The U1A/U2B"/SNF Family of RNA Binding Proteins: Evolution of RNA Binding Specificity and Contributions of Heterotropic Linkage to snRNP Protein Partitioning

Sandra Gisela Williams
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds
Part of the Biology Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
The U1A/U2B″/SNF Family of RNA Binding Proteins:
Evolution of RNA Binding Specificity and Contributions of Heterotropic Linkage to snRNP
Protein Partitioning

by

Sandra Gisela Williams

A dissertation presented to the
Graduate School of Arts and Sciences
Of Washington University in
Partial fulfillment of the
Requirements for the degree
Of Doctor of Philosophy

May 2015

St. Louis, Missouri
# Table of Contents

List of Figures
List of Tables
Acknowledgments
Abstract
Chapter 1. Introduction
  References.
Chapter 2. *Drosophila* SNF protein binds two RNA hairpins
  References.
Chapter 3. Human U2B" Protein Binding to snRNA stemloops
  References.
Chapter 4. Introduction to Phylogenies and Ancestral Sequence Reconstruction
  References.
Chapter 5. Resurrection of an Urbilaterial U1A/U2B"/SNF Protein
  References.
  Supplementary material.
Chapter 6. Evolution of Ancestral URB Protein toward Modern Vertebrate Traits
  References.
  Supplementary table.
Chapter 7. Linkage and Allostery in snRNP Protein/RNA Complexes
  References.
  Supplementary material.
Chapter 8. Binding Affinity and Cooperativity Control snRNA/ U2B"/ RNP formation
  References.
  Supplementary figures.
# List of Figures

<table>
<thead>
<tr>
<th>Figure 1.1.</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.2.</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1.3.</td>
<td>12</td>
</tr>
<tr>
<td>Figure 1.4.</td>
<td>13</td>
</tr>
<tr>
<td>Figure 1.5.</td>
<td>17</td>
</tr>
<tr>
<td>Figure 2.1.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.2.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 2.3.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2.4.</td>
<td>43</td>
</tr>
<tr>
<td>Figure 2.5.</td>
<td>46</td>
</tr>
<tr>
<td>Figure 2.6.</td>
<td>48</td>
</tr>
<tr>
<td>Figure 2.7.</td>
<td>50</td>
</tr>
<tr>
<td>Figure 2.8.</td>
<td>52</td>
</tr>
<tr>
<td>Figure 2.9.</td>
<td>53</td>
</tr>
<tr>
<td>Figure 3.1.</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.2.</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.3.</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3.4.</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3.5.</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3.6.</td>
<td>67</td>
</tr>
<tr>
<td>Figure 4.1.</td>
<td>73</td>
</tr>
<tr>
<td>Figure 4.2.</td>
<td>74</td>
</tr>
<tr>
<td>Figure 4.3.</td>
<td>80</td>
</tr>
<tr>
<td>Figure 5.1.</td>
<td>88</td>
</tr>
<tr>
<td>Figure 5.2.</td>
<td>89</td>
</tr>
<tr>
<td>Figure 5.3.</td>
<td>90</td>
</tr>
<tr>
<td>Figure 5.4.</td>
<td>91</td>
</tr>
<tr>
<td>Figure 5.5.</td>
<td>93</td>
</tr>
<tr>
<td>Figure 5.6.</td>
<td>94</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table 2.1.</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2.</td>
<td>44</td>
</tr>
<tr>
<td>Table 2.3.</td>
<td>48</td>
</tr>
<tr>
<td>Table 2.4.</td>
<td>48</td>
</tr>
<tr>
<td>Table 2.5.</td>
<td>50</td>
</tr>
<tr>
<td>Table 3.1.</td>
<td>65</td>
</tr>
<tr>
<td>Table 3.2.</td>
<td>67</td>
</tr>
<tr>
<td>Table 5.1.</td>
<td>91</td>
</tr>
<tr>
<td>Table 5.2.</td>
<td>92</td>
</tr>
<tr>
<td>Table 5.3.</td>
<td>92</td>
</tr>
<tr>
<td>Table 5.4.</td>
<td>94</td>
</tr>
<tr>
<td>Table 5.5.</td>
<td>98</td>
</tr>
<tr>
<td>Table 6.1.</td>
<td>131</td>
</tr>
<tr>
<td>Table 6.2.</td>
<td>133</td>
</tr>
<tr>
<td>Table 6.3.</td>
<td>141</td>
</tr>
<tr>
<td>Table 6.4.</td>
<td>143</td>
</tr>
<tr>
<td>Table 6.5.</td>
<td>98</td>
</tr>
<tr>
<td>Table 7.1.</td>
<td>169</td>
</tr>
<tr>
<td>Table 8.1.</td>
<td>184</td>
</tr>
<tr>
<td>Table 8.2.</td>
<td>184</td>
</tr>
<tr>
<td>Table 8.3.</td>
<td>187</td>
</tr>
<tr>
<td>Table 8.4.</td>
<td>187</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

