Investigations into RNA-binding proteins involved in eukaryotic gene regulation

Stephanie Helder

A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy

School of Life and Environmental Sciences
Faculty of Science
The University of Sydney
Sydney, Australia

December 2017

Declaration

The work described in this thesis was performed between March 2014 and August 2017 in the School of Life and Environmental Sciences at the University of Sydney. All experiments were conducted by the author unless otherwise specified. This work has not been submitted, in part or in full, for the purpose of obtaining a higher degree at any other institution.

Stephanie Helder

December 2017

Abstract

The flood of RNA-related research in recent decades has revealed RNA to be a structurally and functionally diverse class of molecule, one that generates an intricate network of regulation that has been pivotal to the evolution of complex lifeforms. In order to elucidate how RNA achieves biological function through the formation of ribonucleoprotein complexes, characterisation of RNA recognition by RNA-binding proteins is an essential step. The rules governing the interaction of RNA and RNA-binding proteins have proved difficult to define, and in many instances, it is not understood how specificity is achieved. Knowledge of these rules is crucial to our understanding of RNA-related functions and their role in disease, and requires further in-depth characterisation of a wide variety of ribonucleoprotein complexes.

The research in this Thesis details the RNA-binding behaviour of two reported RNA-binding proteins. Firstly, the RNA-binding behaviour of the *Drosophila* transcription factor bicoid is investigated. For many years it has been believed that the bicoid homeodomain binds the 3′-UTR of the *caudal* mRNA transcript, yet no binding site or specificity determinants have been reported. The work reported here attempts to characterise this interaction. Further, other domains in the protein are examined with a view to understanding how biological specificity might be achieved. Secondly, characterisation of the RNA-binding behaviour of the heterodimeric pair of transcription elongation factors, Spt4 and Spt5, is reported. This heterodimer is known to be an important player in transcription and yet surprisingly little is known about its function. In the present work, the AA-repeat RNA-binding properties of these proteins are investigated, and complex binding behaviour is reported. Overall, it is shown that the elucidation of RNA-binding activity by proteins is often not straightforward, requiring the application of multiple and increasingly sophisticated techniques if we are to grasp the underlying biology.

Acknowledgements

No amount of warning about the difficulties of doing a PhD can really prepare you for the vicissitudes of your own journey. I have many people to thank for their contributions which have helped me through what has been a rather gruelling but exceedingly valuable experience.

Foremost, I would like to thank to my supervisor, Joel Mackay, for giving me the opportunity to work in such a flourishing lab - the success of which is the result of his dedication and ability, and for providing me with projects that fostered my scientific curiosity. I am grateful for the substantial amount of time and expertise that he contributed to my thesis and overall scientific education.

Much appreciation goes to all of the Structural Biology group in G08 for their help and advice. I give special thanks to Ann Kwan for her tireless assistance with all things NMR, to Jason Low for his advice (and detailed protocols) on a wide variety of techniques, to Dorothy Wai for teaching me numerous experimental procedures, to Sandro Ataide, James Walshe and Janine Flores for their assistance with RNA-related matters and for the use of their reagents and equipment, to Ingrid MacIndoe, Ana Silva, Lorna Wilkinson-White and Phillipa Stokes for being general go-to people, and to Taylor Szyszka for her assistance with some NMR acquisitions.

I must also thank my friends and family for helping me in indirect ways. The last eight years of scientific training which has ultimately resulted in the completion of this PhD would not have been possible without the support, both material and emotional, from my mother. I am also indebted to my friend and mentor, Ed Brackenreg, who has likewise been a pillar of support for me; his encouragement and counsel kept me going and always helped me to see my own progress when I could not. Finally I have to thank my partner Pat for his patience and understanding, particularly this year while life has been somewhat on hold; his ability to counteract my general neuroticism helped me immensely and kept me laughing even through the stressful times.

Abbreviations

Ago	Argonaute
Ago2	Argonaute-2
ARM	arginine rich motif
BHD	Bicoid homeodomain construct
bp	base pair
BRE	Bicoid recognition element
BRRM	Bicoid RNA recognition motif
cad	*caudal*
CLIP	cross-linking immunoprecipitation
cryo-EM	cryo-electron microscopy
CTR	C-terminal repeat
DEPC	diethyl pyrocarbonate
dsDNA	double-stranded DNA
dsRBM	double stranded RNA-binding motif
DSS	2,2-dimethyl-2-silapentane-5-sulfonic acid
EMSA	electrophoretic mobility shift assay
FMRP	Fragile X Mental Retardation Protein
HDER	Bicoid homeodomain with extra arginines construct
HSQC	Heteronuclear single quantum coherence
IDR	intrinsically disordered region
IR	infrared
IRP1	iron regulatory protein 1
KH	K homology
KOW	Kyprides, Ouzounis, Woese domain
LC	low complexity
lncRNA	long non-coding RNA
miRNA	microRNA
mRBP	mRNA-binding protein
mRNP	messenger ribonucleoprotein
MST	microscale thermophoresis
ncRNA	non-coding RNA
NGN	NusG N-terminal domain
NS	Number of scans
nt	nucleotide
PABP	Poly-A binding protein
PASR	promoter-associated RNA
P-body	Processing body
PDB	protein data bank
piRNA	PIWI-interacting RNA
PLD	prion-like domain
PP7	Pentaprobe 7
RBD	RNA-binding domain

RBP	RNA-binding protein
RNAi	RNA interference
RNAP	RNA polymerase
RNP	ribonucleoprotein
RRM	RNA-recognition motif
SELEX	Systematic Evolution of Ligands by Exponential enrichment
snRNA	small nuclear RNA
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
TEC	transcription elongation complex
TF	transcription factor
UTR	untranslated region
WD40	tryptophan-aspartic acid domain
zen	*zerkneullt*
ZF	zinc finger

Table of Contents

Chapter 1: Introduction

1.1 Background

The work detailed in this Thesis investigates the RNA-binding capacity of two different proteins. RNA has immense structural and functional diversity, the breadth of which our current knowledge is just beginning to uncover. This makes the study of RNA one of the most challenging yet potentially rewarding fields in modern biochemistry.

1.1.1 The complexity of the eukaryotic transcriptome

The biological processes that yield complex lifeforms from chemical information are extremely intricate. We now appreciate that the majority of DNA within eukaryotic genomes is transcribed into RNA [1-3], and that the vast majority of this is non-coding RNA (ncRNA); that is, it is not translated into a protein. Whilst there have been claims that such pervasive transcription is largely transcriptional noise [4, 5], evidence that conservation can be found outside the sequence level continues to grow [6, 7]. Moreover, the observation that the non-coding portion of transcriptomes correlates consistently with biological complexity has led to suggestions that the proliferation of non-coding RNA has been pivotal to the evolution of developmentally complex lifeforms [8].

RNA function has been increasingly shown to be of a regulatory nature [9], extending far beyond the canonical protein coding messengers, ribosomal components and transfer RNAs. The 1990s saw the first hints of this change in our understanding, starting what would later become burgeoning fields of research: antisense RNA gene silencing was demonstrated [10] and the first long non-coding RNAs (lncRNAs) [11, 12], microRNAs (miRNAs) [13, 14] and small nuclear RNAs (snRNAs) [15] were discovered.

The extent of RNA involvement in gene regulation, and the sophisticated nature of these regulatory pathways, however, has only become apparent this century. Notably, small interfering RNAs (siRNA) and miRNAs, both short RNAs ~21–25 nucleotides (nt) in length, have now been well characterised as part of the RNA interference (RNAi) system, whereby genes are silenced post-transcriptionally through the pairing of sense:antisense RNA [16]. Additionally, miRNAs have been shown to: be implicated in gene silencing at the transcriptional, as well as post-transcriptional, level [17]; in certain instances promote translation [18]; and to be regulated dynamically to yield different post-transcriptional modifications [19] and diverse isomers across tissues and developmental phases [20].

Numerous other small RNAs are involved in gene regulation. Two examples of such small RNAs are 1) PIWI-interacting RNAs (piRNAs) that regulate gene expression at multiple levels; although these were originally discovered in gametogenesis [21], roles for piRNAs in somatic cells continue to be unearthed [22, 23]; and 2) small nucleolar RNAs (snoRNAs) that act as guides for RNA modifications [24]. More small RNAs have been discovered in recent years; for instance, transcription initiation RNAs (tiRNAs) and 3′ splice site RNAs (spliRNAs) are small RNAs of ~18-nt that are derived from these regions and play regulatory functions [25]. Promoter associated RNAs (PASRs) are double-stranded RNAs involved in gene activation and silencing through binding promoter elements of relevant genes [26]. The complexity of eukaryotic transcription along with some of the diverse RNA species transcribed are depicted in Figure 1.1.

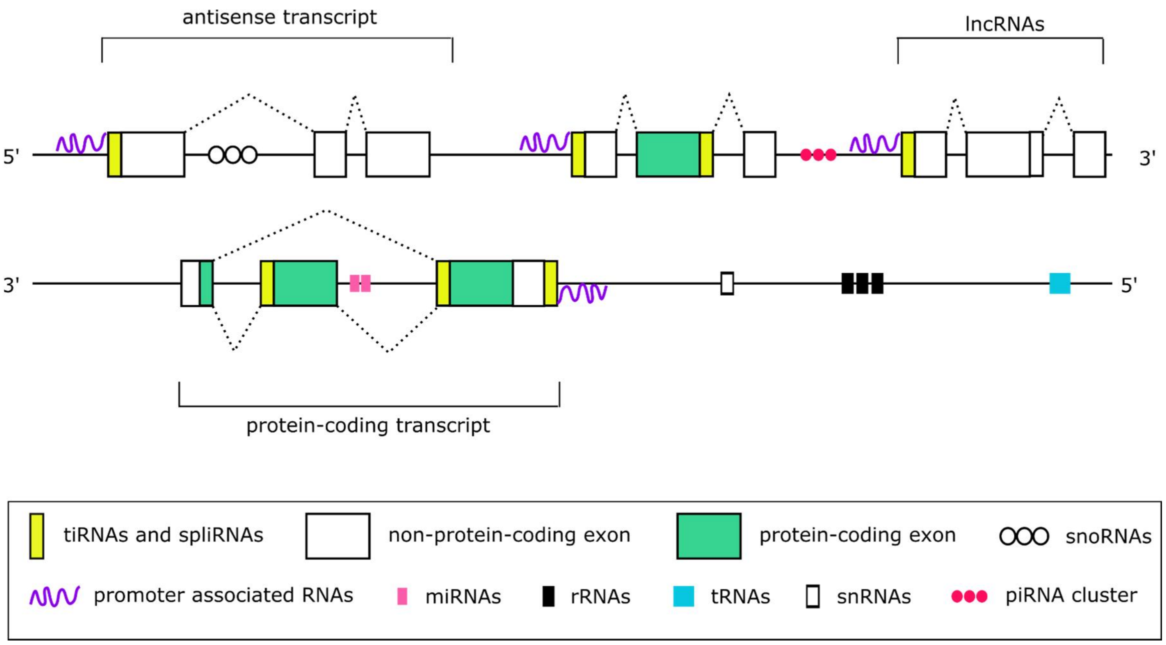


Figure 1.1. The variety of RNA in the eukaryotic transcriptome.
Eukaryotic DNA is pervasively transcribed in both directions in to many different types of RNA species, some of which are represented here. Abbreviations: tiRNAs, transcription initiation RNAs; spliRNAs, splice site RNAs; snoRNAs, small nuceleolar RNAs; miRNAs, microRNAs; rRNAs, ribosomal RNAs; tRNAs, transfer RNAs; snRNAs, small nuclear RNAs; piRNAs, PIWI-interacting RNAs. Protein coding genes encode messenger RNA (mRNA) and non-protein coding genes encode lncRNAs. Splice sites are indicated by dotted lines. This diagram is adapted from Morris and Mattick (2014) [9].

At the larger end of the spectrum, lncRNAs, broadly classified as ncRNAs greater than 200-nt, have been a hot topic of research. lncRNAs are abundant in mammals, with tens of thousands described in humans [27, 28], and they display tissue-specific differential expression [29, 30]. Although the majority of lncRNAs await characterisation, the emerging picture is that lncRNAs are important in both nuclear and cytoplasmic gene regulation, with predominant roles in cell differentiation and development [31]. A recent high-throughput study found that 89% of lncRNAs that modified cell growth acted in a cell-type specific manner [32]. Further, multiple studies have shown that lncRNA expression is more specific to cell type than protein expression [29, 33], supporting the idea that lncRNAs act as cell fate regulators against a backdrop of (relatively) more generic protein composition [9].

Recent data even suggest that mRNAs can have functions beyond their canonical protein coding role. For example, mRNA can act as a competitive binder to sequester miRNAs [34]. Indeed, mRNA regulatory function has proven even more sophisticated than expected. 3′ untranslated regions (UTRs) have been shown, in certain instances, to be translated in isolation from their associated 'parent' transcripts, possibly acting as regulatory molecules in *trans* [35]. Further, protein coding sequences have been shown to contain information apart from codon specification. As illustrations of this multiplicity: an alternatively spliced mRNA molecule has been shown to upregulate translation of its own gene [36], coding sequences can act as enhancers in a tissue-specific manner [37] and human codons have been shown to be often occupied by transcription factors *in vivo* [38].

What emerges is a picture of RNA as a multifarious class of molecules, one that generates an intricate network of regulation that has been pivotal to the evolution of life. Such discoveries highlight both how far we have come in our knowledge of RNA functionality since the central dogma was proposed, and also how much we still have to learn.

1.1.2 RNA-binding proteins

Overwhelmingly, RNA associates with RNA-binding proteins (RBPs) in order to achieve biological function, forming ribonucleoprotein (RNP) complexes. RBPs are often modular, composed of multiple RNA-binding domains (RBDs) as well as a wide variety of effector and other domains. RBDs recognise sequence and/or structural features of RNA and are usually deeply conserved across phyla. Individual RBDs typically bind RNA with insufficient affinity and specificity to stipulate *in vivo* binding sites, but gains in both affinity and specificity are achieved through the use of multiple RBDs [39]. Most RBDs characterised to date are canonical RBDs: discrete, globular domains with well characterised RNA-binding activity. However, an increasing number of non-canonical and dual function domains are being reported as having RNA-binding activity.

1.1.2.1 Canonical RBDs

RNA-recognition motifs (RRMs) constitute the most common RBD in higher vertebrates and exist in all phyla [40]. The RRM family comprises many subclasses due to the diversity seen in their sequence and structural features, with sequence conservation generally only ~30% between RRM subclasses [41]. The typical RRM fold consists of around 90 amino acids with two α-helices packed against a β-sheet made up of four β-strands, connected by loops. Within this core fold, the lengths of the interconnecting loops can all vary, and numbers of both α-helices and β-strands can differ [40]. The most prevalent feature is the presence of two consensus sequences in specific β-strands roughly thirty residues apart, RNP-1 and RNP-2. These degenerate motifs are characterised by a high proportion of hydrophobic residues that make stacking interactions with RNA bases. The single RRM of Fox1 is a typical example, utilising its β-sheet as the RNA-binding interface, with three conserved aromatic residues forming stacking interactions with bases, and specificity conferred by hydrogen bonds from residues located in the loops [42] (Figure 1.2(A)). In some cases, the β-sheet is not used as the binding surface. For example, YxiN utilises the opposite face of its RRM to recognise RNA, with both loop and α-helical residues forming bonding interactions with the RNA bases [43], as shown in Figure 1.2(B). In even more divergent cases, pseudo-RRMs have a conserved heptapeptide in the α1 helix that provides most of the RNA-binding residues [44, 45], and quasi-RRMs use loop regions to contact the RNA target [46]. In RRMs that employ RNP-1 and RNP-2 motifs for RNA binding, these motifs usually do not account for the biological specificities of particular RRMs; target recognition is often supplemented by flanking sequences, as well as loops and additional β-strands that extend the β-sheet [40].

hnRNP K homology (KH) domains (~70 residues) also adopt a two layer α/β sandwich fold similar to that of RRMs and generally recognise 4-nt of single-stranded DNA (ssDNA) or single-stranded RNA (ssRNA) [47]. Interactions with RNA are provided by a binding cleft containing a G-X-X-G motif located between two α-helices. Despite this conserved binding surface, characterised KH domains bind a variety of sequences. Stacking interactions are notably absent; specificity is usually provided by the binding cleft through hydrogen bonding, and lysine residues often contact the sugar-phosphate backbone via electrostatic interactions. The Nova-2 KH domain bound to ssRNA in Figure 1.2(F) illustrates these features, with the binding cleft shown as a small helical portion [48].

Zinc fingers (ZFs) are small domains (~20-60 residues), named after their common feature of coordinating at least one zinc atom that stabilises their compact fold, with different subtypes sometimes identified according to their folds and to the complement of cysteines and histidines coordinating the zinc atom. The classical (C2H2) ZF is most well-known for its DNA-binding role in transcription factors (TFs), but in certain instances can bind RNA as well (the renowned RNA-binding properties of the classical ZFs of TFIIIA are outlined in the upcoming Section 1.2.1). The RanBP2-type ZF, primarily a protein:protein interaction domain, has been shown to bind RNA with a core GGU motif [49]. The most well-characterised RNA-binding ZF domain is the CCCH-type ZF; multiple CCCH-type ZFs have been reported to bind RNA [50-52] and there are three structures of this domain bound to RNA published to date. Like many RBDs, CCCH-type ZF containing RBPs often have multiple copies of the domain with sequence specific recognition achieved through concerted binding. The structure of tandem CCCH-type ZFs from the mRNA-binding protein (mRBP) Tis11d demonstrates sequence specific RNA recognition provided by hydrogen bonds from the protein backbone, stabilised by stacking interactions from aromatic residues, with each ZF recognising four bases in a modular fashion

(Figure 1.2(C)) [53]. In contrast, the three ZFs of Unkempt recognise only four RNA bases in total, facilitated by side and main chain hydrogen bonds (Figure 1.2(D)) [54]. These two structures indicate that there is substantial diversity both in the folds of the domain and their recognition mechanisms.

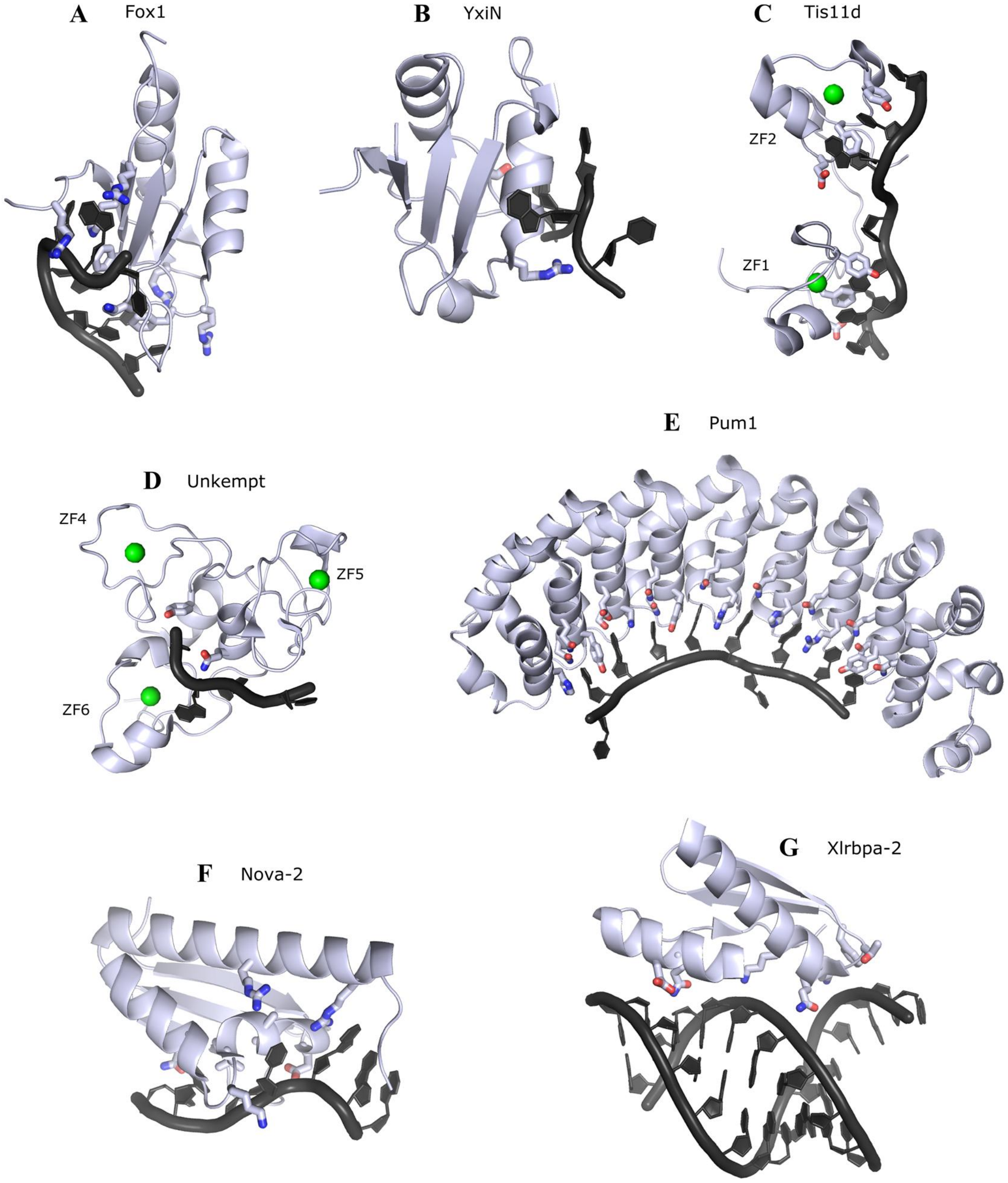


Figure 1.2. RNA recognition by canonical RNA-binding domains.
(A) RRM of Fox-1 bound to ssRNA (PDB: 2ERR). **(B)** RRM of YxiN bound to ssRNA (PDB: 3MOJ). **(C)** ZF1 and ZF2 of TIS11d bound to AU-rich element mRNA (PDB: 1RGO). **(D)** ZF4-6 of Unkempt bound to UAG RNA (PDB: 5ELK). **(E)** Pumilio-1 bound to RNA target with one base flipped out (PDB: 3BSX). **(F)** Nova-2 KH3 domain bound to ssRNA (PDB: 1EC6). **(G)** dsRBM2 of Xlrbpa-2 bound to A-form RNA (PDB: 1DI2). Schematics were made in Pymol. Protein is indicated in *pale blue*, with side chains shown for RNA-binding residues (*blue* indicates nitrogen, *red* indicates oxygen). Coordinated zinc atoms in zinc fingers are shown as *green* spheres.

A number of helical repeat proteins act as ssRNA-binding domains and are distinctive in their large size. PUF and PPR domains fall into this class. PUF domains are typified by a repeating three α-helical unit (~30-40 residues), each of which generally recognises one base [55]. One residue from each repeat stacks between successive RNA bases, with another two creating hydrogen bonds or van der Waals contacts with a single base, creating an overall-characteristic crescent shaped structure. Bases have been observed to tilt away from the binding interface [56], resulting in some flexibility in RNA recognition. These features are demonstrated by the PUF domain of Pum1 bound to ssRNA shown in Figure 1.2(E).

The dsRNA-binding motif (dsRBM), similar to RRMs and KH domains, contains a β-sheet backed by α-helices, named for its role as the best recognised dsRNA-binding protein. RNA-binding is achieved primarily by structural recognition of RNA helices on one side of the helix, however sequence specific contacts in dsRBMs have been reported [57]. Notably, the RNA is not distorted on binding, but there are many examples of dsRBMs that can accommodate bulges, loops or other structural features [58]. Recognition of dsRNA is illustrated by the dsRBM of Xlrbpa-2 in Figure 1.2(G). The domain interacts with 16-bp of RNA; an α-helix and loop region contact successive parts of the minor groove, with another α-helix contacting the enclosed major groove. The helical conformation of the RNA is recognised by side and main chain mediated hydrogen bonds to 2′-OH groups and bases as well as to the phosphodiester part of the backbone [59].

The picture that emerges from currently available RBP structural information is that there is a large degree of variability in binding interfaces, bonding interactions, domain folds and RNA sequences recognised within individual RBD categories which has made the classification of RNA-binding rules difficult, even for canonical RBDs.

1.1.2.2 Non-canonical RBDs

Non-canonical RBDs generally constitute novel domains or sequences that can bind RNA but are somewhat unorthodox in their features compared with the better-defined classical RBDs. The existence of domains that bind RNA that do not fit the conventional categories of canonical RBDs has been known for some time. Early examples include the spliceosomal Sm and Sm-like domains that oligomerise to form ring-like structures and composite RNA-binding sites capable of binding purine-rich RNA [60], and the iron regulatory protein 1 (IRP1) that acts as an aconitase enzyme when iron is plentiful, but also mediates post-transcriptional regulation of iron responsive genes by contacting a specific stem-loop in target mRNA when iron is scarce [61] (Figure 1.3(E&D)). Another example is the YTH domain that recognises N^6-methyladenoise (m^6A) modified mRNA through accommodation of the chemical modification in a deep, hydrophobic binding pocket [62] (Figure 1.3(A)).

Disordered regions, as opposed to globular domains, have also been implicated in RNA-binding. There have been at least several reports of inter-domain linkers mediating RNA-binding interactions directly [63]. One structurally characterised example is the linker between RRM3 and RRM4 of polypyrimidine tract-binding protein 1 (PTBP1) that contributes to a hydrophobic core connecting the two domains as well as mediating RNA-binding [64] (Figure 1.3(C)). Similarly, low-complexity motifs (composed of repeats of a single or small number of amino acids, compared with the broader use of amino acids that constitute folded domains) can have RNA-binding activity. One prominent example is that of RGG-

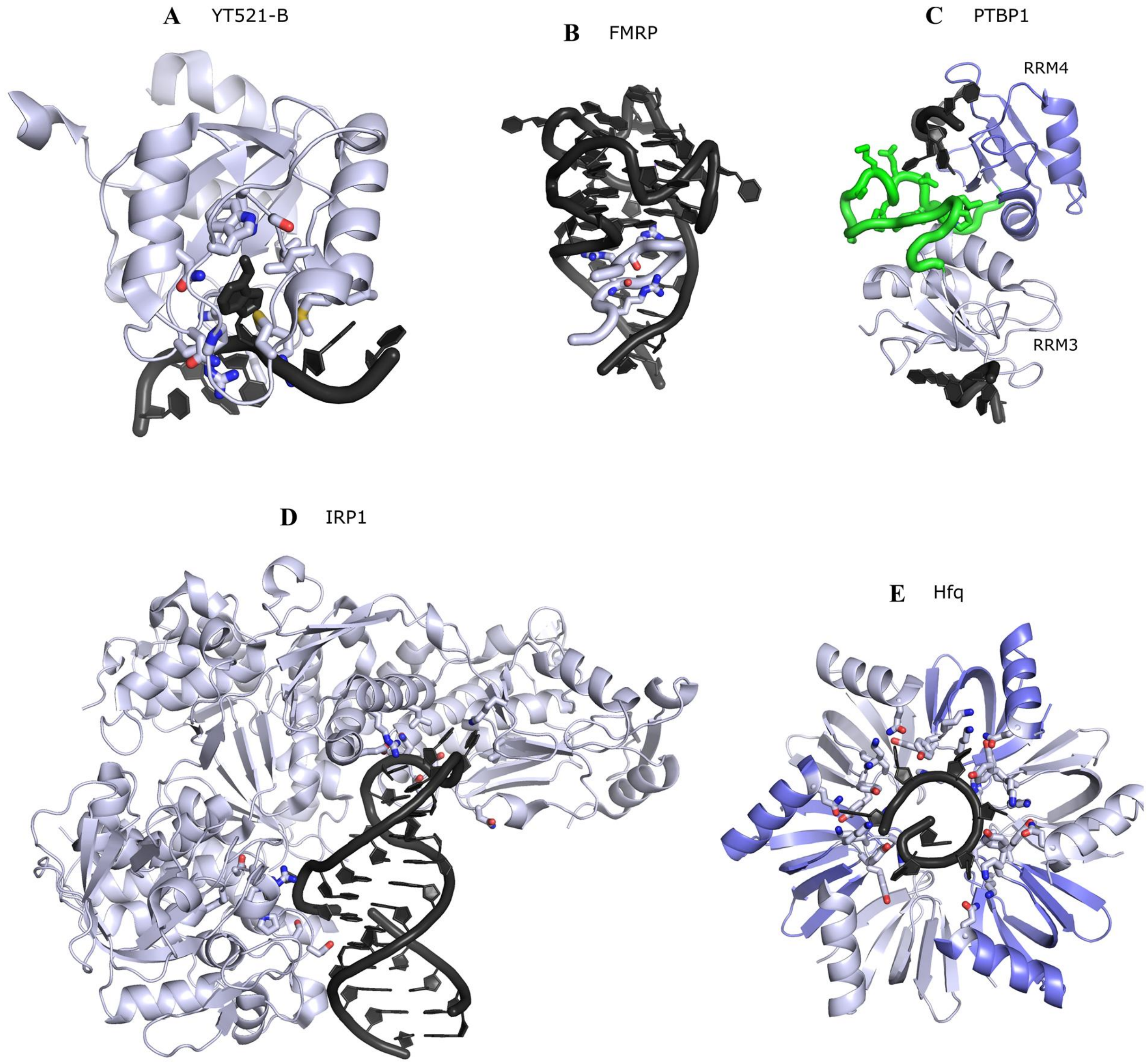


Figure 1.3. Non-canonical RNA-binding domains.
(A) YTH domain of rat YT521-B bound to N^6-methyladenoise RNA (PDB: 2MTV). **(B)** RGG motif of FMRP bound to G-quadruplex RNA (PDB: 5DE5). **(C)** RRM3 (*light blue*), RRM4 (*blue*) and connecting linker (*green*, with selected sidechains involved in RNA-interaction and hydrophobic core indicated in stick representation) of PTBP1 bound to two CU-rich RNAs (PDB: 2ADC). **(D)** IRP1 in complex with ferritin H iron responsive element (PDB: 3SNP). **(E)** Hexamer of Sm-like domains from bacterial protein Hfq bound to A/U-rich RNA (PDB: 1KQ2), alternating proteins indicated by shading). Protein indicated in different shades of *blue* with selected RNA-binding residues indicated in stick representation, and RNA indicated in *black*.

boxes, which consist of repeats of arginines and glycines, with the number of repeats and their spacing varying widely [65]. Few structures have been published of such intrinsically disordered low-complexity motifs bound to RNA; one is the RGG motif of Fragile X Mental Retardation Protein (FMRP). This peptide recognises G-quadruplex RNA through base-specific hydrogen bonds from protein backbone atoms and several arginine sidechains (Figure 1.3(B)) [66].

The prevalence of non-canonical RNA-binding domains has only become clear very recently, facilitated by high-throughput interrogations of the mRNA-bound proteome via cellular crosslinking, oligo(dT) purifications and mass spectrometry. These studies not only confirmed the existence of known non-canonical domains and intrinsically disordered regions (IDRs) in RBPs, but ~10–30% of RBPs identified in these studies were found to harbour neither a known RNA-binding domain (RBD) nor an RNA-related annotation [67-70]. Instead, these proteins have a wide range of roles, including actin binding, phosphorylation, ubiquitination and central carbon metabolism. Surprisingly, a significant number of metabolic enzymes were identified as RNA-binders. In particular, one study in human cells predicted ~1500 human RBPs with ~600 structurally distinct RBDs (many with only a single member) [71]. Taken together, these studies indicate that RBPs are much more diverse than the picture that current structural data paints, and there is clearly much to be learned in this area.

Presently, there are 2250 RNA-protein complex structures deposited in the Protein Data Bank (PDB). Despite this significant amount of data, it is still a small minority of RBPs that have been studied in detail. The rules that govern specificity of RBP:RNA interactions are only understood for a small fraction of RBPs and are proving hard to define, in part due to the huge structural diversity seen in both RNA and RBPs, and the flexibility in their binding interfaces. Understanding the rules that dictate specificity is crucial to our understanding of RNA and RBP function and their role in disease. Reaching of this goal will require in-depth investigation of a wide variety of complexes, particularly newly classified RBPs, which constitute a largely unexplored space.

1.1.3 RNA structure

Like proteins, the structure of RNA is intimately linked to its function. However, RNA is not as chemically diverse as proteins, being made up of only four bases compared to ~20 standard amino acids. As a result, the secondary (base pairing) and tertiary (three dimensional) structures of RNA is used to regulate biological interactions.

RNA rarely consists of sequences that are compatible with long double-helical segments, however short stretches of sequence are often complementary. Therefore, RNA forms much more intricate structures than DNA, folding back on itself to allow base pairing between complementary regions that can be quite distant from each other in the sequence.

One of the most common RNA secondary structural elements is the hairpin loop, consisting of a helical stem section that leads to a terminal loop of unpaired bases as shown in Figure 1.4(A). Both conformational and sequence variation is found in both the terminal loops and base paired helices. The terminal loops of hairpins may be dynamic and sample different conformations. Alternatively, they may form more stable structures such as tetra-loops, consisting of conserved four base motifs that often provide thermodynamic stability and hydrogen bonding potential [72]. Resulting helices can be fully base paired, or exhibit mismatches that manifest as internal loops (Figure 1.4(B)) or bulges (Figure 1.4(C)). Multiloops are formed by the intersection of three or more double helices (Figure 1.4(D)), and pseudoknots consist of interactions between at least two hairpin loops and their stems (Figure 1.4(E)) [73].

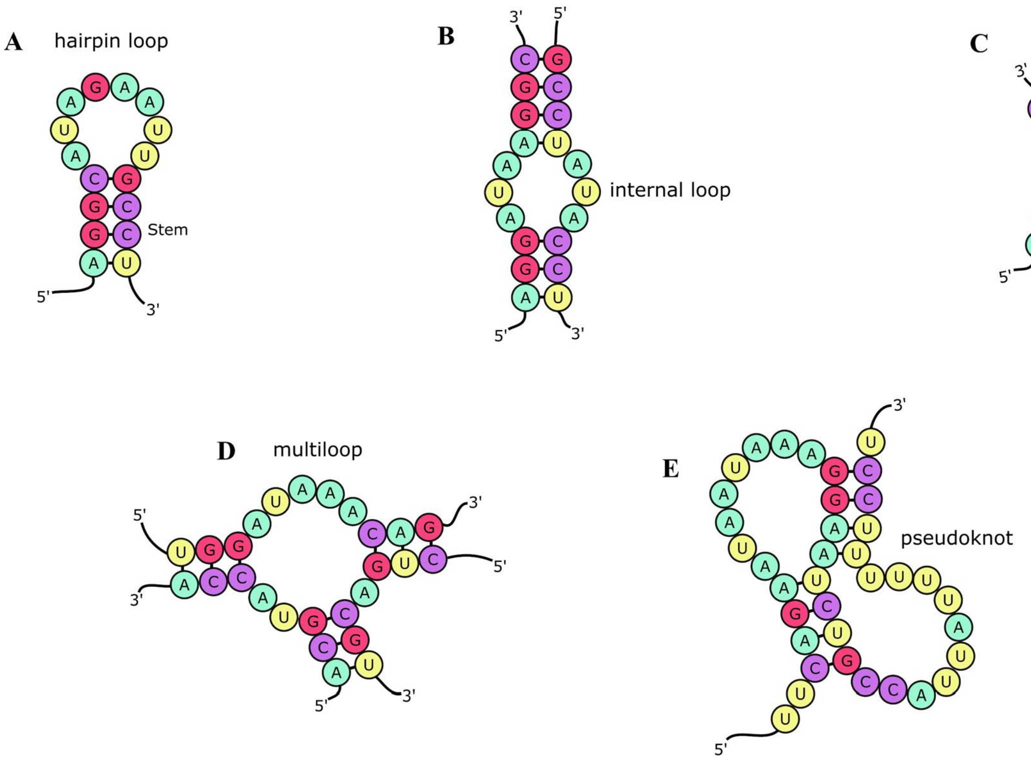


Figure 1.4. RNA structural elements.
(A) Hairpin loops consist of a helical stem section leading to a terminal hairpin loop. **(B)** Internal loops involve unpaired nucleotides with flanking helical segments. **(C)** Bulges consist of unpaired nucleotides on one strand. **(D)** Multiloops are formed when three or more helical segments intersect. **(E)** Pseudoknots contain hydrogen bonding between at least two stem loops.

At the primary sequence level, DNA is as chemically diverse as RNA, however DNA has a propensity to adopt a B-type double helix, and this has been a major factor cementing its role as the genetic information repository [74]. DNA does display some structural variability, also pertinent to its biological functions, forming structures such as cruciforms [75], triplexes [76] and bubbles [77], however, the canonical right-handed double B-type helix overwhelming predominates in the cell. Moreover, the cellular conformation that DNA adopts is usually dictated by the structural constraints of double, triple or quadruple helices. Conversely, RNA is not limited by such restrictions, typically forming intricate three-dimensional structures that allow more diverse functions.

In contrast to the conformational flexibility of ssRNA, double helical RNA (usually A-form) is relatively rigid [78]. The major and minor groove of A-form RNA helices are characterised by a difference in depth to the bases within the helix as opposed to a difference in the width of the helix, as seen in B-form helices. In A-form helices the minor groove is shallow with exposed bases, whilst the major groove is deep with bases buried in the helix (Figure 1.5). In contrast, both the major and minor grooves in DNA B-form helices are deep, but the major groove is wider than the minor groove which allows more access to major groove bases. Additionally, the bases in RNA A-form helices are not perpendicular to the helix axis, unlike DNA B-form helices. These differences create a unique pattern of hydrogen bonding potential for each helix type.

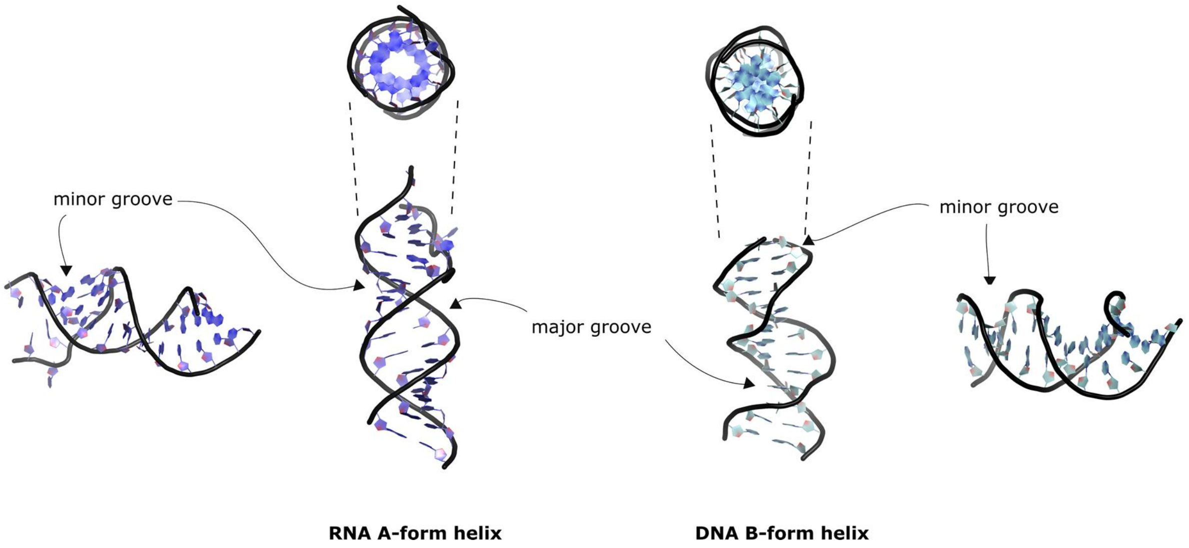


Figure 1.5. DNA and RNA adopt different helical geometries.
RNA canonical A-form helix (left, PDB: 5IEM) and DNA canonical B-form helix (right, PDB: 1ZQ3). Helical RNA is normally A-form, characterised by a wide minor groove with exposed bases. The B-form helix is more tightly packed than the A-form helix, with little space between the bases in the core of the helix. The major groove is wider and shallower than the A-form helix, giving rise to a different pattern of hydrogen bonding potential.

RNA molecules sample a variety of energetically favourable conformations, with the complex three-dimensional structures formed used to regulate target binding. The ability of RNA secondary structure to preclude binding by RBPs has been demonstrated through mutations that disrupted predicted RNA secondary structure, making RBP motifs more accessible, and subsequently these mutated sequences were enriched in pulldowns [79]. Conversely, RBP binding can induce RNA conformational changes [80].

Currently the PDB contains 3556 structures that include RNA, out of 131205 total entries, which is less than 3% of entries. This is a statistic that is out of kilter with biology as there are many more different RNA species in the cell than proteins [81]. Our knowledge of the different types of non-coding RNA in cells is growing considerably, yet our ability to structurally characterise these RNAs is limited at present [82].

The structural versatility of RNA also presents many challenges to the study of RBP:RNA interactions. Considerable effort has been put into the complex task of predicting RNA secondary and tertiary structure in order to assist in experimental design. Programs such as RNAfold [83] and Mfold [84] are well-established RNA-secondary structure prediction programs, and progress is slowly being made in the more complex task of tertiary structure prediction [85].

1.2 RNA-binding properties of the transcription factor Bicoid

Bicoid is a TF that is a master regulator of development in *Drosophila*. It regulates more than ten genes at the transcriptional level, using its homeodomain to bind promoter elements of target genes [86]. Bicoid is also implicated in the post-transcriptional downregulation of one gene, *caudal (cad)*. Interestingly, the homeodomain of bicoid is reported to bind directly and specifically to the *cad* 3′-UTR [87], which to date is the only report of an RNA-binding homeodomain. Explaining how such a small domain can specifically recognise both DNA and RNA is of interest from both a structural and an evolutionary perspective; for instance, how do such dual functions evolve and how common is this behaviour likely to be? Moreover, elucidating the molecular mechanisms of dual function domains will help to unravel the networks of cross-talk that can potentially take place between different regulatory pathways, which is now appreciated to be key to understanding biological specificity in eukaryotes [88].

1.2.1 Transcription factors that bind RNA

There are numerous accounts of TFs that bind RNA, often coupling transcriptional with post-transcriptional gene regulation. Functionally, some examples are well characterised. For example, Smad proteins are cytoplasmic signalling molecules with TF and RBP roles. When phosphorylated, receptor-specific Smads (R-Smads) translocate to the nucleus to regulate target genes [89], including miRNA genes [90, 91]. Interestingly, they are also able to specifically recognise a dsRNA sequence within pri-miRNAs that is similar to its target DNA sequence, and this interaction has been implicated in facilitating processing of the pri-mRNA by the Drosha enzymeto pre-miRNA [92]. Similarly, human glucocorticoid receptor (GR) is a TF that, upon binding the glucocorticoid hormone, moves from the cytoplasm to the nucleus to bind to promoter regions of target genes to drive a range of transcriptional responses. GR also has been shown to bind hundreds of mRNA transcripts [93]; the presence of glucocorticoid effects recruitment of accessory factors to induce decapping and degradation of the bound mRNA [94].

Other examples include (i) NF-90, which both regulates transcription of IL-2 and stabilises the IL-2 mRNA through direct binding [95], (ii) alternatively spliced isoforms of Wilms tumour protein (WT1) that dictate transcriptional or translational regulatory behaviour [96], and (iii) hnRNPs that can bind RNA motifs similar to their promoter DNA targets [97, 98]. TFs have also been shown to bind lncRNAs, for instance SOX2 in complex with lncRNA RMST activates genes involved in neuronal differentiation, whereas SOX2 without RMST activates genes that maintain neural stem cell status [99]. Lastly, a surprising recent study suggested that YY1 promoter occupancy at activated genes is increased through its direct association with RNAs transcribed from promoter and enhancer regions [100]. These examples paint a picture of a considerably intertwined regulatory network.

However, it is interesting to note that the RBP high-throughput studies mentioned in Section 1.1.2 detected only a few percent of the ~1400 TFs encoded in the human genome as having mRNA-binding capacity [67-70]. It may be that RNA-binding behaviour of TFs is relatively rare, or that the experimental conditions in the mRBPome studies were not optimal for detecting TF:RNA interactions.

The only RNA-binding TF for which there is structural data for both DNA and RNA complexes is the ZF protein TFIIIA. In *Xenopus laevis* oocytes, TFIIIA both activates transcription of 5S RNA by binding to an internal control region within the gene, and binds the nascent RNA transcript as it is transported to the cytoplasm, and is therefore sequestered from initiating more transcription of the gene in a negative feedback loop [101].

TFIIIA contains nine classical ZFs. The crystal structure of ZF1-6 bound to 5S DNA indicates that ZFs 1, 2, 3 and 5 insert in to the major groove, dominating the DNA-binding interaction (Figure 1.6(A)) [102]. ZFs 4 and 6 are set back from adjacent minor grooves, with K175 of ZF6 and Q121 and Y135 of ZF4 making contacts with the phosphate backbone. ZF5 makes typical classical ZF interactions with the major groove, with additional interactions provided by L148 to a DNA base, and S150 and K153 contact backbone phosphates. These interactions specify a recognition sequence of NNNAT for the coding strand and GGNNN for the non-coding strand.

In contrast, ZF4-6 constitute the minimal 5S RNA-binding domains [103]; the structure of these fingers bound to 5S RNA is shown in Figure 1.6(B) [104]. ZF4 and the target loop region of 5S RNA are both structurally primed for interaction via sequence specific hydrogen bonds which bear high similarity to typical ZF:DNA interactions. ZF6 binds another loop region of 5S RNA but the induced fit mechanism is more reminiscent of RNA:protein interactions. ZF5 makes no contact with 5S RNA bases, but

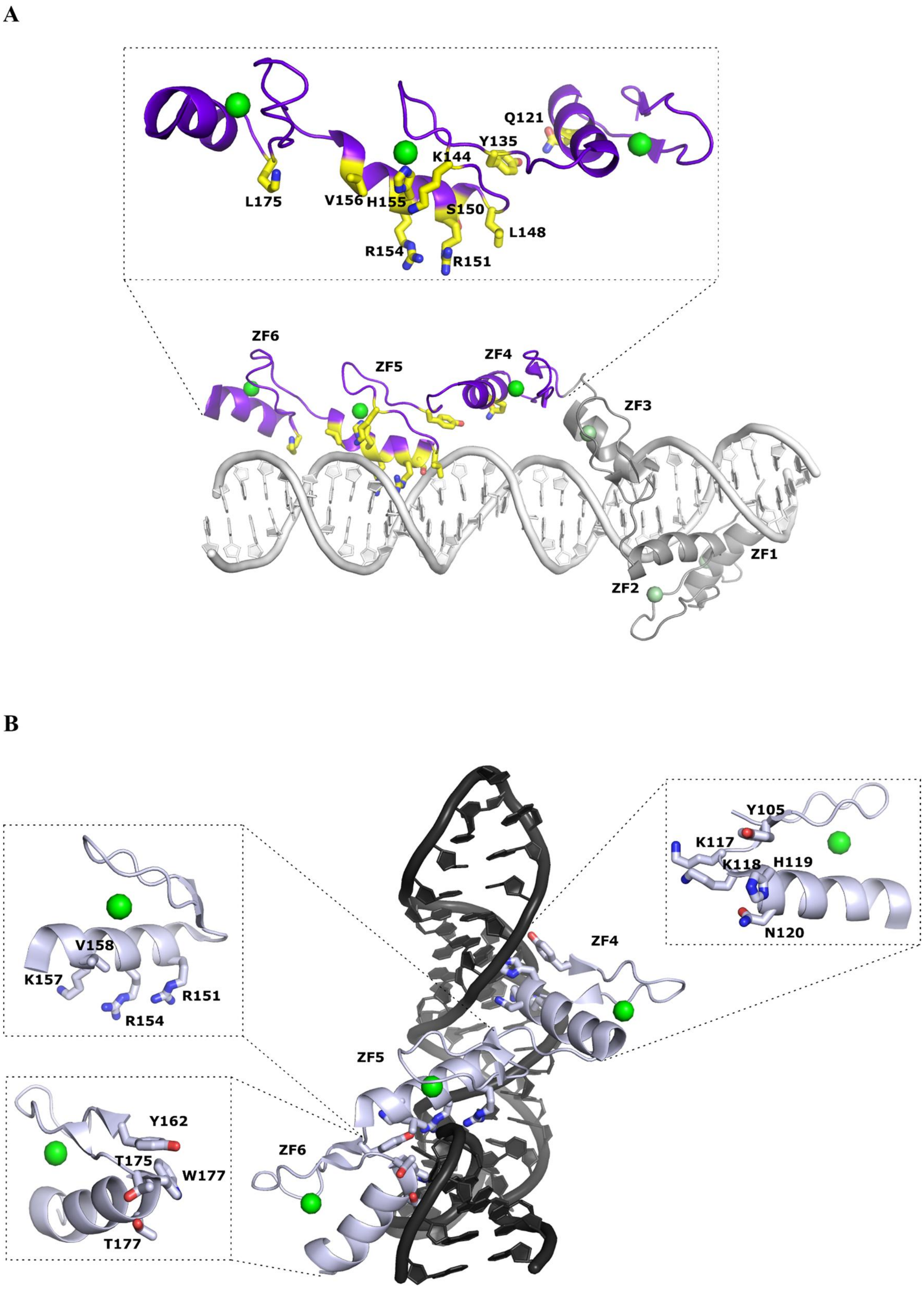


Figure 1.6. ZFs of TFIIIA bind DNA and RNA via different mechanisms.
(A) ZF1-6 of TFIIIA bound to DNA (5S ribosomal RNA gene internal control region) (PDB: 1TF6). ZF1-3 are shown in *grey*, ZF4-6 are highlighted in *purple* for comparison with RNA-bound form in (B). DNA-binding residues from ZF4-6 are labelled and are shown in stick form (*yellow* for carbon, *blue* for nitrogen and *red* for oxygen). Coordinated zinc atoms are shown as *green* spheres. **(B)** ZF4-6 of TFIIIA bound to 5S rRNA (PDB: 2HGH). RNA-binding residues are labelled and are shown in stick form (*light blue* for carbon, *blue* for nitrogen and *red* for oxygen). Coordinated zinc atoms are shown as *green* spheres.

follows the RNA backbone with K157, R151 and R154, and is in position to make electrostatic interactions with sugar and phosphate groups. R151 and R154 appear to be the only common nucleic acid binding residues for both DNA and RNA. In DNA-binding, however, these arginines make base specific major-groove contacts.

Given that there are ~800 classical ZF proteins in the human genome, an outstanding question is how many of these proteins take part in similar dual RNA/DNA-binding activities.

Based on the structural data presented here, it can be concluded that there is considerable plasticity in the ZFs of TFIIIA, recognising DNA and RNA via different molecular mechanisms.

1.2.2 Bicoid – a transcription and translation factor

The *Drosophila melanogaster* protein bicoid is pivotal in establishing the embryonic anterior-posterior axis during early development, by regulating both transcriptional [86] and post-transcriptional gene expression [87, 105] along this axis. *Bicoid* mRNA is anchored to the anterior pole of the oocyte. When translated (around the time of egg fertilisation), bicoid protein is transported through the embryo,

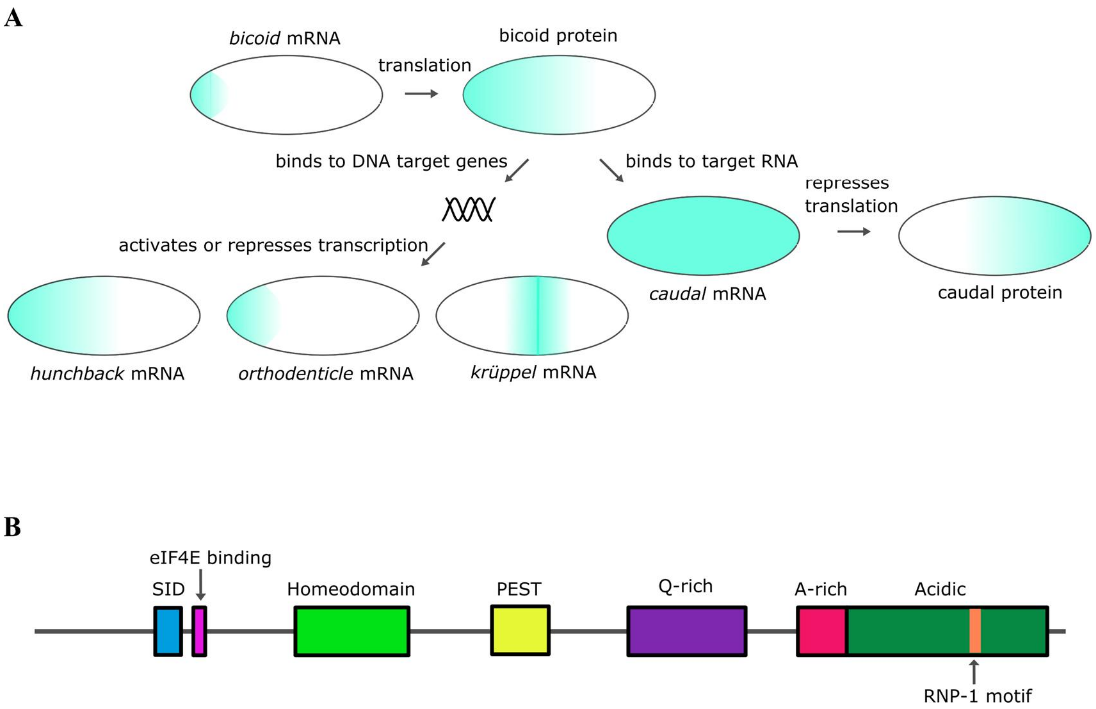


Figure 1.7. Bicoid target gene expression in *Drosophila* embryos.
(A) Schematics of developing *Drosophila* embryos. Localisation of mRNAs and proteins are indicated by shading, with *cyan* indicating high concentration and *white* indicating absence. Examples of DNA target gene products are shown. Embryos are shown anterior to posterior from left to right. **(B)** Known functional domains of bicoid indicated roughly to scale.

forming a concentration gradient that is high at the anterior pole and is detectable to around two thirds of the embryo length [106]. At the DNA level, bicoid directly binds to more than ten genes to activate or repress transcription. The only known RNA target of bicoid is *cad*, an important developmental gene required for tail formation. Bicoid binds to *cad* mRNA and represses its translation, forming a concentration gradient of *cad* inverse to that of bicoid. The developmental gene regulatory function of bicoid is depicted in Figure 1.7(A).

Bicoid is a 54.5 kDa protein with several identified functional domains, represented in Figure 1.7(B). Domains known to be important for transcriptional activation are the acidic region and the Q-rich region [107]. The A-rich region and the self-inhibitory domain (SID) are involved in transcriptional repression: the A-rich region downregulates Q-rich activation [107], whereas the SID recruits the corepressor Sin-3 [108]. The homeodomain is the DNA-interaction domain, introduced in Section 1.2.3.

Domains known to be required for *cad* repression are the homeodomain and PEST domain. The homeodomain is thought to be the RNA-binding domain. PEST domains are known to carry proteolytic signals [109], however this region has also been shown to be required for *cad* repression, but not for transcriptional regulation [110]. The eIF4E binding domain, whilst not required for *cad* repression [111], contributes to repression in splice isoforms that contain this domain by recruiting the eIF4E-related cap-binding protein 4EHP [112].

Bicoid is the product of a gene duplication of the homeobox gene *Hox3* that occurred during the evolution of a particular clade of flies, the Cyclorrhapha (higher dipterans), and as such is only present in these flies [113]. The duplication event allowed *bicoid* to evolve new functionality compared to the ancestral gene, including regulation of *cad* expression (and presumably RNA-binding capacity) and a change in its DNA-binding specificity [114]. Within higher dipterans, the function of bicoid is thought to be largely conserved, despite some co-evolution with TFs [114].

1.2.3 Bicoid homeodomain

The homeodomain is a common DNA, and sometimes protein, interaction domain in eukaryotes, and is well conserved at the sequence and structural level. Over 1600 homeodomains from 32 different organisms have been identified [115], with ~300 in human [116]. The conserved fold of a homeodomain is a 60-amino-acid three helix bundle, with a characteristic helix-turn-helix motif and flexible N-terminal arm (Figure 1.8). On binding DNA, the third ‘recognition’ helix slots in to the major groove, making multiple hydrogen bonds and hydrophobic interactions, and the N-terminal arm becomes more ordered through contacts with bases in the minor groove. Homeodomains typically recognise short adenine and thymine rich sequences (often 5′-TAAT-3′), however sequence discriminating capacity is poor [117], possibly due the high flexibility seen in homeodomain DNA-binding interfaces [118]. HD-containing TFs are involved in a diverse range of processes, including developmental patterning as in the case of *bicoid*, cell-type specification and/or differentiation [119], tumorigenesis [120] and redox regulation [121]. Due to important roles in development and growth, there are almost 200 known instances of human diseases linked to homeodomain proteins, including prostate cancer, autism and diabetes [115].

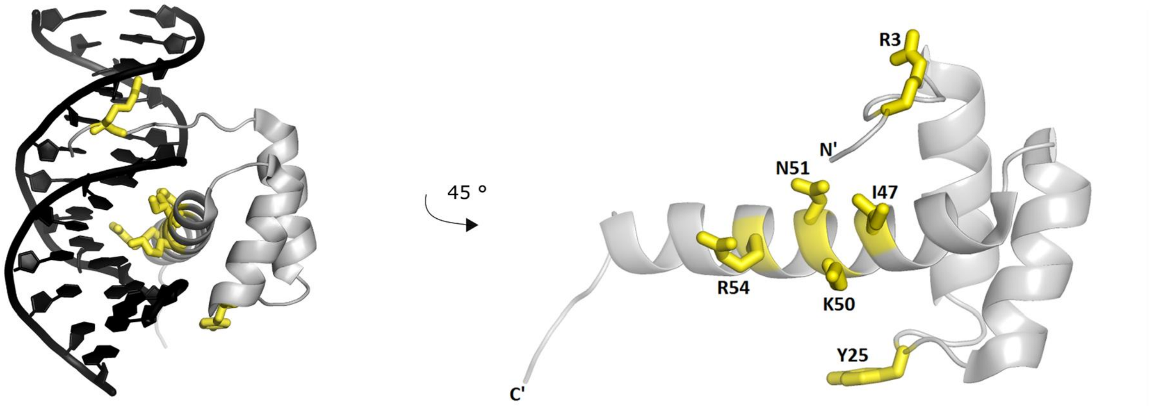


Figure 1.8. Bicoid homeodomain bound to DNA target.
The structure of bicoid homeodomain (*grey*) bound to its DNA target containing consensus site TAATCC (*black*) (PDB: 1ZQ3). DNA-binding residues are shown as *yellow* sticks; the picture on the right is rotated 45 ° along the Y axis with DNA removed and DNA-binding residues labelled.

The structure of the bicoid homeodomain (BHD) bound to target site TAATCC has been solved [122] (Figure 1.8). Whilst displaying the overall topology characteristic of homeodomains, this homeodomain has some unique characteristics. It is the only homeodomain reported to bind RNA, and the only known homeodomain with a K50 and R54 combination. These residue positions are located in the recognition helix and play key roles in distinguishing DNA sequences. The importance of this combination to the distinctive ability of the bicoid homeodomain to bind RNA is demonstrated by a study that showed that K50A mutation abrogates both DNA and RNA target recognition, while R54A just affects RNA target recognition [123].

1.2.4 Evidence that the bicoid homeodomain binds *cad* 3′-UTR directly

Bicoid is involved in silencing *cad* mRNA and this has been well established over decades of research into pattern formation in *Drosophila* development. Maternal effect, segmentation and homeotic genes in *Drosophila* determine axis formation, body plan and the growth of specific body structures. *Bicoid* is an intensely studied maternal effect gene, with many reports implicating bicoid in *cad* silencing; multiple studies have shown *bcd* mutants fail to form a cad gradient [87, 112, 124, 125]. The contribution of the homeodomain to this phenotype was established through mutations to homeodomain residues that disrupted translational repression of *cad* [87, 105]. Early cross-linking [87] and reporter construct [105] experiments indicated that binding of bicoid to the *cad* 3′-UTR was dependent on the homeodomain, although the exact region of *cad* bound by bicoid varied between studies. Subsequently, an electrophoretic mobility shift assay (EMSA) revealed direct and specific binding of GST-tagged bicoid homeodomain to a 343-nt region of the *cad* 3′-UTR, named the bicoid recognition element (BRE) [126] (Figure 1.9(A)).

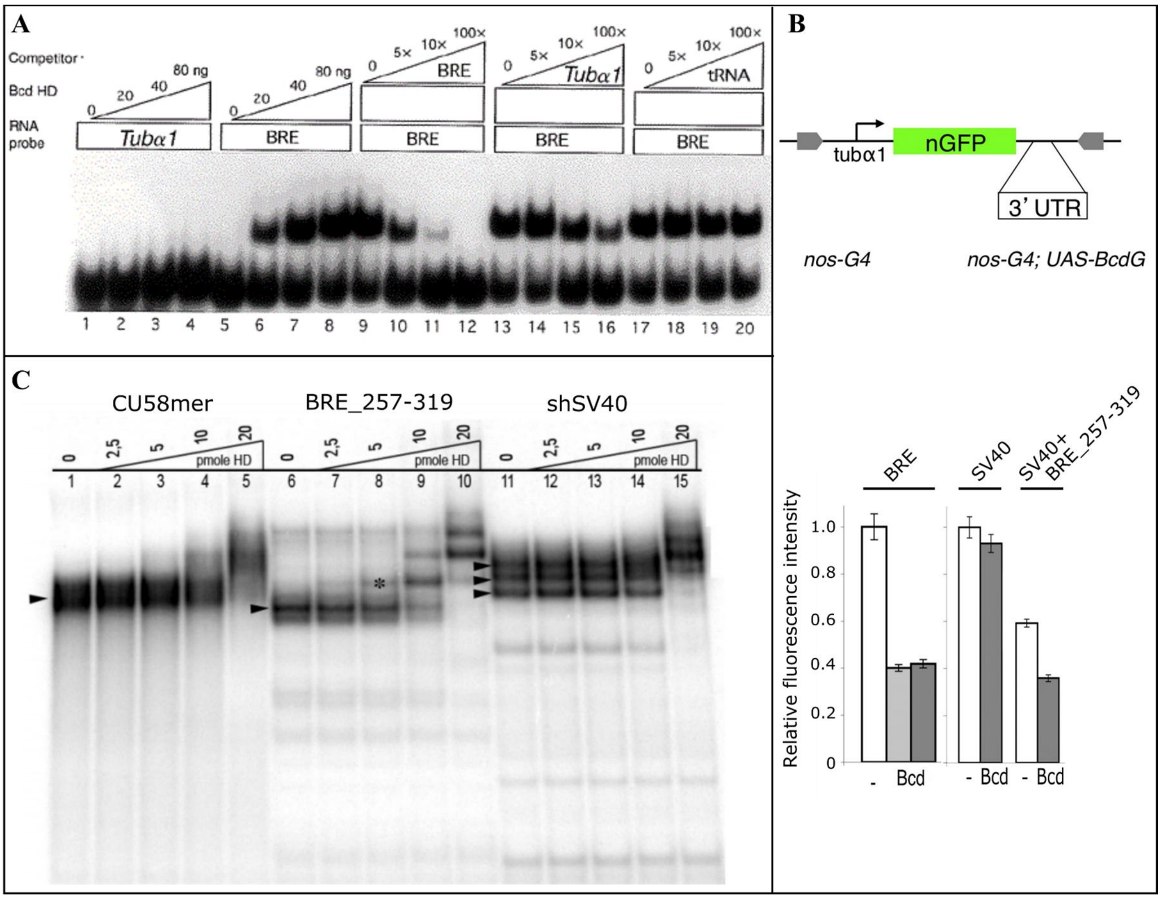


Figure 1.9. The bicoid homeodomain regulates *cad* 3′-UTR at the Bicoid Recognition Element.
(A) Radiolabelled BRE or *Tubα1* 3′-UTRs were incubated with different amounts of GST tagged bicoid homeodomain, with or without cold competitor RNAs as indicated, demonstrating that the bicoid homeodomain binds the BRE but not a fragment of *Tubα1* 3′-UTR. The highest quantity of 80 ng equates to ~ 200 nM GST-tagged bicoid homeodomain. This data is taken from Chan and Struhl (1997). **(B)** BRE and SV40 (control) 3′-UTR sequences were tested for bicoid mediated repression by incorporation of these sequences into the GFP reporter gene as indicated and bicoid was expressed using a nanos-GAL4 system in embryos. The BRE exhibited a drop in relative fluorescence from 1 to 0.4 in the presence of nanos-GAL4 expressed bicoid, whereas SV40+BRE_257-319 exhibited a drop in relative fluorescence from 0.6 to 0.4, indicating that this region partially restores bicoid mediated repression. **(C)** EMSAs indicating that the bicoid homeodomain binds BRE_257-319; the highest quantity of 20 pmole is ~ 2 µM bicoid homeodomain. Binding is also seen to negative control sequences CU58mer and shSV40, albeit weaker. Figures from (B-C) are taken from Rodel *et al.* (2013).

More recently, a region of the *cad* 3′-UTR involved in bicoid-mediated regulation was localised to a 63-nt sequence in the BRE (BRE_257-319); this region contains a hairpin loop that constitutes the most conserved sequence in the *cad* 3′-UTR in drosopholids [111]. This study incorporated GFP reporter constructs at a set location in the *Drosophila* genome, with BRE sequences contained in the 3′-UTR of the sensor (Figure 1.9(B)). Bicoid was expressed under the nanos-GAL4 system. In early fly embryos, the control reporter, consisting of the SV40 early polyadenylation signal, was unresponsive to bicoid. In contrast, the reporter containing BRE demonstrated a reduction in fluorescence with bicoid expression from 1 to 0.4, whereas the SV40+BRE_257-319 reporter demonstrated a reduction from 0.6 to 0.4, indicating that this region of the BRE partially restores bicoid mediated repression. Further, BRE_257-319 is capable of binding the bicoid homeodomain (Figure 1.9(C)). Control sequences consisting of a 58-nt CU sequence (CU58mer), and part of the SV40 3′-UTR (shSV40) were also bound by the bicoid homeodomain, although a higher concentration of protein was required.

Together, these data show that bicoid likely mediates *cad* repression via direct binding of the homeodomain to the BRE, with indications that there may be redundant mechanisms at play within this region.

1.2.5 Bicoid-mediated repression of *cad* translation

The canonical eukaryotic translation initiation process requires binding of the 5′ 7-methyl guanosine (m^7G) cap of an mRNA by an initiation complex containing initiation factor eIF4E; this binding event causes circularisation of mRNA and recruitment of the ribosome. As described in Figure 1.7(B), bicoid contains an eIF4E binding motif that is involved in bicoid-mediated translation silencing of *cad*. An early model of translational repression by bicoid demonstrated that this motif is likely bound by 4EHP, an eIF4E-related cap binding protein. This study demonstrated that 4EHP likely bridges bicoid and the mRNA 5′ m^7G cap as shown in Figure 1.10(A), preventing the translation initiation complex binding the cap. This in turn blocks mRNA circularisation and therefore prevents translation [112].

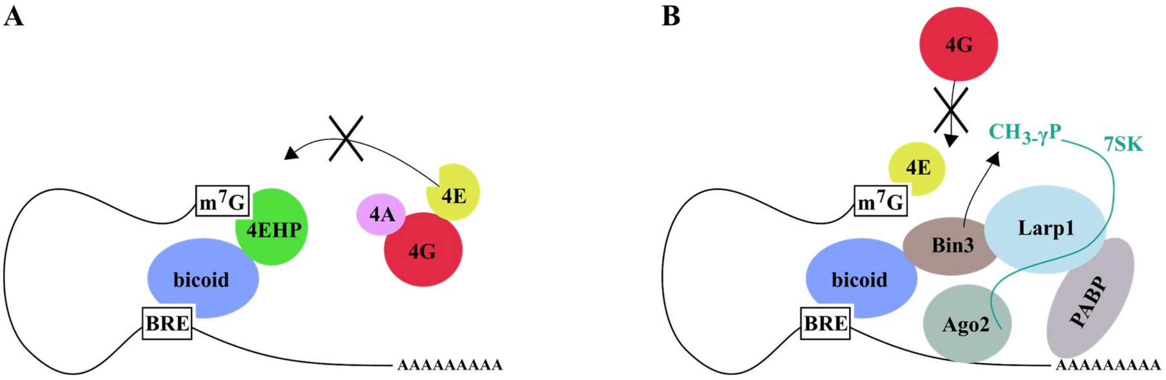


Figure 1.10. Proposed models for bicoid-mediated cad translational repression.
(A) This model proposed by Cho *et al.* (2005) involves 4EHP binding both the 5′ cap of *cad* mRNA and bicoid, preventing the eIF4E initiation complex from binding. **(B)** A more recent model involving bicoid mediated recruitment of Bin3 which methylates 7SK RNA and subsequently effects formation of a repressive complex that includes Ago2, Larp1 and PABP. Abbreviations: 4E: eIF4E, 4G: eIF4G, 4A: eIF4A.

It has since been shown that bicoid isoforms lacking the 4EHP/eIF4E-binding domain are still capable of repressing *cad*, which indicates that there are multiple repression mechanisms utilised by bicoid [111]. The involvement of miRNAs was postulated by the same authors, after noticing that the BRE contains a putative *miR-2* family binding site within the highly conserved hairpin loop sequence of BRE_257-319. Mutations to this region in the BRE that disrupted miRNA binding but preserved the predicted secondary structure of *cad* mRNA stopped bicoid-mediated repression of c*ad* 3′-UTR sensor constructs [111]. Moreover, in the same study, compensatory mutations to the miRNA (that re-established complementarity with the mutated BRE) partially restored bicoid-mediated repression, suggesting that miRNAs play a role in bicoid-mediated repression of *cad*.

A recent report proposes a more complex mechanism of inhibition, involving bicoid-interacting protein 3 (Bin3), and 7SK RNA. Bin3 was initially discovered to interact with bicoid via a yeast-two hybrid screen [127], and has been suggested to be a probable RNA-methyltransferase based on homology to the human protein BCDIN3 which methylates the 5′ γ-phosphate of 7SK RNA [128]. 7SK RNA is a snRNA that is primarily known for its transcriptional role of inhibiting the positive transcription elongation factor (P-TEFb) [129]. Bin3 was found to bind and stabilise 7SK RNA, and both were present in bicoid-immunoprecipitated complexes [130]. Further, in the same study, Bin3 was shown to be required for *cad* translational repression, and genetic interactions were found between *bin3* and the genes for the following translation regulatory proteins: Argonaute-2 (Ago2), poly-A binding protein (PABP), Larp1 and eIF4E. Overall, the data from this study are consistent with a model whereby Bin3 is recruited to *cad* mRNA by bicoid; subsequently Bin3 methylates and stabilises7SK RNA, which serves as a scaffold for the formation of a repressive complex involving the RNA-binding proteins Ago2, poly-A binding protein and Larp (Figure 1.10(B)). Translational repression of *cad* is then thought to occur through preventing binding of essential translation factors such as eIF4G to eIF4E.

At this stage it seems likely that the 4EHP and Bin3 repressive complexes act in a redundant fashion, but further work is needed to clarify the relationships of these pathways. The genetic interaction found between Bin3 and Ago2 may provide a possible mechanism for the miRNA mediated repression proposed discussed above [111].

The current view is that bicoid mediates *cad* repression via direct and specific binding of the homeodomain to regulatory elements within the *cad* 3′-UTR, and that there are redundant mechanisms at play. However, binding motifs within the *cad* 3′-UTR have not been determined, and therefore very little is known about the molecular details of this interaction.

1.3 RNA-binding properties of the transcription elongation factors Spt4&5

Spt4 and Spt5 (Spt4/5) are accessory factors in gene transcription. During transcriptional initiation, RNA polymerases (RNAPs), the enzymes that catalyse transcription, form the transcription initiation complex (TIC) along with general transcription factors (TFIIA, B, D, E, F, H), and this complex identifies and binds promoter DNA and begins transcription [131]. Following the initiation of transcription, the general transcription factors dissociate from the elongating RNAP, and a variety of elongation factors (e.g. Spt4/5, Spt6, FACT, Swi/Snf, TFIIS, Elf1, and Paf1 and TREX complexes) are recruited to form the transcription elongation complex (TEC) along with the double stranded DNA template and the nascent RNA [132]. As well as carrying out transcription elongation, this complex is responsible for initiating co-transcriptional events such as chromatin remodelling [133] and pre-mRNA processing [134] via the recruitment of accessory factors.

