (2005) Translational repression of *C. elegans* p53 by GLD-1 regulates DNA damage-induced apoptosis. Cell **120**, 357–368 CrossRef PubMed

33 Berglund, J.A., Chua, K., Abovich, N., Reed, R. and Rosbash, M. (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell **89**, 781–787 CrossRef PubMed

34 Lin, Q., Taylor, S.J. and Shalloway, D. (1997) Specificity and determinants of Sam68 RNA binding: implications for the biological function of K homology domains. J. Biol. Chem. **272**, 27274–27280 CrossRef PubMed

35 Galarneau, A. and Richard, S. (2009) The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. BMC Mol. Biol. **10**, 47 CrossRef PubMed

36 Ryder, S.P., Frater, L.A., Abramovitz, D.L., Goodwin, E.B. and Williamson, J.R. (2004) RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. Nat. Struct. Mol. Biol. **11**, 20–28

37 Ryder, S.P. (2004) Specificity of the STAR/GSG domain protein Qk1: implications for the regulation of myelination. RNA **10**, 1449–1458 CrossRef PubMed

38 Wright, J.E., Gaidatzis, D., Senften, M., Farley, B.M., Westhof, E., Ryder, S.P. and Ciosk, R. (2011) A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1. EMBO J. **30**, 533–545 CrossRef PubMed

39 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, Jr, M., Jungkamp, A.C., Munschauer, M. et al. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell **141**, 129–141 CrossRef PubMed

40 Musco, G., Stier, G., Joseph, C., Morelli, M.A.C., Nilges, M., Gibson, T.J. and Pastore, A. (1996) Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome. Cell **85**, 237–245 CrossRef PubMed

41 De Boulle, K., Verkerk, A.J., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van den Bos, F., de Graaff, E., Oostra, B.A. and Willems, P.J. (1993) A point mutation in the *FMR-1* gene associated with fragile X mental retardation. Nat. Genet. **3**, 31–35 CrossRef PubMed

42 Lewis, H.A., Musunuru, K., Jensen, K.B., Edo, C., Chen, H., Darnell, R.B. and Burley, S.K. (2000) Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell **100**, 323–332 CrossRef PubMed

43 Liu, Z., Luyten, I., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A. and Sattler, M. (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. Science **294**, 1098–1102 CrossRef PubMed

44 Maguire, M.L., Guler-Gane, G., Nietlispach, D., Raine, A.R.C., Zorn, A.M., Standart, N. and Broadhurst, R.W. (2005) Solution structure and backbone dynamics of the KH-QUA2 region of the *Xenopus* STAR/GSG quaking protein. J. Mol. Biol. **348**, 265–279 CrossRef PubMed

45 Meyer, N.H., Tripsianes, K., Vincendeau, M., Madl, T., Kateb, F., Brack-Werner, R. and Sattler, M. (2010) Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. J. Biol. Chem. **285**, 28893–28901 CrossRef PubMed

46 Beuck, C., Szymczyna, B.R., Kerkow, D.E., Carmel, A.B., Columbus, L., Stanfield, R.L. and Williamson, J.R. (2010) Structure of the GLD-1 homodimerization domain: insights into STAR protein-mediated translational regulation. Structure **18**, 377–389 CrossRef PubMed
47 Beuck, C., Qu, S., Fagg, W.S., Ares, Jr, M. and Williamson, J.R. (2012) Structural analysis of the quaking homodimerization interface. J. Mol. Biol. **423**, 766–781 CrossRef PubMed
48 Teplova, M., Hafner, M., Teplov, D., Essig, K., Tuschl, T. and Patel, D.J. (2013) Structure–function studies of STAR family quaking proteins bound to their *in vivo* RNA target sites. Genes Dev. **27**, 928–940 CrossRef PubMed
49 Cukier, C.D. and Ramos, A. (2010) Creating a twin STAR. Structure **18**, 279–80 CrossRef PubMed
50 Foot, J.N., Feracci, M. and Dominguez, C. (2014) Screening protein–single stranded RNA complexes by NMR spectroscopy for structure determination. Methods **65**, 288–301 CrossRef PubMed

Received 1 April 2014
doi:10.1042/BST20140081