Someone once told me that I would shed sweat, blood, and tears (literally) along the path to completing a thesis. This turned out to be prophetic, but the help and support of many friends, colleagues, and my family meant that I shed far less of each than I might well have.

My family has always been an incredible source of strength and support. This work is dedicated to the memory of my grandmother, who was at my birth and at almost all of the personal milestones that followed. She was an incredible woman: smart, strong-willed, and above all, generous. She had great dreams for my future. While I doubt I can live up to those expectations, they remain a source of encouragement and optimism and a standard for which to strive.

In their own unique ways, my other grandparents were also sources of inspiration and strength. My aunt Mari and uncle Carlos have been a second set of parents; they took me into their home for a few important years and shared both their family and their love of science. My original parents have been endlessly loving and invariably supportive of my endeavors. They tell me that this is what parents do, but I know few people who can say the same. They even followed me to St. Louis, and together we’ve made a home. It’s always a joy to come home to them. My brother David has been a companion throughout almost all of my life. He is a remarkably selfless person with a keen sense of adventure. He has provided wise advice and support over the years, and he has made sure to take me along with him on some of his more tame escapades.

My lab family has also been wonderful. Over the years, we’ve broken up science with
shared laughter and cookies baked in the toaster oven. We’ve watched Star Trek together, decorated a few Christmas trees, and carved a patch-worth of pumpkins. In a special way, I must thank Kim Delaney, who has been a very special friend since she joined the lab.

The content of this dissertation was, of course, a group effort. Chapter 5 would have been absent if not for the tenacity of Mariah Lawler, a rotation student who took on what I considered the tedious task of 50 odd QuikChange mutations. More recently, Kim Delaney has taken up much of the work to advance this project. Greg DeKoster and Changguo Tang have very kindly helped me with NMR-related problems, and Michael Rau very generously made quite a few RNAs for me over the years. Tom Stump always saves the day when equipment or instrumentation is not working as it ought! And I would be remiss not to acknowledge that a few of the experiments presented were actually done by Kathleen.

I first encountered most of my thesis committee during my coursework. They were excellent teachers and continued to be both supportive and critical throughout the years of my thesis work. They made a number of valuable suggestions that improved this dissertation. Michael Harms was my first ever science pen pal. I have yet to meet him, but a talk of his inspired the middle third of this thesis. Mike was incredibly willing to assist me in learning about phylogenetics, and he has been extremely encouraging over the years and provided a lot of guidance for that part of the thesis.

Almost last (but never least), I must thank Kathleen B. Hall, who has been my advisor over the last five years. Kathleen has put up with a wedding cake, my renditions of showtunes, and analyses of both classic telenovelas and the drama surrounding my extended family. She has also
tolerated what is undoubtedly the graver sin: my marked preference for proteins over RNA.

Kathleen cultivates a lab environment that is above all positive; that encourages independent and critical thinking, that fosters personal and scientific growth, and that values creativity and individuality. She has given me a great deal of freedom to take projects in directions that I found interesting and resisted telling me not to bother with experiments she probably felt were doomed. She certainly let me do ‘Midnight Experiments’ in the full light of day. She has tried to make me better at talking to people with whom I am not comfortable. This was mostly unsuccessful, but the blame for that is entirely mine, and I thank her for trying.