The catalytic cores of RNAPs are highly conserved across evolution but their accessory factors that bind during the elongation process are much more divergent. Intriguingly, Spt5 is the only accessory factor that displays the same level of conservation as RNAPs. Further, we know that the Spt5 gene is essential for cell viability in yeast [135]. Whilst Spt4 is non-essential in yeast [136], yeast cells that lack Spt4 demonstrate markedly reduced RNAPII transcription elongation processivity [137]. In *Drosophila* cells, however, loss of either Spt4 or Spt5 is lethal [138]. Taken together, it appears that Spt4 and Spt5 carry out unique and fundamental functions. The details of these functions are still only partially understood, however.

1.3.1 Eukaryotic Spt4 and Spt5 dimerise to carry out transcription related processes

Eukaryotic Spt5 (Figure 1.11) is a ~62 kDa protein that contains a NusG N-terminal domain (NGN, named after the bacterial homolog, NusG) and five Kyprides, Ouzounis, Woese (KOW) domains, flanked by disordered regions on both sides; an N-terminal acidic region and a C-terminal repeat (CTR) region. The NGN domain along with the first KOW domain constitutes the core of Spt5 that is conserved across phyla. Spt4 is an ~11.2 kDa zinc finger containing protein that, in Eukaryotes and Archaea (named RPoE″ in Archaea), dimerises with the NGN domain of Spt5, however bacteria lack an Spt4 homolog. The domain arrangements for Spt4/5 in Eukaryotes, Archaea and Bacteria are shown in Figure 1.11(A).

The NGN domain of Spt5 has a fold that resembles RRM (see Section 1.1.2.1), displaying a βαββαβα topology. The structure of yeast $\mathrm{Spt5_{NGN}}$ in complex with Spt4 demonstrates that these proteins align via their β-sheets, with E338 of Spt5 making an acid-dipole interaction with the helix dipole of α-helix 4 of Spt4 (Figure 1.11(B)) [139].

The conserved core of the heterodimer has been implicated in transcription elongation and RNAP processivity [140, 141], that is, the addition of nucleotides to the nascent transcript as the TEC progresses through the template strand. The NGN domain is thought to mediate associations with RNAPs in all phyla [142]. This interaction has been structurally characterised in Archaea; these data show that Spt5 bridges the RNAP central nucleic acid cleft, which probably enhances the processivity of the transcription elongation complex (TEC) by stabilising it [143, 144]. No structures of eukaryotic Spt4/5 with RNAPs had been published prior to the time of writing, but a model based on both

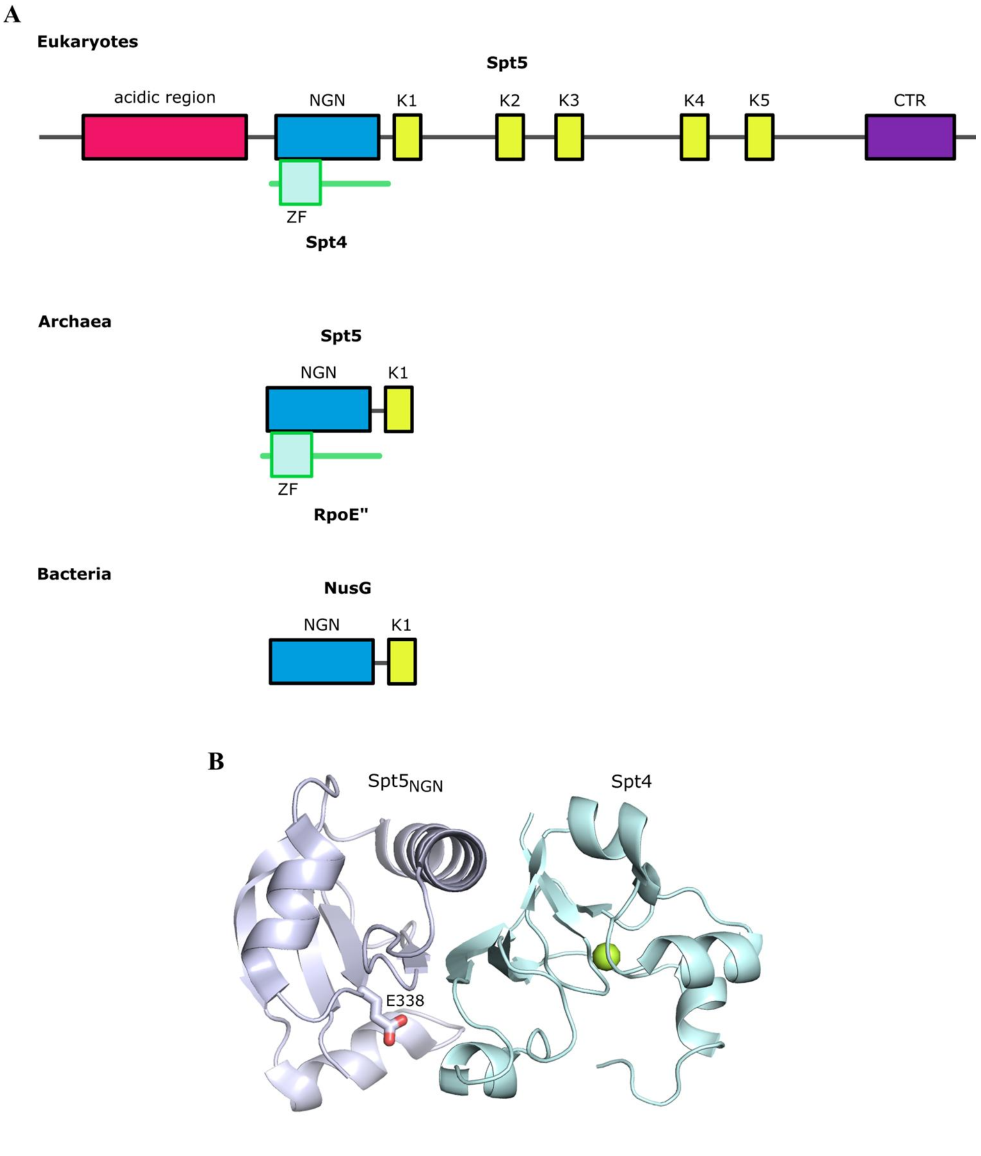


Figure 1.11. Spt4 and Spt5 form a heterodimer.
(A) Schematics illustrating the domain arrangement of Spt4/5 (and homologous proteins) in Eukaryotes, Archaea and Bacteria. Eukaryotes: known functional domains of Spt5 indicated roughly to scale with the full length representing the 62 kDa protein; the 11.2 kDa Spt4 is shown bound to the NGN domain of Spt5. Archaea: Spt5 consists of just the NGN domain and first KOW domain, and the Spt4 homolog RpoE″ dimerises with the NGN domain of Spt5. Bacteria: The bacterial homolog of Spt5 is NusG, consisting of the NGN domain and KOW1. Abbreviations: NGN, NusG; K1-5, KOW domains 1 to 5; CTR, C-terminal repeat region; ZF, zinc finger. **(B)** Structure of yeast Spt4 (*light cyan* with *yellow* Zn^{2+} sphere) and $Spt5_{NGN}$ (*light blue*) (PDB: 2EXU). E338 of $Spt5_{NGN}$, shown in stick form, makes an acid dipole interaction with the adjacent Spt4 α-helix.

crosslinking-coupled mass spectrometry and negative staining electron microscopy indicates that Spt4 and the NGN domain of Spt5 (Spt4/5$_{NGN}$) binds in a similar position to the Archaeal proteins [145].

A sizeable body of work indicates that eukaryotic Spt5 makes extensive use of its large, multi-domain assembly to associate with many accessory partners and execute transcription related processes (reviewed in [146]). The five KOW domains and connecting linkers are certainly important to biological function; the deletion of two or more KOW domains is lethal in yeast and results in reduced affinity of Spt4/5 for RNAPs *in vitro* [142]. Specifically, crosslinking data have demonstrated that KOW domains 4 and 5 contact RNAPII [147]. Moreover, interactions between the KOW domains and connecting linkers with nucleic acid (discussed in the next Section) are also thought to be crucial to the role of Spt4/5 in transcriptional processing. Additionally, the disordered terminal regions provide extra interaction potential. The CTR, for example, is essential for the ability of Spt4/5 to modulate transcriptional processing [148]. This region is phosphorylated to stimulate transcription elongation [149], effecting the recruitment of a variety of accessory factors [134, 150]. No role has yet been ascribed to the N-terminal acidic region.

1.3.2 Nucleic acid binding of Spt4/5

As well as interacting with RNAPs and accessory factors, Spt4/5 is known to interact with nucleic acids as part of its role in transcription. There are multiple reports that suggest Spt5$_{NGN}$ can interact with the non-template DNA strand [151, 152]; this interaction is thought to be required, not for association of Spt4/5 with the TEC, but rather for stopping TEC arrest. The nascent RNA transcript has been shown to be required for optimal binding of Spt4/5 to the TEC [151, 153]. Consistent with this, Spt5 has been shown to bind directly to a 35S rRNA oligonucleotide which constitutes an RNAPI transcript [142]. The first KOW domain in conjunction with the adjacent C-terminal linker has also been shown to be able to bind a variety of DNA and RNA species [154], and there have been two recent reports of KOW5 crosslinking to the nascent RNA transcript [142, 155]. The exact nature of nucleic acid binding, however, and its relevance to the biological role of Spt4/5 remains to be clarified.

With the ultimate aim of better understanding Spt4/5 function at the biochemical level, our collaborators demonstrated that the NGN and KOW domains of Spt5, along with Spt4, preferentially bind RNA over DNA; specifically, single-stranded RNA containing an AA-repeat motif [156]. SELEX (Systematic Evolution of Ligands by Exponential enrichment – a technique whereby a protein is incubated with a random library of RNA oligonucleotides and the tightest binding sequences are identified) was performed on Spt4/5$_{1K}$, Spt4/5$_{2K}$ and Spt4/5$_{5K}$ (domain arrangements are illustrated in Figure 1.12(A)) proteins with a random 24-nt library. The motif with the most significant increase in abundance in all cases was the 14-nt sequence AANAANAANAANAA, where N is any nucleotide.

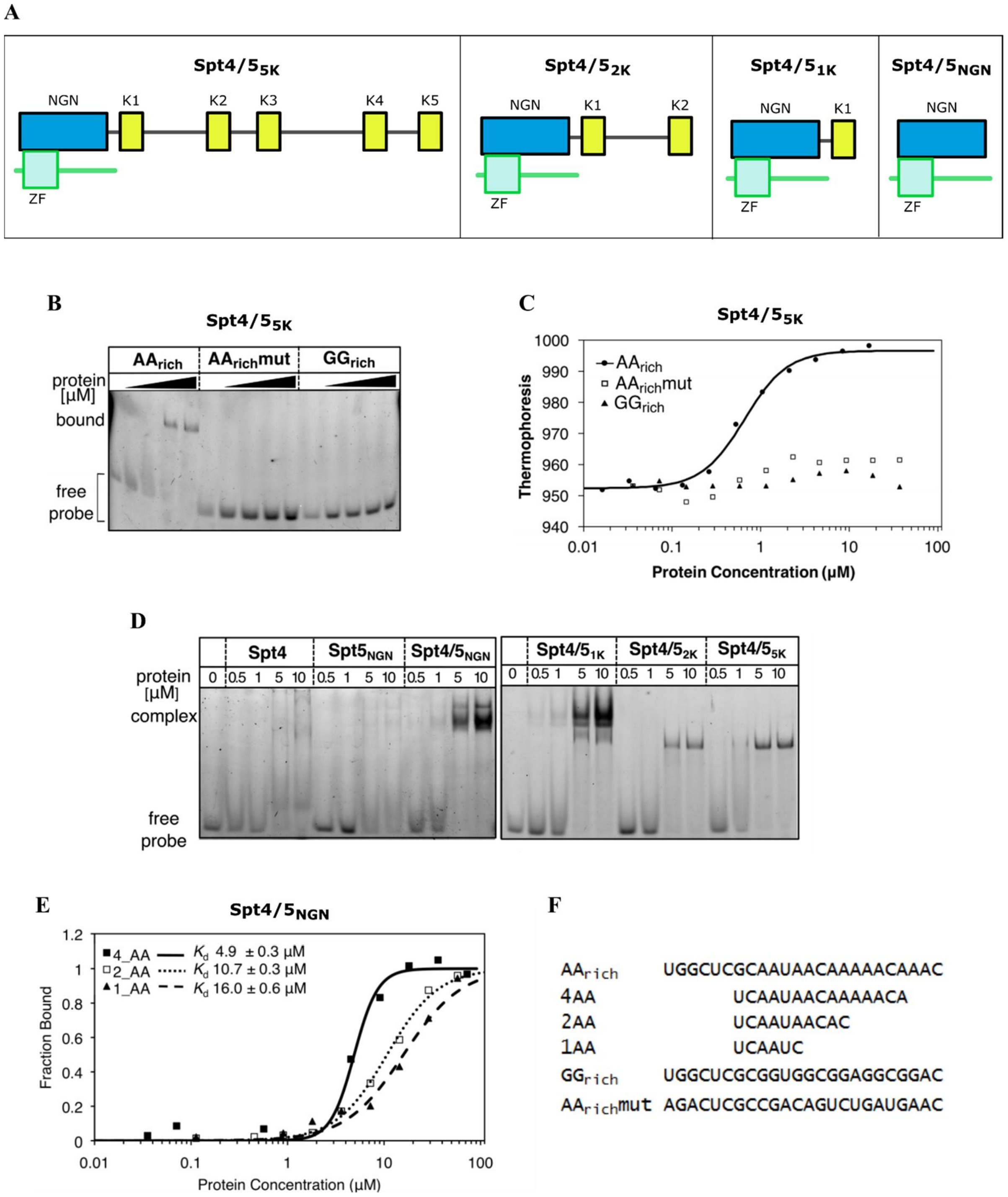


Figure 1.12. Spt4/5 binds AA repeat RNA.

(A) Schematic illustrating domain arrangement for constructs used in this work. **(B)** EMSAs demonstrating that Spt4/5$_{5K}$ binds AA$_{rich}$ but not AA$_{rich}$mut and GG$_{rich}$ RNA. Increasing concentrations of Spt4/5$_{5K}$ were incubated with fluorescently labelled RNA then resolved on a 6% polyacrylamide gel. **(C)** MST assays indicating that Spt4/5$_{5K}$ binds AA$_{rich}$ but not AA$_{rich}$mut or GG$_{rich}$ RNA. Symbols indicate the average of three independent measurements, fitted to a simple 1:1 binding isotherm for AA$_{rich}$ yielding a K$_d$ of 0.65 ± 0.2 μM. **(D)** EMSAs demonstrating that Spt4/5$_{NGN}$ is sufficient for RNA-binding. Increasing concentrations of Spt4, Spt5, Spt4/5$_{NGN}$, Spt4/5$_{1K}$, Spt4/5$_{2K}$ and Spt4/5$_{5K}$ were incubated with fluorescently labelled RNA then resolved on a 6% polyacrylamide gel. Data from (B-D) are taken from Blythe *et al.* (2016). **(E)** Unpublished data performed by Amanda Blythe: MST assays of Spt4/5$_{NGN}$ binding to RNA with varying numbers of AA repeats. Increasing concentrations of Spt4/5$_{NGN}$ were incubated with fluorescently labelled RNA then resolved on a 6% polyacrylamide gel. Binding isotherms were fitted using the Hill method. **(F)** Sequences of RNAs tested.

The rate of enrichment was higher for proteins containing a larger number of KOW domains; significant enrichment was seen for Spt4/5_{5K} in two rounds, Spt4/5_{2K} in four rounds and Spt4/5_{1K} in seven rounds. Specificity of binding to a 24-nt sequence with five AA repeats (AA_{rich}, one of the sequences selected from SELEX) was confirmed by EMSA and Microscale Thermophoresis (MST), whereas no binding was observed to the control oligonucleotides GG_{rich} (the same sequence in which the five AA repeats were mutated to GG repeats) and AA_{rich}mut (five AA repeats mutated to an assortment of non-AA nucleotides) (Figure 1.12(B-C)). Further, Spt4/5_{NGN} was shown to be the minimal RNA-binding region; this polypeptide bound AA_{rich} with an affinity comparable to that of Spt4/5_{5K} (K_d ~2 µM) (Figure 1.12(D)).

MST assays were also performed on shorter RNA oligonucleotides containing one, two and four AA repeats (Figure 1.12(E)). These data indicate that Spt4/5_{NGN} is able to bind RNA with as little as one AA-repeat with a K_d of 16 µM, with each extra AA repeat approximately halving the dissociation constant.

Sequence specific RNA-binding by Spt4/5 is an intriguing possibility; however neither the biological relevance of this interaction nor the molecular basis for it have been established. The demonstrated AA-repeat RNA-binding of Spt4/5 requires further characterisation, to ultimately instruct experiments aimed at determining how this interaction contributes to the transcriptional processing roles of the Spt4/5 heterodimer.

1.4 Aims of this study

The aims of this Thesis are to characterise the RNA-binding properties of what can be considered two non-canonical RNA-binding protein domains, with the ultimate goal of illuminating the molecular underpinnings of the poorly understood RNA-binding behaviour of multifunctional protein domains. Chapters 2 through 4 report on the RNA-binding capacity of bicoid. Firstly, Chapter 2 focuses on determining the RNA-binding specificity of the bicoid homeodomain and *cad* 3′-UTR interaction, which is followed by an examination of the molecular details of this interaction in Chapter 3. Finally, Chapter 4 addresses the features of the full length bicoid protein with a view to illuminating how biological specificity might be achieved. Chapter 5 moves on to investigating and characterising the AA-repeat RNA-binding by Spt4/5_{NGN}.

Chapter 2: Investigation of the RNA-binding properties of the transcription factor Bicoid

2.1 Introduction

As discussed in Chapter 1, the bicoid homeodomain has been reported to bind to the bicoid recognition element (BRE) within the 3′-UTR of *cad* mRNA; no specific binding motif within this RNA has however been identified [111, 126]. The primary aim of this Chapter is to identify the sequence or structural element(s) within the BRE that are required for homeodomain binding.

2.2 Techniques used in this Chapter

2.2.1 Electrophoretic mobility shift assay

The electrophoretic mobility shift assay (EMSA) is a commonly used technique to detect interactions between proteins and nucleic acids. Nucleic acid probes are typically labelled either with a fluorophore or with ^{32}P. A small, constant amount of nucleic acid probe is incubated with increasing amounts of protein, creating a titration series. Samples are equilibrated and then run on a non-denaturing (in order to maintain structures of macromolecules and preserve interactions) polyacrylamide gel. An interaction between the two species is typically indicated by an upward shift of the labelled nucleic acid probe due to the increased size of the complex compared to the unbound nucleic acid probe.

RNA EMSAs are complicated by the structural versatility and electronegative properties of RNA. To help ensure RNA sequences are correctly folded, RNA samples are typically snap cooled before being incubated with protein, and gels are often run at 4 °C in the presence of magnesium to help preserve secondary structure. To assist in reducing non-specific binding, a competitive binder such as heparin is usually included. Heparin is negatively charged and so can compete with RNA for binding to positively charged proteins.

All EMSAs were repeated two to three times in each case; gels shown in this Thesis are representative.

2.2.2 Microscale thermophoresis

Microscale thermophoresis is a recently developed technique that exploits the thermophoretic movement of molecules to quantify biomolecular interactions [157]. When molecules (and other particles) are subjected to a temperature gradient, they move according to a thermophoretic force. The properties of the molecule and of the solution will dictate whether thermophoresis is positive (movement from hot to cold areas of the solution) or negative (movement from cold to hot).

At thermophoretic equilibrium, thermodiffusion is offset by mass balance effects. The ratio of the concentration of molecules in the hot region (C_{HOT}) to the cold region (C_{COLD}) is dictated by difference in temperature (ΔT) and the Soret coefficient (S_T), which is defined as the ratio of the thermal diffusion coefficient to the normal diffusion coefficient:

$$\frac{c_{HOT}}{c_{COLD}} = e^{-S_T \Delta T}$$

$$\text{where } S_T = \frac{\underset{\text{size}}{\boxed{A}}}{kT}\left\{\underset{\text{hydration shell}}{\boxed{-\Delta s_{hyd}}}(T) + \frac{\underset{\text{charge}}{\boxed{\beta\sigma_{eff}^{2}}}}{4\varepsilon\varepsilon_0 T} \times \lambda_{DH}\right\}$$

Equation 2.1

As can be seen in Equation 2.1, the thermophoresis of molecules yields information about their size, charge and hydration shell, which are all potentially perturbed by a binding event. The instrument manufactured by Nanotemper requires that one binding partner be fluorescent (either labelled or intrinsic) in order to quantify thermophoretic movement in the bound and unbound state. A serial dilution of the unlabelled binding partner is executed with the labelled binding partner at a constant, low (typically tens of nanomolar) concentration. Each titration point is loaded into a capillary, and a heat gradient is established by an infrared laser (IR) which heats each sample locally by two to six Kelvin. Fluorescence is detected at the same point in the capillary that was heated by the IR laser. Fluorescence at or near equilibrium is normalised to initial fluorescence to give F_{norm} (Figure 2.1(A)) and affinity is quantified by fitting F_{norm} to a binding model as a function of concentration of the binding partner (Figure 2.1(B)).

Conditions generally need to be optimised to achieve smooth fluorescence curves with good signal to noise ratios. Pertinent parameters include (i) IR power, which establishes the magnitude of the heat gradient, (ii) LED power, which dictates fluorescence excitation, (iii) the buffer composition and (iv) the types of capillaries used, which have varying properties to prevent adsorption by different types of molecules. The sticking of samples to the capillaries is indicated by shoulders in fluorescence peaks before the heat gradient is established, and aggregation can be detected by fluctuations in fluorescence curves due to aggregates moving back to hotter areas by convection flow. The MST curves presented in Figure 2.1 are representative of high quality data.

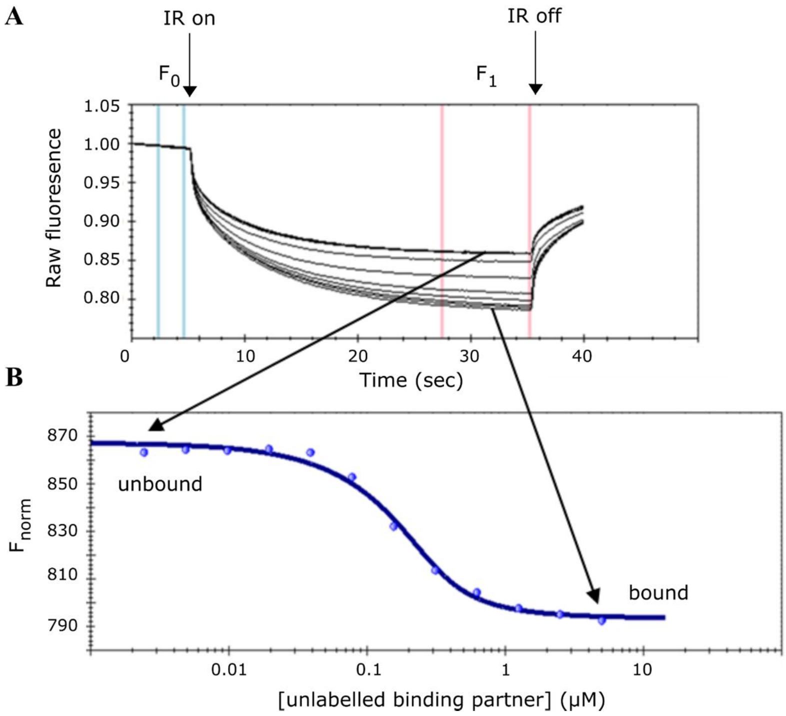


Figure 2.1. Good quality MST data.
(A) Raw fluorescence data from MST. The initial fluorescence F_0 drops when the IR laser is turned on (known as the temperature jump); the magnitude of the drop reflects the temperature sensitivity of the fluorophore. This is followed by diffusion limited thermophoresis, before a steady state of thermophoresis is reached in which diffusion due to thermophoresis is offset by mass balance effects. When the IR laser is turned off, there is an inverse temperature jump and then back diffusion occurs, driven by pure mass diffusion. **(B)** Normalised fluorescence curves. The affinity of an interaction is quantified by normalising the fluorescence after thermodiffusion (area indicated between two red lines) to the initial fluorescence before the heat gradient establishment (area indicated between two blue lines) $\left(F_{norm} = \frac{F_1}{F_0} \right)$ and analysing the change in normalised fluorescence as a function of the concentration of titrated binding partner.

2.2.3 NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is founded in the propensity of magnetic nuclei (isotopes that contain an odd number of protons and/or neutrons) to absorb electromagnetic radiation at discrete and characteristic frequencies when they are subjected to a magnetic field. Valuable structural information about molecules is obtained, primarily because radiation absorbed by each nuclei is dependent on their local chemical environment.

The number of dimensions in NMR experiments reflects the complexity of the information being sought. One dimensional (1D) experiments involving biological molecules such as proteins are convenient as ^{1}H is naturally abundant in biological molecules and so no isotopic labelling procedure is required. Further, ^{1}H spectra are quick to acquire due to the high sensitivity of ^{1}H to the NMR phenomena (about 400 times more sensitive than ^{13}C, one of the other most commonly examined nuclei).

A 1D ^{1}H spectrum allows one to readily assess the folded state of a protein. A well-ordered protein is characterised by sharp and well-dispersed peaks, indicating that individual protons are experiencing somewhat unique chemical environments (a characteristic of a folded protein). With increasing protein size, signal overlap and increases in linewidth reduce the utility of the ^{1}H NMR spectrum somewhat, although it can typically still be used to assess whether a protein is folded or not.

2.3 The search for RNA-binding specificity in the bicoid homeodomain

2.3.1 Cloning, expression and purification of the bicoid homeodomain

Analysis of the RNA-binding specificity of the bicoid homeodomain first required recombinant production and purification of the protein. The homeodomain was cloned into pGEX-6P using full length bicoid, provided by Dr Michalis Averof in the vector pET18b, as a template. The desired sequence (Figure 2.2) was amplified by PCR using primers incorporating *Bam*HI and *Eco*RI restriction sites.

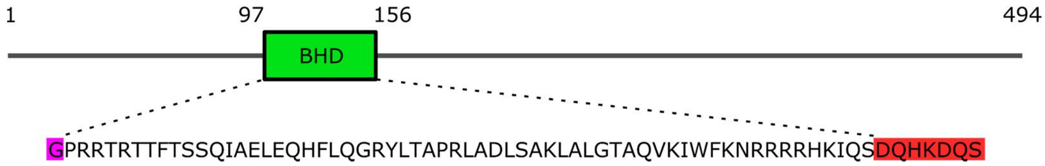

1
97
156
494
BHD
GPRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQS

Figure 2.2. Amino acid sequence of the bicoid homeodomain produced to determine bicoid RNA-binding specificity.
The 60 amino acid homeodomain comprises residues 97 through to 156 of the 494 amino acid bicoid protein. An extra N-terminal glycine residue remains after HRV-3C cleavage of the GST tag (*purple*), and an extra seven C-terminal residues beyond the homeodomain fold were included to increase solubility (*red*). Full length bicoid was provided by Michalis Averof of Institut de Génomique Fonctionnelle de Lyon (IGFL) in pET18b and was used as a template to clone the DNA encoding the above amino acid sequence into pGEX-6P.

This bicoid homeodomain construct, named BHD, would be expressed as a GST fusion, which can be cleaved by HRV-3C to yield a 68 amino acid protein containing the characteristic 60-residue homeodomain with an extra N-terminal glycine remaining from HRV-3C cleavage and seven extra C-terminal residues for solubility [122].

Expression of the recombinant protein in *E. coli* Rosetta(DE3)pLysS cells resulted in good soluble protein yields, which could be purified by GSH affinity chromatography (Figure 2.3). The fusion protein has a theoretical molecular weight of 36.8 kDa, and runs to the expected size on an SDS-PAGE.

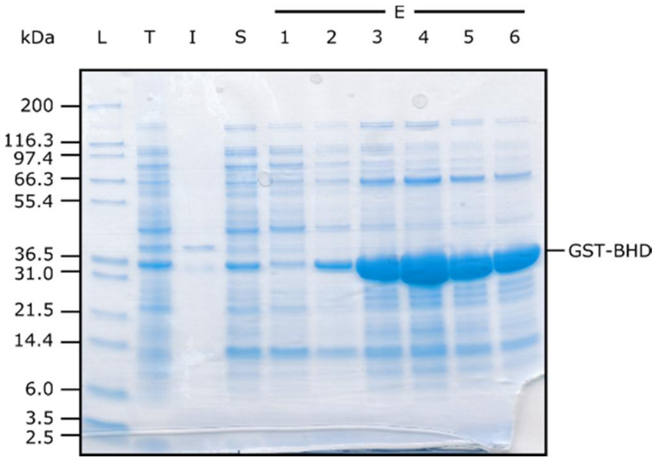


Figure 2.3. Overexpression and affinity purification of GST-BHD.
SDS-PAGE analysis of GST-BHD overexpression and affinity purification. Abbreviations: L, Mark 12 ladder; T, total cell lysate; I, insoluble fraction; S, soluble fraction; E, GSH elutions, numbers as indicated.

The GST tag was cleaved efficiently by HRV-3C protease, and pure BHD was obtained by cation exchange chromatography (Figure 2.4). BHD runs close to its theoretical weight of 7.9 kDa on an SDS-PAGE.

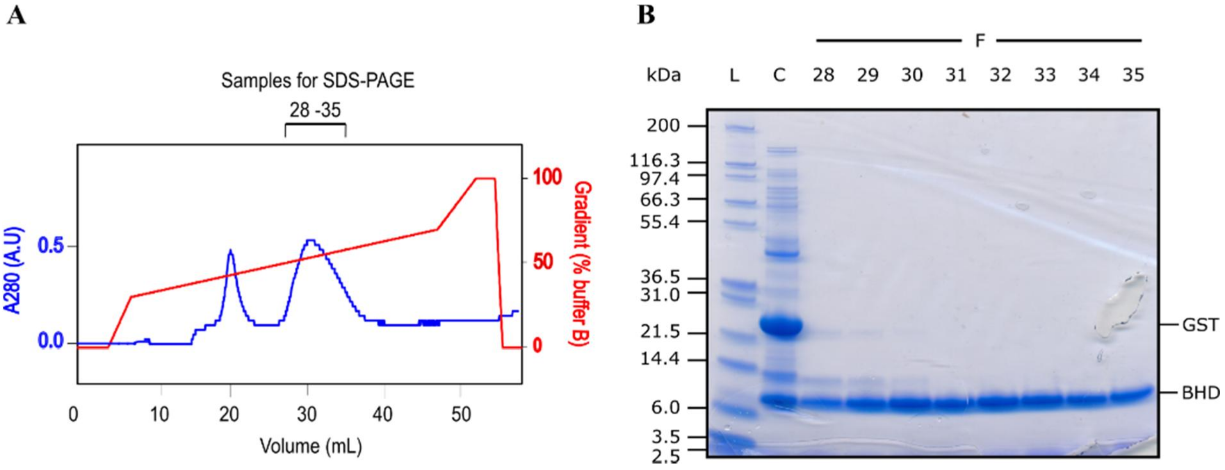


Figure 2.4. Cation exchange chromatography and SDS-PAGE analysis of BHD.
(A) Cation exchange chromatography elution profile showing BHD eluting as the second major peak. Fractions taken for SDS-PAGE are indicated above chromatogram. **(B)** SDS-PAGE analysis of the selected cation exchange chromatography fractions. Abbreviations: L, Mark 12 ladder; C, pooled GST-BHD elutions following HRV-3C cleavage; F, cation exchange chromatography fractions, numbers as indicated.

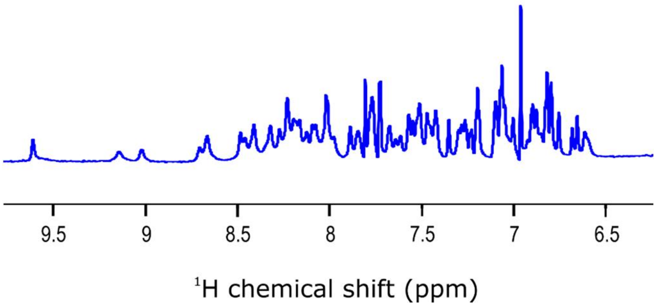


Figure 2.5. Partial 1D ^{1}H NMR spectrum of BHD.
The amide proton region from a 1D ^{1}H NMR spectrum of 150 µM BHD in 20 mM sodium phosphate and 1 mM DTT. Spectrum was recorded at 298 K on a 600 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe.

Fractions 30 to 35 were pooled and concentrated, giving around 1 mg of BHD per litre of culture at >95% purity. The protein was shown to be folded via acquisition of a 1H ^{1}D NMR spectrum, which is shown in Figure 2.5.

2.3.2 Testing bicoid homeodomain binding to *cad* 3′-UTR

In order to confirm the nucleic acid binding capacity of BHD, binding to different nucleic acid probes was tested via EMSA. For DNA-binding validation of the BHD target site, a 12 base pair single-stranded DNA (ssDNA) oligonucleotide containing the known BHD target site TAATC was end labelled with [γ-^{32}P] ATP (sense strand sequence: GCTCTAATCCCG). Binding to this ssDNA oligonucleotide was tested. Separately, this oligonucleotide was annealed with a complementary oligonucleotide to make the dsDNA oligonucleotide. Clear binding of BHD to its dsDNA target site

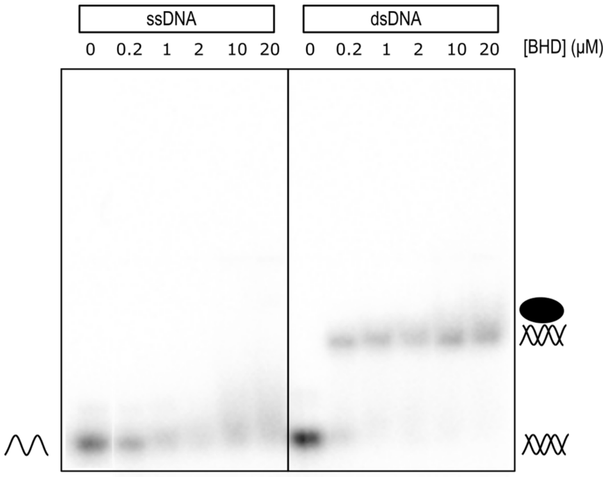


Figure 2.6. DNA target site binding of BHD.
EMSAs confirming that BHD binds to dsDNA, but not ssDNA, containing target site TAATCC. Increasing concentrations of BHD were incubated with ^{32}P-labelled oligonucleotides then resolved on a 6% polyacrylamide native gel.

was confirmed by the observation of a single shifted band, the intensity of which increased as a function of BHD concentration (Figure 2.6). In contrast, negligible binding was seen to the ssDNA oligonucleotide.

In order to confirm binding of BHD to the BRE of *cad*, EMSAs of BHD with ^{32}P-labelled ssRNA were carried out. Templates for transcription were created as oligonucleotides containing T7 promoters with regions complementary to the relevant *cad* sequence; these were amplified from the plasmid pSLfa1180fa (provided by Michalis Averof) which contains the full length *cad* 3′-UTR. Internal labelling of transcripts with ^{32}P was achieved by using ^{32}P UTP in *in-vitro* run off transcription reactions, followed by PAGE purification.

Binding of BHD was tested to two RNA sequences initially: full length *cad* 3′-UTR (855-nt in length) and an 83-nt fragment, BRE(257-319) (see Appendix A1 for RNA sequences tested throughout this Chapter). Nucleotides 257-319 (63-nt) of the *cad* 3′-UTR constitute the highly conserved region identified by Rodel *et al.* [111], however the transcript generated to contain this fragment consists of 83-nt (nucleotides 242-324) in order to conserve the predicted fold. The predicted secondary structures of these sequences (from the software RNAfold [83]) are shown in Figure 2.7(A). BHD bound both probes with comparable apparent affinity, with dissociation constants between 0.15 and 0.7 μM, as seen in Figure 2.7(B). These results are consistent with the published BHD:BRE_257-319 EMSAs detailed

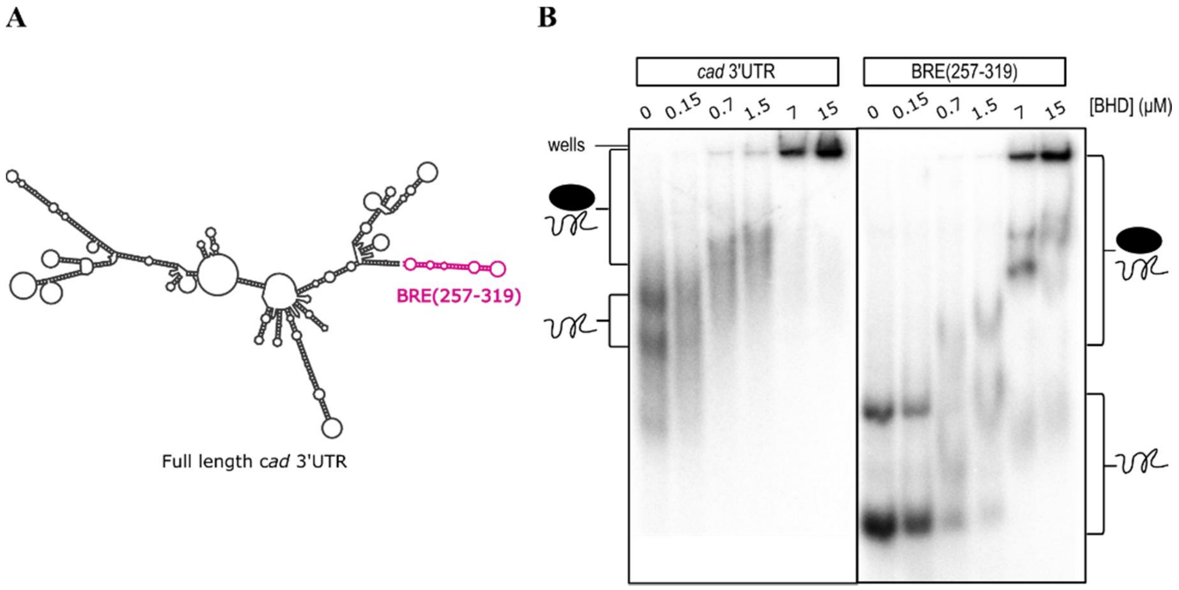


Figure 2.7. Binding of BHD to *cad* 3′-UTR.
(A) Predicted structure of *cad* 3′-UTR, including the BRE(257-319) fragment (*pink*). BRE(257-319) was designed as an 83-nt transcript to conserve the predicted secondary structure. Structures were predicted using RNAfold software. **(B)** EMSAs confirming that BHD binds to both full length *cad* 3′-UTR and BRE(257-319). Increasing concentrations of BHD were incubated with ^{32}P-labelled transcripts then resolved on a 6% native polyacrylamide gel.

in Figure 1.9(C), but not with the tighter association seen between GST-BHD and BRE seen in Figure 1.9(A).

For both free probes, multiple bands were observed, most likely indicating either different conformations or self-association of the transcripts (single bands were excised under denaturing conditions during purification). Multiple shifted bands were also observed, representing either BHD bound to different conformations of the transcripts or a varying number of BHD molecules bound to transcripts, or both. At high protein concentrations (7 μM or more), the RNA transcripts failed to migrate in to the gel, likely due to protein:RNA aggregates that were too large to enter the gel pores.

The affinity of BHD for the full length UTR is around three to four-fold higher than for the 83-nt BRE(257-319) fragment; however the full length UTR is around 10 times longer, and therefore provides around 10 times more protein binding potential. Thus, when accounting for RNA oligonucleotide length, the affinity of BHD for both of the tested oligonucleotides is, overall, quite similar per available binding site. These data indicate that there are no significant sequence elements outside the 83-nt BRE(257-319) that are required for binding.

In order to identify the sequence and/or structural elements of BRE(257-319) that are required for binding, this transcript was split into two overlapping halves in such a manner as to conserve the predicted secondary structure shown in Figure 2.7(A). BRE(257-319) is predicted to form a double stranded structure with multiple loops and abundant base pairing (shown in detail in Figure 2.8(A)). BRE39nt and BRE38nt were designed with five overlapping base pairs, and the 5′ and 3′ overhangs of BRE39nt were removed and the remaining sequence was joined to create a terminal hairpin loop. An unrelated sequence controlling for predicted secondary structure of BRE38nt was also designed and made, named ShapeControl, which is predicted to form base pairs and loops in the same positions but with a different nucleotide sequence.

Again, BHD bound all tested sequences with comparable apparent affinity ($K_d \sim 1$ μM), as seen in Figure 2.8(B). This was an unexpected result given the different sequence composition of the RNAs tested. However, all sequences are predicted to form similar hairpin loop structures, suggesting the possibility that BHD primarily recognises RNA structure rather than sequence. To further assess this possibility, a 32-nt RNA sequence that is predicted to be unstructured (Unst, with the sequence UCGAAGCCCUCUCUCAGUUUGUCAUAUACCCU) was designed and tested for binding to BHD via EMSA (Figure 2.9).

At least two bands appeared for the Unst probe alone; one much fainter than the primary band. This pattern probably indicates the formation of some secondary structure, either within individual probes or between probes. It is very unlikely that this RNA sequence forms a hairpin loop given the base pairing probabilities. Both Unst bands disappear to an equal degree, and the disappearance occurs at a similar protein concentration to BRE(257-319) and BRE38nt, so, assuming that at least one of the Unst bands is unstructured, these data support the idea that BHD is not recognising structural elements of RNA.

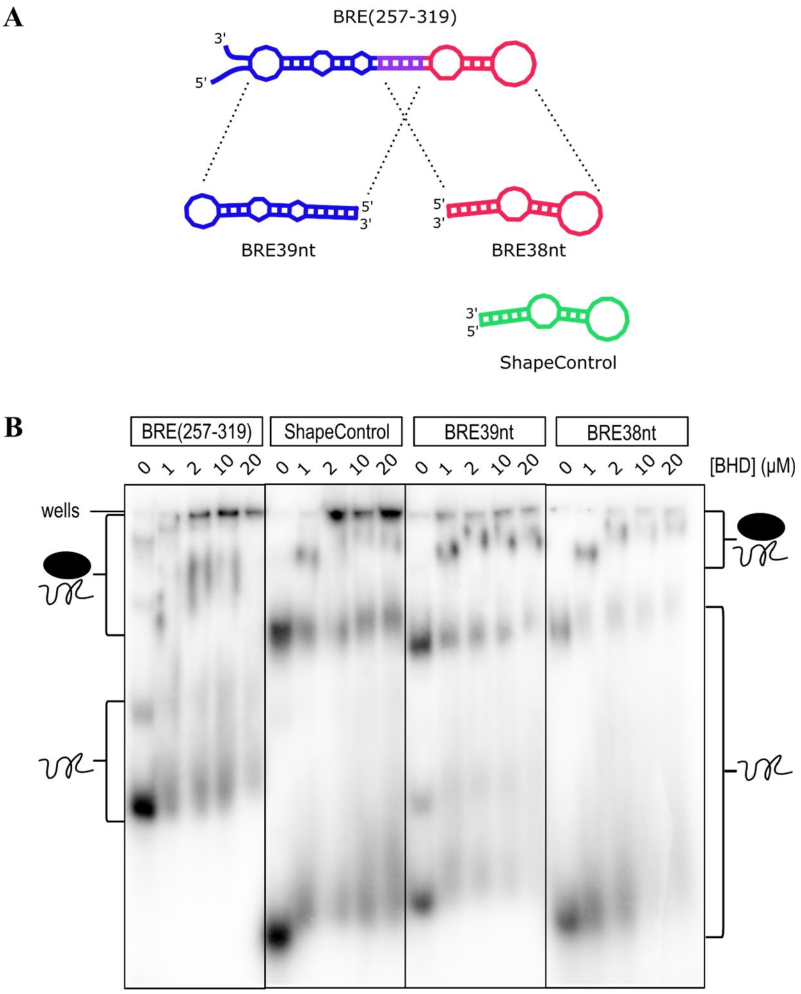


Figure 2.8. Binding of BHD to BRE transcripts.
(A) Design of BRE(257-319) truncated transcripts. Two overlapping halves of BRE(257-319) were designed to conserve the predicted fold (BRE39nt: *blue*, BRE38nt: *red*, overlapping bases: *purple*). ShapeControl (*green*) was designed to display the same secondary structure as BRE38nt in a different sequence context. Structures were predicted using RNAfold software. **(B)** EMSAs demonstrating that BHD binds to BRE(257-319), BRE38nt, BRE39nt and ShapeControl with comparable affinity. Increasing concentrations of BHD were incubated with ^{32}P-labelled transcripts, and then resolved on a 6% polyacrylamide native gel.

Looking at all of the EMSAs presented in this Section collectively, it can be seen that the band pattern of the probes was not consistent across different samples. Multiple bands were often seen for free probes despite single bands being excised under denaturing conditions during PAGE purification.

The multiple bands signify different conformations of the RNA and/or association between RNA molecules. Sometimes shifted bands were seen in samples containing protein; however, this was not consistent between different protein preparations. The inconsistencies observed may be partly due to

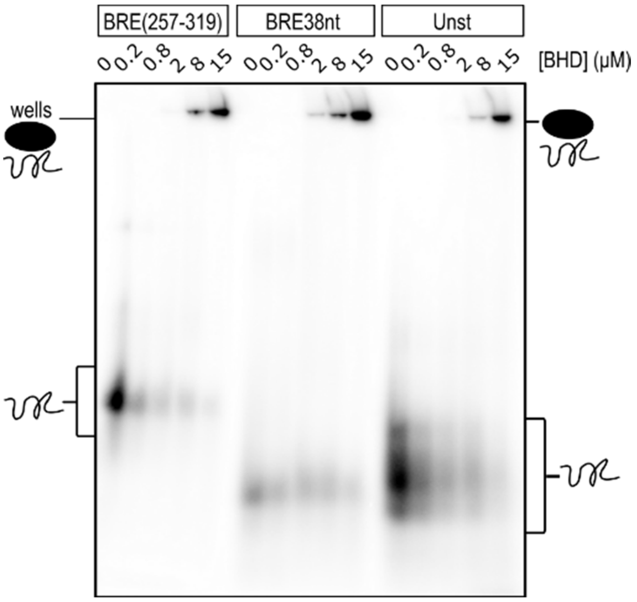


Figure 2.9. Binding of BHD to BRE38nt and Unst, an RNA predicted to be unstructured.
EMSAs assessing the binding of BHD to an RNA sequence that is predicted to be unstructured (right), BRE38nt (middle) and BRE(257-329) (left). The protein-RNA complexes that are formed do not migrate into the gel and appear to form at similar concentrations. Increasing concentrations of BHD were incubated with ^{32}P-labelled transcripts then resolved on a 6% polyacrylamide native gel.

the short half-life of ^{32}P radiation; when the signal is not as strong shifted bands are not visualised. Despite these inconsistencies, the affinity of the interactions remain relatively constant, at least within the limits of what is discernible by radiolabelled EMSAs.

The observation has been made that smearing and a lack of discrete bands in EMSAs for RNP complexes can be associated with higher k_{off} rates [158, 159], which could in turn be a feature of non-specific RNA-binding [160]. Alternatively, smeared bands may signify complexes that are not stable because the protein and/or RNA constructs are not optimal [161]. Taken together, the data presented here indicate either that BHD may require extra elements in order to achieve RNA-binding specificity or that the EMSAs are reporting on non-specific (and perhaps biologically irrelevant) RNA-binding activity.

2.3.3 Investigation of possible miRNA involvement in BHD binding

MicroRNAs are ~22-nt regulatory RNAs that are well known for their involvement in Argonaute (Ago) mediated translational repression and mRNA decay [162]. Primary miRNAs consist of stem-loops that are processed in to mature miRNA strands and then loaded onto the Ago-containing RNA-induced silencing complex (RISC) [163]. The miRNA directs RISC mediated post-transcriptional gene regulation by hybridising with target mRNAs. More than 60% of human protein-coding genes are regulated by miRNAs [164].

Besides this canonical role, new roles of miRNAs continue to be discovered, indicating that there is still more to be learnt about miRNA biology. For example, *miR-122* has been shown to upregulate internal ribosome entry site (IRES) translation via an unknown mechanism [18].

As discussed in Section 1.2.5, miRNAs are likely implicated in bicoid-mediated repression. Given that no specific RNA-binding motif could be located in the *cad* 3′-UTR, EMSAs of BHD and BRE were carried out in the presence of a *mir-2* family member, *mir-308* (as used by Rodel *et al.* [111]), to investigate whether or not miRNAs can contribute to specific BHD binding. As shown in Figure 2.10(A), *miR-308* is imperfectly complementary to the distal hairpin loop of BRE38nt. A negative control sequence was designed to conserve predicted fold, named *miR-Fold* (Figure 2.10(B)).

Perhaps surprisingly, no binding of *miR-308* was seen to BRE(257-319), as shown in the EMSA on the left in Figure 2.10(C). Both *miR-308* and BRE(257-319) are predicted to form hairpin loop structures, and therefore would need to overcome self-association in order to bind each other. Accordingly, RNA samples were incubated together and heated to 70 °C in order to overcome existing secondary structure but this did not facilitate hybridisation of the two RNA species. However, when a constant amount of BHD was included in an EMSA in which BRE(257-319) was titrated with *miR-308*, a distinct difference in migration of the BRE(257-319) probe was observed (Figure 2.10(C), right). The free probe band disappeared and was replaced with a more slowly migrating band (labelled with "?", lower, in Figure 2.10(C)). At higher *miR-308* concentrations, this band became more smeared and an additional band was observed high on the gel (labelled "?", upper, Figure 2.10(C)).

Gel shifts with two RNA species in each sample are rare in the literature but several have been reported; one study showed 50% binding of a ~300-nt RNA to a ~100-nt RNA at the top concentration of 500 nM. When the 11-kDa protein Hfq was added in at a concentration of 1 µM, this resulted in 100% binding at the much lower RNA concentration of 20 nM [165].

When the binding assays were reversed to comprise a constant amount of miRNA and increasing concentrations of BHD, *miR-308* caused the appearance of discrete shifted bands ("?" in Figure 2.10(D)). In order to test for specificity, addition of *mir-Fold* to a BHD-BRE(257-319) EMSA also altered the observed pattern of shifted bands, although the bands were more smeared than those observed with *miR-308*.

If a three-way complex is *not* formed between BHD, BRE and *miR-308*, then it might be expected that BHD would bind separately to both *miR-308* (or *miR-Fold*) as well as to BRE(257-319), given its promiscuity observed in the current work. The miRNAs are at a much higher concentration than BRE(257-319) in these binding assays, so it is reasonable to expect a decrease in the amount of BHD binding to BRE(257-319) as *miR-308* and BRE compete for BHD. A decrease in apparent binding, however, is not observed.

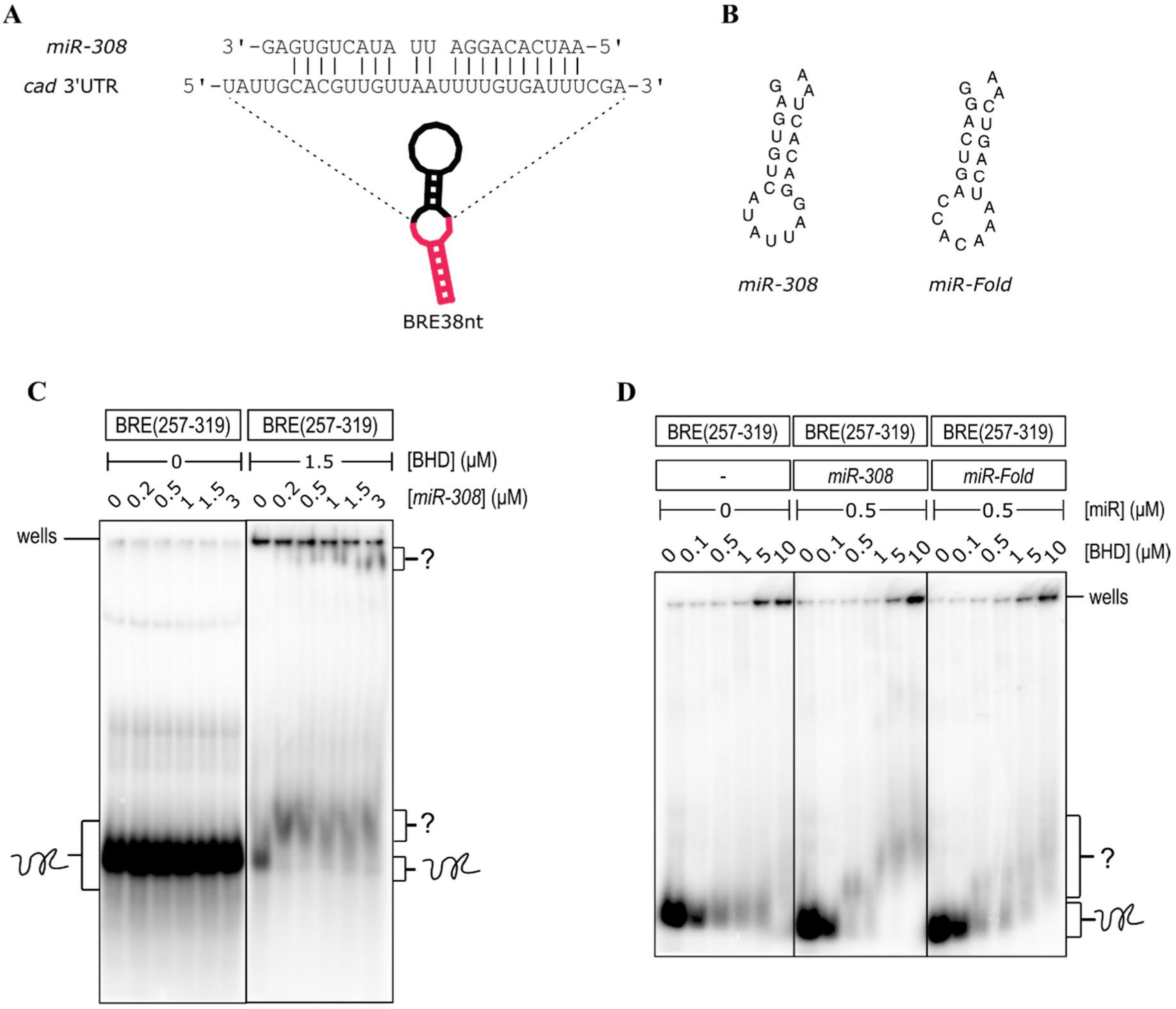


Figure 2.10. BRE contains a putative miRNA binding site.
(A) Sequence alignment of *miR-308* with the predicted miRNA target site contained in BRE. The putative target site is indicated in the predicted structure of BRE38nt in *black*. **(B)** Sequences and predicted folds of miRNA transcripts tested for binding to BHD and BRE. *miR-Fold* was designed to have a different sequence but the same predicted structure as *miR-308*. Structures were predicted using RNAfold software. **(C)** ^{32}P-labelled BRE(257-319) was incubated with an increasing amount of *miR-308*, with and without a constant amount of BHD, right and left respectively, then resolved on a 6% polyacrylamide native gel. **(D)** ^{32}P-labelled BRE(257-319) was incubated with an increasing amount of BHD, without any miRNA (left), with an increasing amount of *miR-308* (middle) and with an increasing amount of *miR-Fold* (right), then resolved on a 6% polyacrylamide native gel.

Alternatively, if *miR-308 does* form a ternary complex with BHD and BRE, then it would be reasonable to expect an increase in affinity when *miR-308* is included in the binding assays. The appearance of discrete bands in EMSAs with miRNAs Figure 2.10(D) does suggest the formation of a complex with a slower off-rate than observed in BRE-BHD EMSAs. Despite the effect being observed for both *miR-308* and *miR-Fold*, it does seem to be slightly more pronounced for *miR-308*.

As far as is known, miRNAs bind their 3′-UTR targets in complex with Ago proteins, as mentioned above. This alliance speeds up the complex job of finding targets within the crowded cellular milieu [166]. Structures of Ago proteins in complex with cognate miRNAs have shown that this association elicits presentation of the highly complementary seed region of miRNAs in an A-form helix, which is optimal for interacting with target mRNA sequences [167, 168]. Moreover, it has been shown that association of Ago proteins with miRNAs changes the base pairing preferences for miRNA-target interactions [169]. Given, however, the results presented here, it seems a possibility that BHD binds to duplex RNA consisting of BRE and miRNA elements, which would represent a novel miRNA 3′-UTR binding mechanism. This would need to be followed up, however it was not pursued further here due to time constraints. Experiments with resolution superior to that of EMSA would be required; NMR-based chemical shift mapping of isotopically labelled BRE, miRNA and BHD might yield helpful data, or x-ray crystallography if crystals of the trimer can be obtained.

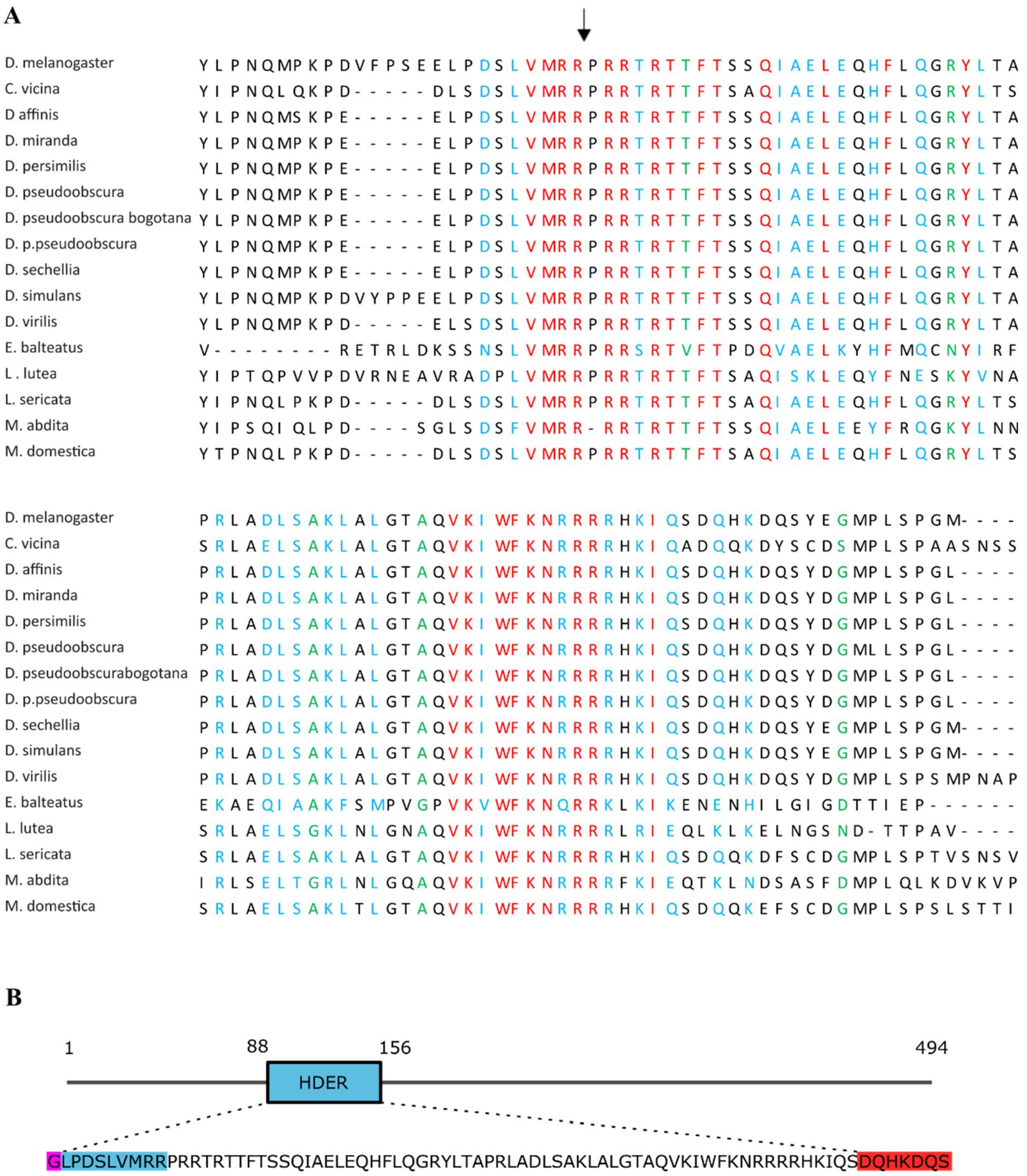
2.3.4 N-terminal extension of homeodomain to include potential arginine-rich motif

As no RNA-binding specificity could be determined for BHD within the canonical boundaries of a homeodomain, and given that smeary bands observed in EMSAs may indicate suboptimal constructs as mentioned in Section 2.3.2, a longer construct was cloned. This construct incorporated two arginine residues immediately N-terminal to the original BHD construct, making a total of five arginines in close proximity in the sequence. It is known that low complexity arginine sequences are overrepresented in the RNA-binding proteome [68]. A common RBD, known as the arginine-rich motif (ARM), is characterised simply by a high local density of arginine residues. A list of ARMs that have been demonstrated to bind RNA is shown in Table 2.1, illustrating the considerable variation in their sequences. ARMs characterised to date tend to bind RNA in a hairpin conformation [170-172].

Table 2.1. List of ARM sequences with characterised RNA targets.

Species	Protein	ARM sequence	RNA target
Human	60S ribosomal protein	ELKIKRLRKKFAQKMLRKARRK	Preference for structured RNA
Human	HEXIM	KKKHRRRPSKKKRHWKPYYKLTWEEKKK	GUAC repeat motif in hairpin loop of 7SK RNA
Phage	P22 N	NAKTRRHERRRKLAIER	GGUGCGCUGACAAAGCGCGCC
Phage	λN	MDAQTRRRERRAEKQAQWKAAN	GGGCCUGAAGAAGGGCCC
Virus	BIV Tat	SGPRPRGTRGKGRRIRR	GGCUCGUGUAGCUCAUUAGCUCCGAGCC
Virus	HIV-1 Tat	SYGRKKRRQRRRPPQ	CCAGAUCUGAGCCUGGGAGCUCUCUGG
Virus	HTLV-1 Rex	MPKTRRRPRRSQRKRP	GCUCAGGUCGAGGTACGCAAGTACCUCCCUUGGAGC

Despite not being required for DNA target recognition, four residues flanking the N-terminal end of the homeodomain are conserved in all organisms that have bicoid (Figure 2.11(A)). In order to determine if these conserved residues and putative ARM contribute to BHD RNA-binding, a longer construct of the bicoid homeodomain was made, named homeodomain with extra arginines (HDER). HDER was cloned to include an extra nine N-terminal residues, as depicted in Figure 2.11(B).



A

```
D. melanogaster              YLPNQMPKPDVFPSEELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
C. vicina                    YIPNQLQKPD-----DLSDSLVMRRPRRTRTTFTSAQIAELEQHFLQGRYLTS
D affinis                    YLPNQMSKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. miranda                   YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. persimilis                YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. pseudoobscura             YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. pseudoobscura bogotana    YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. p.pseudoobscura           YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. sechellia                 YLPNQMPKPE-----ELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. simulans                  YLPNQMPKPDVYPPEELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
D. virilis                   YLPNQMPKPD-----ELSDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTA
E. balteatus                 V--------RETRLDKSSNSLVMRRPRRSRTVFTPDQVAELKYHFMQCNYIRF
L . lutea                    YIPTQPVVPDVRNEAVRADPLVMRRPRRTRTTFTSAQISKLEQYFNESKYVNA
L. sericata                  YIPNQLPKPD-----DLSDSLVMRRPRRTRTTFTSAQIAELEQHFLQGRYLTS
M. abdita                    YIPSQIQLPD----SGLSDSFVMRR-RRTRTTFTSSQIAELEEYFRQGKYLNN
M. domestica                 YTPNQLPKPD-----DLSDSLVMRRPRRTRTTFTSAQIAELEQHFLQGRYLTS

D. melanogaster              PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEGMPLSPGM----
C. vicina                    SRLAELSAKLALGTAQVKIWFKNRRRRHKIQADQQKDYSCDSMPLSPAASNSS
D. affinis                   PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPGL----
D. miranda                   PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPGL----
D. persimilis                PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPGL----
D. pseudoobscura             PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMLLSPGL----
D. pseudoobscurabogotana     PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPGL----
D. p.pseudoobscura           PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPGL----
D. sechellia                 PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEGMPLSPGM----
D. simulans                  PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEGMPLSPGM----
D. virilis                   PRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYDGMPLSPSMPNAP
E. balteatus                 EKAEQIAAKFSMPVGPVKVWFKNQRRKLKIKENENHILGIGDTTIEP------
L. lutea                     SRLAELSGKLNLGNAQVKIWFKNRRRRLRIEQLKLKELNGSND-TTPAV----
L. sericata                  SRLAELSAKLALGTAQVKIWFKNRRRRHKIQSDQQKDFSCDGMPLSPTVSNSV
M. abdita                    IRLSELTGRLNLGQAQVKIWFKNRRRRFKIEQTKLNDSASFDMPLQLKDVKVP
M. domestica                 SRLAELSAKLTLGTAQVKIWFKNRRRRHKIQSDQQKEFSCDGMPLSPSLSTTI
```

B

1
88
156
494
HDER
GLPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQS

Figure 2.11. Amino acid sequence of recombinant BHD extended to include conserved N-terminal residues. **(A)** Sequence alignment of bicoid homeodomain and flanking sequences in all species that have a *bicoid* gene. The start of the canonical homeodomain sequence is indicated by an arrow. *Red* indicates conserved residues, *cyan* shows residues with strongly similar properties and *green* shows residues with with weakly similar properties. Alignment of sequences was done with Clustal Omega. **(B)** Sequence and context of the HDER construct. This 68-residue construct included an extra nine N-terminal residues (*blue*). As per the BHD construct, an extra N-terminal glycine residue (*purple*) remains after HRV-3C cleavage of the GST tag, and an extra seven C-terminal residues (*red*) are included to increase solubility.

Expression and purification of HDER was carried out in a similar fashion to BHD, although protein expression levels were lower for HDER (Figure 2.12), and substantial protein was lost during concentration, resulting in much poorer yields at ~200 µg per litre of culture.

There was no discernible difference between the ability of BHD and HDER to bind to ^{32}P-labelled BRE(257-319) in an EMSA (Figure 2.12(A)). As with the EMSA experiments described above, micromolar protein concentrations resulted in the formation of protein-RNA complexes that did not migrate into the gel. As always, it is difficult to interpret EMSAs that give rise only to complexes that are retained in the wells, other than to conclude that *some* sort of interaction is taking place.

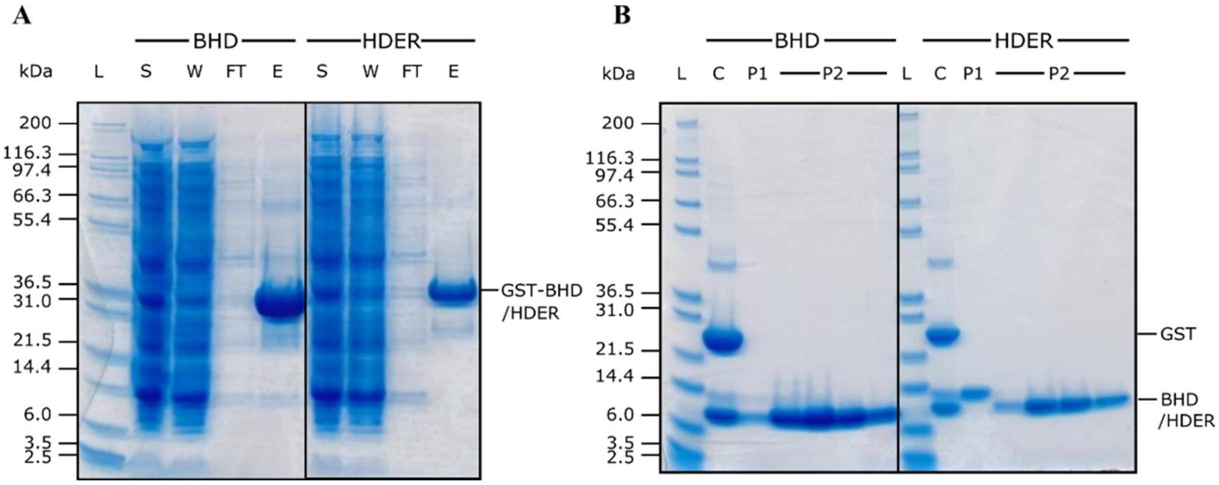


Figure 2.12. Purification of HDER compared with BHD.
(A) SDS-PAGE analysis of GSH affinity purification for BHD and HDER. Abbreviations: L, Mark 12 ladder; S, soluble fraction; W, wash; FT, flow-through; E, GSH elutions 1 to 6 (pooled**). (B)** SDS-PAGE analysis of selected cation exchange chromatography fractions. Abbreviations: L, Mark 12 ladder; C, HRV-3C cleavage of pooled elutions; P1, first peak in cation exchange chromatogram; P2, second peak in cation exchange chromatogram.

As an alternative approach to assessing these interactions, we turned to MST. A shorter RNA was used so that it could be ordered as a fluorescently-tagged oligonucleotide, avoiding lengthy purification and labelling procedures. A 19 base RNA sequence from BRE38nt was chosen (BRE19nt) that encompasses the distal hairpin loop of the BRE (Figure 2.12(B)). Two different salt concentrations were tested in order to gauge the robustness of the interaction.

Good quality, reproducible MST data were obtained. These data were fitted to a simple 1:1 binding model. At either salt concentration, the binding of HDER to BRE19nt was 3–4-fold stronger than the binding of BHD (Figure 2.12(C). Dissociation constants of 1.1 and 3.8 µM for HDER and BHD, respectively (50 mM NaCl), are consistent with typical affinities for single RNA-binding domains for their target. For example, dissociation constants of ~1 µM have been reported for ZFs and KH domains binding to RNA [47, 49]. The higher salt concentration substantially reduced binding affinity for both proteins, by six to eight fold in both cases. Some caution should be given to the binding constants as complete binding curves could not be attained at the higher salt concentration.

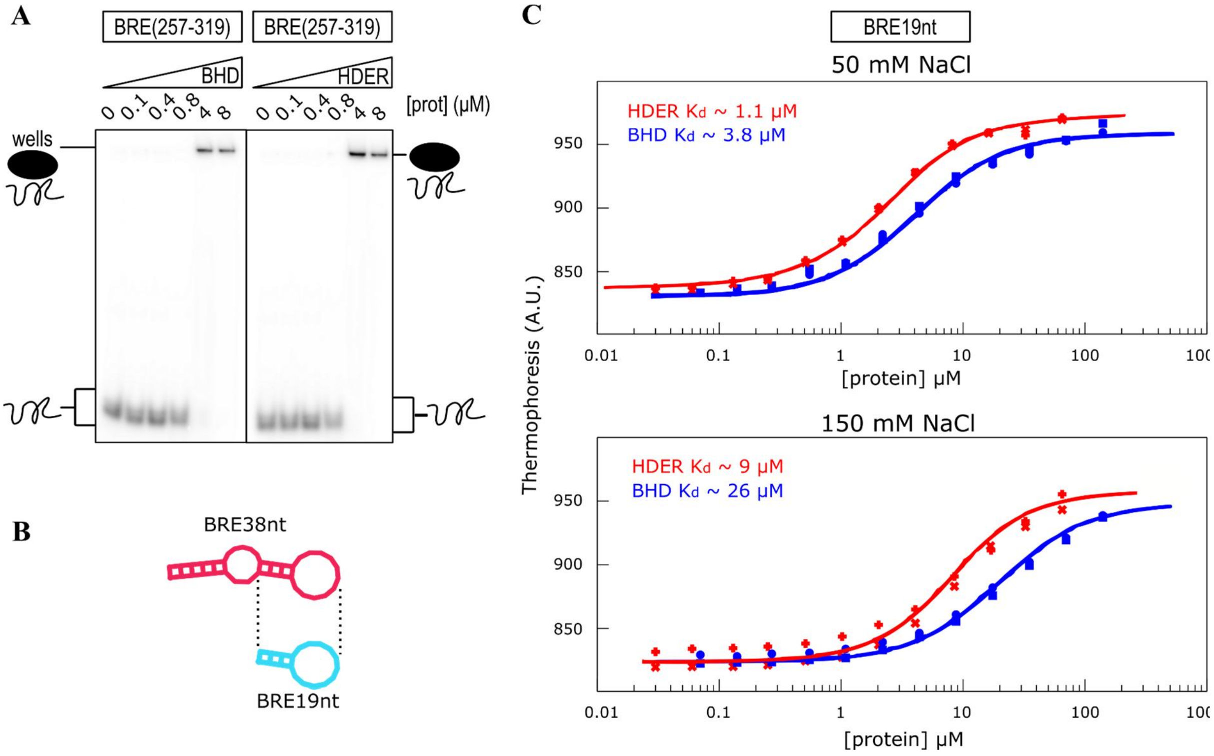


Figure 2.12. Binding of HDER or BHD to BRE.
(A) EMSAs of BHD and HDER binding to BRE(257-319). Increasing concentrations of BHD (left) and HDER (right) were incubated with ^{32}P-labelled transcripts then resolved on a 6% polyacrylamide native gel. **(B)** Design of the RNA sequence for use in MST. BRE19nt (*light blue*) consists of the distal hairpin loop of BRE38nt (*red*). Structures were predicted using RNAfold software. **(C)** MST curves for BHD (*blue*) and HDER (*red*) following titration of these proteins into fluorescently labelled BRE19nt. Titrations were carried out both in 50 mM and 150 mM NaCl. Data points from two independent titrations are shown for each protein and salt combination and fitted to a 1:1 binding isotherm.

