Structural characterisation of the *Saccharomyces cerevisiae* transcription elongation factor Spt4/5 through RNA interactions

AMANDA BLYTHE BSc (HONS)

The thesis is presented for the degree of Doctor of Philosophy of Biochemistry of the University of Western Australia

June 2014
Abstract

The transcription elongation factor Spt4/5 (Spt4/5) tightly associates with RNA polymerase II (RNAPII) regulating elongation and co-transcriptional pre-mRNA processing events; however, the mechanisms by which Spt4/5 acts are poorly understood. Recent studies of the human and *Drosophila* Spt4/5 complexes indicate that they can bind nucleic acids *in vitro* and that this function may be key in understanding the ability of Spt4/5 to exert its control over transcription elongation.

Spt5 is an essential protein and the only universally conserved RNAP-associated transcription elongation factor. A unique feature of eukaryotic Spt5 is the presence of multiple Kyripides, Ouzounis and Woese (KOW) domains. The KOW domains are thought to be responsible for the RNA binding function of Spt4/5 although there is little evidence in the literature to support this theory. Therefore, the main focus of this study was to characterise the RNA binding specificity of *Saccharomyces cerevisiae* Spt4/5.

When expressed in *Escherichia coli*, the Spt4/5 complex is innately insoluble, possibly due to the large unstructured regions connecting the KOW domains to each other and to the Spt5 core. Therefore, the first section of this thesis presents the expression and purification of milligram quantities of three different multi-KOW domain complexes of yeast Spt4/5 through the use of ubiquitin fusion constructs. The multi-KOW domain Spt4/5 complexes were characterised by various biophysical techniques showing that the proteins are folded with an Spt4:Spt5 hetero-dimeric stoichiometry of 1:1.
The second section of this thesis reports novel research characterising the RNA binding specificity of Spt4/5. A new function by way of RNA binding is attributed to a domain of Spt4/5 that associates directly with RNAPII. This result is of unusual significance as this domain has not previously been characterised as a nucleic acid binding domain, therefore providing insight into the molecular interaction of Spt4/5 with RNAPII and raising provocative questions regarding the mechanism underlying transcription elongation and RNA processing. As Spt4/5 is a ubiquitous transcription elongation factor, these results are intriguing as the RNA binding is shown to be sequence specific for multiple AA repeat elements. Therefore, it is possible that this AA repeat motif is a new regulatory element for the regulation of transcription elongation and RNA processing by Spt4/5.

Finally, the determination of the Spt4/5:RNA binding interface was attempted using X-ray crystallography and bioinformatics approaches. Although the structural solution of the protein:RNA complex was not successful, analysis of the known Spt4/5 structure revealed several potential RNA binding sites.

In summary, the characterisation of the RNA binding specificity of Spt4/5 described in this thesis has made a significant contribution towards the understanding of the basic cellular process of transcription elongation. This work makes significant steps towards elucidating the mechanism behind transcriptional control by Spt4/5 through RNA recognition. Future studies can gain from the data presented here in order to investigate the role of Spt4/5 and its molecular interaction with RNAPII.
Declaration

I hereby declare that this thesis comprises only my original work towards this PhD except where indicated in the Statement of Contributions,

Amanda Blythe
Acknowledgements

I would like to express my sincere thanks my supervisor W/Professor Alice Vrielink for providing me with the opportunity to complete my PhD in her laboratory and for her ongoing support, advice and encouragement throughout the entirety of this project. I would also like to thank Professor Joel Mackay for allowing me the full facilities and expertise of his lab at the University of Sydney in order to carry out my RNA binding assays.

I would like to thank all the members, past and present, of the “Happy Corner” for their friendship, help and extensive discussions over the time. My PhD experience would not have been the same without you all.

Finally I would like to thank my husband and family for their love and support throughout my PhD, and for patiently waiting for me to finish. It has been much appreciated.
# Table of Contents

## 1. INTRODUCTION

1.1. TRANSCRIPTION ELONGATION ................................................................. 1

1.2. THE TRANSCRIPTION FACTOR Spt4/5 ......................................................... 3

1.2.1. The domain structure of Spt4/5 .............................................................. 4

1.3. THE FUNCTION OF Spt4/5 ........................................................................... 7

1.3.1. Coupling elongation to mRNA processing ............................................... 9

1.3.2. Spt4/5 interacts directly with RNAPII .................................................. 9

1.3.1. Spt4/5 interacts with nucleic acids ......................................................... 10

1.4. RESEARCH AIMS ......................................................................................... 12

## 2. UBIQUITIN FUSION CONSTRUCTS ALLOW THE EXPRESSION AND PURIFICATION OF MULTI-KOW DOMAIN COMPLEXES OF THE SPT4/5

2.1. ABSTRACT .................................................................................................. 15

2.2. INTRODUCTION .......................................................................................... 16

2.3. MATERIALS AND METHODS ...................................................................... 19

2.3.1. Cloning .................................................................................................. 19

2.3.2. Protein expression .................................................................................. 19

2.3.3. Cell lysis ................................................................................................. 20

2.3.4. Purification ............................................................................................. 21

2.3.5. SDS-PAGE and western blot analysis .................................................... 21

2.3.6. Limited proteolysis ............................................................................... 22

2.3.7. SEC-MALLS ......................................................................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.8. Circular dichroism spectropolarimetry (CD)</td>
<td>23</td>
</tr>
<tr>
<td>2.4. RESULTS AND DISCUSSION</td>
<td>23</td>
</tr>
<tr>
<td>2.5. ACKNOWLEDGMENTS</td>
<td>33</td>
</tr>
<tr>
<td>2.6. AUTHOR CONTRIBUTIONS</td>
<td>33</td>
</tr>
<tr>
<td>3. YEAST TRANSCRIPTION ELONGATION FACTOR SPT4/5 IS A</td>
<td></td>
</tr>
<tr>
<td>SEQUENCE SPECIFIC RNA BINDING PROTEIN</td>
<td>35</td>
</tr>
<tr>
<td>3.1. ABSTRACT</td>
<td>35</td>
</tr>
<tr>
<td>3.2. INTRODUCTION</td>
<td>36</td>
</tr>
<tr>
<td>3.3. METHODS</td>
<td>38</td>
</tr>
<tr>
<td>3.3.1. Cloning, expression and purification</td>
<td>38</td>
</tr>
<tr>
<td>3.3.2. Pentaprobe and oligonucleotide preparation</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3. Gel shift assay</td>
<td>39</td>
</tr>
<tr>
<td>3.3.4. SELEX</td>
<td>40</td>
</tr>
<tr>
<td>3.3.5. High-throughput sequencing and analyses of selected sequences</td>
<td>41</td>
</tr>
<tr>
<td>3.3.6. Genome analysis</td>
<td>41</td>
</tr>
<tr>
<td>3.3.7. Circular dichroism spectroscopy and analysis</td>
<td>42</td>
</tr>
<tr>
<td>3.3.8. Microscale thermophoresis</td>
<td>42</td>
</tr>
<tr>
<td>3.3.9. Structure analysis</td>
<td>43</td>
</tr>
<tr>
<td>3.4. RESULTS</td>
<td>43</td>
</tr>
<tr>
<td>3.4.1. Spt4/5 binds to ssRNA in vitro</td>
<td>43</td>
</tr>
<tr>
<td>3.4.2. Spt4/5 binds RNA with sequence specificity</td>
<td>45</td>
</tr>
<tr>
<td>3.4.3. AAN motifs are enriched in transcripts</td>
<td>46</td>
</tr>
<tr>
<td>3.4.4. The Spt4/5 NGN domain is sufficient for RNA binding.</td>
<td>48</td>
</tr>
<tr>
<td>3.4.5. Spt4/5 hetero-dimer formation and RNA binding are associated with conformational change.</td>
<td>50</td>
</tr>
<tr>
<td>3.5. DISCUSSION</td>
<td>52</td>
</tr>
</tbody>
</table>
3.6. AUTHOR CONTRIBUTIONS ................................................................. 54
3.7. ACKNOWLEDGMENTS ................................................................. 55

4. THE IDENTIFICATION OF A SPECIFIC RNA BINDING MOTIF FOR SPT4/5 ............................................................................................................................ 57
   4.1. INTRODUCTION ........................................................................ 57
   4.2. METHODS ............................................................................. 59
       4.2.1. Protein preparation .......................................................... 59
       4.2.2. Selection .......................................................................... 60
       4.2.3. Monitoring enrichment ..................................................... 60
       4.2.4. Sanger sequencing ............................................................ 60
       4.2.5. Deep Sequencing ............................................................... 60
   4.3. RESULTS AND DISCUSSION ..................................................... 61
       4.3.1. Sequence preparation ....................................................... 65
       4.3.2. MEME analysis ................................................................. 67
       4.3.3. Total dataset analysis ....................................................... 72
       4.3.4. Analysis of the AANAA motif ......................................... 73
       4.3.5. Analysis of the CUCG motif ............................................ 76
       4.3.6. Validation of SELEX results ........................................... 78
   4.4. CONCLUSION ........................................................................ 80

5. TOWARDS THE STRUCTURE OF THE SPT4/5:RNA COMPLEX ........ 83
   5.1. INTRODUCTION ........................................................................ 83
   5.2. METHODS ............................................................................. 87
       5.2.1. Protein preparation .......................................................... 87
       5.2.2. Initial crystallisation trials ................................................. 88
       5.2.3. Crystallisation of Spt4/5284-375 with RNA ....................... 88
# Table of Contents

5.2.4. Seeding ................................................................. 89  
5.2.5. Cryoprotection and crystal storage ............................... 89  
5.2.6. X-ray diffraction and data collection ............................. 90  
5.2.7. Data processing .......................................................... 90  
5.2.8. Molecular replacement ................................................... 90  
5.2.9. Model building ........................................................... 90  
5.2.10. Structure analysis ...................................................... 91  

5.3. Towards the X-ray structure of the Spt4/5:RNA complex ........... 91  
5.3.1. Crystallisation of Spt4/5 284-375 .................................. 91  
5.3.2. Data collection and X-ray diffraction analysis .................. 96  
5.3.3. Crystallisation of Spt4/5 284-375:RNA ............................ 97  
5.3.4. Structure analysis of Spt4/5 284-375:RNA ....................... 100  

5.4. Elucidating a mechanism of RNA binding by Spt4/5 ................. 104  
5.4.1. The NGN domain of Spt4/5 contains an RRM .................. 105  
5.4.2. Other possible modes of RNA binding by Spt4/5 ............... 108  

5.5. Conclusions .................................................................. 111  

6. Final Discussion ................................................................ 113  

7. References ........................................................................ 117  

8. Appendix 1. Oligonucleotide Sequences ............................... 143  

9. Appendix 2. Plasmids generated in this study ......................... 147  

10. Appendix 3. Supplemental information for Chapter 3 ............ 148  

11. Appendix 4. Sequence Alignments ..................................... 152
LIST OF TABLES

List of Tables

Table 2.1. Purification summary for each Spt4/5ubq complex ........................................... 27
Table 4.1. Stringency parameters for each round of SELEX .................................................... 62
Table 5.1. Overview of the number of published protein:RNA structures for each class of RNA binding domain ........................................................................................................... 87
Table 5.2. Optimised crystallisation conditions for Spt4/5284-375 ......................................... 94
Table 5.3. Spt4/5284-375:RNA crystal data collection and processing statistics ............... 98
Table 5.4. Refinement and final model statistics for the final Spt4/5284-375 structure ............ 100

Supplementary table 8.1. Primers used for generation of Spt4 and Spt4/5 constructs .......... 143
Supplementary table 8.2. SELEX primers and oligonucleotides ........................................... 144
Supplementary table 8.3. RNA probes used for REMSA and MST ....................................... 145
Supplementary table 9.1. Overview of cloned plasmids and their corresponding protein products .......................................................................................................................... 147
Supplementary table 10.1. Association of ≥2_AA motifs in UTRs with transcription frequency and UTR length ........................................................................................................... 150
List of Figures

Figure 1.1. Schematic diagram depicting the process of transcription by RNA polymerase (RNAP) ................................................................. 2
Figure 1.2. Universal evolutionary conservation of RNAP and Spt4/5 .................. 4
Figure 1.3. Structure of Spt4/5 ........................................................................ 5
Figure 1.4. All Spt5 homologues associate directly with RNAPII .................... 11
Figure 2.1. Structural organization of the Spt4 and Spt5 proteins ..................... 18
Figure 2.2. Expression and solubility of Spt4/5 constructs from the pET-Duet-1 vector ...................................................................................................................... 25
Figure 2.3. Expression and solubility of Spt4/5ubq284-419 in pHUE .................... 26
Figure 2.4. Visualisation of Spt4 in the purified Spt4/5ubq hetero-dimers .......... 27
Figure 2.5. Analysis of Spt4 doublet ................................................................ 29
Figure 2.6. SEC-MALLS analysis of Spt4/5ubq hetero-dimers ......................... 31
Figure 2.7. Physical properties of multi-KOW Spt4/5ubq hetero-dimers .......... 32
Figure 3.1. Spt4/5 binds to RNA containing AA repeats .................................. 44
Figure 3.2. Distribution of ≥2_AA and ≥2_TT motifs in yeast genes ................ 47
Figure 3.3. Spt4 together with the Spt5 NGN domain is sufficient for RNA binding... 49
Figure 3.4. Structural insight into the RNA binding mechanism of Spt4/5 .......... 50
Figure 4.1. Monitoring the enrichment of the SELEX libraries ......................... 63
Figure 4.2. Alignment of 14 sequences obtained from round 7 of GST-ZRANB2-F12 SELEX ................................................................................................. 65
Figure 4.3. Next generation sequencing of Spt4/5ubq SELEX ......................... 66
Figure 4.4. MEME analysis of the ubiquitin SELEX library after round 7 ......... 69
Figure 4.5. MEME analysis of the SELEX library for each Spt4/5ubq construct after round 7 ......................................................................................................... 71

Figure 4.6. Total library analysis of the top three sequence motifs for each multi-KOW Spt4/5ubq construct...................................................................................... 73

Figure 4.7. The enrichment of increasing numbers of AA repeats for the AA
(2_A A) motif ................................................................................................ 75

Figure 4.8. Specific enrichment of the length of the AA repeat motif .............. 76

Figure 4.9. Analysis of the CUCG motif .......................................................... 77

Figure 5.1 Structures of RNA binding motifs bound to their cognate RNA .......... 85

Figure 5.2. Purification and ubiquitin tag digestion of Spt4/5284-375 .................. 93

Figure 5.3. Spt4/5284-375 crystallisation events observed from the Hampton research Index screen ................................................................................................. 94

Figure 5.4. Optimisation of Spt4/5284-375 crystallisation .................................. 95

Figure 5.5. Seeding improves crystallisation of Spt4/5284-375 .......................... 96

Figure 5.6. Crystallisation of Spt4/5284-375:RNA co-crystals ......................... 97

Figure 5.7. A model of the X-ray structure of Spt4/5284-375 ............................. 101

Figure 5.8. Crystallisation results for Spt4/5284-375:RNA .................................. 104

Figure 5.9. Spt5 NGN contains an RRM ......................................................... 106

Figure 5.10. The electrostatic surface of Spt4/5 NGN......................................... 110

Supplementary figure 10.1. The domain organisation and constructs of the S. cerevisiae Spt4/5 hetero-dimer ......................................................................................... 148

Supplementary figure 10.2. SELEX motif analysis ............................................. 149

Supplementary figure 10.3. KOW domains do not bind RNA in vitro ............... 151

Supplementary figure 11.1. Multiple sequence alignment of Spt5 NGN .......... 152

Supplementary figure 11.2. Multiple sequence alignment of Spt4 ..................... 153
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-ME</td>
<td>β-mercaptoethanol</td>
</tr>
<tr>
<td>C-terminal</td>
<td>Carboxy terminal</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism</td>
</tr>
<tr>
<td>CTR</td>
<td>Carboxy terminal repeat</td>
</tr>
<tr>
<td>DRB</td>
<td>5,6-dichloro-1-β-D-ribofuranosylbenzimidazole</td>
</tr>
<tr>
<td>dsDNA</td>
<td>double stranded DNA</td>
</tr>
<tr>
<td>DSIF</td>
<td>DRB sensitivity inducing factor</td>
</tr>
<tr>
<td>dsRBM</td>
<td>double stranded RNA binding motif</td>
</tr>
<tr>
<td>E-value</td>
<td>expectation value</td>
</tr>
<tr>
<td>E. coli</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>EMSA</td>
<td>Electromobility shift assay</td>
</tr>
<tr>
<td>His-tag</td>
<td>hexahistidine tag</td>
</tr>
<tr>
<td>hnRNP</td>
<td>Heterogeneous ribonucleoprotein particle</td>
</tr>
<tr>
<td>HRP</td>
<td>Horseradish peroxidase</td>
</tr>
<tr>
<td>IMAC</td>
<td>Immobilised metal affinity chromatography</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>K_d</td>
<td>Dissociation constant</td>
</tr>
<tr>
<td>KH</td>
<td>hnRNP K homology domain</td>
</tr>
<tr>
<td>KOW</td>
<td>Kyrpides, Ouzonis and Woese domain</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani</td>
</tr>
<tr>
<td>LLG</td>
<td>Log likelihood gain</td>
</tr>
<tr>
<td>MALLS</td>
<td>Multi angle laser light scattering</td>
</tr>
<tr>
<td>MCS</td>
<td>Multiple cloning site</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>MEME</td>
<td>Multiple EM for motif elicitation</td>
</tr>
<tr>
<td>MST</td>
<td>Microscale thermophoresis</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>N-terminal</td>
<td>Amino terminal</td>
</tr>
<tr>
<td>NELF</td>
<td>Negative elongation factor b</td>
</tr>
<tr>
<td>NGN</td>
<td>NusG N-terminal</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>NusG</td>
<td>N-utilisation substance type G</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PTB</td>
<td>Polypyrimidine tract binding protein</td>
</tr>
<tr>
<td>pTEFb</td>
<td>Positive transcription elongation factor b</td>
</tr>
<tr>
<td>RBD</td>
<td>RNA binding domain</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding protein</td>
</tr>
<tr>
<td>REMSA</td>
<td>RNA electromobility shift assay</td>
</tr>
<tr>
<td>RNAP</td>
<td>RNA polymerase</td>
</tr>
<tr>
<td>RNAPII</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SEC</td>
<td>Size exclusion chromatography</td>
</tr>
<tr>
<td>SEC-MALLS</td>
<td>Size exclusion chromatography coupled to multi angle laser light scattering</td>
</tr>
<tr>
<td>SELEX</td>
<td>Systematic evolution of ligands by exponential enrichment</td>
</tr>
<tr>
<td>SPT</td>
<td>Suppressor of Ty</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single stranded DNA</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ssRNA</td>
<td>single stranded RNA</td>
</tr>
<tr>
<td>TEC</td>
<td>Transcription elongation complex</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>ZF</td>
<td>Zinc finger</td>
</tr>
</tbody>
</table>
CHAPTER 1

1. Introduction

1.1. Transcription elongation

Transcription is the production of an RNA transcript from a DNA template and is the first step in gene expression, making it an important point for cell regulation. In eukaryotes, RNA polymerase II (RNAPII) transcribes protein-encoding genes with the assistance of regulatory transcription factors. Transcription can be broken down into three stages: initiation, elongation and termination. Transcription factors forming a pre-initiation complex bind to the promoter region of the gene and recruit RNAPII, initiating transcription. Elongation commences when RNAPII escapes the promoter and moves along the gene synthesizing the RNA transcript. Transcription termination results from specific sequences in the DNA template or RNA transcript inducing 3’-end formation of the newly transcribed RNA, causing the cleavage and release of the transcript and the destabilisation and dissociation of RNAPII from the DNA template (Figure 1.1).

Transcription elongation is not a continuous process but is punctuated with sustained pauses, stalling, and arrest of RNAPII. For many genes, RNAPII begins transcription but stalls after synthesizing a short stretch of mRNA. This has been termed promoter-proximal pausing and is thought to be an important regulatory step in the rapid and precise control of gene expression (1). It is believed that the presence of a promoter-proximally stalled RNAPII marks a gene for future activation, by keeping transcription factors near the promoter, or by maintaining the chromatin conformation around the
promoter in a transcriptionally conducive form (2). This stalling is commonly seen in genes such as developmental and stimulus response genes where their expression is rapidly switched on and off (3-6).

Figure 1.1. Schematic diagram depicting the process of transcription by RNA polymerase (RNAP). Transcription is initiated when RNAP is recruited to the promoter region of the gene by the pre-initiation complex. RNAP escapes the promoter and the nascent RNA begins to be synthesised by the elongating RNAP. Transcription termination results in the release of the newly synthesised transcript and dissociation of RNAP from the gene.

After promoter-proximal pausing is overcome, elongation is still prone to mishaps such as transcript slippage, where the RNA transcript un-pairs from the DNA template and re-hybridises upstream to produce a longer transcript; backtracking, where RNAPII moves back along the DNA template, causing the 3’ end of the RNA to misalign with the active site of the enzyme; transient pausing, where the template sequence and/or regulatory factors temporarily block the movement of RNAPII; and arrest, where stalled
transcription due to backtracking cannot continue without the aid of transcription factors (7).

Current hypotheses suggest discontinuous transcription occurs to facilitate co-transcriptional processing events such as mRNA capping, where a methylated GMP cap is added to the 5’ end of the nascent mRNA transcript to stabilize the RNA through post-transcriptional processing, transport, and translation (2, 8, 9); aiding chromatin remodelling by removing the nucleosome barrier to allow efficient transcription (10-16); and splicing events, as the rate of transcription influences splicing and alternative splice site choices (17-20). The transient pausing and arrest of RNAPII during elongation requires the assistance of a myriad of proteins, or elongation factors, that help regulate the rate of elongation. Many studies have focused on transcription initiation and termination; however, knowledge is now emerging showing that the transition of transcription initiation into elongation is widely regulated at most genes, highlighting the importance of studying transcription elongation factors.

1.2. The transcription factor Spt4/5

The general transcription factors that assist and direct the activities of RNA polymerase display significant variation in structure and function across the three domains of life. In contrast, the catalytic cores of the multi-subunit polymerases found in all living organisms display deep conservation. There is only one transcription factor known to share this degree of conservation: the Suppressor of Ty 5/ N-utilisation substance type G (Spt5/NusG) family of proteins (Figure 1.2). The Suppressor of Ty (SPT) genes were originally identified through a genetic selection for factors that affect transcription in yeast (21, 22). The conservation of the Spt5/NusG family across the three kingdoms of
life suggests that this family is an ancient and essential regulator of transcription elongation. The Spt5/NusG proteins are essential for life and regulate transcription elongation in bacteria, archaea and eukaryotes (23); however, the mechanisms by which they achieve this regulation and their role in transcription elongation are poorly understood.

Figure 1.2. Universal evolutionary conservation of RNAP and Spt4/5
(Figure modified from (24)) A. The X-ray structures of multi-subunit RNAPII from eukarya (Saccharomyces cerevisiae PDB entry 1NT9 (25)); archaea (Sulfolobus solfataricus PDB entry 2PMZ (26)); and bacteria (Thermus aquaticus PDB entry1I6V (27)). B. The X-ray structures of eukaryotic Spt5 amino terminal NusG (NGN) domain bound to Spt4 (Saccharomyces cerevisiae PDB entry 2EXU (28)), the archaeal Spt5 NGN bound to Spt4 (Methanococcus jannaschii PDB entry 3LPE (29)) and the bacterial NGN domain (Escherichia coli PDB entry 2K06). Spt5 NGN is shown in red and Spt4 in wheat. The Spt5 NGN domain is homologous to NusG, whereas bacteria do not have an Spt4 homologue.