Kathleen, this is for you:

Finally, I (and this thesis) owe a great deal to the intercession of Saints Jude, Albert the Great, and Thomas Aquinas. I am convinced that without their intercession, I would not have gotten through a few difficult moments. One, in particular, was truly an Impossible Cause.

**Financial support:**

This work has been financially supported by the MSTP training grant T32 GM007200, NIH fellowship F31 GM089576, and NIH grant R01 GM096444.
DEDICATION

a la memoria de mi queridísima abuela, Lumida Rodríguez Montijo de Viscal
ABSTRACT OF THE DISSERTATION

The U1A/U2B″/SNF Family of RNA Binding Proteins:
Evolution of RNA Binding Specificity and Contributions of Heterotropic Linkage to snRNP Protein Partitioning

by Sandra Gisela Williams

Doctor of Philosophy in Biology and Biomedical Sciences (Biochemistry)
Washington University in St. Louis, 2015
Professor Kathleen B. Hall, Chair

The U1A/U2B″/SNF is a family of RNA binding proteins that is a highly conserved throughout eukaryotes. These proteins are found in the U1 and/or U2 splicing snRNPs (small nuclear ribonucleoprotein particles). In humans, U1A and U2B″ specifically bind to the U1 and U2 snRNAs, respectively. The Drosophila genome codes for SNF, an essential protein that localizes to both the U1 and U2 snRNP. While a specific splicing functions for these proteins have not been determined, their conserved snRNP localization suggests an important splicing-related function.

The difference in protein number and partitioning between Drosophila and humans suggested that these proteins may use different RNA binding mechanisms to function in their cellular contexts. This work begins by exploring some of the differences amongst human U1A, U2B″, and Drosophila SNF. The thermodynamics of the RNA-protein interactions also reveal substantial differences in the RNA binding mechanisms of these proteins.
Further studies investigate the evolution of this protein family in metazoans. Reconstructing the protein phylogeny permitted resurrection of ancestral proteins. This led to the discovery that the last common ancestor of humans and *Drosophila* had a single U1A/U2B”/SNF family homolog. This protein had RNA binding properties that most closely resemble those of *Drosophila* SNF. Evolution of protein motions and RNA binding specificity toward the defining characteristics of modern vertebrate proteins is also examined.

Finally, linkage effects between protein-protein and protein-RNA interactions are analyzed. U2A’ is a U2 snRNP-specific protein that binds to U2B” in humans and SNF in *Drosophila*. In *Drosophila*, large, positive linkage was only seen between U2A’-SNF and SNF-U2 snRNA binding. The RNA dependence of enhancement for SNF binding to U2A’ can explain the observed protein partitioning of U2A’ *in vivo*. For the more complicated human system, which contains two SNF homologs, substantial contributions to protein partitioning come from differences in both intrinsic RNA-protein binding affinities and differences in protein-U2A’ binding affinities. RNA dependence of the linkage parameter also contributes to protein partitioning. The binding parameters can explain U2A’ protein partitioning, and the presence of U2A’ reinforces U1A and U2B” partitioning to their respective snRNAs. These linkage studies have important implications for the assembly of RiboNucleoprotein Particles, macromolecular complexes that are fundamental to many cellular activities.
Chapter 1.

Introduction
This thesis began with an interest in one protein, *Drosophila* SNF, and how it was able to interact with RNAs. Much was known about SNF in the context of *Drosophila* genetics. The *snf* gene was first identified in classical *Drosophila* mutagenic screens as a protein that is important for both germline and somatic sex determination in flies (1-4). Its name, *sans fille*, is something of a misnomer; flies with particular mutations in the gene are entirely sterile and do not, as the name suggests, only produce male offspring. Mutations in the gene have also been associated with reduced female viability, and some female flies show evidence of sexual transformation (2). A SNF knock-out is embryonic-lethal (4). At the amino acid level, SNF was very clearly and very closely related to the spliceosomal U1A and U2B” proteins that had been identified in humans, potatoes, and yeast (4,5). However, SNF seemed capable of carrying out functions that were usually divided between these proteins; it localized to two snRNPs in the cell, and it also bound the U2A’ protein in the U2 snRNP. If most other organisms had two separate U1A and U2B” proteins, how was the lone SNF able to do everything that it needed to do?