The reduction in affinity with increasing salt concentration indicates that this interaction has a significant electrostatic component. This interpretation is somewhat corroborated by the observation that BHD binds unstructured ssRNA (Figure 2.9) but doesn't substantially bind ssDNA (Figure 2.6). Despite DNA having the same overall charge as RNA, the extra hydroxyl group of RNA provides slightly more hydrogen bonding potential. Moreover, the preference for ssRNA over ssDNA is consistent with the interaction being non-sequence-specific, as non-sequence-specific interactions in RBPs usually comprise electrostatic interactions with the sugar phosphate backbone, as opposed to electrostatic interactions and hydrophobic interactions with the bases typically seen for sequence specific interactions [173]. For example, RIG-I binds to the end of dsRNA in a non-sequence specific fashion mostly through contacts with the sugar phosphate backbone, including hydrogen bonds to ribose

2′-OH groups [174]. Specificity for RNA is also provided to some degree in Ago proteins through interactions with ribose 2′-OH groups [175].

Interestingly, even though the highest protein concentration in MST is an order of magnitude higher than that of the EMSAs, no aggregation of protein and RNA was seen, as evidenced by consistent fluorescent signals (+/- 10%) prior to establishment of the heat gradient (Figure 2.13). The protein:RNA complex formed in EMSA must be somehow affected by the conditions of experiment; it is possible that the complex has reduced solubility in the running buffer, for example.

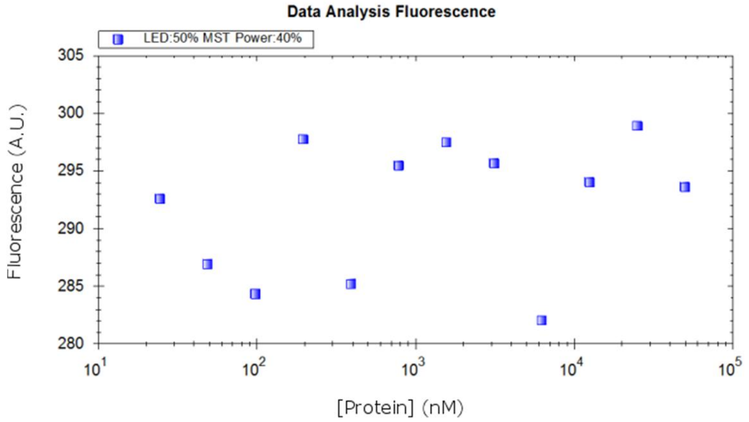


Figure 2.13. Example of MST fluorescence data collected prior to establishment of the temperature gradient. Initial fluorescence data for the BHD:BRE19nt titration, before establishment of the temperature gradient. The absence of a concentration dependant change in fluorescence indicates that there is no aggregation of a protein-RNA complex.

2.4 Discussion

Given that RNA sequence or structural elements of the c*ad* 3′-UTR required for bicoid homeodomain recognition were not able to be defined, two interpretations are possible. The bicoid homeodomain might function *in vivo* as a non-specific (or low-specificity) RBD, or the experimental conditions utilised in this Chapter were not optimal to detect specificity. These scenarios will be discussed below, with reference to current RBP research.

2.5.1 The bicoid homeodomain might be a low-specificity RBD

It is possible that the bicoid homeodomain might be a genuine RBD that binds RNA without much discrimination between sequence and structural elements. Specificity might arise from the action of the full-length protein, or from a binding partner.

Non-specific RBPs have important roles in biology, with recent estimates suggesting that up to 50% of RBPs may be non-specific [81]. Many biological processes necessitate that RBPs interact with a large variety of RNAs. For example, YB-1 downregulates translation initiation broadly through promiscuous binding of mRNAs [176]. Similarly, RNA interference requires promiscuous RNA binding; the PAZ domains of Ago1 and Ago2 bind ssRNA by accommodating the 3′ end in a hydrophobic pocket [177, 178]. Additional non-specific roles of RBPs may be yet to be unearthed. For example, YY1 was recently shown to increase the occupancy of transcription factors at enhancers and promoters by binding to the nascent transcript of the enhancer or promoter [100]. It is an interesting hypothesis that other transcription factors may bind RNA transcripts in a non-specific fashion in order to create positive feedback loops in transcription.

In the case of bicoid, however, it has been shown that bicoid downregulates *cad* translation by forming some sort of inhibitory complex, which necessitates that functional specificity comes from somewhere. How RBPs achieve biological specificity in many cases is still not known [71]. Biological specificity refers to sequence preferences of proteins *in vivo*, contrasted with intrinsic specificity which refers to protein sequence preferences determined *in vitro [81]*.

Despite the early literature report that the bicoid homeodomain binds BRE specifically, as outlined in Chapter 1, we know that RBPs that recognise RNA in a specific manner almost always utilise multiple RBDs to do so. For example, zinc fingers four through six of Unkempt (Figure 1.2(D)) together only specify three bases, 5′-UAG-3′.

The requirement for multiple RBDs to achieve biological specificity complicates the study of these proteins, particularly given that specificities of individual domains are often not the same when part of a full-length protein [173]. An isolated RBD that binds RNA non-specifically, may, in the context of the full length protein, provide sequence discriminating capacity. For example, the mRNA-binding protein TIA-1 contains three RRMs; when analysed as single domains, RRM2 displays high affinity for U-rich RNA, whilst RRM3 binds U- and C-rich RNA weakly. However, as part of the full length protein, RRM3 provides the sequence discriminating capacity to bind target RNAs [179]. Different molecular mechanisms may be at play depending on the number of domains present, as was shown by the observation that an additional N-terminal helix present in RRM3 of TIA-1 was important for binding of double, but not single, RRM polypeptides to similar RNA sequences [180, 181].

Of particular relevance to this Thesis, homeodomains have been shown to have altered specificity for DNA depending on the length of the protein tested. For example, the homeodomain containing protein ultrabithorax (Ubx), in the context of the full length protein, discriminated between several DNA sequences with affinity differences of ~ten-fold. However, when the homeodomain alone was probed for binding, it bound all sequences similarly, with a higher affinity than the full-length protein [182]. Ubx, like many homeodomain-containing proteins, gains affinity and specificity by binding DNA with other homeodomain-containing proteins. This is illustrated by the structure of a ternary complex of Ubx and extradenticle (Exd) bound to a DNA target site (Figure 2.14(A)); Ubx is bound to Exd through the YPWM motif of Ubx. The YPWM motif, as well as disordered regions on both sides of this motif which were not visualised in the crystal structure (six residues N-terminal to YPWM and eight residues C-terminal to YPWM) were shown to have separable effects on the sequence-discriminating capacity of the protein, through modulation of both intra- and inter-protein interactions. These examples, taken together, show that the binding analysis of single domains can often yield results that are not indicative of their function in full length proteins.

Biological and functional specificity is often achieved not only by multiple RBDs in one protein but through the concerted action of multiple RBPs. The requirement for multiple RBPs for RNA target specification is being increasingly documented, particularly for mRNA-binding proteins. For example, the mRNA transport protein She3p, which contains no known RBDs, binds RNA non-specifically in isolation; however synergistic binding between She3p and She2p results in stable, specific mRNA binding [183]. Similarly, Rna15 is an RRM containing 3′-mRNA processing factor that alone displays little sequence discriminating capacity for RNA. Target recognition is achieved through the combined effect of various weak interactions between multiple proteins and RNA; Rna15 interacts with the proteins Hrp1 and Rna14. The binding of Hrp1 to RNA serves to position Rna15 on target sequences, and Rna14 stabilises the two RBPs on the RNA without binding RNA itself [184, 185].

Several structures have been published that exhibit the intricate interplay observed in multi-RBP:RNA complexes. One such structure is that of a pair of RRMs, one each from the widely expressed ASD-1 and the muscle-specific SUP-12; the proteins interact to achieve tissue specific splicing, creating a cleft that accommodates a guanine base [186], as shown in Figure 2.14(B). Another example that underscores the complex network of interactions that can arise is the structure of an RRM from the *Drosophila* female-specific Sex-lethal (Sxl) bound to the cold shock domain of Upstream-of-N-ras (Unr). This dimer creates an intricate ternary arrangement with *msl2* mRNA [187] that is illustrated in Figure 2.14(C).

As well as providing biological specificity to non-specific RBPs, the association of multiple RBPs can also relax the intrinsic specificity of RBPs. This is illustrated in an elegant example that also demonstrates how functional specificity can arise through the interplay of cellular RBP gradients. In *Drosophila*, the non-specific double ZF-containing RBP Nanos (Nos) is concentrated at the posterior end of the developing embryo, whilst Pumilio (Pum) and *hb* mRNAs are spread throughout the embryo. Nos and Pum form a ternary complex with *hb* mRNA, repressing translation of *hb* in the posterior of the embryo. This ternary complex is shown in Figure 2.15(A), and involves a cytosine base that is flipped out from the Pum binding interface [188]. In the same study, the ability of Nos to modulate the RNA-binding specificity of Pum was illustrated. Pum does not bind *CycB* RNA in isolation, however Nos is able to stabilise Pum on *CycB* RNA. This stabilisation produces a change in the method that Pum recognises bases that are located away from the Pum and Nos binding interface. Instead of the base flip method seen in Figure 2.15(A), recognition of *CycB* RNA involves the 1:1 (base to helix repeat) method, as seen in Figure 2.15(B) [188]. This change in recognition mechanism results in suboptimal contacts with an adenine base, and the disordering of a terminal base.

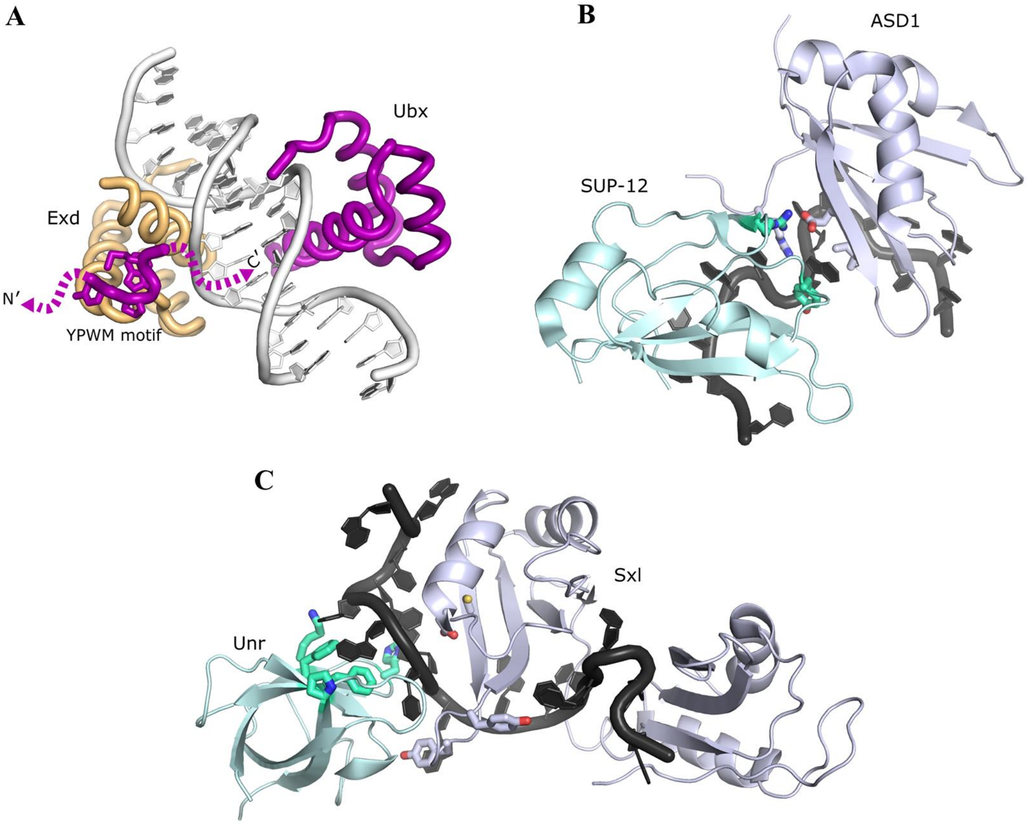


Figure 2.14. Ternary complexes of proteins and nucleic acid.
(A) Ubx (*purple*) and Exd (*orange*) bound to DNA target site (*white*) (PDB: 1B8I). Sidechains of the YPWM motif of Ubx are shown as sticks. Disordered residues flanking the YPWM motif that are not visualised in this structure, but that act as specificity determinants, are indicated by dotted purple lines. **(B)** RRMs from ASD1 (*light blue*) and SUP-12 (*light green*) bound to each other and RNA target site (*black*) sandwiching a guanine base in the middle (PDB: 2MGZ). **(C)** Cold shock domain of Unr (*light green*) and two RRMs of Sxl (*light blue*) bound to RNA target site (*black*) (PDB: 4QQB). Selected RNA-binding residues are shown in stick format (*blue*: nitrogen, *red*: oxygen, *yellow*: sulphur).

The molecular details of the complexes that bicoid forms whilst repressing *cad* translation are still lacking. As discussed in Section 1.2.5, there are several proposed models for bicoid-mediated *cad* silencing. Both Bin3 (a probable methyltransferase, discovered as a bicoid-interacting protein via a yeast two-hybrid screen) and 7SK RNA (a snRNA primarily known for its transcriptional role of inhibiting the positive transcription elongation factor (P-TEFb)) have been shown to stabilise bicoid at the BRE, and other proteins that could possibly interact with bicoid and regulate binding to *cad* include Ago2, poly-A binding protein, Larp1, eIF4E [130] and 4EHP [112]. The studies outlined in the paragraphs above indicate that, in order to determine how bicoid specifically recognises *cad* mRNA, further characterisation of the protein interaction network of bicoid might be required.

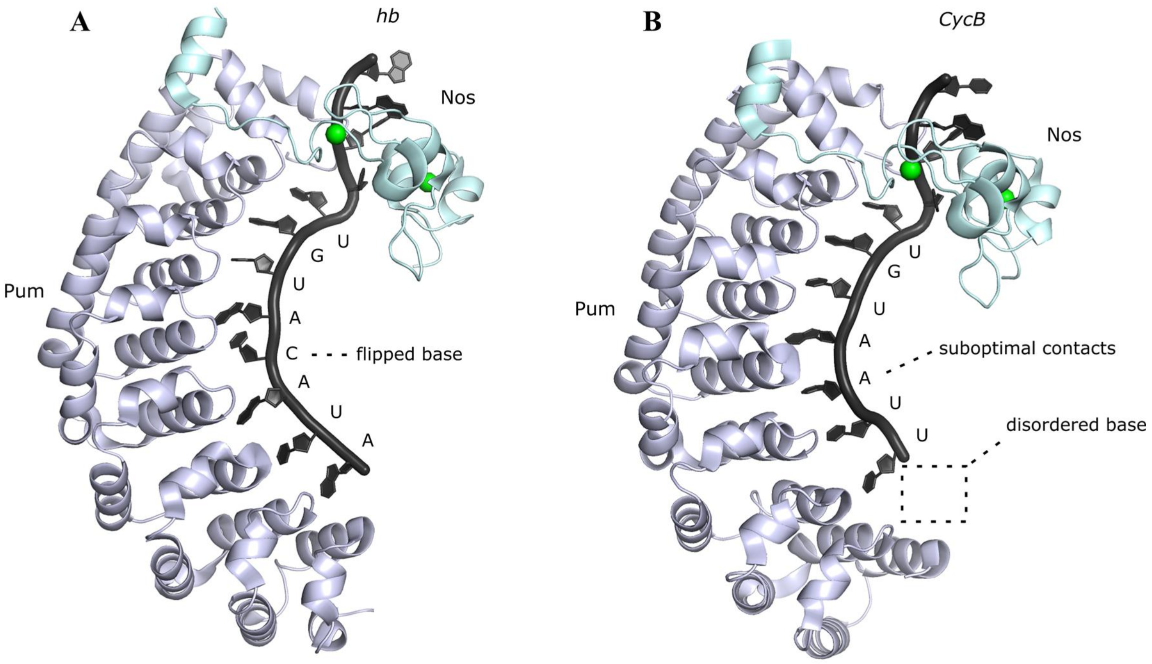


Figure 2.15. Nos modulates Pum specificity to associate with different mRNAs.
(A) Ternary complex of Nos (*light green*, coordinated zincs as *green* spheres) and Pum (*light blue*) bound to *hb* RNA (PDB: 5KL1). A cytosine base that is flipped away from Pum is indicated. **(B)** Ternary complex of Nos (*light green*, coordinated zincs as *green* spheres) and Pum (*light blue*) bound to *CycB* RNA (PDB: 5KL8). Suboptimal binding interactions are observed between Pum and the specified adenine base, and the disordered terminal base not observed in the crystal structure is indicated.

The reduction in BHD binding of BRE19nt induced by an increase in salt concentration to 150 mM may be an indication of the role of the domain in the full length protein. The binding of the cold shock domain of Lin28 with RNA is disrupted by ionic strength (although at a much higher concentration than 150 mM), whereas when the ZnK domain of the same protein is present, ionic strength no longer perturbs the interaction. The proposal is that the CSD samples the transcriptome through transient electrostatic interactions [189].

2.5.2 Experimental conditions might not have been optimal to detect specificity

An alternative explanation to the bicoid homeodomain being a non-specific RBD is that the domain does indeed bind RNA specifically but that the current work was not able to detect this specificity. There are several possible reasons why this may be the case.

First, it is possible that the bicoid homeodomain recognises a short redundant motif that is present in all RNA sequences tested. We know that most RBDs recognise sequences of only two to eight nucleotides, with considerable sequence variation frequently permissible [190]. For example, the optimal binding motif of the RRM from the splicing factor hnRNP G is a 5′-AA-3′ motif, however the domain can also bind 5′-CCC-3′ and 5′-CCA-3′ motifs [191]. In the work described in this Chapter, care was taken to design unique RNA transcripts as controls, but it is difficult to design sequences utilising all four bases that don't contain instances of the same trinucleotides. To illustrate this point, all tested sequences include at least one instance of 5′-GAA-3′ and 5′-GUU-3′ trinucleotides.

It is also possible that the bicoid homeodomain binds non-target RNA non-specifically, but we have failed to find its target sequence because it was not contained in the RNA sequences tested in this Chapter. Given the experimental procedure for making radiolabelled RNA, it is difficult to accurately estimate the size of the RNA product (in contrast to the situation for non-radioactive RNAs). Therefore, if transcription of the full length *cad* 3′-UTR was partially truncated, a critical element of the bicoid homeodomain target might have been omitted. Alternatively, the target sequence might not be contained in the *cad* 3′-UTR. Reports of the bicoid homeodomain target sequence within the *cad* 3′-UTR have varied as previously mentioned in Section 1.2.4, and it has even been reported that the binding site might extend upstream in to the coding region of *cad* [87]. Further, 7SK RNA is reported to be part of a repressive complex formed with bicoid to inhibition *cad* translation [130], which foreseeably could be directly bound by the homeodomain.

It is also possible that assay conditions might not have been optimal for detecting a specific interaction between the bicoid homeodomain and RNA. Many reasons for this possibility arise due to the structural flexibility of RNA and RBPs that was introduced in Chapter 1. For example, the RNA might not be folded correctly, or the RNA or homeodomain might be missing required post-transcriptional or post-translational modifications.

RNA secondary structure complicates the study of RBP:RNA interactions as binding interfaces can become exposed or sequestered depending on the length of RNA used in one's experiments [192]. An excellent example of this phenomenon is the bacterial global translation repressor protein CsrA/RsmE that exhibits modulation in affinity to a GGA motif from ~10 nM to 3 mM; the highest affinities are observed when the A(N)GGAX motif is at the top of a hairpin loop, and the affinity is reduced when the motif is obstructed due to base pairing [193]. Further, the hairpin loop is a preferred target compared to the same motif in ssRNA, due to the lower entropic cost of binding to a hairpin. This study highlights how *in vitro* results can be distorted from the underlying biology if the RNA is not optimally folded.

The structural flexibility of RNA and RBPs can also mean that there is often plasticity in their binding interfaces. Transcriptome wide studies of RBPs have shown binding of individual RBPs to multiple

RNA targets with extensive sequence and/or structural variation [194, 195]. The accommodation of various RNA sequences has been shown via structural adaptations of the protein [56, 196] and RNA [80, 197]. As a result, it is being increasingly understood that domain arrangements and conformational dynamics of RBPs have important implications for the recognition of biological targets [173]. This is of particular relevance to the bicoid homeodomain, with multiple studies indicating that sidechain conformational heterogeneity in the domain's DNA-binding residues (that is, the sidechains have been observed to adopt a variety of conformations when bound to different DNA sequences) leads to the homeodomain's unique nucleic acid binding properties [122, 198, 199]. Such flexibility presumably makes finding biologically relevant targets more difficult.

Post-transcriptional and post-translational modifications of proteins and RNA further complicate the *in vitro* study of their interactions. For example, phosphorylation has been shown to up-regulate binding of the Ezh2 domain of PRC2 to HOTAIR lncRNA [200]. Some RBPs recognise PTMs specifically. PIWI proteins, for example, recognise piRNAs that are 2′-O methylated at their 3′ ends [201], and piRISCs bind RNAs that are 2′-O methylated at 3′ ends. When proteins are expressed outside of their host organism, and RNA is transcribed *in vitro*, any relevant PTMs will be missing.

Whilst *in vitro* quantification of binding interactions often provides useful information, the inherent affinity an interaction is only one factor determining biological specificity. Often specificity is not an inherent property of RBPs but rather it is dictated by the cellular context in which RNP interactions take place [81]. The concentration and accessibility of each binding partner (including the accessibility of the RNA-binding site as discussed above), as well as the concentration of competitive binders, to a large extent dictates what interactions will take place. A demonstration of the effect of cellular context on specificity is the inhibition of PRC2 histone methyltransferase activity made at the vestigial locus in *Drosophila*. *In vitro*, inhibition can occur by both forward and reverse noncoding RNAs; however, in cells only the latter appears to bind PRC2, indicating that biological specificity is being driven by something other than affinity, such as availability of RNAs [202].

Moreover, the highest affinity interactions do not always indicate biological targets. This was demonstrated by an innovative, high-throughput kinetics approach which showed that the biological substrates of C5 (a component of the tRNA processing RNAse P in *Escherichia coli*) are not the sequences it binds tightest but rather those near the middle of the affinity distribution [203].

An excellent review detailing the complexity often encountered *in vivo* for RNA:RBP interactions has been published recently, highlighting how the categorisation of these interactions as inherently specific or non-specific is often not overly meaningful to the underlying biology [81]. Notably, this review outlines the impact of the kinetic context in which a binding event takes place. Strikingly, the kinetics of reactions that are separate to, but preceding or following, binding interactions can counterbalance the inherent affinity of a protein for an RNA. The potential relevance of this analysis to bicoid is clear; as depicted in Figure 1.7(A), the developing *Drosophila* embryo is patterned precisely by differential concentrations of protein and RNA molecules, creating unique kinetic contexts for bicoid throughout the embryo. Bicoid binds tens of DNA targets as well as *cad* mRNA, so the differing concentration of the protein (and binding partners and competitive binders) throughout the embryo is likely instructive to its biological RNA-binding specificity.

2.4 Summary

This Chapter has demonstrated that BHD binds RNA promiscuously *in vitro*. The strategy employed was to test binding of BHD to various RNA sequences in an attempt to observe differential binding, as a starting point to be able to define sequence or structural elements that were both necessary and sufficient for binding. However, BHD was observed to bind equally well to RNAs of differing sequence composition, and with both the presence and absence of predicted secondary structure. The involvement of a miRNA complementary to a conserved hairpin loop in BRE was investigated, with results indicating the possibility that BHD might bind a miRNA:mRNA duplex. The N-terminal boundary of the homeodomain was then extended by nine residues to include a potential ARM; this extended domain, HDER, had a four-fold higher affinity for the conserved distal loop of BRE than that of BHD. The ability of physiological salt concentration to reduce the RNA-binding affinity of both BHD and HDER by six to eight-fold indicates that this interaction is largely electrostatic.

The RBP research presented in the previous Section highlights that determining how RBPs find their cellular targets is a highly complex task, requiring *in-vivo* and *in-vitro* techniques to work side by side. The involvement of the bicoid homeodomain in the specific repression of *cad* has been postulated based on previous *in-vivo* and *in-vitro* research [87, 123, 126], however, the *in-vitro* work presented in this Chapter was unable to define specificity determinants of the interaction. Further characterisation of the homeodomain's interaction with RNA using techniques that yield superior molecular resolution may help to resolve its role in *cad* translational repression.

Chapter 3: NMR analysis of BHD and the interaction between BHD and RNA

3.1 Introduction

The previous Chapter addressed the RNA-binding specificity of the bicoid homeodomain but no clear specificity could be discerned. However, the analysis at the end of the Chapter highlighted the point that ascertaining how RBPs achieve biological specificity is a very complex question. We therefore sought to characterise the BHD-RNA interaction in more detail, in the hope that knowledge of the molecular basis for the interaction would be of value in understanding (a) how proteins can recognise both DNA and RNA and (b) how proteins find their cellular targets. The bicoid homeodomain is of particular interest in these regards as it is the only homeodomain that is currently known to bind both DNA and RNA.

Accordingly, this aims of this Chapter are to characterise the molecular details of RNA-binding by BHD and to further interrogate its RNA-binding specificity.

3.2 Techniques used in this Chapter

3.2.1 Nuclear magnetic resonance spectroscopy

The NMR data presented in this Chapter go beyond the simple 1D ^{1}H NMR spectra presented in the previous Chapter, and employ uniform ^{15}N and ^{13}C isotopic labelling of the protein to both improve resolution and to permit assignment of signals to specific atoms in the protein.

3.2.1.1 Two dimensional NMR spectroscopy

At the two-dimensional (2D) level, the introduction of isotopic labelling results in less signal overlap by spreading signals over another dimension, providing the resolution required to distinguish signals for individual residues. As the size of a protein increases, so too does the chance of signal overlap, which makes the task of distinguishing signals progressively more difficult. The 2D experiment used most widely in protein biochemistry is the ^{15}N-^{1}H heteronuclear single quantum coherence (^{15}N-HSQC) spectrum. ^{15}N-HSQC experiments are set up such that only directly bonded N/H groups give rise to signals, yielding signals for each amide group of the protein backbone as well as any amide groups contained in sidechains.

The chemical shift and intensity of each signal is dependent on the chemical environment of the amide group, and therefore ^{15}N-HSQC experiments can provide extremely valuable structural information. For example, information about the flexibility or secondary structure of residues is garnered as those that are contained in rigid sections of secondary structure will show reasonably strong peak intensities, whereas flexible residues will often show significantly more intense signals or, conversely, are sometimes not observed. Further, as the chemical environments of magnetic nuclei are altered by binding events, the ability to observe binding induced changes, known as chemical shift mapping, makes it possible to characterise the molecular details of binding interactions. For suitable proteins, these changes can be observed by ^{15}N-HSQC experiments. The contribution of specific residues to a binding interaction can be quantified by chemical shift, linewidth or intensity differences between the free and bound state.

Chemical shift change is the most commonly used parameter for assessing interactions. As mentioned, a nucleus perturbed by a binding interaction will display a different chemical shift in the free and complexed form. The signal(s) that this nucleus gives rise to in an NMR spectrum depend on both the rate of exchange between the free and complexed form, and also the difference in the chemical shift between these two forms. A fast exchange regime occurs when the exchange between the two forms is faster than the difference in frequency (chemical shift) between the nucleus in the two (or more) states, and generates a single signal with a population weighted average chemical shift. A slow exchange regime occurs when the exchange between the two forms is slower than the frequency difference, and gives rise to two individual signals for the free and complexed state. An intermediate exchange regime lies in the middle, and results in signal broadening, to the extent that resolution can become so poor that signals are not observed.

In order to quantify the chemical shift changes upon RNA addition, the following equation is applied to the data to calculate weighted average chemical shift changes; this equation factors in empirically determined differences in resonance sensitivity of amide nitrogen and protons [204]:

$$\Delta\delta_{total} = \sqrt{(\delta_{NH}W_{NH})^2 + (\delta_N W_N)^2} \qquad \text{Equation 3.1}$$

where $\Delta\delta_i$ is the chemical shift change between nucleus/nuclei i in the free and complexed state and W_i is empirically determined weighting factor for each nucleus; in this thesis W_{NH} = 1 and W_N = 0.154 [205]. $\Delta\delta_{total}$ is then plotted as a function of residue number. Residues for which $\Delta\delta_{total}$ is greater than the average $\Delta\delta_{total}$ plus one standard deviation are deemed to be significantly influenced by RNA binding.

3.2.1.2 Three dimensional NMR spectroscopy

Three dimensional (3D) protein NMR spectra typically are acquired on proteins that are ^{13}C as well as ^{15}N labelled. Many different spectra can be acquired that correlate the absorption frequencies of three (or more) nuclei; often two of these are the amide proton and nitrogen. In the most commonly used set of experiments, amide groups are correlated with ^{13}Cα/^{13}Cβ or ^{13}CO atoms from the same and/or preceding residues.

Like other nuclei, different backbone carbon atoms display characteristic chemical shifts: carbonyl carbons display different chemical shifts to that of α and β carbons, and α and β carbon resonance ranges vary for each amino acid. Particular residues such as alanine, serine and threonine display diagnostic Cα and Cβ chemical shifts. Spectra are acquired in pairs so that one spectrum will obtain ^{13}C signals for both the same residue as the amide signal (r_i) and the residue preceding the amide signal (r_{i-1}); the paired spectrum will correlate the amide group just with the carbon(s) of residue for r_{i-1}. This linkage between pairs of spectra, coupled with knowledge of the protein sequence and characteristic carbon shifts, facilitates the sequential assignment of signals to protein residues. Therefore, ^{13}C/^{15}N triple resonance experiments allow the connection to be made between NMR signals and protein residues. This information can be used in conjunction with 2D chemical shift mapping to identify which residues are involved in a binding event.

3.3 NMR analysis of bicoid homeodomain and RNA interaction

This section details NMR experiments aimed at further validating and characterising the bicoid homeodomain and its interaction with RNA. ^{15}N-HSQC experiments were carried out to map chemical shift changes of BHD upon titration with RNA to gain further insight in to the nature of the interaction, and ^{15}N/^{13}C triple resonance experiments were collected to allow assignment of protein residues in order to determine which residues are important for binding RNA.

Despite HDER having a higher affinity for RNA than BHD as discussed in Section 2.3.4, the lower yield obtained from overexpression and purification deems it a poor candidate for detailed NMR studies. Therefore the BHD construct was chosen for further binding analysis.

3.3.1 BHD ^{15}-N HSQC spectrum

Initially, a two dimensional ^{15}N-HSQC spectrum of BHD was acquired. As introduced above (Section 3.2.1.1), a ^{15}N-HSQC spectrum contains signals that correlate the absorption frequencies of directly bonded H-N atom pairs. Signals are thus observed for backbone amides and for some sidechains.

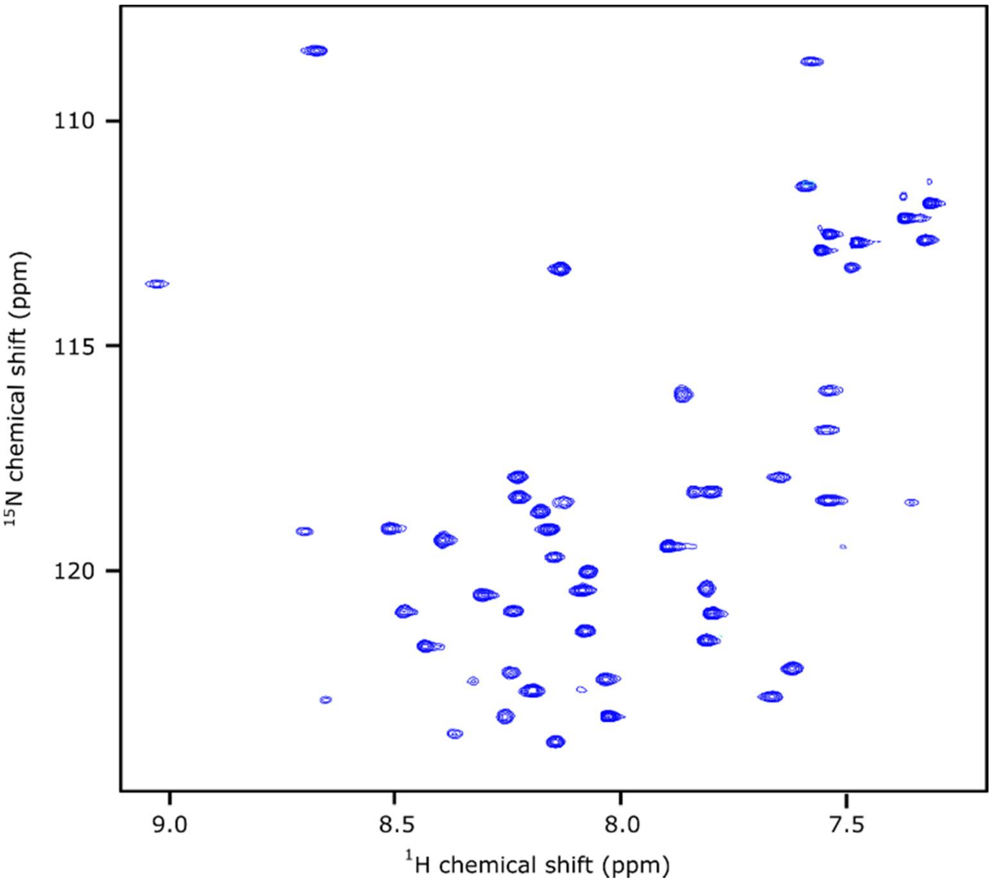


Figure 3.1. ^{15}N-HSQC spectrum of ^{15}N-labelled BHD.
^{15}N-HSQC spectrum of 200 µM BHD in 50 mM HEPES pH 7.3, 50 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP. Spectrum was recorded on an 800 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe (NS = 6; 1TD = 256) at 298 K.

^{15}N-labelled BHD was expressed in minimal medium (Section 7.2.2.4) and purified as per unlabelled protein (Section 2.3.1). ^{15}N-labelled BHD expression and purification gave a similar yield to unlabelled BHD, at around 1 mg per litre of culture and >95% purity. A ^{15}N-HSQC spectrum of BHD alone was acquired; the domain gave rise to a spectrum with good line widths and well-defined peaks. Signals were observed for 43 out of 65 residues expected to give rise to an HSQC signal, shown in Figure 3.1. As for 1D ^{1}H spectra, well-defined and dispersed peaks indicate an ordered protein. Correspondingly, this spectrum indicates that BHD is well-folded, as expected. According to the structure of the bicoid homeodomain bound to DNA, there are 46 α-helical residues with the remainder of the residues located in loops or terminal arms [122]. It is perhaps likely that many of the missing signals are from these more flexible regions that are not located in the α-helices of the domain. Amide protons that are not located in stable structure can often undergo millisecond to microsecond timescale dynamics. These dynamics can give rise to the intermediate exchange phenomenon described above, which can have the effect of broadening the signals beyond detection.

3.3.2 Assignment of BHD residues

In order to make the backbone assignments for the domain, BHD was overexpressed as ^{13}C/^{15}N-labelled protein in minimal medium (Section 7.2.2.4) and purified as per ^{15}N and unlabelled protein (Section 2.3.1). ^{13}C/^{15}N-labelled BHD expression and purification gave a lower yield compared with that of ^{15}N and unlabelled BHD, at around 500 μg per litre of culture and >95% purity. Expression was scaled up to two litres to enable ~200 mM samples to be made. A series of 3D, triple resonance spectra were acquired on the protein alone, namely HNCACB, CBCA(CO)NH, HN(CA)CO and HNCO. Correlations of the amide proton and nitrogen with ^{13}Cα and ^{13}Cβ nuclei are obtained from the HNCACB (r_i and r_{i-1}) and CBCA(CO)NH (r_{i-1}) spectra, and ^{13}CO correlations are obtained from the HN(CA)CO (r_i and r_{i-1}) and HNCO (r_{i-1}) spectra, as depicted in Figure 3.2(A). The strategy employed for assignment of the protein backbone residues is illustrated in Figure 3.2(B), which shows assignment of residues D129 to A132. These spectra illustrate the linkage of signals between residues (indicated by dotted lines), specifically demonstrating the low ^{13}Cβ shift that is diagnostic of alanine, and the high ^{13}Cβ shift that is diagnostic of serine. Residues with diagnostic shifts are typically identified first, and carbon signals are allocated to the previous residue if they are observed in the CBCA(CO)NH spectrum, then connections between adjacent residues are made by tracing signals through the HNCACB spectrum. The CO resonances assist in assignment by connecting carbonyl signals to amide signals in an analogous manner from the HN(CA)CO and HNCO spectra, which can be helpful if there are missing or faint Cα and Cβ signals.

Some undesirable features are also shown, namely; signal overlap between D129 and L130 (for both ^{13}Cα and ^{13}CO), and a missing ^{13}CO signal for S131 in the S131 strip of the HN(CA)CO. The frequency of signal overlap and missing signals depends on the individual protein; the probability of the former increases with the size of the protein, and the latter depends in part on the quality of the spectra and on a variety of other factors, including the presence of chemical exchange processes. When spectral quality is less than ideal, multiple pairs of spectra may need to be acquired to facilitate assignment, as was the case here for BHD.

Assignments of amide N/H resonances were made for 44 residues. This included two separate situations in which the ^{15}N-HSQC signals for two residues were coincident peak overlaps, resulting in assignment of 42 out of the total count of 43 separable peaks in the spectrum. The assigned ^{15}N-HSQC spectrum is shown in Figure 3.3(A).

The chemical shifts obtained from this data set were used to predict the secondary structure of the domain by MICS (Motif Identification by Chemical Shift) program [206]. This program uses an artificial neural network algorithm to predict the likelihood that each residue will be located in particular types of secondary structure. The domain is predicted to be primarily α-helical as expected, and the

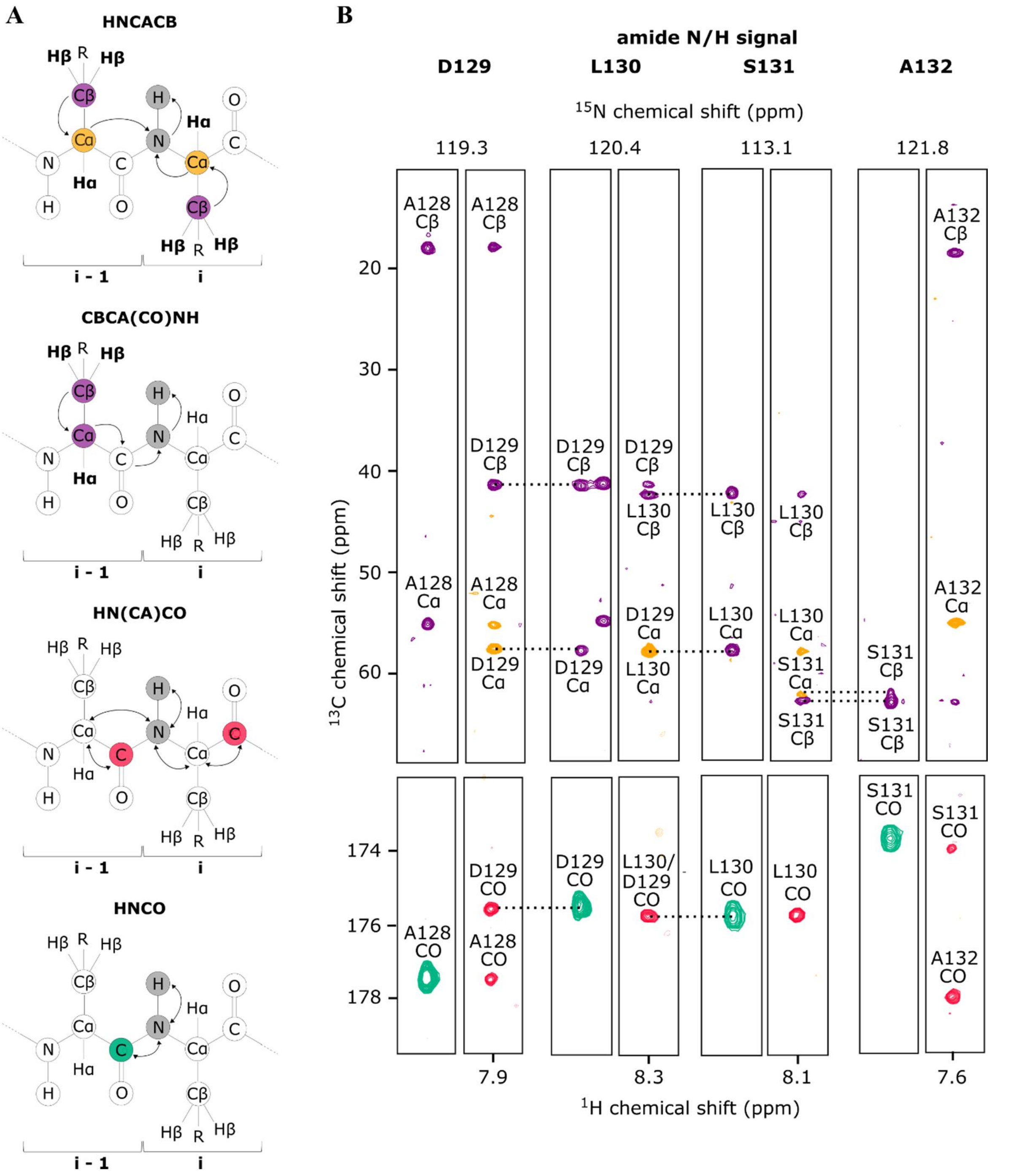


Figure 3.2. Triple resonance spectral assignment of BHD residues.

(A) Schematics indicating magnetisation transfer for each spectrum. HNCACB: magnetisation is transferred via α and β hydrogens from r_i and r_{i-1} (indicated in bold text) to both α and β carbons respectively, with β carbon magnetisation transferred to the attached α carbon, then α carbon magnetisation from both r_i and r_{i-1} is transferred to the amide nitrogen of r_i before being transferred to the amide hydrogen. CABCA(CO)NH: magnetisation is transferred via α and β hydrogens from r_{i-1} (indicated in bold) to the attached β and then α carbon, then to the carbonyl carbon (where it is not evolved hence no signal), then to the nitrogen and hydrogen of r_i. HN(CA)CO: magnetisation is transferred from the amide hydrogen of r_i to the attached amide nitrogen, then to both α carbons (not evolved) and then onto attached carbonyl carbons. Magnetisation is then transferred back via the same route for detection. HNCO: magnetisation is transferred from the amide hydrogen of r_i to the amide nitrogen and then to the attached carbonyl carbon, before being transferred back via the same route for detection. Abbreviation: r, residue. Shaded circles indicate a signal is obtained for that atom. **(B)** Strips from nitrogen planes for HNCACB, CBCA(CO)NH, HNCACO and HNCO showing signal links (indicated by dotted lines) between spectra that allow sequential assignment of BHD residues. Signals are coloured to indicate identity consistent with schematics in (A).

predicted helical regions largely correspond to the helical regions identified by the solution structure of the bicoid homeodomain bound to DNA [122] (Figure 3.3(C)). This analysis suggests that BHD purified for this work likely adopts a similar secondary structure to the domain when bound to DNA.

Residues that were not assigned mapped to a number of regions throughout the protein, and have been highlighted on the structure of the domain (when bound to DNA) in Figure 3.3(B). Six out of eight N-terminal residues were unassigned, indicating that this region is probably undergoing some conformational exchange. Following DNA binding, it has been shown that this region becomes ordered by making contacts with the DNA minor groove [122]. Eight out of 22 residues from helix three were not observed. This helix is known to display a high degree of conformational flexibility [122, 198]. Recently, thermal denaturation studies indicate that the flexibility of this helix is attributable to the availability of multiple folded conformations with similar stability [199]; this is a likely explanation for why a large number of residues were not observed in this region. The majority of the remaining missing residues are located in the loop region between helix one and helix two, a region that is also likely to exhibit conformational flexibility.

3.3.3 Mapping the RNA-binding residues on BHD

The chemical shift assignments made in the previous Section were used to determine which residues in BHD are involved in RNA binding. The first RNA target tested was BRE19nt, the conserved distal hairpin loop of BRE shown in Figure 2.13(B); BHD binds BRE19nt with a K_d ~ 3.8 µM. In initial experiments, BRE19nt was titrated into a solution of ^{15}N-labelled BHD in 0.2 molar equivalent increments. However, in all cases precipitation was observed at lower ratios of RNA and protein. In the aim of avoiding precipitation, the protein was added to 0.2 molar equivalent increments of RNA (instead of 0.2 molar equivalents of RNA added to the protein), however some precipitation still occurred. Finally, precipitation was avoided was by adding two molar equivalents of RNA to the protein at once.

Figure 3.4(A) shows a ^{15}N-HSQC spectrum of BHD alone and in the presence of two molar equivalents of BRE19nt. Figure 3.5(A) shows the magnitudes of chemical shift changes upon RNA addition, which were calculated using Equation 3.1.

As a titration was unable to be carried out, signal changes could not be tracked unambiguously. Instead, the nearest neighbour method was used [207], whereby new signals are assigned to the nearest residue as outlined in Section 7.2.7.4. In this method, the assignments of the bound state are based on proximity to the nearest peak in the free protein state, with a correction of 1/7 used for ^{15}N resonances in order to

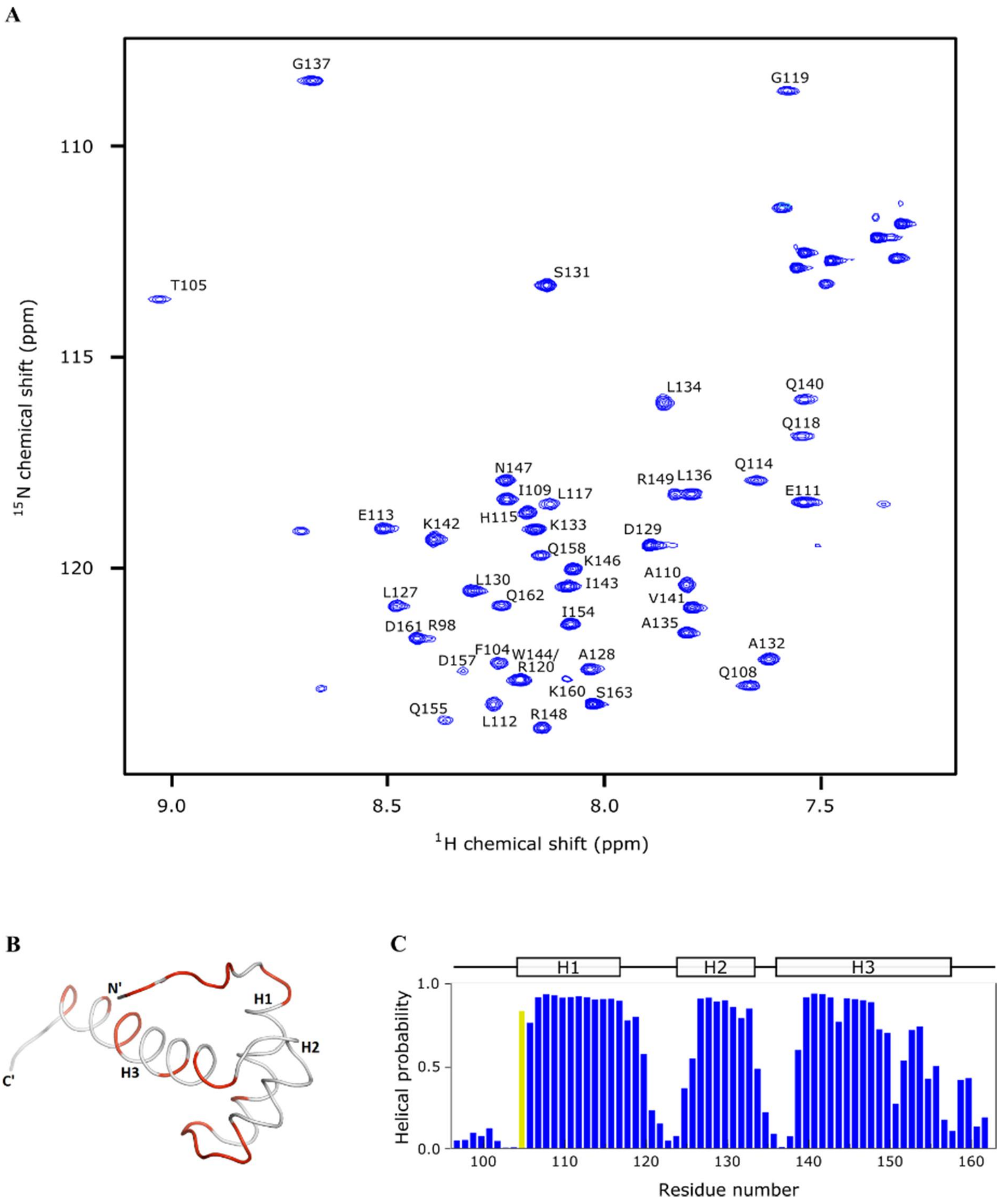


Figure 3.3. Assigned ^{15}N-HSQC spectrum of BHD.

(A) ^{15}N-HSQC spectrum of BHD alone (blue). Starting concentration of BHD was 200 µM, in 50 mM HEPES pH 7.3, 50 mM KCl, 1 mM MgCl$_2$, 1 mM TCEP. Spectrum was recorded on an 800 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe (NS = 6, 1TD = 256) at 298 K. **(B)** Residues missing from the assigned ^{15}N-HSQC spectrum are mapped onto the structure of bicoid homeodomain (PDB: 1ZQ3). Residues that could not be assigned are shown in *red* and assigned residues are shown in *grey*. N′ and C′ terminal ends are indicated, and helices one, two and three are indicated as H1, H2 and H3 respectively. **(C)** Predicted likelihood that BHD residues are located in an α-helix, obtained using the algorithm MICS and based on analysis of available ^{1}H, ^{15}N and ^{13}C (α, β, CO) chemical shifts. Residues known to be helical from the structure of bicoid homeodomain bound to DNA (PDB: 1ZQ3) are indicated above the graph. The yellow bar indicates the likelihood of this residue being an N-terminal helix capping motif.

give a roughly equal weighting to the ^{1}H signals. Moreover, assignment of new signals that appeared in the spectrum with RNA was not possible; this would have required the acquisition of more triple resonance spectra of BHD in the presence of BRE19nt.

All signals moved at least slightly upon RNA addition, with a $\Delta\delta_{mean}$ of 0.05 ppm. There were more new signals than signals that disappeared (12 versus two, respectively), indicating that the domain becomes more ordered on binding RNA, or at least that more amide protons are protected from exchange with the solvent water. The extent of signal change implies that the interaction with RNA is widespread across the domain, and/or that some conformational change takes place upon binding.

The observed chemical shift changes for the BHD:RNA interaction are relatively small with significant perturbations ranging between 0.09 and 0.16 ppm. Chemical shift perturbations of this magnitude have been reported for non-sequence specific RNA-binding interactions. Examples include the non-sequence specific binding of RNA by both the YY1 ZFs (which resulted in significant chemical shift perturbations of 0.06–0.14 ppm [160, 208]) and ZF1 of JAZ (0.05–0.11 ppm [160]). In contrast, some larger chemical shift perturbations of non-sequence specific interactions have been reported. For instance, ZF3 from JAZ, like ZF1, binds dsRNA without regard to sequence (except the requirement of A-form RNA), however significant shifts were between 0.12 and 0.55 ppm [160]. Certainly, the magnitude of the perturbations observed for the BHD:RNA interaction are smaller than those commonly seen for sequence specific RBP:RNA interactions. As illustrations of these larger chemical shift perturbations, ZF2 of ZRANB2 displayed significant shifts in the range of 0.2–1.2 ppm [49] and the RRM of SRp20 displayed significant shifts between 0.4–2 ppm [209] upon binding RNA sequence specifically. These larger chemical shifts are generally indicative of base-specific hydrogen bonding interactions, which effect large downfield shifts [210] or the intercalation of aromatic rings between bases [211]. There is a notable absence of significant downfield ^{1}H shifts in the BHD:BRE19nt spectrum, perhaps indicating an absence of specific hydrogen bonds, at least with the protein backbone. Taken together, the chemical shift changes observed for this complex are likely indicative of a non-specific interaction.

Residues that displayed the largest chemical shifts changes upon RNA addition are shown on the bicoid homeodomain structure in Figure 3.6(A). The residues that undergo the largest perturbation map to helix three (the DNA-binding recognition helix) and to the loop between helices one and two, which is in broad agreement with the regions most perturbed by DNA-binding (DNA-binding residues are highlighted in Figure 3.6(B)) [122].

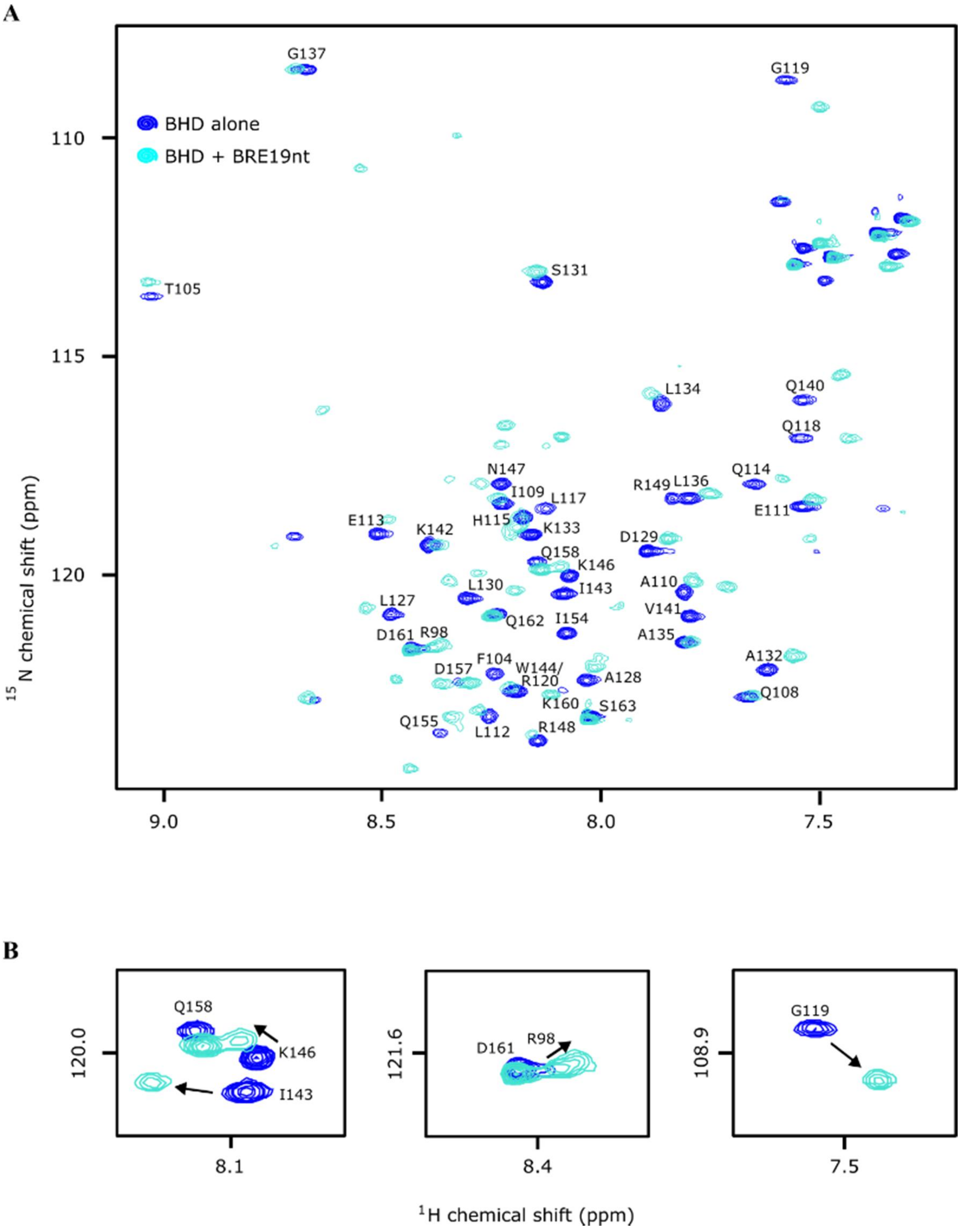


Figure 3.4. Overlay of ^{15}N-HSQC spectra of BHD in the presence and absence of BRE19nt.
(A) ^{15}N-HSQC spectra of 200 μM BHD alone (*blue*) and with 2 molar equivalents of BRE19nt (*cyan*). Buffer composition was 50 mM HEPES pH 7.3, 50 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP. Spectra were recorded on an 800 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe (NS = 3 for protein alone and NS = 6 for 2 molar equivalents BRE19nt, 1TD = 256) at 298 K. **(B)** Specific residues of interest; likely direction of change is indicated by arrows.

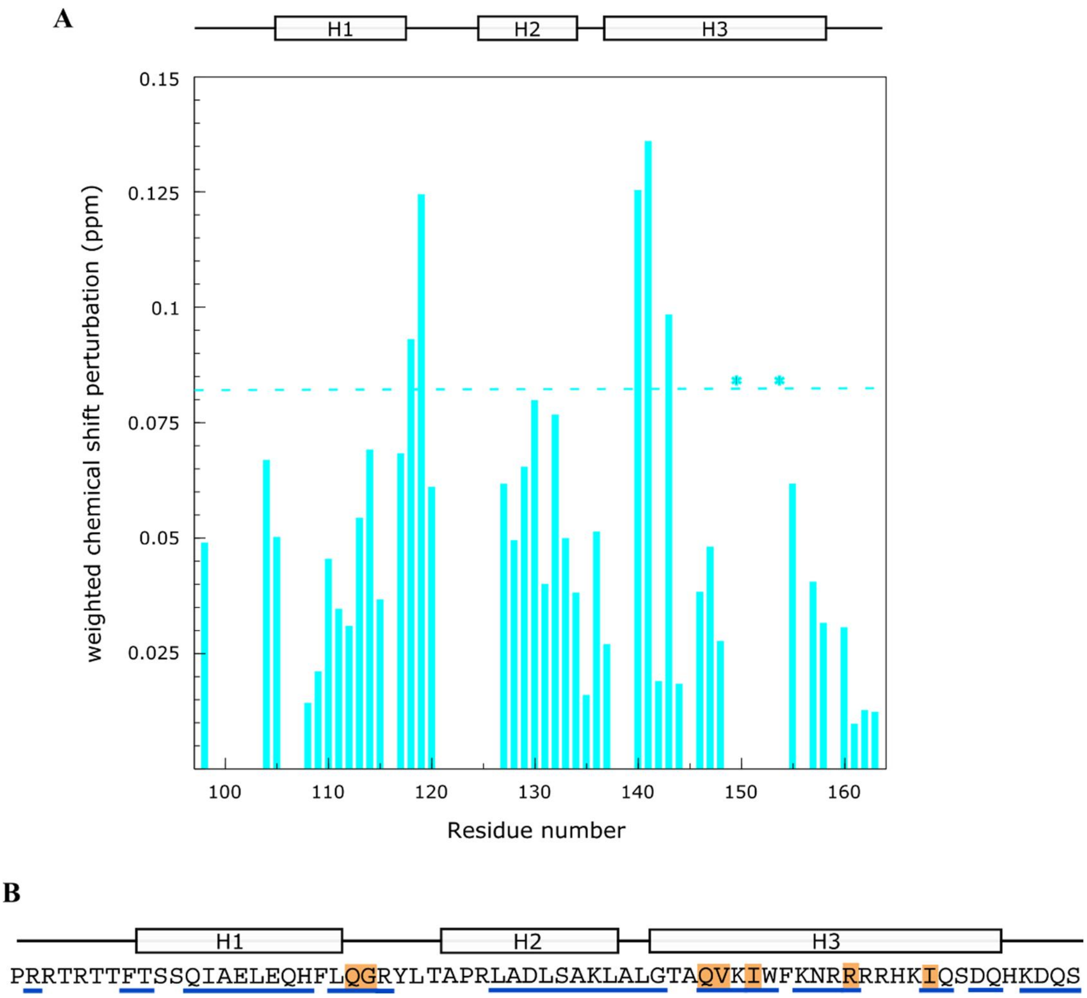


Figure 3.5. BRE19nt RNA-induced chemical shift changes of BHD backbone amides.
(A) Weighted chemical shift changes of BHD residues plotted as a function of residue number following the addition of 2 molar equivalents of BRE19nt. Horizontal dashed line at ~0.082 indicates the threshold level of significance at one standard deviation above the average chemical shift change. Asterisks indicate residues which disappeared with BRE19nt addition. α-helices of BHD (when bound to DNA) are indicated above the graph, and are derived from the published structure of the BHD-DNA complex (PDB: 1ZQ3). **(B)** Amino acid sequence of BHD with assigned residues underlined in *blue* and significantly perturbed residues highlighted in *orange*. α-helices of BHD (when bound to DNA) are indicated above the sequence.

Hydrophobic interactions are likely important for the interaction between BHD and RNA; three hydrophobic residues in helix three displayed significant shifts (V141, I143 and I154). Three charged residues were also observed to be key to this interaction; Q118, Q140 and R149, as well as a single glycine, G119.

Some degree of conformational change upon binding is indicated by these spectra. Firstly, movement of the helices relative to one another is suggested by a number of hydrophobic core residues (as determined in the structure of the domain bound to DNA [122]) displaying shift changes over 0.05 ppm. These residues consist of F104, L130 and L136 as well as the significantly perturbed V141 (with a

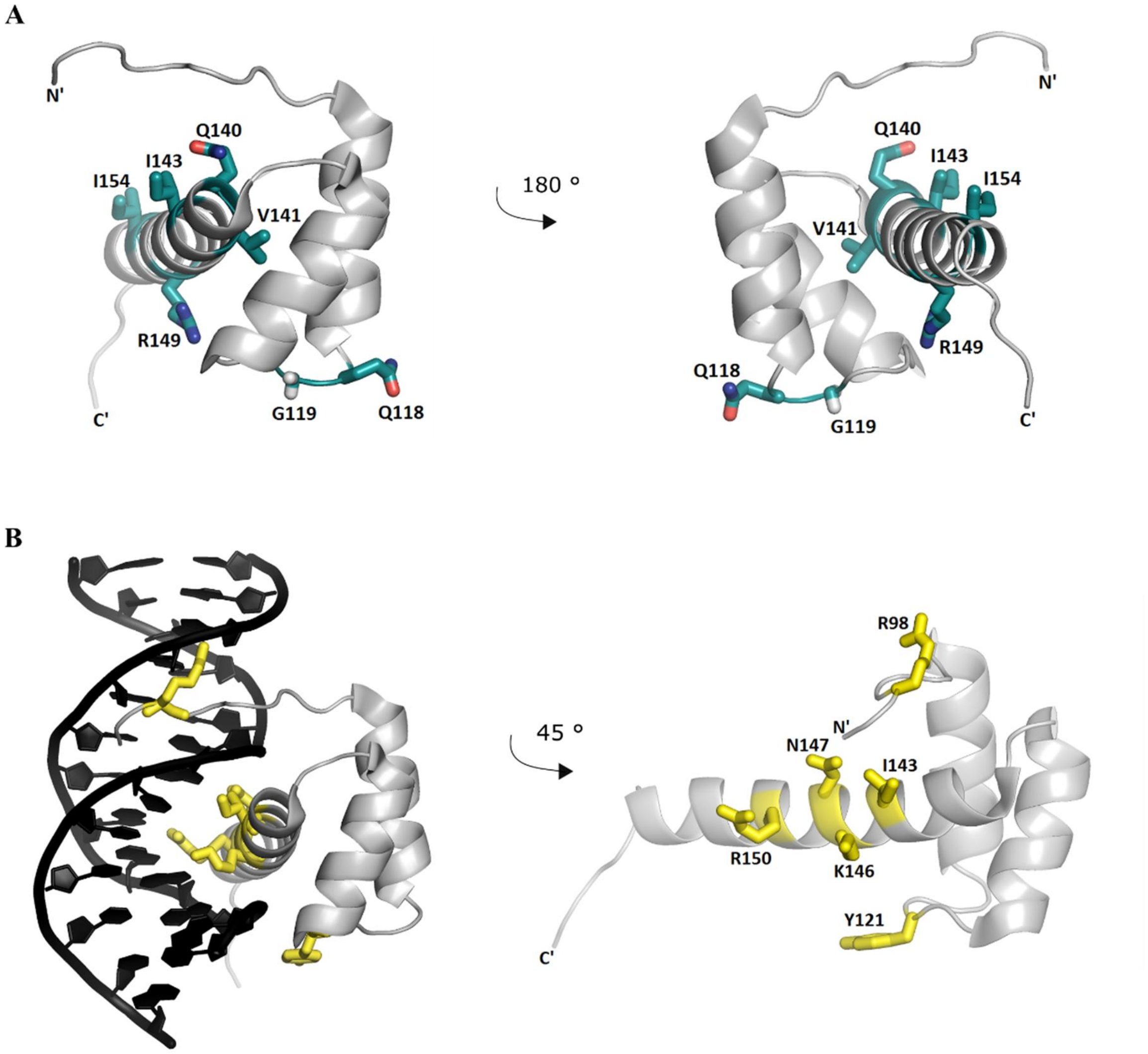


Figure 3.6. BRE19nt RNA-induced chemical shift changes of BHD backbone amide residues.
(A) Structure of BHD, taken from the BHD-DNA structure (PDB:1ZQ3). RNA-binding residues (identified by weighted average backbone chemical shift changes between free and RNA-bound BHD) are mapped on to the structure as sticks (*teal* indicates carbon, *blue* indicates nitrogen, *red* indicates oxygen, *white* indicates hydrogen (hydrogen only indicated for G118). Structures are rotated 180° along the y-axis relative to each other. **(B)** Structure of BHD bound to DNA consensus sequence (PDB:1ZQ3), with DNA-binding residues mapped on to structure as sticks. The right-hand image has been rotated 45° along the y-axis, relative to the left-hand image.

weighted chemical shift change of 0.14 ppm). Given more limited solvent accessibility, shifts in internal regions of proteins are affected primarily by bond geometries and packing interactions [212]. Secondly, Baird-Titus *et al*., in their analysis of the BHD:DNA interaction, make the point that it is unusual to have a glycine in position 119, and speculate that this residue allows helix one to move closer to helix two than is typical for homeodomains [122]. A prominent shift in the signal of G119 of 0.12 ppm (Figure 3.4(B), right)) may imply that there is movement of helix one relative to helix two upon binding RNA. This is corroborated by chemical shifts greater than 0.05 ppm observed for Q114, L117 and Q118, residues which are also key to this 2.5-Å deviation of helix one [122].

The broad regions of secondary structure utilised by the domain to bind nucleic acid appear to be conserved between DNA and RNA recognition, with the residues that are significantly perturbed by both DNA and RNA-binding located primarily in helix three as well as the loop between helices one and two, and the N-terminal arm (Figure 3.6).

However, identification of the specific residues involved in RNA recognition cannot be determined unequivocally. The residues that contact DNA directly are I143, K146 and R150 from helix three, Y121 located in the loop between helix one and two and R98 from the N-terminal arm. The data presented here provides evidence that I143 and R98 are also involved in RNA recognition. I143 displayed a significant chemical shift of ~0.1 ppm. R98 was one of only two N-terminal arm residues that were visible in the protein alone spectrum. It is notable that this signal was very faint in the protein-alone spectrum, but that a more intense signal appears close by after RNA addition (Figure 3.4(B), middle)). Y121 and R150 were not observed, and therefore their role in RNA recognition cannot be confirmed. Lastly, K146 was assigned to a slightly shifted signal ($\Delta\delta_{K146}$ = 0.04 ppm) using the nearest neighbour method of assignment. As can be seen in the left schematic of Figure 3.4(B), there is some ambiguity with this assignment. It is possible that the closely shifted signal could be from I143, with K146 exhibiting the large ^{1}H downfield shift attributed to I143; however, further studies either involving optimisation of the conditions for triple resonance spectral acquisition, or triple resonance spectral assignment of BHD bound to BRE19nt, will be required to confirm this. In any case, given that it is the lysine sidechain that contacts DNA, a lack of substantial chemical shift change observed for the backbone amide cannot rule out a sidechain interaction.

3.3.4 Further BHD:RNA chemical shift mapping

The results from Chapter 2 demonstrated that BHD possibly binds RNA non-specifically, in a primarily electrostatic fashion. In order to corroborate these assertions, further chemical shift mapping involving both a new RNA sequence and increased salt concentration was utilised in conjunction with the assignments made in Section 3.3.2.

3.3.4.1 Molecular characterisation of the binding of BHD to an unrelated RNA sequence

To gain further insight into how BHD recognises different RNA sequences, binding to an unrelated RNA transcript, 12AG (AAGGGAAAGGAA), was tested by ^{1}H-^{15}N chemical shift mapping. This RNA is predicted to be unstructured, unlike BRE19nt which is predicted to form a hairpin loop. ^{15}N-HSQC spectra were acquired of BHD alone and with two molar equivalents of 12AG. These spectra are shown annotated with residue assignments and overlaid with BHD:BRE19nt spectrum in Figure 3.7(A). Chemical shift changes upon RNA addition were quantitated using Equation 3.1, and plotted as a function of residue number. These spectra are shown overlaid with chemical shifts calculated for BHD:BRE19nt in Figure 3.8(A).