1.2.1. The domain structure of Spt4/5

Eukaryotic Spt5 is essential for life in all species. It is a large multi-domain protein composed of an N-terminal acidic region, a NusG N-terminal (NGN) domain, multiple Kyripides, Ouese and Woese (KOW) domains, and a C-terminal repeat (CTR) region.
The domain structure across eukaryotes is similar; however, the archaeal Spt5 and the bacterial homolog NusG consist only of the NGN domain and a single KOW domain (Figure 1.3A).

Figure 1.3. Structure of Spt4/5
A. Domain structure of Spt4/5 across the three kingdoms of life. Spt5 contains an N-terminal acidic region, the amino-terminal NusG (NGN) domain, five Kyprides, Ouese and Woese (KOW) domains and a C-terminal repeat (CTR) region. Spt4 contains an N-terminal zinc finger (Zn). B. Crystal structure of the Spt4/5 hetero-dimer from Pyrococcus furiosus (PDB entry 3P8B). The structure shows the β-sheet formed by the alignment of the β-strands from Spt4 and Spt5, and the unstructured linker connecting the KOW domain to the NGN domain. Spt4 is shown in blue, Spt5NGN in purple, the KOW domain in green and the flexible linker in black. The zinc atom of Spt4 is shown as a grey sphere coordinated by four cysteine residues coloured red. C. (Figure taken from (28)). Superimposition of five bacterial (Aquifex aeolicus) NusG structures and the yeast Spt4/5 NGN complex. Each structure depicts the KOW domain in different positions indicating a degree of mobility or flexibility in the linker region.

The NGN domain of Spt5 contains a conserved sequence that is the basis of interaction with the zinc-finger protein Spt4 (RpoE” in archaea), and though the domain has been conserved in all species, bacteria do not contain an Spt4 homologue (28, 30, 31). The
crystal structure of *S. cerevisiae* Spt4/5 characterised the hetero-dimer interface as an anti-parallel β-sheet formed by the alignment of the β-strands from the Spt4 and Spt5 monomers (Figure 1.3B) (28). Spt4 is not essential for life in yeast (32) but is essential in higher eukaryotes. It is a small protein, 102 amino acids long, containing an N-terminal zinc-finger domain co-ordinated by four cysteine residues (33). Spt4 appears to be an obligate binding partner of Spt5 since no free Spt4 exists in HeLa cells or yeast cell extracts (28, 34). Mutations of the yeast SPT4 gene show that it forms strong genetic interactions with Spt5 (32, 35), suggesting that Spt4 is critical for normal function of the Spt4/5 complex; however, the biological role of Spt4 in the hetero-dimer is largely unknown.

KOW domains are small domains, approximately 27 residues in length, and arranged into 5 alternating blocks of hydrophobic and hydrophilic residues with the hydrophobic blocks forming β-sheets (36). This motif was originally thought to be shared between members of the ribosomal protein family and the NusG protein family. The strict conservation of KOW domains in several ribosomal proteins resulted in their classification as an RNA binding domain (36), and the domain has now been found in several other RNA binding proteins (37, 38).

In all the proteins containing the KOW motif, it is interesting to note that only the eukaryotic Spt5 contains multiple KOW domains. Linker regions connect the KOW domains to each other and the NGN domain. This raises the question of whether these domains act as separate entities or in concert with each other. It bears mentioning that to date there are no structures of the Spt4/5 hetero-dimer containing multiple KOW domains. The linker regions connecting the KOW domains are of varying length, and have no predicted secondary structure (Figure 1.3). The crystal structures of archaeal
Spt4/5 and NusG show discrepancy in the position of the KOW domain relative to the NGN domain (Figure 1.3C) (28). This would suggest there is flexibility in the linker regions between the KOW and NGN domain, and also between KOW domains. Thus, the arrangement of the C-terminal portion of Spt5 relative to the NGN domain is unknown.

The carboxy terminal repeat (CTR) region of *S. cerevisiae* Spt5 is composed of 15 copies of a six amino acid repeat (SA/TWGCA/Q) and is critical for Spt5 function (32). Comparatively, the CTR region of human Spt5 has only seven copies of a different six amino acid repeat (K/RTPA/MYG/D/E), which has also been found in 20 copies in *C. elegans* Spt5 (39). The CTR region has been shown to be a target for phosphorylation by the positive transcription elongation factor b (pTEFb) (34, 35) and has been suggested to be a docking site for other transcriptional regulators (9, 40-43).

1.3. The function of Spt4/5

Spt4/5 was originally identified in a genetic selection for factors that affect transcription in yeast (21, 22). Mutations in the SPT4 and SPT5 genes affect transcription of some histone genes (44), leading to the theory that Spt4/5 regulated transcription through its effect on chromatin. Initial studies suggested that Spt4/5 negatively affects transcription by maintaining chromatin in a repressed state until remodelling occurs by the chromatin remodelling complex, Swi/Snf (32). In contrast, mutational studies suggested Spt4/5 positively affects transcription as the probability of RNAPII pausing or arresting when it encounters a nucleosome is increased when Spt4/5 function is impaired (35).
This led to the proposal that Spt4/5 helps RNAPII to overcome the nucleosome barrier, by either remodelling chromatin to allow the passage of RNAPII or by reassembling chromatin after the passage of RNAPII. To date, there has been no conclusive evidence showing that Spt4/5 directly associates with histones, chromatin or chromatin remodelling factors, so evidence that the protein complex modulates its effects on transcription through chromatin remains elusive.

Studies on the human and *Drosophila* Spt4/5 homologues, 5,6-dicholoro-1-β-D-ribofuranosylbenzimidazole (DRB) sensitivity inducing factor (DSIF) (39), support theories that the complex plays both a positive and negative role in transcription elongation. DSIF stimulates transcription in the presence of pTEFb and under certain conditions such as limiting concentrations of nucleoside triphosphate, but inhibits transcription when pTEFb is absent or inhibited by the drug DRB (39, 45). Studies also show that DSIF stimulates transcription elongation across the HIV genome through association with the HIV Tat protein (17, 18). This function is analogous to the role of NusG which stimulates transcription across the lambda genome through association with the N protein of the lambda phage (46).

In contrast to its positive role, DSIF has been shown to inhibit elongation through its interaction with the negative elongation factor (NELF) complex (47-50). Phosphorylation of both DSIF and NELF causes the latter to dissociate from the transcription elongation complex (TEC) allowing productive elongation (7). Sequence and functional conservation between yeast Spt4/5 and human DSIF suggest that they carry out a similar role (29, 30); however, *Saccharomyces cerevisiae* (S. cerevisiae) lacks a NELF homologue (49) and whether the yeast Spt4/5 complex can also negatively affect transcription elongation remains unclear.
In addition to its regulation of transcription elongation, recent observations suggest that Spt4/5 may couple the activities of the TEC to pre-mRNA processing. SPT5 mutational analysis shows an effect on splicing, polyadenylation and nuclear export of mRNA (51-54). Additionally, biochemical and proteomic studies show that Spt5 associates with a wide variety of 5’ and 3’ RNA processing factors including RNA capping enzymes, polyadenylation and RNA cleavage factors (9, 40, 51, 55). The recruitment of these processing factors is thought to occur through the differentially phosphorylated forms of the CTR in conjunction with the C-terminal domain of RNAPII (56). The interaction of Spt4/5 with these processing factors suggests that some kind of cross-talk occurs between the newly synthesized transcript, RNAPII, Spt4/5 and other processing factors; however, the exact mechanism of this cross-talk is yet to be elucidated.

1.3.2. Spt4/5 interacts directly with RNAPII

All Spt5 homologues have been shown to associate directly with RNAPII, providing a platform for the protein’s regulatory effect on transcription elongation (35, 46, 57). Recent structural studies revealed that the NGN domain of archael Spt4/5 binds directly to the clamp domain of RNAPII (58-60). The binding of Spt4/5 to RNAPII serves to bridge the central cleft of RNAPII, closing the DNA binding channel and holding the DNA in the elongation complex. Identification of a conserved RNAP binding surface on Spt5 allowed the modelling of NusG and the yeast Spt4/5 bound to their cognate RNAPIIs (58). The conserved binding surface suggests a similar mechanism of interaction for Spt5 in all domains of life (Figure 1.4).
Spt4/5 influences transcription elongation through its direct interaction with RNAPII and likely exerts its positive effects on elongation by preventing disengagement of the DNA template from the transcribing RNAP. This would reduce the likelihood of pausing or stalling, and enhance the processivity of transcription elongation. Spt4/5 is situated in such a way that it could allow the allosteric modulation of the RNAPII active site, as well as directly interacting with nucleic acids in the TEC and other regulatory elongation factors. The mechanisms of such regulation are unknown, thus further structural and functional analyses are required for clarification of these processes.

1.3.1. Spt4/5 interacts with nucleic acids

Recent studies have shown that eukaryotic Spt4/5 is recruited to the TEC when the nascent transcript is longer than 18-nt (61, 62), just long enough for the 5’ end of the RNA to protrude beyond the surface of RNAP (63). *Drosophila* Spt5 cross-links directly to this nascent RNA in the TEC (61), NusG proteins have been shown to interact with nucleic acids (64, 65), and the plant Spt5 homolog (KTF1) binds RNA both in vitro and in vivo (38). These data suggest that Spt4/5 may exert its effect on elongation or processing through direct contacts with the nascent transcript.
It is a widely accepted theory that the KOW domains of Spt5 mediate RNA binding as they have been suggested to make RNA contacts in several other RNA binding proteins (37, 64, 69, 70). The current model, based on the cryo electron microscopy structure of archael Spt4/5 bound to RNAPII, proposes that the first KOW domain contacts the downstream DNA whilst the second to fifth KOW domains contact the nascent RNA (58). Although KOW domain containing proteins have been implicated in RNA binding, there has been no conclusive evidence for this interaction. Thus, the actual mechanism of the KOW domain-RNA interaction is unknown.
1.4. Research Aims

Recent structures of Spt4/5 have provided detailed insight into the mechanism of Spt5 binding to Spt4, and have provided possible roles for the KOW domains in the complex; however, important questions are still left unanswered. How does Spt4 contribute to the function of the Spt4/5 complex? How are the KOW domains arranged in the complex and what is the biological significance of the multiple KOW domains in eukaryotic Spt5? Finally, what is the mechanism of RNA binding by the Spt4/5 complex?

Spt4/5 tightly associates with RNA polymerase II and regulates elongation and co-transcriptional pre-mRNA processing events; however, the mechanisms by which Spt4/5 acts are poorly understood. Like its bacterial homolog, NusG, recent studies of the human and Drosophila Spt4/5 complexes indicate that they can bind nucleic acids in vitro, yet there has been no conclusive evidence to date to support this theory. Therefore, the first aim of this study was to confirm the nucleic acid binding specificity of the yeast Spt4/5.

As the yeast Spt4/5 contains predicted RNA binding domains, the second aim of this study was to identify a specific RNA binding partner for Spt4/5 using de novo methods. The information would be used to identify a potential RNA retention motif that would be a possible cellular RNA target of Spt4/5. Characterising the RNA binding specificity of Spt4/5 would pave the way for future work studying the TEC.

A comprehensive understanding of the structure-function relationship of proteins is essential to clarify their role, and an important key step towards elucidating their mechanism of action. Therefore, the final aim of this study was to determine the molecular structure of Spt4/5 bound to its specific RNA target using X-ray
crystallography or nuclear magnetic resonance (NMR). It was anticipated that the structural information would provide fundamental insights into the mechanism of the RNA binding by Spt4/5, and its role in transcription elongation overall.

The research component of this thesis is presented in three distinct chapters: [1] The cloning, protein expression, purification and biophysical characterisation of multi-KOW domain Spt4/5 complexes, [2] the characterisation of the RNA binding specificity of Spt4/5, and [3] structural investigation of the Spt4/5-RNA binding interface. All chapters work towards the same goal of characterising the RNA binding specificity of the yeast transcription elongation factor Spt4/5, in order to elucidate the mechanism of how Spt4/5 is able to regulate transcription elongation.
2. Ubiquitin fusion constructs allow the expression and purification of multi-KOW domain complexes of the *Saccharomyces cerevisiae* transcription elongation factor Spt4/5

Amanda Blythe\textsuperscript{a}, Sanjika Gunasekara\textsuperscript{a}, James Walshe\textsuperscript{a}, Joel P. Mackay\textsuperscript{b}, Grant A. Hartzog\textsuperscript{c}, Alice Vrielink\textsuperscript{a}

\textsuperscript{a}School of Chemistry and Biochemistry, University of Western Australia, 35 Stirling Highway, Crawley, WA, 6009 Australia

\textsuperscript{b}School of Molecular Bioscience, University of Sydney, Sydney, NSW, 2006, Australia

\textsuperscript{c}Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA, 95064 USA

2.1. Abstract

Spt4/5 is a hetero-dimeric transcription elongation factor that can both inhibit and promote transcription elongation by RNA polymerase II (RNAPII). However, Spt4/5’s mechanism of action remains elusive. Spt5 is an essential protein and the only universally-conserved RNAP-associated transcription elongation factor. The protein contains multiple Kyrpides, Ouzounis and Woese (KOW) domains. These domains, in other proteins, are thought to bind RNA although there is little direct evidence in the literature to support such a function in Spt5. This could be due, at least in part, to difficulties in expressing and purifying recombinant Spt5. When expressed in
Escherichia coli (E. coli), Spt5 is innately insoluble. Here we report a new approach for the successful expression and purification of milligram quantities of three different multi-KOW domain complexes of Saccharomyces cerevisiae Spt4/5 for use in future functional studies. Using the E. coli strain Rosetta2 (DE3) we have developed strategies for co-expression of Spt4 and multi-KOW domain Spt5 complexes from the bi-cistronic pET-Duet vector. In a second strategy, Spt4/5 was expressed via co-transformation of Spt4 in the vector pET-M11 with Spt5 ubiquitin fusion constructs in the vector pHUE. We characterized the multi-KOW domain Spt4/5 complexes by western blot, limited proteolysis, circular dichroism, SDS-PAGE and size exclusion chromatography-multiangle light scattering and found that the proteins are folded with a Spt4:Spt5 hetero-dimeric stoichiometry of 1:1. These expression constructs encompass a larger region of Spt5 than has previously been reported, and will provide the opportunity to elucidate the biological function of the multi-KOW containing Spt5.

2.2. Introduction

Gene transcription is a highly dynamic and regulated process. Eukaryotes employ three multi-subunit RNA polymerases (RNAP I, II and III) to produce distinct classes of RNA. These RNAPs are related by a common evolutionary history, as well as common structures and mechanisms by which they transcribe RNA in a DNA-template-dependent manner (reviewed in (24)). As RNAPII transcribes protein-coding genes, it is the most studied of the three enzymes. In general, RNAPII mediated transcription can be broken down into three stages: initiation, elongation and termination. In contrast to initiation, which has been intensely studied for nearly five decades, elongation has only recently garnered significant attention. Recent studies have shown that during elongation, RNAPII is the target of many important regulators (reviewed in (71-73)).
RNAPII depends on a large number of accessory factors to promote efficient transcription in vivo. One such factor is the protein complex Spt4/5 (30, 35). Spt4/5 is the only known RNAPII-associated factor that has been conserved across all three domains of life, strongly suggesting that the complex plays an important role in transcription elongation. Functions such as transcription elongation processivity (35, 39, 49), RNA processing (9, 17, 18, 40), chromatin regulation (10, 74), transcription-coupled DNA repair (43, 75, 76) and developmental regulation (76, 77) have all been attributed to Spt4/5, yet the mechanism underlying these diverse roles is poorly understood.

Spt5 comprises a N-terminal acidic domain, a conserved NusG N-terminal (NGN) domain, a C-terminal repeat region and five proposed RNA binding domains (called KOW domains after the authors Kyrpides, Ouzounis and Woese (36), Figure 2.1). In archaea and eukaryotes, hydrophobic interactions promote binding of the NGN domain to Spt4, a 12-kDa zinc finger protein (28, 30). Spt4 bound to the NGN domain of Spt5 has been shown to associate with RNAPII, completely encircling the DNA in the RNAPII binding channel during transcription (58, 59). The first KOW domain of Spt5 has been proposed to stabilise the upstream DNA as it emerges from the transcription bubble, and the second to fifth KOW domains are thought to contact the nascent RNA as it emerges from RNAPII (59).

The KOW domain was first described in 1996 as a novel protein motif shared between the prokaryotic transcription elongation factor NusG and ribosomal proteins (36). The domain contains 27 residues with a conserved glycine at position 11, and is arranged into 5 alternating blocks of hydrophobic and hydrophilic residues, with the hydrophobic
blocks forming beta sheets \((36)\). The domains are strictly conserved in several ribosomal proteins suggesting they play an integral role in RNA binding \((36)\).

**Figure 2.1. Structural organization of the Spt4 and Spt5 proteins**

A. FoldIndex prediction of the probability of foldedness in yeast Spt5. FoldIndex predictions were made using a 51 residue window. B. Domain structures of *S. cerevisiae* Spt4 showing the zinc finger (Zn) and Spt5 showing the acidic N-terminus, the Kyprides, Ouzounis and Woese (KOW) domains and the C-terminal repeat (CTR) region. C. Schematic diagram showing domain truncations of the Spt5 constructs made for the expression and purification of Spt4/5 hetero-dimers.

In order to better understand the function of the multiple KOW domains of *Saccharomyces cerevisiae* (*S. cerevisiae*) Spt4/5, we cloned, expressed and purified three different hetero-dimeric complexes containing the full length Spt4 and truncations of Spt5 containing the NGN domain plus the first, first and second, and first to fifth KOW domains. Through biophysical characterisation of each Spt4/5 complex we found the proteins to have substantial secondary structure and 1:1 hetero-dimeric stoichiometry. These proteins are therefore tools for use in further studies.
2.3. Materials and methods

2.3.1. Cloning

Cloning was carried out using standard PCR based methods. Insert sequences coding for the wild type *S. cerevisiae* *SPT4* and *SPT5* genes were prepared by PCR from plasmids using primers incorporating appropriate restriction enzyme sites. Initially, co-expression of *S. cerevisiae* Spt4 and Spt5 in *E. coli* was achieved using the pET-Duet-1 expression vector. In all cases SPT4 was cloned into the first multiple cloning site (MCS) and SPT5 constructs into the second MCS. All Spt5 constructs contained a carboxy terminal (C-terminal) hexahistidine tag (His tag). The boundaries of each Spt5 construct are shown in Figure 2.1C.

The same *SPT5* sequences were cloned into the expression vector pHUE to form N-terminal His-tagged ubiquitin fusion constructs (Spt5ubq). To facilitate Spt4/5ubq co-expression, the full length *SPT4* gene was cloned into pET-M11. This construct contains no affinity tag. For the over-expression of Spt4 alone, the full length *SPT4* gene was cloned into pET28b (Novagen) and contained an amino terminal (N-terminal) His-tag. The primers and restriction enzymes used for the cloning of each construct are listed in Appendix 1. A list of all plasmids cloned in this study and their corresponding protein products are shown in Appendix 2. All constructs were verified by DNA sequencing.

2.3.2. Protein expression

Chemically competent *E. coli* Rosetta2 (DE3) cells (78) were transformed with the appropriate pET-Duet-1 plasmid or co-transformed with pETM11-SPT4 and the appropriate pHUE-SPT5 plasmid. pET-28b-SPT4 was transformed into chemically competent *E. coli* BL21 (DE3) cells.
Luria-Bertani (LB) medium containing the appropriate antibiotics was inoculated with a single colony of transformed Rosetta2 (DE3) or BL21 (DE3) *E. coli* and incubated overnight, shaking at 37 °C. Each overnight culture was used to inoculate 1 L of LB medium containing appropriate antibiotics to an optical density (600 nm; OD) of 0.1. The antibiotic resistance and concentrations used during expression are outlined in Appendix 2. The cultures were shaken at 37 °C until the OD$_{600}$ reached 0.6-0.8 at which point they were supplemented with 0.3 mM ZnCl$_2$, which was necessary for optimal expression of Spt4, and protein expression was induced with 1 mM IPTG at 30 °C for 3 h (pET-Duet constructs) or at 22 °C overnight (pHUE and Spt4 constructs). Culture samples were taken prior to induction (T$_0$) and at the time of harvest (T$_F$) and expression levels analysed by SDS-PAGE. To ensure equal SDS-PAGE loading, the samples were normalised to comparable OD$_{600}$ levels. Cells were pelleted by centrifugation (5000 g) for 15 min before being flash frozen in liquid nitrogen.

### 2.3.3. Cell lysis

Cell pellets from protein over-expression in pET-Duet were resuspended in pET-Duet lysis buffer (10 mM Hepes pH 7.5, 10 mM imidazole, 10% glycerol, 0.1% Triton X-100, 0.5 M KCl, 1 mM β-mercaptoethanol (β-ME)). The cells were disrupted by sonication on ice for 30 s at 20% amplitude using a microtip (Sonifier B-12, Branson). Sonication was repeated twice. The lysate was then clarified by centrifugation (20,000 g, 4 °C, 40 min) and the protein of interest purified from the soluble fraction.

Cell pellets from protein over-expression in pHUE were resuspended in pHUE lysis buffer (50 mM sodium phosphate pH 7.4, 20 mM imidazole, 20% glycerol, 2 M KCl, 10 mM β-ME). The cells were disrupted with an Emulsiflex C5 high-pressure
homogeniser (Avestin), the lysate clarified by centrifugation (20,000 g, 4 °C, 40 min) and the protein of interest purified from the soluble fraction.

SDS-PAGE was used to assess protein solubility. A fraction of the lysate was separated into the soluble (S) and insoluble fraction (I) by centrifugation at 15,000 rpm. To ensure equal loading of the SDS-PAGE gels, each sample was made up to 25 μL with SDS-PAGE loading dye.

2.3.4. Purification

Proteins were purified by nickel-affinity chromatography performed on an ÄKTA Purifier (GE Healthcare) using a 1 ml Ni-NTA superflow cartridge (QIAGEN) and the proteins eluted with an imidazole gradient (50 mM sodium phosphate pH 7.4, 0.3 M KCl, 20% glycerol, 20-500 mM imidazole). Protein elution was monitored spectrophotometrically by absorbance at 280 nm and peak fractions were pooled, concentrated and further purified by size exclusion chromatography on a HiLoad 16/60 Superdex™ 200 or Superdex™ 75 preparative-grade gel filtration column (GE Healthcare) in 50 mM sodium phosphate pH 7.4, 0.3 M KCl, 20 mM imidazole and 10% (w/v) glycerol. Data were collected and analysed using the Unicorn™ 5.11 software (GE Healthcare). Protein purity was assessed by SDS-PAGE, and fractions were concentrated to 3–10 mg mL⁻¹ as estimated from the absorbance at 280 nm. Extinction coefficients were calculated using ProtParam (79).

2.3.5. SDS-PAGE and western blot analysis

20 μL protein samples were prepared by the addition of 5 μL of SDS-PAGE loading dye (0.125 M Tris-HCl pH 6.8, 10% (w/v) glycerol, 10% β-ME, 2% SDS, 0.1% bromophenol blue) and heat-treated at 95 °C for 5 min prior to electrophoresis. The samples were resolved on 15% glycine SDS-PAGE mini gels in Tris/glycine buffer (25
mM Tris, 250 mM glycine, 0.1% SDS). Gels were electrophoresed using a Hoefer Mighty Small II system at 150 V for 80 minutes and stained with 0.1% Coomassie brilliant blue dye (Sigma) or transferred to AmershamTM HybondTM-ECL membranes (GE Healthcare) at 100 V, 170 mA for 70 min. Antisera against *S. cerevisiae* Spt4 (35) and anti-rabbit IgG antibody conjugated to horseradish peroxidase (HRP; Santa Cruz Biotechnology) were used at 1:2000 dilutions. HRP was activated using Amersham™ ECL Prime Western Blotting Detection Reagent (GE Healthcare) and western blots exposed for 1 min on a FujiFilm LAS-1000 luminescent image analyser (FujiFilm).