The underlying hypothesis has been that differences in the intrinsic RNA and protein binding properties of SNF (compared to U1A and U2B”) can explain its ability to perform functions executed by two separate proteins in other organisms. We determined relatively early that SNF does indeed possess unique RNA binding specificity for its *in vivo* targets. In attempting to understand the mechanistic basis for these properties of SNF, we became interested in how these functions emerged historically. Investigating this latter problem led to the conclusion that SNF is not really unique at all; it is not, as had been thought, a U1A/U2B” chimera. Rather, SNF-like proteins are riddled across the metazoan branch of life, whereas U1A and U2B” uniquely belong to jawed vertebrates. Novel protein functions emerged following a
gene duplication on the vertebrate branch of life and not, as had been thought, in an ancestor of flies.

The problem of understanding SNF was always a comparative problem of understanding variation in the larger protein family. The work presented in subsequent chapters compares proteins that are very highly related but yet have different intrinsic properties and RNA binding mechanisms. This is a small part of a much larger problem, that of understanding RNA-protein interactions, which appear to be as ancient as Biology and as varied as Nature.

To begin my life with the beginning of my life, I record that I was born.

-or- RNA: The Hero of Life?

The central dogma of molecular biology, as Marshall Nirenberg very succinctly put it, is that “DNA makes RNA makes protein.” In this paradigm, RNA is simply a conduit between the information-carrying DNA and proteins, the molecules that power life. However, RNA can both store genetic information (like DNA) and be capable of catalytic activity, as first established with Group 1 self-splicing introns from *Tetrahymena* and bacterial RNase P. Like proteins, RNAs can be regulated by cellular metabolites to affect a number of downstream activities, including transcription termination and translation (riboswitches). They can also function in pathways that regulate gene expression (microRNAs) and even as an adaptive immune system, capable of recognizing and destroying foreign DNA (CRISPRs). RNA is clearly much more than just an intermediate between DNA and protein, regardless of whether or not the RNA world hypothesis (which posits that RNA was the first biotic molecule, precursor to both proteins and DNA) is correct. While RNAs can perform some of these functions on their own, RNAs are most frequently found complexed with proteins. The ribosome is essentially a ribozyme; the peptidyl
The transferase reaction is catalyzed by ribosomal RNA (6). However, translation is highly dependent on a host of proteins. CRISPR and microRNA biogenesis and activity are dependent on proteins, as are many aspects of post-transcriptional RNA regulation. The spliceosome, which will be discussed later, is one of the most complex macromolecular machines, with five non-coding RNAs and well over 150 proteins that together coordinate pre-mRNA splicing. Thus, RNA-protein interactions play critical roles in many of the most fundamental activities of a cell as well as many of the most specialized, regulatory cell events.

Happy families are all alike; every unhappy family is unhappy in its own way.

-or- RRM: an Unhappy Family?

Unlike DNA, which is found predominantly as a double-stranded B-form duplex, RNA is highly heterogeneous in structure. This allows RNA to be a highly versatile molecule. Double-stranded regions of RNA form an A-form duplex, which is characterized by a much deeper, narrower major groove than that found in B-form DNA. This makes sequence-specific recognition of major groove nucleobases (a common binding mechanism for DNA binding proteins) an unfeasible strategy for sequence-specific RNA binding. Indeed, most double-stranded RNA binding proteins are not sequence-specific.