Chemical shift changes of BHD upon binding of 12AG were overwhelmingly in the same direction but of a lesser magnitude than the changes observed upon titration with BRE19nt, indicating that overall the interaction mode is likely to be very similar but that the interaction with 12AG is likely to be weaker in affinity and/or the resulting complex less well ordered. Overall, the mean chemical shift change,

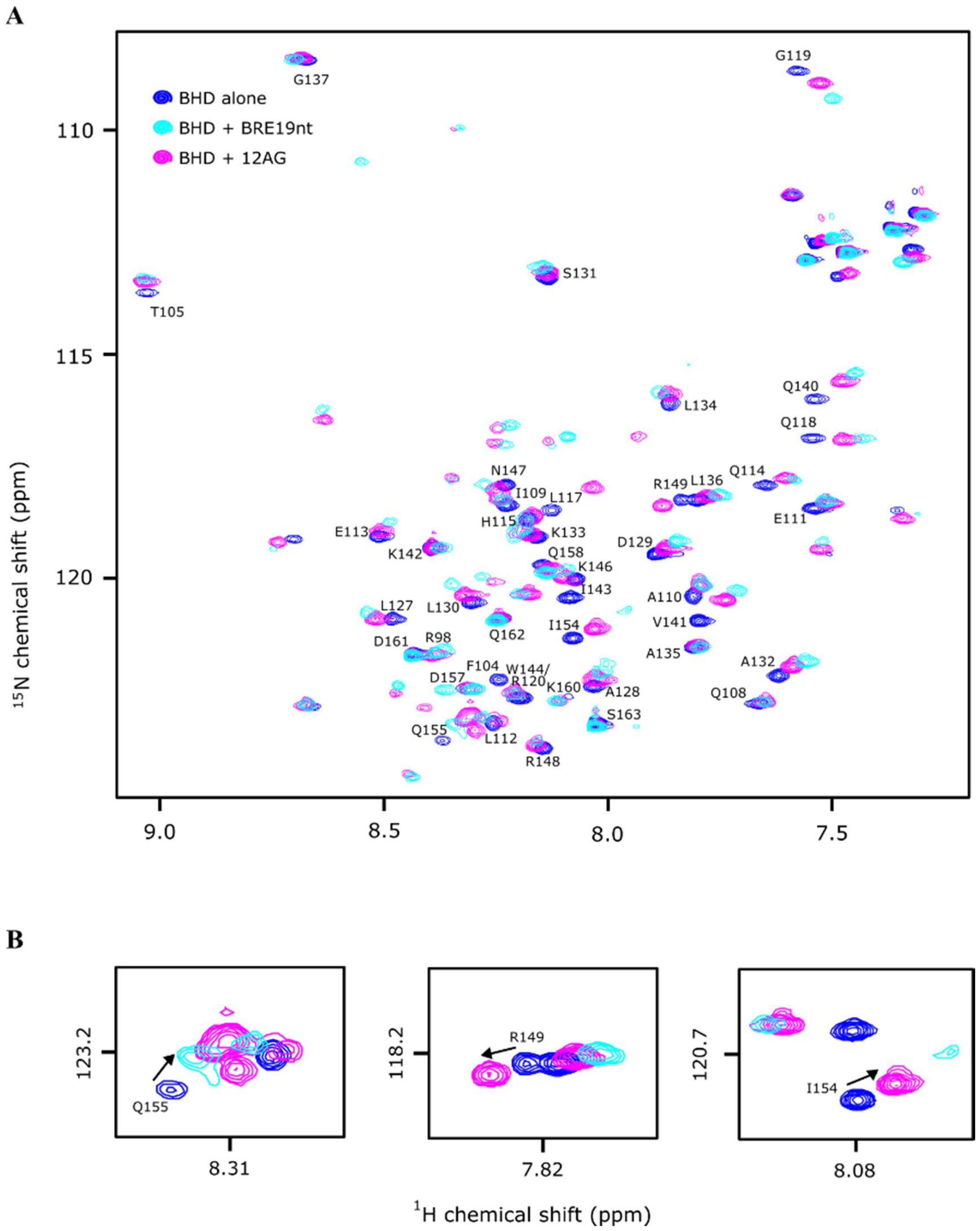
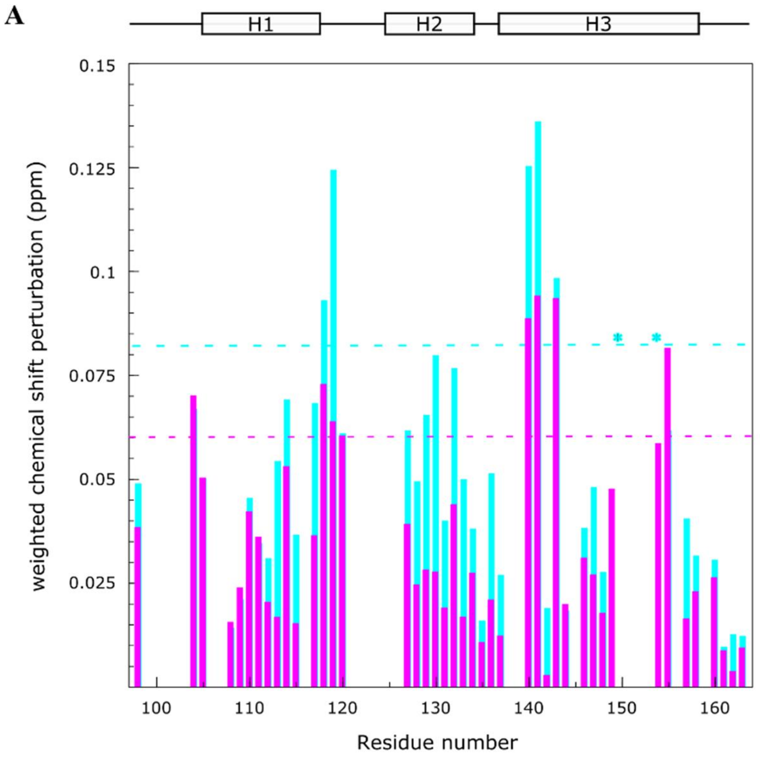


Figure 3.7. Overlay of ^{15}N-HSQC spectra of in the presence and absence of RNA.
(A) Overlay of ^{15}N-HSQC spectra of BHD alone (*blue*), BHD with 2 molar equivalents of BRE19nt (*cyan*) and BHD with 2 molar equivalents of 12AG (*magenta*). Starting concentration of BHD was 200 µM, in 50 mM HEPES pH 7.3, 50 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP. Spectra were recorded on an 800 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe (NS = 3 for protein alone and NS = 6, 1TD = 256) at 298 K. **(B)** Specific residues of interest; likely direction of change is indicated by arrows.

$\Delta\delta_{mean}$, for 12AG was 0.035 ppm, compared to 0.051 ppm for BRE19nt. Signals in the BHD:12AG ^{15}N-HSQC spectrum were often more intense compared with the corresponding signals in the ^{15}N-HSQC of



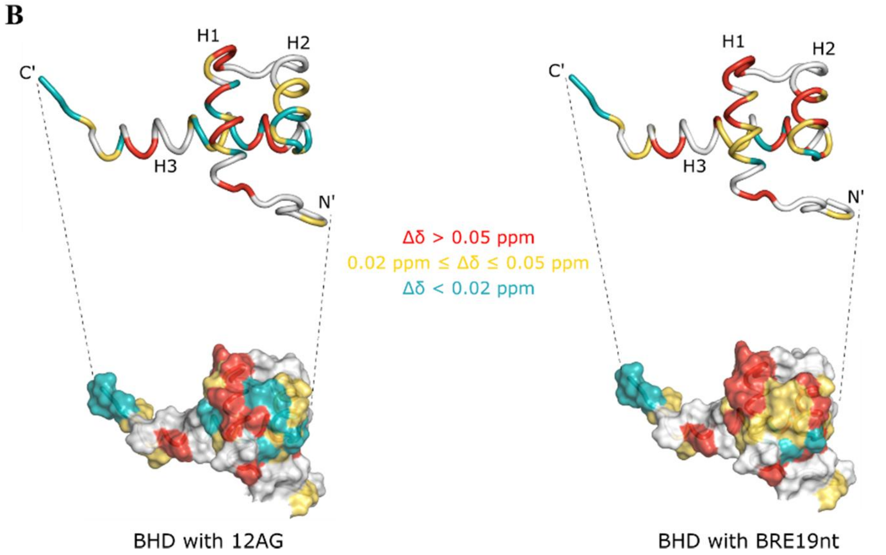


Figure 3.8. 12AG versus BRE19nt induced chemical shift changes of BHD backbone amide residues.
(A) Weighted chemical shift changes of BHD residues plotted as a function of residue number following the addition of 2 molar equivalents of 12AG (*magenta*) or BRE19nt (*cyan*). Horizontal dashed lines indicate the threshold level of significance at one standard deviation above the average chemical shift change. α-helices of BHD (when bound to DNA) are indicated above the graph, and are derived from the published structure of the BHD-DNA complex (PDB: 1ZQ3). **(B)** BHD with RNA-induced chemical shift changes for 12AG (left) and BRE19nt (right) mapped on to structure BHD (PDB:1ZQ3), colour coded by weighted average chemical shift ($\Delta\delta > 0.05$ ppm, *red*; 0.02 ppm $\leq \Delta\delta \leq$ 0.05 ppm, *yellow*; $\Delta\delta < 0.02$ ppm, *teal*; residues not observed in unbound state, *grey*).

the BHD:BRE19nt complex, indicating that there is a difference in the nature of the exchange processes between the samples.

It can be seen in Figure 3.8 that the regions of the domain that display the largest chemical shift changes are conserved between the 12AG and BRE19nt complexes, namely helix three and the loop between helices one and two. The biggest differences between the two complexes are observed for helix two, with no residues in this region in the 12AG complex displaying a chemical shift greater than 0.05 ppm. Helix two contains four leucine residues that contribute to the hydrophobic core of the domain [122], two of which display a shift change greater than 0.05 ppm upon formation of the BRE19nt complex. This may suggest that BHD does not undergo the same degree of conformational change or tightening up upon formation of a complex with 12AG as with BRE19nt.

It is notable that although both R149 and I154 disappear with BRE19nt addition (implicating them in RNA binding), the observed changes for these residues were much less pronounced following treatment of BHD with 12AG. Both signals did undergo some chemical shift change ($\Delta\delta_{I154} = 0.058$ ppm, $\Delta\delta_{R149} = 0.047$ ppm) but were still very intense (Figure 3.7(B), middle and right)). Further, Q155 in helix three is notable in that it displayed a greater shift for 12AG ($\Delta\delta_{Q155} = 0.081$ ppm) than for BRE19nt ($\Delta\delta_{Q155} = 0.061$ ppm) (Figure 3.7(B), left)). These differences are perhaps an indication that there may be some degree of flexibility in recognition mechanisms for different RNA sequences.

3.3.4.2 Molecular characterisation of BHD:BRE19nt under physiological salt conditions

In Chapter 2 it was shown that increasing the salt concentration in the binding buffer from 50 mM to 150 mM increased the dissociation constant of the BHD and BRE19nt interaction from ~4 μM to ~26 μM. This is a marked reduction in affinity and calls into question the biological relevance of the interaction, given that 150 mM salt is no higher than physiological concentration. In order to gain further insight into the ionic strength dependence of the BHD:RNA interaction, ^{15}N-HSQC spectra of BHD in the presence and absence of two molar equivalents of BRE19nt were acquired in 150 mM KCl. These spectra are shown in Figure 3.9(A).

There were fewer amide signals in the protein alone spectrum in 150 mM salt compared to the spectrum recorded in 50 mM salt (38 versus 44, respectively), indicating that some signals have been lost due to line broadening. An increase in salt concentration reduces the strength of electrostatic interactions and may be altering the rate of dynamic exchange processes taking place within the protein.

Signal changes upon RNA addition were modest compared with samples in 50 mM KCl; only a small number of signals shifted and no new signals appeared. Most of the residues that shift are hydrophobic, namely I154, L117, V141, F104 and I143. The chemical shift changes observed for non-hydrophobic residues were significantly smaller, with Q118, Q140, R149 and G119 displaying markedly reduced

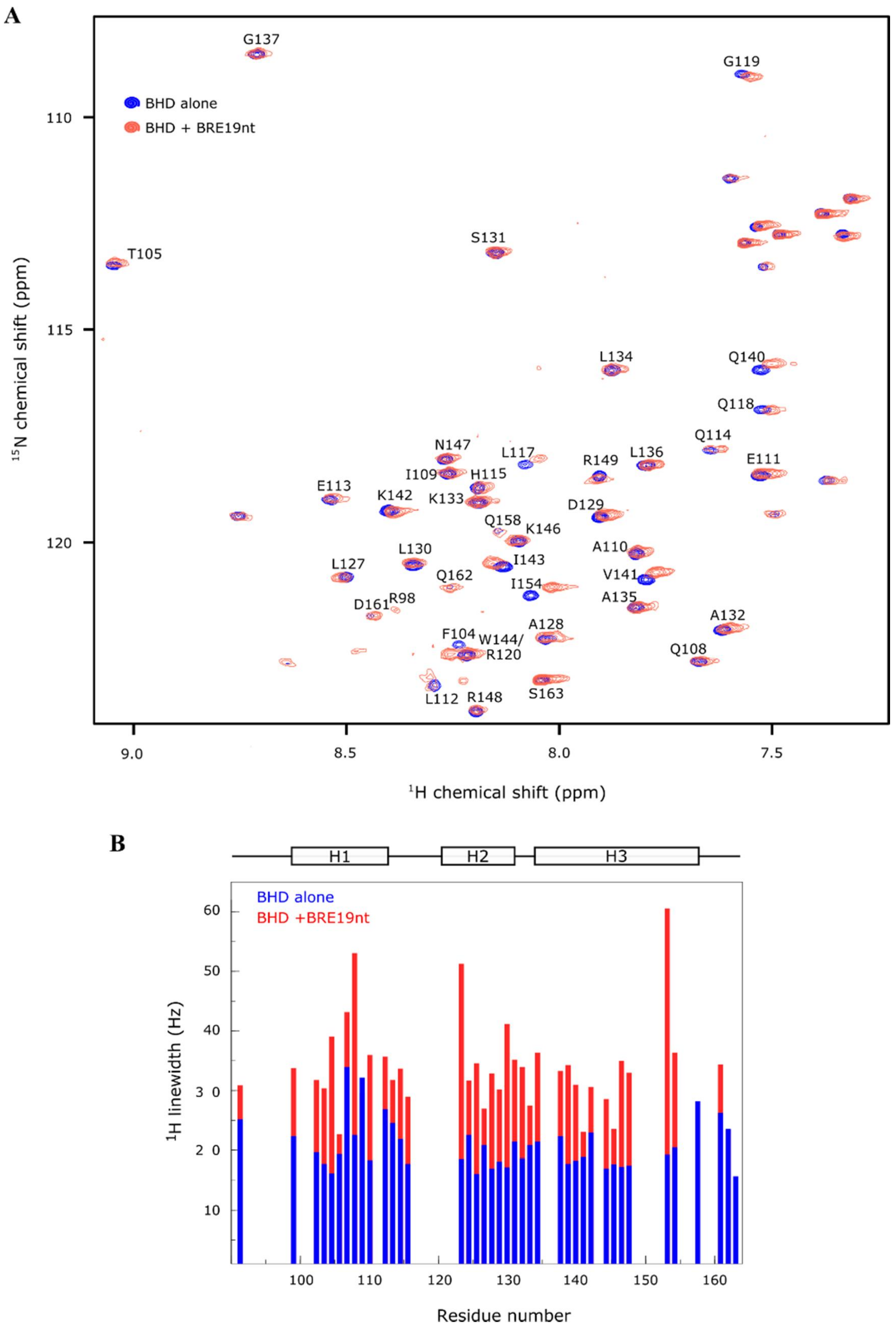


Figure 3.9. ^{15}N-HSQC spectral changes of BHD in the presence and absence of BRE19nt in 150 mM KCl. Overlay of ^{15}N-HSQC spectra of BHD alone (*blue*) and BHD with 2 molar equivalents of BRE19nt (*orange*). Starting concentration of BHD was 300 µM, in 50 mM HEPES pH 7.3, 150 mM KCl, 1 mM MgCl$_2$, 1 mM TCEP. Spectra were recorded on an 800 MHz Bruker Avance III NMR spectrometer equipped with a cryoprobe (NS = 3 for protein alone and NS = 6, 1TD = 256) at 298 K. **(B)** ^{15}N-HSQC ^{1}H linewidths of BHD with (*red*) and without (*blue*) BRE19nt in 150 mM NaCl, plotted as a function of residue number. α-helices of BHD (bound to DNA) are indicated above the graph, and are derived from the published structure of the BHD-DNA complex (PDB: 1ZQ3).

changes. This observation is consistent with electrostatic attraction between BHD and BRE19nt constituting a major driving force for the interaction.

Of note, there is both global and selective line broadening observed upon RNA addition. Linewidths were obtained in Sparky by integrating the signals, and plotted as a function of residue number, shown in Figure 3.9(B). These changes indicate that an interaction is definitely taking place between BHD and BRE19nt, albeit a substantially weaker one.

3.4 Summary and Discussion

This Chapter provides evidence that BHD is able to bind different RNA sequences in a primarily electrostatically driven manner. The backbone resonances of the domain were assigned using standard triple resonance NMR experiments, which facilitated the identification of residues that are likely to be involved in binding RNA. Overall, the regions of secondary structure of the domain used to recognise RNA appear to be the same as those in DNA recognition. BHD recognition of DNA involves the loop between helices one and two, the N-terminal arm and helix three. The data here show that this is also likely to be the case for RNA recognition. Helix three contains the most perturbed residues, presumably reflecting its role as the primary recognition helix, followed by the loop between helices one and two.

Some residues were implicated in the binding of both RNA and DNA. However, given that some residues were unable to be assigned, future studies should aim to obtain more complete assignments of BHD in the RNA-bound state in order to make a full characterisation and comparison of the residues involved. If complete BHD assignments in the unbound state cannot be obtained through the trialling of different sample conditions such as a lowering of temperature, then assignment of BHD residues in the RNA-bound state should be carried out, achieved by acquiring triple resonance spectra of BHD bound to RNA.

BHD was then shown to recognise a less chemically diverse RNA in a similar but weaker fashion. The ability of BHD to bind different RNA sequences, as well as the observation that ~30% of signals were missing in the BHD alone ^{15}N-HSQC spectra, may be reflections of the conformational dynamics of this domain and the flexibility it displays in its nucleic acid binding interface [122, 198, 199]. The nuances observed in RNA recognition mechanisms by BHD might not manifest as large differences in *in vitro* affinities, but are likely to be significant if BHD does indeed bind RNA *in vivo*. Further analysis, particularly solving structures of RBDs bound to different sequences, will contribute to a more comprehensive understanding of how RBPs recognise their biological targets in the presence of many ‘decoy sequences’ in the cell.

Finally, the effects on binding affinity of BHD for RNA produced by an increase in salt concentration was observed at the residue level, manifesting as a marked reduction in chemical shift changes with less contribution from charged residues, providing corroborating evidence that this interaction is driven primarily by electrostatic attraction. Further, the substantial line broadening observed indicates that the salt ions are significantly altering the rate of exchange between BHD and RNA.

Overall, these data provide more evidence that BHD is a promiscuous binder of RNA *in vitro*, however, the biological relevance of the observed RNA-binding of this domain remains unproven. Based on the data presented so far in this Thesis, it appears that BHD does not bind RNA in a specific manner in isolation. The next Chapter will consider the domain in the context of the full length protein in order to determine if there are other elements within the protein that might contribute to the biological RNA-binding specificity of this domain.

Chapter 4: Further investigations into the RNA-binding capacity of bicoid

4.1 Introduction

Data presented previously in this Thesis suggest the possibility that bicoid requires elements outside the homeodomain that have not been considered so far in order to achieve biological specificity in its RNA-binding capability. This Chapter considers features of the full-length protein to determine if specificity might be provided in this manner. The role of potential dimerisation of bicoid is investigated using EMSA, and a putative RRM in bicoid is produced with the aim of assessing its RNA-binding potential in a similar manner. Finally, the potential role of intrinsic disorder and repetitive regions present in bicoid is analysed in the context of the recent scientific literature, and particular consideration is given to RBPs that are components of phase separated bodies and how this affects the *in-vitro* study of such molecules.

4.2 Techniques used in this Chapter

4.2.1 Cy5 labelled RNA EMSAs

EMSAs were introduced in Section 2.2.1. Unlike the EMSAs presented in Chapter 2, however, the EMSAs described in this Chapter were carried out using fluorescently labelled RNA instead of ^{32}P labelled RNA. The fluorophore label is Cy5; production of the labelled RNA involved an *in-vitro* transcription reaction in which a fraction of the total UTP (~10%) in the reaction is replaced by 5-[3-aminoallyl]-2'-uridine-5'-triphosphate (aminoallyl UTP). Amine groups thereby incorporated into RNA transcripts are then reacted with amine-reactive Cy5.

Given the relatively larger size of the Cy5 label compared with ^{32}P, this technique is suited to longer RNA transcripts where the fraction of UTPs that incorporate a Cy5 label can be kept low, in order to reduce potential interference of the label with the RNA structure and/or binding of the RNA to partners.

This Chapter uses an RNA Pentaprobe in EMSAs to assess RNA-binding potential. Pentaprobes are a set of twelve 100-nt RNA sequences that were designed to together encompass all possible five-base combinations, making them a tool for the rapid detection of potential RNA-binding capacity of proteins [213]. Given the effects of RNA secondary structure, Pentaprobes will not contain every five-base sequence in every structural context, but because RBPs can often bind suboptimal binding sites, there is a high chance that most single-stranded RBPs will recognise one or more of the Pentaprobes. For example, the RRM of Fox-1 was shown to bind all twelve Pentaprobes [213].

4.3 Does bicoid homeodomain bind RNA as a dimer?

As discussed in Section 1.2.4, there is ample evidence that the bicoid homeodomain is involved in *cad* translational silencing, but the only evidence that the domain alone binds *cad* with appreciable specificity is the EMSA published by Chan and Struhl (1997) [126], which is reproduced in Figure 4.1. This EMSA shows a tight interaction between GST-tagged bicoid homeodomain and the 343-nt BRE RNA (K_d of ~ 50 nM). Because GST is known to dimerise [214], it is therefore possible that this interaction comprises a dimer of the homeodomain bound to BRE.

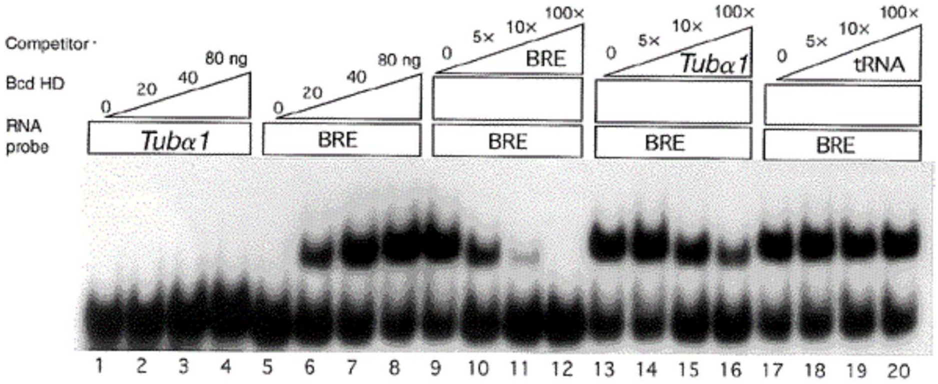


Figure 4.1. GST-BHD binds specifically to BRE.
Radiolabelled BRE or *Tubα1* 3′-UTRs were incubated with different amounts of GST-BHD, with or without cold competitor RNAs as indicated, demonstrating that BHD binds BRE but not *Tubα1*. The highest quantity of 80 ng equates to ~200 nM GST-BHD. This data is taken from Chan and Struhl (1997).

Supporting this possibility is evidence that bicoid binds DNA cooperatively in yeast and *in vitro*. It binds as a monomer to high affinity sites, and forms pairwise cooperative interactions once bound to DNA to recruit another bicoid molecule to a lower affinity site [215, 216] (Figure 4.2(A)). The homeodomain alone does not bind DNA cooperatively; however, the homeodomain is required for cooperative interactions, and the homeodomain residues S106, A124, S131, and K153 have been shown to be important for cooperativity. Binding isotherms for A124T, S131T, K153R and the wild type homeodomain with DNA containing either either three strong sites (SSS), two strong sites surrounding a weak site (SWS), or two strong sites with a non-binding spacer sequence (SXS) are shown in Figure 4.2(A)) [216]. These residues involved in cooperativity are shown mapped onto the structure of BHD bound to DNA in Figure 4.2(B). Sequences flanking either side of the homeodomain are also required for DNA-binding cooperativity, likely to facilitate intermolecular bicoid interactions [217].

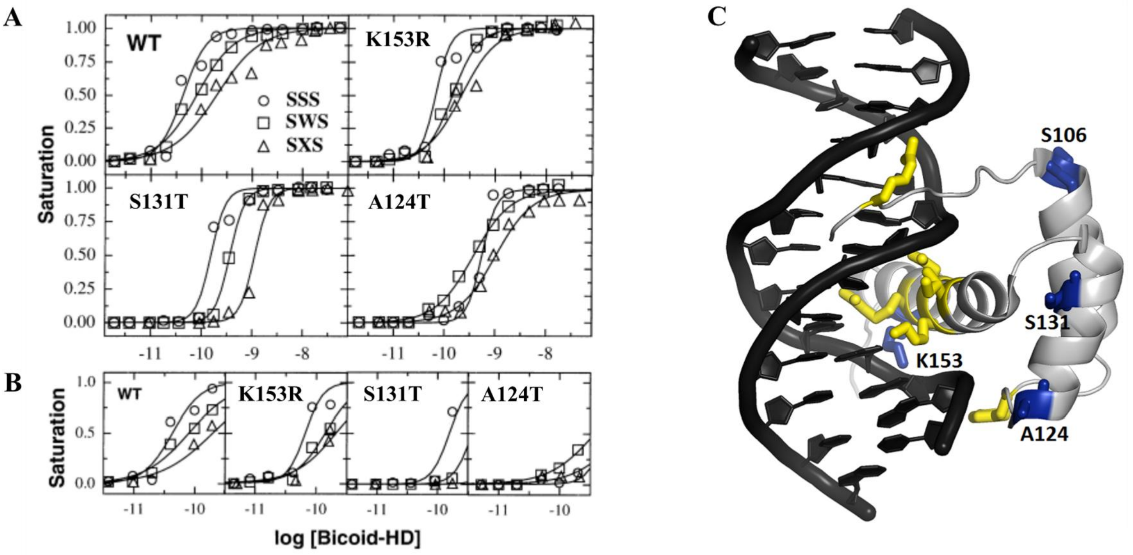


Figure 4.2. Bicoid binds to DNA cooperatively.
(A) Binding isotherms (calculated from EMSA data) of bicoid homeodomain and mutants binding to either three strong sites (SSS), two strong sites surrounding a weak site (SWS), or two strong sites with a non-binding spacer sequence (SXS). This data, as well as beta-galactosidase activity assays that incorporated these binding sites into a lacZ reporter gene, indicate bicoid homeodomain cooperativity of binding, as well as a reduction of this cooperativity by the S106C, A124T, S131T and K153R mutants. These data were reported by Burz and Hanes (2001). **(B)** Expansion of the first part of the binding curves from (A). **(C)** The homeodomain residues involved in cooperativity are shown mapped on to the BHD structure bound to DNA (PDB: 1ZQ3). Binding residues are shown in yellow and cooperativity residues are shown in blue and numbered according positions in the full length protein.

In order to determine whether or not dimerisation of bicoid might be facilitating the high-affinity interaction between GST-BHD and BRE observed in Figure 4.1, GST-BHD was purified and tested for binding to the *cad* 3′-UTR by Cy5 labelled RNA EMSAs.

4.3.1 Expression and purification of GST-BHD

GST-BHD was expressed and purified in the same manner as for the cleaved protein in Section 2.3.1 up to the GSH affinity purification step, at which point the stringency of the wash buffer and number of wash steps were increased in order to optimise the purity of GST-BHD. Elutions two to seven were pooled (Figure 4.3) and dialysed into EMSA buffer (20 mM HEPES pH 7.5, 50 mM KCl, 1 mM DTT) ready for use in EMSAs; ~1.2 mg of purified GST-BHD was obtained per litre of culture at >95% purity.

4.3.2 Testing binding of GST-BHD to *cad* 3′-UTR

In order to determine if GST-BHD binds the *cad* 3′-UTR with a greater affinity than BHD, EMSAs were carried out with GST-BHD and Cy5 labelled RNA.

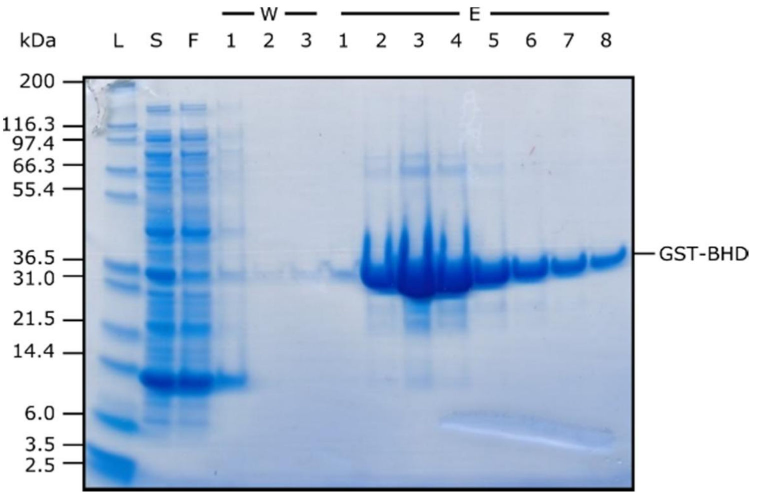


Figure 4.3. Overexpression and affinity purification of GST-BHD.
SDS-PAGE analysis of GST-BHD purification. Abbreviations: L, Mark 12 ladder; S, soluble fraction; F, flow through from GST-beads; W, wash steps (numbers as indicated); E, GSH elutions (numbers as indicated).

A different approach to *cad* 3′-UTR production was taken in comparison to that presented in Chapter 2. This time the *cad* 3′-UTR was divided into three overlapping fragments (Figure 4.4(A)), named according to the number of nucleotides in each sequence. The reasons for this were twofold, and founded on discussion points raised in Section 2.5.2. Firstly, there were some concerns over obtaining a full length *cad* 3′-UTR from *in-vitro* transcription and purification methods, and there was no straightforward method by which to determine the size of this RNA, as is commonly done in non-radiolabelled gels. RNA species of >800 nt in length could be too large to enter the pores of polyacrylamide gel matrices, and therefore it is possible that the full length *cad* 3′-UTR purified may have been a truncation. Secondly, given that there has been some uncertainty regarding the location of the specific target site within *cad* (see Section 1.2.4), if the bicoid homeodomain does indeed bind outside the BRE then these interactions may have been missed as all tested sequences (except for the full length 3′-UTR) were contained within the BRE. The sequences designed were: cad411, which contains most of the bicoid binding region defined by Rivera-Pomar *et al.* [87]; cad525, which contains the BRE; and cad186, which contains three repeats of an AU-rich element (AUUUA). AU-rich elements are sequence motifs that have been implicated in AU-mediated mRNA decay [218]. A small amount (~20 bases) of the 5′ end of the *cad* 3′-UTR was not included, due to primer design considerations.

Internal Cy5 labelling of RNA was performed instead of the ^{32}P labelling used in Chapter 2, as noted above. This procedure allowed the transcripts to be sized on a gel, and it is also a more practical labelling technique given that the RNAs can be stored for use for longer periods of time without the signal from the label deteriorating. Templates for transcription were amplified through the use of primers containing a T7 promoter with regions complementary to the relevant *cad* sequence, as per Chapter 2. Internal labelling of transcripts with Cy5 was achieved by incorporating amino allyl UTP (~1:10 molar ratio of amino allyl UTP to normal UTP) in *in-vitro* run-off transcription reactions; the product was extracted, labelled with Cy5 and then purified using a PureLink® RNA Mini Kit.

Binding of GST-BHD was tested to cad411, cad525 and cad186 by EMSA (Figure 4.4(B)). GST-BHD bound all probes similarly, with the majority of the bound complex failing to either migrate out of the wells or be resolved as distinct bands within the gel. Instead, protein-bound RNA was observed in the wells, or as smears indicative of binding in the gel. Despite using the same binding conditions as Chan and Struhl (1997) [126], their GST-BHD EMSA results were not able to be replicated here. Instead, GST-BHD binds *cad* RNA transcripts with a similar affinity to BHD as presented in Chapter 2.

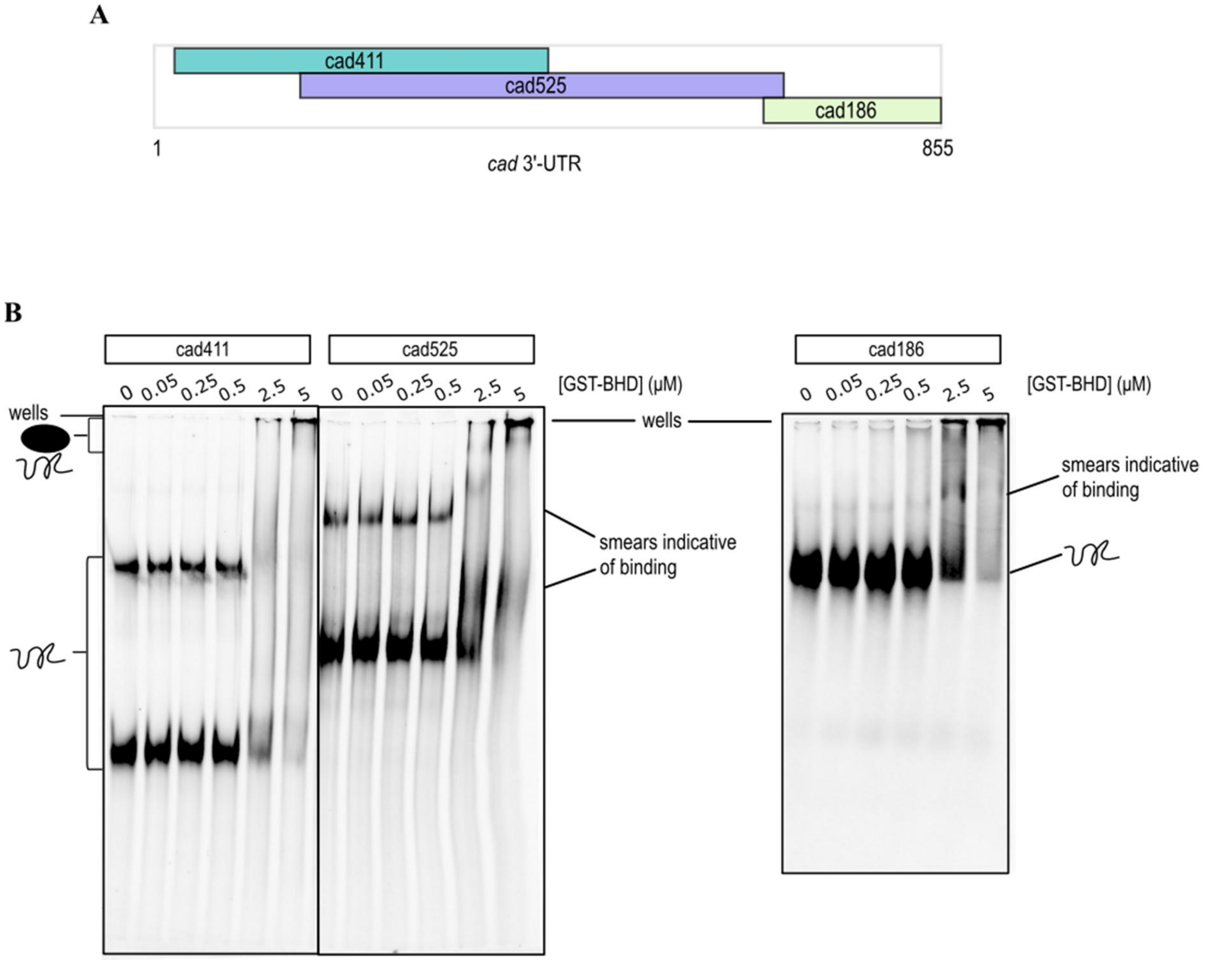


Figure 4.4. EMSAs of GST-BHD with *cad* 3′-UTR sequences.
(A) The 855-nt *cad* 3′-UTR was divided into three overlapping fragments, as indicated here. Numbers indicate length of RNA transcripts. **(B)** EMSAs demonstrating GST-BHD binding of cad411, cad525 and cad186. Increasing concentrations of GST-BHD were incubated with Cy5-labelled RNAs then resolved on a 6% polyacrylamide native gel.

It is difficult to explain the inconsistencies between these results. Chan and Struhl (1997) expressed their homeodomain construct in *E. coli* as is the case here, although they did use a construct that encoded the 60 amino acid homeodomain plus three N-terminal amino acids (Met, Gly and Arg). The BHD construct used in this Thesis consists of the 60 amino acid homeodomain plus seven C-terminal amino acids added for solubility, as per Figure 2.2. Ubx was shown in Section 2.5.1 to have unstructured regions that increase the selectivity of its homeodomain for DNA. It is therefore possible, but perhaps unlikely, that the seven C-terminal amino acids contained in BHD have the effect of reducing the affinity of the domain for RNA.

Chan and Struhl (1997) did not describe their method of RNA production in detail, besides noting that the RNA sequence contained the 343-nt BRE and was produced by *in-vitro* transcription. Unsuccessful attempts were made here to amplify just the BRE sequence as a template for *in-vitro* transcription. Failure to do so was likely attributable to suboptimal primer sequences.

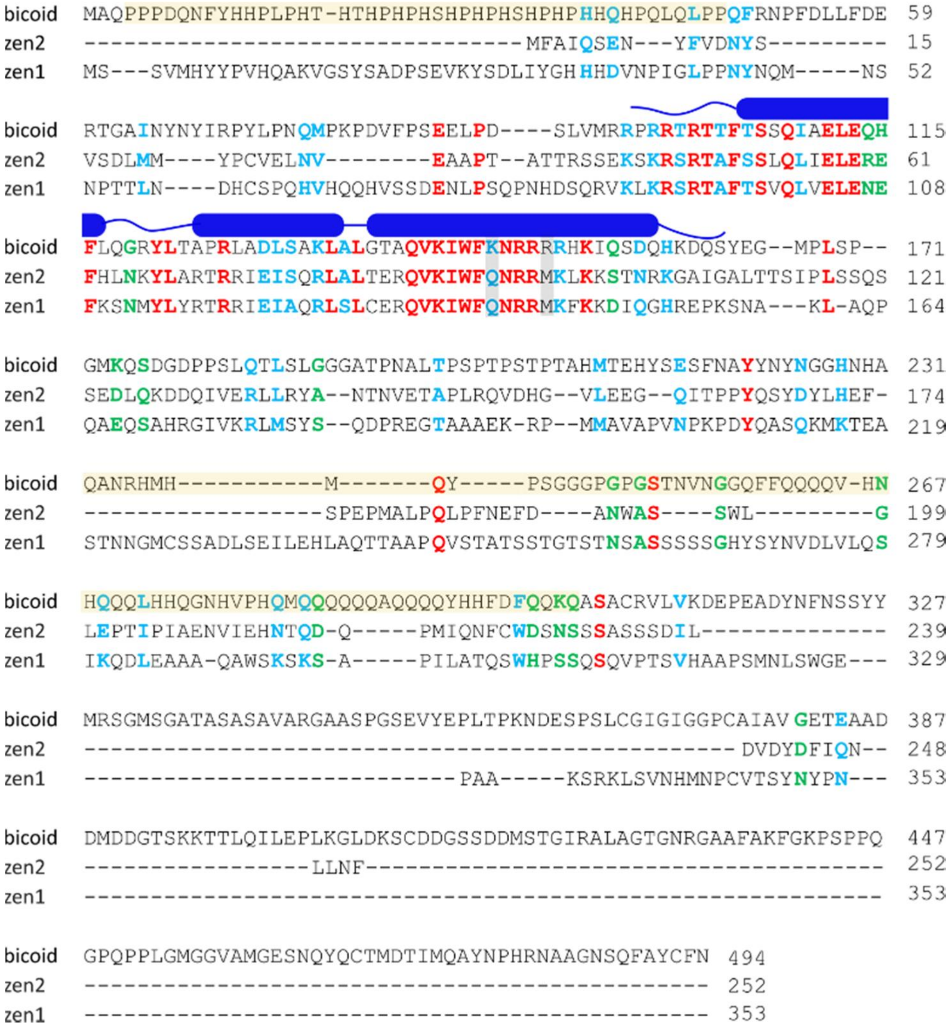
4.4 Bioinformatic analysis of bicoid

The inability to reproduce the only published work that demonstrates specific binding of the bicoid homeodomain to *cad* calls into question the validity of this result. At this point, the possibility that *in vivo* specificity is provided by elements outside the homeodomain must be considered. The presence of extra specificity determinants outside the homeodomain does not contradict previous results that indicate bicoid-mediated regulation of *cad* is dependent on elements contained within the homeodomain [87, 105]. To examine the possibility that other protein features may affect specificity, a bioinformatic analysis of bicoid will be presented here.

4.4.1 Sequence features of full length bicoid

As mentioned in Section 1.2.2, *bicoid* is the product of a gene duplication event. Specifically, the homeobox gene *Hox3* was duplicated in higher dipterans, resulting in the paralogs *bicoid* and *zerkneullt* (*zen)* [219]. In lower dipterans, the single *Hox3* gene is closer in sequence to *zen* than *bicoid*. In *D. melanogaster*, the *zen* gene duplicated again, and therefore there are two *zen* genes. The homeodomain is the only region that displays high conservation overall, as shown in Figure 4.5. However, within the homeodomain there have been two key mutations that have been pivotal to bicoid acquiring both RNA-binding capability (the M54R mutation) [123] and a change in DNA-binding specificity (the Q50K mutation) [220]. Despite the requirement of R54 for bicoid-mediated repression of *cad* translation [123], there still may be other elements outside of the homeodomain that are required for RNA target recognition.

Notably, bicoid has acquired two regions of low-complexity sequence: a histidine/proline repeat region at the N-terminus, and a glutamine/glycine repeat region C-terminal to the homeodomain (Figure 4.5). Low-complexity motifs are known to be enriched in both transcription factors and RBPs [221]. The high-throughput poly-A capture methods introduced in Section 1.1.2.2 have shown that mammalian mRNA-binding proteins (mRBPs) contain an overrepresentation of low-complexity, repetitive motifs [68, 69]. The same has also been shown recently for *Drosophila*, with glycine, glutamine and asparagine, in particular, shown to be common in repeat regions of RBPs [222, 223]. It is notable that bicoid was not detected as an RBP in either of these high-throughput RBP-ome studies in *Drosophila*.



```
bicoid  MAQPPPDQNFYHHPLPHT-HTHPHPHSHPHPHSHPHPHHQHPQLQLPPQFRNPFDLLFDE  59
zen2    ---------------------------------MFAIQSEN---YFVDNYS---------  15
zen1    MS---SVMHYYPVHQAKVGSYSADPSEVKYSDLIYGHHHDVNPIGLPPNYNQM-----NS  52

bicoid  RTGAINYNYIRPYLPNQMPKPDVFPSEELPD----SLVMRRPRRTRTTFTSSQIAELEQH  115
zen2    VSDLMM----YPCVELNV--------EAAPT--ATTRSSEKSKRSRTAFSSLQLIELERE  61
zen1    NPTTLN----DHCSPQHVHQQHVSSDENLPSQPNHDSQRVKLKRSRTAFTSVQLVELENE  108

bicoid  FLQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEG--MPLSP--  171
zen2    FHLNKYLARTRRIEISQRLALTERQVKIWFQNRRMKLKKSTNRKGAIGALTTSIPLSSQS  121
zen1    FKSNMYLYRTRRIEIAQRLSLCERQVKIWFQNRRMKFKKDIQGHREPKSNA---KL-AQP  164

bicoid  GMKQSDGDPPSLQTLSLGGGATPNALTPSPTPSTPTAHMTEHYSESFNAYYNYNGGHNHA  231
zen2    SEDLQKDDQIVERLLRYA--NTNVETAPLRQVDHG--VLEEG--QITPPYQSYDYLHEF-  174
zen1    QAEQSAHRGIVKRLMSYS--QDPREGTAAAEK-RP--MMAVAPVNPKPDYQASQKMKTEA  219

bicoid  QANRHMH-----------M-------QY-----PSGGGPGPGSTNVNGGQFFQQQQV-HN  267
zen2    ------------------SPEPMALPQLPFNEFD----ANWAS----SWL---------G  199
zen1    STNNGMCSSADLSEILEHLAQTTAAPQVSTATSSTGTSTNSASSSSSGHYSYNVDLVLQS  279

bicoid  HQQQLHHQGNHVPHQMQQQQQQAQQQQYHHFDFQQKQASACRVLVKDEPEADYNFNSSYY  327
zen2    LEPTIPIAENVIEHNTQD-Q-----PMIQNFCWDSNSSSASSSDIL--------------  239
zen1    IKQDLEAAA-QAWSKSKS-A-----PILATQSWHPSSQSQVPTSVHAAPSMNLSWGE---  329

bicoid  MRSGMSGATASASAVARGAASPGSEVYEPLTPKNDESPSLCGIGIGGPCAIAVGETEAAD  387
zen2    -------------------------------------------------DVDYDFIQN--  248
zen1    ----------------------------PAA-----KSRKLSVNHMNPCVTSYNYPN---  353

bicoid  DMDDGTSKKTTLQILEPLKGLDKSCDDGSSDDMSTGIRALAGTGNRGAAFAKFGKPSPPQ  447
zen2    -----------------LLNF---------------------------------------  252
zen1    ------------------------------------------------------------  353

bicoid  GPQPPLGMGGVAMGESNQYQCTMDTIMQAYNPHRNAAGNSQFAYCFN  494
zen2    -----------------------------------------------  252
zen1    -----------------------------------------------  353
```

Figure 4.5. Sequence alignment of *D. melanogaster* proteins bicoid, zen1 and zen2.
Conserved residues are shown in *red*, residues with strongly similar properties shown in *cyan* and residues with weakly similar properties shown in *green*. Q50K and M54R mutations, important in the evolution of DNA-binding specificity and RNA-binding capability respectively, are shaded *grey*. Low-complexity repeat regions are shaded *yellow*. The three α-helices, loop regions and terminal arms of the homedomain are mapped above the sequence in *blue*. Protein residue numbers are indicated on the right of each row. Alignment of sequences was done with Clustal Omega.

The roles of these low-complexity motifs are beginning, in some instances, to be elucidated. In certain cases, it appears that such repeat motifs can bind RNA directly. For example, RGG repeat motifs, introduced in Section 1.1.1.2, are a type of non-canonical RBD. Aside from directly mediating RNA contacts, low-complexity motifs are being increasingly implicated in RNP bodies [224]. These cellular bodies are a diverse group of membrane-less, phase-separated granules that are enriched in IDR containing RBPs and RNA, and will be discussed in further detail in Section 4.6.

The second low-complexity region in bicoid is a ~60 residue glutamine-rich domain that is part of a longer ~80 amino acid stretch that is predicted to be prion-like by PLAAC, a hidden-Markov model that detects probable prion-like sequences [225] (Figure 4.6). Prion-like behaviour in proteins encompass a range of properties that involve self-templating conformational change [226]. Prion-like domains (PLDs) are also prevalent in paraspeckle (another type of phase-separated RNP complex) proteins [227], and the relevance of these domains to paraspeckle phase separation is just beginning to be elucidated. For example, the prion-like domains of RBM14 and FUS have been shown to be required for paraspeckle formation [228].

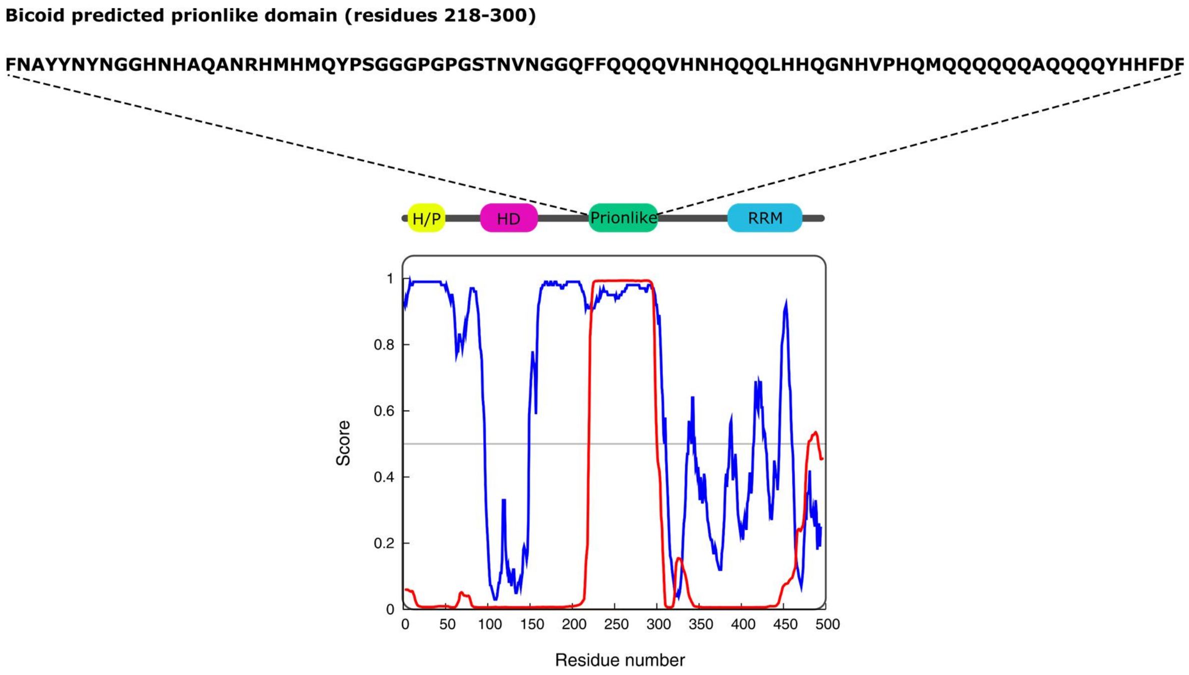
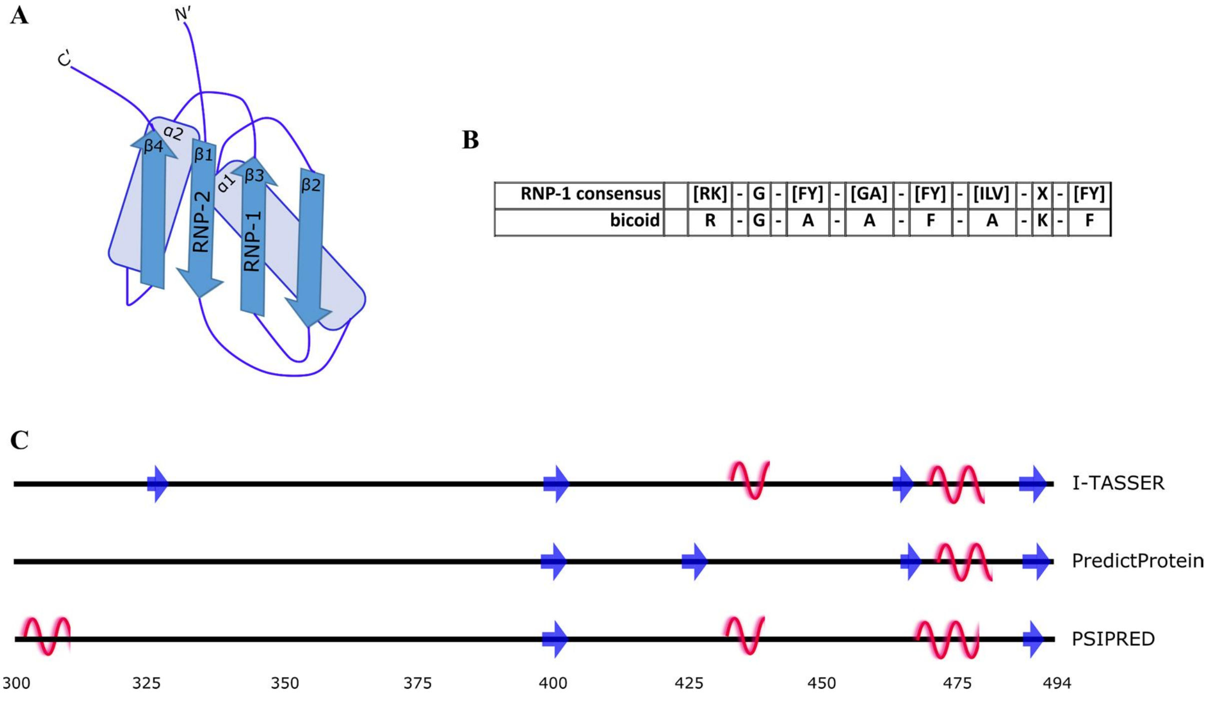


Figure 4.6. The predicted PLD within bicoid maps to a region of disorder.
The DISOPRED3 disorder prediction of bicoid is shown graphed in *blue* and the PLAAC prion-like domain prediction is shown in *red*. The threshold of 0.5 (indicated by a *grey* line) signifies that these amino acids are either likely to be disordered (*blue*) or part of a PLD (*red*). Low complexity sequences and putative RBDs of bicoid are indicated above the graph, as well as the amino acid sequence of the predicted PLD. Abbreviations: H/P, histidine/proline repeat region; HD, homeodomain; RRM, putative RNA-recognition motif (introduced in the next Section).

Both the bicoid PLD and the histidine/proline repeat map to regions of predicted disorder. Figure 4.6 shows a disorder prediction made by the DISOPRED3 algorithm [229]. There is now ample evidence that divergent IDRs are not only associated with RNP granules as mentioned above, but are specifically involved in the assembly of these granules [224, 230-235], particularly poly-glutamine repeats [236-238] and glutamine/asparagine rich sequences [239-241]. Further, both histidine repeats [242] and proline repeats [232] have been implicated in RNP granule formation. Therefore, it is possible that these predicted disordered regions in bicoid may be involved in the formation of some sort of phase-separated, bicoid-containing RNP complex.

4.4.2 Bicoid contains a putative RRM

RRMs were introduced in Section 1.1.2.1 as the most common RBD. The typical RRM fold, βαββαβ, with RNP-1 and RNP-2 motifs located on β-strands 3 and 1, respectively, is depicted in Figure 4.7(A). However, with only ~30% sequence conservation among RRMs [41] and the many different varieties of atypical domain folds [40], they are also possibly the most divergent RBD.



RNP-1 consensus	[RK]	-	G	-	[FY]	-	[GA]	-	[FY]	-	[ILV]	-	X	-	[FY]
bicoid	R	-	G	-	A	-	A	-	F	-	A	-	K	-	F

Figure 4.7. Bicoid contains a putative RRM.
(A) The typical RRM fold is depicted schematically with two α-helices packed against a four strand β-sheet; RNP-1 and RNP-2 consensus motifs are located on the two central β-strands. **(B)** The RNP-1 consensus motif is shown in comparison to the similar motif found in bicoid. **(C)** A range of secondary structure prediction tools (I-TASSER, PredictProtein and PSIPRED) were utilised to predict the structure of residues 300 – 494 of bicoid. Arrows indicate predicted β-strands, and α-helices are indicated by curves.

This diversity in sequence and structures of RRMs has made the discovery of RRM variants complicated. Pairwise alignments of divergent RRM sequences have often proven insufficient to detect RRMs due to the highly varying nature of sequence and structural elements that are recognised in the domain [44]. As a result, several different methods have been used to identify potential RRMs. Initially, searches were based on conservation of the RNP-1 and RNP-2 motifs, but given the degeneracy and sometimes absence of these motifs, the criteria were broadened to consider domain topology [243], for instance by looking for the presence of conserved residues in positions relative to subdomains, such as hydrophobic core residues or loop residues [44]. Even today, search criteria for subclasses of RRMs are not well defined [244], and variant domains that fit into the broad category of RRMs continue to be discovered. For example, xRRMs, with an atypical βαββαββα fold and a unique RNA-binding mechanism, have recently been classified; these domains are likely to be present solely in La and LARP7 proteins that bind RNAPIII transcribed ncRNAs [245].

It was first postulated that bicoid contains an RRM at its C-terminus (see Figure 4.6) several decades ago [246], given similarity to an RRM consensus sequence that was devised through the alignment of known RRMs at the time, and subsequent observation of conserved amino acid properties in certain positions, allowing for conservative substitutions [247]. This consensus sequence is shown in Figure 4.8 and is compared with some known RRMs; bicoid fits the consensus at 14 out of 32 highly conserved positions, and 10 out of 16 positions that are dictated by one or two specific residues. In particular, bicoid contains an eight-residue motif in this region that is similar to the RNP-1 motif (Figure 4.7(B)).

A range of secondary structure prediction tools (I-TASSER [248], PredictProtein [249], and PSIPRED [229]) predict some secondary structure in this region of bicoid, as shown in Figure 4.7(C). The secondary structure elements predicted by these software programs, however, do not indicate a likelihood that this protein sequence forms enough secondary structure to form an RRM fold.

The putative bicoid RRM (BRRM) does not readily appear to fit any current subclasses of RRMs based on primary sequence or predicted secondary structure. It may be that this region of bicoid constitutes a novel RRM. Alternatively, the domain may bind RNA without fitting the classification of an RRM, given that other proteins such as cold shock domains [250] and the bacteriophage protein T4gp32 [251] have the RNP-1 motif but not the RRM fold and have been shown to bind RNA.

Despite its identification some decades ago, no published studies have reported on the structure or the RNA-binding properties of this region of bicoid; it was therefore decided to determine if this domain might play a role in forming the bicoid complex with the *cad* 3′-UTR.

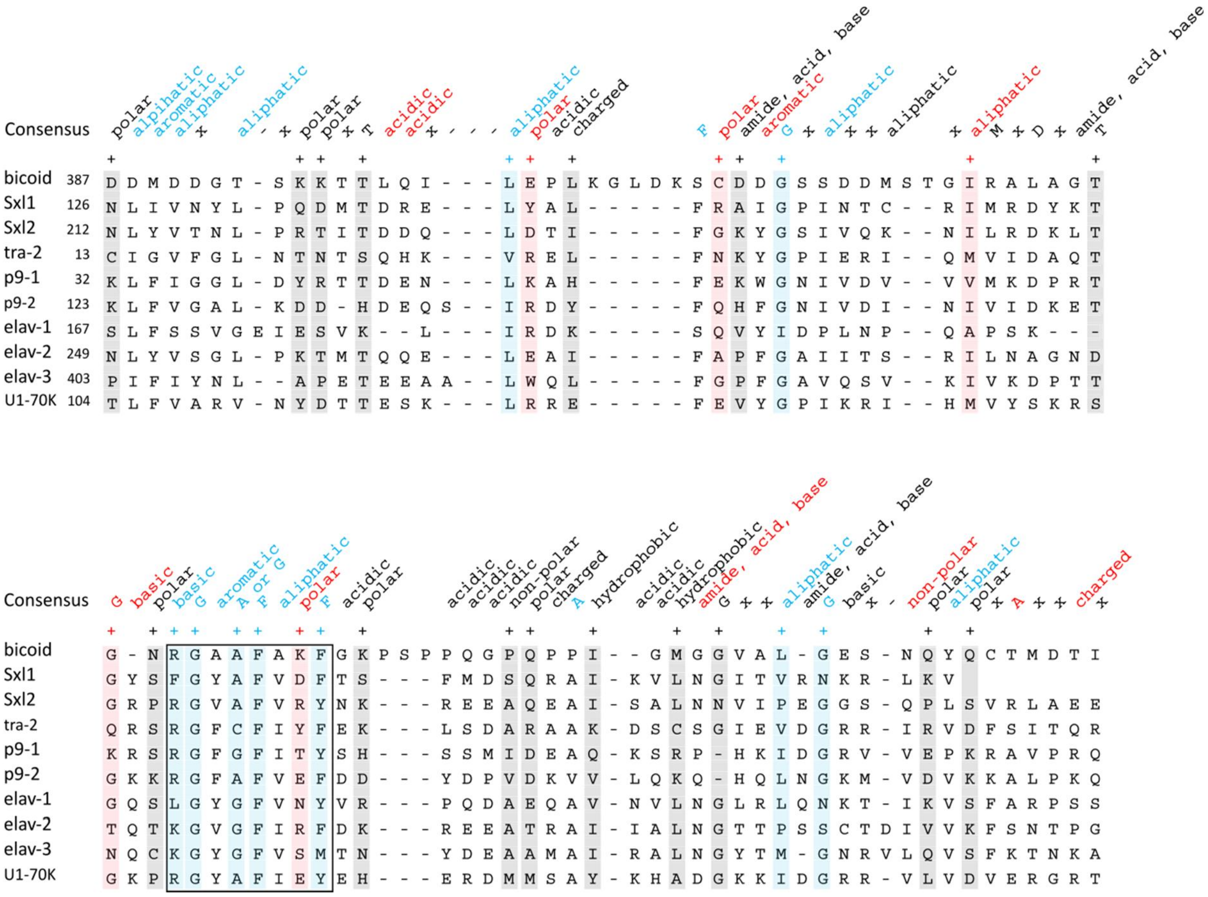

Consensus
polar alpihatic aromatic aliphatic x aliphatic - x polar polar x T acidic acidic x - - - aliphatic polar acidic charged F polar amide, acid, base aromatic G x aliphatic x x aliphatic x aliphatic M x D x amide, acid, base T
bicoid 387 DDMDDGT-SKKTTLQI---LEPLKGLDKSCDDGSSDDMSTGIRALAGT
Sxl1 126 NLIVNYL-PQDMTDRE---LYAL-----FRAIGPINTC--RIMRDYKT
Sxl2 212 NLYVTNL-PRTITDDQ---LDTI-----FGKYGSIVQK--NILRDKLT
tra-2 13 CIGVFGL-NTNTSQHK---VREL-----FNKYGPIERI--QMVIDAQT
p9-1 32 KLFIGGL-DYRTTDEN---LKAH-----FEKWGNIVDV--VVMKDPRT
p9-2 123 KLFVGAL-KDD-HDEQS--IRDY-----FQHFGNIVDI--NIVIDKET
elav-1 167 SLFSSVGEIESVK--L---IRDK-----SQVYIDPLNP--QAPSK---
elav-2 249 NLYVSGL-PKTMTQQE---LEAI-----FAPFGAIITS--RILNAGND
elav-3 403 PIFIYNL--APETEEAA--LWQL-----FGPFGAVQSV--KIVKDPTT
U1-70K 104 TLFVARV-NYDTTESK---LRRE-----FEVYGPIKRI--HMVYSKRS
Consensus
G basic polar basic G aromatic A or G F aliphatic polar F acidic polar acidic acidic acidic non-polar polar charged A hydrophobic acidic acidic hydrophobic amide, acid, base G x x aliphatic amide, acid, base G basic x - non-polar polar aliphatic polar x A x x charged x
bicoid G-NRGAAFAKFGKPSPPQGPQPPI--GMGGVAL-GES-NQYQCTMDTI
Sxl1 GYSFGYAFVDFTS---FMDSQRAI-KVLNGITVRNKR-LKV
Sxl2 GRPRGVAFVRYNK---REEAQEAI-SALNNVIPEGGS-QPLSVRLAEE
tra-2 QRSRGFCFIYFEK---LSDARAAK-DSCSGIEVDGRR-IRVDFSITQR
p9-1 KRSRGFGFITYSH---SSMIDEAQ-KSRP-HKIDGRV-VEPKRAVPRQ
p9-2 GKKRGFAFVEFDD---YDPVDKVV-LQKQ-HQLNGKM-VDVKKALPKQ
elav-1 GQSLGYGFVNYVR---PQDAEQAV-NVLNGLRLQNKT-IKVSFARPSS
elav-2 TQTKGVGFIRFDK---REEATRAI-IALNGTTPSSCTDIVVKFSNTPG
elav-3 NQCKGYGFVSMTN---YDEAAMAI-RALNGYTM-GNRVLQVSFKTNKA
U1-70K GKPRGYAFIEYEH---ERDMMSAY-KHADGKKIDGRR-VLVDVERGRT

Figure 4.8. Comparison of conserved features of putative bicoid RRM with other RRMs.
RRM consensus (shown above sequence alignment) devised by Query *et al.* (1989) through the alignment of RRMs and subsequent observation of conserved amino acid properties in certain positions, allowing for conservative substitutions. *Blue* indicates highly conserved positions, *red* indicates well conserved with the remainder being less well conserved, + indicates that bicoid conforms to the consensus (with residue positions shaded accordingly), single capital letters in the consensus header indicate a conserved amino acid based on standard single letter amino acid abbreviations, x indicates that there is no consensus feature for that position. All RRMs are from *Drosophila* except for U1-70K which is from human. The conserved RNP-1 octapeptide is outlined in a box.

4.5 Investigation of the RNA-binding properties of the putative RRM in bicoid

4.5.1 Expression and purification of BRRM

In order to ascertain the RNA-binding capacity of BRRM, recombinant protein production and purification was required. To assist in construct design, secondary structure elements were predicted by a variety of online tools, as shown in Figure 4.7(C) (for secondary structure prediction of the full length protein, see Appendix A.5). No reliable tertiary structures were predicted by I-TASSER [248], when various fragments of the C-terminal portion of bicoid were used as input (data not shown). The construct was designed to include all predicted secondary structural elements in this region of the protein, with eight N-terminal residues in addition to the consensus sequence (Figure 4.9). BRRM was cloned from full-length bicoid contained in the vector pET16b, provided by Michalis Averof of Institut de Génomique Fonctionnelle de Lyon (IGFL), into pGEX-6P. BRRM was thus expressed as a GST fusion, and then cleaved from the GST tag with HRV-3C to yield a 127 amino acid protein.

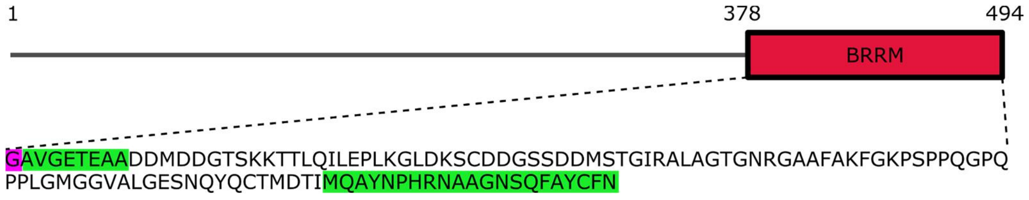

1
378
494
BRRM
GAVGETEAADDMDDGTSKKTTLQILEPLKGLDKSCDDGSSDDMSTGIRALAGTGNRGAAFAKFGKPSPPQGPQ
PPLGMGGVALGESNQYQCTMDTIMQAYNPHRNAAGNSQFAYCFN

Figure 4.9. Amino acid sequence of BRRM produced to test RNA-binding.
The 127 amino acid recombinant protein comprises residues 378 to 494 of the 494 amino-acid protein bicoid. This construct contains residues 387 to 473, which is the sequence that bears resemblance to an RRM, plus an extra eight N-terminal residues and 21 extra C-terminal residues highlighted in green. An N-terminal glycine (purple) will remain following HRV-3C cleavage.

Expression of the recombinant protein in *E. coli* Rosetta(DE3)pLysS cells resulted in a large amount of soluble protein, which was purified by GSH affinity purification. The recombinant GST-fusion protein appeared to run as two different bands on an SDS-PAGE, as seen in Figure 4.10.

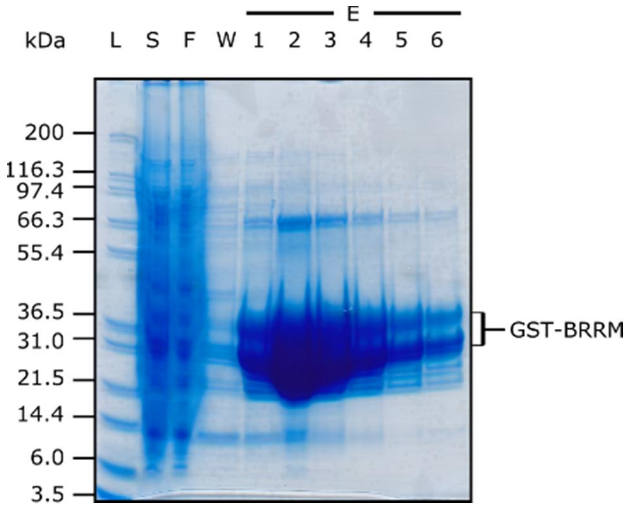


Figure 4.10. Overexpression and affinity purification of GST-BRRM.
SDS-PAGE analysis of GST-BRRM overexpression and affinity purification. Abbreviations: L, Mark 12 ladder; S, soluble fraction; F, flow through from GST beads; W, wash; E, GSH elutions, numbers as indicated.

GSH elutions one to six were pooled, and the GST tag was cleaved efficiently by HRV-3C (Figure 4.11(A)). The resulting cleaved protein was purified by sizing exclusion chromatography on a

Superdex 75 column. BRRM contains no tryptophan residues, and therefore has a low extinction coefficient of 4470 M^{-1} cm^{-1}. As a result, the protein did not appear as a clear peak on the size exclusion chromatograms. Analysis of chromatography fractions showed that BRRM elutes directly after the GST tag (Figure 4.11(B)). The protein runs as quite a diffuse band on SDS-PAGE, running close to its theoretical molecular weight of 13.2 kDa.

Fractions 47 to 51 were pooled and concentrated. Around 700 µg of protein was obtained per litre of culture at >95% purity, judging by bands on SDS-PAGE (Figure 4.11). A 1D ^{1}H NMR spectrum of the protein (Figure 4.12) indicated this protein does not appear to form a stable, well-folded domain.

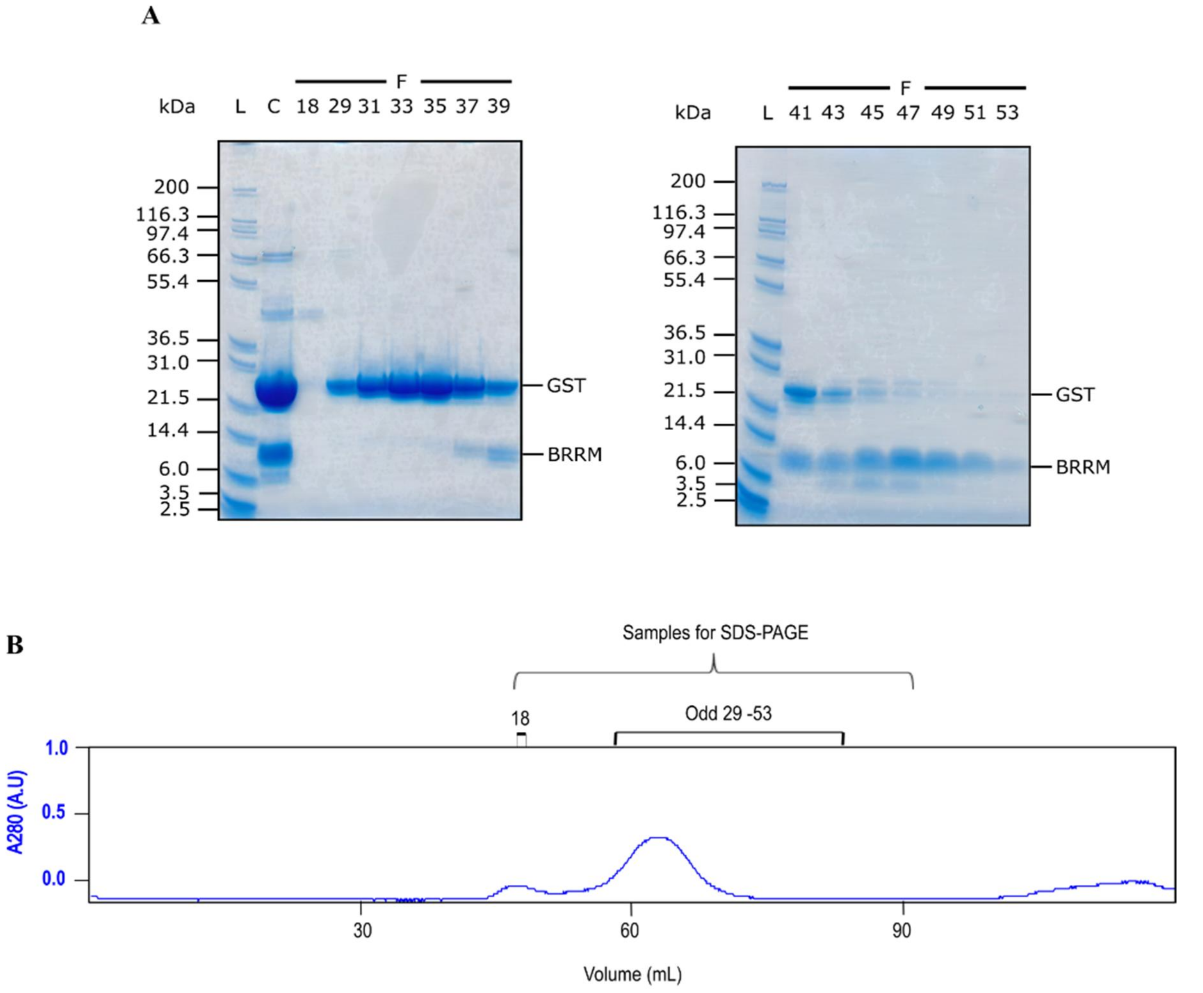


Figure 4.11. Size exclusion chromatography and SDS-PAGE analysis of BRRM.
(A) SDS-PAGEs of the selected size exclusion chromatography fractions. Abbreviations: L, Mark 12 ladder; C, HRV-3C cleavage of pooled GSH elutions; F, size exclusion chromatography fractions, numbers as indicated. **(B)** Size exclusion chromatography elution profile. Fractions taken for SDS-PAGE are indicated above the chromatogram.