### 2.3.6. Limited proteolysis

His-tagged Spt4 at 0.2 mg mL⁻¹ was incubated overnight at 37 °C with trypsin, chymotrypsin or subtilisin (Sigma) at a protein to enzyme ratio of 50:1. 10 μL samples were taken at 0, 1, 3 and 20 h and analysed by SDS-PAGE on 15% Tris-glycine gels.

### 2.3.7. SEC-MALLS

Size estimates and oligomeric states of purified Spt4/5ubq hetero-dimer complexes were measured by size exclusion chromatography followed by in-line multi-angle laser light scattering (SEC-MALLS). Proteins were separated by size exclusion chromatography (SEC) with Superose™ 12 10/30 or Superdex™ 75 10/30 gel filtration columns (Amersham-Pharmacia) in 50 mM sodium phosphate pH 7.4, 0.3 M KCl, 20 mM imidazole and 10% (w/v) glycerol and monitored at 215 nm, 260 nm and 280 nm on an ÄKTA purifier (GE Healthcare). The eluent was further monitored by an in-line miniDawn Tristar multi-angle light scattering instrument (MALLS) containing a laser light source at 690 nm and three detector angles in combination with an Optilab DSP refractive index detector (Wyatt Technology Corp.). Voltage/light scattering intensity was calibrated against a standard sample of HPLC grade toluene, yielding a calibration constant of $8.461 \times 10^{-6}$ V cm⁻¹ and scattering detectors normalised with respect to the
90° detector using monomeric BSA (Sigma). Data were collected and analysed using Astra software (Wyatt Technology) and molecular masses calculated using the specific dN/dC (refractive index increment) values for protein (0.190 mL g⁻¹).

### 2.3.8. Circular dichroism spectropolarimetry (CD)

Far-UV spectra were measured at 20 °C on a CD spectrophotometer (Jasco J-720) using a 200 μm path length quartz cuvette. Data were collected every 1 nm with a 1 nm bandwidth using an integration time of 5 s per step. Protein samples were diluted to 0.1 mg mL⁻¹ in 100 mM KF and the CD spectra measured between 260 and 195 nm. The resulting spectra represent the average of three accumulations and are buffer-baseline corrected. Thermal melt analysis was carried out at 220 nm from 20–80 °C. Melting temperatures were calculated from the inflection point at which the fraction of unfolded protein was 0.5.

### 2.4. Results and Discussion

We designed recombinant *S. cerevisiae* Spt4/5 complexes in the bicistronic plasmid pET-Duet-1 containing full length Spt4 and His-tagged derivatives of Spt5. The protein boundaries of Spt5 in these expression constructs were chosen to avoid large regions that were predicted to have no secondary structure (Figure 2.1A) and so the Spt5 polypeptides do not contain the acidic N-terminus or C-terminal repeats. Additionally, the role(s) of the five KOW domains is unknown; therefore, the expression constructs were designed with differing numbers of KOW domains. The resulting Spt5 derivatives encompass the NGN domain plus one, two or five KOW domains (amino acids 284-419 (Spt4/5284-419), 284-567 (Spt4/5284-567) and 284-839 (Spt4/5284-839), respectively (Figure 2.1B and C)). The pET-Duet-1 vector was chosen to allow the co-expression of each gene in a 1:1 stoichiometric ratio, and an N-terminal His-tag allows for simple purification.
Initial attempts at expressing the hetero-dimer complexes resulted in low expression levels for all three constructs. Additionally, for Spt4/5\textsubscript{284-419} and Spt4/5\textsubscript{284-567} the proteins expressed were insoluble (Figure 2.2). Following SDS-PAGE analysis, only bands for Spt5 could be visualised. This observation is discussed in the following sections. Subsequent purification of these complexes by nickel affinity chromatography resulted in low yields (0.1 mg from 2 g cell pellet) of protein with a low level of purity (~40%, data not shown). Over-expression of Spt4/5\textsubscript{284-839} could not be detected by SDS-PAGE analysis and while it appears that this complex was soluble, the final yield of the purified protein was negligible and also subject to degradation that was not inhibited by any of the commonly available protease inhibitors (data not shown).

A common approach to increase the solubility of a protein is to fuse the protein to another highly soluble protein. Eukaryotic ubiquitin fusion constructs have been demonstrated to increase both the solubility and expression of previously insoluble and poorly expressed proteins (80, 81). The cloning of a target protein into the multiple cloning site of the vector pHUE enables the construction of a cleavable His-tagged ubiquitin fusion construct (80). As the insertion of only a single gene is possible using the pHUE vector, we subcloned Spt5 into this vector and co-expression of recombinant Spt4/5ubq hetero-dimers was achieved through the co-transformation of pHUE-SPT5 and pETM-11-SPT4 in Rosetta 2 (DE3).

We assessed the solubility of over-expressed Spt4/5ubq\textsubscript{284-419} in \textit{E. coli} lysates using SDS-PAGE. Comparison of the soluble and insoluble protein fractions indicated that approximately 50% of the protein was soluble (Figure 2.3). Although Spt4/5ubq\textsubscript{284-419} was not completely soluble, the expression of the protein as an ubiquitin fusion clearly
yielded a significant improvement over the solubility of this construct when expressed from the pET-Duet vector.

By obtaining soluble protein we were able to successfully purify Spt4/5ubq284-419 under native conditions using immobilised metal affinity chromatography (IMAC). Following the elution of the protein from a nickel-NTA column, assessment of purity by SDS-PAGE revealed that the protein was approximately 80% pure (Figure 2.3). His-tags are among the most commonly used affinity tags for high throughput protein purification; however, the high yield attainable from the use of His-tags is often compromised by a low purity due to contaminant host proteins having a natural affinity for metal ions, or containing cysteine, tryptophan and/or histidine rich regions (82). The impurity of Sp4/5ubq284-419 following IMAC can partially be attributed to the co-purification of free his-tagged ubiquitin. Premature ubiquitin processing is a common anomaly in the expression of ubiquitin-fusion constructs, however is negligible due to the large amounts of intact fusion proteins frequently isolated (80, 83).
Figure 2.3. Expression and solubility of Spt4/5ubq284-419 in pHUE

SDS-PAGE analysis of protein expression pre (T₀) and post (Tᵢ) induction, the soluble (S) and insoluble (I) fractions following cell lysis, subsequent purification by nickel affinity chromatography (His) and molecular weight markers (M). Positions of Spt4 and Spt5 are indicated. Dashed line indicates where lanes have been removed from the image. The final concentration of β-ME in each sample was 0.3 M (2%).

Common procedures to reduce contaminant proteins include increasing the imidazole concentration or adjusting the ionic strength of the wash buffers (84). Furthermore, in many cases, a second purification step is often required to achieve the desired purity. To improve the purity of Spt4/5ubq complexes, we increased the concentration of KCl to 2 M in the lysis and binding buffer. Additionally, a subsequent size exclusion chromatography step was added to the purification process. Size exclusion chromatography was successful in separating the free ubiquitin tag and other impurities from the Spt4/5ubq complexes, and using this purification protocol, each Spt4/5ubq fusion construct was successfully purified under native conditions resulting in milligram quantities of soluble and homogeneous protein (Figure 2.4). A summary of the purification showing the optimal yield obtained for each fusion construct is outlined in Table 2.1.
Figure 2.4. Visualisation of Spt4 in the purified Spt4/5ubq hetero-dimers
A. 15% SDS-PAGE gel of Spt4 alone and Spt4/5ubq hetero-dimers. Positions of Spt4 and Spt5ubq proteins and molecular weight markers (M) are indicated. B. Western blot of the same gel using an antibody targeted to Spt4. Dashed line indicates where lanes have been removed from the image. The final concentration of β-ME in each sample was 0.3 M (2%).

Table 2.1. Purification summary for each Spt4/5ubq complex

<table>
<thead>
<tr>
<th>Protein complex</th>
<th>Cell pellet (g)</th>
<th>His (mg)*</th>
<th>SEC (mg)#</th>
<th>mg/g cell pellet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spt4/5ubq\textsubscript{284-419}</td>
<td>3.5</td>
<td>14.3</td>
<td>2.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Spt4/5ubq\textsubscript{284-567}</td>
<td>3.7</td>
<td>35</td>
<td>11.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Spt4/5ubq\textsubscript{284-839}</td>
<td>3.4</td>
<td>52</td>
<td>14</td>
<td>4.1</td>
</tr>
</tbody>
</table>

* Yield given as milligrams of soluble protein purified from Ni-affinity chromatography (His).
# Yield given as milligrams of soluble protein following size exclusion chromatography (SEC).

SDS-PAGE analyses of the Spt4/5ubq hetero-dimers commonly showed apparent stoichiometry differences between Spt4 and Spt5ubq, and often Spt4 could not be visualised at all. To confirm the presence of weakly stained Spt4 in the purified hetero-dimers, two identical 15% SDS-PAGE gels were electrophoresed with 150 pmol of Spt4
alone and each Spt4/5 construct. One gel was stained with Coomassie and the second gel was transferred to nitrocellulose membrane for western blotting with an antibody targeted to Spt4. The Coomassie stained gel showed noticeable differences for all constructs in terms of the relative amounts of Spt4 versus Spt5 in the hetero-dimer (Figure 2.4A). Additionally, even though equal amounts of all constructs were loaded in each lane, the amounts of Spt4 across the samples appeared comparatively different. The western blot confirmed the presence of Spt4 in each purified hetero-dimer and level of signal across the gel also indicated that there are equal amounts of the protein in all samples. It is possible that the additional basic histidine residues conferred on Spt4 alone via the his-tag contributes to the enhanced Coomassie staining ability of this protein compared to the non-tagged Spt4 in the Spt4/5ubq complexes. These results suggest that the apparent stoichiometric differences seen for Spt4/5 in SDS-PAGE gels are simply an anomaly of Coomassie staining.

Even though the theoretical molecular weight estimation for Spt4 is 12.1 kDa, the protein often ran as a doublet with a lower molecular weight (MW) band at the predicted size of 12 kDa and a higher MW band at 14 kDa (Figure 2.4A). The anti-Spt4 antibody recognised both of these proteins, providing positive identification of each band as *S. cerevisiae* Spt4 (Figure 2.4B). The presence of two different MW bands for Spt4 has been seen previously in the purification of human Spt4 (31, 61). The authors suggested that the lower band is a degradation product containing C-terminal truncations of the protein.

To investigate the possibility of proteolytic cleavage, Spt4 alone was subjected to limited proteolysis using trypsin, chymotrypsin and subtilisin over a time period of 20 h. SDS-PAGE analysis showed that none of the enzymes could reduce the higher MW
band to the size of the lower MW band, and that all three enzymes were able to digest Spt4 in its entirety (Figure 2.5A). Additionally, the no-enzyme control did not show any time-dependent degradation of the higher MW species into the species of the lower MW. These data suggests that both Spt4 species are stable, and that the higher mobility species is not a product of proteolytic cleavage during the expression and purification process.

Figure 2.5. Analysis of Spt4 doublet

A. Limited proteolysis of Spt4 analysed by SDS-PAGE followed by Coomassie staining. His-tagged Spt4 was digested with three common proteases: trypsin, subtilisin and chymotrypsin. Samples were taken at 0, 1, 3, and 20 h. B. The effect of β-ME and heat on the gel mobility of Spt4. His-tagged Spt4 was treated with or without β-ME and/or heat denaturation at 95 °C, separated by SDS-PAGE and then detected by anti-Spt4 western blotting. For A and B, the final concentration of β-ME in each sample was 0.3 M (2%). C. β-ME has a concentration dependent effect on Spt4. His-tagged Spt4 was treated with increasing amounts of β-ME and heat denatured before electrophoresis on 15% SDS-PAGE gels.

To further examine the occurrence of multiple Spt4 species, we investigated the effects of reduction and heat denaturation on the protein using SDS-PAGE. When no β-ME was present, Spt4 migrated through a 15% gel as a single band at the predicted MW of 12.1 kDa. Denaturing the sample by boiling prior to electrophoresis had no effect on the migration of Spt4; however, the addition of β-ME to the sample resulted in the
appearance of the higher MW species (Figure 2.5B). This indicates that the reduction of the cysteine residues in Spt4 causes the slower motility of the protein through an SDS-PAGE gel.

To determine if the effect of β-ME on Spt4 was concentration dependent, Spt4 was incubated with increasing concentrations of β-ME and electrophoresed on a 15% SDS-PAGE gel. The higher MW band appeared at β-ME concentrations in excess of 50 mM; however, the shift of the lower MW Spt4 band to the higher MW band was not complete even at a β-ME concentration of 1 M (Figure 2.5C). We are unable to explain why a high concentration of β-ME results in such a specific MW species, or why there is only partial shifting to the higher MW. However, we conclude that the higher MW species seen for Spt4 is an SDS-PAGE specific anomaly. The typical concentration of reducing agents used in protein purification is between 1 and 20 mM. As the concentration dependent shift of Spt4 did not occur until β-ME concentrations were in excess of 50 mM, we conclude that the 10 mM β-ME used in the preparation of Spt4 or Spt4/5 hetero-dimers is unlikely to affect the downstream applications in the functional or structural analysis of Spt4/5.

The difference in the Coomassie staining of Spt4 compared to Spt5 has been seen in other S. cerevisiae and Pyrococcus furiosis Spt4/5 constructs (58, 61); and previously an Spt4/5 complex from yeast could only be crystallised when fused and produced as a single polypeptide. This is suggestive of an uncertain stoichiometry for yeast Spt4/5 produced in bacteria. Therefore, in order to assess the oligomeric state of the Spt4/5ubq hetero-dimers in solution, the protein complexes were subjected to SEC-MALLS analysis. The proteins were applied to a Superose 12 column with an in-line MALLS detector and stoichiometry was estimated by determining the weight-averaged
molecular weight in solution. Each Spt4/5ubq complex eluted with a molecular mass of comparable value to the theoretical molecular mass as calculated using ProtParam (79) (Figure 2.6). The MALLS analysis conclusively identify a 1:1 stoichiometry for each Spt4/5 hetero-dimer complex.

Figure 2.6. SEC-MALLS analysis of Spt4/5ubq hetero-dimers
A. The absorbance (continuous line) and light-scattering (dotted line) are shown for each Spt4/5ubq hetero-dimeric complex. B. Summary of the theoretical monomeric and experimentally determined molecular weight of each Spt4/5ubq complex.

<table>
<thead>
<tr>
<th>Protein Complex</th>
<th>Theoretical Molecular Weight</th>
<th>MALLS Calculated Weight (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spt4/5ubq284-419</td>
<td>37.6</td>
<td>38 ± 0.4%</td>
</tr>
<tr>
<td>Spt4/5ubq284-567</td>
<td>60.9</td>
<td>50 ± 2.7%</td>
</tr>
<tr>
<td>Spt4/5ubq284-839</td>
<td>83.9</td>
<td>79 ± 3%</td>
</tr>
</tbody>
</table>

Often, expression and purification of recombinant proteins does not result in stably folded protein. To assess the conformational properties of the Spt4/5ubq hetero-dimers, circular dichroism (CD) spectroscopy was performed on each protein complex. Each resulting spectrum showed characteristic absorption minima for a protein containing both β-strands and α-helices (Figure 2.7A). These data indicated that the isolated hetero-dimers exhibited secondary structure.
CHAPTER 2

Figure 2.7. Physical properties of multi-KOW Spt4/5ubq hetero-dimers

A. CD spectra for each hetero-dimer construct. Each spectrum was buffer baseline corrected and smoothed using a 15-point Savitzky-Golay filter.  
B. Thermal denaturation at 220 nm of each hetero-dimer showing the fraction of unfolded protein (F\textsubscript{unfolded}) versus increasing temperature.

CD thermal melts were carried out at 220 nm on the same protein samples to assess the stability of the Spt4/5 complexes. Melting temperatures of 52 °C, 66 °C and 64 °C were obtained for Spt4/5ubq\textsubscript{284-419}, Spt4/5ubq\textsubscript{284-567}, and Spt4/5ubq\textsubscript{284-839} respectively (Figure 2.7B). The relatively high inflection points for these proteins indicate that these complexes have a stable secondary structure. Overall, the CD analysis show that each Spt4/5ubq hetero-dimer exhibits stable secondary structure, making these complexes useful tools for further functional and structural characterisation.

We successfully established a fast and simple expression and purification protocol for multi-KOW domain Spt4/5 hetero-dimers from S. cerevisiae using E. coli as a
heterologous expression host. Expression of the Spt5 derivatives as ubiquitin fusion constructs allowed significant amounts of pure and soluble protein to be produced for use in functional and structural studies. Biophysical characterisation has shown that each Spt4/5ubq complex is well folded with 1:1 hetero-dimeric stoichiometry, thus being of sufficient quality for future downstream applications.

2.5. Acknowledgments

We acknowledge funding support for this work from the Raine Medical Research Foundation. We thank Prof. Charlie Bond (University of Western Australia, Australia) for providing the pET-Duet vector and Dr. Chris Wanty (Australian National University, Australia) for providing the pHUE vector. We thank Dr. Phillipa Stokes (University of Sydney, Australia) for her assistance with the SEC-MALLS work.

2.6. Author contributions

A.B., S.G., G.A.H. and A.V. designed research; A.B., S.G, and J.W. performed research; A.B., J.P.M. and A.V. analysed data and A.B., G.H., J.P.M. and A.V. wrote the paper.

The authors declare no conflict of interest
3. Yeast transcription elongation factor Spt4/5 is a sequence specific RNA binding protein.

Amanda Blythe\textsuperscript{a}, Berra Yazar-Klosinski\textsuperscript{b}, Michael W Webster\textsuperscript{c}, Eefei Chen\textsuperscript{d}, Marylène Vandevenne\textsuperscript{c}, Katerina Bendak\textsuperscript{e} Joel P Mackay\textsuperscript{c}, Grant A. Hartzog\textsuperscript{b}, Alice Vrielink\textsuperscript{a,1}

\textsuperscript{a}School of Chemistry and Biochemistry, University of Western Australia, Crawley, Western Australia, 6009 Australia
\textsuperscript{b}Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, California, 95064 USA
\textsuperscript{c}School of Molecular Bioscience, University of Sydney, Sydney, New South Wales, 2006 Australia
\textsuperscript{d}Department of Chemistry and Biochemistry, University of California, Santa Cruz, California, 95064 USA

\textsuperscript{1}To whom correspondence should be addressed.

3.1. Abstract

The transcription elongation factor Spt4/Spt5 (Spt4/5) tightly associates with RNAPII regulating elongation and co-transcriptional pre-mRNA processing events; however, the mechanisms by which Spt4/5 acts are poorly understood. Recent studies of the human and Drosophila Spt4/5 complexes indicate that they can bind nucleic acids \textit{in vitro}. We demonstrate here that yeast Spt4/5 can bind nucleic acids \textit{in vitro} in a sequence specific
manner to single stranded RNA containing multiple AA repeats. Furthermore, we show that the major protein determinant for binding is Spt4 together with the NGN domain of Spt5 and that the KOW domains are not required for the specific RNA interaction. Genomic analysis shows that AA motifs are enriched in the 5’ untranslated regions of transcribed sequences and are associated with an increased transcription frequency, providing a functional rationale for the observed binding specificity. Finally, we propose that the RNA binding activity of Spt4/5 is mediated through the non-canonical use of a previously unrecognized RNA recognition motif in Spt5.

3.2. Introduction

Transcription is a highly dynamic and regulated process which in eukaryotes is carried out by three multi-subunit RNA polymerases (RNAPI, II and III) to produce distinct classes of RNA. These RNAPs are related through their common evolutionary histories, their structures and the mechanisms by which they transcribe RNA in a DNA-template-dependent manner. The catalytic cores of multi-subunit polymerases found in all living organisms display deep conservation (24). In contrast, the general regulatory proteins that assist and direct the activities of these RNAPs generally exhibit significant variation in structure and function across the three kingdoms of life and between functional classes of polymerase. However, a single family of transcriptional regulators displays the same degree of conservation as seen in RNAP, namely the Spt5/NusG family (24). These proteins are essential for life and are known to regulate transcription elongation in eukaryotes, archaea and bacteria. Their strict conservation across all domains of life suggests that they carry out an ancient, core function. The details of that function are, however, largely unknown.
Eukaryotic Spt5 is a large multi-domain protein consisting of an N-terminal acidic domain, a NusG N-terminal (NGN) domain, several Kyprides, Ouzounis, Woese (KOW) domains and a set of C-terminal repeats (CTRs) whose sequence varies across species (85, 86) (Supplementary figure 10.1A). The conserved core of Spt5, comprising the NGN domain and a single KOW domain, is found in archael Spt5 and the bacterial homolog NusG. In eukaryotes and archaea (but not bacteria), Spt5 forms a non-covalent complex with a small zinc finger protein, Spt4 (RpoE” in archaea), via its NGN domain (28, 30, 31).

The mechanism by which the Spt4/5 hetero-dimer regulates elongation is not well understood; however, recent structural studies have suggested that the NGN domain of NusG and Spt4/5 binds directly to RNAP bridging the central cleft of the RNAP (58-60). This arrangement effectively seals the DNA into the elongation complex and may prevent the disengagement of the DNA template from the transcribing RNAP, reducing the likelihood of pausing or stalling and enhancing the processivity of transcription elongation. This model also places Spt4/5/NusG in a location where it can allosterically modulate the RNAP active site and potentially interact with nucleic acids in the transcription elongation complex (TEC).

Several observations indicate that, in addition to regulating elongation, Spt4/5 may couple the activities of the TEC to pre-mRNA processing. The CTR region of Spt5 is required for normal elongation control (87, 88) and genetically interacts with the CTD of RNAPII (56, 89). Through regulated phosphorylation, the CTR of Spt5 may serve as a scaffold for the cooperative assembly of transcription and RNA processing factors (56, 88, 90-93). Biochemical and proteomic studies show that Spt5 associates with a wide variety of 5’ and 3’ RNA processing factors including RNA capping enzymes,
polyadenylation factors and RNA cleavage factors \((9, 40, 41, 51, 55, 94)\). Additionally, \textit{spt5} mutations affect splicing, polyadenylation and nuclear export of mRNA \((51-54, 95)\).

Recent evidence has suggested that Spt4/5 may exert its effects on elongation or processing through direct contacts with the nascent transcript. NusG and Spt4/5 can interact with nucleic acids \((47, 64, 65, 96)\). Furthermore, \textit{in vitro} transcription studies suggest that Spt4/5 associates with nascent transcripts soon after they emerge from the elongating polymerase, and that efficient association of Spt4/5 with RNAPII may depend on transcript binding \((61, 62)\). Several observations have led to the suggestion that the KOW domains of Spt5 may mediate RNA binding. KOW domains in ribosomal proteins and rRNA processing factor Mtr4 have been observed to directly contact RNA, but the contribution of the KOW domain to the nucleic acid binding activity of Spt4/5 is unknown \((37, 97, 98)\).

We show here that Spt4/5 has sequence specific RNA binding activity, and that the Spt5 NGN-Spt4 hetero-dimer, in the absence of KOW domains, is responsible for this functionality. We further show that Spt4/5 displays specificity for RNA bearing multiple AA repeat elements, and we propose a model by which Spt4/5 uses this motif to exert its regulatory effects on transcription elongation.

### 3.3. Methods

#### 3.3.1. Cloning, expression and purification.

The Spt4 and multi-KOW domain Spt5 constructs used in this study were cloned, expressed and purified as described previously in Chapter 2 \((99)\). The Spt5\textsubscript{284-375}
construct was cloned into the pHUE vector, expressed as an ubiquitin fusion construct and purified in an identical manner.