Important in the structural heterogeneity of RNA is the existence of single-stranded stretches. These can serve as binding sites for proteins, but they are also highly susceptible to degradation. Much of ‘single-stranded’ RNA is in fact found within the context of a defined secondary structure, such as a stem-loop, an internal loop, a bulge, or a junction. The geometrical constraints imposed by the secondary structure can play an important role in protein recognition.
A number of domain types exist to recognize single-stranded RNA. These include K Homology domains, zinc fingers, Pumilio and FBF (PUF) homology proteins, and RNA recognition motifs (RRMs). For PUF proteins, which are made up of repeating units (much like ankyrin and leucine-rich repeat proteins), there appears to be a relatively straightforward code for RNA recognition. Each 36 residue, 3 α-helix repeat recognizes a single nucleotide, determined by the amino acids at positions 12 and 16 of the repeating unit (reviewed in (7)). However, for RRM, by far the most abundant eukaryotic RNA binding motif, no such code exists. Rather, relatively small variations in structure within a simple, 80 amino acid protein domain, result in highly variable RNA binding properties, which are not readily predictable from the protein sequence.

RRMs are essentially a ferredoxin protein fold with two consensus sequences called RNP (RiboNucleoproteinParticle) motifs. These sequences are found in the middle strands of the β-sheet surface, where single-stranded RNA will bind (although there are exceptions to this, as well). The RNP motifs are of the form:

RNP1: [RK]-G-[FY]-[GA]-[FY]-[ILV]-X-[FY] (on β3)

RNP2: [ILV]-[FY]-[ILV]-X-N-L (on β1)

The aromatic residues are solvent-exposed, and they can (and do) stack with RNA nucleobases, providing a general platform for RNA binding. Specificity is achieved through interactions (which can include direct and water-mediated hydrogen bonding, salt bridges, and hydrophobic packing) between the RNA and protein moieties, with contributions from nearby sidechains or the peptide backbone. Thus, domain loops, tails, and extensions to the basic ferredoxin fold
often form the basis of the proteins’ RNA binding specificity. Electrostatics play an important role in binding. Most RRMs have very significant positive charge, a characteristic that is intuitively important for binding polyanionic RNA.

Figure 1. Canonical RRM fold. Location of RNP motifs and Loop 3 are indicated.

U1A RRM1 has been studied for years as a prototypical RRM, and a great deal has been learned about the specifics of RNA-RRM interactions from this model protein. However, it has also become clear that U1A RRM1 is, in many ways, an unusual RRM, binding to its target RNA (stem-loop II of U1 snRNA) with extremely high affinity and specificity. Below, I will highlight binding features of three RRM-containing proteins, which will serve to emphasize some of the commonalities as well as some of the source of variability in RNA-protein interactions within this protein family.

U1A. This was one of the first RRM-containing proteins to be investigated, and it remains the best-studied. U1A contains two RRMs separated by a long (~100 amino acid) interdomain linker, which is predicted to be intrinsically disordered. While the second domain folds into a typical RRM structure, it does not appear to bind RNA (8). Its RNP2 motif is canonical, whereas the first three residues of RNP1 are HDI, which clearly deviate from the
consensus sequence ([R/K]-[G]-[F/Y] at these sites). This change replaces one of the canonical aromatic side chains with a large hydrophobic residue. The aspartate substitution both confers undesirable charge characteristics to the RNA binding surface and restricts motions in Loop 3, which can be important for RNA binding. Outside the RNPs, the electrostatic potential surface of the β-sheet surface includes both positive and negative patches, and the overall charge of the domain is relatively neutral. This is in contrast to most RNA binding proteins, which have a significant net positive charge. RRM2 is not unique in its failure to bind RNA. Whole classes of RRM2s have now been identified that appear to bind proteins rather than RNAs (9), and it is possible that RRM2, which is highly conserved in eukaryotes, is important for as yet unknown interactions with other proteins.