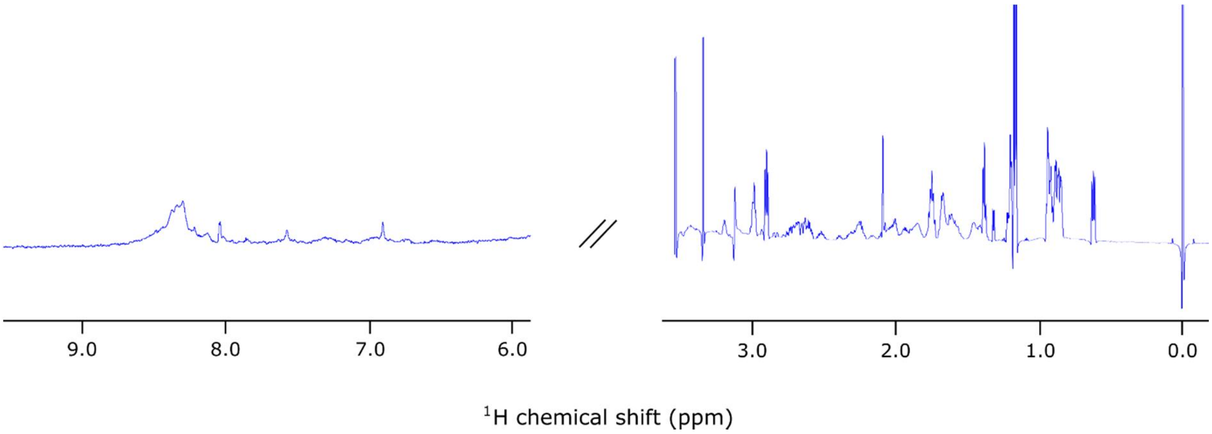


Figure 4.12. 1D ^{1}H NMR spectrum of BRRM.
The amide (left) and methyl (right) proton region of 200 µM BRRM in 20 mM sodium phosphate and 1 mM DTT at 298K.

Binding of BRRM to RNA was tested despite the fact that the domain was not highly ordered, given the presence of the peptide sequence very similar to the RNA-binding RNP-1 motif (Figure 4.7(B)), which alone might display RNA-binding capability. Further, given the remarkable demonstrations concerning the role of disorder in RBPs in recent years (see Section 4.4), it was possible that either (a) this domain might bind RNA without taking up a well-defined fold in the unbound state or (b) that it might fold upon binding, into an RRM (or another) fold. By way of example, a ~500 residue disordered basic domain of the microtubule scaffolding protein APC has been shown to bind mRNA [252]. Additionally, there are many examples of intrinsically disordered proteins that fold upon binding [253]. Indeed, disorder to order transitions have been observed in some RBPs upon binding RNA [172]. This has even been seen for certain secondary structural elements in RRMs (albeit not the whole domain); for example, an additional α-helix 3 of the RRM from the splicing factor Snu17p only folds upon ternary complex formation [254].

4.5.2 Testing binding of BRRM to *cad* 3′-UTR and Pentaprobes

In order to determine if BRRM displays any RNA-binding capacity, Cy5 labelled RNA was made by *in-vitro* transcription as described in Section 4.3.2. Binding of BRRM was tested to cad411, cad525, cad186 as well as Pentaprobe 7 (PP7) by EMSA. A Pentaprobe was included in addition to *cad* 3′-UTR sequences in order to increase the diversity of RNA sequences tested. BHD was used as a positive control.

No substantial binding of BRRM to any of the tested RNA sequences was seen (Figure 4.13). There is a faint shifted band observed for the cad411 probe, but given that it is only a minor shift at the highest concentration of 15 µM, the interaction is very weak. BHD, consistent with its behaviour in the preceding sections of this Thesis, bound all sequences to a similar extent. It might be that BRRM requires extra N-terminal amino acids in order to fold properly. Nevertheless, these results demonstrate that the RNP-1 motif in bicoid is unlikely to bind RNA in the context of the isolated domain.

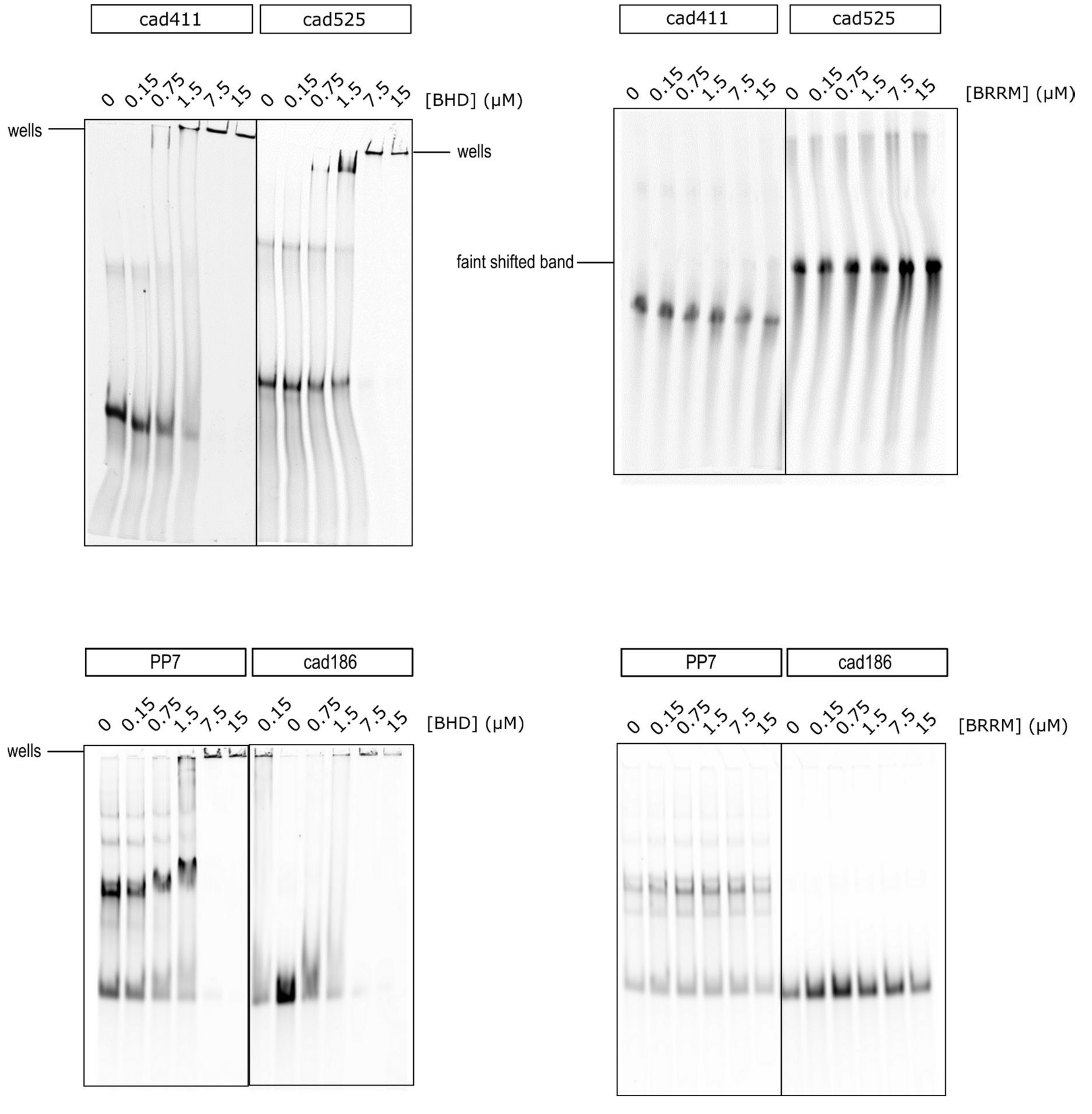

cad411
cad525
0 0.15 0.75 1.5 7.5 15 0 0.15 0.75 1.5 7.5 15 [BHD] (μM)
wells
wells
cad411
cad525
0 0.15 0.75 1.5 7.5 15 0 0.15 0.75 1.5 7.5 15 [BRRM] (μM)
faint shifted band
PP7
cad186
0 0.15 0.75 1.5 7.5 15 0.15 0 0.75 1.5 7.5 15 [BHD] (μM)
wells
PP7
cad186
0 0.15 0.75 1.5 7.5 15 0 0.15 0.75 1.5 7.5 15 [BRRM] (μM)

Figure 4.13. EMSAs of BRRM with *cad* 3′-UTR fragments and PP7.
EMSAs of BRRM and BHD binding to cad411, cad525, cad186 and PP7. Increasing concentrations of BRRM (right) and BHD (left) were incubated with Cy5-labelled RNAs then resolved on a 6% polyacrylamide native gel.

4.6 Phase-separated RNP granules

In the final part of this Chapter, the possible role of disordered regions of bicoid (Section 4.4.1) in the formation of phase-separated RNP granules is discussed.

It is notable that phase-separated RNP bodies are prevalent in *Drosophila* development [255]. Thus, 30% (143 of 476) of all mRBPs identified in the early fly embryo were also found in RNA granules in *Drosophila* S2 cells [222]. Even *bicoid* mRNA itself is known to form phase-separated bodies during development together with the RBP Staufen, localising the mRNA to the anterior pole [256]. Dimerisation of *bicoid* mRNA is required for Staufen recognition, which probably involves structural read out of the RNA by multiple dsRBMs of Staufen [257]. Another characterised example is that of the *Drosophila* protein Bruno, which binds to Bruno recognition elements within the *oskar* 3′-UTR, and this can result in oligomerisation of *oskar* mRNA into large, translationally silenced RNP particles [258].

Indeed, it appears that translational silencing often occurs in phase separated bodies; characterised examples of mRNP silencing granules in eukaryotic cells include P-bodies [259, 260], GW-bodies [261], stress granules [262] and germ granules [263]. Translation repression and the formation of phase separated RNP bodies may even be directly coupled in some instances, as demonstrated by the ability of the DDX6-4-ET complex to effect both de-novo P-body assembly and miRNA-dependent translation repression [264].

The prevalence and variety of these RNP granules that have been discovered recently in biology is changing the way some RNA-protein complexes need to be studied, and fundamentally changing our understanding of how proteins and RNA function in the cell. Therefore, a brief analysis of recent biochemical research involving RNP granules will be given here.

4.6.1 What do we know about RNP granules?

RNP granules are an assortment of membrane-less, phase-separated granules that are enriched in IDR containing RBPs and RNA. These granules, which are found in eukaryotic cells, act as liquid or hydrogel droplets that are capable of distortion, budding and fusion [265]. Phase separation is thought to be driven by the large-scale effects of weak, multivalent interactions [232], primarily by highly dynamic interactions between RBDs, IDRs and low-complexity regions within proteins, and modulated through interactions with RNA [233]. This type of network of interactions is depicted in Figure 4.14(B), and shown in contrast to a monomeric RNP complex in which one protein molecule is complexed with RNA via multiple RBDs and disordered regions (Figure 4.14(A)).

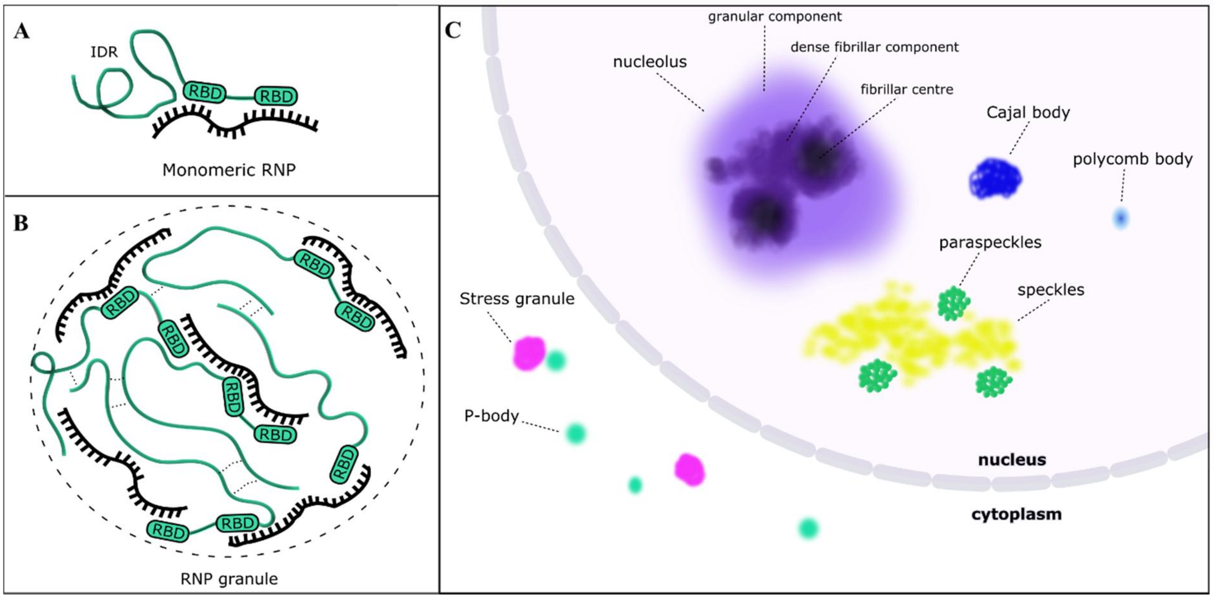


Figure 4.14. Phase separated RNP bodies.
(A) A typical monomeric RNP complex with one protein bound to an RNA molecule through multiple RBDs and an IDR that becomes ordered on binding. **(B)** A phase-separated RNP complex consisting of a variety of RNA species and proteins with multiple RBDs and IDRs. Multivalent interactions, which can be dynamic and transitory, and involve low complexity sequences, drive RNP granule formation. Protein is shown in green and RNA in black. **(C)** Representation of some common nuclear and cytoplasmic RNP granules. Descriptions of these bodies are contained in the text.

A variety of common cellular RNP granules are depicted in Figure 4.14(C). The largest RNP granule is the nucleolus, which forms around regions of ribosomal DNA in chromosomes and organises protein translation machinery [266]. The nucleolus has been shown to consist of multiple phase-separated sub-compartments with unique properties; fibrillar centres are located inside dense fibrillar components which are contained within the less dense granular component [267]. Cajal bodies are another common type of RNP granule, forming on regions of active snRNA loci and likely functioning to assemble spliceosomal snRNPs [268]. Other nuclear bodies include speckles and paraspeckles which both contain mRBPs, mRNA and lncRNAs, and display distinct, granular morphologies. Nuclear speckles are foci that contain mRNA production and processing factors [269], and paraspeckles are thought to be involved in the retention of certain RNA species and transcription factors within the nucleus [270, 271]. Cytoplasmic bodies include stress granules [272], and P-bodies, two different types of mRNA silencing granules [273, 274].

The formation of these structures appears in general to be reversible. Under *in vitro* conditions, temperature, salt concentration, post translational modifications and RNA-binding capacity all can affect granule formation, leading to suggestions that the formation and disassembly of structures is functionally regulated [234, 275-277].

Both RBDs and IDRs are important in granule formation. The presence of both have been shown to be required for RNA-induced granule formation in the cases of hnRNPA1 (stress granules) [234] and Whi3 (microorganism RNP body protein) [238]. In some instances, the specific role of these protein domains has been elucidated. For example, the low-complexity R/G domain of FIB1 has been shown to drive phase separation of the dense fibrillar components of nucleoli, and, whilst not required for droplet formation, the RBD of FIB1 (RNA methyltransferase domain (MD)) prevents mixing of the fibrillary component and the granular component [267].

General rules for RNP granule assembly are still lacking. However, the latest evidence suggests that multivalent interactions between low-complexity sequences, RBDs and RNA likely create scaffolds for the formation of granules, which can then exchange lower-valency binding partners (that is, components with a lesser number of potential binding sites) on free scaffold binding sites [278].

Specific amino acids are enriched in the low-complexity regions that are involved in phase separation, namely glycine, glutamine, asparagine, tyrosine, serine, glutamic acid, aspartic acid and phenylalanine [241]. Brangwynne *et al.* (2015) propose in their analysis of recent data that there is a ‘hierarchical interplay’ of interactions that result in the formation of these bodies. That is, there are long-range electrostatic interactions that are supplemented by short-range, directional dipolar interactions between glycine, glutamine, glutamic acid, asparagine and serine, or cation-π interactions between the positively charged arginine and the aromatic sidechains of phenylalanine and tyrosine [241].

How RNP granules regulate their composition is also not well understood in most cases [234, 235]. There is some evidence, however, that low-complexity motifs can target proteins to granules; for instance the progressive mutation of [G/S]Y[G/S] motifs in FUS resulted in reduced recruitment of the protein to stress granules [279].

A recent study elegantly highlighted how RNP germ granules can form according to the interplay of RNA and RBP concentration gradients. All sexually reproducing organisms contain germ granules within their germ cells, and these granules appear to function broadly in post-transcriptional control of gene expression [263]. P-granules in *C. elegans* are a well-studied type of germ granule. They are named due to their posterior localisation that materialises over repeated cell divisions during zygotic development (notwithstanding their role in zygotic development, they are classified as germ granules as they are passed directly from mother to daughter). The primarily disordered RBP MEG-3 has no known RBDs and is required for P-granule formation in embryos. The protein phase-separates *in vitro* in a concentration dependent manner (at micromolar concentrations) and this behaviour is stimulated by RNA (Figure 4.15(A)). Another RBP, MEX-5, contains two RNA-binding zinc fingers and competitively binds RNA that is required for this phase transition, thereby inhibiting granule formation. *In vivo*, MEX-5 is localised at high concentration at the anterior end of the zygote, blocking granule

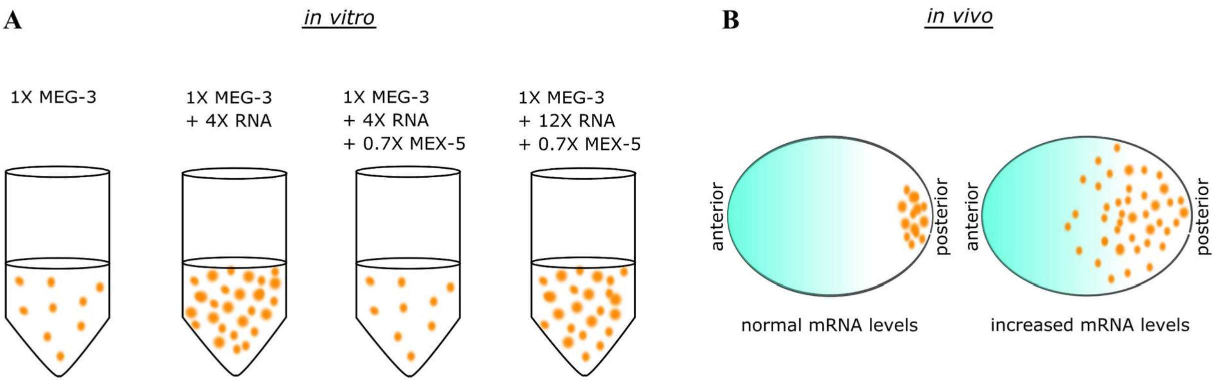


Figure 4.15. Germ granules can form according to the interplay of RNA and RBP gradients.
(A) The RBP MEG-3 forms phase-separated droplets (*orange*) *in vitro* at micromolar concentrations, and this separation is stimulated by RNA and inhibited by MEX-5. **(B)** Schematic of developing zygotes. MEX-5 is at a high concentration at the anterior end of the zygote (MEX-5 concentration is indicated by *cyan* shading) and competes for binding to mRNA with MEG-3. In normal zygotes, MEG-3 granules (*orange*) form only at the posterior pole of the zygote (left schematic). When mRNA turnover is blocked, MEG-3 granule formation spreads into the anterior end of the zygote (right schematic).

formation at this end of the cell. RNAi-mediated depletion of LET-711 (a scaffolding component of the primary mRNA deadenylase that functions during early development) prevents mRNA turnover, resulting in the formation of MEG-3 granules extending further into the anterior end of the zygote [280] (Figure 4.15(B)).

Both MEG-3 and MEX-5 display little RNA-binding specificity *in vitro* [280, 281], and specificity determinants dictating which mRNAs are bound by these proteins have not been identified.

4.6.2 Bicoid might repress *cad* in an RNP translationally silenced granule

Given the presence of the low-complexity sequences in bicoid detailed in Section 4.4.1, and the prevalence of translationally silenced mRNP granules [282] it seems possible that bicoid might repress *cad* in a phase-separated granule.

Supporting this possibility is new evidence that bicoid binding events in the nucleus occur in localised hubs, as visualised by single molecule fluorescence in developing embryos (Figure 4.16) [283]. In the posterior region of the embryo where bicoid concentration is low, the majority of bicoid binding events are localised to these foci. This heterogeneous spatial distribution of bicoid binding events in the nucleus involves another protein, Zelda, which contains a high proportion of low-complexity sequence, including multiple stretches of glutamine repeats. The authors surmise that these foci, by creating high local concentrations of bicoid, enable regulation of bicoid target genes at posterior locations where the overall concentration of bicoid is very low (~2 nM). The molecular details of bicoid foci formation are not known, however the process may involve a change in chromatin state given that both bicoid and

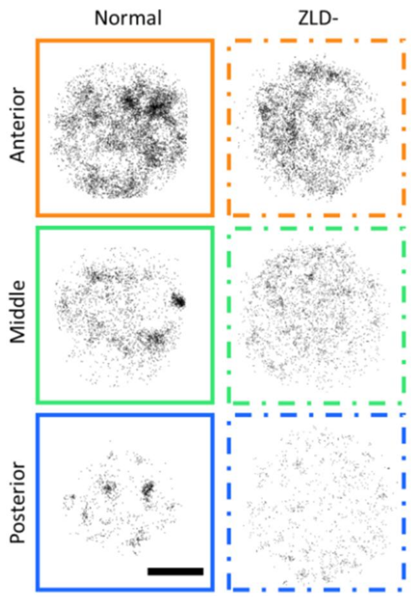


Figure 4.16. Distribution of bicoid binding events in normal and *Zelda*⁻ mutants in the nucleus. Representative data for the distribution of nuclear bicoid binding events in normal and *Zelda*-null (ZLD-) embryos, visualised using single molecule fluorescence of eGFP-tagged bicoid. The length of the bar represents 2.5 µM. The image in this Figure is taken from Mir *et al.* (2017).

Zelda have been previously shown to increase chromatin accessibility [284]. In the case of bicoid, this capacity to increase chromatin accessibility requires the presence of the C-terminal portion of the protein that includes the predicted PLD and the putative RRM. Moreover, given the high proportion of low-complexity sequence in Zelda, it is foreseeable that these foci might involve multivalent, protein-protein interactions between disordered segments in bicoid and Zelda. Indeed, it has recently been proposed that the cooperativity in binding interactions that phase separation can elicit might be a crucial factor in transcriptional gene regulation [285].

The sequestration of non-translating mRNAs in phase-separated P-bodies is pivotal in *Drosophila* development [286-288]. How proteins and RNAs are targeted to P-bodies is not well understood, however aggregation prone poly-glutamine and glutamine/asparagine rich domains are believed to play a role in P-body localisation and assembly given the prevalence of such sequences in P-body components such as Edc3, Lsm4 and GW182 [240].

Interestingly, P-bodies have been reported to contain RISC components such as Ago2 [289, 290], and therefore bicoid may be associated with P-bodies because Ago2 has been shown to genetically interact with bicoid, and miRNAs have been implicated in bicoid mediated repression of *cad* (both detailed in Section 1.2.5). Ago2 in Drosophila, as well as in many other invertebrates, contains a long poly-glutamine repeat region at its N-terminus [291]. The localisation of RISC components as well as the mRNA degradation machinery such as the deadenylase CCR4-NOT complex and decapping enzymes within P-bodies may increase the efficiency of translation inhibition due to the increased local concentration of required components. It may be that bicoid, through its low-complexity domains, targets *cad* to repressive bodies such as P-bodies.

Alternatively, a repressive RNP granule specific to the bicoid:*cad* interaction might form, with contributions from other disordered proteins similar to Zelda (involved in the formation of bicoid nuclear foci, introduced above). More generally, given the overrepresentation of low-complexity sequences in RBPs, it might be that phase separation is a common strategy for the repression of particular mRNAs and that these phase-separated granules have been below the limit of detection by microscopes.

Such spatial regulation may be particularly important to *Drosophila* during development, firstly, because the *Drosophila* embryo is a syncytium (consisting of many nuclei that exist in the shared cytoplasm and are not segregated by individual cell membranes), and secondly, because there is a high concentration of both maternally deposited and zygotic mRNAs, the expression of which is required to be tightly regulated in space and time. Accordingly, the further division of the syncytial cytoplasm into non-membrane bound partitions may be an essential feature of the embryo. For example, removing non-translating mRNA species from the bulk cytosol might result in increased translation efficiency by lowering the concentration of binding partners that can compete for limiting translation machinery.

In order to explore this possibility, existing *in situ* immunofluorescence data of bicoid and *cad* in *Drosophila* embryos in the scientific literature were examined, because RNP granules can sometimes be observed as distinct foci using this technique. The bicoid and cad protein gradients in an early stage *Drosophila* embryo are shown in Figure 4.17, in work by Rivera-Pomar *et al.* (1996) [87]. The green, spherical cad-containing bodies are syncytial nuclei. Cad is a transcription factor and is therefore found in the nucleus. Bicoid, being a transcription and translation factor, is found in both the nucleus and

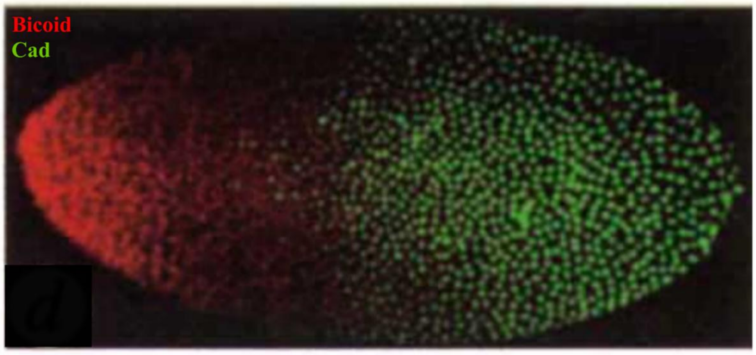


Figure 4.17. Bicoid and cad protein gradients in the early *Drosophila* embryo.
Anti-cad and anti-bicoid monoclonal antibodies were used to visualise the overlapping bicoid (*red*) and cad (*green*) gradients in the developing *Drosophila* embryo. Bicoid localisation is diffuse due to its presence in the cytoplasm as well as nuclei of the embryo, but no indication of phase separation is visualised. The image in this Figure is taken from Rivera-Pomar *et al.* (1996).

cytoplasm, and as a result is not localised to distinct spherical nuclei like cad; its distribution is more diffuse.

No clear evidence that bicoid forms RNP granules with *cad* was found in the literature. Given, however, the high spatiotemporal variability in protein and RNA distribution in the *Drosophila* embryo due to the complexity of embryonic gene regulation, *in-vitro* and *in-situ* experiments that test specifically for the phase separation of bicoid with *cad* will be required to more rigorously assess this possibility.

4.6.3 Specificity in RNP granules?

Determining specificity in phase-separated bodies is unchartered territory. Current determinants of specificity that have, by and large, served well for more well-defined, conventional macromolecular assemblies will likely need to be reassessed in mixed-phase cell biology.

Weak, multivalent binding drives droplet formation, as discussed above. This means that RBP:RNA interactions that have traditionally been classed as weak may be acting in aggregate to elicit phase compartmentalisation of the cell to establish localised, dynamic environments. For example, the IDR of the mRNA decapping protein Edc3 has been shown to have micromolar affinity for RNA and this interaction is sufficient to induce phase separation [292]. It is therefore possible that the prevalence of RNP granules goes some way to explaining the commonly observed weak binding in RNA:protein complexes [39]. If bicoid does indeed inhibit *cad* in phase-separated granules in the cytoplasm, this may rationalise the weak *in-vitro* affinity of BHD for RNA observed in this Thesis (K_d of BHD and BRE19nt ~3.8 μM, Section 2.3.4).

Moreover, RNP granules result in a high concentration of a particular subset of proteins and RNAs in a restricted area, with binding partners often at a much higher concentration than their dissociation constants [266]. Therefore, granule formation constitutes a type of spatial regulation of the cell which can foreseeably effect functional specificity in intermolecular interactions. Indeed, localised high concentrations of reactants can increase the kinetics of reactions. For example, mathematical modelling based on available empirical data have shown that the localisation of U4 and U6 snRNP components in Cajal bodies increases the rate of snRNP assembly by at least an order of magnitude [293].

The research presented in this Section highlights that there is much more biophysical analysis needed to define the kinetics and thermodynamics covering the formation and maintenance of RNP granules. This will involve determining the repertoire of macromolecules involved in phase separated bodies and measurement of the constituent interactions. Further, we will require a better understanding of how the compositions of these bodies are regulated, as well as the rules dictating interactions under what are quite different physiochemical conditions.

4.7 Summary

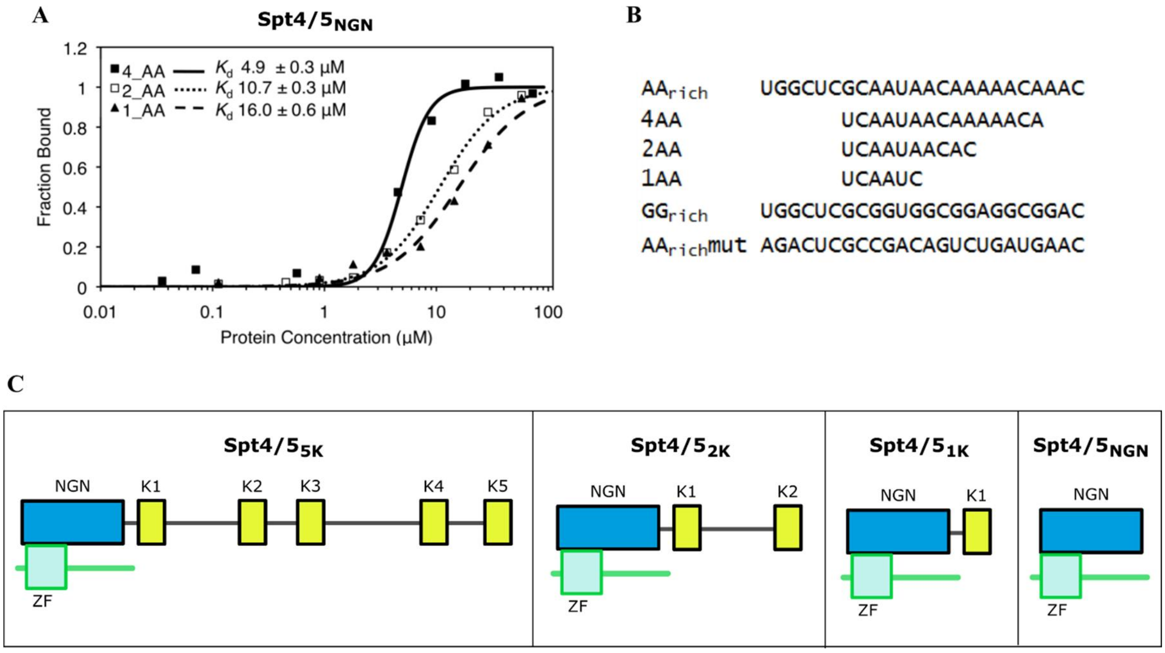
This Chapter investigated whether dimerisation of the bicoid homeodomain might be responsible for achieving RNA-binding specificity, but no increase in affinity of GST-tagged BHD was seen for RNA. A putative RRM in bicoid was investigated and produced, but this protein domain was not folded and little RNA-binding activity was detected. A bioinformatic analysis of full-length bicoid was presented, with a focus on low-complexity, disordered regions present in the protein. The potential role of these domains in the RNA-binding functionality of bicoid was discussed in the context of the latest research on RNP phase-separated bodies. Finally, the implications of phase separation to functional and biological specificity was considered, highlighting a need for innovative biochemistry research that will inform the reassessment of specificity determinants in these poorly understood cellular conditions.

Chapter 5: Investigation of AA-repeat RNA-binding by the transcription elongation factors Spt4 and Spt5

5.1 Introduction

As introduced in Chapter 1, Spt4/5$_{\mathrm{NGN}}$ from *Saccharomyces cerevisiae* has been described by our collaborators (Alice Vrielink's laboratory at the University of Western Australia) to bind in a sequence specific manner to ssRNA containing one or more repeats of the dinucleotide AA. Given these data, it was decided to characterise the structural basis for this interaction in order to ultimately further understand the important role that the Spt4/5 heterodimer plays in transcription.

To summarise the results presented in Chapter 1, SELEX was performed on Spt4/5$_{\mathrm{5K}}$ and the heterodimer was observed to enrich AA-repeat RNA. The sequence preference of Spt4/5$_{\mathrm{5K}}$ for AA-repeat RNA was confirmed through EMSAs and MST. The conserved core of the heterodimer, Spt4/5$_{\mathrm{NGN}}$, was observed to bind AA-repeat RNA with a similar affinity to Spt4/5$_{\mathrm{5K}}$, and MST assays demonstrated that Spt4/5$_{\mathrm{NGN}}$ could also bind shorter RNA oligonucleotides with one, two and four AA repeats (Figure 5.1).



AA$_{rich}$	UGGCUCGCAAUAACAAAAACAAAC
4AA	UCAAUAACAAAAACA
2AA	UCAAUAACAC
1AA	UCAAUC
GG$_{rich}$	UGGCUCGCGGUGGCGGAGGCGGAC
AA$_{rich}$mut	AGACUCGCCGACAGUCUGAUGAAC

Figure 5.1. Spt4/5$_{\mathrm{NGN}}$ binds AA-repeat RNA.
(A) MST assays of Spt4/5$_{\mathrm{NGN}}$ binding to RNA with varying numbers of AA-repeats. Increasing concentrations of Spt4/5$_{\mathrm{NGN}}$ were incubated with fluorescently labelled RNA then resolved on a 6% polyacrylamide gel. Binding isotherms were fitted using the Hill method. Data were acquired by Amanda Blythe. **(B)** Sequences of RNAs tested. **(C)** Domain arrangements of Spt4/5 constructs.

This Chapter describes the use of NMR spectroscopy and EMSAs to investigate and characterise the molecular underpinnings of the AA-repeat RNA-binding of Spt4/5$_{NGN}$.

5.2 Techniques used in this Chapter

5.2.1 NMR spectroscopy

This Chapter employs both 2D and 3D NMR methods, which were introduced in Section 3.2.1 as tools that can be used to identify residues involved in binding events. These techniques are used with the same aim in this Chapter. Some brief coverage of theoretical concepts relevant to the application of these techniques in this Chapter will be given here.

5.2.1.1 Sources of line broadening in NMR

The nuclear magnetisation that gives rise to an NMR signal decays with time, a process known as relaxation (one particular form of relaxation – so-called transverse relaxation – is most relevant to line broadening effects). This decay process results in a loss of signal intensity and, in general terms, the more rapid the relaxation is, the broader the NMR signals are in the resulting spectrum. Therefore, slower relaxation rates are desirable for good quality NMR spectra. There are many factors that affect the relaxation properties of nuclei, but when analysing changes in spectral appearance due to a binding event, line broadening due to two different scenarios is usually considered.

The first of these is line broadening caused by chemical exchange processes, introduced in Section 3.2.1.1. For a nucleus undergoing a chemical exchange process (whether it is folding-unfolding, binding, or simply conformational dynamics), the question of whether the nucleus appears as a single signal or as two (or more) discrete signals is dictated by the difference in the rate of exchange between the different forms (k_{ex}), and by the chemical shift of each of these states ($\Delta\omega$). The observed effects of exchange processes on NMR spectra fall on a continuum defined by the limits of fast ($k_{ex} >> \Delta\omega$) and slow ($k_{ex} << \Delta\omega$) exchange, where either one or two signals is/are observed per nucleus, respectively. Intermediate exchange lies in the middle of these two extremes ($kex \sim \Delta\omega$), and often results in signals failing to be resolved due to broadened signal lines. The signals obtained for each of these exchange regimes are depicted in Figure 5.2.

The other main binding-induced source of line broadening arises from slower molecular tumbling, a common occurrence when a multimer or biomolecular complex is substantially larger than the monomer/unbound protein. Generally, the tumbling time of a molecule increases with molecular weight, but it is also dependent on shape. Broadly speaking, the magnetised spins of molecules that

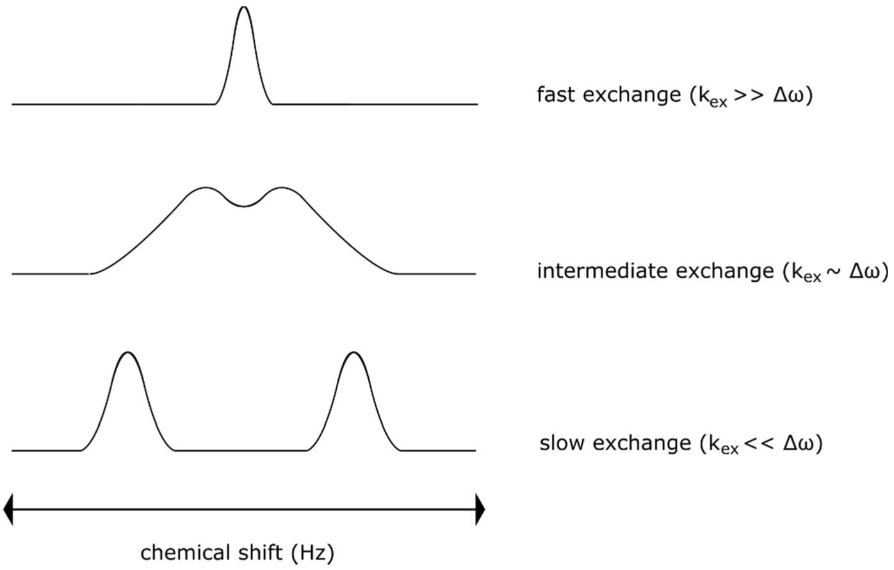

fast exchange (k_{ex} >> Δω)
intermediate exchange (k_{ex} ~ Δω)
slow exchange (k_{ex} << Δω)
chemical shift (Hz)

Figure 5.2. NMR chemical exchange regimes.
NMR chemical exchange regimes fit in to three broad categories. Fast exchange, where $k_{ex} >> \Delta\omega$, results in one signal at a weighted-average frequency. Slow exchange, where $k_{ex} << \Delta\omega$, results in a separate signal for each populated state. In between these two states, where $k_{ex} \sim \Delta\omega$, line broadening results in a regime intermediate to these two extremes and an inability to resolve signals.

tumble more slowly have a higher susceptibility to the transient local magnetic fields as the relatively slower movement means that the two are aligned longer. This means that phase coherence of magnetised spins of nuclei are lost more quickly. Therefore, transverse relaxation is more efficient for molecular systems with a slower tumbling time, resulting in broader signals.

In general, slower molecular tumbling will affect all signals. In contrast, in the case of chemical exchange, Δω varies amongst nuclei within a protein, and therefore the line broadening effects will be different for different signals in the NMR spectrum.

5.2.2 MST

MST was introduced in Section 2.2.2 as a recently developed technique to quantify biomolecular binding interactions. This Chapter presents MST data acquired by collaborators of the Mackay Laboratory that initiated this study. Some good quality sample MST data are shown in Figure 2.1, demonstrating smooth thermophoresis curves with good signal-to-noise ratio; these data yielded a sigmoidal binding curve and are indicative of a simple 1:1 binding event. Another important requirement for good quality MST data that is not illustrated in sample data in Figure 2.1 is the requirement that the fluorescence in each capillary (which contains the same concentration of the fluorescently tagged partner with different concentrations of the untagged partner), before the temperature gradient is established, be within 10% of each other. Figure 5.3 shows some pre-MST fluorescence data that fit this requirement. If these initial fluorescence intensities are not within 10% of each other, a test involving protein denaturation by sodium dodecyl sulfate (SDS) is required in order to rule out loss of the fluorescent molecule due to aggregation. If the change in fluorescence is due to a binding event (which is acceptable for good data) this can be confirmed through denaturation of the protein with SDS as this will disrupt interactions between the protein and binding partner and restore fluorescence. Alternatively, if aggregation is responsible, fluorescence will not be restored as these aggregates are removed during centrifugation prior to the addition of SDS. In this case, experimental conditions need to be optimised in order to eliminate aggregation.

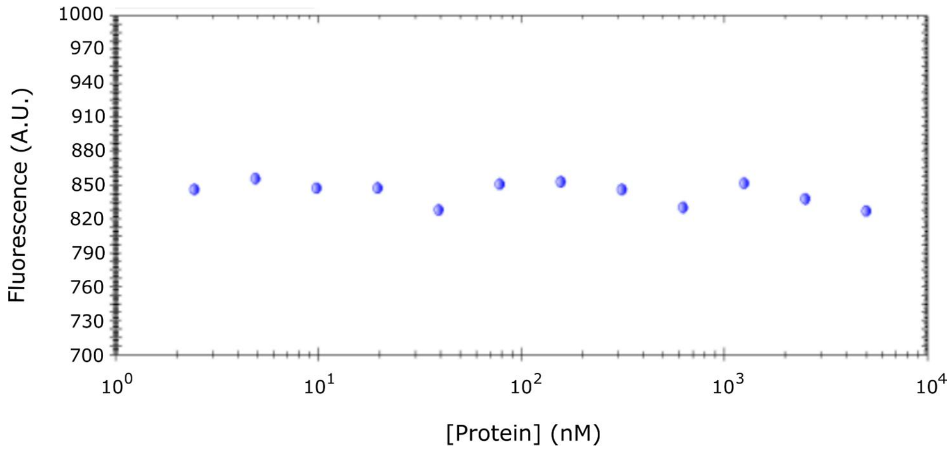


Figure 5.3. Fluorescence prior to heat gradient establishment in MST.
Good quality MST data require that the fluorescence intensity in each capillary prior to establishment of the MST temperature gradient be within 10% of each other. If they are not, an SDS-test needs to be done in order to determine the nature of the protein dependent loss in fluorescence.

5.3 Investigation of sequence specific binding of Spt4/5$_{\text{NGN}}$ to AA-repeat RNA

5.3.1 Expression and purification of Spt4/5$_{\text{NGN}}$

The production of soluble *Saccharomyces cerevisiae* Spt4/5$_{\text{NGN}}$ in bacteria is not straightforward. An effective system was established by Amanda Blythe from the Vrielink laboratory, whereby soluble protein was obtained through the co-expression of a His$_6$-ubiquitin Spt5 fusion in the vector pHUE with untagged Spt4 in the vector pETM11 [294]. pHUE-Spt5$_{\text{NGN}}$ was engineered to incorporate a TEV cleavage site by Jason Low in the Mackay laboratory because it was suspected that there was a bacterial protease that was cleaving the ubiquitin tag and decreasing yields.

For NMR experiments, ^{15}N and $^{15}N/^{13}C$ labelled heterodimer was expressed in minimal media containing $^{15}NH_4Cl$ or $^{15}NH_4Cl$ and ^{13}C-D-glucose respectively (see Section 7.2.2.4). For EMSAs, unlabelled protein was expressed as per Section 7.2.2.3.

Co-expression of Spt4/5$_{NGN}$ in *E. coli* Rosetta(DE3)pLysS cells resulted in modest soluble protein yields, judging from the SDS-PAGE bands observed following purification by nickel affinity chromatography (Figure 5.4). Curiously, Spt4 runs as a double band when the β-mercaptoethanol concentration is above ~ 10 mM [294].

Elutions one to eight (Figure 5.4) were pooled and the His$_6$-Ubq tag was cleaved by the addition of TEV protease. This step required deviation from Amanda Blythe's optimised protocol as TEV protease failed to cleave in 0.5 M imidazole (unlike the Usp2cc deubiquitinase used in the original protocol). Substantial optimisation was necessary as dialysis of the pooled elutions to reduce the imidazole concentration prior to cleavage resulted in precipitation of Spt4/5$_{NGN}$. The use of a desalting column avoided precipitation however concentration of the subsequently diluted protein solution to concentrations of ~50 µM resulted in up to 50% loss of protein due to binding to the centrifugal filter membrane. Low yields of high purity were obtained by size exclusion chromatography; ~500 µg per litre of culture at close to 100% purity (Figure 5.5(A&B)).

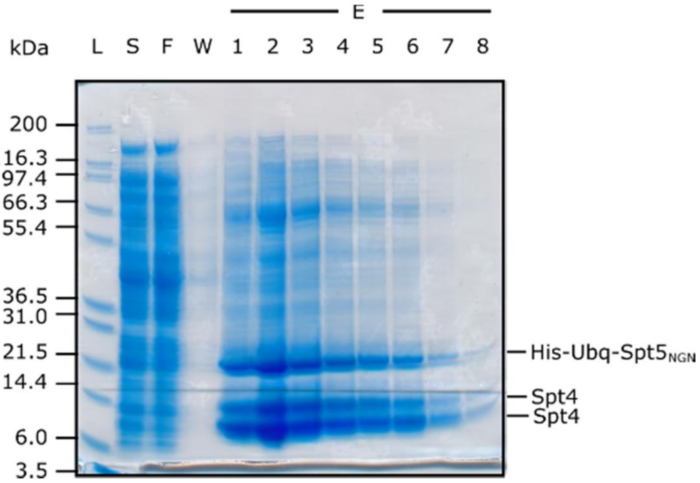


Figure 5.4. Overexpression and affinity purification of His-Ubq-Spt5$_{NGN}$ and Spt4.
SDS-PAGE analysis of Overexpression and affinity purification of His-Ubq-Spt5$_{NGN}$ and Spt4. Abbreviations: L, Mark 12 ladder; S, soluble fraction; F, flow through from nickel beads; W, wash from nickel beads; E, imidazole elutions, numbers as indicated.

Another strategy that was trialled involved using a lower concentration of imidazole in the elution buffer (0.2 M instead of 0.5 M). A higher TEV concentration was however required to get the same level of cleavage which resulted in poor separation of Spt4/5$_{NGN}$ from the TEV enzyme during size exclusion and consequently resulted in yields of ~1 mg per litre of culture at around 90% purity (Figure 5.5(A&C)).

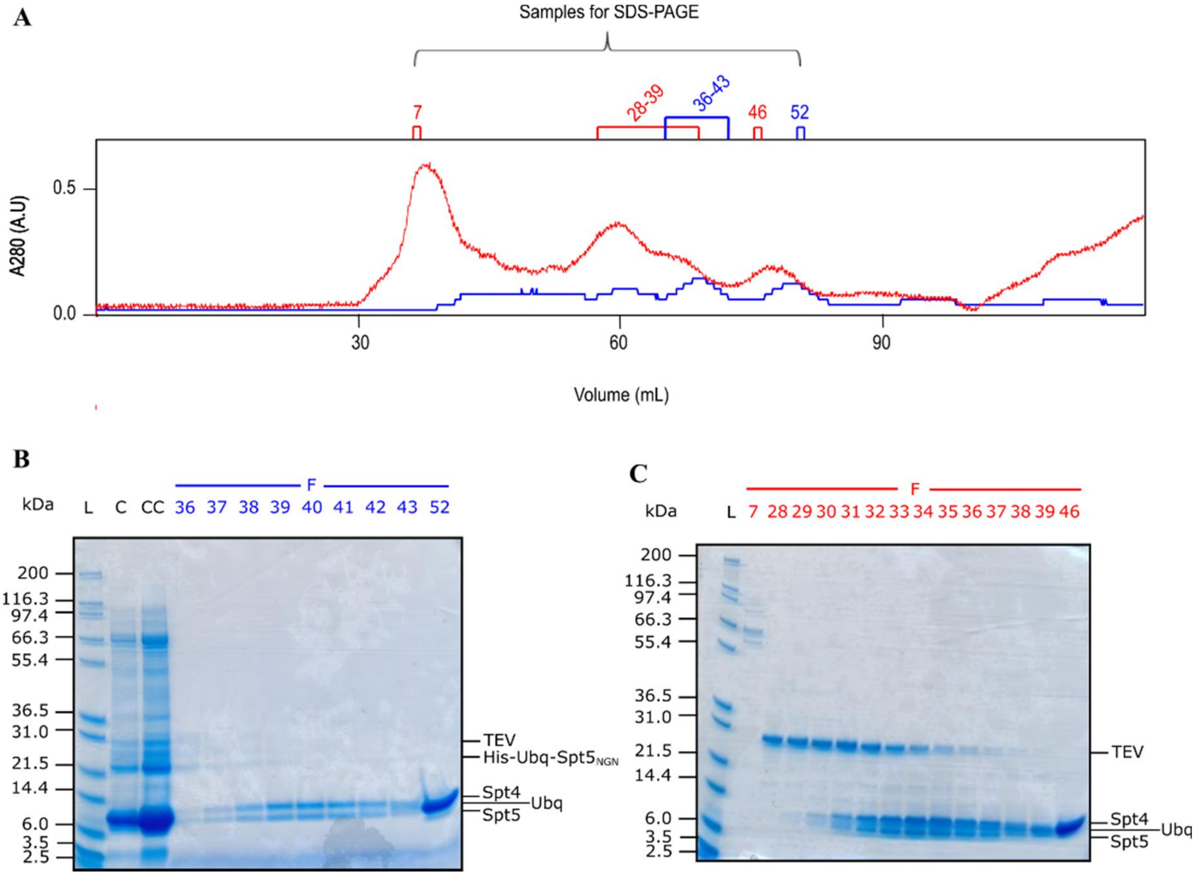


Figure 5.5. Size exclusion chromatography and SDS-PAGE analysis of Spt4/5$_{NGN}$.
(A) Size exclusion chromatography elution profiles of Spt4/5$_{NGN}$. The *blue* trace was from a preparation using a desalting column to dilute the imidazole, and the *red* trace was from a preparation using a reduced imidazole concentration of 0.2 M in elution buffer. Fractions taken for SDS-PAGE are indicated above chromatogram, with correspondence to each trace indicated by colouring. **(B)** SDS-PAGE analysis of the selected size exclusion chromatography fractions from the desalting column purification technique (*blue*). Abbreviations: L, Mark 12 ladder; C, TEV cleavage of pooled His$_6$-Ubq-Spt4/5$_{NGN}$ elutions; CC, cleavage solution of pooled His$_6$-Ubq-Spt4/5$_{NGN}$ elutions, concentrated to 5 mL for size exclusion chromatography; F, size exclusion chromatography fractions, numbers as indicated. **(C)** SDS-PAGE analysis of the selected size exclusion chromatography fractions from the reduced imidazole concentration purification technique (*red*). Abbreviations: L, Mark 12 ladder; F, size exclusion chromatography fractions, numbers as indicated.

Overall, Spt4/5$_{NGN}$ was poorly behaved. There was significant batch to batch variation in protein behaviour, resulting in varying yields. The proteins often precipitated at one or more of several steps throughout the purification, and had to be kept in at least 150 mM KCl. Protein behaviour deteriorated with isotopic labelling (^{15}N/^{13}C Spt4/5$_{NGN}$ being the most poorly behaved); batches were therefore scaled up accordingly to increase protein yields.

5.3.2 ^{15}N-HSQC analysis of Spt4/5$_{\text{NGN}}$ binding to AA-repeat RNA

NMR spectroscopy was used to characterise the RNA-binding properties of Spt4/5$_{\text{NGN}}$. This technique excels at characterising the molecular details of binding interactions, because individual residues can be identified in the spectra and their response to the addition of the binding partner can be directly observed. The first goal of this analysis was to collect ^{15}N-HSQC spectra of Spt4/5$_{\text{NGN}}$ and to assign the spectrum both in the absence and presence of AA-repeat RNA in order to identify the complement of residues involved in binding.

Recombinant ^{15}N-labelled Spt4/5$_{\text{NGN}}$ was overexpressed and purified as per Section 5.3.1 and ^{15}N-HSQC spectra were acquired with Spt4/5$_{\text{NGN}}$ alone and in the presence of RNA.

The initial spectrum of Spt4/5$_{\text{NGN}}$ (Figure 5.6(A)) shows approximately 90% of the expected number of signals (~164/183). This spectrum closely resembles ^{15}N-HSQC spectra previously acquired by Amanda Blythe (data not shown), indicating that the fold of the protein is essentially unaltered by the change in the purification protocol. The line shapes look reasonable given the experimental conditions (only ~50 μM protein, and a relatively high salt concentration of 150 mM KCl) and for a heterodimer of this size (21.6 kDa). The signals are widely dispersed, consistent with a structure containing significant amounts of β-sheet. This is in agreement with the known structure of the heterodimer (see structure in Figure 1.12(B)). A degree of variation in signal intensity is, however, observed (shown plotted for each residue Figure 5.6(B)), suggesting the presence of a conformational exchange process.

Initially, an RNA oligonucleotide with the sequence AACCAA (T2AA) was selected for binding analysis. This was 4-nt shorter than the 2AA sequence previously tested by Amanda Blythe, and was designed to minimise the length of the RNA oligonucleotide. Shorter oligonucleotides are preferred if possible because, as discussed in Section 5.2.1.1, larger molecules have a longer tumbling time which results in signal broadening and consequently a reduction in both sensitivity and resolution. ^{15}N-HSQC spectra of Spt4/5$_{\text{NGN}}$ alone and with T2AA RNA were acquired and are shown in Figure 5.7.

Unexpectedly, no substantial changes were observed upon RNA addition indicating that Spt4/5$_{\text{NGN}}$ does not bind AACCAA RNA with an appreciable affinity. This sequence was designed based on the hypothesis that a double AA motif might be the minimal Spt4/5$_{\text{NGN}}$ binding sequence; one possible explanation for the lack of binding observed in the ^{15}N-HSQC experiment is that the AACCAA repeat alone is insufficient for binding and that Spt4/5$_{\text{NGN}}$ may require sequence flanking the AA-repeats in order to bind. Given that MST binding data indicates that Spt4/5$_{\text{NGN}}$ binds a six nucleotide sequence with one AA-repeat (1AA) with an affinity of 16 μM (Figure 5.1(A)), the oligonucleotide 1AA was next tested by NMR for its ability to bind the heterodimer. ^{15}N-HSQC spectra of Spt4/5$_{\text{NGN}}$ alone and with 1AA RNA were acquired and are shown in Figure 5.8.

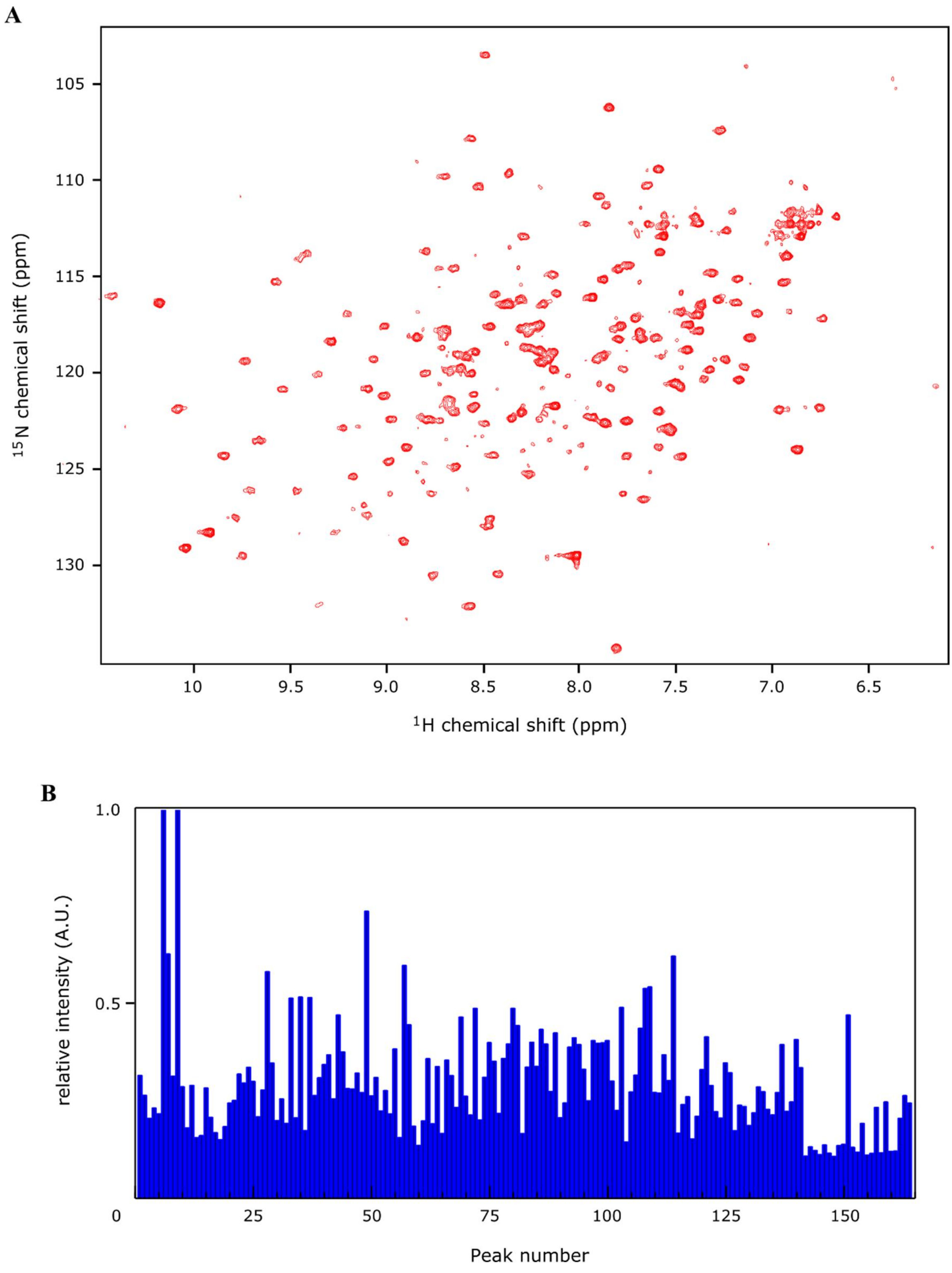


Figure 5.6. ^{15}N-HSQC spectrum of Spt4/5$_{\text{NGN}}$ in 150 mM KCl.

(A) ^{15}N-HSQC spectrum of 50 µM Spt4/5$_{\text{NGN}}$ alone. Spectrum was recorded at 298 K in 50 mM sodium phosphate, 150 mM KCl, 1 mM MgCl$_2$, 10 mM β-mercaptoethanol, pH 7.4. **(B)** Plot of relative intensity for each signal. Peaks are numbered by Sparky according to positions as they are unassigned at this point. Peak heights were obtained from Sparky.

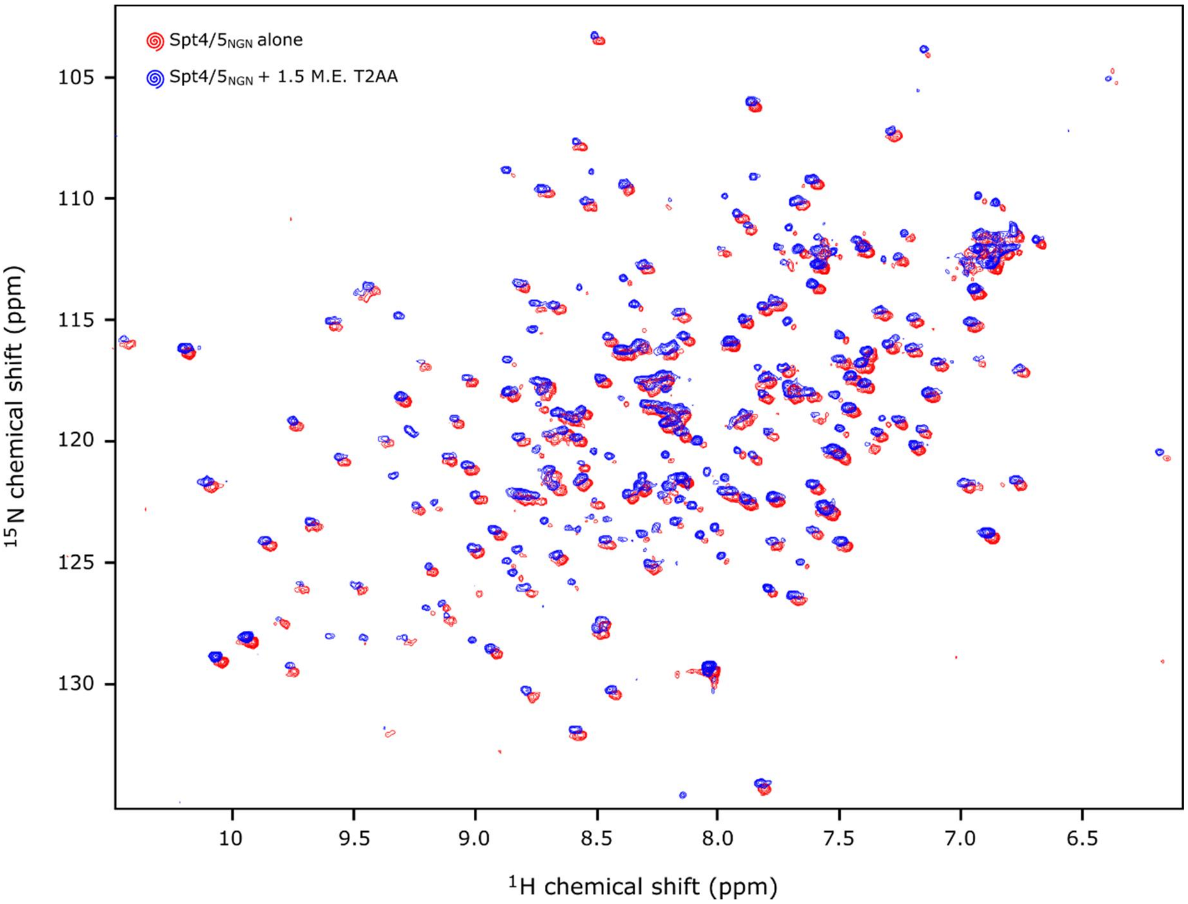


Figure 5.7. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ with and without T2AA RNA in 150 mM KCl. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 1.5 molar equivalents of T2AA RNA, offset slightly for clarity (*blue*). Starting concentration of Spt4/5$_{NGN}$ was ~50 µM, in 50 mM sodium phosphate, 150 mM KCl, 1 mM $MgCl_2$, 10 mM β-mercaptoethanol, pH 7.4. Spectra were recorded at 298K.

Again, no signal changes were seen upon addition of 1AA RNA to Spt4/5$_{NGN}$. Assuming that the protein and RNA in this assay have been accurately quantified, this rules out any binding with a dissociation constant of around 100 µM or stronger. Examination of the MST data for the interaction with this oligonucleotide (Figure 5.1(A)) shows that a complete MST curve for Spt4/5$_{NGN}$:1AA was not obtained. It is possible that the fitted dissociation constant of 16 µM is inaccurate, but even if it was 10-fold weaker, changes would still be expected in the HSQC spectrum under the conditions used here. As a next step, 4AA was tested for binding to Spt4/5$_{NGN}$, given that more AA-repeats should strengthen the affinity of the interaction. Again however, no signal changes were seen upon the addition of 2.5 molar equivalents of 4AA RNA (Figure 5.9).

Finally, the 24-nt AA$_{rich}$ – the full sequence obtained from SELEX enrichment, was tested for binding to Spt4/5$_{NGN}$. This is the only RNA sequence for which there is both EMSA and MST binding data, with a clearly shifted band visible in EMSAs (Figure 1.12(D)). Whilst less favourable for structural

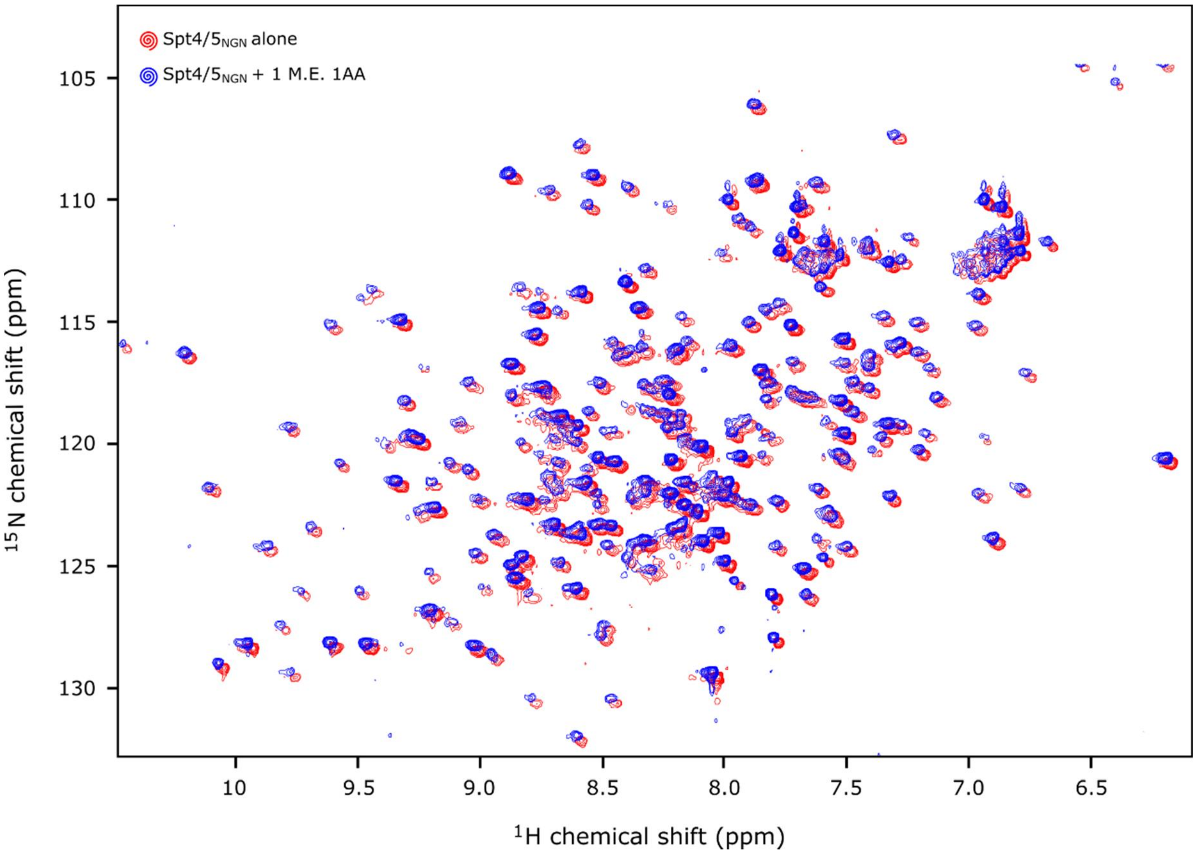


Figure 5.8. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ with and without 1AA RNA in 150 mM KCl. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 1 molar equivalent of 1AA RNA, offset slightly (*blue*). Starting concentration of Spt4/5$_{NGN}$ was ~60 µM, in 50 mM HEPES, 150 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP, pH 7.4. Spectra were recorded at 298K.

studies due to the size of the RNA, binding was tested via NMR in order to try to reconcile the current and previously acquired data.

The first time binding was assessed to AA$_{rich}$, the protein precipitated upon RNA addition, suggesting an interaction between Spt4/5$_{NGN}$ and AA$_{rich}$. In an effort to circumvent the precipitation issue, the binding buffer was changed to 50 mM MOPS (instead of 50 mM sodium phosphate or HEPES, which had been trialled previously), and the protein was added to the RNA instead of the reverse. This time, substantial changes in the HSQC spectrum were observed, with 71 of the ~220 signals disappearing (Figure 5.10(A)). No new signals were observed.

The intensity of each peak was measured in Sparky by integrating with Gaussian fitting to obtain fit heights. The differences between relative peak intensities in the bound and unbound state are shown in Figure 5.10(B). It can be seen that there is considerable variation between peak intensities, both within

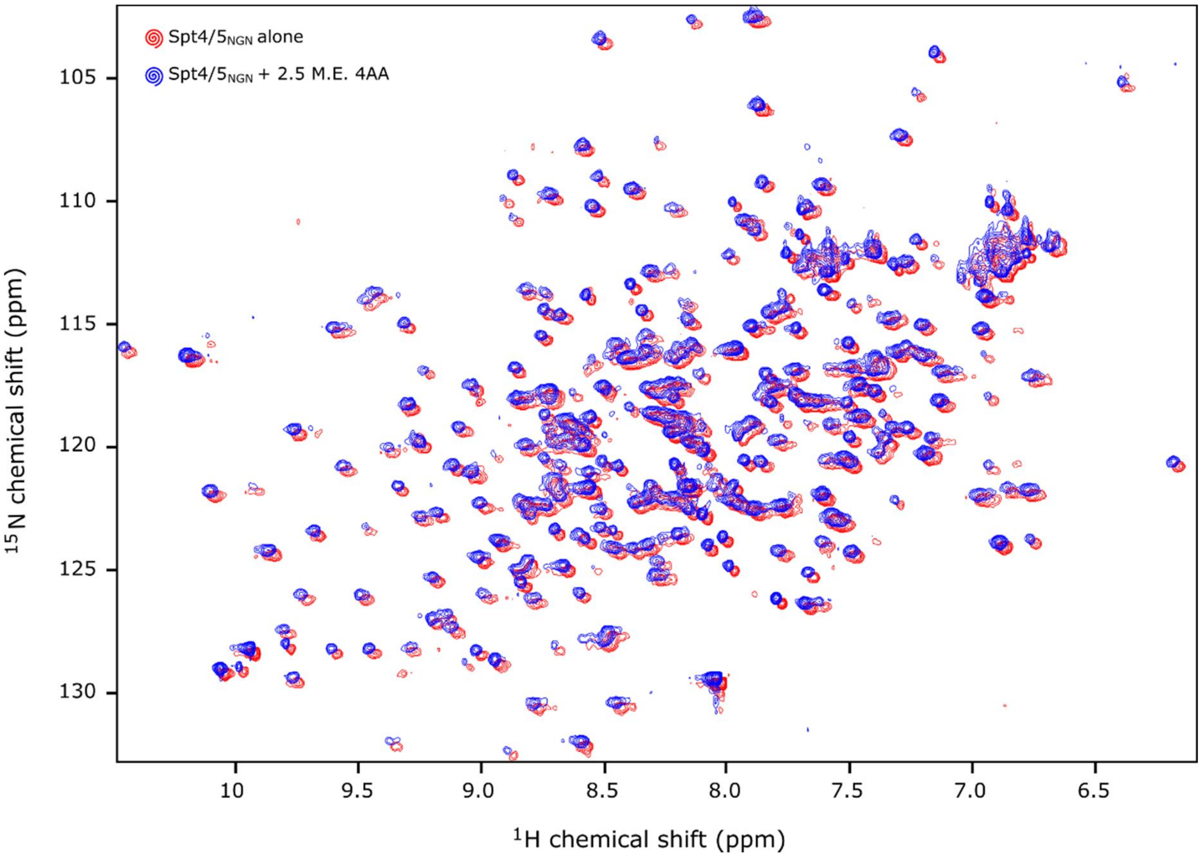


Figure 5.9. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{\mathrm{NGN}}$ with and without 4AA RNA in 150 mM KCl. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{\mathrm{NGN}}$ alone (*red*), Spt4/5$_{\mathrm{NGN}}$ with 2.5 molar equivalents of 4AA RNA, offset slightly (*blue*). Starting concentration of Spt4/5$_{\mathrm{NGN}}$ was ~65 µM, in 50 mM sodium phosphate, 150 mM KCl, 1 mM $MgCl_2$, 10 mM β-mercaptoethanol, pH 7.4. Spectra were recorded at 298K.

and across the two samples. On the whole, peaks that are present in the bound state are less intense than the same peak in the unbound state, as expected.

This experiment was repeated once more and a similar pattern of spectral change was observed (Figure 5.11).