The individual and tandem KOW domains constructs were cloned into the vector pGEX-6p1 (GE Healthcare) and expressed as GST-fusion constructs in the E. coli strain Rosetta 2 (DE3) (Novagen). Primers used are listed in Appendix 1. E. coli transformants containing the recombinant plasmids were incubated in Luria-Bertani (LB) growth medium supplemented with 50 μg ml⁻¹ ampicillin and 34 μg ml⁻¹ chloramphenicol at 37 °C. When the optical density at 600 nm reached 0.6–0.8, protein expression was induced with 0.2 mM IPTG. The induced culture was incubated at 25 °C overnight, harvested by centrifugation (4000 x g, 4 °C, 15 min), the bacterial pellets resuspended in 50 mM Tris pH 8, 0.2 M NaCl, 20 mM imidazole, 1% Triton-X and lysed using an Emulsiflex C5 high pressure homogeniser (Avestin). The lysate was clarified by centrifugation (20,000 x g, 4 °C, 45 min) before loading onto a 1 ml GSTrap column (GE Healthcare) and washing with 50 mM Tris pH 8, 300 mM NaCl, 10% glycerol. Affinity chromatography was performed on an ÄKTA purifier (GE Healthcare) and recombinant protein eluted with glutathione (50 mM Tris pH 8, 150 mM NaCl, 0.005% Triton-X, 20 mM glutathione).

3.3.2. Pentaprobe and oligonucleotide preparation.

Pentaprobes were prepared as described in (100). Synthetic oligonucleotides (Dharmacon) were 2’-deprotected, according to the manufacturers instructions, prior to 5’ end-labelling with [γ-³²P] ATP (PerkinElmer) using T4 polynucleotide kinase.

3.3.3. Gel shift assay.

Radiolabelled nucleic acid probes were incubated with protein samples in gel shift buffer (10 mM MOPS pH 7.0, 50 mM KCl, 5 mM MgCl₂, 10% glycerol and 0.03 mg
ml⁻¹ heparin) at 4 °C for 30 min. Heparin is a negatively charged, sugar-containing macromolecule that bears structural similarity to RNA and DNA. Therefore, it is commonly used as a competitor in RNA or DNA binding assays to reduce nonspecific and background binding ([100, 101]). The binding reactions were electrophoresed on a pre-equilibrated 6% native acrylamide/bisacrylamide gel (19:1) in 0.5X Tris-borate buffer at 200 V for 2 h at room temperature. Gels were visualized on a Phosphor screen with a Typhoon PhosphoImager™ FLA 9000 (GE Healthcare).

5’ 56FAM (Integrated DNA Technologies) labelled RNA probes (5 nM) were incubated with protein samples in gel shift buffer at 4 °C for 30 min. The binding reactions were electrophoresed on a pre-equilibrated 6% native acrylamide/bisacrylamide gel (19:1) in 0.5X Tris-Hepes buffer at 65 mA for 1 h at 4 °C. Gels were visualized with a Typhoon Trio Variable Mode Imager (Amersham Biosciences). The sequences of the RNA probes are listed in Appendix 1.

3.3.4. SELEX.

The initial SELEX library was prepared by treating a synthetic oligonucleotide pool containing a 25-nt random sequence flanked by two primer binding sites with DNA polymerase I large (Klenow) fragment (New England Biolabs) at 37 °C for 30 min and purified using a PCR clean up kit (QIAGEN). The library was transcribed using the T7-RiboMAX™ Large Scale RNA production system (Promega). Unincorporated nucleotides were removed with Sephadex® G-25 Quick Spin™ columns (Roche) and RNA extracted by phenol/chloroform and ethanol-precipitation. Binding reactions were carried out in SELEX Buffer (40 mM MOPS pH 7.0, 20 mM KCl, 10 mM MgCl₂, 10% glycerol, 0.2% Triton X-100, 0.2 mM PMSF, 2 mM DTT). Each 150 μl binding reaction contained 5–80 pmol of protein immobilised on MagneHis™ Ni-Particles
(Promega) containing 0.06–1.6 mg ml\(^{-1}\) heparin sulphate and 80 pmol RNA, and mixed for 60 min at 4 °C. Unbound RNA was removed and the beads were washed 5 times with SELEX buffer (500 μl). Bound RNA was eluted from the immobilized protein in water by vigorous mixing at 95 °C for 15 min. The selected RNA was ethanol precipitated and reverse-transcribed using a complementary primer, then amplified by 10 or 15 cycles of PCR with Pfu DNA polymerase. The PCR products were transcribed into RNA and applied to fresh protein coupled MagneHis\(^{TM}\) Ni-Particles and the cycle was repeated. A total of seven rounds of SELEX were completed. The primers and SELEX oligonucleotide are listed in Appendix 1.

### 3.3.5. High-throughput sequencing and analyses of selected sequences.

Purified PCR products from each round of SELEX were re-amplified with barcoded primers and sequenced on the HiSeq2000 (Illumina) sequencing platform using a single-end, 50 nucleotide sequencing protocol at the University of California, Terasaki Life Sciences Building. The primers used for the addition of the barcodes are listed in Appendix 1. Each barcode contained between 2.4-5.5 million reads. A workable data set of 10,000 sequences per barcode was taken, and barcodes and duplicate sequences removed. 2000 of these unique sequences were randomly selected and subjected to analysis with the MEME suite. RNA secondary structure analysis was performed using the online RNA secondary structure prediction tool Context Fold (102).

### 3.3.6. Genome analysis.

The sequence of the yeast genome, its ORFs (SacCer3) and its 5’ and 3’ UTRs (103) were downloaded from the Saccharomyces Genome Database (SGD; (104)). Motifs were located within the yeast genome using the EMBOSS tool fuzznuc (105). Genome wide transcription rate data, derived from array analysis of metabolically pulse-labeled RNA, are described in (106). Overlapping motifs were first merged and then mapped to
binned yeast genes using BedOps (107). Statistical analysis and data plotting were performed using R 3.0.1(108) and GraphPad (109).

3.3.7. **Circular dichroism spectroscopy and analysis.**

For Spt4, Spt5\textsubscript{284-839} and Spt4/5\textsubscript{284-839} analysis, far-UV spectra were measured on a CD spectrophotometer (AVIV 60DS, Lakewood, NJ) using a 200 \(\mu\)m path length quartz cuvette. Data were collected every 1 nm with a 1 nm bandwidth in the 180–320 nm wavelength region using an integration time of 8 s per step. The far-UV spectra represent the average of 15 scans for Spt5\textsubscript{284-839} and spt45 and 22 scans for Spt4. The data sets were collected for three different sample preparations (~20 \(\mu\)M) each for Spt4, Spt5, and Spt4/5. CD spectra were measured at room temperature.

For analysis of Spt4/5\textsubscript{284-839} with and without RNA, far-UV spectra were measured at 20 \(^\circ\)C on a CD spectrophotometer (Jasco J-720) using a 200 \(\mu\)m path length quartz cuvette (Hellma). Protein samples were diluted to 0.1 mg ml\(^{-1}\) in 20 mM sodium phosphate pH 7.4 and 100 mM KF and the data collected every 1 nm with a 1 nm bandwidth between 195 and 260 nm using an integration time of 5 s per step. A 5’ 56-FAM labelled RNA oligonucleotide (IDT) with the sequence UGGCUCGCAAUAACAAAAACAAAC was added to the protein at a 1:1 molar ratio and data collected as above. The spectra recorded represent the average of three accumulations and are buffer-baseline corrected.

3.3.8. **Microscale thermophoresis.**

MST experiments were performed on a Monolith NT.115 system (NanoTemper Technologies) using 95% LED and 40% IR-laser power. Laser on and off times were set at 30 s and 5 s respectively. A 12-point twofold dilution series was prepared for unlabelled protein in 50 mM Hepes pH 7.4, 0.15 M KCl, 0.01% Tween-20 and 0.005% RNase inhibitor with 5’ 56-FAM labelled RNA oligonucleotides (IDT) at a final
concentration of 50 or 100 nM. Measurements were performed in standard treated capillaries (NanoTemper Technologies) and the data from three replicate measurements were combined and analysed using the implemented fitting software NT Analysis (NanoTemper technologies). Binding isotherms were fitted to a simple 1:1 binding isotherm.

3.3.9. Structure analysis.
Electrostatic surfaces were calculated with APBS (110) and all figures were generated with PyMOL (111).

3.4. Results

To address RNA binding properties of Spt4/5, we first tested the ability of a recombinant yeast Spt4/5 complex containing the Spt5 NGN and all five KOW domains (Spt4/5284-839) (Figure 3.1A) to bind double stranded DNA (dsDNA), single stranded DNA (ssDNA) and single stranded RNA (ssRNA) pentaprobes using electrophoretic mobility shift assays (EMSAs). Pentaprobes are overlapping 100-nt oligonucleotides encompassing all possible 5-nt sequences (100, 112). As shown in Figure 3.1A, Spt4/5284-839 was strongly selective for ssRNA, which is evident from the protein concentration dependent shift of the ssRNA probe. No binding to the dsDNA probe was observed and only a small shift of the ssDNA probe could be seen at the highest protein concentrations. These data show that the Spt4/5 complex binds selectively to ssRNA.
**Figure 3.1. Spt4/5 binds to RNA containing AA repeats**

A. Increasing concentrations of protein were incubated with $^{32}$P labelled dsDNA, ssDNA and RNA pentaprobe and electrophoresed on 6% polyacrylamide gels. The dotted line distinguishes where lanes have been cut between the probe alone and other samples for each gel. B. The sequence logo derived from MEME analysis of 2000 unique sequences obtained from a SELEX experiment of Spt4/5. The motif encompasses 86% of the input sequences. C. Gel shift showing sequence specificity of Spt4/5. RNA probes 24 nt in length were electrophoresed in the presence of increasing concentrations of Spt4/5. The sequence of 5K1 is a representative 24 nt sequence obtained from SELEX; it harbours 5 AA repeats separated by a single nucleotide. 5K1_mut is similar to 5K1 with the AA repeats mutated to non-A nucleotides. 5K1_GG comprises the same sequence as 5K1 with the AA repeats replaced by GG. D. Microscale thermophoresis (MST) data measuring the change in thermophoresis of fluorescently labelled RNA sequences in the presence of increasing concentrations of Spt4/5. Typical datasets and a fit of the
5K1 data to a simple 1:1 binding isotherm are shown with a derived $K_d$ of 1.85 ± 0.2 μM. No significant change in thermophoresis was observed when Spt4/5 was incubated with 5K1_mut or 5K1_GG. E. The interaction of fluorescently labelled RNA sequences containing different numbers of AA repeats with increasing concentrations of Spt4/5 as determined by MST. Symbols show actual data points while the curves represent a fit of the binding data giving $K_d$ estimations. Each data point is the average of three replicate measurements.

3.4.2. Spt4/5 binds RNA with sequence specificity.

To determine whether Spt4/5 recognizes ssRNA with any sequence or structural specificity, we carried out systematic evolution of ligands by exponential enrichment (SELEX) (113, 114). We found that the most significantly enriched motif (5_AA) is 14 nt long and consists of AANAANAAANAAANAA, where N denotes any nucleotide (Figure 3.1B and Supplementary figure 10.2. See Chapter 4 for further details on the SELEX data). Structural analyses of sequences containing this motif revealed a distinct lack of secondary structure (data not shown).

To determine whether the repetitive AA motif was selected for its ability to bind Spt4/5 or as a result of PCR or other bias, RNA EMSA (REMSA) and microscale thermophoresis (MST) was carried out using ssRNA probes with and without the AA repeats (Figure 3.1C and 3.1D). Strong selectivity for the AA repeat sequence is observed, with an affinity of $K_d$ 1.85 ± 0.2 μM, while there was no measurable binding to sequences without the AA repeats. Thus, the specificity of Spt4/5 for ssRNA is sequence based.

The repetitive nature of the 5_AA motif means that shorter sequences, such as AANAA (2_AA), were also clearly enriched; however, the longest enriched sequence was clearly 5_AA (Supplementary figure 10.2C). MST measurements of ssRNA probes containing fewer AA repeats showed that Spt4/5 is able to bind to a short sequence containing a
single AA (1_AA) with a K_d of 26 ± 2 μM; however, each additional AA repeat approximately doubles the affinity (Figure 3.1E). These data suggest that the motif is not strictly defined as 5_AA but is the culmination of an additive affinity for AA repeats.

### 3.4.3. AAN motifs are enriched in transcripts

To address the biological significance of the interaction between Spt4/5 and AA motifs in vivo, we mapped the locations of these motifs in the open reading frames of the yeast genome. We observed a clear and statistically significant enrichment (chi-square test, p<2.2e-16) of sequences containing at least the AANAA motif or longer (≥2_AA) on the transcribed strand relative to the expected distribution if the motifs were randomly distributed across the genome (Figure 3.2). Furthermore, ≥2_TT motifs (i.e., ≥2_AA motifs on the opposite strand) were observed significantly less frequently than expected (chi-square test, p<2.2e-16), suggesting that the enrichment of ≥2_AA repeats does not merely reflect a preference for A/T rich regions (Figure 3.2). Therefore, AA motifs are specifically enriched within ORFs.

Because the above observations may reflect the constraints of protein sequence on the nucleotide sequence of genes, we focused on untranslated regions of genes (UTRs), which are free of coding constraints (Figure 3.2). We observed that both AA and TT motifs are significantly enriched over the expected value at both 5’ and 3’ UTRs (chi-square test, p<0.0001 in both cases); however, in 5’ UTRs, ≥2_AA motifs are significantly more abundant than ≥2_TT motifs. Thus, AA motifs are also preferentially enriched in 5’ UTRs of mRNA.
We next asked if the presence of ≥2_AA motifs in UTRs is associated with altered rates of transcription. We focused on the subset of UTRs that have either ≥2_AA or ≥2_TT motifs (but not both), as well as UTRs that lack both types of motif. We observed that genes with ≥2_AA motifs in their 5’ UTRs are transcribed ~27% more frequently than those with ≥2_TT motifs, and ~14% more frequently than genes lacking both motifs (e.g. 17.11 transcripts per cell cycle/15.00 transcripts per cell cycle; Supplementary table 10.1). In contrast, the transcription rates of genes with either the ≥2_AA or ≥2_TT motif in their 3’ UTRs, are indistinguishable from each other. Thus, these data are consistent with the possibility that recognition of ≥2_AA motifs in 5’ UTRs of mRNAs by Spt4/5 leads to an elevated rate of transcription.
UTRs with $\geq 2$ AA or $\geq 2$ TT motifs are longer than those that lack them (Supplementary table 10.1), raising the possibility that our observations could be explained by a systematic effect of UTR length on transcription rate. However, we found that gene transcription rate and UTR length are uncorrelated (Pearson’s $r=0.003$ in the case of 5’ UTRs and 0.012 for 3’ UTRs). Overall, our analysis demonstrates that the increased frequency of $\geq 2$ AA motifs in 5’ UTRs is clearly associated with elevated transcription rates.

### 3.4.4. The Spt4/5 NGN domain is sufficient for RNA binding.

To determine the smallest region of Spt4/5 that can bind ssRNA, we carried out REMSAs using a sequence containing the 5_AA motif as the target probe and a series of Spt4/5 complexes encompassing (i) Spt4 with just the NGN domain of Spt5 (Spt4/5_{284-375}) or (ii) Spt4 with the NGN domain of Spt5 plus one, two or five KOW domains (Spt4/5_{284-419}, Spt4/5_{284-567}, Spt4/5_{284-839}, respectively; Figure 3.3). Surprisingly, the complex composed only of Spt4 and the isolated NGN domain (Spt4/5_{284-375}) bound RNA avidly, indicating that the KOW domains are not required for RNA binding. Additionally, MST experiments do not show any significant increase in binding affinity for Spt4/5 complexes when one or more KOW domains are included in the protein (Supplementary figure 10.3A), and no RNA binding was observed for KOW domains alone or in tandem (Supplementary figure 10.3B). These results indicate that KOW domains have no significant affinity for RNA containing multiple AA repeats, contradicting the prevailing idea that the KOW domains are responsible for the nucleic acid binding capabilities of the Spt4/5 complex.
It was observed that the smaller protein complexes (Spt4/5\textsubscript{284-375}, Spt4/5\textsubscript{284-419}) had a lower mobility than the larger protein complexes (Spt4/5\textsubscript{284-567}, Spt4/5\textsubscript{284-839}; Figure 3.3). It is possible that this phenomenon is due to the formation of higher order protein:RNA complexes as a result of a non 1:1 binding stoichiometry. The use of isothermal titration calorimetry would be useful for determining the stoichiometry of the Spt4/5:RNA complexes and any differences in stoichiometry between the different protein complexes.

Addition of Spt4 or Spt5NGN alone causes the disappearance of the unbound probe, indicating that each possesses a degree RNA binding activity (Figure 3.3A). However, there is no corresponding appearance of a stable shifted species, suggesting that both have a reduced specificity or a greater rate of dissociation than the Spt4/5 hetero-dimer.

**Figure 3.3.** Spt4 together with the Spt5 NGN domain is sufficient for RNA binding

0.5nM 5K1 RNA was electrophoresed in the presence of increasing concentrations of **A.** Spt4 alone, Spt5\textsubscript{284-375} alone, and the Spt4/5\textsubscript{284-375} hetero-dimer containing only the NusG homology region and **B.** Spt4/5 complexes containing the NusG homology region plus one (Spt4/5\textsubscript{284-419}), two (Spt4/5\textsubscript{284-567}), or five (Spt4/5\textsubscript{284-839}) KOW domains.
3.4.5. Spt4/5 hetero-dimer formation and RNA binding are associated with conformational change.

Since neither Spt4 nor Spt5 bound RNA strongly in isolation, we considered the possibility that a conformational change of one or both subunits is required for high-affinity binding. We tested this hypothesis by comparing far-UV circular dichroism (CD) spectra for Spt4, Spt5_{284-839} and Spt4/5_{284-839} (Figure 3.4A). Spectra characteristic of a structured protein (containing a mixture of α-helices and β-strands) were obtained for the individual Spt4 and Spt5_{284-839} proteins, as expected from the known structures of the two proteins (28). Interestingly, the spectrum of the Spt4/5_{284-839} complex was distinct from an appropriately weighted average of the spectra of Spt4 and Spt5_{284-839}. Thus, the formation of the Spt4/5 hetero-dimer is associated with a conformational change in one or both subunits.

Figure 3.4. Structural insight into the RNA binding mechanism of Spt4/5
A. CD spectra of Spt4, Spt5_{284-839} Spt4/5_{284-839} compared to the theoretical sum of Spt4 + Spt5_{284-839} shows a conformational change in one or both subunits upon formation of the hetero-dimer. B. Spt5 displays the typical \( \beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4 \) topology of an RRM, with Spt4 contributing an extra two \( \beta \)-strands to the \( \beta \)-sheet surface (PDB ENTRY 2EXU). The three canonical aromatic residues (green sticks) are shown on the surface of the \( \beta \)-sheet; however, access to these residues is blocked by an additional C-terminal \( \alpha \)-helix. The N and C termini of Spt5 are shown. Spt5 is coloured purple, and Spt4 in blue. C. Analysis of the structural B-factors shows the C-terminal \( \alpha \)-helix of Spt5 displays some structural flexibility. D. Raw CD spectra of RNA alone, and Spt4/5_{284-839} with and without RNA shows the conformation of Spt4/5 changes upon the binding of RNA. Protein was incubated with RNA in a 1:1 molar ratio. The molar concentrations of both the protein and RNA are equal across the three spectra.

The NGN domain of Spt5 can be considered to be a variant RNA Recognition Motif (RRM), one of the most common RNA binding domains. Interestingly, in archaea, human and yeast Spt5, the \( \beta \)-sheet of the RRM is extended by the hetero-dimerisation interface with Spt4 (28) Figure 3.4B). Such an arrangement of a hetero-dimeric RRM has never before been reported. Thus, the conformational change in either Spt4 or Spt5 upon formation of the hetero-dimer is likely mediated by the arrangement of this novel inter-protein RRM and is consistent with our observations that both Spt4 and Spt5 are required for the interaction with RNA.

The \( \beta \)-sheet of the yeast Spt4/5 RRM lacks the canonical sequences, although it does contain three aromatic residues (F326 and Y336, Spt5 and Y74, Spt4) that typically underpin the interaction with RNA in canonical RRMs. Of these aromatics, only Y336 and Y74 are conserved, however the human and archaeal Spt5 contain a third aromatic residue elsewhere on the \( \beta_3 \) strand (Supplementary figure 11.1). The presence of the aromatics is suggestive of a canonical RRM-like RNA interaction; however, this surface is predominantly blocked by an additional \( \alpha \)-helix at the C-terminus of Spt5 (Figure 3.4B), precluding a typical RRM-RNA interaction. This helix displays some flexibility (Figure 3.4C), and therefore it is possible that the helix could be displaced upon RNA
binding. Such a mechanism has been shown in the case of the RRM s of the human U1A spliceosomal protein (115, 116).

To determine whether the structure of Spt4/5 is altered upon RNA binding, we measured CD spectra for Spt4/5 with and without RNA (Figure 3.4D). The shape of the CD spectrum of Spt4/5\textsubscript{284-375} changed significantly in the presence of RNA. These data are consistent with the idea that the binding of RNA to Spt4/5 induces a change in the structure of the hetero-dimer, perhaps exposing the β-sheet of the Spt4/5 RRM to facilitate RNA binding.

3.5. Discussion

Our \textit{in vitro} data show that yeast Spt4/5 is a sequence specific RNA binding protein that recognizes multiple AA repeat sequences with an affinity that is consistent with the values of many other sequence specific RNA binding proteins (117). We were surprised to discover that Spt4/5 binds RNA in a sequence specific manner, given that the distribution of Spt4/5 on genes closely mirrors that of RNAPII (118). We further observed that AA motifs are strongly enriched at 5’ UTRs, and are associated with elevated gene transcription frequencies. Together, these data support the idea that recognition by Spt4/5 of AA motifs in the 5’ UTR of an mRNA increases production of the remainder of the transcript. Intriguingly, a recent study demonstrated that Spt4 is required for normal transcription of genes containing trinucleotide repeats and that genes with (AAA)\textsubscript{N} repeats showed the strongest dependence on Spt4 (119). Thus, the RNA binding activity of Spt4/5 may facilitate expression of genes with A-rich sequence motifs \textit{in vivo}.  

52
We show that together, Spt4 and Spt5 make significant RNA contacts. This raises the question of whether the specific RNA binding activity we observed for yeast Spt4/5 is conserved in NusG, which lacks an Spt4 homolog. NusG has been shown to interact with both DNA and RNA (64), and although we cannot rule out the possibility that Spt4/5 may bind DNA in the context of transcription elongation complexes, we did not observe significant DNA binding activity for yeast Spt4/5, suggesting that its primary nucleic acid binding activity is directed toward RNA.

Previously, the KOW domains of NusG/Spt5 have been proposed to mediate RNA binding (58, 59, 64); however, we show here that the NGN domain alone is sufficient to bind RNA and that the isolated KOW domains were unable to interact with RNA. Thus, although the KOW domains are important for Spt5 function in yeast (96), they are unlikely to contribute to the specific RNA binding activity of Spt4/5. We propose instead that the KOW domains mediate protein-protein interactions. Consistent with this hypothesis, fragments of Spt5 containing KOW domains but lacking the NGN or CTR domains are capable of binding polymerase in pulldown assays (96, 120, 121), and the KOW 4-5 domains have been shown to make extensive contacts with the dissociable subunit of RNAP, Rpb4/7 (122). It remains to be determined if the KOW domains are involved in Spt5-RNAPII binding in vivo, or if they have other binding partners besides RNAPII.

We have shown that Spt4/5 undergoes a conformational change upon RNA binding potentially through the displacement of the C-terminal α-helix of Spt5 exposing a canonical RRM RNA binding surface. However, this helix contains some key residues which are involved in the RNAPII-Spt5 interface. Therefore, the shifting of the helix to accommodate RNA binding would significantly change the way Spt4/5 interacts with
RNAP. This change could provide a mechanism whereby RNA binding by Spt4/5 triggers the allosteric modulation of RNAPII activity, and also raises the question of whether Spt4/5 binds RNAPII and RNA simultaneously.

Our data suggest two possible mechanisms based on the RNA binding properties of Spt4/5. Firstly, RNA binding may promote recruitment of Spt4/5 to RNAP to stabilize the transcription elongation complex, as several prior in vitro studies have demonstrated increased association of Spt4/5 with elongation complexes that contain RNA transcripts long enough to protrude beyond the polymerase (61, 62). Alternatively, because Spt4/5 also associates with pre-mRNA processing and regulatory factors (9, 40, 51, 55, 94), RNA binding by Spt4/5 may facilitate pre-mRNA processing by enhancing recruitment of pre-mRNAs to their processing machineries. Such a function would explain observations of RNA processing defects in spt4 and spt5 mutants. Further experiments to elucidate the cross-talk between RNA-bound Sp4/5 and the rest of the TEC will allow these possible mechanisms to be distinguished.

3.6. Author contributions

3.7. Acknowledgments

We would like to thank the sequencing facility at UCLA run by Prof. Matteo Pellegrini for running the deep sequencing.
4. The identification of a specific RNA binding motif for Spt4/5

Chapter 3 described the elucidation of the RNA binding specificity of Spt4/5 using REMSA, SELEX and MST. The SELEX experiment outlined in the previous chapter only reports the results for the Spt4/5 construct containing 5 KOW domains (Spt4/5ubq284-839); however, the SELEX experiment carried out encompassed each multi-KOW Spt4/5 hetero-dimer plus a positive and negative control. This chapter will provide a comprehensive overview of the SELEX experiment through a more detailed analysis of the results.

4.1. Introduction

In order to determine whether Spt4/5 possesses any sequence or structural specificity for RNA, a Systematic Evolution of Ligands by Exponential Enrichment (SELEX) experiment was conducted. SELEX is an *in vitro* method developed by Tuerk and Gold in 1990, for rapidly identifying the binding sequence of a target molecule from a pool of random sequences (113). An advantage of SELEX over other methods that characterise protein-nucleic acid interactions is the simultaneous screening of a large pool of nucleic acid sequences that cover every possible combination of nucleotides; the number of individual oligonucleotides that can be simultaneously tested by this method is much larger than can be directly tested *in vivo* (113). Furthermore, SELEX requires no prior knowledge of the target RNA sequence to which the protein binds.

In this study, SELEX was used to find the specific RNA sequence that binds to Spt4/5. The SELEX method is based on the 3 main criteria for evolution: variation, selection
and replication (114). The variation is obtained via an oligonucleotide, in the case of this study, RNA, that contains a region of set length with a completely random sequence. The random region is flanked on either side by constant regions, whose sequence is known and are used as primer binding sites for PCR. One of the constant regions contains the recognition sequence for transcription by T7 RNA polymerase. Having a pool of this oligonucleotide essentially provides one with an RNA strand for every possible combination of nucleotides in the random region. This pool then becomes the basis for selection through selective binding of unique sequences to the molecule of interest, i.e. Spt4/5. The pool is incubated with the protein, and the RNA-bound protein is separated from the unbound RNA, usually through affinity chromatography. The resulting pool is now enriched for sequences that bind to the protein of interest. The selected RNA is reverse transcribed and then amplified by PCR, completing one round of SELEX. As the original oligonucleotide pool is highly complex it is expected to contain only a small fraction of sequences that are specific for the protein. Therefore, iterative cycles of SELEX are performed (typically 7-15) with each cycle reducing the complexity of the pool and enriching the target-binding candidates. Each round of enrichment results in the exponential increase of the strongest binding RNA sequence, until they dominate the population of sequences and the pool reaches a level of specificity for the protein where there can be no more enrichment (114). The final step of SELEX is to sequence the enriched library in preparation for bioinformatics analysis.

Traditionally, Sanger sequencing was employed to sequence the final enriched DNA pool. This would encompass cloning the PCR products into ‘easy-clone’ vectors, transforming strains of E. coli with the vector and extracting the plasmids from each individual colony. A common problem encountered when using this method is in
deciding how many clones to sequence. This decision results in a trade-off between a number that will provide a statistically significant result, and the time and/or cost involved in preparing that number of plasmids to be sequenced. The actual number of clones sequenced varies across published results but typically range between 10 – 50. The sequencing of only a sample population of the total SELEX pool runs the risk of a potential target sequence being potentially overlooked as a bona fide motif, as it may not have been represented by enough copies in the sequenced population. Due to the limitations in the viability of Sanger sequencing large numbers of sequences, next-generation sequencing (NGS), or deep-sequencing, of the final enriched pool is fast becoming the norm for SELEX sequence analysis.

For this study, three different multi-KOW domain Spt4/5ubq complexes, plus a positive and ubiquitin-only negative control were subjected to seven rounds of SELEX using a ssRNA oligonucleotide containing a random region of 25-nt.

4.2. Methods

4.2.1. Protein preparation

The multi-KOW domain Spt4/5 hetero-dimers were prepared as described in Chapter 2 (99). His-tagged ubiquitin was over-expressed and purified from the empty pHUE vector as described by Catanzariti, A. M., et al., 2004 (80). The vector containing the gene for GST-tagged ZRANB2-F12 (1-95) (123) was kindly provided by Prof. Joel Mackay (University of Sydney, Australia) and was prepared as described by O’Connell et al., 2012 (123).
4.2.2. Selection

The selection process is described in section Chapter 3 (section 3.3.4). Ubiquitin was immobilised in MagneHis\textsuperscript{TM} Ni-NTA particles (Promega), and ZRANB2-F12 was immobilised on MagneGST\textsuperscript{TM} particles (Promega).

4.2.3. Monitoring enrichment

REMSA was used to monitor the enrichment of each round of SELEX. Radio-labelled ssRNA probes were prepared by transcription of the initial dsDNA SELEX library and the PCR products from rounds 2, 4, 6 and 7 of each multi-KOW domain construct and the ubiquitin control, using the Ribomax\textsuperscript{TM} Large Scale RNA production system –T7 (Promega) in the presence of $[^\gamma-{}^{32}\text{P}]}$ UTP (PerkinElmer). The transcription reactions were DNase treated for 30 min at 37 °C to remove any dsDNA template. The resulting radio-labelled nucleic acid probes were incubated with 10 μM protein samples in gel shift buffer (10 mM MOPS pH 7.0, 50 mM KCl, 5 mM MgCl\textsubscript{2}, 10% glycerol and 0.03 mg ml\textsuperscript{-1} heparin) at 4 °C for 30 min. The binding reactions were electrophoresed on pre-equilibrated 6% native acrylamide/bisacrylamide (19:1) gels in 0.5X Tris-borate buffer at 200 V for 2 h at room temperature. Gels were visualized on a Phosphor screen with a Typhoon PhosphoImager\textsuperscript{TM} FLA 9000 (GE Healthcare).

4.2.4. Sanger sequencing

For the positive control, sequences recovered after the final round of SELEX (round 7) were cloned using the Zero Blunt\textsuperscript{®} TOPO\textsuperscript{®} PCR cloning kit (Invitrogen\textsuperscript{TM}) and used to transform \textit{E. coli} DH5α competent cells. 15 single colonies were selected and the plasmids extracted using a QIAGEN\textsuperscript{®} Plasmid Mini kit (QIAGEN) before being Sanger sequenced.

4.2.5. Deep Sequencing

Deep sequencing was carried out as described in Chapter 3 (section 3.3.5).
4.3. Results and Discussion

The transcribed ssRNA library was incubated with the Spt4/5ubq constructs or His-tagged ubiquitin (Ubq) as a negative control. As an additional positive control, to ensure the protocol was working as intended, the ssRNA pool was incubated with GST-ZRANB2-F12 (123). GST-ZRANB2-F12 (1-95) has previously been reported to select specifically for a double-AGGUAA motif in ssRNA using a similar 25-nt ssRNA library (124).

After incubating the ssRNA pool with the bead-coupled proteins and washing extensively with SELEX buffer, any bound RNA was eluted from the protein by heat denaturation. RNA was purified using phenol:chloroform extraction and ethanol precipitation. The purified RNA was reverse transcribed, PCR amplified and transcribed into ssRNA again for use as the input library for a second round of selection with fresh bead-coupled protein. To select for high affinity binders, the binding stringency was increased from rounds 3 - 7. Firstly, the ratio of RNA to protein was increased in the binding reaction through the rounds (Table 4.1). This was achieved by decreasing the amount of protein relative to the amount of RNA, which serves to increase the competition for highly selective sequences by decreasing the number of available binding sites. Although an attempt was made to keep the total amount of RNA added to the binding reactions consistent, in some cases the amount of RNA transcribed from the products of a prior round was a limiting factor. In these cases the maximum possible amount of RNA was used. The amounts of RNA used in the binding reactions were determined individually for each protein construct and every round depending on the yield of RNA obtained from the previous round and the transcription efficiency of the resulting cDNA pool.
Table 4.1. Stringency parameters for each round of SELEX

<table>
<thead>
<tr>
<th>SELEX Round</th>
<th>Heparin [mg ml⁻¹]</th>
<th>Protein (pmol)</th>
<th>Ubq RNA (pmol)</th>
<th>Sp₄/₅₂₈₄₋₄₁₉ RNA (pmol)</th>
<th>Sp₄/₅₂₈₄₋₅₆₇ RNA (pmol)</th>
<th>Sp₄/₅₂₈₄₋₈₃₉ RNA (pmol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>80</td>
<td>70</td>
<td>1:0.9</td>
<td>70</td>
<td>1:0.9</td>
</tr>
<tr>
<td>2</td>
<td>0.13</td>
<td>80</td>
<td>300</td>
<td>1:3.5</td>
<td>300</td>
<td>1:3.5</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>80</td>
<td>80</td>
<td>1:1</td>
<td>80</td>
<td>1:1</td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>40</td>
<td>80</td>
<td>1:2</td>
<td>80</td>
<td>1:2</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>20</td>
<td>80</td>
<td>1:4</td>
<td>27</td>
<td>1:1.5*</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
<td>10</td>
<td>80</td>
<td>1:8</td>
<td>80</td>
<td>1:8</td>
</tr>
<tr>
<td>7</td>
<td>1.6</td>
<td>5</td>
<td>40</td>
<td>1:8*</td>
<td>100</td>
<td>1:20</td>
</tr>
</tbody>
</table>

This table shows the amounts of RNA, protein and heparin utilised for each round of SELEX. Additionally the molar ratio (Mᵣ) of protein:RNA is shown.

*For some samples the Mᵣ of protein and RNA varies from round to round depending on the yield of RNA from the previous round.

A second strategy for increasing the stringency of the binding reactions was to increase concentrations of the competitor, heparin, in the binding reaction. Heparin is a mixture of oligosaccharides that is commonly used to block non-specific interactions between nucleic acid and proteins. Each sugar moiety of heparin contains a negatively charged sulphate group giving it a broad resemblance to the negatively charged sugar-phosphate backbone of nucleic acid. Increasing the amounts of competitor in the binding reaction serves to select for higher affinity binders. Table 4.1 lists the concentrations of heparin used for each round.

A total of seven selection rounds were completed. The enrichment of the SELEX library for each protein was monitored after every second round using REMSA (Figure 4.1). The enrichment of the library was defined by the complete protein dependent shift of the library for any given SELEX round. Using this method, enrichment of the multi-KOW Sp₄/₅ubq hetero-dimers after seven rounds of selection in comparison to the Ubq alone negative control was clearly evident.
**Figure 4.1. Monitoring the enrichment of the SELEX libraries**

RNA electromobility shift assays of the initial library (0) and round 2, 4, 6 and 7 for each multi-KOW domain Spt4/5 construct and ubiquitin only control. The probes were incubated with (+) or without (-) 10 μM protein and electrophoresed on a pre-equilibrated 6% TBE polyacrylamide gel. The dashed lines delineate each round.

Before analysis of the entire SELEX experiment, and to ensure the SELEX protocol was working as intended, sequences recovered after round 7 for the positive control GST-ZRANB2-F12 were cloned, transformed into *E. coli*, and 14 single colonies selected for plasmid extraction and Sanger sequencing. Interestingly, only four out of the 14 sequences contained a random region of 25-nt in length. The other 10 sequences contained random regions ranging from 41-89-nt in length. Changes in the length of the random region is a common side-effect of the SELEX process due to the nature of PCR amplification from a randomised ssDNA library template (125, 126). The shorter or longer sequences are often non-specific amplification products arising from the presence of secondary structures in the ssDNA templates disturbing the polymerisation reaction; or, from product-product or primer-product annealing and extension (125-128).
A sequence alignment, using the online sequence alignment tool Clustal Ω (129), of the total 14 sequences revealed 11 of these contained an identical 25-nt sequence, UGCUGUGAAUGACGCGUUUCCC (Figure 4.2A). This sequence did not contain the anticipated double GGU binding motif of ZRANB2-F12, nor did it encompass any part of the constant regions of the SELEX oligonucleotide. Exclusion of this 25-nt motif from the sequences resulted in oligonucleotides with random regions between 15 and 39-nt in length. Subsequent alignment of the residual sequences revealed that 13 oligonucleotides contained the expected double GGU motif (Figure 4.2B). This result confirmed that, regardless of the addition of 25-nt motif being incorporated into the random region of the SELEX oligonucleotide, the library was enriched for the anticipated binding specificity for GST-ZRANB2-F12. This result verified that the selection process was working.

For the multi-KOW Spt4/5ubq hetero-dimers and the ubiquitin control, NGS of the enriched sequences was employed to provide a more comprehensive overview of possible consensus sequences. This technique has been shown to be highly successful in determining the binding motifs for a multitude of RNA binding proteins (130, 131) and it is possible that even a single round of in vitro selection is sufficient to show a distinct, enriched consensus sequence (130, 132, 133).

To prepare sequences for NGS, the final products from the enriched SELEX library were PCR amplified to incorporate specific barcodes and sequencing adapters. The use of barcodes allows the parallel sequencing of multiple samples. The initial library, each round of SELEX for the multi-KOW Spt4/5ubq hetero-dimers, and round 7 of the Ubq control were sequenced. The sequences of the primers used to incorporate the unique barcode for each SELEX library are listed in Appendix 1.
Figure 4.2. Alignment of 14 sequences obtained from round 7 of GST-ZRANB2-F12 SELEX
A. Sequence alignment using Clustal Ω. The resulting aligned motif is shown underlined and in bold font. B. The motif found in A was removed from the sequences, and the remaining sequences re-aligned. Every sequence contained the expected double GGU motif, shown in red.

4.3.1. Sequence preparation

The sequences obtained from NGS were deconvoluted into their respective barcodes for analysis. As shown in Figure 4.3A, between 0.5-4.5 million sequences were obtained per barcode. The libraries were then filtered to remove any sequences present in multiple copies, as well as sequences which contained a longer or shorter than 25-nt random region, so only the unique sequences of the correct length remained. The
average number of unique 25-nt sequences obtained per barcode was between 5-15,000 (Figure 4.3B). An overall enrichment profile for the entire SELEX experiment was obtained by looking at the proportion of unique sequences in the total pool over each round of SELEX (Figure 4.3C). As expected, the number of unique sequences in the multi-KOW domain Spt4/5ubq libraries decreased over the rounds indicating enrichment of the SELEX pool for a specific RNA binding motif. Comparatively, the ubq-only control showed a very low proportion of unique sequences in the final round of SELEX. This confirmed that the ubiquitin tag was not binding RNA in a specific manner and therefore not contributing to the enrichment of the multi-KOW domain SELEX libraries. Interestingly, each multi-KOW domain library showed a different rate of enrichment over the rounds of SELEX; however, the total enrichment converged at the final round.

![Figure 4.3. Next generation sequencing of Spt4/5ubq SELEX](image)

**Figure 4.3.** Next generation sequencing of Spt4/5ubq SELEX  
A. The total number of sequences obtained per barcode. B. The number of unique sequences obtained per barcode. C. The enrichment profile of the SELEX library over the 7 rounds of SELEX. The library becomes enriched as the proportion of unique sequences decreases.
4.3.2. MEME analysis

To analyse the sequences obtained and detect any enriched sequence motifs, the online program MEME (Multiple EM for Motif Elicitation) was used (134). MEME is a tool for discovering motifs in a group of related sequences. Suppling a list of sequences to the program enables the user to determine the most statistically significant (‘best’) motifs in the list of provided sequences. The statistical relevance is given by an expectation value (E-value), which is based on the length of the discovered motif, the number of sequences contributing to the construction of the motif, and the number of sequences in the input. The E-value represents the probability of the observed number of occurrences of the motif versus the probability of it occurring in a set of random sequences (with each position being treated independently with letters chosen according to background frequencies). The lower the E-value, the higher the statistical significance; however, no threshold for significance is given, leaving it up to the user to decide on the significance of a motif. A typical value for well-defined motifs is 1 x 10^-4 or less (135).

The output of MEME is in the form of sequence logos containing stacks of letters at each position of the motif. The height of each letter depicted in the motif represents the proportion of that nucleotide found for that position. Additionally, the total height of all letters in a given position provides a measure of the information content at that position. For example, if only a single nucleotide is ever found at a given position the height of that nucleotide will be 2 bits. Conversely, if all four nucleotides were represented equally, that position provides no specificity and the total height displayed would be 0. MEME is primarily used for DNA or protein sequences; therefore, all uracil nucleotides in the sequences are represented as thymine in the MEME outputs.
As the MEME web server has a maximum capacity of 2000 input sequences, the first 2000 sequences in the list of unique sequences obtained from each SELEX library was used for MEME analysis. Figure 4.4 shows the top three best motifs, as calculated by MEME, generated from 2000 sequences obtained after seven rounds of selection for the ubiquitin only negative control. Similar to the GST-ZRANB2-F12 Sanger sequencing, the 25-nt motif UGCUGUGAAUGACAGCCGUUUCCC was detected. The motif was found in 4 out of the 2000 input sequences with a calculated E-value of 1.7 x 10^{-9}. Although the E-value is representative of a well-defined motif, the sequence was only present in 0.2% of the input sequences. Therefore, this motif does not represent a significant proportion of the SELEX pool and is not a good candidate for an enriched motif. Likewise, the top and bottom motif shown in Figure 4.4 were present in less than 2% of the input sequences indicating that these sequences are not strong motifs. The lack of a discernable motif for the ubq dataset is consistent with the conclusion that the ubq tag does not interact with RNA, and therefore would not contribute to any RNA binding specificity shown by the Spt4/5ubq constructs. This is in agreement with results observed in the REMSA (Figure 4.1) and as such, Ubq can serve as a robust negative control in this experiment.
Figure 4.4. MEME analysis of the ubiquitin SELEX library after round 7
2000 input sequences were analysed and the top three motifs shown with their corresponding E-value. The height of each letter depicted represents the proportion of that nucleotide found in each position. All motifs were generated using the MEME online server.

Figure 4.5 shows the MEME motifs generated from 2000 sequences obtained after 7 rounds of selection using each multi-KOW Spt4/5ubq hetero-dimer. As the E-value did not seem to be an accurate representation for the enrichment of a motif in the ubiquitin SELEX pool, only the number of sites for sequence occurrence was considered as a measure of enrichment. For all multi-KOW constructs, the best motif found was an AA repeat sequence. The motif consists of multiple AA repeat units separated by a single nucleotide that doesn’t show any sequence preference. The AA repeat motif was seen in over 50% of the dataset for each Spt4/5ubq construct and was therefore likely to be a good candidate for an enriched motif. However, as this motif was also seen in the Ubq control (albeit in significantly less of the total sequence population) it is possible that the enrichment of this motif is the result of PCR bias.
The next best motif found was also in each multi-KOW construct and contained the sequence AGAUC. This motif was found in an average of 130 sites for all three Spt4/5ubq constructs, representing less than 7.5% of the input sequences for each construct. Although 7.5% seems to be a small population of the total pool, the motif was not seen in the ubq control and could therefore represent a potential binding motif.

To determine any other possible consensus motifs, MEME analysis was performed on 2000 randomly selected sequences from each round of SELEX for each of the multi-KOW Spt4/5ubq constructs. Additionally, as the AA repeat sequences dominated the libraries for each protein construct, MEME analysis was performed on 2000 sequences from each round not containing the AA repeat motif. One other possible consensus motif was found in round 5 for Spt4/5ubq_{284-839}. The motif consisted of 4-nt, CUCG, and was present in 420 sites representing 21% of the input sequences (Supplementary Figure 10.2).
Figure 4.5. MEME analysis of the SELEX library for each Spt4/Subq construct after round 7
The top three motifs are shown from the analysis of 2000 input sequences with the corresponding E-value. The height of each letter depicted represents the proportion of that nucleotide found in each position. All motifs were generated using the MEME online server (134, 136).
MEME’s statistics are designed for detecting only barely enriched motifs and therefore provides a good basis for the elucidation of potential enrichment motifs. However, as MEME is limited by its input capacity of 2000 sequences, it was necessary to use a different approach in analysing enriched motifs in the total deep sequencing dataset.

4.3.3. Total dataset analysis

The potential motifs uncovered in the MEME analysis were used as the input sequences for Perl script programs designed to count the prevalence of a motif in large datasets. The programs were designed and provided by Mr. Michael Webster (University of Sydney, Australia). By counting the prevalence of each of the three MEME motifs (AANAA, where N is any nucleotide, (2_AA), AGAUC and CUCG) in each SELEX library, it was possible to see the enrichment profile of each motif throughout the selection, for each protein construct (Figure 4.6). The use of 2-AA as a search motif, served to identify all sequences containing any number of AA repeats.

At the final round of enrichment, the AANAA motif is present in over 70% of the total pool for all three Spt4/5ubq constructs. Such a high level of occupancy suggests a highly specific binding motif. The prevalence of the motif AGAUC is less than 15% of the total pool for any of the protein constructs; therefore it is unlikely to represent a bona fide motif. The motif CUCG is represented in 45% of the total SELEX library in round 7 for Spt4/5ubq_{284-839}, but in less than 15% for the other two multi-KOW domain constructs. This suggests that CUCG motif is binding to Spt4/5 somewhere between the 2^{nd} - 5^{th} KOW domain.
Figure 4.6. Total library analysis of the top three sequence motifs for each multi-KOW Spt4/5ubq construct

The enrichment of AANAA (where N is any nucleotide), CUCG and AGAUC is shown as a proportion of the total deep sequenced library for each round. Motifs containing AA repeats are clearly the most enriched sequences for each construct, and the ubiquitin control shows no enrichment for any motif.

4.3.4. Analysis of the AANAA motif

For Spt4/5ubq_284-839, the AANAA consensus sequence reached its maximum occupancy within the total pool at round 3, and stayed consistent until round 7 (Figure 4.6). For Spt4/5ubq_284-567, the AA NAA motif reached its maximum occupancy by round 5, whereas the occupancy of the same motif for Spt4/5ubq_284-419 does not reach a plateau by round 7. Thus, whether the occupancy of this motif for the Spt4/5ubq_284-419 library reflects it’s maximum enrichment level is unknown. The rate of enrichment of the AANAA motif seems to be a direct result of the number of KOW domains in the protein; i.e. the more KOW domains present, the faster the motif is enriched. This
would suggest that the KOW domains do play a role in the selection of the motif; however, as shown in Chapter 3, the KOW domains are not directly responsible or even involved in the RNA binding capabilities of Spt4/5. Why the multi-KOW domain constructs enrich at different rates for the AA₅AA is unknown, as there is no significant difference in the affinity of each construct for this RNA binding motif (Supplementary figure 10.3A). Therefore, the different enrichment rates seen are most likely to be an anomaly of the selection process.