The failure to observe new signals could be due to either an intermediate exchange regime or a slower molecular tumbling time of the complex. These concepts were introduced in Section 5.2.1.1; both scenarios result in line broadening but via different mechanisms. To reiterate, an intermediate exchange regime describes the situation where differences in signal frequencies between the bound and unbound state is similar to the exchange rate between the two states. Such situations are common for interactions with dissociation constants in the micromolar range. The EMSA data for Spt4/5$_{\mathrm{NGN}}$ also indicates that there may be more than one heterodimer bound due to multiple shifted bands (Figure 1.12(D)), in which case multiple chemical exchange events would be taking place at the same time. In addition, the protein-RNA complex has a larger molecular weight than the protein alone and signal broadening resulting

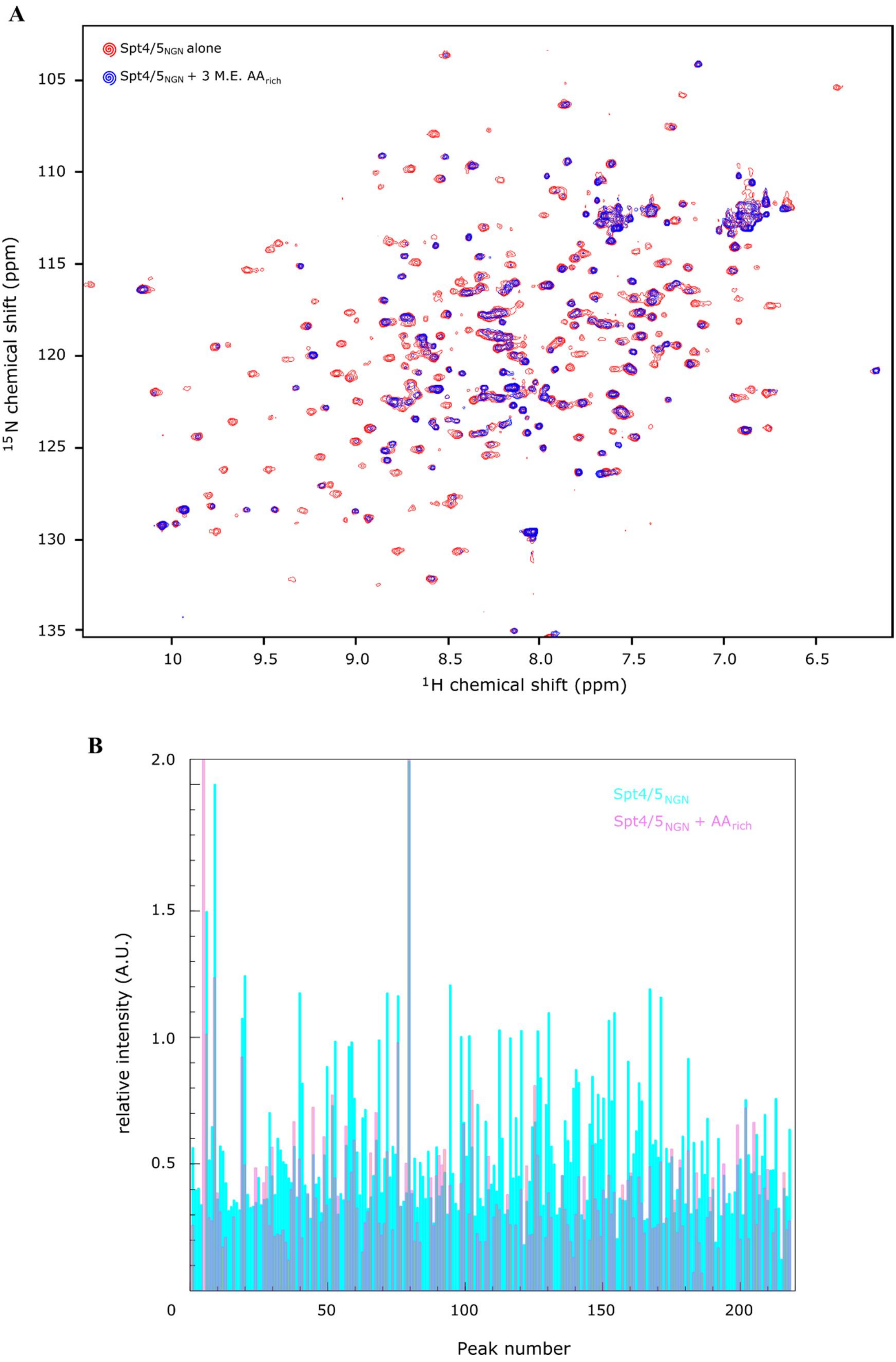


Figure 5.10. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ with and without AA$_{rich}$ RNA in 150 mM KCl. **(A)** Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 3 molar equivalents of AA$_{rich}$ RNA (*blue*). Starting concentration of Spt4/5$_{NGN}$ was 50 µM, in 50 mM MOPS, 150 mM KCl, 1 mM MgCl$_2$, 1 mM TCEP, pH 7.4 buffer. Spectra were recorded at 298K. **(B)** Plot of intensity for each signal: Spt4/5$_{NGN}$ alone (*cyan*), Spt4/5$_{NGN}$ + AA$_{rich}$ (*pink*). Fit heights were obtained in Sparky by integrating with Gaussian fitting. Peaks are numbered by Sparky according to positions as they are unassigned at this point.

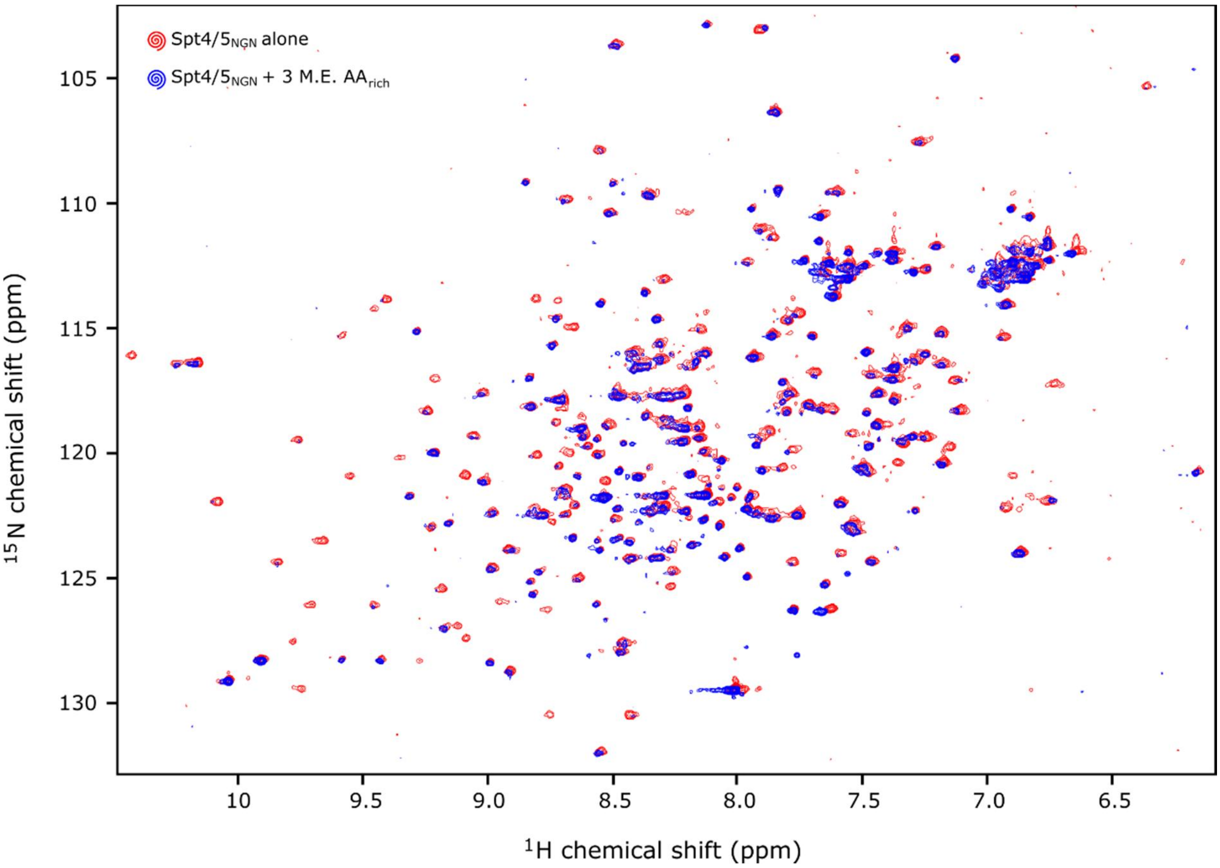


Figure 5.11. Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ with and without AA$_{rich}$ RNA in 150 mM KCl. Repeat of ^{15}N-HSQC spectra Spt4/5$_{NGN}$ with (*red*) and without 3 molar equivalents of AA$_{rich}$ (*blue*). Starting concentration of Spt4/5$_{NGN}$ was 50 µM, in 50 mM MOPS, 150 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP, pH 7.4 buffer. Spectra were recorded at 298K.

from slower molecular tumbling will also contribute to drops in signal intensity across the spectrum. In this situation, flexible (disordered) regions of the protein can retain sharp lines as they can effectively reorient (tumble) faster than the whole molecule by undergoing local motions.

5.3.3 Reconciling ^{15}N-HSQC spectra with MST data

The ^{15}N-HSQC spectrum for Spt4/5$_{NGN}$ shows substantial signal changes only for the 24-nt RNA sequence which was enriched in SELEX; no binding to any of the shorter AA-repeat RNA sequences was observed, which is in contrast to the MST data presented in Figure 1.12.

Upon closer inspection of the MST data, it can be seen that the change in thermophoresis for AA$_{rich}$ is at least five-fold bigger than the shorter RNA sequences (Figure 5.12(A)). MST guidelines stipulate that there needs to be a minimum change in thermophoresis between the bound and the unbound state of more than five units to obtain reliable data. For 1AA, 2AA and 4AA the change in thermophoresis is 10-18 units, but it should be noted that complete binding curves were not obtained. In contrast, the change in thermophoresis for AA_{rich} is over 100 units.

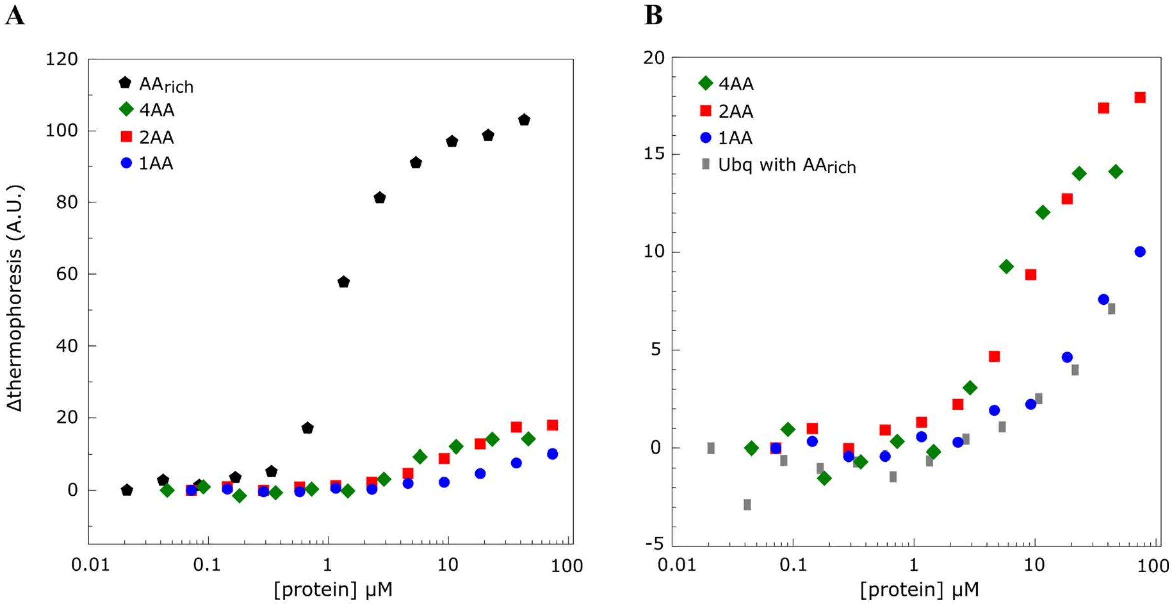


Figure 5.12. MST data of Spt4/5_{NGN} with AA-repeat RNA.
(A) MST curves of ubiquitin-tagged Spt4/5_{NGN} with 1AA, 2AA, 4AA and AA_{rich} fluorescently labelled RNA oligonucleotides and ubiquitin with fluorescently labelled AA_{rich}. Data points represent the average from three independent titrations. **(B)** The same MST curves of Spt4/5_{NGN}Ubq with 1AA, 2AA and 4AA as in (A) this time plotted without AA_{rich}. Data points represent the average from three independent titrations. Data were acquired by Amanda Blythe.

Examination of the fluorescence signals before the establishment of the MST heat gradient reveals protein dependent loss in fluorescence for 1AA, 2AA and 4AA, and in contrast, an enhancement of fluorescence for AA_{rich} (Figure 5.13). The enhancement in fluorescence is consistent with Spt4/5_{NGN} gel shifts with AA_{rich}, where there is a marked increase in the fluorescence of the shifted band compared to the probe alone (Figure 1.12(D)).

In MST, the concentration of the fluorescently labelled RNA should be equal for each titration point per sample and thus, as outlined above, variation in fluorescence prior to the establishment of the heat gradient should be within 10% for good quality data (as outlined in Section 5.2.2). A protein dependent reduction in fluorescence prior to heat gradient formation suggests material loss of the fluorescent

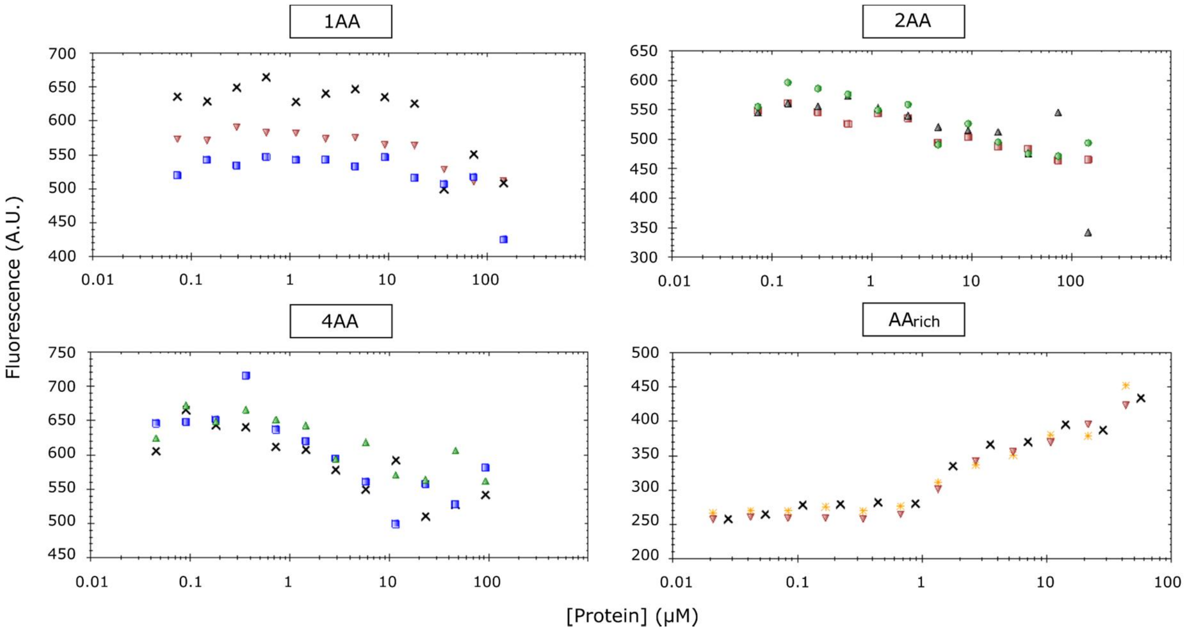


Figure 5.13. Pre-thermophoretic fluorescence of Spt4/5$_{NGN}$ with AA-repeat RNA.
Fluorescence from ubiquitin-tagged Spt4/5$_{NGN}$ MST data prior to establishment of the heat gradient for AA-repeat RNA as indicated. 1AA, 2AA and 4AA all exhibit a loss in fluorescence with increasing protein concentration. In contrast, there is a protein dependent enhancement in fluorescence seen for AA$_{rich}$. Fluorescence is shown in arbitrary units (A.U.).

molecule due to aggregation or adsorption. Because the SDS test described above has not been carried out for these oligonucleotides, it therefore is possible that the binding data presented for 1AA, 2AA and 4AA constitute false positive results. At the very least, there is clearly a difference in behaviour in the MST experiments between Spt4/5$_{NGN}$ with the shorter AA-repeat RNAs and AA$_{rich}$.

Given the ^{15}N-HSQC spectra presented above look very similar to spectra of Spt4/5$_{NGN}$ acquired by Amanda Blythe, it is reasonable to assume that the protein is behaving similarly under both the MST and NMR conditions.

Differences in experimental conditions for the MST and NMR experiments may have contributed to the contradictory data. Spt4/5$_{NGN}$ is His$_6$-ubiquitin tagged in the MST assays. Ubiquitin only controls were carried out with AA$_{rich}$ in order to rule out the tag contributing to RNA-binding. Ubiquitin alone does elicit some small change in thermophoresis upon addition to the labelled oligonucleotide; this change is around seven units at the highest ubiquitin concentration, which is very similar to the change seen for Spt4/5$_{NGN}$ with 1AA that is shown in Figure 5.12(B). Additionally, the RNA utilised in MST is fluorescein labelled. The fluorescein tag is relatively hydrophobic and so the possibility exists for a non-specific interaction to take place between the tag and the protein.

The existence of EMSA, MST and ^{15}N-HSQC data that show binding of Spt4/5$_{\mathrm{NGN}}$ to AA$_{\mathrm{rich}}$ provides a significant level of confidence that this interaction is real. In contrast, the MST data for Spt4/5$_{\mathrm{NGN}}$ binding to 1AA, 2AA and 4AA is of poorer quality due to the small changes in thermophoresis (<20 units), the incomplete binding curves obtained, the unverified nature of the protein dependent loss of fluorescence and, perhaps most clearly, the lack of an observed interaction in the ^{15}N-HSQC experiments presented in this Chapter.

Another possible explanation is that AA$_{\mathrm{rich}}$ is forming some sort of transient secondary structure that Spt4/5$_{\mathrm{NGN}}$ recognises. Most RNA secondary structure prediction tools, including RNAfold, are based on canonical Watson-Crick (A:U and G:C) and Wobble (G:U) base pairs, however, non-canonical base pairing is commonly observed in RNA structures [295, 296]. Moreover, non-canonical base pairing has been shown to be important for more accurate detection of RNA structural motifs [297] and to give better inference of RNA structure from sequence [298]. MC-fold, an RNA structure prediction tool that takes into account non-canonical base pairing, indicates that AA$_{\mathrm{rich}}$ may potentially sample a variety of transient secondary structures [298]. It is therefore possible that Spt4/5$_{\mathrm{NGN}}$ requires some structural elements of AA$_{\mathrm{rich}}$ that do not form in any of the truncated sequences tested.

5.3.4 Investigation of binding determinants for Spt4/5$_{\mathrm{NGN}}$ and AA$_{\mathrm{rich}}$ RNA by NMR

The data presented in this Chapter so far indicate that Spt4/5$_{\mathrm{NGN}}$ binds AA$_{\mathrm{rich}}$ RNA but not shortened AA-repeat RNA sequences. In order to determine which elements of the RNA are required for binding, further HSQC spectra were acquired of the proteins in the presence of truncated versions of AA$_{\mathrm{rich}}$.

AA$_{\mathrm{rich}}$ was first divided into two overlapping 13-nt halves, 1HAA$_{\mathrm{rich}}$ and 2HAA$_{\mathrm{rich}}$ (Figure 5.14(A)), and ^{15}N-HSQC spectra were acquired in the presence and absence of these RNA sequences. No binding was seen to either sequence (Figure 5.14(B&C)). Next, three bases were trimmed off the 5′ end and two bases trimmed off the 3′ end (TrimAA$_{\mathrm{rich}}$), and perhaps surprisingly, no binding was seen to this sequence either (Figure 5.14(D)).

These results suggest the possibility that the RNA-binding to Spt4/5$_{\mathrm{NGN}}$ might require the formation of structure that can only be formed by AA$_{\mathrm{rich}}$ and not by any of the shorter sequences. However, AA$_{\mathrm{rich}}$ is predicted to be unstructured by a variety of RNA structure prediction tools, such as RNAfold [83]. If AA$_{\mathrm{rich}}$ is indeed unstructured, it is surprising that we cannot observe any binding at all – even at reduced affinity - of Spt4/5$_{\mathrm{NGN}}$ to the truncated AA$_{\mathrm{rich}}$ sequences. The size of the Spt4/5$_{\mathrm{NGN}}$ heterodimer suggests that it is unlikely to require the full 24-nt of RNA for binding if the RNA is not structured. A possible explanation is that Spt4/5$_{\mathrm{NGN}}$ binds AA$_{\mathrm{rich}}$ cooperatively as a dimer of heterodimers, and requires the entirety of (or close to) the AA$_{\mathrm{rich}}$ sequence in order to do so, or just that both sets of terminal bases are required together. However, there is no evidence currently that the Spt4/5 heterodimer forms a 2:2 tetramer.

A

AA_{rich}	UGGCUCGCAAUAACAAAAACAAAC
$1HAA_{rich}$	UGGCUCGCAAUAA
$2HAA_{rich}$	AACAAAAACAAAC
$TrimAA_{rich}$	CUCGCAAUAACAAAAACA

B

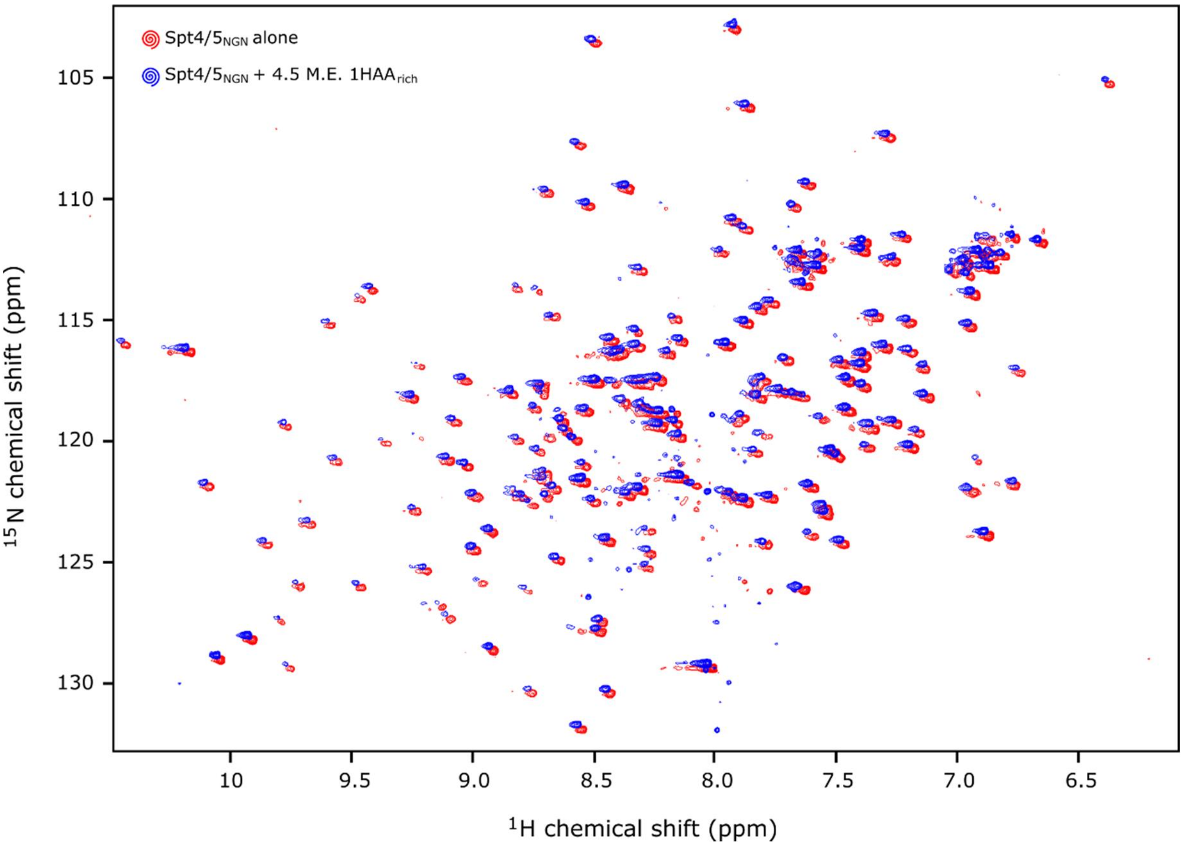
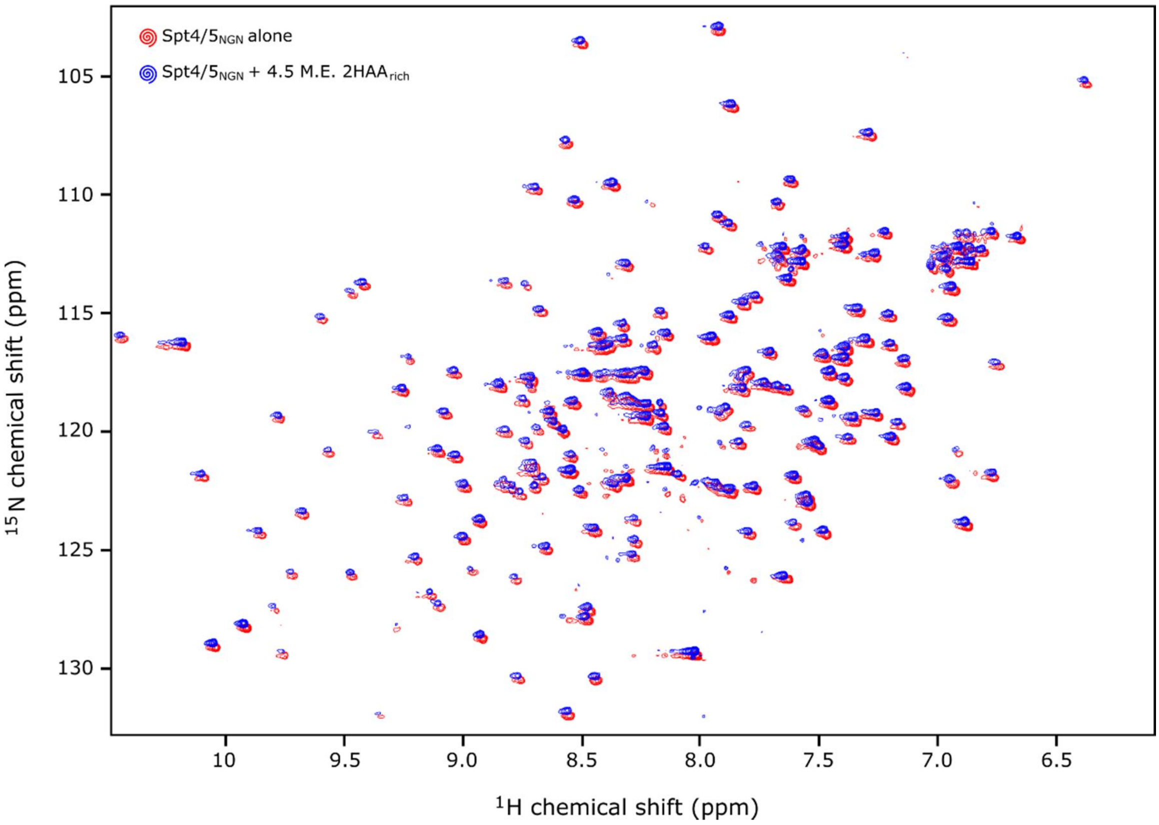


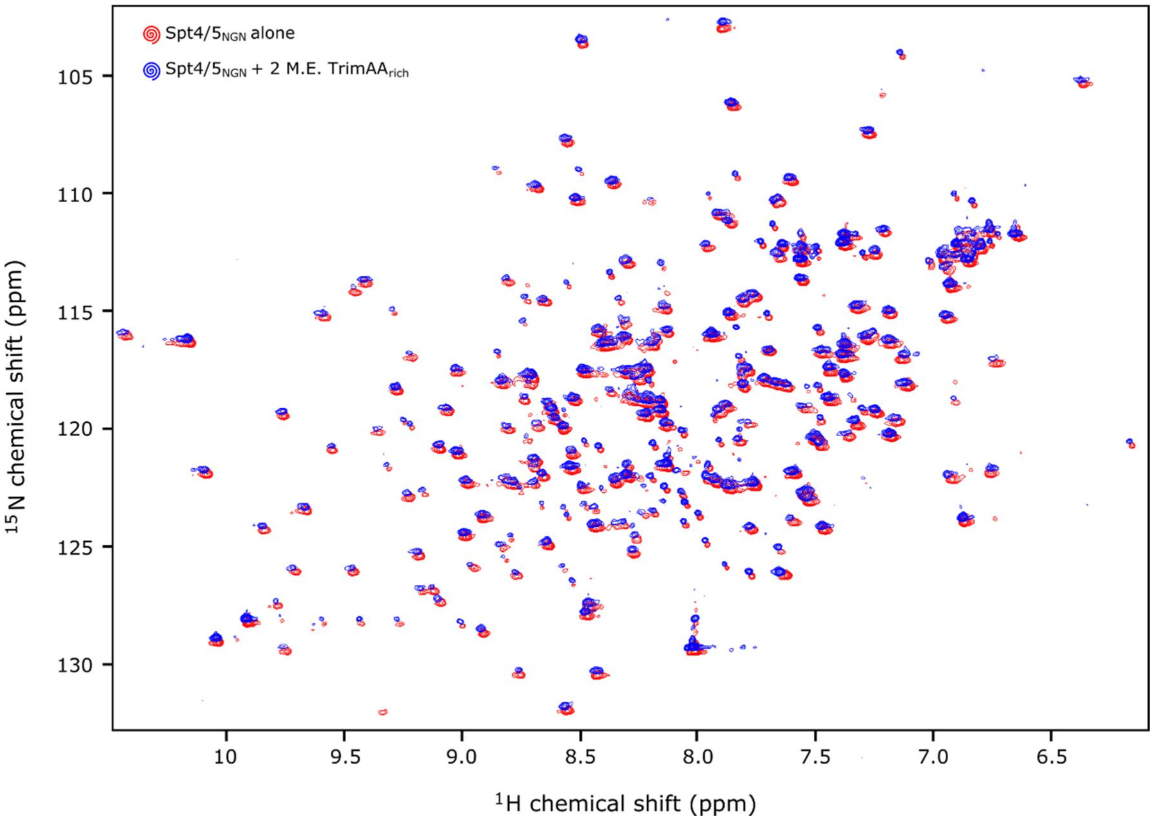
Figure 5.14. Overlays of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ with and without AA$_{rich}$ derived RNA oligonucleotides in 150 mM KCl.

(A) Sequences of AA$_{rich}$ derived RNAs tested for binding to Spt4/5$_{NGN}$. **(B)** Overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 4.5 molar equivalents of 1HAA$_{rich}$ RNA (*blue*). Starting concentration of Spt4/5$_{NGN}$ was 40 µM, in 50 mM MOPS, 150 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP, pH 7.4. Spectra were recorded at 298K. **(C)** Figure on the next page - overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 4.5 molar equivalents of 2HAA$_{rich}$ RNA (*blue*). Starting concentration of Spt4/5$_{NGN}$ was 40 µM, in 50 mM MOPS, 150 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP, pH 7.4. Spectra were recorded at 298K. **(D)** Figure on the next page - overlay of ^{15}N-HSQC spectra of Spt4/5$_{NGN}$ alone (*red*), Spt4/5$_{NGN}$ with 2 molar equivalents of TrimAA$_{rich}$ RNA (*blue*). Starting concentration of Spt4/5$_{NGN}$ was 80 µM, in 50 mM MOPS, 150 mM KCl, 1 mM $MgCl_2$, 1 mM TCEP, pH 7.4. Spectra were recorded at 298K.

C


Spt4/5$_{NGN}$ alone
Spt4/5$_{NGN}$ + 4.5 M.E. 2HAA$_{rich}$
105
110
115
120
125
130
^{15}N chemical shift (ppm)
10
9.5
9.0
8.5
8.0
7.5
7.0
6.5
^{1}H chemical shift (ppm)

D


Spt4/5$_{NGN}$ alone
Spt4/5$_{NGN}$ + 2 M.E. TrimAA$_{rich}$
105
110
115
120
125
130
^{15}N chemical shift (ppm)
10
9.5
9.0
8.5
8.0
7.5
7.0
6.5
^{1}H chemical shift (ppm)

5.3.5 Investigation of binding determinants for Spt4/5$_{NGN}$ and AA$_{rich}$ RNA by EMSA

In order to determine what elements of AA$_{rich}$ are required for binding, Spt4/5$_{NGN}$ was tested for binding to a series of mutants of AA$_{rich}$ (Figure 5.15(A)). AA5 is designed to test whether the three 5′-nt are required, whereas A4 through to A1 progressively replaces AA-repeats with CC repeats.

Spt4/5$_{NGN}$ was expressed and purified as per Section 5.3.1. Internal labelling of RNA transcripts with ^{32}P was carried out using ^{32}P UTP in *in-vitro* run off transcription reactions, followed by PAGE purification, as detailed in Section 7.2.4.5. Templates for transcription consisted of annealed oligonucleotides containing a T7 promoter.

Binding of Spt4/5$_{NGN}$ to AA$_{rich}$ and mutated sequences was tested via EMSA as shown in Figure 5.15(B-D). There was no discernible difference in behaviour for any of the sequences. At high protein concentration and without the presence of heparin (a non-specific negatively charged competitor), the RNA fails to move out of the wells, indicating that whatever complex is forming is insoluble and/or too large to enter the wells (Figure 5.15(B)). Heparin was then added to a concentration of 0.03 mg/mL, in line with Amanda Blythe's EMSA data, and this abrogated binding completely (Figure 5.15(C)). The heparin concentration was reduced to 0.01 mg/ml (Figure 5.14(D)), which lessened some of the complex formation in the wells but a clearly shifted band is still not observed, in contrast to Amanda Blythe's data for AA$_{rich}$ that are shown in Figure 1.12(D)). There is, however, loss of the free RNA and concentration-dependent smearing that is indicative of binding. The pattern of smearing is similar for all oligonucleotides.

The experiment was repeated, this time to include two control sequences that Spt4/5$_{NGN}$ is not expected to bind, GG$_{rich}$ and TrimAA$_{rich}$. ^{15}N-HSQC data presented in Figure 5.14(D) failed to detect an interaction between Spt4/5$_{NGN}$ and TrimAA$_{rich}$. Similarly, Amanda Blythe's data shows Spt4/5$_{5K}$ does not bind GG$_{rich}$, by either EMSA or MST analysis (Figure 1.13(A&B)).

As shown in Figure 5.16, Spt4/5$_{NGN}$ interacts with all RNA sequences tested in an analogous manner. In this instance, the acrylamide concentration of the gel was increased from 6% as in Figure 5.15 to 9% and electrophoresis was run for an extra hour in order to try to get better resolution. However, once again, no clearly shifted band was visualised for Spt4/5$_{NGN}$ with AA$_{rich}$. Unexpectedly, the RNA probes for GG$_{rich}$ and TrimAA$_{rich}$ shift from resolved bands to smears with increasing Spt4/5$_{NGN}$ concentration, indistinguishable from the pattern with AA$_{rich}$.

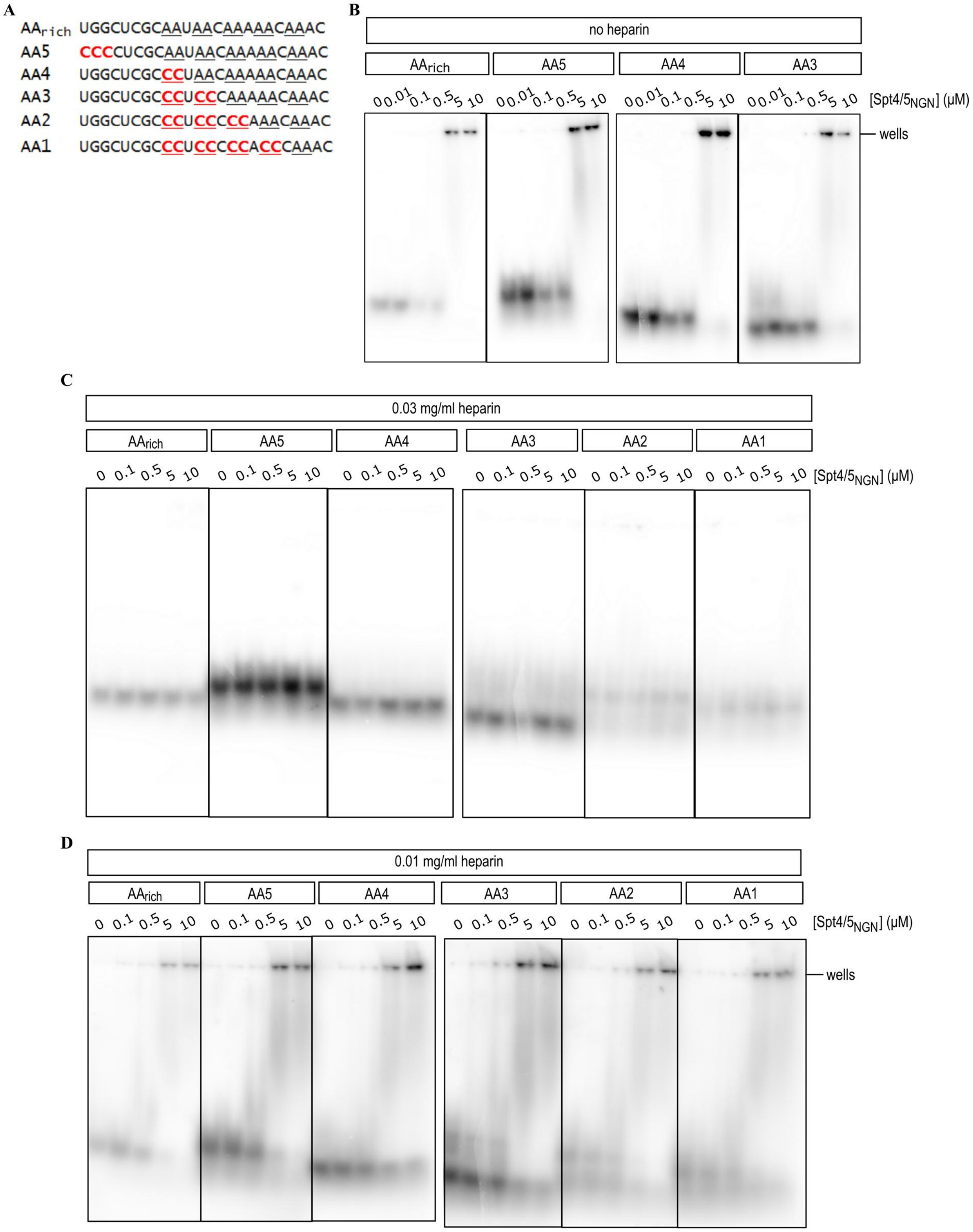


Figure 5.15. EMSAs aimed at determining sequence elements required for Spt4/5$_{NGN}$ binding of AA$_{rich}$.
(A) Sequences of RNA transcripts made by ^{32}-P *in-vitro* transcription to test binding to Spt4/5$_{NGN}$. AA-repeats are underlined, and mutations are indicated in *red*. **(B)** EMSAs of Spt4/5$_{NGN}$ and AA$_{rich}$ and AA5, AA4 and AA3 with no heparin. Increasing concentrations of Spt4/5$_{NGN}$ were incubated with ^{32}P-labelled transcripts then resolved on a 6% polyacrylamide native gel. **(C)** EMSAs of Spt4/5$_{NGN}$ and AA$_{rich}$ and AA5 through AA1 with 0.03 mg/ml heparin. Increasing concentrations of Spt4/5$_{NGN}$ were incubated with ^{32}P-labelled transcripts then resolved on a 6% polyacrylamide native gel. **(D)** EMSAs of Spt4/5$_{NGN}$ and AA$_{rich}$ and AA5 through AA1 with 0.01 mg/ml heparin. Increasing concentrations of Spt4/5$_{NGN}$ were incubated with ^{32}P-labelled transcripts then resolved on a 6% polyacrylamide native gel.

Discrepancies between the EMSA data detailed in this Section and Amanda Blythe's EMSA data presented in Figure 1.12 can perhaps be explained by the protein construct used. Spt4/5$_{\mathrm{NGN}}$ is ubiquitin tagged in Figure 1.13, which helps to keep the protein soluble. In contrast, the data presented in this Section utilises untagged protein, which has always been observed to precipitate at salt concentrations less than 150 mM. Binding conditions for these EMSAs included 150 mM salt, however no salt was contained in the gel or running buffer as this would cause too high a current to achieve electrophoretic motion of the protein and RNA. Therefore, it is likely that Spt4/5$_{\mathrm{NGN}}$ becomes insoluble when loaded on to the gel. This scenario could result in RNA being caught up with the insoluble protein non-specifically due to electrostatic attraction.

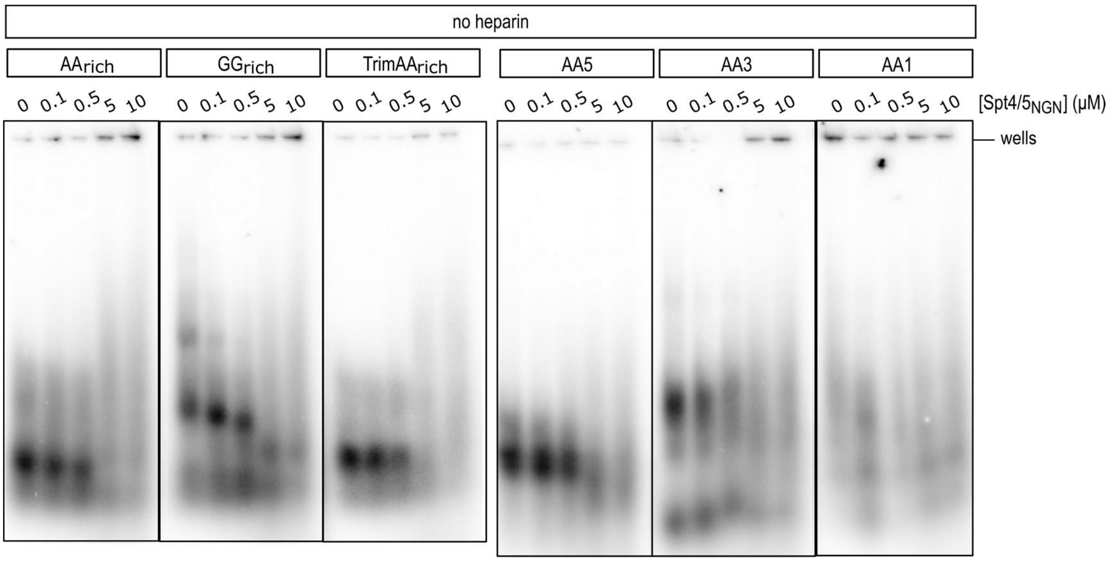


Figure 5.16. Further EMSAs aimed at determining sequence elements required for Spt4/5$_{\mathrm{NGN}}$ binding of AA$_{\mathrm{rich}}$.
EMSAs of Spt4/5$_{\mathrm{NGN}}$ and AA$_{\mathrm{rich}}$ and AA5, AA3 and AA1 with no heparin. Increasing concentrations of Spt4/5$_{\mathrm{NGN}}$ were incubated with ^{32}P-labelled transcripts then resolved on a 9% polyacrylamide native gel.

It is also possible that TrimAA$_{\mathrm{rich}}$ has some extra bases from *in-vitro* transcription, which might explain the different results observed between the MST results in this Section and the NMR binding analysis in Figure 5.14(D), which contained TrimAA$_{\mathrm{rich}}$ that was purchased as chemically synthesised RNA-oligonucleotides.

Nevertheless, failure to resolve a shifted band of untagged Spt4/5$_{\mathrm{NGN}}$ with AA$_{\mathrm{rich}}$, compared with Ubq-Spt4/5$_{\mathrm{NGN}}$, which demonstrates clearly shifted bands in the gel (Figure 1.12(D)), means that EMSAs of untagged Spt4/5$_{\mathrm{NGN}}$ must be deemed a poor technique for binding analysis due to the unfavourable experimental conditions.

5.3.6 Analysis of Spt4/5_{NGN} RNA-binding specificity

Data presented in this Chapter indicate that Spt4/5_{NGN} binds AA_{rich} RNA. The length of the RNA appears to be important given that no chemical shift changes were observed upon addition of TrimAA_{rich}, an oligonucleotide that has three and two bases trimmed off the 5′ and 3′ termini, respectively.

At this stage, it is unclear whether Spt4/5_{NGN} interacts with GG_{rich} as this oligonucleotide was not tested by NMR. Certainly, it seems that Spt4/5_{5K} does not interact with GG_{rich} based on the EMSA and MST data presented in Figure 1.12. In regards to Spt4/5_{NGN}, however; a mixture of this polypeptide with GG_{rich} does exhibit an increase in fluorescence at higher protein concentrations in the MST experiment prior to the establishment of the temperature gradient (Figure 5.17(A)). This is similar to what is seen for AA_{rich}, however the enhancement of fluorescence for GG_{rich} is smaller in magnitude than that seen for AA_{rich}. The increase in pre-MST fluorescence for the Spt4/5_{NGN}:AA_{rich} mixture is around 1.8 fold increase, whereas for Spt4/5_{NGN}:GG_{rich} it is 1.5 fold increase. Curiously, in the case of GG_{rich}, there is initially a decrease in fluorescence of the type seen for 1AA, 2AA and 4AA (Figure 5.13) with increasing Spt4/5_{NGN} concentration, then at [Spt4/5_{NGN}] >1 µM fluorescence is enhanced. AA_{rich} exhibits little change in fluorescence until [Spt4/5_{NGN}] >1 µM after which there is an enhancement of fluorescence.

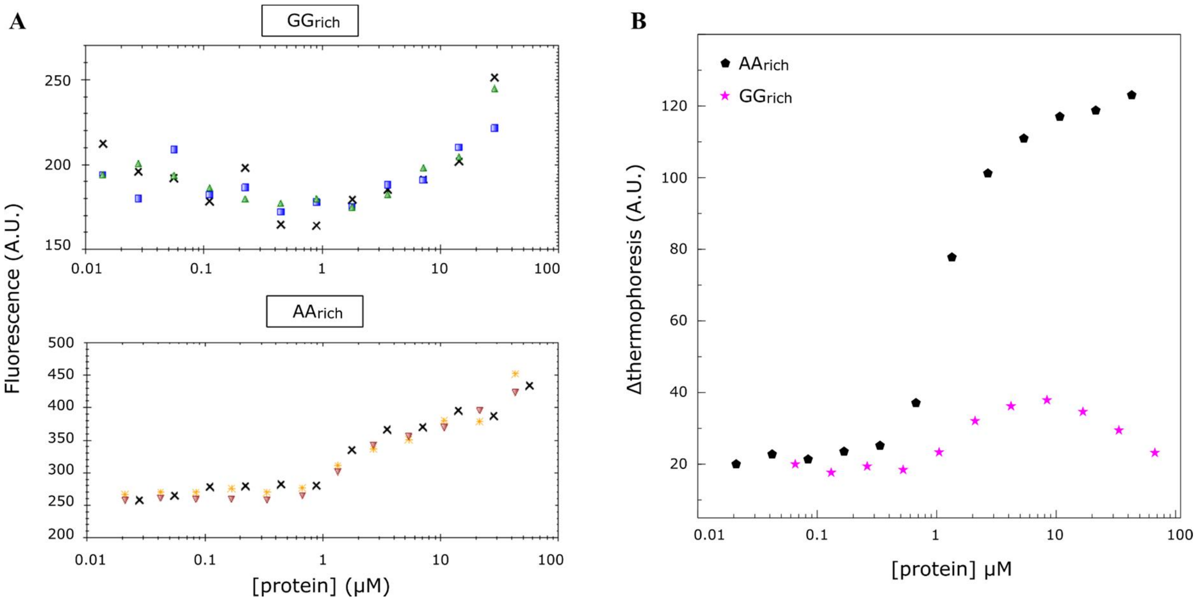


Figure 5.17. MST data for Spt4/5_{NGN} with AA_{rich} and GG_{rich}.
(A) Pre-thermophoretic fluorescence of Spt4/5_{NGN} protein with GG_{rich} and AA_{rich} as indicated. Fluorescence is shown in arbitrary units (A.U.). **(B)** Thermophoresis curves for Spt4/5_{NGN} with AA_{rich} and GG_{rich}. Data points represent the average from three independent titrations. Data were acquired by Amanda Blythe.

There is not, however, the marked change in thermophoresis for GG_{rich} as seen for AA_{rich} (Figure 5.17(B)). For Spt4/5_{NGN}:GG_{rich}, the Δthermophoresis changes direction from an increase in thermophoresis for $1\ \mu M \leq [\text{Spt4/5}_{\text{NGN}}] \leq 10\ \mu M$ to a decrease in thermophoresis for $[\text{Spt4/5}_{\text{NGN}}] > 10\ \mu M$. This behaviour is indicative of multiple events occurring. It is possible that multiple binding events are occurring for Spt4/5$_{\text{NGN}}$:AA$_{\text{rich}}$ as well but they are masked by the same directional change in thermophoresis. This is hinted at by the shape of the MST curve seen in Figure 5.17(B), which is steeper than a sigmoidal curve generated by a 1:1 binding event.

Multiple binding events between Spt4/5$_{\text{NGN}}$ and AA$_{\text{rich}}$ are also indicated by the MST data presented in Figure 1.12(D). Here we see multiple shifted bands, and the shifted bands display a lower electrophoretic mobility than the Spt4/5$_{\text{5K}}$:AA$_{\text{rich}}$ band, indicating that that the species formed between Spt4/5$_{\text{NGN}}$ and AA$_{\text{rich}}$ may be larger than that of Spt4/5$_{\text{5K}}$ and AA$_{\text{rich}}$.

5.3.7 Spt4/5$_{\text{NGN}}$ triple resonance data sets

The RNA-binding behaviour exhibited by Spt4/5$_{\text{NGN}}$ is somewhat unusual. It is difficult to explain the nature of ~30% ^{15}N-HSQC signals disappearing upon AA$_{\text{rich}}$ addition, with no signal changes observed when binding was tested to any of the other AA-repeat sequences tested.

With the aim of resolving some of the unanswered questions about the RNA-binding properties of Spt4/5$_{\text{NGN}}$, resonance assignments of the protein backbone residues were sought using 3D NMR methods, as described for BHD in Section 3.3.2. With assignments in hand, the molecular details of the interaction could begin to be addressed. This required the production of ^{13}C/^{15}N-labelled protein (Section 7.2.2.4), which was unfortunately very poorly behaved. Individual batches were scaled up to three litres in order to compensate for poor behaviour and solubility; however protein concentrations only around 150 μM were obtained. Double labelled protein was expressed three times, with the following data sets collected over the three batches; HNCA, HNCO, HN(CO)CA, HNCACB and CBCACONH (Table 5.1). Each sample was subjected to spectral acquisition for around a week given the low protein concentration.

Unfortunately, these experiments did not yield data sets of sufficient quality to assign the backbone residues of Spt4/5$_{\text{NGN}}$ (Table 5.1). HNCO, HN(CO)CA and HNCA gave the most complete data sets with 80-100% of the expected peaks observed (Figure 5.18(A)). These are typically the most sensitive of the triple resonance experiments. The HNCACB and CBCA(CO)NH datasets, on the other hand, had low levels of completeness (44% and 47% respectively) and were therefore not able to be used.

Sample strips for HNCACB, CBCA(CO)NH, HNCA and HN(CO)CA are shown in Figure 5.18(B). These strips illustrate two problematic features of the data sets which prevented the sequential assignment of Spt4/5$_{\text{NGN}}$ backbone residues. Firstly, due to the relatively large size of the protein by NMR standards (21.6 kDa), there is substantial signal overlap which makes the task of distinguishing signals very difficult. An example of signal overlap is demonstrated in the upper strips of Figure 5.18(B). Secondly, signal intensity for HNCACB and CBCA(CO)NH spectra was usually poor, and often expected signals were not present (expected signals not observed are indicated by empty squares in upper and lower strips of Figure 5.18(B)).

Table 5.1. Triple resonance data sets acquired of Spt4/5$_{\text{NGN}}$ and their completeness. The expected number of peaks was calculated from the number of relevant peaks in the HSQC. As there were more peaks in the HSQC compared with the number of expected peaks based on the heterodimer sequence (225 compared with 183) it is likely that the extra peaks are due to contamination or slow exchange processes.

Experiment	Number of scans	Expected number of peaks	Observed number of peaks	Completeness (%)
HNCA	144	225	176	78
HN(CO)CA	144	225	243	100
HNCO	48	225	268	100
HNCACB	88	900	399	44
CBCA(CO)NH	176	450	213	47

The inability to obtain more complete data sets was due to inadequate protein concentration, coupled with the low signal-to-noise ratio. Additionally, as mentioned above, the lack of resolution in parts of the spectra due to a large number of signals in these areas further complicates assignment.

The diminishing signal-to-noise returns that are available with longer spectrometer experiment time, coupled with the likelihood of protein degradation, means that this strategy is unlikely to yield spectra of sufficient quality to sequentially assign the backbone residues of Spt4/5$_{\text{NGN}}$.

In order to improve the quality of the spectra, the next logical step is to produce perdeuterated, triple labelled protein (^{2}H/^{15}N/^{13}C Spt4/5$_{\text{NGN}}$). The use of deuterium labelling can achieve narrower linewidths by reducing the relaxation of the nuclei and therefore increase signal-to-noise in the spectra. It is costly to produce triple labelled, perdeuterated protein, and an application to have the protein made by the Australian Nuclear Science and Technology Organisation (ANTSO) was submitted and approved. Due to problems encountered by ANSTO with the transformation of BL21 cells with the pETM11-Spt4 and

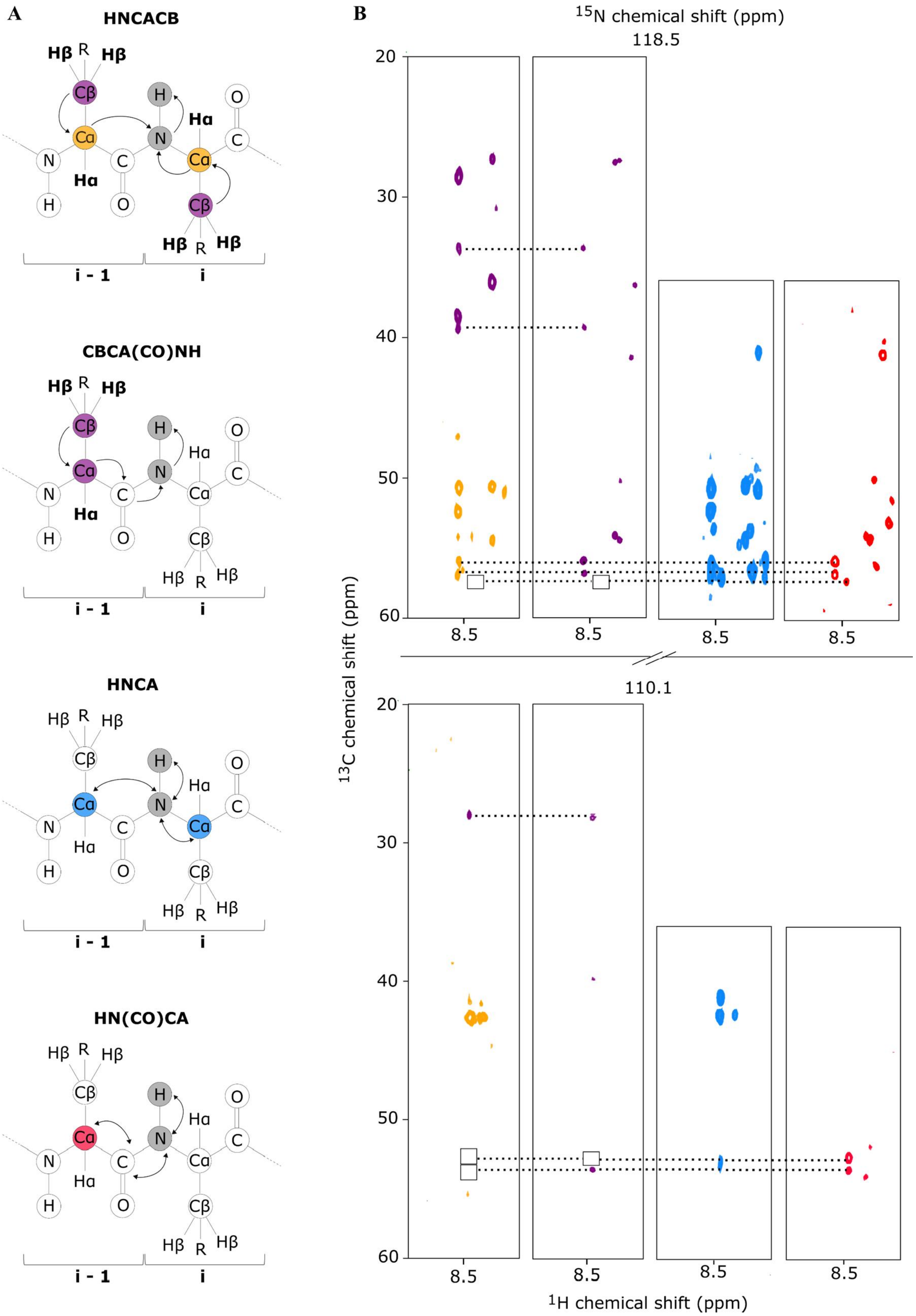


Figure 5.18. Spt4/5NGN triple resonance spectra.
(A) Schematics illustrating the magnetisation transfer steps for each spectrum (indicated by arrows) and atoms which give rise to an NMR signal (shaded). **(B)** HNCACB, CBCA(CO)NH, HNCA and HN(CO)CA strips shown for two nitrogen planes (upper, 118.5 ppm; lower, 110.1 ppm). Dotted lines join the same signal in different spectra, and squares indicate where signals are missing.

pHUE-Spt5$_{NGN}$ plasmids, a significant delay was encountered and the protein was unfortunately unable to be produced before the writing of this Thesis.

5.4 Summary and Discussion

The data presented in this Chapter provides evidence that Spt4/5$_{NGN}$ binds AA-repeat RNA that is longer than 19-nt. NMR-based chemical shift perturbation analysis showed binding only for the 24-nt AA$_{rich}$ oligonucleotide; no binding was observed to AA-repeat RNA that was 19-nt or less in length (Section 5.3.2). These results suggest that the MST binding curves to 4AA, 2AA and 1AA-repeat RNA sequences possibly constitute false positive results that arise due to loss of the fluorescent molecule upon protein addition.

The RNA-binding behaviour of Spt4/5$_{NGN}$ appears to be somewhat complex or unusual on a number of grounds. It is unusual that no binding was visualised by chemical shift mapping to any of a variety of overlapping truncations of the 24-nt AA$_{rich}$ RNA sequence (Section 5.3.4), given that RBPs will exhibit at least some binding to suboptimal sequences [299]. This result indicates that the length of the RNA is important; however given the compact, globular fold of the heterodimer it is difficult to rationalise how more than 19 bases is required to observe any binding at all. It may be that the RNA is forming some sort of transient secondary structure that the heterodimer recognises (and then stabilises), or that binding requires a minimum length of RNA to facilitate binding of a dimer.

Analysis of the stoichiometry of the complex is not straightforward, in part because data acquired to date are not consistent between different techniques. The steep MST binding curve for Spt4/5$_{NGN}$ and AA$_{rich}$ interaction (Figure 5.17(B)) suggest that the interaction is not a simple 1:1 binding event. However, when binding is analysed by NMR, the selective disappearance of ~30% of the heterodimer's signals in the HSQC spectra upon AA$_{rich}$ addition (seen in Figures 5.10 and 5.11) indicates that this complex is unlikely to consist of more than one heterodimer bound to each RNA molecule. This is because, as discussed in Section 5.2.1.1, line broadening due to a slower tumbling time generally affects all residues. If the complex consisted of two heterodimers bound to one AA$_{rich}$ molecule, the molecular weight would be over 55 kDa, and one would reasonably expect all signals to disappear due to the slower tumbling time and faster relaxation of this large complex. This indicates that the perhaps more likely scenario is line-broadening on a 1:1 complex formation due to an intermediate exchange regime.

Contributions from the ubiquitin and fluorescein tags to the MST binding behaviour, however, cannot be ruled out. There are indications to have some concerns about both. Firstly, there is a big increase in fluorescence seen in both MST (Figure 5.13) and EMSA data (Figure 1.12(D)) upon Spt4/5$_{NGN}$ addition. Secondly, ubiquitin alone, when added to AA$_{rich}$, gives rise to a similar change in thermophoresis to

1AA when treated with Spt4/5$_{\mathrm{NGN}}$ (Figure 5.12(B)). Also, a discrete band for a Spt4/5$_{\mathrm{NGN}}$ (without the ubiquitin tag) and AA$_{\mathrm{rich}}$ complex was not observed by EMSA (Section 5.3.5).

Moreover, based on MST data presented in Figure 5.18 it appears likely that Spt4/5$_{\mathrm{NGN}}$ interacts with GG$_{\mathrm{rich}}$ RNA to some degree. In order to further probe the nature and specificity of this binding event, binding to other RNA sequences longer than 19 bases, including GG$_{\mathrm{rich}}$, should be tested by chemical shift mapping.

Taken together, these results indicate that a description of the RNA-binding capacity of Spt4/5$_{\mathrm{NGN}}$ as being specific to AA-repeat RNA is likely to be inaccurate. It remains possible, however, that the longer Spt4/5$_{\mathrm{5K}}$ construct might exhibit more specificity and this issue is worth addressing in additional binding studies.

5.4.1 Recently published structures of eukaryotic RNAPII in complex with Spt4/5

As noted in Section 1.3.1, no structures of eukaryotic Spt4/5 in complex with RNAPs had been published prior to the writing of this thesis. At the time of writing, structures of RNAPII TECs containing Spt4/5 from the yeast *Komagataella pastoris* were published [300], and these data are relevant to the work detailed in this Chapter.

Firstly, Ehara *et al*. (2017) provide evidence that the KOW5 domain from Spt4/5 is the sole component required for the transcription elongation role of Spt4/5 *in vitro*, and based on this information, the co-crystal structure of the *K. pastoris* RNAPII with only this domain from Spt5 was solved (Figure 5.19(A)). In this structure, KOW5 completes the channel where the RNA is funnelled out of the TEC by bridging RNAPII proteins Rpb1 and Rpb2, as well as interacting electrostatically with surrounding RNAPII proteins. These interactions are formed by residues that are frequently conserved between *K. pastoris* and *S. cerevisiae* and *H. sapiens* (Figure 5.19(B)). The location of KOW5 and its proximity to the emerging RNA is consistent with previous results showing crosslinking of KOW5 to the nascent RNA transcript (Section 1.3.2).

Secondly, a cryo-electron microscopy (cryo-EM) structure of the RNAPII TEC in complex with Spt4/5$_{\mathrm{5K}}$ was also reported, and this structure indicates that the NGN domain of Spt5 fills a U-shaped cavity created by Rpb2, and also contacts the Rpb1 clamp (Figure 5.20(A)). Through these interactions the Spt4/5 dimer completes the formation of the DNA exit channel. In doing so, the NGN domain contacts both the departing DNA double helix and the non-template strand, and this is corroborated by reports of the NGN domain contacting DNA [151, 152] as described in Section 1.3.2. The regions of the NGN domain implicated in binding RNAPII and DNA are indicated in Figure 5.20(B&C).

KOW1 is observed to bind to the coiled coil clamp of Rpb1 and thereby also forms part of the DNA exit channel on the side where the RNA exit channel is located. KOW4 is also located near the RNA exit channel, and the authors propose that KOW1, KOW4, and KOW5 extend the RNA exit point into a funnel-like structure. The location of KOW5 is consistent between this structure and the co-crystal structure, providing some confidence that the structures are correct. KOW2 and KOW3 domains from Spt5 were not visualised due to structural heterogeneity, but given the location of the other KOW domains it is reasonable to hypothesise that they could form part of this extended RNA exit channel.

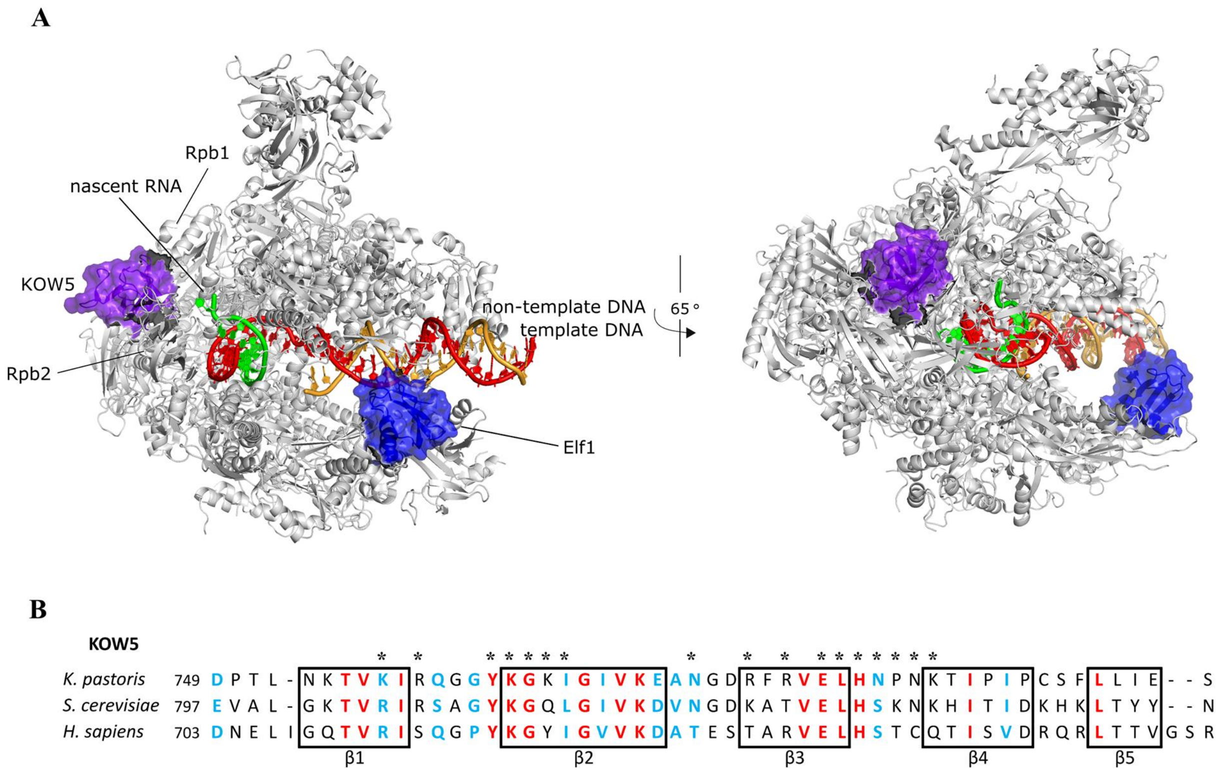

A
Rpb1
nascent RNA
KOW5
Rpb2
non-template DNA
template DNA
65 °
Elf1
B
KOW5
* * * * * * * * * * * * * * * * *
K. pastoris 749 D P T L - N K T V K I R Q G G Y K G K I G I V K E A N G D R F R V E L H N P N K T I P I P C S F L L I E - - S
S. cerevisiae 797 E V A L - G K T V R I R S A G Y K G Q L G I V K D V N G D K A T V E L H S K N K H I T I D K H K L T Y Y - - N
H. sapiens 703 D N E L I G Q T V R I S Q G P Y K G Y I G V V K D A T E S T A R V E L H S T C Q T I S V D R Q R L T T V G S R
β1
β2
β3
β4
β5

Figure 5.19. Co-crystal structure of RNAPII in complex with KOW5.
(A) KOW5 co-crystal structure of RNAPII TEC (PDB: 5XOG). RNAPII subunits are shown in cartoon format in *grey*, the KOW5 domain of Spt5 is shown in surface representation in *purple*, and the elongation factor 1 (Elf1) is shown in surface representation in *blue*. The RNA sequence 5′-UUUUUUUAUCGAGAGGU-3′ is shown in *green*, template DNA (5′-CACTCTACCGATAAGCAGAGCTACCTCTCGATTTTTGGT-3′) is shown in *red* and non-template DNA (5′-AATGGTTTGGCTCTGCTTATCGGTAGAGTG-3′) in *orange*. Regions of Rpb1 and Rpb2 that are contacted by KOW5 are labelled; Rpb1 and Rpb2 are large proteins (over 1000 residues) that extend throughout areas of the RNAPII TEC. **(B)** Sequence alignment of Spt5 KOW5 domains from *K. pastoris*, *S cerevisiae* and *H. sapiens*. Conserved resides are shown in *red* and residues with conserved properties are shown in *blue*, β-strands are outlined and numbered. *K. pastoris* Spt5 KOW5 residues that were observed to interact with RNAPII proteins are indicated above the sequence by asterisks. Data that instructed this figure are from Ehara *et al.* (2017).

5.4.2 Analysis of AA-repeat RNA-binding of Spt4/5 in light of this new structural information

The structural data presented in the above Section provides further evidence that some of the KOW domains are RNA-binding proteins. This is consistent with the data presented in this Chapter and by Blythe *et al.* (2016) that indicate the likelihood that Spt4/5$_{5K}$ binds AA-repeat RNA sequence specifically but not Spt4/5$_{NGN}$; this core heterodimer exhibits less sequence specificity (detailed in Section 5.3.6) but requires a minimum length of between 20 and 24 RNA bases in order to bind. It is therefore likely that the specificity of AA-repeat RNA-binding is provided, at least in part, by the KOW domains.

These new RNAPII TEC structures, however, call in to question whether the Spt4/5$_{NGN}$ core binds RNA as part of its transcription elongation function in the TEC. In the cryo-EM structure, Spt4/5$_{NGN}$ is located away from the RNA exit channel; rather than binding RNA it contacts the upstream DNA duplex and the non-template DNA strand. If Spt4/5$_{NGN}$ indeed binds RNA as part of its role in transcription elongation, then the RNA would need to wrap around this distal face of the heterodimer.

The cryo-EM structure of Spt4/55K in complex with the RNAPII TEC shows that Spt4/51K has many surface exposed basic residues when in complex with the TEC, particularly KOW1 and Spt5NGN, as highlighted in Figure 5.21(A). Many of these basic residues in Spt5NGN are conserved in in S. cerevisiae (shown in Figure 5.21(B)). These basic residues might contribute to Spt4/5NGN RNA-binding in a non-sequence specific fashion, with sequence selectivity achieved by KOW5 and perhaps KOW4. An instructive experiment would be to test Spt4/51K binding of RNA and DNA simultaneously to see if the heterodimer can accommodate both nucleic acids at the same time by using different surfaces. Initially, a traditional EMSA should be carried out (that is, one with an increasing protein concentration and a constant nanomolar concentration of labelled RNA). The protein concentration that lies around the dissociation constant of the interaction would then be chosen, and another EMSA carried out with this protein concentration and the same concentration of labelled RNA (both at a constant level throughout the titration series). In this EMSA an increasing amount of unlabelled DNA target sequence would be added such that the titration sees the ratio of DNA to RNA increase from ~0.1 to 10X. By comparing these EMSAs, the DNA concentration at which RNA-binding is reduced can be assessed and this will be instructive as to whether it is likely that the heterodimer can bind both oligonucleotides simultaneously.