As the best MEME motif contained a longer AA repeating unit than just AA₅AA, the enrichment of longer AA repeat motifs was investigated. The prevalence of AA₅AA (2_AA) was compared to the prevalence of AA₅AA₅AA (3_AA), and so forth, up to AA₅AA₅AA₅AA₅AA₅AA (7_AA) for each multi-KOW Spt4/5ubq construct across each round of SELEX (Figure 4.7). The enrichment of 6_AA and 7_AA is similar and falls below 20% at round 7 for all three protein constructs, suggesting that the motif containing 5_AA repeats was the longest motif selected for by Spt4/5ubq. The 2_AA motif appears at the same time and level as the 3_AA motif suggesting that two AA repeats alone are not sufficient for selection, and that the base unit of selection is at least three AA repeats. The 4_AA, and 5_AA motifs are less prevalent in the total pool at the same round and over the following rounds, which suggests that the extra AA repeats are not entirely necessary for binding.
It is possible that the longer AA repeat motifs may have been enriched as a consequence of the 3-AA selection, as all sequences containing four or five AA repeats also contain two or three of the 3-AA motif respectively. Therefore, to determine the specific enrichment of these longer motifs, the proportion of sequences containing only the motif of a defined length (i.e. not as a subset of a longer motif) was examined in all the sequences containing at least 2-AA (Figure 4.8). Interestingly, for all Spt4/5ubq constructs, the proportion of sequences containing exactly the 2-AA motif dominated the libraries at the beginning of selection and then decreased dramatically as the selection proceeded. Additionally, the proportion of sequences containing exactly the 3-AA motif was consistent across the seven rounds of selection for all Spt4/5ubq constructs.
Figure 4.8. Specific enrichment of the length of the AA repeat motif
For each multi-KOW Spt4/5ubq construct, the proportion of each AA repeat sequence containing a motif of the exact defined length is shown for each of the 7 rounds of selection.

These results suggests that neither the 2_AA or 3_AA motif was being enriched, and the high prevalence of both motifs in the total SELEX libraries was due to their being a subset of the longer 4_AA or 5_AA motifs. Both the 4_AA and 5_AA motif showed enrichment to a higher level than the 3_AA motif for the Spt4/5ubq284-567 and Spt4/5ubq284-839 protein construct. The specific enrichment of the longer motifs for the Spt4/5ubq284-419 construct is ambiguous, which could be due to the fact that for this protein seven rounds of selection were not sufficient to complete the enrichment process. Regardless, this analysis suggests that the 4_AA or 5_AA motif is likely to represent the true high affinity target.

4.3.5. Analysis of the CUCG motif
Total library analysis found that the CUCG motif was prevalent in 45% of the SELEX pool after seven rounds of selection for Spt4/5ubq284-839 (Figure 4.6). The fact that the
AANAA motif dominates the sequence pools makes it difficult to determine the significance of the CUCG motif. Therefore, sequences containing AANAA were removed and the remaining sequences analysed for the prevalence of CUCG (Figure 4.9A). This analysis showed that CUCG was enriched in 45% of sequences by round 7, irrespective of whether the AANAA was present. Furthermore, position analysis of the CUCG motif in the random region showed that the motif was prevalent at the 5’ region of the oligonucleotide, also regardless of the presence of the AANAA motif (Figure 4.9B). Together, these data classify CUCG as an independently selected motif.

![Figure 4.9. Analysis of the CUCG motif](image)

**A.** CUCG is enriched in 45% of the sequences by round 7 irrespective of the presence of the AANAA motif.  
**B.** The CUCG motif is found at the 5’ end of the 25-nt SELEX oligonucleotide’s random region, irrespective of the presence of the AANAA motif.

The CUCG motif has been found to bind to the zinc finger of the nucleocapsid protein of Maloney murine leukemia virus (MLV)(137). It is possible that the enrichment of the CUCG motif may be due to interactions with the zinc finger of Spt4, which supports the data presented in Chapter 3, suggesting that Spt4 may be capable of interacting with RNA (see section 3.4.4 and Fig 3.3A). If this was the case, the CUCG motif should be enriched in all Spt4/5uq libraries; however, even though CUCG was present in the NGS libraries for each Spt4/5ubq construct, the motif was only significantly enriched for
Therefore, it is more likely that the latter KOW domains are responsible for the interaction with this motif.

In summary, two potential consensus sequences were enriched during the SELEX of Spt4/5ubq: CUCG and the AANAA repeat sequence. PCR bias and enrichment of non-specific binding oligonucleotides during SELEX is often observed; therefore, validation of these results by other methods is necessary.

### 4.3.6. Validation of SELEX results

For this study, the SELEX results were validated using REMSA and MST. The results of this validation have already been described in Chapter 3, however for the sake of completeness I will provide a brief summary of those results here in the context of this chapter.

To validate the sequence specificity of Spt4/5 for the SELEX sequence motifs, REMSA was performed using an RNA probe containing the 5_AA repeat motif (5K1), the same probe containing GG repeats in place of the AA repeats (5K1_GG) and a probe containing CUCG and no AA repeats (5K1_mut) (Figure 3.1C). The REMSA showed clear binding of Spt4/5ubq to the AA repeat motif sequence, but not to the sequence containing only the CUCG motif or the GG repeat motif. Additionally, MST on the same probes showed an identical result (Figure 3.1D), and a binding affinity (Kd) of 1.8 ± 0.2 μM for the binding of Spt4/5ubq to the 5K1 probe.

As the MEME and total sequence analysis suggested the enrichment of longer AA repeat motifs, the affinity of Spt4/5 for RNA probes containing the 5_AA, 4_AA,
3_AA, 2-AA and also a 1-AA motif were measured (Figure 3.1E). Unfortunately the data for the 5_AA and 3_AA probes was not of sufficient quality to provide a reproducible or accurate Kd, and time constraints during this data collection meant that the measurements for these sequences could not be repeated. Therefore, only the data for the 4_AA, 2_AA and 1_AA probe are shown. Regardless, the MST results show that Spt4/5 binds preferentially to sequences containing a longer AA repeat motif.

The fact that no binding was observed to the CUCG motif using REMSA or MST suggests that this motif was an artefact of the selection procedure; however, the clear binding of the AA_NAA motif is representative of a true high affinity target. The genomic analysis described in Chapter 3 suggests the AA repeat motifs are enriched in the UTRs of transcripts, and that the presence of this motif in the 5’UTR results in an increased level of transcription. However, whether the AA_NAA binding motif is representative of the RNA binding specificity of Spt4/5 in vivo remains to be seen.

Unfortunately, time constraints did not allow the validation of the SELEX results in vivo, therefore, this remains an important future direction for this project. Techniques such as individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) are useful to determine the RNA binding specificity of RNA binding proteins in vivo (138, 139), and have been used to determine specific binding motifs in a variety of proteins (140-143). An experiment incorporating iCLIP could provide in vivo evidence of Spt4/5 binding to repetitive AA repeat motifs in the UTR of transcripts.

Chapter 3 discussed the observation that the core region of the Spt4/5, containing no KOW domains, is responsible for the RNA binding ability of the protein. This core region is highly conserved across Spt4/5 homologues in all domain of life, suggesting
that RNA interactions could be an important and conserved function of the protein.
Performing SELEX on other Spt4/5 homologs would determine whether the sequence
specificity observed for \textit{S. cerevisiae} Spt4/5 is species specific or conserved throughout
evolution and could potentially identify the A\textsubscript{N}A\textsubscript{N}AA motif as a conserved sequence
element.

\textbf{4.4. Conclusion}

This chapter, in combination with chapter 3, has shown that Spt4/5ubq can bind to
single stranded nucleic acids and can select repetitive AA motifs from a random RNA
library. The binding was validated using REMSA and MST showing that Spt4/5ubq
bound preferentially to sequences containing the A\textsubscript{N}A\textsubscript{N}AA motif. Furthermore, these
experiments have shown that the ubiquitin tag does not interfere with, or contribute to,
the RNA binding. Therefore, Spt4/5 does appear to be a sequence specific RNA
binding protein.

The A\textsubscript{N}A\textsubscript{N}AA motif is a previously uncharacterised RNA binding motif; however,
genomic analysis showed that A\textsubscript{N}A\textsubscript{N}AA motifs were enriched in the 5’ and 3’ UTRs of
yeast genes. Furthermore, the analysis suggested that genes containing these motifs are
transcribed more frequently compared to genes without the motif. Therefore, the
A\textsubscript{N}A\textsubscript{N}AA motif may represent a new regulatory sequence embedded in transcripts, the
binding of which may serve as a mechanism for the control of transcription elongation
by Spt4/5.

None of these experiments give any indication of the basis of the molecular interaction
between Spt4/5 and the A\textsubscript{N}A\textsubscript{N}AA motif. Therefore, to characterise this interaction we
sought to solve the macromolecular structure of the Spt4/5:RNA complex using X-ray
crystallography. Chapter 5 describes the techniques employed to determine the RNA binding interface of Spt4/5 with the AA$_N$AA motif in order to provide information on the molecular basis of the interaction.
5. Towards the Structure of the Spt4/5:RNA complex

Chapters 3 and 4 determined that Spt4/5 binds specifically to ssRNA probes containing a multiple AA repeat motif. This chapter describes approaches to determine the macromolecular structure of *S. cerevisiae* Spt4/5 in complex with RNA in order to ascertain the RNA binding interface of the protein. By determining the molecular mechanism by which Spt4/5 binds RNA, insight can be gained into how the RNA binding function of Spt4/5 contributes to mechanism by which the protein controls transcription elongation.

### 5.1. Introduction

Each step of RNA metabolism is regulated by RNA binding proteins (RBPs). These proteins influence the processing, transport, localisation, translation and stability of mRNAs and also non-coding RNA. Some RBPs associate with RNA from transcription right through to the RNA degradation, while others recognise and transiently bind to RNA. Structural biology has been an indispensible method in providing essential information for the understanding of RBP functions and their mechanisms of interactions with RNA. RBPs contain distinct structural elements, RNA binding domains (RBDs), which are responsible for the specific recruitment of an RNA target to the protein. The four main RBD families include RNA recognition motifs (RRMs), zinc-fingers, heterogeneous ribonucleoprotein particle (hnRNP) K homology (KH) domains and ds-RNA binding motifs (dsRBMs). Each domain is structurally distinct and recognises its cognate RNA in a specific way.
The most extensively studied RBD is the RRM as this is the most abundant RBD in the human genome, being present in about 0.5% - 1% of human genes (144, 145). An RRM adopts a conserved fold consisting of a four-stranded β-sheet packed against two α-helices (146). Typically, RNA binding occurs through three surface-exposed aromatic residues on the β-sheet surface that form stacking interactions with RNA bases (147). Additionally, several other basic residues on the same surface interact with the phosphate backbone (145) (Figure 5.1A). RRM s display great plasticity in their RNA binding mechanisms as they can also interact with RNA through the loop regions of the β-sheet, as well as the N and C terminal regions of the domain. These RRMs are often found in multiple copies in RBPs in order to increase their affinity and specificity for RNA.

The Zinc finger (ZnF) is classically about 30 amino acids long, adopting a ββα topology. There are several types of ZnF that are classified based on the amino acids interacting with a Zn²⁺ ion (e.g., CCHH, CCCH or CCCC), with CCHH being the most common type. The ZnF is a versatile domain being able to interact with proteins, DNA, ssRNA and dsRNA; and can be found alone, as a repeated domain or in combination with other types of RBDs. Zinc fingers use amino acids embedded in the protein loops and α-helices to interact with RNA in a highly diverse way, through binding to the backbone of an RNA double-helix (148), hydrogen bonding (149, 150) and aromatic base stacking interactions (124, 151) (Figure 5.1B). This domain is able to achieve sequence specificity through its ability to adopt different folds.
Figure 5.1 Structures of RNA binding motifs bound to their cognate RNA

A. The RRM of Human Fox-1 (green) binds to ssRNA (cyan) in a canonical way through three conserved aromatic residues (pink sticks) on the β-sheet surface. (PDB entry 2ERR (152))

B. The zinc finger of ZRANB2 (green) binds to ssRNA (cyan) through side-chain hydrogen bonds and the formation of a guanine-tryptophan-guanine “ladder” (PDB entry 3G9Y (124)). The cysteines (yellow sticks) coordinating the Zn$^{2+}$ atom (grey sphere), and the residues involved in making RNA contacts are shown as stick representation.

C. The Type I KH domain of the Nova protein (green) recognising an RNA stem loop (cyan). The secondary structure elements that form the RNA binding cleft (left), are shown as a surface representation binding to an RNA stem loop (right; PDB entry 1EC6 (153)).

D. The dsRBM of ADAR2 (green) is shown bound to dsRNA (orange phosphate backbone with green/blue bases; PDB entry 2L3C (154)).
KH domains are known for interacting with RNA or ssDNA in a variety of proteins involved in splicing, transcriptional regulation and translational control. All KH domains contain a GXXG loop, however the domain has two distinct types of folds: types I and II found in eukaryotes and prokaryotes, respectively. Type I adopts fold consisting of an antiparallel $\beta$-sheet packed against three $\alpha$-helices whereas the type II fold contains a characteristic $\beta$-sheet in which the central strand ($\beta$2) is parallel to $\beta$3 and antiparallel to $\beta$1 (155, 156). The KH domain usually accommodates four bases in a binding cleft formed by the $\alpha$1 helix, GXXG loop, $\alpha$2 helix and variable loop (Figure 5.1C)(156). Nucleic acid recognition occurs through the hydrophobic centre of the binding cleft with a variety of additional interactions contributing to the specificity of the interaction (156). Individually, these domains bind their targets with weak affinity; however achieve greater affinity and specificity through cooperative binding of multiple copies of KH domains within the protein (157-159).

dsRBMs recognise RNA shape rather than sequence. These domains typically fold into a three-stranded, anti-parallel $\beta$-sheet packed along two $\alpha$-helices (Figure 5.1D). These domains are often found as multiple copies within a single protein chain and bind the cognate RNA target in a cooperative fashion.

It is clear that there is great variability and plasticity in the way RBDs interact with their cognate RNA targets, and through the use of structural biology, we are slowly beginning to gain a greater understanding of these mechanisms. Of the known RBPs across all domains of life, there are relatively few structures of RBD:RNA complexes (Table 5.1), highlighting the crucial need for more research in this area in order to understand the way in which these RBPs control gene expression.
Table 5.1. Overview of the number of published protein:RNA structures for each class of RNA binding domain

<table>
<thead>
<tr>
<th>RNA binding domain</th>
<th>Number of proteins containing the domain*</th>
<th>Number of protein structures containing the domain†</th>
<th>Number of RNA:protein structures‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRM</td>
<td>2508</td>
<td>312</td>
<td>49</td>
</tr>
<tr>
<td>ZnF</td>
<td>45902^</td>
<td>1743§</td>
<td>68</td>
</tr>
<tr>
<td>KH</td>
<td>6406</td>
<td>306</td>
<td>194</td>
</tr>
<tr>
<td>dsRBD</td>
<td>1985</td>
<td>59</td>
<td>25</td>
</tr>
</tbody>
</table>

* Total number of proteins from the genomic database of SMART (160).
† Structures deposited in the Protein Data Bank (161).
‡ Includes all classes of ZnF that bind to DNA, RNA and protein.

Spt4/5 has long been classified as an RBP yet the specific mode of RNA interaction and the regions of the protein involved in RNA interaction have not yet been identified. Chapters 3 and 4 present evidence that Spt4/5 binds RNA with sequence specificity and that Spt4 together with the NGN domain of Spt5 is the major protein determinant for RNA binding. This chapter describes the approaches used to determine the specific mode of RNA binding by Spt4/5 through X-ray crystallography and bioinformatics analysis.

5.2. Methods

5.2.1. Protein preparation

Spt4/5ubq284-375 was prepared as described in Chapter 2 (99). In order to facilitate crystallisation, the ubiquitin tag was removed using the enzyme Usp-2cc. Usp-2cc was prepared as described by Catanzariti, A. M., et al., 2004 (80). The ubiquitin tag was cleaved from the purified Spt4/5ubq284-375 through the addition of Usp-2cc in a 10:1 molar ratio. The reaction was incubated at 4 °C overnight concurrent with dialysis into cleavage buffer (50 mM sodium phosphate, pH 7.4, 0.3 M KCl, 20 mM imidazole, 20
% glycerol) to remove high concentrations of imidazole. As both Usp-2cc and the ubiquitin tag contain a His-tag, the cleaved protein was purified from the enzyme, ubiquitin tag, and any uncleaved protein using reverse nickel affinity chromatography. Phosphate buffers are typically avoided for crystallisation as they readily form salt crystals with the divalent cations in many of the common crystallisation protocols. Thus the final step in the purification of Spt4/5\textsubscript{284-375} utilised size exclusion chromatography to perform a buffer exchange into a more suitable crystallisation buffer (50 mM Hepes, 0.3 M KCl, 20 % glycerol, pH 7.4). The resulting protein was concentrated to between 9-11 mg mL\textsuperscript{-1}, before being flash frozen in liquid nitrogen and storage at -80°C.

5.2.2. Initial crystallisation trials

Initial crystallisation trials of Spt4/5\textsubscript{284-375} were performed using commercial crystallisation screens (Index, PEG/Ion and Natrix (Hampton Research)) in a 96-well sitting drop vapour diffusion format using the Art Robbins Phoenix robot (Art Robbins Scientific). Reservoir solution was mixed with the protein sample (9-11 mg mL\textsuperscript{-1} Spt4/5\textsubscript{284-375} in 50 mM Hepes pH 7.4, 0.3 M KCl) in a 1:1, 1:2 and 2:1 (protein:reservoir) ratio yielding drops of 0.4 \( \mu \text{l} \) and 0.3 \( \mu \text{l} \), with equilibration against 80 \( \mu \text{l} \) of reservoir solution at 20 °C. Crystallisation leads were obtained from multiple conditions and optimised utilising the vapour diffusion method with the 24-well plate hanging drop format. Drop sizes of 2 \( \mu \text{l} \) or 3 \( \mu \text{l} \) containing protein to reservoir solution ratios of 1:1 and 1:2, respectively, were equilibrated against 1 mL of reservoir solution at 20 °C.

5.2.3. Crystallisation of Spt4/5\textsubscript{284-375} with RNA

Co-crystallisation of Spt4/5\textsubscript{284-375} (10 mg mL\textsuperscript{-1} in 50 mM Hepes, 0.3 M KCl, 20 % glycerol, pH 7.4) with 3\_AA RNA (UCAAUAACACAA; 30 mM in 50 mM Hepes, 0.3
M KCl, pH 7.4 and RNase inhibitor, Promega) was attempted by incubating the RNA and protein at a 10:1 molar ratio (RNA:protein) on ice for 30 min prior to crystallisation. Drops sizes of 2 µl containing protein/RNA solution to reservoir solution ratios of 1:1, were equilibrated against 1mL of reservoir solution at 20°C in the 24-well plate hanging drop format. Reservoir solutions contained 0.1 M Bis-Tris pH 5.5-8.0; 0.2 mM lithium sulfate, ammonium sulfate, or ammonium chloride; and polyethylene glycol (PEG) 3,350 (18-30%). Where RNA soaking was attempted, 0.2 µL of RNA (30 mM) was added to the drops 3 hours before cryoprotection and freezing.

Additionally, protein:RNA co-crystallisation trials were performed using commercial crystallisation screens (Index, PEG/Ion and Natrix (Hampton Research)) in 96 well sitting drop vapour diffusion format using the Art Robbins Phoenix robot (Art Robbins Scientific). 3_AA RNA was incubated with Spt4/5284-375 in a molar ratio of 1.5:1 (RNA:protein), to give a final protein concentration between 9-11 mg mL\(^{-1}\). The mixture was incubated for 30 minutes on ice prior to crystallisation.

### 5.2.4. Seeding
Where seeding was attempted, micro-seed solutions were prepared using the entire sample drop of interest. The crystals were crushed in 100 µl of the mother liquor by vortexing using teflon seed beads (Hampton Research) \((162, 163)\), and a dilution series of the seeds was prepared at 1/50, 1/100, 1/250 and 1/500 dilutions. The dilution series was screened through the addition of 0.2 µl of the micro-seed solution to drops containing suitable crystals.

### 5.2.5. Cryoprotection and crystal storage
Crystals were harvested from the sample drops using nylon loops mounted on a CrystalCap CopperTM base (Hampton Research), flash frozen by rapid immersion in
liquid nitrogen and stored in a 96 crystal sample cassette (CrystalPositioningSystems) under liquid nitrogen until required for data collection. For crystals where a cryoprotectant was used the crystal was placed into a 1 μl drop of the cryoprotectant solution immediately prior to being frozen in liquid nitrogen. Cryoprotectant solutions used were: Paratone oil (Hampton Research), ethylene glycol, or mother liquor containing 20 % glycerol, 30 % glycerol or 70% D-sorbitol. Diffraction quality crystals that gave high-resolution data sets were obtained from the crystals where Paratone was used as a cryoprotectant.

5.2.6. X-ray diffraction and data collection

X-ray diffraction data sets were collected on a micro focus beamline (MX2) at the Australian Synchrotron with an ADSC Quantum 315r detector. Where crystal annealing was trialled, the nitrogen stream was interrupted for 2 seconds prior to X-ray exposure.

5.2.7. Data processing

Data processing was performed using the XDS suite of software (164) and the data merged and scaled using AIMLESS (165, 166) as part of the CCP4 suite of software (167).

5.2.8. Molecular replacement

Phases were obtained by molecular replacement methods with PHASER (168) using the S. cerevisiae Spt4/5 hetero-dimer (PDB entry 2EXU; (28)) as the search model.

5.2.9. Model building

The Spt4/5	extsubscript{284-375} and Spt4/5	extsubscript{284-375}:RNA models were modelled using the graphical program COOT (169, 170). Refinement was performed with REFMAC (171, 172) and
PHENIX \((173)\) with non-crystallographic symmetry restraints employed at all stages. The refinement progress was monitored using \(R_{\text{free}} \) \((174)\) (5% of the reflections), COOT validation tools \((169, 170)\), as well as MOLPROBITY \((175)\). Figures were prepared using PyMOL \((111)\).

5.2.10. Structure analysis

Structural superpositions were performed using PyMOL or SSM superepose in COOT (based on the secondary structure homologies) and graphical representations of electrostatic surface potentials were generated using APBS \((110)\). Structures were analysed and figures generated using PyMOL.

5.3. Towards the X-ray structure of the Spt4/5:RNA complex

Understanding the intermolecular interactions required to facilitate efficient transcription elongation requires detailed structural knowledge of the component proteins and how they interact with each other, with the RNA transcript and with RNA polymerase. A crystallographic approach was employed as a key step towards this goal, to determine the precise mechanism of RNA binding by Spt4/5.

5.3.1. Crystallisation of Spt4/5\textsubscript{284-375}

The X-ray structure of the yeast Spt4/5 NGN domain has been solved previously (hereby referred to by its PDB ID, 2EXU); however, to achieve crystallisation, Spt4 and Spt5 were fused with an engineered 2-residue linker \((28)\). Thus, the crystallisation conditions for the native, unfused Spt4/5\textsubscript{284-375} hetero-dimer were likely to be different from the published protocol.

In order to facilitate crystallisation, the ubiquitin tag was cleaved from nickel-purified Spt4/5\textsubscript{ubq284-375} by digestion with the enzyme Usp-2cc. Analysis by SDS-PAGE
showed the complete cleavage of the ubiquitin tag and the effective purification of the cleaved product using reverse nickel affinity chromatography (Fig 5.2A and B). Cleaved Spt4/5\textsubscript{284-375} eluted from a size exclusion column as a single peak indicating the protein to be of suitable quality for crystallisation (Fig 5.2C).