One possible hypothesis that is consistent with the data presented in this Chapter, as well as these recently published structures, is that Spt4/5 binds sequence specifically to nascent mRNA transcripts in order to prompt mRNA processing events. It is known that Spt5 interacts with an assortment of accessory factors such as mRNA capping enzymes [301, 302], and termination and 3′-end processing

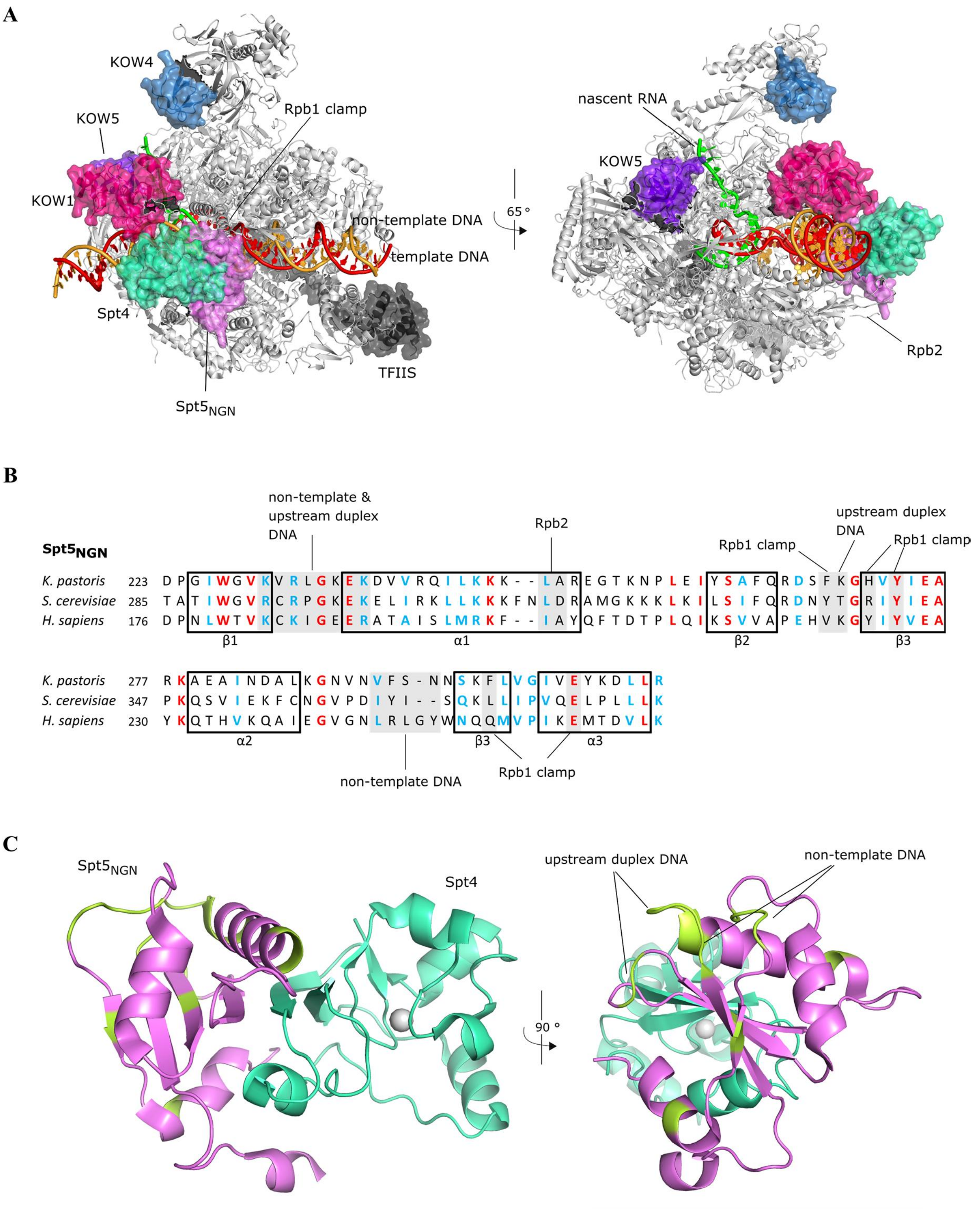

A
KOW4
KOW5
Rpb1 clamp
KOW1
non-template DNA
template DNA
Spt4
TFIIS
Spt5NGN
65 °
nascent RNA
KOW5
Rpb2
B
non-template &
upstream duplex
DNA
Rpb2
upstream duplex
DNA
Rpb1 clamp
Rpb1 clamp
Spt5NGN
K. pastoris 223 D P G I W G V K V R L G K E K D V V R Q I L K K K - - L A R E G T K N P L E I Y S A F Q R D S F K G H V Y I E A
S. cerevisiae 285 T A T I W G V R C R P G K E K E L I R K L L K K K F N L D R A M G K K K L K I L S I F Q R D N Y T G R I Y I E A
H. sapiens 176 D P N L W T V K C K I G E E R A T A I S L M R K F - - I A Y Q F T D T P L Q I K S V V A P E H V K G Y I Y V E A
β1
α1
β2
β3
K. pastoris 277 R K A E A I N D A L K G N V N V F S - N N S K F L V G I V E Y K D L L R
S. cerevisiae 347 P K Q S V I E K F C N G V P D I Y I - - S Q K L L I P V Q E L P L L L K
H. sapiens 230 Y K Q T H V K Q A I E G V G N L R L G Y W N Q Q M V P I K E M T D V L K
α2
β3
α3
non-template DNA
Rpb1 clamp
C
Spt5NGN
Spt4
90 °
upstream duplex DNA
non-template DNA

Figure 5.20. Cryo-EM structures of RNAPII in complex with Spt4 and Spt5$_{5K}$.
(A) Cryo-EM density map of Spt4/5 in complex with RNAPII TEC (PDB: 5XON). RNAP subunits are shown in cartoon format in *grey*, the remaining proteins are shown in surface representation (Spt5$_{NGN}$, *light pink*; Spt5 KOW1 domain, *dark pink*; Spt5 KOW4 domain, *blue*; Spt4, *green-blue*; TFIIS, *black*). The RNA sequence 5′-AUCUUGAAUCUAUUUCUUUUAUCGAGAGGU-3′ is shown in *green*, template DNA (5′-CACTCTACCGATAAGCAGACGTACCTCTCGAC CCTGTGCTAGAC ACGG-3′) is shown in *red* and non-template DNA (5′-CCGTGTCTAGCACAGGGAAAT GGTTTGTGTCTG CTTATCGGTAGAGTG -3′) in *orange*. Regions of Rpb1 and Rpb2 that are contacted by KOW5 are labelled. **(B)** Sequence alignment of Spt5$_{NGN}$ from *K. pastoris*, *S cerevisiae* and *H. sapiens*. Conserved resides are shown in red and residues with conserved properties are shown in *blue*, β-strands are outlined and numbered. Interactions between *K. pastoris* Spt5$_{NGN}$ and Rpb1, Rpb2 and DNA are indicated by *grey* shading. **(C)** Cryo-EM structures of Spt4 (*green-blue* with zinc atom as a *grey* sphere) and Spt5$_{NGN}$ (*magenta*) bound to RNAPII TEC (hidden) (PDB: 5XON). Regions of Spt5 implicated in binding RNAP subunits and DNA are coloured in *limon*. Data that instructed this figure are from Ehara *et al.* (2017).

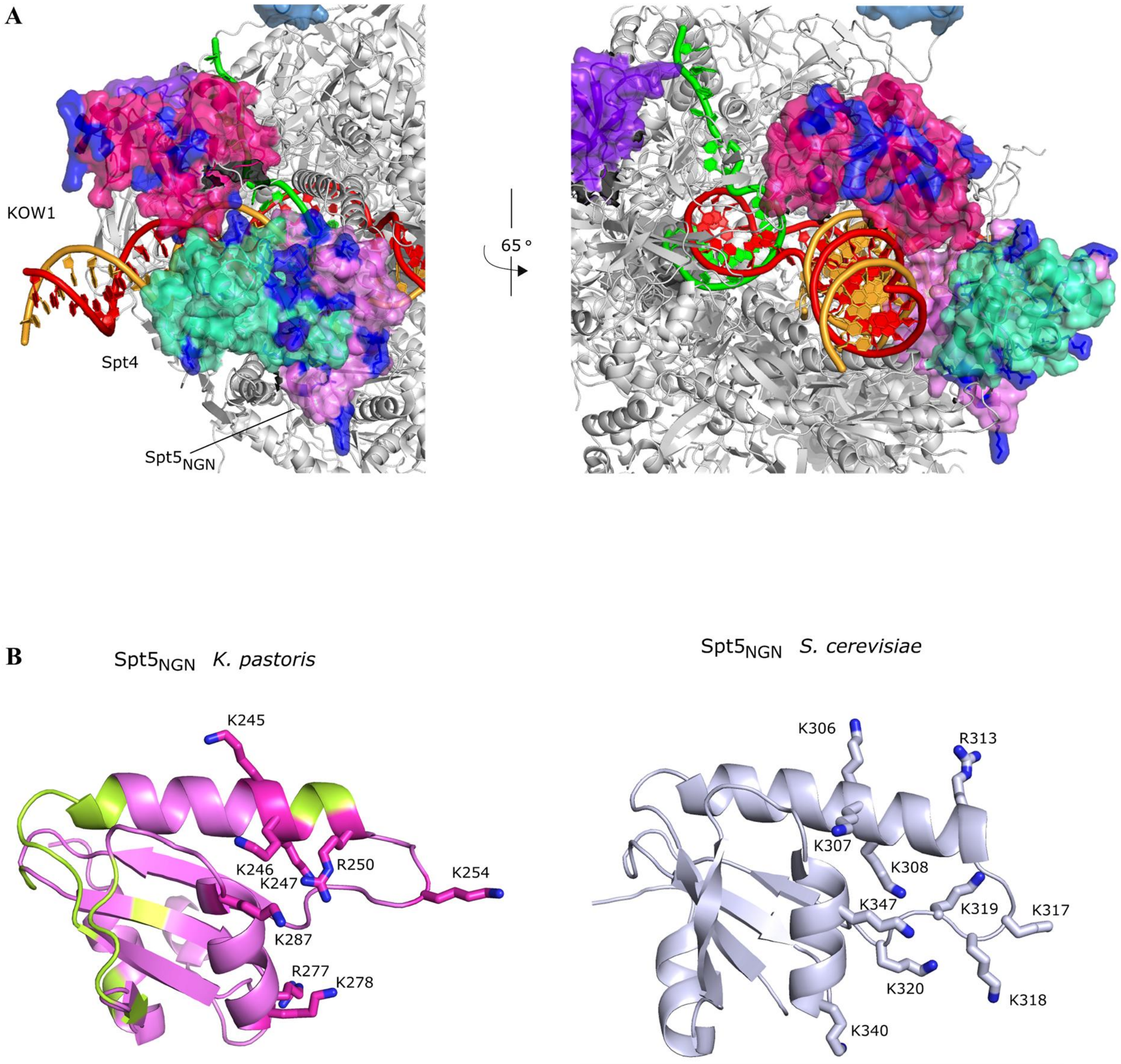


Figure 5.21. NGN domains from both *K. pastoris* and *S. cerevisiae* contain basic residues that point away from the TEC.
(A) Cryo-EM structure of Spt4/5 in complex with RNAPII TEC (PDB: 5XON) from Figure 5.19(A) with surface exposed basic residues of Spt4/5$_{1K}$ in *blue*. **(B)** A high proportion of basic residues (shown in stick format) are conserved on one face of Spt5$_{NGN}$ between *K. pastoris* and *S. cerevisiae*. Residues from *K. pastoris* involved in binding DNA and RNAPII subunits are shown in *limon*, as per Figure 5.19(C).

factors [134, 150]. Spt4/5, through the concerted action of the NGN domain and all five KOW domains, may recognise particular RNA sequences emerging from the TEC to begin downstream processes. For example, the proteins may recognise a polyA or another terminal signal to help disengage a nascent RNA from the complex and destabilise the TEC, or facilitate some other event through the recruitment of accessory factors.

In summary, the data presented in this Chapter indicates complex RNA-binding behaviour of Spt4/5$_{\mathrm{NGN}}$. Although the new published structures of Spt4/5 in complex with the TEC support a role for DNA-binding for this core heterodimer in transcription elongation, it is very unlikely that these structures are representative of its full functionality. The completion of the backbone residue assignments of Spt4/5$_{\mathrm{NGN}}$, as attempted in this Chapter, will determine which regions of the heterodimer are involved in binding RNA, and will help to instruct experiments aimed at elucidating this functionality.

Chapter 6: Concluding discussion

6.1 Introduction

The aims of this Thesis were to elucidate the molecular recognition principles underlying two reported protein-RNA interactions in order to better understand the mechanisms by which these proteins regulate gene expression. These aims were not achieved in full; the inability to reproduce prior data, coupled with the limits of what *in-vitro* experiments can accomplish, saw only partial characterisation of the RNA-binding behaviours of bicoid and Spt4/5. This Chapter summarises and analyses the findings of this Thesis and addresses the challenges facing this field.

6.2 The RNA-binding behaviour of bicoid

6.2.1 BHD is a promiscuous RBD *in vitro*

The studies of the RNA-binding behaviour of the bicoid homeodomain outlined in this Thesis were instructed by reports that the domain bound specifically to the BRE of *cad* mRNA. This result was unable to be reproduced; instead, the domain bound in an indistinguishable fashion to transcripts of varying sequence composition and predicted secondary structure, as observed by EMSA, and the relevant biological *cad* mRNA-binding site could not be ascertained.

With the superior molecular resolution provided by NMR spectroscopy, it was observed that the recognition mechanism used by BHD to bind RNAs of differing base composition and predicted structure is largely similar. Thus it can be concluded that BHD displays a considerable amount of plasticity in recognising RNA; however, the differences in the magnitudes of the chemical shift changes observed for different RNA targets indicate that the domain does indeed have some sequence and/or structural preferences. Additionally, the NMR data showed that α-helix 3 of the domain (the DNA-binding recognition helix) also plays a dominant role in RNA-binding. Thus, it appears likely that the binding mode employed by BHD to bind DNA and RNA bears some similarity, at least to the extent that the same broad regions of the domain are involved in binding both targets. This situation is in contrast to the sole RNA-binding TF for which there is structural information for both DNA and RNA-binding (TFIIIA, see Section 1.2.1), which displays distinct recognition mechanisms for DNA and RNA. Overall, the RNA-binding data of BHD presented in this Thesis are consistent with RNA-binding capacity evolving, in part, through the reported conformational flexibility of residues in this recognition helix [122, 198, 199, 303].

The ability of physiological salt concentration to reduce the affinity of the RNA-binding interaction by around seven-fold, as well as to overwhelmingly abrogate binding induced chemical shift changes, are indications that the domain's RNA-binding ability is driven to a significant extent by electrostatic attraction, and possibly largely constitutes non-specific interactions with RNA phosphate groups. The promiscuous RNA-binding properties of the domain coupled with the salt dependency of complex formation indicates that a biologically specific interaction between bicoid and *cad* would likely require elements outside the homeodomain. This idea adds to the growing amount of data reporting low inherent specificity of RBDs [81], indicating that biological specificity of RBPs for RNA often requires elements additional to individual RBDs.

The failure to observe clear binding preferences for BHD *in vitro* is not overly surprising, firstly, given what we know about RNA recognition by RBPs (see Section 2.5), and secondly, because of the conformational flexibility observed in homeodomains, which results in loose DNA-binding specificity (see Section 1.2.3). Extra DNA-binding specificity is often conferred to homeodomain-containing proteins by additional DNA-binding domains (including ZF domains [304], additional homeodomains [305] or even less common DNA-binding domains such as paired domains [306]), other domains that effect oligomerisation (for example, the ubiquitin-like domain [307]) or additional binding partners (such as other homeodomain proteins as outlined in Section 2.5.1, or other DNA-binding proteins such as c-Jun [308]).

Given that BHD does display some sequence preferences, it seems likely that it binds semi-selectively to RNA with binding sites that can vary substantially, and that target specification is achieved by multiple domains and/or binding partners. Analysis of the RNA-binding properties of the homeodomain in the context of the relevant biological complex may see the salt dependency of the interaction reduced due to the synergistic effects of cooperative binding.

Further complicating the elucidation of RNA-binding specificity of the bicoid homeodomain is the fact that bicoid functions in embryos that are patterned differentially by RBP and RNA gradients. Therefore, specificity in RNA-binding might be modulated by these unique molecular contexts encountered by bicoid throughout the embryo. There are a variety of possibilities for how such specificity might be accomplished; examples include the presence or absence of binding partners dictating which RNA species are bound (see Section 2.5.1), or distinct kinetic contexts which could override an inherently non-specific binding regime (see Section 2.5.2). Furthermore, the interplay of RBP and RNA gradients can determine where phase separated granules occur (see Section 4.6.1), and little is currently known about how specificity is achieved under these special biological conditions.

6.2.2 Reports that bicoid contains an RRM are likely inaccurate

After failing to observe a specific interaction between the bicoid homeodomain and *cad* mRNA, reports of bicoid containing an RRM were investigated. Recombinant production of this domain resulted in a peptide that was not well-ordered, and did not convincingly bind RNA. These results indicate that, excluding the possibility that additional N-terminal sequence is required in order for the domain to fold correctly, this domain is not an RRM, and its constituent RNP-1 motif does not readily bind RNA.

Based on disorder predictions and the disordered nature of this peptide in isolation, it is likely that this region of the protein exhibits low conformational stability in its biological roles. The roles of disorder in RBPs and proteins more broadly are beginning to be elucidated (for instance, their role in phase-separated granules, see Sections 4.4.1 and 4.6) but many questions remain [309]. Given the presence of the RNP-1 motif, it may be that this region of the protein contributes to the formation of *cad* containing RNP granules through weak binding of RNA. Disorder to order transitions necessitate an entropic payoff and therefore other regions of the protein or binding partners may be required in order to observe substantial folding and concomitant RNA binding.

6.2.3 Future directions

6.2.3.1 Determination of the bicoid RNA-binding site

To determine what RNA sequences are bound by bicoid *in vivo*, cross-linking immunoprecipitation (CLIP) methods coupled to high-throughput sequencing should be used in early embryos [310]. These methods involve UV cross-linking of nucleic acids to proteins, followed by partial RNA digestion, immunoprecipitation of the protein of interest, reverse transcription and then sequencing of the resulting transcripts. These techniques are more arduous than *in-vitro* binding assays but given that specificity could not be determined by the latter, application of *in-vivo* techniques will likely be required to determine the domain's biological RNA target sites. By virtue of the nature of *in-vivo* techniques, the effects of post-transcriptional and post-translational modifications as well as the accessibility and concentration of binding sites and that of competitive binders are accounted for. Further, information about what regions of the protein are involved in RNA-binding can be garnered which will be helpful to determine regions additional to the homeodomain that contribute to RNA binding.

6.2.3.2 Analysis of intrinsic specificity

Once RNA sequence or structural motifs that mediate biologically relevant binding by bicoid are identified, the contribution of the homeodomain to RNA binding can be quantified through chemical shift mapping of the domain and RNA target site. The structure of the domain bound to an RNA target site, as well as suboptimal RNA-binding sites, will yield valuable information to help understand how such a small domain can bind both DNA and RNA, as well as illuminating the molecular mechanisms of binding site discrimination.

The sequence preferences of bicoid (if the full-length protein can be successfully overexpressed and purified) and its homeodomain can also be more thoroughly characterised through global profiling methods that result in quantitative affinity distributions (as was done for the *E. coli* protein C5, see Section 2.5.2). These techniques, such as the high-throughput sequencing analysis of equilibrium binding (HiTS-EQ) approach [311] and RNA Bind-n-Seq [312] are more sophisticated than traditional, binary binding assays as they yield superior, quantitative binding information that accounts for the effects of RNA secondary structure. These studies will help to determine whether the intrinsic specificities of the homeodomain determine the RNA sequences bound by the full-length protein *in vivo*, or the extent of other factors at play. Such a comprehensive description of bicoid RNA-binding specificity will be of help not only to understanding its gene regulatory role, but more broadly, how RBPs find their cellular targets in an intertwined regulatory network.

6.2.3.3 How common is dual DNA and RNA-binding behaviour in homeodomains?

The conformational flexibility of the recognition helix of the bicoid homeodomain discussed in Section 6.2.1 seems to be common to homeodomains; there are many studies that report on the dynamics of homeodomain side chain residues which give rise to adaptability in DNA-binding [118, 313-316]. Given this malleability in binding interfaces observed in homeodomains, it might be considered somewhat surprising that no other homeodomains have been reported to bind RNA. It would be interesting to test the RNA-binding potential of other homeodomains through the use of RNA Pentaprobes in binding assays to see if this RNA-binding capacity is indeed unique to bicoid. Also, given that RNA binding by the bicoid homeodomain has been attributed to it being the only known homeodomain with a K50 and R54 combination (see Section 1.2.3), it would be useful to test other K50 homeodomains, as well as BHD with an R54A mutation, in order to ascertain the contribution of individual amino acids to the domain's RNA-binding capacity.

6.2.3.4 Does bicoid effect cad degradation in RNP granules?

The observation that bicoid was not detected as an mRBP in either high-throughput, poly-A capture study in *Drosophila* embryos (see Section 4.4.1) may indicate that *cad* mRNA was degraded or de-adenylated at the time of crosslinking. Indeed, the presence of a conserved *miR-2* binding site in *cad* (as outlined in Section 2.3.3) attests to this possibility, because miRNAs are implicated in mRNA decapping, de-adenylation and degradation. Moreover, recent research has established that this miRNA-mediated silencing can occur in cytoplasmic P-bodies [317] (see Section 4.6.2). The genetic interaction observed between bicoid and Ago2 (see Section 1.2.5), the propensity of Ago2 to localise to P-bodies

[318] as well as the presence of a poly-glutamine domain in bicoid (see Section 4.4.1) may be indications that bicoid-mediated degradation of *cad* occurs in P-bodies, or another type of RNP granule.

As a first step to see if bicoid might form phase separated granules, *in-situ* hybridisation of *cad* in early *Drosophila* embryos using highly sensitive methods developed by Ali-Murthy and Kornberg (2016) [106] could help to visualise the localisation of *cad* mRNA to see if it condenses into granules. Further, single molecule fluorescence studies of GFP-tagged bicoid *in vivo* could help to determine bicoid distribution throughout the syncytial cytoplasm.

6.2.3.5 A possible model for miRNA involvement in bicoid-mediated cad repression

The possibility of *miR-308* directly mediating specificity between BHD and *cad* was investigated in Section 2.3.3, based on a report by Rodel *et al.* (2013) that miRNAs are involved in bicoid-mediated *cad* repression. However, the results were ambiguous; despite the observation of a unique shifted band pattern in EMSAs, no synergy of binding between *mir-308*, *cad* and BHD was observed.

Instead of miRNAs directly mediating specificity between BHD and *cad*, it might be that bicoid interacts with some component/s of RISC, as discussed above. Alternatively, an indirect interaction between bicoid and RISC is also possible, whereby bicoid binds *cad* and induces a structural rearrangement of the mRNA which facilitates RISC binding of the miRNA target site, or vice versa. Such a mechanism was hypothesised to explain the ability of Pumilio to enhance the miRNA mediated repression of the transcription factor E2F3 [319]. Determination of the bicoid binding site on *cad* will help to design experiments that could test the validity of this model and will help to understand the observed cross-talk and synergism between mRBPs and miRNAs [320].

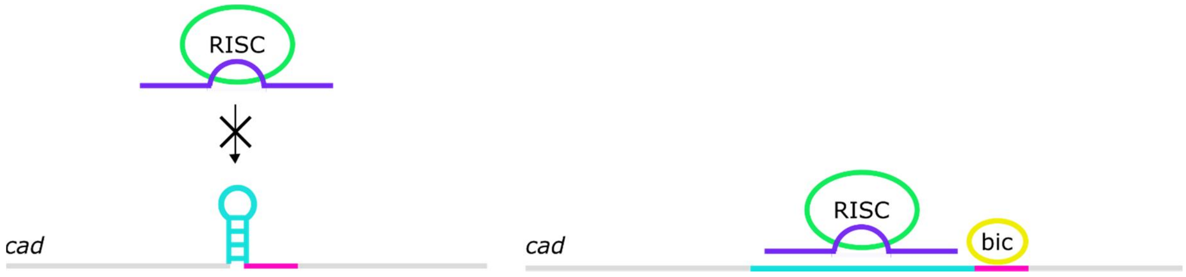


Figure 6.1 A model for miRNA involvement in bicoid mediated *cad* repression.
In this model, RISC is obstructed from binding the *cad* 3′-UTR due to the presence of secondary structure. Bicoid facilitates RISC binding by binding a separate site that disrupts the secdonary structure of the RISC binding site.

6.3 The RNA-binding properties of Spt4/5$_{\text{NGN}}$

6.3.1 Spt4/5$_{\text{NGN}}$ binds RNA of a minimum length

The investigations into RNA-binding by Spt4/5$_{\text{NGN}}$ outlined in this Thesis indicate that a better description of the specificity of the interaction is that the heterodimer binds RNA of a minimum length of somewhere between 20 and 24-nt, rather than short AA-repeat RNA sequences. The chemical shift mapping data for the interaction of Spt4/5$_{\text{NGN}}$ with AA$_{\text{rich}}$ indicate that it is unlikely there is more than one heterodimer bound to each RNA, and this raises the question of how a small heterodimer could require this length of RNA to display any binding at all. The assignment of the ^{15}N-HSQC spectrum would have gone some way to answering this question, and this should be completed in future experiments using perdeuterated, triple-labelled protein in order to increase the signal-to-noise ratio of the acquired 3D spectra.

6.3.2 Future directions

If Spt4/5 does indeed bind RNA sequence specifically *in vivo*, elucidating the potential biological function is likely to be a complex task. It is probable that Spt4/5 binds nucleic acids in a dynamic and transitory fashion, given the nature of transcription and the movement of the transcription elongation complex (TEC) relative to processed nucleic acids. This conclusion is corroborated by the observation that Spt4/5 crosslinks to various transcribed regions, including downstream of polyadenylation sites [321]. A sequence-specific interaction between Spt4/5 and RNA would be expected to be persistent compared with the more transient interactions during transcriptional processing, and therefore may serve to induce a conformational change in the complex or recruit accessory factors.

The characterisation of Spt4/5$_{\text{5K}}$ is more complicated than Spt4/5$_{\text{NGN}}$ due to its larger size (~73 kDa c.f. ~22 kDa). However, the data presented in this Thesis indicates that Spt4/5$_{\text{NGN}}$ does not form a specific 1:1 complex with AA-repeat RNA. This knowledge, coupled with experiments that included KOW domains in binding assays (introduced in Figure 1.12), together indicate that the KOW domains are likely key to the formation of this complex. Further, the recently published crosslinking and structural data implicate the KOW5 domain, in particular, in RNA binding. Taken together, characterisation of Spt4/5 RNA-binding activity – with the inclusion of the five KOW domains – would be of significant interest. As Spt4/5$_{\text{5K}}$ appears to form a specific complex with AA$_{\text{rich}}$ RNA, X-ray crystallography may be a good technique to probe the interaction structurally. Determining which residues are involved in recognising AA-repeat RNA specifically will help to instruct experiments designed to assess the biological relevance of these data.

The RNA-binding capacity of Spt4/5 is likely modulated in some fashion by the TEC and other binding partners, necessitating binding studies in the context of its relevant biological complexes. Newly developed, sophisticated techniques such as single molecule FRET, which has been applied to elucidating the molecular dynamics of RNP complexes such as the ribosome [322], telomerase [323], and splicing RNPs [324], could help to achieve the goal of uncovering the biological relevance of the presented RNA-binding data.

6.4 The prospects for studying RBPs

The work outlined in this Thesis has demonstrated the challenges inherent in examining the interactions between RBPs and their RNA targets, and some of these challenges will be discussed briefly here.

6.4.1 Potential problems with experimental techniques

Both projects in this Thesis were instructed by reports of proteins that exhibited specific RNA-binding behaviour. In the case of bicoid, multiple reports showed a direct interaction between bicoid and *cad*, including cross-linking and 3′-UTR reporter construct data, and an EMSA which demonstrated specific binding of the bicoid homeodomain to *cad* [87, 123, 126]. For Spt4/5, high-affinity RNA-binding sites were identified through the *in-vitro* technique, SELEX, with the biological relevance to be elucidated after characterisation of the interaction.

The aim of both projects in this Thesis was to use *in-vitro* techniques to further define and characterise the RNA-binding determinants of the selected RBPs. In both cases, the techniques applied yielded some information, but specificity determinants were unable to be defined. In the case of the bicoid homeodomain this was because the domain bound RNA promiscuously, and in the case of Spt4/5$_{\mathrm{NGN}}$, because the heterodimer displayed complex binding behaviour that was generally not consistent between techniques.

Comparing the bicoid homeodomain MST data with the EMSA data in Chapter 2, it can be seen that the EMSAs presented are perhaps less sensitive to affinity differences than other methods. The MST data shows a dissociation constant of 3.8 μM for BHD:BRE19nt, and 1.1 μM for HDER:BRE19nt, whereas the EMSA data shows an indistinguishable binding pattern for these two interactions (Figure 2.13). Such differences in affinity could have important biological implications because small differences in binding preferences *in vitro* can have large differences *in vivo* [325].

The limited discriminating power of the EMSAs presented in this Thesis is in part a reflection of the failure of RNA-protein complexes to migrate out of the wells. Given the prevalence of this occurrence throughout this Thesis, it may have been due to a problem with experimental technique. Substantial time and effort was invested in unsuccessful attempts to resolve these RNP complexes. Specifically, the purified, precipitated RNA was resuspended in MQW as well as buffer containing salt (to help the RNA fold correctly), and transcribed RNAs were snap-cooled, and not snap-cooled. Further, different running buffers, voltages and temperatures (room temperature or 4 °C) were trialled. Some differences were observed on occasion but no conditions resulted in the consistent removal of protein-bound RNA species in the wells.

It may be that these species are a result of non-specific interactions between protein and RNA, perhaps resulting in complexes that are too large to enter the wells. Alternatively, it could also be due to some RNA electrostatic effect brought about by the gel conditions, as this effect was never observed in DNA EMSAs.

As always in biochemistry, the application of multiple techniques is required to provide corroborating evidence for outcomes. In the case of bicoid, both NMR and MST techniques showed that an interaction does indeed occur *in vitro*. These techniques provided more information than EMSAs: MST gave detail on the stoichiometry of the interaction (a 1:1 binding event) and NMR spectroscopy yielded resolution at the individual residue level to allow determination of residues and regions of the protein involved in RNA-binding. The situation is a little less clear for Spt4/5, as conflicting data were provided by different techniques. As discussed in Section 5.4, these discrepancies might be attributed to the influence of ubiquitin and fluorescein tags, the inclusion of which create experimental conditions which are one step further removed from biology.

6.4.2 The difficulty in identifying bona fide RBP binding sites

The determination of *in vitro* binding affinities will be sufficient to instruct biological binding sites only in a fraction of instances. This is because, as has been described throughout this Thesis, inherent affinity is only one factor determining biological specificity in RNP interactions. Indeed, many current studies are reporting little inherent RNA-binding specificities for RBPs [208, 326-329]. For example, the two RRMs and connecting linker of hnRNP A1 specifically recognise the core sequence 5′-AG-3′, with the surrounding structural and sequence features determining binding preferences; however, non-AG containing RNAs can also compete for binding [326]. In instances of low reported RNA-binding specificities, proteins can either act as non-specific RBPs or specificity is provided by the cellular context in which the interactions take place. Where RNA-binding specificity cannot be ascertained from *in-vitro* techniques, experiments that take into consideration the cellular conditions that give rise to these interactions will be required.

CLIP techniques, as discussed in Section 6.2.3.1, are at present the most promising method of ascertaining biological RNA-binding sites due to their *in vivo* nature. Such techniques are not without drawbacks, however. For instance, it has been reported that uridines crosslink preferentially [330] and that some proteins don't crosslink well due to the absence of aromatic amino acids near the nucleic acid binding site [331]. Such problems can be addressed to some extent in the experimental protocol, for example by optimising crosslinking for each protein individually, but it is possible that such techniques will not be suitable for all proteins.

Overall, elucidating the determinants of RNA-binding specificity in many instances will be a complex task, and will often require a multifaceted and integrated biochemical approach. Recently developed, sophisticated techniques are likely going to be required. For the most sensitive analysis of intrinsic specificities of RBPs, high-throughput profiling methods (Section 6.2.3.2) should be used. Further, single molecule fluorescence techniques will be increasingly applied as they allow both the visualisation of RNP molecular dynamics (Section 6.3.2) as well as the determination of spatiotemporal regulation of macromolecules under *in vivo* conditions (Section 4.6.2), factors which are now known to be key to RNP interactions. Moreover, the structural characterisation of native RNP complexes is growing, and techniques such as NMR spectroscopy, X-ray crystallography and cryo-EM (Section 5.4) are facilitating this end. Such innovative biochemistry will be required if we are to grasp the underlying biology of our RBP binding data.

6.5 Concluding remarks

The work described in this Thesis has investigated the bicoid homeodomain, and has demonstrated that it is a promiscuous binder of RNA *in vitro*, with the regions of secondary structure used by the domain consistent between RNA and DNA binding. The bicoid homeodomain was shown to have some RNA sequence and/or structural preferences, and whilst further work will be required to prove biological relevance, the data presented in this Thesis are consistent with the domain recognising semi-instructive RNA-binding sites, with other regions of the protein and/or binding partners contributing to its specific translational repression of *cad*. In particular, recent RNP research has highlighted that disordered segments in bicoid may contribute to effecting biological RNA-binding specificity.

This Thesis also detailed the AA-repeat RNA-binding properties of Spt4/5$_{\mathrm{NGN}}$, and has shown that a more accurate description of the RNA binding of this heterodimer is that it binds RNA of a minimum length of 20 or more bases. This binding behaviour is not easily rationalised, and additional experiments aimed at determining the repertoire of residues involved in binding RNA will help to resolve unanswered questions about this interaction.

Chapter 7: Materials and Methods

7.1 Materials

7.1.1 Consumables and reagents

A list of materials used for this Thesis and their suppliers is detailed in Table 7.1.

Table 7.1 Materials and suppliers

Item	Supplier
^{13}C D-glucose	Cambridge Isotope Laboratories (Andover MA, USA)
$^{15}NH_4Cl$	Cambridge Isotope Laboratories (Andover MA, USA)
2,2-dimethyl-2-silapentane-5-sulfonate (DSS)	Fluka A. G. (Buchs, Switzerland)
[α-^{32}P] uridine-5?-triphosphate (UTP)	Perkin Elmer (Melbourne, VIC)
[γ-^{32}P] adenosine triphosphate (ATP)	Perkin Elmer (Melbourne, VIC)
Amicon centrifugal concentrators	Merck Millipore (Bayswater, VIC)
CelluSep®H1 1 kDa MWCO dialysis tubing	Membrane Filtration Products (Seguin TX, USA)
cOmplete™ EDTA-free protease inhibitor tablets	Sigma (Castle Hill, NSW)
Deuterium oxide	Sigma (Castle Hill, NSW)
Dithiothreitol	Quantum Scientific (Milton, QLD)
DNA ladders (2-log, 100 bp)	New England Biolabs (Beverly, MA, USA)
DNA oligonucleotides	Integrated DNA Technologies (Baulkham Hills, NSW)
DNase I	Roche Applied Science (Mannheim, Germany)
dNTPs	New England Biolabs (Beverly MA, USA)
Ethidium bromide	Bio-Rad (Regents Park, NSW)
Glutathione-Sepharose® 4B beads	Amersham Biosciences (Castle Hill, NSW)
Isopropyl β-D-thiogalactopyranoside	Quantum Scientific (Milton, QLD)
Lysozyme	Sigma (Castle Hill, NSW)
Mark12™ Unstained Protein Standards	Thermo Fisher Scientific (Mulgrave, VIC)
Mini Quick Spin RNA columns	Roche Diagnostics (Castle Hill, NSW)
Nickel-NTA Agarose resin	Invitrogen/Thermo Fisher Scientific (Mulgrave, VIC)
PD-10 pre-packed desalting columns	GE Healthcare (Silverwater, NSW)
Phenylmethylsulfonyl fluoride (PMSF)	Sigma (Castle Hill, NSW)
QIAprep® spin miniprep kit	QIAGEN (Doncaster, VIC)
QIAquick® gel extraction kit	QIAGEN (Doncaster, VIC)
QIAquick® PCR purification kit	QIAGEN (Doncaster, VIC)
Quick-Stick ligase	(Bioline, Alexandria, NSW)
Restriction enzymes	New England Biolabs (Arundel, QLD)
RiboSafe RNAse inhibitor	Bioline (Alexandria, NSW)
RNA oligonucleotides	Integrated DNA Technologies (Baulkham Hills, NSW)
RQ1 RNase-Free DNase	Promega (Alexandria, NSW)
T4 DNA ligase	Fermentas (Ontario, Canada)
Triton X-100	Progen (Darra, QLD)
Tween® 20	Astral Scientific (Taren Point, NSW)

7.1.2 Plasmids and bacterial strains

7.1.2.1 Plasmids

Bicoid constructs (BHD, HDER and BRRM): Full length bicoid in the vector pET18b was provided by Michalis Averof (Institut de Génomique Fonctionnelle de Lyon, France), and this was used as a template to make all bicoid constructs cloned into the vector pGEX6P.

Spt4/5: Spt5 in the vector pHUE and Spt4 in the vector pETM11 were provided by Amanda Blythe (University of Western Australia, WA). pHUE-Spt5$_{\mathrm{NGN}}$ was engineered to incorporate a TEV cleavage site by Jason Low in the Mackay laboratory (University of Sydney, NSW).

Cad transcripts: The full length *cad* 3′-UTR gene in the vector pSLfa1180fa was provided by Michalis Averof of Institut de Génomique Fonctionnelle de Lyon (IGFL), and this was used to make templates for *cad* transcription (*cad* 3′-UTR, BRE(257-319), BRE39nt, BRE38nt).

Pentaprobes: Pentaprobe sequences were previously cloned by Fionna Loughlin in the Mackay laboratory (University of Sydney, NSW) and used to make templates for Pentaprobe 7 production.

7.1.2.2 Bacterial strains

DH5α (used for cloning and plasmid propagation): *supE44*, Δ*lac*U169, [Φ80*lacZΔM15*], *hsdR17*($r_K^- m_K^+$), *recA1*, *endA11*, *gyrA1*, *thi-1*, *relA1* (Bethesda Research Laboratories, Gaithersburg, Maryland, USA).

Rosetta(DE3)pLysS (used for protein overexpression): F- *ompT hsdS$_B$* ($r_B^- m_B^-$) *gal dcm* (DE3) pLysSRARE (Cam^R) (Novagen®).

7.1.3 Equipment and suppliers

A list of equipment and their suppliers is contained in Table 7.2.

Table 7.2 Equipment and suppliers

Equipment	Supplier
Bolt^{TM} Mini Gel Tank	ThermoFisher Scientific, Scoresby VIC
Biometra T3000 thermocycler	Biometra, Goettingen, Germany
UNO S1 column	Bio-Rad Laboratories, Hercules CA, USA
Superdex® 75 HiLoad 16/60 column	GE Healthcare, Parramatta, NSW
ProtParam web server	Swiss Institute of Bioinformatics, http://web.expasy.org/protparam
Nanodrop® ND-1000 UV-Vis spectrophotometer	ThermoFisher Scientific, Wilmington DE, USA
Typhoon FLA900 scanner	GE Healthcare, Parramatta NSW
Hoefer SE400 Sturdier™ vertical slab gel unit	GE Healthcare, Parramatta NSW
Monolith NT.115 instrument	NanoTemper Technologies GmbH, München, Germany
Shigemi NMR tubes	Shigemi, Tokyo, Japan
Bruker Avance III NMR spectrometers	Bruker, Karlsruhe, Germany
TOPSPIN3	Bruker, Karlsruhe, Germany

7.2 Methods

7.2.1 Cloning

7.2.1.1 Polymerase chain reaction (PCR)

All sequences for insertion were made by PCR. The reaction mixture consisted of 1 U/50 µL *Pfu* DNA polymerase in Pfu buffer (Table 7.3) with 10% [v/v] DMSO, 0.4 µM forward and reverse primers, 0.1 mM dNTPs and ~1 ng/uL plasmid template. PCR was carried out on a Biometra T3000 thermocycler. PCR programs employed an initial denaturation step of 2 minutes at 95 °C and final extension step of 4-8 minutes at 72 °C, with 30 cycles of the following in between: 30 seconds at 95 °C for primer denaturation, 30 seconds at 47-60 °C for primer annealing, and 2-4 minutes extension at 72 °C.

7.2.1.2 Template and vector processing

PCR products were subject to PCR clean up using a QIAquick® PCR purification kit as per the manufacturer's instructions. Milli-Q® was used to elute PCR products from the spin columns.

Purified PCR products (~1-3 µg) and vectors for insertion (~ 3 µg) were then digested with 10-30 U of both EcoRI and BamHI high-fidelity restriction enzymes (supplied in CutSmart® buffer, Table 7.3) at 37 °C for at least 2 hours.

7.2.1.3 Agarose gel electrophoresis

Restriction digest reactions were mixed 5:1 [v/v] with DNA loading dye and then loaded onto 1-2% [w/v] agarose gels containing 1 µg/mL ethidium bromide, and electrophoresed in TAE running buffer (Table 7.3) at 100 V for ~45 minutes. Gel-separated products were visualised with ultraviolet light and then the relevant bands were excised. DNA was separated from the agarose gel casing using a QIAquick® gel extraction kit as per the manufacturer's protocol.

7.2.1.4 Ligation and transformation

Digested vectors and PCR inserts were then combined in molar ratios ranging from 1:3 through to 1:10 respectively with 0.05% [v/v] Quick-Stick ligase in the supplied buffer and incubated at room temperature for at least 20 minutes. This reaction (~ 5 µL) was then used to transform competent DH5α cells (20 µL of cells with a half volume of KCM). The reaction mixture was heat shocked for 45 seconds at 42 °C, and then recovered at 37 °C for an hour, before being plated out onto LB-agar plates containing 50 µg ampicillin and incubated overnight at 37 °C.

7.2.1.5 Colony PCR

Single colonies were used as templates for PCR reactions by adding cells directly to the PCR mixture with a pipette tip. These PCR reactions contained the forward primer from the vector and the reverse primer from the insert in order to determine if inserts had been successfully ligated in each colony. PCR was carried out as per Section 7.2.1.1, with the reactions then electrophoresed as per Section 7.2.1.3. Colonies which yielded a band of the expected size were selected for plasmid propagation and purification.

7.2.1.6 Plasmid propagation and purification

Selected colonies were used to inoculate ~10 mL of LB (Table 7.3) supplemented with 50 µg ampicillin and incubated overnight with shaking at 180 x g and 37 °C. The cells were pelleted by centrifugation (5000 x g, 5 minutes) and plasmid DNA was purified using a QIAprep® spin miniprep kit according to the manufacturer's instructions. Successful insertions were confirmed by Sanger sequencing, carried out by the Australian Genome Research Facility (Westmead, NSW). Milli-Q® was used to elute the DNA from the spin column, and plasmids were then stored at -20 °C.

Table 7.3. Cloning solutions

Solution	Composition
Pfu buffer	20 mM Tris, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% [v/v] Triton X-100, pH 8.8
KCM	100 mM KCl, 30 mM $CaCl_2$, 50 mM $MgCl_2$
LB	1.0% [w/v] casein peptone, 0.5% [w/v] yeast extract, 0.5% [w/v] NaCl, pH 7.0
LB-agar	LB with 1.5% [w/v] agar
CutSmart® buffer	50 mM potassium acetate, 20 mM Tris-acetate pH 7.9, 10 mM magnesium acetate, 100 µg/mL BSA
TAE running buffer	40 mM Tris, 40 mM glacial acetic acid, 2 mM EDTA

7.2.2 Protein overexpression

7.2.2.1 Transformations

For each transformation, 50 µL of competent *E. coli* Rosetta (DE3) pLysS cells were thawed on ice for at least 20 minutes and 30 µL of KCM (Table 7.3) and ~ 30 ng of plasmid DNA was added and the mixture was left to rest on ice for 30 minutes and then heat shocked at 42 °C for 45 seconds, before the addition of 200 µL of sterile LB. This mixture was incubated with shaking at 180 x g and 37 °C for at least 45 minutes, and then streaked onto an LB-agar plate containing the appropriate antibiotics and incubated at 37 °C overnight.

7.2.2.2 Sequential transformations with calcium chloride competency step

In the special case of Spt4/5$_{NGN}$ expression, which required the transformation of two plasmids, pHUE-Spt5$_{NGN}$ was transformed first as described in the previous Section, and then a colony was selected to inoculate 10 mL of LB and this culture was grown to an optical density measured at 600 nm not in excess of 0.4. The culture was rested on ice for 10 minutes, and then the cells were pelleted by centrifugation (5000 x g, 5 mins, 4 °C). The supernatant was decanted off and the cells were gently resuspended in cold 0.1 M $CaCl_2$ and then incubated on ice for 20 minutes. The cells were then collected by centrifugation again, and then resuspended in 1 mL of cold 0.1 M $CaCl_2$. 500 µL of this solution was added to 30 ng of pETM11-Spt4 and incubated on ice for an hour, before being heat shocked at 42 °C for 45 seconds, and then recovered at 37 °C for an hour. The transformation solution was then streaked onto an agar plate containing 50 µg ampicillin, 35 µg chloramphenicol and 15 µg kanamycin and incubated at 37 °C overnight.

7.2.2.3 Culture growth, induction, expression and harvest

Starter cultures were made by inoculating 10 mL LB (containing the relevant antibiotics) with single colonies from transformation plates, and incubating this culture at 37 °C with shaking (150 x g) until the optical density at 600 nm reached induction values (Table 7.4). Protein expression was induced by the addition of isopropyl β-D-thiogalactopyranoside (IPTG), and expression was carried out for ~17 hours at 18-22 °C with shaking (150 x g). Spt4/5$_{NGN}$ expression required the addition of ZnSO$_4$ (3 mM) at induction as Spt4 is a zinc finger protein. Cells were pelleted by centrifugation at 5000 x g for 15 minutes and either purified immediately or flash frozen in liquid nitrogen and stored at -20 °C.

7.2.2.4 Expression of isotopically labelled protein

Table 7.4. Protein expression details by construct

Construct	Vector	Induction OD$_{600nm}$	Induction conditions	[IPTG] (mM)	Antibiotics
BHD	pGEX6P	0.8	20 °C, ~17 hours	0.5	50 µg/mL Amp, 34 µg/mL Cam
HDER	pGEX6P	0.8	20 °C, ~17 hours	0.5	50 µg/mL Amp, 34 µg/mL Cam
BRRM	pGEX6P	0.6	18 °C, ~17 hours	1	50 µg/mL Amp, 34 µg/mL Cam
Spt4	pETM11	0.6	22 °C, ~17 hours	1	50 µg/mL Amp, 34 µg/mL Cam, 30 µg/mL Kan
Spt5$_{NGN}$	pHUE				

For the expression of ^{15}N or ^{15}N/^{13}C labelled protein, 2 L of culture for every 1 L of unlabelled protein expression was prepared as described in the previous Section, up to the point of reaching an optical density at 600 nm of 0.6-0.8. At this point the cells were harvested by centrifugation at 4000 x g at room temperature, resuspended gently in minimal media (1 L minimal media per 1 L culture, recipe detailed in Table 7.5) before being harvested by centrifugation again and resuspended in 1 L of fresh minimal media supplemented with 1 g of ^{15}NH$_4$Cl for ^{15}N labelled protein, as well as 3 g of ^{13}C D glucose for ^{15}N/^{13}C labelled protein. The cultures were then incubated with shaking at the induction temperature for one hour to clear unlabelled metabolites, and then induction, expression and harvest were carried out as per unlabelled protein.

7.2.3 Protein purification

7.2.3.1 Lysis

Cell pellets from 1 L of expression media were resuspended in 30 mL lysis buffer (see Table 7.6 for recipes for each protein).

In the case of BHD, HDER and BRRM, the resuspended solution was sonicated for 30 seconds, before being incubated at 4 °C for 30 minutes with gentle rocking. DNAse I (100 µg/mL) and MgCl$_2$ (100 µM) were then added and the solution was incubated for another hour at 4 °C with gentle rocking, and then sonicated three to four times for 30 seconds. The insoluble material was pelleted by centrifugation at 10,000 x g for 30 minutes at 4 °C and the soluble fraction was kept for affinity purification.

Table 7.5. Minimal media recipe for the production of ^{15}N and $^{15}N/^{13}C$ labelled proteins

Salts recipe	**g/L**
KH_2PO_4	13
K_2HPO_4	10
Na_2HPO_4	9
K_2SO_4	2.4
Trace metal recipe	**g/L**
$FeCl_2.7H_2O$	6
$CaCl_2.2H_2O$	6
$MnCl_2.4H_2O$	1.2
$COCl_2.6H_2O$	0.8
$ZnSo_4.7H_2O$	0.7
$CuCl_2.2H_2O$	0.3
H_3BO_3	0.02
$(NH_4)_6Mo_7O_{24}.4H_2O$	0.25
EDTA	5
Minimal media recipe	**/L**
Salts recipe	970 mL
Trace metal recipe	10 mL
1 M $MgCl_2$	10 mL
0.1 g/L yeast extract	50 µL
5 mg/mL thiamine	6 mL
$^{15}NH_4Cl$	1 g
glucose*	3 g

*glucose consists of ^{12}C for ^{15}N labelled protein and ^{13}C for $^{15}N/^{13}C$ labelled protein

For Spt4/5$_{NGN}$, the resuspended solution was flash frozen in liquid nitrogen, and then crunched three times with a French press. DNAse I (100 µg/mL) and $MgCl_2$ (100 µM) were then added, and the solution was incubated at 4 °C (no rocking), before the insoluble material was removed by centrifugation at 10,000 x g for 45 minutes at 4 °C.

7.2.3.2 Affinity purification

Glutathione Sepharose® 4B beads (1.5 mL per litre of culture; BHD, HDER and BRRRM) or nickel-nitrilotriacetic acid (nickel-NTA) beads (0.5 mL per litre of culture; Spt4/5$_{NGN}$) were washed with lysis buffer and then incubated with the soluble fraction at 4 °C for one hour, before being washed with

50 mL wash buffer (Table 7.6). Tagged proteins were eluted in six to eight 1 mL fractions with elution buffer (Table 7.6). Purity of eluted fractions was assessed by SDS-PAGE (Section 7.2.3.3).

Table 7.6 Lysis, wash and elution buffers used in protein purification

Lysis buffer	Wash buffer	Elution buffer
BHD and HDER		
50 mM Tris-Cl	50 mM Tris-Cl	50 mM Tris-Cl
500 mM NaCl	300 mM NaCl	150 mM NaCl
1% [v/v] triton X100	1 mM dithiothreitol	1 mM dithiothreitol
1 mM dithiothreitol	5% [v/v] glycerol	50 mM glutathione
2 mM PMSF	pH 8	pH 8
1 mg/mL lysozyme		
pH 8		
BRRM		
50 mM Tris-Cl	50 mM Tris-Cl	50 mM Tris-Cl
1 M NaCl	500 mM NaCl	150 mM NaCl
1% [v/v] triton X100	1 mM dithiothreitol	1 mM dithiothreitol
1 mM dithiothreitol	10% [v/v] glycerol	50 mM glutathione
2 mM PMSF	pH 8	pH 8
1 mg/mL lysozyme		
pH 8		
Spt4/5$_{NGN}$		
50 mM sodium phosphate	50 mM sodium phosphate	50 mM sodium phosphate
2 M KCl	2 M KCl	0.3 M KCl
10% glycerol	10% glycerol	10% glycerol
20 mM imidazole	20 mM imidazole	0.2 M imidazole
1 mM dithiothreitol	1 mM dithiothreitol	1 mM dithiothreitol
1 cOmplete™ tablet	pH 7.4	pH 7.4
pH 7.4		

7.2.3.3 SDS-PAGE

Protein samples were mixed with 1 X LDS (Table 7.7) and heated at 80 °C for 3 minutes prior to loading onto precast Bolt® 4-12% bis-tris polyacrylamide gels along with Mark12^{TM} unstained protein standards in 1 X NuPAGE® MES buffer (Table 7.7) on a Bolt^{TM} Mini Gel Tank at 180 V for 20-22 minutes. Gels were stained with Coommassie Brilliant Blue G-250 (0.1 g/L in MQW) and destained in ROW.

7.2.3.4 Proteolytic cleavage

Samples selected for pooling were identified by SDS-PAGE.

GST-BHD and GST-HDER were dialysed into cation exchange buffer A (Table 7.7) along with HRV-3C protease to remove the N-terminal GST tag at 4 °C overnight.

GST-BRRM was incubated with HRV-3C protease to remove the N-terminal GST tag at 4 °C overnight.

His_6-Ubq-Spt5NGN along with Spt4 were incubated with TEV at 4 °C overnight to remove the His_6-Ubq tag.

7.2.3.5 Cation exchange chromatography

BHD and HDER were purified by cation exchange chromatography. Dialysed protein solutions were filtered and then loaded at 1 ml/min on to a UNOS1 column with cation exchange buffer A and then eluted with a step wise salt gradient with increasing concentration of cation exchange buffer B (Table 7.7). Purity of fractions was assessed by SDS-PAGE (section 7.2.3.3) and relatively pure fractions were pooled.

7.2.3.6 Size exclusion chromatography

Cleavage solutions of BRRM and $Spt4/5_{NGN}$ were filtered and then run over a Superdex® 75 HiLoad 16/60 column at a rate of 0.5 mL/min with the relevant size exclusion buffer (Table 7.7). Purity of fractions was assessed by SDS-PAGE (section 7.2.3.3) and relatively pure fractions were pooled.

Table 7.7. Protein purification solutions

Buffer	Composition
1 X LDS	0.065 M Tris-HCl pH 6.8, 2% [w/v] SDS, 10% [v/v] glycerol, 5% [v/v], β-mercaptoethanol, 0.1% [w/v] bromophenol blue
1 X NuPAGE® MES buffer	50 mM MES, 50 mM Tris pH 7.3, 0.1% [w/v] SDS, 1 mM EDTA
cation exchange buffer A	50 mM Tris-Cl, 50 mM NaCl, 1 mM dithiothreitol pH 8
cation exchange buffer B	50 mM Tris-Cl, 1 M NaCl, 1 mM dithiothreitol pH 8
size exclusion buffer $Spt4/5_{NGN}$	50mM sodium phosphate, 150 mM KCl, 1 mM dithiothreitol pH 7.4
size exclusion buffer BRRM	50 mM Tris-Cl, 150 mM NaCl, 1 mM dithiothreitol pH 8

7.2.3.7 Protein quantification

Protein concentration was assessed by absorbance at 280 nm (A_{280}) using ε_{280nm} estimated from the primary sequence using the ProtParam web server [332]. ε_{280nm} that were used are listed in Table 7.8.

Table 7.8. Protein extinction coefficients at 280 nm

Protein	ε_{280nm} M^{-1} cm^{-1}
GST-BHD	49850
BHD	6990
HDER	6990
Spt4	13980
Spt5$_{NGN}$	9970

7.2.4 RNA preparation

7.2.4.1 Working with RNA

In order to avoid RNAse contamination, appropriate buffers and solutions for RNA work were treated with 0.1% [v/v] diethyl pyrocarbonate (DEPC) for 30 minutes and then autoclaved to remove residual DEPC. Swab solution (Table 7.10) was used to decontaminate bench surfaces and gloves, and equipment such as pipettes and dialysis buttons. RNAse free plasticware was used where possible.

7.2.4.2 Nucleic acid quantification

The concentration of unlabelled nucleic acids was assessed by absorbance at 260 nm (A_{260}) using the supplier's web server ε_{260nm} calculator (http://sg.idtdna.com/calc/analyzer). ε_{260nm} that were used are listed in Table 7.9.

Table 7.9. RNA extinction coefficients at 260 nm

RNA	ε_{260nm} M^{-1} cm^{-1}	RNA	ε_{260nm} M^{-1} cm^{-1}
BRE(257-319)	580400	1AA	55600
BRE38nt	343600	1HAA$_{rich}$	124500
BRE39nt	347800	2HAA$_{rich}$	144400
ShapeControl	377700	TrimAA$_{rich}$	183900
Unst	277200	GG$_{rich}$	220700
mir-308	221200	AA5	235300
mir-Fold	219500	AA4	230500
AA$_{rich}$	241500	AA3	219900
4AA	162000	AA2	210300
T2AA	66200	AA1	200700

The concentration of fluorescently labelled RNAs was determined by A_{260} and adjusted for fluorescein absorbance using the following equation:

$$[RNA] = \frac{A_{260}\,.c_{493}^{Fl} - A_{493}\,.c_{260}^{Fl}}{\varepsilon_{260}^{RNA}\,.\varepsilon_{493}^{Fl}} \qquad \text{Equation 7.1}$$

where ε_{260}^{Fl}= 26000 $M^{-1}cm^{-1}$ and ε_{493}^{Fl}= 74600 $M^{-1}cm^{-1}$.

7.2.4.3 Sample preparation

RNA oligonucleotides less than 20-nt in length were purchased either fluorescein labelled or unlabelled. The lyophilised RNAs were resuspended in DEPC MQW to concentrations between 0.5 and 1 mM.

7.2.4.4 Oligonucleotide annealing

Unlabelled RNA oligonucleotides greater than 25-nt in length and all ^{32}P-labelled RNA were createdby *in-vitro* transcription, using annealed DNA oligonucleotides incorporating a T7 promoter. DNA oligonucleotides were purchased, and the lyophilised DNA oligonucleotides were resuspended in annealing buffer (Table 7.10) to ~200 µM. Equal concentrations of sense and antisense strands were heated to 90 °C for two minutes and then allowed to cool to room temperature slowly.

7.2.4.5 In-vitro transcription

Templates for transcription were made by PCR with primers incorporating a T7 promoter, using either annealed oligonucleotides or plasmid DNA. The following components were assembled at room temperature in the following order: 0.5X transcription buffer, 4 mM DTT, 500 µM rNTPs, DNA template ~0.1 µg/µL, 1 µg/mL pyrophosphatase, 0.5 U/uL RNAse inhibitor, 100 µg/mL T7 RNA polymerase (recombinantly produced in-house). The reaction mixture was incubated for four hours at 37 °C. RQ1 RNAse-free DNAse (1U) was added and the reaction was incubated for a further 30 minutes. RNA was purified by applying the reaction mixture to a Mini Quick Spin RNA column for four minutes at 1000 g, followed by extraction with an equal volume of phenol saturated with 0.1 M citrate, pH 4.3, and then centrifugation at 17000 g for two minutes. The aqueous phase was separated

Table 7.10. Nucleic acid and EMSA preparatory solutions

Solution	Composition
swab solution	0.1 M NaOH and 1 mM EDTA
annealing buffer	10 mM Tris, 50 mM KCl, 1 mM EDTA, pH 7.5
transcription buffer	80 mM HEPES-KOH, 70 mM $MgCl_2$, 1 mg/ml RNAse free BSA, pH 7.5
T4 polynucleotide kinase buffer	70 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT
1 X RNA loading dye	0.016% [w/v] bromophenol blue, 0.04% [w/v] xylene cyanol, 5% [w/v] Ficoll™
1 X TBE buffer	90 mM Tris, 90 mM boric acid, 2.5 mM EDTA
1 X TB buffer	90 mM Tris, 90 mM boric acid, 5 mM $MgCl_2$
EMSA binding buffer	10 mM MOPS pH 7.0, 50 mM KCl, 5 mM $MgCl_2$, 10% [v/v] glycerol, 1 mM DTT
MST binding buffer	50 mM Tris-HCl, 50-150 mM NaCl, 10 mM $MgCl_2$, 0.05 % [v/v] Tween-20, 1 mM DTT, pH 7.5

and vortexed with an equal volume of 24:1 chloroform:isoamyl alcohol and centrifuged again, before RNA precipitation with 0.1 volume of 3 M sodium acetate pH 5.4 and 2.5 volumes of 100% ethanol at -20 °C overnight. The RNA was then pelleted by centrifugation at 17000 g for 45 minutes, washed with 70% [v/v] ethanol and resuspended in RNAse-free MQW and stored at -20 °C.

7.2.5 Electrophoretic mobility shift assays

7.2.5.1 Production of ^{32}P-labelled DNA probes

^{32}P-labelled DNA probes were produced by 5′ end-labelling 10 µM ssDNA with [γ-^{32}P] ATP (10 mCi/mL) with 0.25 U T4 polynucleotide kinase in the supplied buffer (Table 7.10) at 37 °C for one hour. The reaction mixture was then applied to a Mini Quick Spin RNA column for four minutes at 1000 g, before ethanol precipitation. Double-stranded DNA probes were made by labelling one strand with ^{32}P, and then annealing with a four-fold molar excess of the unlabelled complementary strand in annealing buffer (Table 7.10) at 90 °C for two minutes and cooled to room temperature slowly.

7.2.5.2 Production of ^{32}P-labelled RNA probes

^{32}P-labelled RNA probes were produced by *in-vitro* transcription. The same protocol was used as detailed in Section 7.2.4.5, with the concentration of rUTP reduced to ~50 µM with the rest replaced by the addition of [α-^{32}P] UTP. After transcription, the reactions were mixed with 1X RNA loading dye (Table 7.10) before heating at 70 °C for two minutes. The samples were then loaded onto a pre-cast 6% Tris-boric acid urea gel and electrophoresed at 200 V for 20-30 minutes in 1 X TBE buffer (Table 7.10). Gels were exposed on a phosphor screen and imaged on a Typhoon FLA900 scanner. Bands were excised and the RNA was extracted by macerating the gel pieces in MQW and incubating at 37 °C overnight. Gel pieces were pelleted by centrifugation at 17000 g for 30 minutes, the supernatant was removed and 2.5 volumes of 100% ethanol was added and the RNA was precipitated overnight at -20 °C, before being pelleted by centrifugation (17000 g, 45 minutes). The RNA was resuspended in RNAse-free MQW and stored at -20 °C.

7.2.5.3 ^{32}P Native PAGE

Polyacrylamide gels were cast in 1 X TB buffer (Table 7.10) and run at 200 V for 30 minutes in 0.5 X TB buffer. Protein samples were incubated with ~ 10 counts per second of ^{32}P-labelled probes in EMSA binding buffer (Table 7.10) at 4 °C for 30 minutes before being loaded onto the gel. Electrophoresis was carried out at 30-50 mA for one to six hours at 4 °C on a Hoefer SE400 Sturdier™ vertical slab gel unit. Gels were exposed on a phosphor screen and imaged on a Typhoon FLA900 scanner.

7.2.6 Microscale thermophoresis

7.2.6.1 Sample preparation

Protein samples were concentrated in 1 kDa cut-off Microsep centricons. Serial dilutions of protein samples were done to give 12 samples. Fluorescein-labelled RNA was added to each protein dilution to a final concentration of ~50 nM and samples were loaded into capillaries.

7.2.6.2 Data acquisition

Assays were run on a Monolith NT.115 instrument with LED power at 50% (excitation 460-480 nm, emission 515-530 nm) and MST power at 20% (infrared laser for heating of sample, 1480 nm) and the relative fluorescence for each point of the titration was plotted as a function of protein concentration. The data were fitted using Nanotemper software to the model:

$$f(x) = \frac{x+B+K_D-\sqrt{(x+B+K_D)^2-4B}}{(2B)} \qquad \text{Equation 7.2}$$

Where x = [titrated binding partner], B = [constant binding partner], K_D = dissociation constant of first binding phase.

7.2.7 Nuclear magnetic resonance spectroscopy

7.2.7.1 Sample preparation

Protein samples were concentrated in 1 kDa cut-off Microsep centricons or 3 kDa cut-off Vivaspin centricons and then filtered using 0.22 µm spin columns. Deuterium oxide (5-10% [v/v]) and 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS, 200 µM) were added to protein samples before being loaded into 3 or 5 mm Shigemi NMR tubes.

7.2.7.2 Spectral acquisition

Spectra were acquired on either a 600 or 800 MHz Bruker Avance III NMR spectrometers, fitted with cryogenic TCI probes, at 298 K.

The number of scans was increased upon RNA addition to account for protein dilution according to the following equation:

$$NS_{i+1} = NS_i \left(\frac{[P]_i}{[P]_{i+1}}\right)^2 \qquad \text{Equation 7.3}$$

Where NS = number of scans, i = experiment number and [P] = protein concentration.

7.2.7.3 *Spectral processing*

All spectra were processed with TOPSPIN3 and 2D and 3D spectra were analysed in NMRFAM-SPARKY [333]. The ^{1}H frequency scale was referenced using the DSS signal set to 0 ppm. ^{15}N and ^{13}C reference values were calculated from the ^{1}H frequency using the ratios provided by the Biological Magnetic Resonance Data Bank (^{15}N ratio = 0.101329118; ^{13}C ratio = 0.251449530) [334].

7.2.7.4 *Assignment of spectra*

HSQC spectra of RNA-bound states were assigned using the nearest neighbour method [207]. In this method, the assignments of the bound state are used based on proximity to the nearest peak in free protein state, with a correction of 1/7 used for ^{15}N resonances in order to give a roughly equal weighting to the ^{1}H signals.

References

1. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW *et al*: **Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs**. *PLoS Genet* 2006, **2**(4):e62.

2. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al*: **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**(5740):1559-1563.

3. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.

4. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "dark matter" transcripts are associated with known genes**. *PLoS Biol* 2010, **8**(5):e1000371.

5. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A *et al*: **The reality of pervasive transcription**. *PLoS Biol* 2011, **9**(7):e1000625; discussion e1001102.

6. Smith MA, Gesell T, Stadler PF, Mattick JS: **Widespread purifying selection on RNA structure in mammals**. *Nucleic Acids Res* 2013, **41**(17):8220-8236.

7. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP: **Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution**. *Cell* 2011, **147**(7):1537-1550.

8. Taft RJ, Pheasant M, Mattick JS: **The relationship between non-protein-coding DNA and eukaryotic complexity**. *Bioessays* 2007, **29**(3):288-299.

9. Morris KV, Mattick JS: **The rise of regulatory RNA**. *Nat Rev Genet* 2014, **15**(6):423-437.

10. van der Krol AR, Mur LA, de Lange P, Mol JN, Stuitje AR: **Inhibition of flower pigmentation by antisense CHS genes: promoter and minimal sequence requirements for the antisense effect**. *Plant Mol Biol* 1990, **14**(4):457-466.

11. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S: **The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus**. *Cell* 1992, **71**(3):515-526.

12. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF: **The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus**. *Cell* 1992, **71**(3):527-542.

13. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.

14. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans**. *Nature* 2000, **403**(6772):901-906.

15. Guthrie C, Patterson B: **Spliceosomal snRNAs**. *Annu Rev Genet* 1988, **22**:387-419.

16. Morris KV, Chan SW, Jacobsen SE, Looney DJ: **Small interfering RNA-induced transcriptional gene silencing in human cells**. *Science* 2004, **305**(5688):1289-1292.

17. Kim DH, Saetrom P, Snove O, Jr., Rossi JJ: **MicroRNA-directed transcriptional gene silencing in mammalian cells**. *Proc Natl Acad Sci U S A* 2008, **105**(42):16230-16235.

18. Mengardi C, Limousin T, Ricci EP, Soto-Rifo R, Decimo D, Ohlmann T: **microRNAs stimulate translation initiation mediated by HCV-like IRESes**. *Nucleic Acids Res* 2017.

19. Blow MJ, Grocock RJ, van Dongen S, Enright AJ, Dicks E, Futreal PA, Wooster R, Stratton MR: **RNA editing of human microRNAs**. *Genome Biol* 2006, **7**(4):R27.

20. Fernandez-Valverde SL, Taft RJ, Mattick JS: **Dynamic isomiR regulation in Drosophila development**. *RNA* 2010, **16**(10):1881-1888.

21. Pal-Bhadra M, Bhadra U, Birchler JA: **RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila**. *Mol Cell* 2002, **9**(2):315-327.

22. Rajasethupathy P, Antonov I, Sheridan R, Frey S, Sander C, Tuschl T, Kandel ER: **A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity**. *Cell* 2012, **149**(3):693-707.

23. Zhong F, Zhou N, Wu K, Guo Y, Tan W, Zhang H, Zhang X, Geng G, Pan T, Luo H *et al*: **A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes**. *Nucleic Acids Res* 2015, **43**(21):10474-10491.

24. Bachellerie JP, Cavaille J, Huttenhofer A: **The expanding snoRNA world**. *Biochimie* 2002, **84**(8):775-790.

25. Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE *et al*: **Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans**. *Nat Struct Mol Biol* 2010, **17**(8):1030-1034.

26. Yan BX, Ma JX: **Promoter-associated RNAs and promoter-targeted RNAs**. *Cell Mol Life Sci* 2012, **69**(17):2833-2842.

27. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**(5725):1149-1154.

28. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T *et al*: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**(7493):462-470.

29. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM *et al*: **Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome**. *Genome Res* 2006, **16**(1):11-19.

30. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain**. *Proc Natl Acad Sci U S A* 2008, **105**(2):716-721.

31. Fatica A, Bozzoni I: **Long non-coding RNAs: new players in cell differentiation and development**. *Nat Rev Genet* 2014, **15**(1):7-21.

32. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y *et al*: **CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells**. *Science* 2016.

33. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes Dev* 2011, **25**(18):1915-1927.

34. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP: **A coding-independent function of gene and pseudogene mRNAs regulates tumour biology**. *Nature* 2010, **465**(7301):1033-1038.

35. Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI *et al*: **Expression of distinct RNAs from 3' untranslated regions**. *Nucleic Acids Res* 2011, **39**(6):2393-2403.

36. Hashimoto K, Ishida E, Matsumoto S, Shibusawa N, Okada S, Monden T, Satoh T, Yamada M, Mori M: **A liver X receptor (LXR)-beta alternative splicing variant (LXRBSV) acts as an RNA co-activator of LXR-beta**. *Biochem Biophys Res Commun* 2009, **390**(4):1260-1265.

37. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L *et al*: **Coding exons function as tissue-specific enhancers of nearby genes**. *Genome Res* 2012, **22**(6):1059-1068.

38. Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM *et al*: **Exonic transcription factor binding directs codon choice and affects protein evolution**. *Science* 2013, **342**(6164):1367-1372.

39. Lunde BM, Moore C, Varani G: **RNA-binding proteins: modular design for efficient function**. *Nat Rev Mol Cell Biol* 2007, **8**(6):479-490.

40. Afroz T, Cienikova Z, Clery A, Allain FH: **One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence**. *Methods Enzymol* 2015, **558**:235-278.

41. Manival X, Ghisolfi-Nieto L, Joseph G, Bouvet P, Erard M: **RNA-binding strategies common to cold-shock domain- and RNA recognition motif-containing proteins**. *Nucleic Acids Res* 2001, **29**(11):2223-2233.

42. Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, Allain FH: **Molecular basis of RNA recognition by the human alternative splicing factor Fox-1**. *EMBO J* 2006, **25**(1):163-173.

43. Hardin JW, Hu YX, McKay DB: **Structure of the RNA binding domain of a DEAD-box helicase bound to its ribosomal RNA target reveals a novel mode of recognition by an RNA recognition motif**. *J Mol Biol* 2010, **402**(2):412-427.

44. Birney E, Kumar S, Krainer AR: **Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors**. *Nucleic Acids Res* 1993, **21**(25):5803-5816.

45. Clery A, Sinha R, Anczukow O, Corrionero A, Moursy A, Daubner GM, Valcarcel J, Krainer AR, Allain FH: **Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition**. *Proc Natl Acad Sci U S A* 2013, **110**(30):E2802-2811.

46. Dominguez C, Fisette JF, Chabot B, Allain FH: **Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs**. *Nat Struct Mol Biol* 2010, **17**(7):853-861.

47. Valverde R, Edwards L, Regan L: **Structure and function of KH domains**. *FEBS J* 2008, **275**(11):2712-2726.

48. Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, Burley SK: **Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome**. *Cell* 2000, **100**(3):323-332.

49. Loughlin FE, Mansfield RE, Vaz PM, McGrath AP, Setiyaputra S, Gamsjaeger R, Chen ES, Morris BJ, Guss JM, Mackay JP: **The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences**. *Proc Natl Acad Sci U S A* 2009, **106**(14):5581-5586.

50. Peng X, Zhao Y, Cao J, Zhang W, Jiang H, Li X, Ma Q, Zhu S, Cheng B: **CCCH-type zinc finger family in maize: genome-wide identification, classification and expression profiling under abscisic acid and drought treatments**. *PLoS One* 2012, **7**(7):e40120.