Crystallisation conditions for Spt4/5\textsubscript{284-375} were obtained by testing a variety of commercially available screens and crystallisation events were observed in multiple conditions. These conditions produced a mixture of small thin plates and needle shaped crystals, and visualisation under UV light suggested that the crystals contained protein (Figure 5.3). Optimisation of these conditions was carried out in a 24-well hanging drop format. The resulting crystal conditions (Table 5.2) produced larger plate shaped crystals diffracting to 3.5 Å (Figure 5.4). In an attempt to improve the diffraction quality of these crystals, annealing was trialled through the interruption of the cyrostream during data collection; however, there was no visible improvement in the resolution of diffraction.
Figure 5.2. Purification and ubiquitin tag digestion of Spt4/5284-375

A. A 15% SDS PAGE gel analysis of the nickel affinity purified Spt4/5ubq284-375 (His) and the sample after digestion with the deubiquitinating enzyme Usp2cc (Digest). Using reverse nickel affinity chromatography, the cleaved protein was effectively purified (FT) from the enzyme and cleaved ubiquitin tag (elution). Relative positions of the proteins in each sample are indicated and sizes (kDa) of the molecular weight marker (M) are shown. The dashed line indicates lanes that have been omitted from the gel. B. A table showing the predicted MW of each protein represented in the gel. C. The size exclusion chromatography elution profile using a HiLoad 16/600 Superdex 75 pg (GE Healthcare) column of the nickel purified Spt4/5284-375 following ubiquitin tag digestion.

In many cases significant improvement in crystal morphology and size can be obtained using various seeding techniques (176). In the case of Spt4/5284-375, the small crystals obtained were utilised as micro-seeds to improve the quality of the protein crystals. Of the dilution series of micro-seeds added to the Spt4/5284-375 drops, the 1/50 and 1/250 micro-seed dilutions yielded a significant improvement in crystal quality, as assessed by their diffraction to better than 3 Å (Figure 5.5).
**Table 5.2. Optimised crystallisation conditions for Spt4/5\textsubscript{284-375}**

<table>
<thead>
<tr>
<th>Crystal condition</th>
<th>Protein: buffer ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 M sodium formate pH 7.0</td>
<td>1:1 and 1:2</td>
</tr>
<tr>
<td>0.2 M ammonium sulfate 0.1 M BIS-TRIS pH 5.5 25% PEG 3,350</td>
<td>1:1</td>
</tr>
<tr>
<td>0.2 M lithium sulfate monohydrate 0.1 M BIS-TRIS pH 5.5 25% PEG 3,350</td>
<td>1:1, 1:2 and 2:1</td>
</tr>
<tr>
<td>0.2 M ammonium acetate 0.1 M BIS-TRIS pH 5.5 25% PEG 3,350</td>
<td>1:1 and 1:2</td>
</tr>
<tr>
<td>sodium citrate tribasic dihydrate 20% PEG 3,350</td>
<td>1:1 and 1:2</td>
</tr>
</tbody>
</table>

**Figure 5.3.** Spt4/5\textsubscript{284-375} crystallisation events observed from the Hampton research Index screen
Figure 5.4. Optimisation of Spt4/5 crystallisation

A. Crystal conditions were optimised to produce plate shaped crystals (top) that fluoresce under UV light (bottom). B. The crystals from panel A diffracted to 3.5 Å. The image was taken with a detector distance of 300 mm. C. A zoomed in region of the diffraction in panel B showing a spot at the resolution edge of 3.5 Å.
Figure 5.5. Seeding improves crystallisation of Spt4/5_{284-375}
A. Optimised crystallisation conditions for Spt4/5_{284-375} were seeded to produce larger crystals (top) that fluoresced under UV light (bottom). B. The crystals from panel A diffracted to 2.61 Å. The image was taken with a detector distance of 600 mm. C. A zoomed in region of the diffraction in panel B showing a spot at the resolution edge of 2.6 Å.

5.3.2. Data collection and X-ray diffraction analysis
A complete native dataset in the resolution range 64.2 Å – 2.6 Å was collected for Spt4/5_{284-375} crystals at the microfocus beamline (MX2) of the Australian Synchrotron from rotation of the crystal through 360° in the resolution range with 1 s exposure per 1° rotation image. The crystal was indexed on a primitive lattice in space group P1 with unit cell parameters, a = 62.7 Å, b = 68.6 Å c = 99.6 Å, a = 71.8°, b = 76.7°, c = 76.9°; however, the data was not of sufficient quality to be used for structural solution (data statistics not shown).
Although the crystallisation of Spt4/5\textsubscript{284-375} did not yield a structural solution, the crystallisation attempts for the Spt4/5\textsubscript{284-375} protein alone were carried out concurrently with the co-crystallisation of Spt4/5\textsubscript{284-375} with RNA (Spt4/5\textsubscript{284-385}:RNA), which did yield a solution.

5.3.3. Crystallisation of Spt4/5\textsubscript{284-375}:RNA.

The optimised Spt4/5\textsubscript{284-375} crystallisation conditions (Table 5.2) were used to co-crystallise Spt4/5\textsubscript{284-375} with a 12-nt RNA probe (UCAAUAACAACA) containing three AA repeats (3\_AA). The resulting Spt4/5\textsubscript{284-375}:RNA crystals were significantly larger than the protein only crystals of the same conditions, with diffraction to 3 Å (Figure 5.6).

![Figure 5.6. Crystallisation of Spt4/5\textsubscript{284-375}:RNA co-crystals](image)

A. Optimised crystallisation conditions for Spt4/5\textsubscript{284-375} was used to co-crystallise the hetero-dimer producing large plate shaped crystals (top) that fluoresced under UV light (bottom). B. The crystal from panel A. diffracted to 3 Å. The image was taken with a detector distance of 300 mm. C. A zoomed in region of the diffraction in panel B showing a spot at the resolution edge of 3Å.
A complete native dataset in the resolution range 40 – 3 Å was collected for Spt4/5\textsubscript{284-375}:RNA crystals at the microfocus beam line (MX2) of the Australian Synchrotron, from the rotation of the crystal through 360° of data in the resolution range, with 1 s exposure per 1° rotation image. The crystal was indexed on a monoclinic lattice in space group I\textsubscript{1}2\textsubscript{1} with unit cell parameters, a= 89.5 Å, b= 47.6 Å c= 255.9 Å, a= 90.0°, b= 92.5°, c= 90.0°. The data collection statistics are listed in Table 5.4. The data quality statistics were acceptable indicating that the data were of good quality with a high-resolution limit of 3Å. Furthermore, the L-test to detect twinning (Xtriage, PHENIX) gave a multivariate Z-score of 1.36. Good or reasonable data typically have a Z-score lower than 3.5, therefore, no twinning was suspected in the dataset.

<table>
<thead>
<tr>
<th>Table 5.3. Spt4/5\textsubscript{284-375}:RNA crystal data collection and processing statistics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values in parentheses correspond to data in the highest resolution shell.</td>
</tr>
<tr>
<td>Space group</td>
</tr>
<tr>
<td>Cell dimensions</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>(R_{merge})</td>
</tr>
<tr>
<td>(R_{meas})</td>
</tr>
<tr>
<td>(R_{p.i.m.})</td>
</tr>
<tr>
<td>Number of observations</td>
</tr>
<tr>
<td>Number of unique reflections</td>
</tr>
<tr>
<td>(I/\sigma(I))</td>
</tr>
<tr>
<td>Completeness (%)</td>
</tr>
<tr>
<td>Multiplicity</td>
</tr>
</tbody>
</table>

\[
R_{merge} = \frac{\sum_{h} \sum_{l} |I_{hl} - \langle I_h \rangle|}{\sum_{h} \sum_{l} \langle I_{hl} \rangle} \sum_{h} \sum_{l} (I_{hl})
\]

\[
R_{meas} = \left( \frac{n_h}{n_h + 1} \right)^{1/2} \frac{\sum_{h} \sum_{l} |I_{hl} - \langle I_h \rangle|}{\sum_{h} \sum_{l} \langle I_{hl} \rangle} \sum_{h} \sum_{l} (I_{hl})
\]

\[
R_{p.i.m.} = \left( \frac{1}{n_h - 1} \right)^{1/2} \frac{\sum_{h} \sum_{l} |I_{hl} - \langle I_h \rangle|}{\sum_{h} \sum_{l} \langle I_{hl} \rangle} \sum_{h} \sum_{l} (I_{hl})
\]

where I is the intensity of an individual measurement of the reflection, \(\langle I_h \rangle\) the mean intensity of that reflection and n is the number of observations of reflection.

Molecular replacement was used to obtain phase information, using 2EXU as a search model. To prepare the 2EXU structure for use as a search model, the Zn\textsuperscript{2+} atom of Spt4
and the engineered linker between Spt4 and Spt5 was removed. Additionally, Spt4 was
delineated in the hetero-dimer as chain A and Spt5 as chain B. To account for any
variability between the structures, a range of ensembles were trialled for molecular
replacement, including: 2EXU, described above, as an individual ensemble, Spt4 alone
(from the modified 2EXU structure), Spt5 alone (from the modified 2EXU structure),
Spt4 and Spt5 (from the modified 2EXU structure) as separate ensembles, and one or
both sub-units as poly-alanine models. The log likelihood gain (LLG) of the solutions
gives an indication of how much “better” the solution is compared to a random solution,
and was used to monitor the success of molecular replacement. A unique solution with
an LLG of 40 or greater usually suggests a correct solution has been obtained, although
the current version of PHASER aims for an LLG of 120. Signal to noise is judged by a
TF Z-score, which indicates how many standard deviations a solution is above the
mean, and a TF Z-score of greater than 6 often represents a correct solution.

The top molecular replacement solution from PHASER for Spt4/5\textsubscript{284-375}:RNA came
from using the modified 2EXU Spt4/5 structure as a single ensemble, present in 4
copies in the asymmetric unit, giving a TF Z-score of 11.8 and an LLG of 304. Based
on this score it was assumed that a successful molecular replacement solution to the
phase problem had been obtained. The solution was used to compute electron density
maps which were visually inspected in COOT. The resulting density was interpretable
and there were no unfavourable steric clashes between neighboring molecules in the
lattice. Difference density was observed for the Zn\textsuperscript{2+} atoms, loop regions and side
chains, and these were modelled manually using COOT, and refined using REFMAC
and PHENIX. Iterative model building was carried out using COOT and
PHENIX/REFMAC to further improve the phases of the model. The refinement
statistics of the completed model are listed in Table 5.4.
Table 5.4. Refinement and final model statistics for the final Spt4/5_{284-375} structure

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>45.3-3 (3.1-3)</td>
</tr>
<tr>
<td>Total reflections</td>
<td>43194</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>21637</td>
</tr>
<tr>
<td>$R_{\text{work}}/R_{\text{free}}$</td>
<td>0.22/0.28</td>
</tr>
<tr>
<td>Number of atoms</td>
<td>5337</td>
</tr>
<tr>
<td>Protein</td>
<td>5333</td>
</tr>
<tr>
<td>Ligand molecules</td>
<td>4</td>
</tr>
<tr>
<td>B-factors (Å$^2$)</td>
<td>54.5</td>
</tr>
<tr>
<td>Protein</td>
<td>54.5</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>39.8</td>
</tr>
<tr>
<td>RMS deviations</td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.01</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.6</td>
</tr>
<tr>
<td>Ramachandran favoured (%)</td>
<td>91</td>
</tr>
<tr>
<td>Ramachandran outliers (%)</td>
<td>2.9</td>
</tr>
</tbody>
</table>

$$R_{\text{work}} = \left( \frac{\sum \sum_h \left| F_{\text{obs}}(h) - k F_{\text{calc}}(h) \right|}{\sum_h |F_{\text{obs}}(h)|} \right)$$

$R_{\text{free}}$ is the $R$-factor for a randomly selected subset (5%) of the data that have not been used for minimisation of the crystallographic residual.

5.3.4. Structure analysis of Spt4/5_{284-375}:RNA

The structure shows an arrangement of four Spt4/5_{284-375} molecules per asymmetric unit, and contains four Zn$^{2+}$ atoms (Figure 5.7A). Size exclusion chromatography does not indicate formation of a higher order complex (Figure 5.2B); therefore, the multiple copies of Spt4/5 in the asymmetric unit are most likely a function of crystal packing rather than biological significance.
Figure 5.7. A model of the X-ray structure of Spt4/5_{284-375}

A. The structure contains four Spt4/5_{284-375} molecules in the asymmetric unit. The structure is shown as a ribbon representation showing the secondary structure elements and is coloured by chain. The Zn^{2+} atoms of Spt4 are shown as grey spheres, and the positions and chain denominations (A-H) of Spt4 and Spt5 in each hetero-dimer are indicated. B. Superimposition of each hetero-dimer in the asymmetric unit showing the flexible loop region between \( \alpha_1 \) and \( \beta_2 \). Molecules are coloured as in panel A. C. Superimposition of Spt5 from yeast (purple; PDB entry 2EXU (28)) and human (grey; 3H7H (31)) showing the variability in the orientation of the loop region between \( \alpha_1 \) and \( \beta_2 \).

The backbone of each hetero-dimer in the asymmetric unit are superimposable with the exception of a single loop region between \( \beta \)-strand 2 (\( \beta_2 \)) and \( \alpha \)-helix 1 (\( \alpha_1 \)) of Spt5 (Figure 5.7B). This loop is shown to adopt multiple conformations, and the average temperature factor for this region is 96.8 compared to a total Spt5 temperature factor of 69.8 (data not shown). These results suggest a degree of flexibility for this region. Alignment of the yeast (2EXU) and human Spt4/5 structures also show variability in the position of this loop (Figure 5.7C); however, there is little sequence homology across
the species for this loop region (Supplementary figure 11.1), and it is not involved in making crystal contacts in any of the structures. Thus, whether its flexibility is relevant to the function of Spt5 is unlikely.

For the 2EXU structure, formation of the Spt4/Spt5 hetero-dimer in the crystal lattice resulted from the binding of Spt4 of one fused monomer to the Spt5 NGN domain of an adjacent fused monomer. Alignment of the 2EXU structure with Spt4/5\textsubscript{284-375} showed an identical arrangement of the Spt4/Spt5 interface, supporting their model that that the interface of Spt4 and Spt5 NGN domains from distinct monomers represents the normal mode of interaction between Spt4 and Spt5 \textit{in vivo}.

Unfortunately, despite the fact that RNA was added to the crystallization conditions of the protein, after refinement of the structure was complete, no extra density was observed to indicate the presence of RNA in the crystal. As the crystal packing showed large regions of empty space, crystal soaks with fresh RNA prior to crystal freezing were attempted; however, these soaks and subsequent data collection and analysis of the electron density maps failed to reveal the presence of RNA. Interestingly, even though there was no RNA present in the electron density, the presence of RNA during crystallisation clearly had an effect on producing better diffraction quality crystals that were much larger than when the protein was crystallised alone. It is possible that the high concentration of RNA facilitated improved contacts, producing a more ordered lattice; however, further experiments are needed to determine if using a lower concentration of RNA has the same effect, or if the improved crystallisation is due to an overcrowding effect.
Data presented in Chapter 3 suggested a conformational change upon RNA binding; therefore, it stands to reason that the crystal conditions for a protein:RNA complex could be entirely different from the protein only crystal conditions. Therefore a variety of conditions were tested using the Hampton Index, PEG/ION and NATRIX screen using a 96-well sitting drop format to identify new crystal forms. Crystallisation events were observed in conditions identical to the protein only conditions (Figure 5.3); however, multiple crystallisation events were also observed in conditions that differed significantly from the protein only conditions (Figure 5.8). These conditions also produced thin plate shaped crystals, similar to the morphology of the protein-only crystals, with one exception. The crystals from the condition, sodium thiocyanate and 20% PEG 3,350, produced small spherical crystals.

All crystals obtained fluoresced under UV light indicating that they contained protein. Not all of the crystals obtained were tested for diffraction as they were very small and difficult to extract from a 96-well plate using a mounted cryoloop. Of the crystals successfully frozen, the diffraction was tested using the MX2 beamline at the Australian Synchrotron. There was no diffraction for the spherical crystals, and the other crystals had very weak diffraction to 9-12Å (Figure 5.8). A few datasets were collected from these in order to attempt to determine a space group for the crystal, however, the data was not of high enough resolution to pursue further analysis (data statistics not shown). The fact that these crystals diffracted at all indicates that the crystals are indeed protein, and the fact that they were obtained in significantly different conditions from the protein-only crystals is promising, as it suggests that they may contain a protein:RNA complex. Proteinase K or RNase A digestion of the crystals could confirm the presence both RNA and protein in the crystals, and further optimisation of the conditions will be needed to improve the size and diffraction quality of the crystals for structure solution.
CHAPTER 5

Figure 5.8. Crystallisation results for Spt4/5<sub>284-375</sub>:RNA

Multiple crystallisation events were observed from various Hampton research screens in conditions that differed from the protein only crystals.

5.4. Elucidating a mechanism of RNA binding by Spt4/5

In the absence of an X-ray Spt4/5:RNA complex structure, a bioinformatics approach was used to identify a putative RNA interaction site on Spt4/5. Chains A (Spt4) and B (Spt5<sub>284-375</sub>) from the Spt4/5<sub>284-375</sub> structure were extracted from the asymmetric unit of Spt4/5<sub>284-375</sub> (Figure 5.7) and used for the subsequent analysis.
5.4.1. The NGN domain of Spt4/5 contains an RRM

Comparison of the topology between the NGN domain of Spt5 and other RNA binding modules revealed a remarkable similarity to the RNA recognition motif (RRM) (Figure 5.9A). This similarity has been observed previously after the determination of the crystal structure of NusG (64), however, seems to have been overlooked in the description of the eukaryotic and archaeal complexes (28, 29, 59, 121). The NGN domain of yeast Spt5 contains the typical RRM arrangement of four antiparallel β-strands packed against two α-helices; however, the α1-helix is extended and there is an additional C-terminal α-helix (α3) (Figure 5.9A).

Spt4 interacts with the NGN domain of Spt5 through the alignment of their β-strands (28). In the context of an RRM, this would serve to extend the β-sheet surface by an additional two β-strands (Figure 5.9B). Such an extension of the β-sheet has been seen in proteins containing tandem RRMs, such as the poly(A)-binding protein (177) and the polypyrimidine tract binding protein (178) to help accommodate longer RNA motifs and achieve higher affinity for the RNA target; however, this is the first report to date of a hetero-dimeric RRM. The extension of the potential RNA binding surface by Spt4 would explain why both Spt4 and Spt5 are required as a complex to make significant RNA interactions.
Figure 5.9. Spt5 NGN contains an RRM

A. Comparison of the NGN domain of Spt5\textsubscript{284-375} (purple) with the canonical RRM domain of NonO (green; PDB entry 2CPJ) showing the extended $\alpha$1 helix and the additional C-terminal $\alpha$3 helix of Spt5. The N and C terminus of each protein is shown. B. The structure of the yeast Spt4/5\textsubscript{284-375} hetero-dimer showing that Spt4 (blue) contributes two additional $\beta$-strands to the RRM of Spt5. The three canonical aromatics on the surface of the extended $\beta$-sheet are shown (stick representation); Only Y336 and Y74 are conserved; however, the archaeal and human Spt5 contain an additional aromatic at position R334. C. The cartoon structure of yeast Spt4/5 coloured by B-factors shows that the C-terminal region of Spt5 and portions of the C-terminus of Spt4 display flexibility.

Most canonical RRMs interact with ssRNA through three highly conserved aromatic rings located on the $\beta$-sheet surface. Spt5 NGN only has two aromatic residues on this surface: F326 on $\beta$2 and Y336 on $\beta$3; however, the extension of the RRM by Spt4 contributes a third aromatic residue to the $\beta$-sheet surface: Y74 (Figure 5.9B). Of these three aromatic residues, only Y336 and Y74 are conserved; however, archaeal and human Spt5 have a third aromatic residue at what would be positioned identically to R334 in yeast (Figure 5.9B and Supplementary figure 11.1). The presence of three
aromatic residues on the surface of the β-sheet in Spt4/5 from yeast, human and archaea suggests that Spt4/5 may bind RNA in a canonical RRM-like manner.

In all three Spt4/5 homologs however, the β-sheet surface is occluded by the position of the additional C-terminal α3 helix (Figure 5.9B). This is a feature that has also been seen in other RRM containing proteins (115, 179-181). Analysis of the temperature factors for the Spt4/5 structure shows that the C-terminal region of Spt5 (from residue K340-Q366) connecting the C-terminal helix to the RRM displays a higher flexibility in comparison to the rest of the structure (Figure 5.9C). Furthermore, the average B-factor for this region is 39.4 compared to the average B-factor for the whole Spt4/5 structure of 19.7 (data not shown). Therefore, it would be possible for this region to accommodate a conformational change in the Spt5 structure upon RNA binding.

Typically RRM’s do not undergo a conformational change upon RNA binding; however, a similar mechanism for RNA binding through the displacement of the C-terminal helix has been observed for the RRM's of the human U1A spliceosomal protein (115, 116). Additionally, in Tra2-β1, the unstructured N and C-termini of regions of the free RRM undergo a significant structural change, becoming ordered upon RNA binding (182, 183). The CD results described in Chapter 3 indicate Spt4/5 undergoes a conformational change upon RNA binding (see section 3.4.5 and Figure 3.4D). Therefore, it is possible that the α3 helix is displaced upon RNA binding, making the aromatic residues on the β-sheet surface available for interaction with RNA. This has been suggested as an RNA binding mechanism for NusG (64), although to date there is no experimental evidence supporting this.
5.4.2. Other possible modes of RNA binding by Spt4/5

The repetitive AA motif bound by Spt4/5 is at least 5-nt long and potentially up to 14-nt in length; however, the β-sheet of an RRM typically only accommodates between two to four nucleotides. The extension of the binding surface by Spt4 could potentially accommodate additional nucleotides, similar to the RNA binding of the polypyrimidine tract binding protein (PTB), where the RRMs contain a fifth β-strand allowing the binding of one or two additional nucleotides (184, 185); however, the specific RNA binding motif of Spt4/5 is still potentially much longer than the available RNA binding surface provided by the β-sheet surface. This suggests residues outside of the canonical β-sheet surface are involved in the RNA recognition, or that there may be another mode of RNA binding.

The results presented in Chapter 3 suggest that Spt4 could be involved in (but not directly responsible for) the RNA binding properties of Spt4/5. Spt4 is a rare type of ZnF containing a Zn$^{2+}$ ion coordinated by four Cys residues. There is only a single report of this class of ZnF binding RNA, the human splicing factor ZRANB2 (124). ZRANB2 binds RNA through two short β-hairpins of the ZnF that make the specific nucleic acid contacts with the RNA motif GGU (124). The RNA sequence specificity of ZRANB2 is mediated by base stacking interactions with a tryptophan residue and hydrogen bonding to arginine, asparagine and aspartate residues.

Like the C-terminal helix of Spt5, the disordered C-terminus of Spt4 is also shown to be highly flexible (Figure 5.9C). Interestingly this region contains a number of residues that could be involved in making both specific and non-specific RNA contacts, and which are also conserved in human Spt4 (supplementary figure 11.2). The C-terminus of archael Spt4 is significantly shorter than its yeast or human counterparts and doesn’t
contain this flexible region. Therefore, consistent with the knowledge that zinc fingers are able to bind RNA, as well as a conformational change in the Spt4/5 hetero-dimer upon RNA binding, Spt4 could be involved in making RNA contacts, possibly through the re-orientation of the C-terminus in an RNA binding event.

RNA recognition can also be achieved by other planar side-chains such as Arg, Asn, Asp or His acting in a similar way to the aromatics. Such mechanisms for RNA interactions have been observed for the RRM s of hnRNP A1 (186) and SRp20 (187). Furthermore, in many other RRM containing proteins, the loops between the β-strands and α-helices can be involved in RNA recognition. For example, the RRM of the human Fox-1 protein uses the loop regions to bind to the first four nucleotides of an RNA heptamer, whilst the last three nucleotides bind to the RRM in a canonical way (152). In this way, the RNA binding surface is extended to more than double that of a canonical binding surface. Additionally, the quasi RRM s (named so for their distinct lack of conserved aromatic residues) of the heterogeneous nuclear ribonucleoprotein F (hnRNP F) use only the loops and not the β-sheet to bind RNA (181).

As the loop regions of RRM containing proteins are highly variable in length as well as sequence, the accurate prediction of an RNA binding mode for Spt4/5 by this mechanism is difficult. Furthermore, the loop regions of the Spt5 RRM do not contain any conserved residues that are known to form RNA interactions. This suggests that these loops are not involved in RNA binding, or that the mode of RNA binding by this mechanism is not conserved in other Spt4/5 homologs.

Analysis of the electrostatic surface potential of RNA binding proteins can often provide insight into the RNA binding mechanism of an RBP. The electrostatic surface
of the Spt4/5 hetero-dimer shows Spt5 to be predominantly positively charged, while Spt4 is predominantly negatively charged (Figure 5.10). Furthermore, Spt5 contains a highly positively charged groove, which is commonly seen in RNA binding proteins where it interacts with the phosphate backbone of an RNA molecule. The positively charged patch on Spt5 could represent a likely RNA interaction site on Spt4/5. This would suggest that Spt5 is mainly responsible for making direct contact with the RNA, and that Spt4 is facilitating RNA binding through contributing structural stability to the hetero-dimer.

![Figure 5.10. The electrostatic surface of Spt4/5 NGN](image)

The positive and negative charged regions of the surface of the Spt4/5\textsubscript{284-375} hetero-dimer are shown in blue and red respectively, and the positively charged groove of Spt5 is indicated. The ribbon structure for each orientation of the surface representation is also shown.

Overall, analysing the structure of Spt4/5 using a bioinformatics approach uncovers possible sites that could facilitate interaction with RNA. Determination of the level of
conservation of the RNA binding specificity observed for yeast Spt4/5 between species could be useful in the identification of a likely RNA binding surface; however, structural validation of the models using X-ray crystallography or other structural techniques is needed to elucidate the exact mechanism of RNA interaction by Spt4/5.

5.5. Conclusions

This chapter has described the approach towards the determination of the Spt4/5:RNA binding interface. Although the X-ray structure of Spt4/5\textsubscript{284-375} in complex with its RNA target was unattainable within the time constraints of this study, good progress was made towards this goal. Therefore, with some optimisation of the crystal conditions, a protein:RNA structure could be realistically attainable in the near future. Through bioinformatic analysis, the possibility of Spt4/5 being a non-canonical RRM and analysis of potential modes of RNA binding were investigated through comparisons with observations in the literature. Eventual structural solution of an Spt4/5:RNA complex will elucidate the exact binding mode of this protein.

Chapter 3 provided evidence that the NGN domain was the major protein determinant for RNA binding; however, this domain has also been well characterised for its direct interaction with RNAPII (59). The genomic analysis presented in chapter 3 showed that the AA repeat motif is enriched in the 5’ UTR regions of transcripts, and studies have shown that Spt4/5 associates with RNAP only when the nascent transcript is 18 nt in length, just long enough for the 5’ region of the transcript to reach the surface of the RNAPII (61). Therefore, recognition of the AA motif in the 5’ UTR of the nascent RNA by Spt4/5 may serve as the mechanism for the recruitment of Spt4/5 to RNAPII.
Alternatively, RNA binding by Spt4/5 may occur post-RNAPII binding. The additional C-terminal helix of Spt5 contains conserved residues that are a key component of the Spt4/5:RNA polymerase interaction (59). Therefore, RNA binding resulting in the displacement of the C-terminal helix would significantly affect the interaction of Spt4/5 with RNAP. However, cross-linking evidence shows further amino acids on Spt5, other than the ones already characterised for the RNAP-Spt4/5 interface, are able to interact with the surface of RNAPII (122). This suggests that the interaction between Spt4/5 and RNAPII is dynamic. Therefore, a conformational change in Spt4/5 due to an RNA binding event, which resulted in a change of the Spt4/5:RNAP interface, could serve to allosterically modulate RNAPII activity providing a mechanism for the transcriptional control of RNAPII by Spt4/5. It is possible that the presence of RNAPII could have an effect on the affinity of the Spt4/5:RNA interaction and whether Spt4/5 is able to bind RNAPII and RNA simultaneously remains to be seen. Further investigation of this would help shed light on the sequence of events of the RNA binding activity of Spt4/5 in vivo.

Given that RNA binding seems to be associated with a conformational change in Spt4/5, in combination with a potentially dynamic Spt4/5:RNAPII interface, any putative RNA binding surfaces identified in this chapter may likely change when in the context of the TEC. Therefore, validation of the binding models presented in this chapter requires further biochemical and structural studies to characterise the RNA binding capacity of Spt4/5 in a biological context.
6. Final Discussion

Spt4/5 is a transcription elongation factor that regulates the processivity of RNAP. It has been theorised that it does this through the association of the predicted RNA binding KOW domains with the nascent transcript as it emerges from the polymerase. The primary aim of this work was to gain a greater understanding of the transcription elongation factor Spt4/5 through its interactions with nucleic acid in order to gain insight into the mechanism of how it regulates transcription elongation. Specifically, I planned to show definitively that *S. cerevisiae* Spt4/5 is able to bind nucleic acids, in particular RNA. Furthermore, I aimed to determine the binding specificity of the RNA target and kinetically define this interaction. Finally, I intended to obtain a crystal structure of Spt4/5 bound to the specific RNA target in order to determine the protein:RNA binding interface.

In this work, I utilised SELEX and REMSA to show unequivocally that Spt4/5 binds to RNA preferentially over single or double stranded DNA, with specificity for sequences containing multiple AA repeats. Furthermore, I determined that the KOW domains were not important for the sequence specific protein-RNA contact; and, while both Spt4 and Spt5 alone were capable of forming a weak interaction with RNA, it was Spt4 together with the NGN domain of Spt5 that was responsible for the strong molecular interaction.
While there is a pre-conception in the literature that the KOW domains are RNA binding domains, this study showed that the KOW domains of yeast Spt4/5 were not involved in the sequence specific recognition of RNA. Furthermore, there was no evidence to suggest that the multiple KOW domains of Spt5 were contributing in any way to the affinity of the RNA interaction. These results suggest that KOW domains are not specific RNA binding domains; however, the possibility that they can interact non-specifically with nucleic acids in the context of the TEC cannot be ruled out.

Alternatively, KOW domains share close structural homology with, and have been suggested to be a subset of, Tudor domains (64, 86). Tudor domains mediate protein-protein interactions (typically through binding methylated arginine or lysine residues) with proteins that are directly bound to RNA or contain RNA binding motifs themselves and have been found to be involved in RNA processing, stability and translation (188). Therefore, it is possible that the KOW domains of Spt4/5 may similarly mediate protein-protein interactions. As such, the single KOW domain of NusG is speculated to facilitate interactions with protein factors rho, NusA and NusE (64). Furthermore, recent evidence suggests that Spt5 interacts extensively with the Rpb4/7 subunit of RNAPII through its KOW 4-5 domains (122). Further characterisation of the KOW domains of Spt4/5 is required to elucidate their interaction partners.

This study showed that the Spt4 together with the Spt5 NGN domain was the minimum protein determinant for RNA binding. These results were intriguing as RNA binding is not a function that has been previously attributed to this domain of the protein. Upon comparing the structure of the Spt5 NGN domain with other known RNA binding
proteins, it became evident that the protein yielded a structural fold similar to an RRM domain. Furthermore, the RRM of Spt5 is extended by the hetero-dimerisation with Spt4. The RRM surface formed by Spt4/5 contains three aromatic residues that typically interact with RNA in canonical RRM’s; therefore, Spt4/5 may bind RNA in an RRM-like manner. Such a hetero-dimeric extension of an RRM has not previously been reported and the contribution of Spt4 to the RRM fold may explain why both Spt4 and Spt5 are required to make significant RNA contacts.

The observation that both Spt4 and Spt5 are required to make significant RNA contacts raises the question of whether the sequence specific RNA binding observed for yeast Spt4/5 is conserved in other species, as bacterial species lack an Spt4 homologue. RNA interactions have been suggested for both Drosophila and human Spt4/5 (47, 61, 62), and both the human and archaeal Spt4/5 hetero-dimer contain three aromatics on the surface of the canonical RRM binding surface. Therefore, other eukaryotic Spt4/5 homologues may have a similarly specific RNA interaction as shown for the yeast hetero-dimer. In contrast, NusG has been shown to interact with both DNA and RNA (64), which may reflect the absence of any specificity in nucleic acid recognition that could otherwise be provided by the presence of Spt4. Further studies characterising the nucleic acid binding specificity of other Spt4/5 homologues are needed to elucidate the conservation of the sequence specific RNA binding observed for the yeast Spt4/5 hetero-dimer. Additionally, structural characterisation of the Spt4/5:RNA binding interface will be crucial to elucidate the exact contribution of both Spt4 and Spt5 to the RNA interaction.
As Spt4/5 is a ubiquitous transcription factor, it was interesting to find that Spt4/5 bound with sequence specificity to a single stranded RNA sequence encompassing multiple AA repeats. Furthermore, AA repeat motifs were found to be enriched in the 5’ UTR of yeast transcripts, suggesting that the AA repeat sequence motif may be a newly defined regulatory element embedded in RNA transcripts. Additionally, *in vitro* genomic analysis found that transcripts containing the AA repeat motifs were transcribed more frequently than those without. Therefore, recognition of the AA motif in the 5’ UTR of the nascent RNA by the RNAP-bound Spt4/5 may serve as a signalling event to recruit RNA processing machineries that have been shown to associate with Spt4/5 (9, 40, 51, 55, 94).

Alternatively, recognition of the AA motif in nascent transcripts may serve to recruit Spt4/5 to the TEC as several studies on human and *Drosophila* Spt4/5 indicate that its recruitment to RNAP is facilitated by RNA binding (61, 62). However, archaeal Spt4/5 is able to associate with RNAP regardless of nucleic acid binding (30). This again brings in to question whether the sequence specific RNA binding is conserved across all Spt4/5 homologues. Nevertheless, the binding of Spt4/5 to AA repeat motifs in 5’UTRs of nascent transcripts could provides a mechanism for how Spt4/5 is able to exert its effect on transcription elongation.

This thesis demonstrates new mechanistic insight into the basic cellular process of transcription elongation as it changes the paradigm for future studies regarding the role of Spt4/5 and its molecular interaction with RNAPII and the nascent RNA transcript.
7. References


REFERENCES


REFERENCES


109. GraphPad Software. La Jolla California, USA, [www.graphpad.com](http://www.graphpad.com).


111. Shroedinger, LLC. The PyMOL Molecular Graphics System, Version 1.5.0.4.


REFERENCES


REFERENCES

REFERENCES


8. Appendix 1. Oligonucleotide sequences

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spt4 (pETM-11, pET28b, pET-Duet-1)</td>
<td>AGCATCTAGAGCAAGGAGATATACCATGTCTAGTGAAAGAGCCTGTATGC</td>
<td>CTGAAGCTTGTACTCAACTTGACTGCCATCCCTC</td>
</tr>
<tr>
<td>pET-Duet-4/5</td>
<td>ATTCCATATGACAGCTACCATTGGGGTG</td>
<td></td>
</tr>
<tr>
<td>pET-Duet-4/5_1K</td>
<td>GGGCTCGAGTTAGTGGTGGTGGTGGGCTGAGGAAACATTTTCAGCATACCTTC</td>
<td></td>
</tr>
<tr>
<td>pET-Duet-4/5_2K</td>
<td>GGGCTCGAGTTAGTGGTGGTGGTGGTGGTGAATTTATTTGATAGTACAAATCTTTCGTGG</td>
<td></td>
</tr>
<tr>
<td>pET-Duet-4/5_5K</td>
<td>GGGCTCGAGTTAGTGGTGGTGGTGGTGGTGTGTAATGTGTTTGCTTCAGTGG</td>
<td></td>
</tr>
<tr>
<td>pHUE-Spt5</td>
<td>GATGGGATCCACAGCTACCATTTGGGGTG</td>
<td></td>
</tr>
<tr>
<td>pHUE-Spt5_NGN</td>
<td>GGCCATGGTTAACGAGGAACAATTTTCAGCAT</td>
<td></td>
</tr>
<tr>
<td>pHUE-Spt5_1K</td>
<td>GGCCATGGTTAACGAGGAACAATTTTCAGCAT</td>
<td></td>
</tr>
<tr>
<td>pHUE-Spt5_2K</td>
<td>GGGCTCGAGTTAGTGGTGGTGGTGGTGGTGTGTAATGTGTTTGCTTCAGTGG</td>
<td></td>
</tr>
<tr>
<td>pHUE-Spt5_5K</td>
<td>GGGCTCGAGTTAGTGGTGGTGGTGGTGGTGTGTAATGTGTTTGCTTCAGTGG</td>
<td></td>
</tr>
<tr>
<td>pGEX-K1</td>
<td>GTCAAGGATCCGATGTGTGTTTGG</td>
<td>CCTGGAATTTCCATGCAGAAATTT</td>
</tr>
<tr>
<td>pGEX-K2</td>
<td>ACACGGATCCGCTGCAAGGGTCACTTTCC</td>
<td>CGTCGGCTCGAGTTAATTTAGTTTGAATAGCATACAT</td>
</tr>
<tr>
<td>pGEX-K4</td>
<td>CTCAGGATCCCGTAATGAAATCA</td>
<td>ACCAGGCAATTCTTAGTTAACAACAAAAACC</td>
</tr>
<tr>
<td>pGEX-K5</td>
<td>GAGAGGATCCGAGTTTGCCAC</td>
<td>GAGCAGCGGCGGCTTATTGATAGTATGTA</td>
</tr>
</tbody>
</table>
### Supplementary table 8.2. SELEX primers and oligonucleotides

<table>
<thead>
<tr>
<th>SELEX_library</th>
<th>TCCCCTCGTCGTC(T(N&lt;sub&gt;25&lt;/sub&gt;)CCGATCGCTCCCGCTCCTCCCT*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELEX_T7For</td>
<td>GAAATTAATACGACTCACTATAGGAGGACGATCGG</td>
</tr>
<tr>
<td>SELEX_Rev</td>
<td>TCCCGCTCGTGCTCT</td>
</tr>
<tr>
<td>Deep sequencing primer</td>
<td>CAAGCAGAAGACGCCATACGAGCTCTTCGATCTTCCCGCTGTCGTCCT</td>
</tr>
<tr>
<td>Deep sequencing primer (Bns_f)</td>
<td>AATGATACGGCGACCACCGAGATCTACACTCTTTCTCCCTACACGCGCTCTTCGCCGACGCTCT</td>
</tr>
<tr>
<td>Deep sequencing barcode 1</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 2</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 3</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 4</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 5</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 6</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 7</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 8</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 9</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 10</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 11</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 12</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 13</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 14</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 15</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 16</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 17</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
<tr>
<td>Deep sequencing barcode 18</td>
<td>ACGACGCTCTCTCCGATCTGACATCAGGGGAGGACGATCGG</td>
</tr>
</tbody>
</table>
Deep sequencing barcode 19 Forward ACGACGCTCTTCCGATCTGCGTGAAAGGGAGGACGATGCGG
Deep sequencing barcode 20 Forward ACGACGCTCTTCCGATCTGCGTTGGGGAGGACGATGCGG
Deep sequencing barcode 21 Forward ACGACGCTCTTCCGATCTGCGTTTCGGGGAGGACGATGCGG
Deep sequencing barcode 22 Forward ACGACGCTCTTCCGATCTGCACGGGGAGGACGATGCGG
Deep sequencing barcode 23 Forward ACGACGCTCTTCCGATCTGCAGTGGGGGAGGACGATGCGG
*N represents any random nucleotide

Supplementary table 8.3. RNA probes used for REMSA and MST*

<table>
<thead>
<tr>
<th></th>
<th>RNA probes used for REMSA and MST*</th>
</tr>
</thead>
<tbody>
<tr>
<td>5K1</td>
<td>56-FAM-UGGCUCGCAAAUACAAAAACAAAC</td>
</tr>
<tr>
<td>5K1_mut</td>
<td>56-FAM-AGACUCGGCCGACAGUCAGAAGAC</td>
</tr>
<tr>
<td>5K1_GG</td>
<td>56-FAM-UGGCUCGGCGUUGGCGAGGCGAC</td>
</tr>
<tr>
<td>4_AA</td>
<td>56-FAM-UCAAUAACAAAAACA</td>
</tr>
<tr>
<td>2_AA</td>
<td>56-FAM-UCAAUAACAC</td>
</tr>
<tr>
<td>1_AA</td>
<td>56-FAM-UCAAUUUC</td>
</tr>
</tbody>
</table>

* The probes used for REMSA and MST are 5’ fluorescein labelled (56-FAM)
9. Appendix 2. Plasmids generated in this study

Supplementary table 9.1. Overview of cloned plasmids and their corresponding protein products

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Restriction enzymes used for gene insertion</th>
<th>Protein product</th>
<th>Antibiotic resistance and concentration used for expression</th>
<th>Affinity tag location</th>
</tr>
</thead>
<tbody>
<tr>
<td>pET-Duet-4/5_1K</td>
<td>NdeI / XhoI</td>
<td>Spt4/5284-419</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>Spt5 C-terminal (his)</td>
</tr>
<tr>
<td>pET-Duet-4/5_2K</td>
<td>NdeI / XhoI</td>
<td>Spt4/5284-567</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>Spt5 C-terminal (his)</td>
</tr>
<tr>
<td>pET-Duet-4/5_5K</td>
<td>NdeI / XhoI</td>
<td>Spt4/5284-839</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>Spt5 C-terminal (his)</td>
</tr>
<tr>
<td>pET28b-Spt4</td>
<td>XbaI / HindIII</td>
<td>H6Spt4</td>
<td>Kanamycin (30 μg ml⁻¹)</td>
<td>N-terminal (his)</td>
</tr>
<tr>
<td>pET-M11-Spt4</td>
<td>XbaI / HindIII</td>
<td>Spt4</td>
<td>Kanamycin (15 μg ml⁻¹)</td>
<td>None</td>
</tr>
<tr>
<td>pHUE-Spt5_NGN</td>
<td>BamH1 / NcoI</td>
<td>Spt5ubq284-375</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (his)</td>
</tr>
<tr>
<td>pHUE-Spt5_1K</td>
<td>BamH1 / NcoI</td>
<td>Spt5ubq284-419</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (his)</td>
</tr>
<tr>
<td>pHUE-Spt5_2K</td>
<td>BamH1 / Kpn1</td>
<td>Spt5ubq284-567</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (his)</td>
</tr>
<tr>
<td>pHUE-Spt5_5K</td>
<td>BamH1 / Kpn1</td>
<td>Spt5ubq284-839</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (his)</td>
</tr>
<tr>
<td>pGEX-K1</td>
<td>BamH1 / EcoRI</td>
<td>KOW1378-567</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K2</td>
<td>BamH1 / XhoI</td>
<td>KOW2429-567</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K4</td>
<td>BamH1 / EcoRI</td>
<td>KOW4666-714</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K5</td>
<td>BamH1 / Not1</td>
<td>KOW5787-849</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K12</td>
<td>BamH1 / XhoI</td>
<td>KOW1278-567</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K23</td>
<td>BamH1 / EcoRI</td>
<td>KOW23429-626</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K34</td>
<td>BamH1 / EcoRI</td>
<td>KOW3478-714</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
<tr>
<td>pGEX-K45</td>
<td>BamH1 / Not1</td>
<td>KOW45666-849</td>
<td>Ampicillin (50 μg ml⁻¹)</td>
<td>N-terminal (GST)</td>
</tr>
</tbody>
</table>
10. Appendix 3. Supplemental Information for Chapter 3

**Supplementary figure 10.1. The domain organisation and constructs of the *S. cerevisiae* Spt4/5 hetero-dimer**

**A.** The domains of Spt4 showing the zinc finger (Zn) and Spt5 showing the acidic N-terminus, the five Kyrvides, Ouzounis and Woese (KOW) domains and the C-terminal repeat (CTR) region are delineated.

**B.** The names and domain truncations of the Spt4/5 constructs utilised in this study. The amino acid (aa) boundaries of each construct are shown.
Supplementary figure 10.2. SELEX motif analysis

A. The sequence logos of the second and third most significant sequence motif derived from MEME analysis of 2000 unique sequences. The CTCG motif exists in 21% of the input sequences. The AGATC motif exists in 4% of the input sequences. B. The enrichment profile of the top three motifs clearly shows the AA repeat motif was the most highly enriched. C. The enrichment profile of the AA motif containing different numbers of AA repeat units clearly show the longest motif enriched for by SELEX contains five AA repeats (5_AA).
Supplementary table 10.1. Association of $\geq 2$ AA motifs in UTRs with transcription frequency and UTR length

<table>
<thead>
<tr>
<th></th>
<th>5’ UTR</th>
<th></th>
<th>3’ UTR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA*</td>
<td>TT*</td>
<td>neither*</td>
<td>AA*</td>
</tr>
<tr>
<td>median UTR length (bp)</td>
<td>47</td>
<td>62</td>
<td>29</td>
<td>80</td>
</tr>
<tr>
<td>p-value *</td>
<td>N/A</td>
<td>&lt; 2.20e-16</td>
<td>&lt; 2.20e-16</td>
<td>N/A</td>
</tr>
<tr>
<td>median transcription frequency</td>
<td>17.11</td>
<td>13.50</td>
<td>15.00</td>
<td>15.80</td>
</tr>
<tr>
<td>p-value *</td>
<td>N/A</td>
<td>3.84e-05</td>
<td>6.33e-05</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* AA, TT and neither refer to UTRs with $\geq 2$ AA motifs, $\geq 2$ TT motifs and those that lack both motifs.

+ Median number of transcripts produced per gene per cell cycle

# Wilcoxon rank sum test, comparing AA to the TT or neither categories.
Supplementary figure 10.3. KOW domains do not bind RNA in vitro

A. The number of KOW domains in Spt5 does not affect the affinity of Spt4/5 for RNA. The interaction of fluorescently labelled 5_AA RNA with increasing concentrations of Spt4/5 encompassing the NGN domain (284-375) plus one (284-419), two (284-567) or five (284-839) KOW domains, as determined by MST. Symbols show actual data points while the curves represent a fit of the binding data giving $K_d$ estimations. Each data point is the average of 3 replicate measurements.

B. 0.5 nM 5_AA RNA was electrophoresed in the presence of increasing concentrations of each indicated protein. There is no RNA binding by either the single or tandem KOW-domain proteins.
11. Appendix 4. Sequence alignments

Supplementary figure 11.1. Multiple sequence alignment of Spt5 NGN

The alignment was performed using Clustal 2.1 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). The residues are coloured by their degree of similarity with completely conserved residues in red. The residue number and secondary structure of *S. cerevisiae* Spt5284-375 (SPT5_YEAST) is marked above the sequences. All sequences were derived from UniProt (www.uniprot.org/).
**Supplementary figure 11.2. Multiple sequence alignment of Spt4**

The alignment was performed using Clustal 2.1 ([http://www.ebi.ac.uk/Tools/msa/clustalw2/](http://www.ebi.ac.uk/Tools/msa/clustalw2/)). The residues are coloured by their degree of similarity with completely conserved residues in red. The residue number and secondary structure of *S.cerevisiae* Spt4 (SPT4_YEAST) is marked above the sequences. All sequences were derived from UniProt ([www.uniprot.org/](http://www.uniprot.org/)).