51. Hurt JA, Obar RA, Zhai B, Farny NG, Gygi SP, Silver PA: **A conserved CCCH-type zinc finger protein regulates mRNA nuclear adenylation and export**. *J Cell Biol* 2009, **185**(2):265-277.

52. Gao G, Guo X, Goff SP: **Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein**. *Science* 2002, **297**(5587):1703-1706.

53. Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE: **Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d**. *Nat Struct Mol Biol* 2004, **11**(3):257-264.

54. Murn J, Teplova M, Zarnack K, Shi Y, Patel DJ: **Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt**. *Nat Struct Mol Biol* 2016, **23**(1):16-23.

55. Wang X, McLachlan J, Zamore PD, Hall TM: **Modular recognition of RNA by a human pumilio-homology domain**. *Cell* 2002, **110**(4):501-512.

56. Koh YY, Opperman L, Stumpf C, Mandan A, Keles S, Wickens M: **A single C. elegans PUF protein binds RNA in multiple modes**. *RNA* 2009, **15**(6):1090-1099.

57. Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB *et al*: **The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove**. *Cell* 2010, **143**(2):225-237.

58. Tian B, Bevilacqua PC, Diegelman-Parente A, Mathews MB: **The double-stranded-RNA-binding motif: interference and much more**. *Nat Rev Mol Cell Biol* 2004, **5**(12):1013-1023.

59. Ryter JM, Schultz SC: **Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA**. *EMBO J* 1998, **17**(24):7505-7513.

60. Achsel T, Stark H, Luhrmann R: **The Sm domain is an ancient RNA-binding motif with oligo(U) specificity**. *Proc Natl Acad Sci U S A* 2001, **98**(7):3685-3689.

61. Walden WE, Selezneva AI, Dupuy J, Volbeda A, Fontecilla-Camps JC, Theil EC, Volz K: **Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA**. *Science* 2006, **314**(5807):1903-1908.

62. Theler D, Dominguez C, Blatter M, Boudet J, Allain FH: **Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA**. *Nucleic Acids Res* 2014, **42**(22):13911-13919.

63. Dyson HJ: **Roles of intrinsic disorder in protein-nucleic acid interactions**. *Mol Biosyst* 2012, **8**(1):97-104.

64. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL *et al*: **Structure of PTB bound to RNA: specific binding and implications for splicing regulation**. *Science* 2005, **309**(5743):2054-2057.

65. Thandapani P, O'Connor TR, Bailey TL, Richard S: **Defining the RGG/RG motif**. *Mol Cell* 2013, **50**(5):613-623.

66. Vasilyev N, Polonskaia A, Darnell JC, Darnell RB, Patel DJ, Serganov A: **Crystal structure reveals specific recognition of a G-quadruplex RNA by a beta-turn in the RGG motif of FMRP**. *Proc Natl Acad Sci U S A* 2015, **112**(39):E5391-5400.

67. Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M *et al*: **The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts**. *Mol Cell* 2012, **46**(5):674-690.

68. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM *et al*: **Insights into RNA biology from an atlas of mammalian mRNA-binding proteins**. *Cell* 2012, **149**(6):1393-1406.

69. Kwon SC, Yi H, Eichelbaum K, Fohr S, Fischer B, You KT, Castello A, Krijgsveld J, Hentze MW, Kim VN: **The RNA-binding protein repertoire of embryonic stem cells**. *Nat Struct Mol Biol* 2013, **20**(9):1122-1130.

70. Beckmann BM, Horos R, Fischer B, Castello A, Eichelbaum K, Alleaume AM, Schwarzl T, Curk T, Foehr S, Huber W *et al*: **The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs**. *Nat Commun* 2015, **6**:10127.

71. Gerstberger S, Hafner M, Tuschl T: **A census of human RNA-binding proteins**. *Nat Rev Genet* 2014, **15**(12):829-845.

72. Thapar R, Denmon AP, Nikonowicz EP: **Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins**. *Wiley Interdiscip Rev RNA* 2014, **5**(1):49-67.

73. Staple DW, Butcher SE: **Pseudoknots: RNA structures with diverse functions**. *PLoS Biol* 2005, **3**(6):e213.

74. Travers A, Muskhelishvili G: **DNA structure and function**. *FEBS J* 2015, **282**(12):2279-2295.

75. Brazda V, Laister RC, Jagelska EB, Arrowsmith C: **Cruciform structures are a common DNA feature important for regulating biological processes**. *BMC Mol Biol* 2011, **12**:33.

76. Frank-Kamenetskii MD, Mirkin SM: **Triplex DNA structures**. *Annu Rev Biochem* 1995, **64**:65-95.

77. Yan J, Marko JF: **Localized single-stranded bubble mechanism for cyclization of short double helix DNA**. *Phys Rev Lett* 2004, **93**(10):108108.

78. Pan Y, MacKerell AD, Jr.: **Altered structural fluctuations in duplex RNA versus DNA: a conformational switch involving base pair opening**. *Nucleic Acids Res* 2003, **31**(24):7131-7140.

79. Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS, Bazile CA, Burge CB: **RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation**. *Mol Cell* 2016, **64**(2):294-306.

80. Daubner GM, Clery A, Allain FH: **RRM-RNA recognition: NMR or crystallography...and new findings**. *Curr Opin Struct Biol* 2013, **23**(1):100-108.

81. Jankowsky E, Harris ME: **Specificity and nonspecificity in RNA-protein interactions**. *Nat Rev Mol Cell Biol* 2015, **16**(9):533-544.

82. Blythe AJ, Fox AH, Bond CS: **The ins and outs of lncRNA structure: How, why and what comes next?** *Biochim Biophys Acta* 2016, **1859**(1):46-58.

83. Gruber AR, Bernhart SH, Lorenz R: **The ViennaRNA web services**. *Methods Mol Biol* 2015, **1269**:307-326.

84. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Res* 2003, **31**(13):3406-3415.

85. Miao Z, Westhof E: **RNA Structure: Advances and Assessment of 3D Structure Prediction**. *Annu Rev Biophys* 2017, **46**:483-503.

86. Driever W, Nusslein-Volhard C: **The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner**. *Cell* 1988, **54**(1):95-104.

87. Rivera-Pomar R, Niessing D, Schmidt-Ott U, Gehring WJ, Jackle H: **RNA binding and translational suppression by bicoid**. *Nature* 1996, **379**(6567):746-749.

88. Rowland MA, Deeds EJ: **Crosstalk and the evolution of specificity in two-component signaling**. *Proc Natl Acad Sci U S A* 2014, **111**(15):5550-5555.

89. Morikawa M, Koinuma D, Miyazono K, Heldin CH: **Genome-wide mechanisms of Smad binding**. *Oncogene* 2013, **32**(13):1609-1615.

90. Kong W, Yang H, He L, Zhao JJ, Coppola D, Dalton WS, Cheng JQ: **MicroRNA-155 is regulated by the transforming growth factor beta/Smad pathway and contributes to epithelial cell plasticity by targeting RhoA**. *Mol Cell Biol* 2008, **28**(22):6773-6784.

91. Qin W, Chung AC, Huang XR, Meng XM, Hui DS, Yu CM, Sung JJ, Lan HY: **TGF-beta/Smad3 signaling promotes renal fibrosis by inhibiting miR-29**. *J Am Soc Nephrol* 2011, **22**(8):1462-1474.

92. Davis BN, Hilyard AC, Nguyen PH, Lagna G, Hata A: **Smad proteins bind a conserved RNA sequence to promote microRNA maturation by Drosha**. *Mol Cell* 2010, **39**(3):373-384.

93. Ishmael FT, Fang X, Houser KR, Pearce K, Abdelmohsen K, Zhan M, Gorospe M, Stellato C: **The human glucocorticoid receptor as an RNA-binding protein: global analysis of glucocorticoid receptor-associated transcripts and identification of a target RNA motif**. *J Immunol* 2011, **186**(2):1189-1198.

94. Cho H, Park OH, Park J, Ryu I, Kim J, Ko J, Kim YK: **Glucocorticoid receptor interacts with PNRC2 in a ligand-dependent manner to recruit UPF1 for rapid mRNA degradation**. *Proc Natl Acad Sci U S A* 2015, **112**(13):E1540-1549.

95. Shi L, Godfrey WR, Lin J, Zhao G, Kao PN: **NF90 regulates inducible IL-2 gene expression in T cells**. *J Exp Med* 2007, **204**(5):971-977.

96. Bor YC, Swartz J, Morrison A, Rekosh D, Ladomery M, Hammarskjold ML: **The Wilms' tumor 1 (WT1) gene (+KTS isoform) functions with a CTE to enhance translation from an unspliced RNA with a retained intron**. *Genes Dev* 2006, **20**(12):1597-1608.

97. Abdul-Manan N, Williams KR: **hnRNP A1 binds promiscuously to oligoribonucleotides: utilization of random and homo-oligonucleotides to discriminate sequence from base-specific binding**. *Nucleic Acids Res* 1996, **24**(20):4063-4070.

98. Tomonaga T, Levens D: **Heterogeneous nuclear ribonucleoprotein K is a DNA-binding transactivator**. *J Biol Chem* 1995, **270**(9):4875-4881.

99. Ng SY, Bogu GK, Soh BS, Stanton LW: **The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis**. *Mol Cell* 2013, **51**(3):349-359.

100. Sigova AA, Abraham BJ, Ji X, Molinie B, Hannett NM, Guo YE, Jangi M, Giallourakis CC, Sharp PA, Young RA: **Transcription factor trapping by RNA in gene regulatory elements**. *Science* 2015, **350**(6263):978-981.

101. Pelham HR, Brown DD: **A specific transcription factor that can bind either the 5S RNA gene or 5S RNA**. *Proc Natl Acad Sci U S A* 1980, **77**(7):4170-4174.

102. Nolte RT, Conlin RM, Harrison SC, Brown RS: **Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex**. *Proc Natl Acad Sci U S A* 1998, **95**(6):2938-2943.

103. Clemens KR, Wolf V, McBryant SJ, Zhang P, Liao X, Wright PE, Gottesfeld JM: **Molecular basis for specific recognition of both RNA and DNA by a zinc finger protein**. *Science* 1993, **260**(5107):530-533.

104. Lee BM, Xu J, Clarkson BK, Martinez-Yamout MA, Dyson HJ, Case DA, Gottesfeld JM, Wright PE: **Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA**. *J Mol Biol* 2006, **357**(1):275-291.

105. Dubnau J, Struhl G: **RNA recognition and translational regulation by a homeodomain protein**. *Nature* 1996, **379**(6567):694-699.

106. Ali-Murthy Z, Kornberg TB: **Bicoid gradient formation and function in the Drosophila pre-syncytial blastoderm**. *Elife* 2016, **5**.

107. Janody F, Sturny R, Schaeffer V, Azou Y, Dostatni N: **Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway**. *Development* 2001, **128**(12):2281-2290.

108. Fu D, Zhao C, Ma J: **Enhancer sequences influence the role of the amino-terminal domain of bicoid in transcription**. *Mol Cell Biol* 2003, **23**(13):4439-4448.

109. Rechsteiner M, Rogers SW: **PEST sequences and regulation by proteolysis**. *Trends Biochem Sci* 1996, **21**(7):267-271.

110. Niessing D, Dostatni N, Jackle H, Rivera-Pomar R: **Sequence interval within the PEST motif of Bicoid is important for translational repression of caudal mRNA in the anterior region of the Drosophila embryo**. *EMBO J* 1999, **18**(7):1966-1973.

111. Rodel CJ, Gilles AF, Averof M: **MicroRNAs act as cofactors in bicoid-mediated translational repression**. *Curr Biol* 2013, **23**(16):1579-1584.

112. Cho PF, Poulin F, Cho-Park YA, Cho-Park IB, Chicoine JD, Lasko P, Sonenberg N: **A new paradigm for translational control: inhibition via 5'-3' mRNA tethering by Bicoid and the eIF4E cognate 4EHP**. *Cell* 2005, **121**(3):411-423.

113. Stauber M, Jackle H, Schmidt-Ott U: **The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene**. *Proc Natl Acad Sci U S A* 1999, **96**(7):3786-3789.

114. McGregor AP: **How to get ahead: the origin, evolution and function of bicoid**. *Bioessays* 2005, **27**(9):904-913.

115. Moreland RT, Ryan JF, Pan C, Baxevanis AD: **The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family**. *Database (Oxford)* 2009, **2009**:bap004.

116. Holland PW, Booth HA, Bruford EA: **Classification and nomenclature of all human homeobox genes**. *BMC Biol* 2007, **5**:47.

117. Affolter M, Slattery M, Mann RS: **A lexicon for homeodomain-DNA recognition**. *Cell* 2008, **133**(7):1133-1135.

118. Gutmanas A, Billeter M: **Specific DNA recognition by the Antp homeodomain: MD simulations of specific and nonspecific complexes**. *Proteins* 2004, **57**(4):772-782.

119. Hadrys T, DeSalle R, Sagasser S, Fischer N, Schierwater B: **The Trichoplax PaxB gene: a putative Proto-PaxA/B/C gene predating the origin of nerve and sensory cells**. *Mol Biol Evol* 2005, **22**(7):1569-1578.

120. Dardaei L, Longobardi E, Blasi F: **Prep1 and Meis1 competition for Pbx1 binding regulates protein stability and tumorigenesis**. *Proc Natl Acad Sci U S A* 2014, **111**(10):E896-905.

121. Tao G, Kahr PC, Morikawa Y, Zhang M, Rahmani M, Heallen TR, Li L, Sun Z, Olson EN, Amendt BA *et al*: **Pitx2 promotes heart repair by activating the antioxidant response after cardiac injury**. *Nature* 2016, **534**(7605):119-123.

122. Baird-Titus JM, Clark-Baldwin K, Dave V, Caperelli CA, Ma J, Rance M: **The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site**. *J Mol Biol* 2006, **356**(5):1137-1151.

123. Niessing D, Driever W, Sprenger F, Taubert H, Jackle H, Rivera-Pomar R: **Homeodomain position 54 specifies transcriptional versus translational control by Bicoid**. *Mol Cell* 2000, **5**(2):395-401.

124. Driever W, Ma J, Nusslein-Volhard C, Ptashne M: **Rescue of bicoid mutant Drosophila embryos by bicoid fusion proteins containing heterologous activating sequences**. *Nature* 1989, **342**(6246):149-154.

125. Mlodzik M, Gehring WJ: **Hierarchy of the genetic interactions that specify the anteroposterior segmentation pattern of the Drosophila embryo as monitored by caudal protein expression**. *Development* 1987, **101**(3):421-435.

126. Chan SK, Struhl G: **Sequence-specific RNA binding by bicoid**. *Nature* 1997, **388**(6643):634.

127. Zhu W, Hanes SD: **Identification of drosophila bicoid-interacting proteins using a custom two-hybrid selection**. *Gene* 2000, **245**(2):329-339.

128. Jeronimo C, Forget D, Bouchard A, Li Q, Chua G, Poitras C, Therien C, Bergeron D, Bourassa S, Greenblatt J *et al*: **Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme**. *Mol Cell* 2007, **27**(2):262-274.

129. Peterlin BM, Brogie JE, Price DH: **7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription**. *Wiley Interdiscip Rev RNA* 2012, **3**(1):92-103.

130. Singh N, Morlock H, Hanes SD: **The Bin3 RNA methyltransferase is required for repression of caudal translation in the Drosophila embryo**. *Dev Biol* 2011, **352**(1):104-115.

131. Wade JT, Struhl K: **The transition from transcriptional initiation to elongation**. *Curr Opin Genet Dev* 2008, **18**(2):130-136.

132. Sims RJ, 3rd, Belotserkovskaya R, Reinberg D: **Elongation by RNA polymerase II: the short and long of it**. *Genes Dev* 2004, **18**(20):2437-2468.

133. Zhou K, Kuo WH, Fillingham J, Greenblatt JF: **Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5**. *Proc Natl Acad Sci U S A* 2009, **106**(17):6956-6961.

134. Mayer A, Schreieck A, Lidschreiber M, Leike K, Martin DE, Cramer P: **The spt5 C-terminal region recruits yeast 3' RNA cleavage factor I**. *Mol Cell Biol* 2012, **32**(7):1321-1331.

135. Swanson MS, Malone EA, Winston F: **SPT5, an essential gene important for normal transcription in Saccharomyces cerevisiae, encodes an acidic nuclear protein with a carboxy-terminal repeat**. *Mol Cell Biol* 1991, **11**(8):4286.

136. Malone EA, Fassler JS, Winston F: **Molecular and genetic characterization of SPT4, a gene important for transcription initiation in Saccharomyces cerevisiae**. *Mol Gen Genet* 1993, **237**(3):449-459.

137. Mason PB, Struhl K: **Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo**. *Mol Cell* 2005, **17**(6):831-840.

138. Jennings BH: **Pausing for thought: disrupting the early transcription elongation checkpoint leads to developmental defects and tumourigenesis**. *Bioessays* 2013, **35**(6):553-560.

139. Guo G, Gao Y, Zhu Z, Zhao D, Liu Z, Zhou H, Niu L, Teng M: **Structural and biochemical insights into the DNA-binding mode of MjSpt4p:Spt5 complex at the exit tunnel of RNAPII**. *J Struct Biol* 2015, **192**(3):418-425.

140. Chen H, Contreras X, Yamaguchi Y, Handa H, Peterlin BM, Guo S: **Repression of RNA polymerase II elongation in vivo is critically dependent on the C-terminus of Spt5**. *PLoS One* 2009, **4**(9):e6918.

141. Burova E, Hung SC, Sagitov V, Stitt BL, Gottesman ME: **Escherichia coli NusG protein stimulates transcription elongation rates in vivo and in vitro**. *J Bacteriol* 1995, **177**(5):1388-1392.

142. Viktorovskaya OV, Appling FD, Schneider DA: **Yeast transcription elongation factor Spt5 associates with RNA polymerase I and RNA polymerase II directly**. *J Biol Chem* 2011, **286**(21):18825-18833.

143. Martinez-Rucobo FW, Sainsbury S, Cheung AC, Cramer P: **Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity**. *EMBO J* 2011, **30**(7):1302-1310.

144. Hirtreiter A, Damsma GE, Cheung AC, Klose D, Grohmann D, Vojnic E, Martin AC, Cramer P, Werner F: **Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif**. *Nucleic Acids Res* 2010, **38**(12):4040-4051.

145. Bernecky C, Herzog F, Baumeister W, Plitzko JM, Cramer P: **Structure of transcribing mammalian RNA polymerase II**. *Nature* 2016, **529**(7587):551-554.

146. Hartzog GA, Fu J: **The Spt4-Spt5 complex: a multi-faceted regulator of transcription elongation**. *Biochim Biophys Acta* 2013, **1829**(1):105-115.

147. Li W, Giles C, Li S: **Insights into how Spt5 functions in transcription elongation and repressing transcription coupled DNA repair**. *Nucleic Acids Res* 2014, **42**(11):7069-7083.

148. Yamaguchi Y, Wada T, Watanabe D, Takagi T, Hasegawa J, Handa H: **Structure and function of the human transcription elongation factor DSIF**. *J Biol Chem* 1999, **274**(12):8085-8092.

149. Yamada T, Yamaguchi Y, Inukai N, Okamoto S, Mura T, Handa H: **P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation**. *Mol Cell* 2006, **21**(2):227-237.

150. Wier AD, Mayekar MK, Heroux A, Arndt KM, VanDemark AP: **Structural basis for Spt5-mediated recruitment of the Paf1 complex to chromatin**. *Proc Natl Acad Sci U S A* 2013, **110**(43):17290-17295.

151. Crickard JB, Fu J, Reese JC: **Biochemical Analysis of Yeast Suppressor of Ty 4/5 (Spt4/5) Reveals the Importance of Nucleic Acid Interactions in the Prevention of RNA Polymerase II Arrest**. *J Biol Chem* 2016, **291**(19):9853-9870.

152. Yakhnin AV, Murakami KS, Babitzke P: **NusG Is a Sequence-specific RNA Polymerase Pause Factor That Binds to the Non-template DNA within the Paused Transcription Bubble**. *J Biol Chem* 2016, **291**(10):5299-5308.

153. Cheng B, Price DH: **Analysis of factor interactions with RNA polymerase II elongation complexes using a new electrophoretic mobility shift assay**. *Nucleic Acids Res* 2008, **36**(20):e135.

154. Meyer PA, Li S, Zhang M, Yamada K, Takagi Y, Hartzog GA, Fu J: **Structures and Functions of the Multiple KOW Domains of Transcription Elongation Factor Spt5**. *Mol Cell Biol* 2015, **35**(19):3354-3369.

155. Qiu Y, Gilmour DS: **Identification of Regions in the Spt5 Subunit of DSIF That Are Involved in Promoter Proximal Pausing**. *J Biol Chem* 2017.

156. Blythe AJ, Yazar-Klosinski B, Webster MW, Chen E, Vandevenne M, Bendak K, Mackay JP, Hartzog GA, Vrielink A: **The yeast transcription elongation factor Spt4/5 is a sequence-specific RNA binding protein**. *Protein Sci* 2016, **25**(9):1710-1721.

157. Seidel SA, Dijkman PM, Lea WA, van den Bogaart G, Jerabek-Willemsen M, Lazic A, Joseph JS, Srinivasan P, Baaske P, Simeonov A *et al*: **Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions**. *Methods* 2013, **59**(3):301-315.

158. Brown KA, Sharifi S, Hussain R, Donaldson L, Bayfield MA, Wilson DJ: **Distinct Dynamic Modes Enable the Engagement of Dissimilar Ligands in a Promiscuous Atypical RNA Recognition Motif**. *Biochemistry* 2016, **55**(51):7141-7150.

159. Laniel MA, Beliveau A, Guerin SL: **Electrophoretic mobility shift assays for the analysis of DNA-protein interactions**. *Methods Mol Biol* 2001, **148**:13-30.

160. Burge RG, Martinez-Yamout MA, Dyson HJ, Wright PE: **Structural characterization of interactions between the double-stranded RNA-binding zinc finger protein JAZ and nucleic acids**. *Biochemistry* 2014, **53**(9):1495-1510.

161. Yadav DK, Lukavsky PJ: **NMR solution structure determination of large RNA-protein complexes**. *Prog Nucl Magn Reson Spectrosc* 2016, **97**:57-81.

162. Iwakawa HO, Tomari Y: **The Functions of MicroRNAs: mRNA Decay and Translational Repression**. *Trends Cell Biol* 2015, **25**(11):651-665.

163. Czech B, Hannon GJ: **Small RNA sorting: matchmaking for Argonautes**. *Nat Rev Genet* 2011, **12**(1):19-31.

164. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**(1):92-105.

165. Moller T, Franch T, Hojrup P, Keene DR, Bachinger HP, Brennan RG, Valentin-Hansen P: **Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction**. *Mol Cell* 2002, **9**(1):23-30.

166. Klein M, Chandradoss SD, Depken M, Joo C: **Why Argonaute is needed to make microRNA target search fast and reliable**. *Semin Cell Dev Biol* 2017, **65**:20-28.

167. Elkayam E, Kuhn CD, Tocilj A, Haase AD, Greene EM, Hannon GJ, Joshua-Tor L: **The structure of human argonaute-2 in complex with miR-20a**. *Cell* 2012, **150**(1):100-110.

168. Schirle NT, Sheu-Gruttadauria J, MacRae IJ: **Structural basis for microRNA targeting**. *Science* 2014, **346**(6209):608-613.

169. Salomon WE, Jolly SM, Moore MJ, Zamore PD, Serebrov V: **Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides**. *Cell* 2015, **162**(1):84-95.

170. Bayer TS, Booth LN, Knudsen SM, Ellington AD: **Arginine-rich motifs present multiple interfaces for specific binding by RNA**. *RNA* 2005, **11**(12):1848-1857.

171. Lebars I, Martinez-Zapien D, Durand A, Coutant J, Kieffer B, Dock-Bregeon AC: **HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements**. *Nucleic Acids Res* 2010, **38**(21):7749-7763.

172. Jarvelin AI, Noerenberg M, Davis I, Castello A: **The new (dis)order in RNA regulation**. *Cell Commun Signal* 2016, **14**:9.

173. Mackereth CD, Sattler M: **Dynamics in multi-domain protein recognition of RNA**. *Curr Opin Struct Biol* 2012, **22**(3):287-296.

174. Stowell JA, Webster MW, Kogel A, Wolf J, Shelley KL, Passmore LA: **Reconstitution of Targeted Deadenylation by the Ccr4-Not Complex and the YTH Domain Protein Mmi1**. *Cell Rep* 2016, **17**(8):1978-1989.

175. Ipsaro JJ, Joshua-Tor L: **From guide to target: molecular insights into eukaryotic RNA-interference machinery**. *Nat Struct Mol Biol* 2015, **22**(1):20-28.

176. Svitkin YV, Evdokimova VM, Brasey A, Pestova TV, Fantus D, Yanagiya A, Imataka H, Skabkin MA, Ovchinnikov LP, Merrick WC *et al*: **General RNA-binding proteins have a function in poly(A)-binding protein-dependent translation**. *EMBO J* 2009, **28**(1):58-68.

177. Ma JB, Ye K, Patel DJ: **Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain**. *Nature* 2004, **429**(6989):318-322.

178. Lingel A, Simon B, Izaurralde E, Sattler M: **Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain**. *Nat Struct Mol Biol* 2004, **11**(6):576-577.

179. Cruz-Gallardo I, Aroca A, Gunzburg MJ, Sivakumaran A, Yoon JH, Angulo J, Persson C, Gorospe M, Karlsson BG, Wilce JA *et al*: **The binding of TIA-1 to RNA C-rich sequences is driven by its C-terminal RRM domain**. *RNA Biol* 2014, **11**(6):766-776.

180. Wang I, Hennig J, Jagtap PK, Sonntag M, Valcarcel J, Sattler M: **Structure, dynamics and RNA binding of the multi-domain splicing factor TIA-1**. *Nucleic Acids Res* 2014, **42**(9):5949-5966.

181. Cruz-Gallardo I, Aroca A, Persson C, Karlsson BG, Diaz-Moreno I: **RNA binding of T-cell intracellular antigen-1 (TIA-1) C-terminal RNA recognition motif is modified by pH conditions**. *J Biol Chem* 2013, **288**(36):25986-25994.

182. Liu Y, Matthews KS, Bondos SE: **Internal regulatory interactions determine DNA binding specificity by a Hox transcription factor**. *J Mol Biol* 2009, **390**(4):760-774.

183. Muller M, Heym RG, Mayer A, Kramer K, Schmid M, Cramer P, Urlaub H, Jansen RP, Niessing D: **A cytoplasmic complex mediates specific mRNA recognition and localization in yeast**. *PLoS Biol* 2011, **9**(4):e1000611.

184. Noble CG, Walker PA, Calder LJ, Taylor IA: **Rna14-Rna15 assembly mediates the RNA-binding capability of Saccharomyces cerevisiae cleavage factor IA**. *Nucleic Acids Res* 2004, **32**(11):3364-3375.

185. Leeper TC, Qu X, Lu C, Moore C, Varani G: **Novel protein-protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and Hrp1**. *J Mol Biol* 2010, **401**(3):334-349.

186. Kuwasako K, Takahashi M, Unzai S, Tsuda K, Yoshikawa S, He F, Kobayashi N, Guntert P, Shirouzu M, Ito T *et al*: **RBFOX and SUP-12 sandwich a G base to cooperatively regulate tissue-specific splicing**. *Nat Struct Mol Biol* 2014, **21**(9):778-786.

187. Hennig J, Militti C, Popowicz GM, Wang I, Sonntag M, Geerlof A, Gabel F, Gebauer F, Sattler M: **Structural basis for the assembly of the Sxl-Unr translation regulatory complex**. *Nature* 2014, **515**(7526):287-290.

188. Weidmann CA, Qiu C, Arvola RM, Lou TF, Killingsworth J, Campbell ZT, Tanaka Hall TM, Goldstrohm AC: **Drosophila Nanos acts as a molecular clamp that modulates the RNA-binding and repression activities of Pumilio**. *Elife* 2016, **5**.

189. Peters DT, Fung HK, Levdikov VM, Irmscher T, Warrander FC, Greive SJ, Kovalevskiy O, Isaacs HV, Coles MC, Antson AA: **Human Lin28 forms a high-affinity 1:1 complex with the 106~363 cluster miRNA miR-363**. *Biochemistry* 2016.

190. Mitchell SF, Parker R: **Principles and properties of eukaryotic mRNPs**. *Mol Cell* 2014, **54**(4):547-558.

191. Moursy A, Allain FH, Clery A: **Characterization of the RNA recognition mode of hnRNP G extends its role in SMN2 splicing regulation**. *Nucleic Acids Res* 2014, **42**(10):6659-6672.

192. Mustoe AM, Brooks CL, Al-Hashimi HM: **Hierarchy of RNA functional dynamics**. *Annu Rev Biochem* 2014, **83**:441-466.

193. Duss O, Michel E, Diarra dit Konte N, Schubert M, Allain FH: **Molecular basis for the wide range of affinity found in Csr/Rsm protein-RNA recognition**. *Nucleic Acids Res* 2014, **42**(8):5332-5346.

194. Milek M, Wyler E, Landthaler M: **Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing**. *Semin Cell Dev Biol* 2012, **23**(2):206-212.

195. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y: **Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts**. *Genome Res* 2009, **19**(3):381-394.

196. Thickman KR, Sickmier EA, Kielkopf CL: **Alternative conformations at the RNA-binding surface of the N-terminal U2AF(65) RNA recognition motif**. *J Mol Biol* 2007, **366**(3):703-710.

197. Kotik-Kogan O, Valentine ER, Sanfelice D, Conte MR, Curry S: **Structural analysis reveals conformational plasticity in the recognition of RNA 3' ends by the human La protein**. *Structure* 2008, **16**(6):852-862.

198. Dave V, Zhao C, Yang F, Tung CS, Ma J: **Reprogrammable recognition codes in bicoid homeodomain-DNA interaction**. *Mol Cell Biol* 2000, **20**(20):7673-7684.

199. Adhikary R, Tan YX, Liu J, Zimmermann J, Holcomb M, Yvellez C, Dawson PE, Romesberg FE: **Conformational Heterogeneity and DNA Recognition by the Morphogen Bicoid**. *Biochemistry* 2017, **56**(22):2787-2793.

200. Kaneko S, Li G, Son J, Xu CF, Margueron R, Neubert TA, Reinberg D: **Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA**. *Genes Dev* 2010, **24**(23):2615-2620.

201. Siomi MC, Sato K, Pezic D, Aravin AA: **PIWI-interacting small RNAs: the vanguard of genome defence**. *Nat Rev Mol Cell Biol* 2011, **12**(4):246-258.

202. Herzog VA, Lempradl A, Trupke J, Okulski H, Altmutter C, Ruge F, Boidol B, Kubicek S, Schmauss G, Aumayr K *et al*: **A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element**. *Nat Genet* 2014, **46**(9):973-981.

203. Guenther UP, Yandek LE, Niland CN, Campbell FE, Anderson D, Anderson VE, Harris ME, Jankowsky E: **Hidden specificity in an apparently nonspecific RNA-binding protein**. *Nature* 2013, **502**(7471):385-388.

204. Ayed A, Mulder FA, Yi GS, Lu Y, Kay LE, Arrowsmith CH: **Latent and active p53 are identical in conformation**. *Nat Struct Biol* 2001, **8**(9):756-760.

205. Seavey BR, Farr EA, Westler WM, Markley JL: **A relational database for sequence-specific protein NMR data**. *J Biomol NMR* 1991, **1**(3):217-236.

206. Shen Y, Bax A: **Identification of helix capping and b-turn motifs from NMR chemical shifts**. *J Biomol NMR* 2012, **52**(3):211-232.

207. Williamson RA, Carr MD, Frenkiel TA, Feeney J, Freedman RB: **Mapping the binding site for matrix metalloproteinase on the N-terminal domain of the tissue inhibitor of metalloproteinases-2 by NMR chemical shift perturbation**. *Biochemistry* 1997, **36**(45):13882-13889.

208. Wai DC, Shihab M, Low JK, Mackay JP: **The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity**. *Nucleic Acids Res* 2016, **44**(19):9153-9165.

209. Hargous Y, Hautbergue GM, Tintaru AM, Skrisovska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH: **Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8**. *EMBO J* 2006, **25**(21):5126-5137.

210. Wüthrich K: **NMR of Proteins and Nucleic Acids**. New York: Wiley; 1986.

211. Loughlin FE, Lee M, Guss JM, Mackay JP: **Crystallization of a ZRANB2-RNA complex**. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2008, **64**(Pt 12):1175-1177.

212. Skinner AL, Laurence JS: **High-field solution NMR spectroscopy as a tool for assessing protein interactions with small molecule ligands**. *J Pharm Sci* 2008, **97**(11):4670-4695.

213. Bendak K, Loughlin FE, Cheung V, O'Connell MR, Crossley M, Mackay JP: **A rapid method for assessing the RNA-binding potential of a protein**. *Nucleic Acids Res* 2012, **40**(14):e105.

214. Parker MW, Lo Bello M, Federici G: **Crystallization of glutathione S-transferase from human placenta**. *J Mol Biol* 1990, **213**(2):221-222.

215. Burz DS, Rivera-Pomar R, Jackle H, Hanes SD: **Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo**. *EMBO J* 1998, **17**(20):5998-6009.

216. Burz DS, Hanes SD: **Isolation of mutations that disrupt cooperative DNA binding by the Drosophila bicoid protein**. *J Mol Biol* 2001, **305**(2):219-230.

217. Yuan D, Ma X, Ma J: **Sequences outside the homeodomain of bicoid are required for protein-protein interaction**. *J Biol Chem* 1996, **271**(35):21660-21665.

218. Helfer S, Schott J, Stoecklin G, Forstemann K: **AU-rich element-mediated mRNA decay can occur independently of the miRNA machinery in mouse embryonic fibroblasts and Drosophila S2-cells**. *PLoS One* 2012, **7**(1):e28907.

219. Stauber M, Jäckle H, Schmidt-Ott U: **The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene**. *Proceedings of the National Academy of Sciences* 1999, **96**(7):3786-3789.

220. Lynch J, Desplan C: **Evolution of development: beyond bicoid**. *Curr Biol* 2003, **13**(14):R557-559.

221. Rado-Trilla N, Alba M: **Dissecting the role of low-complexity regions in the evolution of vertebrate proteins**. *BMC Evol Biol* 2012, **12**:155.

222. Wessels HH, Imami K, Baltz AG, Kolinski M, Beldovskaya A, Selbach M, Small S, Ohler U, Landthaler M: **The mRNA-bound proteome of the early fly embryo**. *Genome Res* 2016, **26**(7):1000-1009.

223. Sysoev VO, Fischer B, Frese CK, Gupta I, Krijgsveld J, Hentze MW, Castello A, Ephrussi A: **Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila**. *Nat Commun* 2016, **7**:12128.

224. Courchaine EM, Lu A, Neugebauer KM: **Droplet organelles?** *EMBO J* 2016, **35**(15):1603-1612.

225. Lancaster AK, Nutter-Upham A, Lindquist S, King OD: **PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition**. *Bioinformatics* 2014, **30**(17):2501-2502.

226. Newby GA, Lindquist S: **Blessings in disguise: biological benefits of prion-like mechanisms**. *Trends Cell Biol* 2013, **23**(6):251-259.

227. Stanek D, Fox A: **Nuclear bodies: news insights into structure and function**. *Curr Opin Cell Biol* 2017, **46**:94-101.

228. Hennig S, Kong G, Mannen T, Sadowska A, Kobelke S, Blythe A, Knott GJ, Iyer KS, Ho D, Newcombe EA *et al*: **Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles**. *J Cell Biol* 2015, **210**(4):529-539.

229. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT: **Scalable web services for the PSIPRED Protein Analysis Workbench**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W349-357.

230. Uversky VN, Kuznetsova IM, Turoverov KK, Zaslavsky B: **Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates**. *FEBS Lett* 2015, **589**(1):15-22.

231. Kato M, McKnight SL: **Cross-beta Polymerization of Low Complexity Sequence Domains**. *Cold Spring Harb Perspect Biol* 2016.

232. Li P, Banjade S, Cheng HC, Kim S, Chen B, Guo L, Llaguno M, Hollingsworth JV, King DS, Banani SF *et al*: **Phase transitions in the assembly of multivalent signalling proteins**. *Nature* 2012, **483**(7389):336-340.

233. Elbaum-Garfinkle S, Kim Y, Szczepaniak K, Chen CC, Eckmann CR, Myong S, Brangwynne CP: **The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics**. *Proc Natl Acad Sci U S A* 2015, **112**(23):7189-7194.

234. Lin Y, Protter DS, Rosen MK, Parker R: **Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins**. *Mol Cell* 2015, **60**(2):208-219.

235. Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD *et al*: **Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles**. *Mol Cell* 2015, **57**(5):936-947.

236. Lee C, Zhang H, Baker AE, Occhipinti P, Borsuk ME, Gladfelter AS: **Protein aggregation behavior regulates cyclin transcript localization and cell-cycle control**. *Dev Cell* 2013, **25**(6):572-584.

237. Hubstenberger A, Cameron C, Noble SL, Keenan S, Evans TC: **Modifiers of solid RNP granules control normal RNP dynamics and mRNA activity in early development**. *J Cell Biol* 2015, **211**(3):703-716.

238. Zhang H, Elbaum-Garfinkle S, Langdon EM, Taylor N, Occhipinti P, Bridges AA, Brangwynne CP, Gladfelter AS: **RNA Controls PolyQ Protein Phase Transitions**. *Mol Cell* 2015, **60**(2):220-230.

239. Reijns MA, Alexander RD, Spiller MP, Beggs JD: **A role for Q/N-rich aggregation-prone regions in P-body localization**. *J Cell Sci* 2008, **121**(Pt 15):2463-2472.

240. Decker CJ, Teixeira D, Parker R: **Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in Saccharomyces cerevisiae**. *J Cell Biol* 2007, **179**(3):437-449.

241. Brangwynne CP, Tompa P, Pappu RV: **Polymer physics of intracellular phase transitions**. *Nat Phys* 2015, **11**(11):899-904.

242. Salichs E, Ledda A, Mularoni L, Alba MM, de la Luna S: **Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment**. *PLoS Genet* 2009, **5**(3):e1000397.

243. Kenan DJ, Query CC, Keene JD: **RNA recognition: towards identifying determinants of specificity**. *Trends Biochem Sci* 1991, **16**(6):214-220.

244. Loerch S, Kielkopf CL: **Dividing and Conquering the Family of RNA Recognition Motifs: A Representative Case Based on hnRNP L**. *J Mol Biol* 2015, **427**(19):2997-3000.

245. Singh M, Choi CP, Feigon J: **xRRM: a new class of RRM found in the telomerase La family protein p65**. *RNA Biol* 2013, **10**(3):353-359.

246. Rebagliati M: **An RNA recognition motif in the bicoid protein**. *Cell* 1989, **58**(2):231-232.

247. Query CC, Bentley RC, Keene JD: **A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein**. *Cell* 1989, **57**(1):89-101.

248. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: **The I-TASSER Suite: protein structure and function prediction**. *Nat Methods* 2015, **12**(1):7-8.

249. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Honigschmid P, Schafferhans A, Roos M, Bernhofer M *et al*: **PredictProtein--an open resource for online prediction of protein structural and functional features**. *Nucleic Acids Res* 2014, **42**(Web Server issue):W337-343.

250. Landsman D: **RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain**. *Nucleic Acids Res* 1992, **20**(11):2861-2864.

251. Bandziulis RJ, Swanson MS, Dreyfuss G: **RNA-binding proteins as developmental regulators**. *Genes Dev* 1989, **3**(4):431-437.

252. Preitner N, Quan J, Nowakowski DW, Hancock ML, Shi J, Tcherkezian J, Young-Pearse TL, Flanagan JG: **APC is an RNA-binding protein, and its interactome provides a link to neural development and microtubule assembly**. *Cell* 2014, **158**(2):368-382.

253. Wright PE, Dyson HJ: **Linking folding and binding**. *Curr Opin Struct Biol* 2009, **19**(1):31-38.

254. Wysoczanski P, Schneider C, Xiang S, Munari F, Trowitzsch S, Wahl MC, Luhrmann R, Becker S, Zweckstetter M: **Cooperative structure of the heterotrimeric pre-mRNA retention and splicing complex**. *Nat Struct Mol Biol* 2014, **21**(10):911-918.

255. Kato Y, Nakamura A: **Roles of cytoplasmic RNP granules in intracellular RNA localization and translational control in the Drosophila oocyte**. *Dev Growth Differ* 2012, **54**(1):19-31.

256. Ferrandon D, Koch I, Westhof E, Nusslein-Volhard C: **RNA-RNA interaction is required for the formation of specific bicoid mRNA 3' UTR-STAUFEN ribonucleoprotein particles**. *EMBO J* 1997, **16**(7):1751-1758.

257. Wagner C, Palacios I, Jaeger L, St Johnston D, Ehresmann B, Ehresmann C, Brunel C: **Dimerization of the 3'UTR of bicoid mRNA involves a two-step mechanism**. *J Mol Biol* 2001, **313**(3):511-524.

258. Chekulaeva M, Hentze MW, Ephrussi A: **Bruno acts as a dual repressor of oskar translation, promoting mRNA oligomerization and formation of silencing particles**. *Cell* 2006, **124**(3):521-533.

259. Jain S, Parker R: **The discovery and analysis of P Bodies**. *Adv Exp Med Biol* 2013, **768**:23-43.

260. Parker R, Sheth U: **P bodies and the control of mRNA translation and degradation**. *Mol Cell* 2007, **25**(5):635-646.

261. Eystathioy T, Jakymiw A, Chan EK, Seraphin B, Cougot N, Fritzler MJ: **The GW182 protein colocalizes with mRNA degradation associated proteins hDcp1 and hLSm4 in cytoplasmic GW bodies**. *RNA* 2003, **9**(10):1171-1173.

262. Moser JJ, Fritzler MJ: **Relationship of other cytoplasmic ribonucleoprotein bodies (cRNPB) to GW/P bodies**. *Adv Exp Med Biol* 2013, **768**:213-242.

263. Voronina E, Seydoux G, Sassone-Corsi P, Nagamori I: **RNA granules in germ cells**. *Cold Spring Harb Perspect Biol* 2011, **3**(12).

264. Kamenska A, Simpson C, Vindry C, Broomhead H, Benard M, Ernoult-Lange M, Lee BP, Harries LW, Weil D, Standart N: **The DDX6-4E-T interaction mediates translational repression and P-body assembly**. *Nucleic Acids Res* 2016, **44**(13):6318-6334.

265. Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoege C, Gharakhani J, Julicher F, Hyman AA: **Germline P granules are liquid droplets that localize by controlled dissolution/condensation**. *Science* 2009, **324**(5935):1729-1732.

266. Brangwynne CP, Mitchison TJ, Hyman AA: **Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes**. *Proc Natl Acad Sci U S A* 2011, **108**(11):4334-4339.

267. Feric M, Vaidya N, Harmon TS, Mitrea DM, Zhu L, Richardson TM, Kriwacki RW, Pappu RV, Brangwynne CP: **Coexisting Liquid Phases Underlie Nucleolar Subcompartments**. *Cell* 2016, **165**(7):1686-1697.

268. Strzelecka M, Trowitzsch S, Weber G, Luhrmann R, Oates AC, Neugebauer KM: **Coilin-dependent snRNP assembly is essential for zebrafish embryogenesis**. *Nat Struct Mol Biol* 2010, **17**(4):403-409.

269. Spector DL, Lamond AI: **Nuclear speckles**. *Cold Spring Harb Perspect Biol* 2011, **3**(2).

270. Anantharaman A, Jadaliha M, Tripathi V, Nakagawa S, Hirose T, Jantsch MF, Prasanth SG, Prasanth KV: **Paraspeckles modulate the intranuclear distribution of paraspeckle-associated Ctn RNA**. *Sci Rep* 2016, **6**:34043.

271. Hirose T, Virnicchi G, Tanigawa A, Naganuma T, Li R, Kimura H, Yokoi T, Nakagawa S, Benard M, Fox AH *et al*: **NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies**. *Mol Biol Cell* 2014, **25**(1):169-183.

272. Wippich F, Bodenmiller B, Trajkovska MG, Wanka S, Aebersold R, Pelkmans L: **Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling**. *Cell* 2013, **152**(4):791-805.

273. Bruno I, Wilkinson MF: **P-bodies react to stress and nonsense**. *Cell* 2006, **125**(6):1036-1038.

274. Weil TT, Parton RM, Herpers B, Soetaert J, Veenendaal T, Xanthakis D, Dobbie IM, Halstead JM, Hayashi R, Rabouille C *et al*: **Drosophila patterning is established by differential association of mRNAs with P bodies**. *Nat Cell Biol* 2012, **14**(12):1305-1313.

275. Yang Y, Wen L, Zhu H: **Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein**. *Cell & Bioscience* 2015, **5**(1):1-10.

276. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynov S, Mahamid J, Saha S, Franzmann TM *et al*: **A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation**. *Cell* 2015, **162**(5):1066-1077.

277. Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, Mittag T, Taylor JP: **Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization**. *Cell* 2015, **163**(1):123-133.

278. Banani SF, Rice AM, Peeples WB, Lin Y, Jain S, Parker R, Rosen MK: **Compositional Control of Phase-Separated Cellular Bodies**. *Cell* 2016, **166**(3):651-663.

279. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J *et al*: **Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels**. *Cell* 2012, **149**(4):753-767.

280. Smith J, Calidas D, Schmidt H, Lu T, Rasoloson D, Seydoux G: **Spatial patterning of P granules by RNA-induced phase separation of the intrinsically-disordered protein MEG-3**. *Elife* 2016, **5**.

281. Pagano JM, Farley BM, McCoig LM, Ryder SP: **Molecular basis of RNA recognition by the embryonic polarity determinant MEX-5**. *J Biol Chem* 2007, **282**(12):8883-8894.

282. Buchan JR: **mRNP granules. Assembly, function, and connections with disease**. *RNA Biol* 2014, **11**(8):1019-1030.

283. Mir M, Reimer A, Haines JE, Li XY, Stadler M, Garcia H, Eisen MB, Darzacq X: **Dense Bicoid hubs accentuate binding along the morphogen gradient**. *Genes Dev* 2017, **31**(17):1784-1794.

284. Hannon CE, Blythe SA, Wieschaus EF: **Concentration dependent chromatin states induced by the bicoid morphogen gradient**. *Elife* 2017, **6**.

285. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA: **A Phase Separation Model for Transcriptional Control**. *Cell* 2017, **169**(1):13-23.

286. Weil TT: **mRNA localization in the Drosophila germline**. *RNA Biol* 2014, **11**(8):1010-1018.

287. Lin MD, Jiao X, Grima D, Newbury SF, Kiledjian M, Chou TB: **Drosophila processing bodies in oogenesis**. *Dev Biol* 2008, **322**(2):276-288.

288. Patel PH, Barbee SA, Blankenship JT: **GW-Bodies and P-Bodies Constitute Two Separate Pools of Sequestered Non-Translating RNAs**. *PLoS One* 2016, **11**(3):e0150291.

289. Eulalio A, Huntzinger E, Izaurralde E: **GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay**. *Nat Struct Mol Biol* 2008, **15**(4):346-353.

290. Lian SL, Li S, Abadal GX, Pauley BA, Fritzler MJ, Chan EK: **The C-terminal half of human Ago2 binds to multiple GW-rich regions of GW182 and requires GW182 to mediate silencing**. *RNA* 2009, **15**(5):804-813.

291. Palmer WH, Obbard DJ: **Variation and Evolution in the Glutamine-Rich Repeat Region of Drosophila Argonaute-2**. *G3 (Bethesda)* 2016, **6**(8):2563-2572.

292. Schutz S, Noldeke ER, Sprangers R: **A synergistic network of interactions promotes the formation of in vitro processing bodies and protects mRNA against decapping**. *Nucleic Acids Res* 2017.

293. Klingauf M, Stanek D, Neugebauer KM: **Enhancement of U4/U6 small nuclear ribonucleoprotein particle association in Cajal bodies predicted by mathematical modeling**. *Mol Biol Cell* 2006, **17**(12):4972-4981.

294. Blythe A, Gunasekara S, Walshe J, Mackay JP, Hartzog GA, Vrielink A: **Ubiquitin fusion constructs allow the expression and purification of multi-KOW domain complexes of the Saccharomyces cerevisiae transcription elongation factor Spt4/5**. *Protein Expr Purif* 2014, **100**:54-60.

295. Leontis NB, Westhof E: **Geometric nomenclature and classification of RNA base pairs**. *RNA* 2001, **7**(4):499-512.

296. Xin Y, Olson WK: **BPS: a database of RNA base-pair structures**. *Nucleic Acids Res* 2009, **37**(Database issue):D83-88.

297. Lescoute A, Leontis NB, Massire C, Westhof E: **Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments**. *Nucleic Acids Res* 2005, **33**(8):2395-2409.

298. Parisien M, Major F: **The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data**. *Nature* 2008, **452**(7183):51-55.

299. Li X, Kazan H, Lipshitz HD, Morris QD: **Finding the target sites of RNA-binding proteins**. *Wiley Interdiscip Rev RNA* 2014, **5**(1):111-130.

300. Ehara H, Yokoyama T, Shigematsu H, Yokoyama S, Shirouzu M, Sekine SI: **Structure of the complete elongation complex of RNA polymerase II with basal factors**. *Science* 2017.

301. Pei Y, Shuman S: **Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5**. *J Biol Chem* 2002, **277**(22):19639-19648.

302. Wen Y, Shatkin AJ: **Transcription elongation factor hSPT5 stimulates mRNA capping**. *Genes Dev* 1999, **13**(14):1774-1779.

303. Zhao C, Dave V, Yang F, Scarborough T, Ma J: **Target selectivity of bicoid is dependent on nonconsensus site recognition and protein-protein interaction**. *Mol Cell Biol* 2000, **20**(21):8112-8123.

304. Xiong D, Wang Y, Deng C, Hu R, Tian C: **Phylogenic analysis revealed an expanded C(2)H(2)-homeobox subfamily and expression profiles of C(2)H(2) zinc finger gene family in Verticillium dahliae**. *Gene* 2015, **562**(2):169-179.

305. Hench J, Henriksson J, Abou-Zied AM, Luppert M, Dethlefsen J, Mukherjee K, Tong YG, Tang L, Gangishetti U, Baillie DL *et al*: **The Homeobox Genes of Caenorhabditis elegans and Insights into Their Spatio-Temporal Expression Dynamics during Embryogenesis**. *PLoS One* 2015, **10**(5):e0126947.

306. Vorobyov E, Horst J: **Getting the proto-Pax by the tail**. *J Mol Evol* 2006, **63**(2):153-164.

307. Wang Z, Yang X, Guo S, Yang Y, Su XC, Shen Y, Long J: **Crystal structure of the ubiquitin-like domain-CUT repeat-like tandem of special AT-rich sequence binding protein 1 (SATB1) reveals a coordinating DNA-binding mechanism**. *J Biol Chem* 2014, **289**(40):27376-27385.

308. Schaefer LK, Wang S, Schaefer TS: **Functional interaction of Jun and homeodomain proteins**. *J Biol Chem* 2001, **276**(46):43074-43082.

309. Varadi M, Zsolyomi F, Guharoy M, Tompa P: **Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins**. *PLoS One* 2015, **10**(10):e0139731.

310. Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y: **Design and bioinformatics analysis of genome-wide CLIP experiments**. *Nucleic Acids Res* 2015, **43**(11):5263-5274.

311. Jankowsky E, Harris ME: **Mapping specificity landscapes of RNA-protein interactions by high throughput sequencing**. *Methods* 2017, **118-119**:111-118.

312. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB: **RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins**. *Mol Cell* 2014, **54**(5):887-900.

313. Chaney BA, Clark-Baldwin K, Dave V, Ma J, Rance M: **Solution structure of the K50 class homeodomain PITX2 bound to DNA and implications for mutations that cause Rieger syndrome**. *Biochemistry* 2005, **44**(20):7497-7511.

314. Billeter M, Guntert P, Luginbuhl P, Wuthrich K: **Hydration and DNA recognition by homeodomains**. *Cell* 1996, **85**(7):1057-1065.

315. Tsao DH, Gruschus JM, Wang LH, Nirenberg M, Ferretti JA: **The three-dimensional solution structure of the NK-2 homeodomain from Drosophila**. *J Mol Biol* 1995, **251**(2):297-307.

316. Cox M, van Tilborg PJ, de Laat W, Boelens R, van Leeuwen HC, van der Vliet PC, Kaptein R: **Solution structure of the Oct-1 POU homeodomain determined by NMR and restrained molecular dynamics**. *J Biomol NMR* 1995, **6**(1):23-32.

317. Macfarlane LA, Murphy PR: **MicroRNA: Biogenesis, Function and Role in Cancer**. *Curr Genomics* 2010, **11**(7):537-561.

318. Sen GL, Blau HM: **Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies**. *Nat Cell Biol* 2005, **7**(6):633-636.

319. Miles WO, Tschop K, Herr A, Ji JY, Dyson NJ: **Pumilio facilitates miRNA regulation of the E2F3 oncogene**. *Genes Dev* 2012, **26**(4):356-368.

320. Ciafre SA, Galardi S: **microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer**. *RNA Biol* 2013, **10**(6):935-942.

321. Kim M, Ahn SH, Krogan NJ, Greenblatt JF, Buratowski S: **Transitions in RNA polymerase II elongation complexes at the 3' ends of genes**. *EMBO J* 2004, **23**(2):354-364.

322. Blanchard SC, Gonzalez RL, Kim HD, Chu S, Puglisi JD: **tRNA selection and kinetic proofreading in translation**. *Nat Struct Mol Biol* 2004, **11**(10):1008-1014.

323. Stone MD, Mihalusova M, O'Connor C M, Prathapam R, Collins K, Zhuang X: **Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein**. *Nature* 2007, **446**(7134):458-461.

324. Lamichhane R, Daubner GM, Thomas-Crusells J, Auweter SD, Manatschal C, Austin KS, Valniuk O, Allain FH, Rueda D: **RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression**. *Proc Natl Acad Sci U S A* 2010, **107**(9):4105-4110.

325. Davidovich C, Wang X, Cifuentes-Rojas C, Goodrich KJ, Gooding AR, Lee JT, Cech TR: **Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA**. *Mol Cell* 2015, **57**(3):552-558.

326. Jain N, Lin HC, Morgan CE, Harris ME, Tolbert BS: **Rules of RNA specificity of hnRNP A1 revealed by global and quantitative analysis of its affinity distribution**. *Proc Natl Acad Sci U S A* 2017.

327. Wang X, Schwartz JC, Cech TR: **Nucleic acid-binding specificity of human FUS protein**. *Nucleic Acids Res* 2015, **43**(15):7535-7543.

328. Tauchert MJ, Fourmann JB, Luhrmann R, Ficner R: **Structural insights into the mechanism of the DEAH-box RNA helicase Prp43**. *Elife* 2017, **6**.

329. Rensburg GV, Mackedenski S, Lee CH: **Characterizing the Coding Region Determinant-Binding Protein (CRD-BP)-Microphthalmia-associated Transcription Factor (MITF) mRNA interaction**. *PLoS One* 2017, **12**(2):e0171196.

330. Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J: **Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions**. *Genome Biol* 2012, **13**(8):R67.

331. Wang Z, Tollervey J, Briese M, Turner D, Ule J: **CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo**. *Methods* 2009, **48**(3):287-293.

332. Gasteiger E, Hoogland, C., Gattiker, A., Duvaud, S.e., Wilkins, M.R., Appel, R.D. and , Bairoch A: **The Proteomics Protocols Handbook**; 2005.

333. Lee W, Tonelli M, Markley JL: **NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy**. *Bioinformatics* 2015, **31**(8):1325-1327.

334. Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD: **1H, 13C and 15N chemical shift referencing in biomolecular NMR**. *J Biomol NMR* 1995, **6**(2):135-140.

Appendices

Appendix A.1 RNA oligonucleotide sequences

Table A.1 RNA oligonucleotide sequences

RNA	Sequence
BRE(257-319)	GCCGUUGCACCUGGAAUAUUGCACGUUGUUAAUUUUUGUGAUUGUAUAUUCCU GGUUUCGACACGC
BRE38nt	GAAUAUUGCACGUUGUUAAUUUUGUGAUUGUAUAUUC
BRE39nt	UAUUCCUGGUUUCGACACGCGGCCGUUGCACCUGGAAUA
ShapeControl	UGGUUUAUGUGGAACAAUUAAAACCACUAACAAAACCA
Unst	UCGAAGCCCUCUCUCAGUUUGUCAUAUACCCU
mir-308	GAGUGUCAUAUUAGGACACUAA
mir-Fold	GGACUGACCACAAAUCAGUCAA
AA_{rich}	UGGCUCGCAAUAACAAAAACAAAC
4AA	UCAAUAACAAAAACA
T2AA	AACCAA
1AA	UCAAUC
$1HAA_{rich}$	UGGCUCGCAAUAA
$2HAA_{rich}$	AACAAAAACAAAC
$TrimAA_{rich}$	CUCGCAAUAACAAAAACA
AA5	CCCCUCGCAAUAACAAAAACAAAC
AA4	UGGCUCGCCCUAACAAAAACAAAC
AA3	UGGCUCGCCCUCCCAAAAACAAAC
AA2	UGGCUCGCCCUCCCCCAAACAAAC
AA1	UGGCUCGCCCUCCCCCACCCAAAC

Appendix A.2 DNA oligonucleotides sequences

Table A.2 DNA oligonucleotide sequences

Use	Oligonucleotide	Sequence
Cloning	BHDFwd	GCGGATCCCCACGTCGCACCCGCAC
	BHDRev	GCGAATTCCTATCATTAGGACTGGTCCTTGTGCTGATC
	HDERFwd	GCGGATCCCTGCCCGACTCTCTGGTGATG
	HDERRev	GCGAATTCTCATTAGGACTGGTCCTTGTGCTGATC
	BRRM Fwd	GCGGATCCGCCGTTGGCGAGACG
	BRRMRev	GCGAATTCTCATTACTAATTGAAGCAGTAGGCAAAC
In-vitro transcription	Cad3UTRFwd	GCGGATCCTAATACGACTCACTATAGGGAGACACGACCATTCCTGTTATGCGG
	Cad3UTRRev	GCGAATTCCCGCTGAGCAATAACTAGCGAGTTGCTTTATCTATGGTGTTCATATTTTA
	BRE(257-319)Fwd	TAATACGACTCACTATAGGGAGAATGGAGGACTTGGCGGCCGTTG
	BRE(257-319)Rev	GCGCGTGTCGAAACCAGGAATATACAAT
	BRE38ntFwd	TAATACGACTCACTATAGGGAGAGAATATTGCACGTTGTTAATTTTTGTGATTGTATATTC
	BRE38ntRev	GAATATACAATCACAAAAATTAACAACGTGCAATATTCTCTCCCTATAGTGAGTCGTATTA
	BRE39ntFwd	TAATACGACTCACTATAGGGAGATATTCCTGGTTTCGACACGCGGCCGTTGCACCTGGAATA
	BRE39ntRev	TATTCCAGGTGCAACGGCCGCGTGTCGAAACCAGGAATATCTCCCTATAGTGAGTCGTATTA
	ShapeControlFwd	TAATACGACTCACTATAGGGAGATGGTTTATGTGGAACAATTAAAACCACTAACAAAACCA
	ShapeControlRev	TGGTTTTGTTAGTGGTTTTAATTGTTCCACATAAACCATCTCCCTATAGTGAGTCGTATTA
	UnstFwd	TAATACGACTCACTATAGGGAGATCGAAGCCCTCTCTCAGTTTGTCATATACCCT
	UnstRev	AGGGTATATGACAAACTGAGAGAGGGCTTCGATCTCCCTATAGTGAGTCGTATTA
	Cad525Fwd	GAAATTAATACGACTCACTATAGGGCTTGGACTTGGCTTAACCCTTA
	Cad525Rev	GCTCGAAGAGTGCGTTACAT
	Cad411Fwd	GAAATTAATACGACTCACTATAGGGCGGCGACAGTAACAACTACA
	Cad411Rev	ACTTACTACTGCTTACGAGCTATTC
	CadEndFwd	GAAATTAATACGACTCACTATAGGGATGTAACGCACTCTTCGAGC
	CadEndRev	GAGTTGCTTTATCTATGGTGTTCATA
	AA$_{\text{rich}}$Fwd	GAAATTAATACGACTCACTATAGGGTGGCTCGCAATAACAAAAACAAAC
	AA$_{\text{rich}}$Rev	GTTTGTTTTTGTTATTGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	AA5Fwd	GAAATTAATACGACTCACTATAGGGCCCCTCGCAATAACAAAAACAAAC
	AA5Rev	GTTTGTTTTTGTTATTGCGAGGGGCCCTATAGTGAGTCGTATTAATTTC
	AA4Fwd	GAAATTAATACGACTCACTATAGGGTGGCTCGCCCTAACAAAAACAAAC
	AA4Rev	GTTTGTTTTTGTTAGGGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	AA3Fwd	GAAATTAATACGACTCACTATAGGGTGGCTCGCCCTCCCAAAAACAAAC
	AA3Rev	GTTTGTTTTTGGGAGGGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	AA2Fwd	GAAATTAATACGACTCACTATAGGGTGGCTCGCCCTCCCCCAAACAAAC
	AA2Rev	GTTTGTTTGGGGGAGGGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	AA1Fwd	GAAATTAATACGACTCACTATAGGGTGGCTCGCCCTCCCCCACCCAAAC
	AA1Rev	GTTTGGGTGGGGGAGGGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	GG$_{\text{rich}}$Fwd	GAAATTAATACGACTCACTATAGGGUGGCUCGCGGUGGCGGAGGCGGAC
	GG$_{\text{rich}}$Rev	GTCCGCCTCCGCCACCGCGAGCCACCCTATAGTGAGTCGTATTAATTTC
	TrimAA$_{\text{rich}}$Fwd	GAAATTAATACGACTCACTATAGGGCUCGCAAUAACAAAAACA
	TrimAA$_{\text{rich}}$Rev	TGTTTTTGTTATTGCGAGCCCTATAGTGAGTCGTATTAATTTC

Appendix A.3 Properties of protein constructs

Table A.3 Protein construct properties

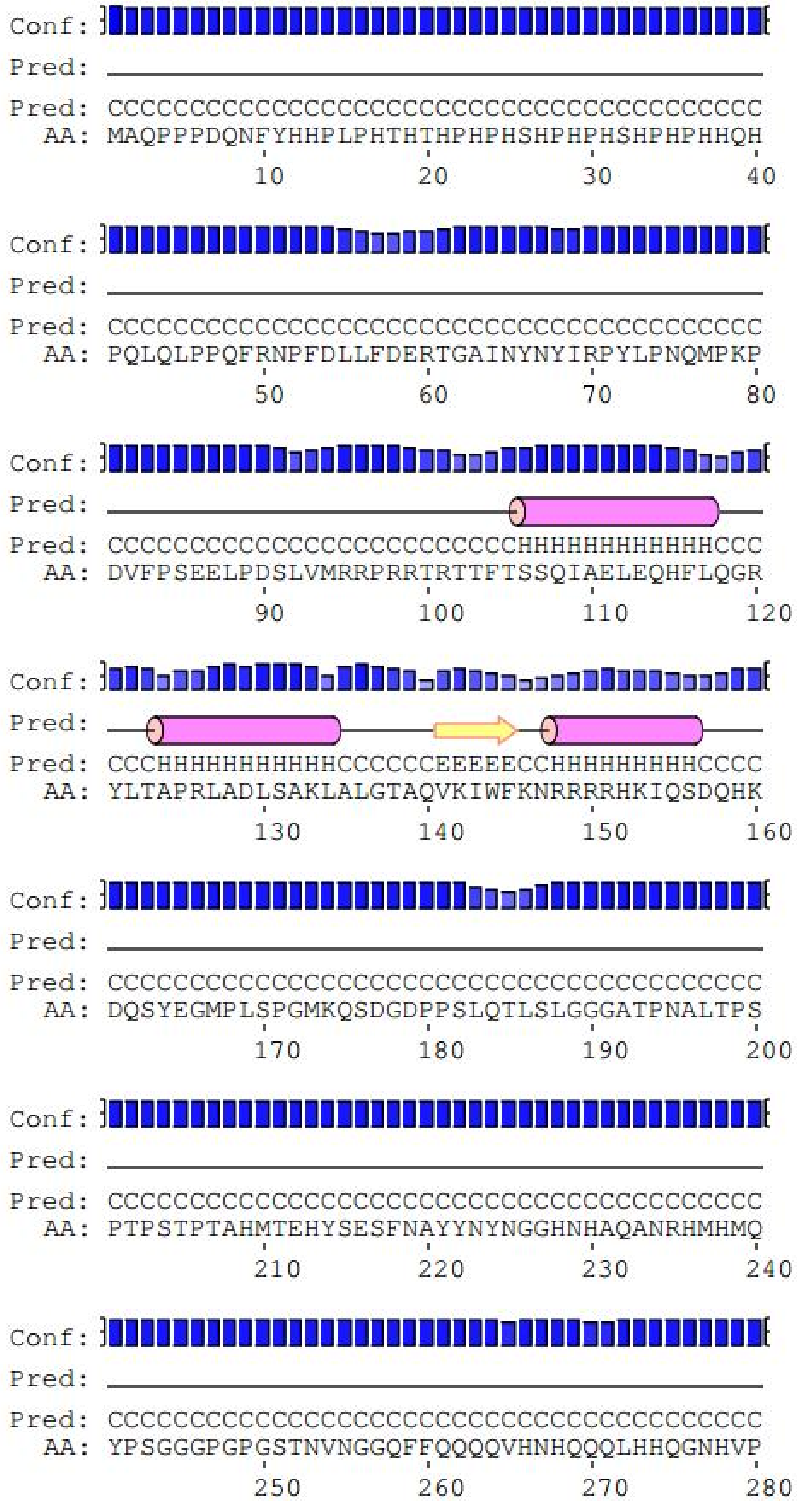
Construct	Protein residue numbers	molecular weight (Da)	pI
GST-BHD	GST + 97-163	36785.1	6.0
BHD	97-163	7888.9	11.5
HDER	88-163	9014.3	11.6
BRRM	378-494	13228.4	4.3
Spt4	1-101	11157.7	5.0
Spt5_{NGN}	284-375	10484.6	10.0

Amino acid sequence of GST (HRV-3C cleavage site indicated in red):

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGD
VKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSK
LPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAI
PQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD**LEVLFQGP**LGS

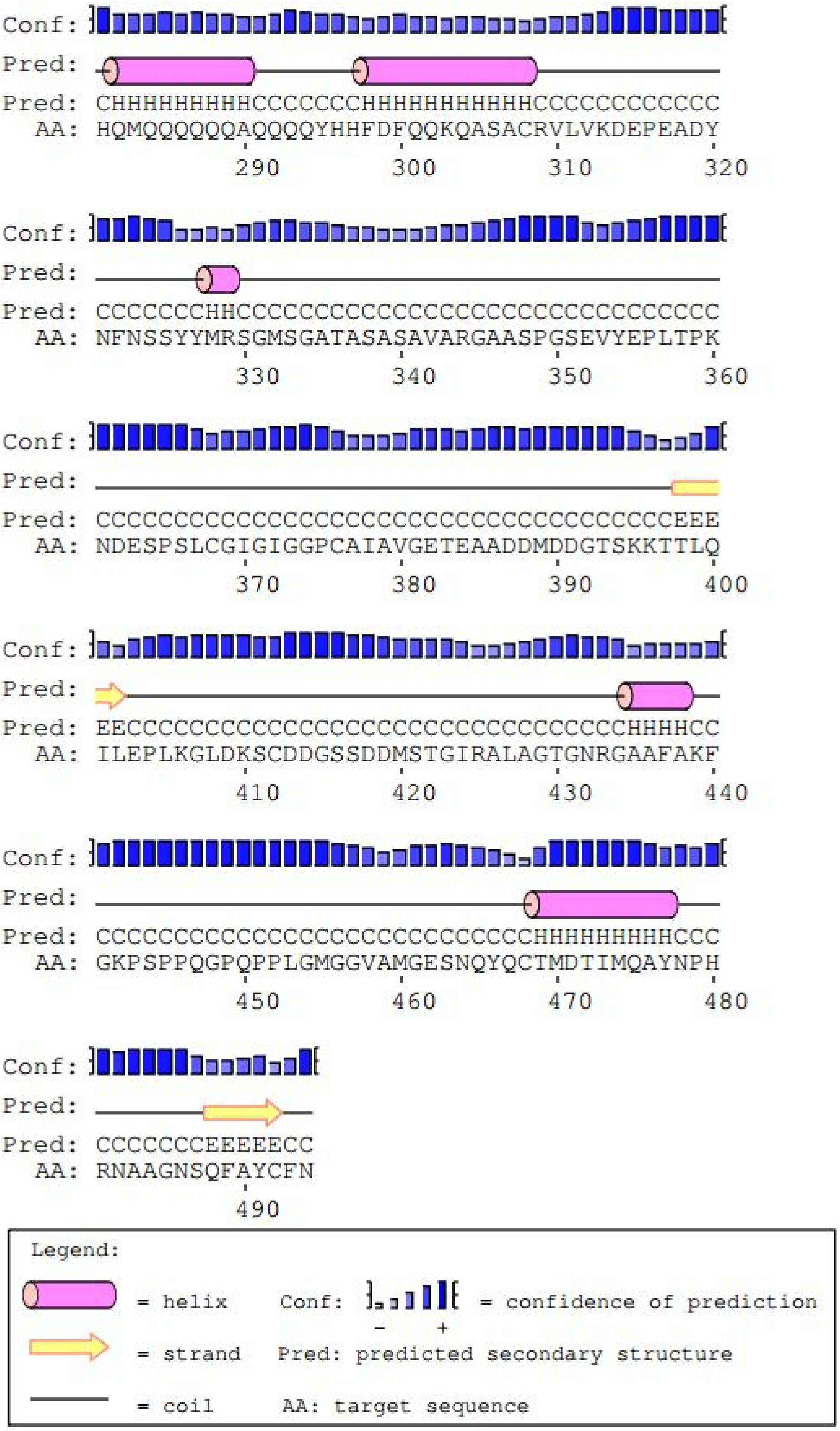
Appendix A.4 PSIPRED Secondary structure prediction of bicoid

Figure A.1 Bicoid PSIPRED secondary structure prediction



```
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: MAQPPPDQNFYHHPLPHTHTHPHPHSHPHPHSHPHPHHQH
                10        20        30        40

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: PQLQLPPQFRNPFDLLFDERTGAINYNYIRPYLPNQMPKP
                50        60        70        80

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCC
  AA: DVFPSEELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGR
                90       100       110       120

Conf:
Pred:
Pred: CCCHHHHHHHHHHHCCCCCCEEEEECCHHHHHHHHHCCCC
  AA: YLTAPRLADLSAKLALGTAQVKIWFKNRRRRHKIQSDQHK
               130       140       150       160

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: DQSYEGMPLSPGMKQSDGDPPSLQTLSLGGGATPNALTPS
               170       180       190       200

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: PTPSTPTAHMTEHYSESFNAYYNYNGGHNHAQANRHMHMQ
               210       220       230       240

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: YPSGGGPGPGSTNVNGGQFFQQQQVHNHQQQLHHQGNHVP
               250       260       270       280
```

Figure A.1 Continued



```
Conf:
Pred:
Pred: CHHHHHHHHHCCCCCCCHHHHHHHHHHHCCCCCCCCCCCC
  AA: HQMQQQQQQAQQQQYHHFDFQQKQASACRVLVKDEPEADY
                290       300       310       320

Conf:
Pred:
Pred: CCCCCCCHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: NFNSSYYMRSGMSGATASASAVARGAASPGSEVYEPLTPK
                330       340       350       360

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEE
  AA: NDESPSLCGIGIGGPCAIAVGETEAADDMDDGTSKKTTLQ
                370       380       390       400

Conf:
Pred:
Pred: EECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCC
  AA: ILEPLKGLDKSCDDGSSDDMSTGIRALAGTGNRGAAFAKF
                410       420       430       440

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHCCC
  AA: GKPSPPQGPQPPLGMGGVAMGESNQYQCTMDTIMQAYNPH
                450       460       470       480

Conf:
Pred:
Pred: CCCCCCCEEEEECC
  AA: RNAAGNSQFAYCFN
                490
```

Legend:

= helix Conf: = confidence of prediction
- +

= strand Pred: predicted secondary structure

= coil AA: target sequence