The first RRM of U1A binds very specifically to SLII of the U1 snRNA. The protein-RNA interaction is shown in Figure 2a (10), and its position within the U1 snRNP is shown in Figure 2b (11). Apart from binding to the U1 snRNA, it can also bind to two almost adjacent sequences in the 3′ untranslated region (UTR) of its own pre-mRNA. Binding of the 3′ UTR blocks the polyadenylation machinery from being able to access the pre-mRNA. Polyadenylation is important in stabilizing the transcript and plays a role in mRNA export to the cytosol and translation. Binding by U1A to its own pre-mRNA therefore acts as a negative regulatory feedback on its own translation.
U1A binds with very high specificity to the following consensus sequence:

AUUGCACXXX. Any mutations to this sequence results in at least a 10x loss of binding affinity. The association rate is salt-dependent and faster than the diffusion limit for neutrally charged particles; this rapid association rate is most likely the result of favorable electrostatic interactions between the molecules (12, 13). The geometric positioning of the consensus sequence is critically important for protein recognition. Loop 3 specifically recognizes the loop-closing base-pair and protrudes through the RNA loop, which continues to stack as an A-form duplex in the absence of the protein. U1A therefore disrupts the base stacking interactions and
any potential hydrogen bonding interactions between bases. A single nucleotide insertion
between the end of the stem and the 5′ A of the consensus sequence causes a reduction of affinity
of 3 orders of magnitude (14). Entirely single-stranded display of the consensus sequence is
devastating to binding. And while the last three nucleotides of the consensus sequence are not
specifically recognized and can be replaced by a long ethylene glycol linker with no effect on
binding, their deletion results in greater than five orders of magnitude loss in binding affinity.
The loop therefore needs to be large enough at its 3′ end for proper accommodation on the RNA
binding surface of the protein.

As with many other RRMs, the aromatic sidechains of the RNP motifs stack with loop
nucleobases. Protein sidechain and backbone contacts are made with the RNA consensus
sequence. Residues from all of the β strands (including both RNP motifs) as well as Loop 3 and
Loop 6 are involved in recognizing the consensus sequence. While most of these contacts are
with the nucleobases, a few are with the RNA backbone. The domain features an additional C-
terminal α-helix, which packs against the β-sheet in crystal structures but which is probably not
constrained to this position in solution. The helix does not contact the RNA, but trimming these
residues does decrease RNA binding affinity. Pairwise coupling analysis showed that the C-
terminal helix is indirectly coupled to the β-sheet surface when RNA is bound (15). These
residues are important in stabilizing the interactions between Loop 6 and the RNA and can be
substituted with other residues that have helical propensity; they are important but not sequence-
specific. Recently, a crystal structure of the U1 snRNP was obtained following purification and
limited proteolysis of U1 snRNPs from HeLa cells (11). This structure shows that SLII extends
far from the rest of the snRNP. A larger U1A fragment (containing an extra 10 residues from the
end of the C-terminal helix) was crystallized and showed the protein’s C-terminus folding over
the tip of the RNA loop and contacting the opposite face of the RNA (from that bound by the β sheet). It is certainly possible that the interdomain linker at least transiently contacts the RNA stem, which extends far out from the core of the snRNP.

Pairwise coupling analysis also showed that the aromatic residues in the RNP motifs are coupled to each other and to Loop 3, both in the absence and presence of RNA. Most likely, a complex hydrogen bonding network links these sites and is critical for RNA recognition (16). Molecular dynamics simulations showed fast correlated motions in the free RRM as well as in both the RRM and RNA upon complex formation, which may be important to the observed thermodynamic coupling (17, 18). In U1A RRM1, the first aromatic residue of RNP1 (which lies in Loop 3) is replaced by a glutamine. This is a feature shared with other members of the U1A/U2B”/SNF family. Comparisons across over 12,000 RRMs (19) shows this site to be almost universally as tyrosine or phenylalanine; a glutamine substitution is extremely uncommon. Mutations of this residue in U1A to an asparagines, glutamate, or the phylogenetically more common phenylalanine are devastating to SLII binding, suggesting that this modification of the RNP motif is critical in determining the specificity of RNA binding in this protein family.

**Sex-lethal.** The sex-lethal (SXL) protein is found in *Drosophila* and is important for sex determination. The protein contains two RRMs with a short interdomain linker and long N- and C-terminal tails. The RRMs appear to be sufficient for binding to the their RNA targets, polypyrimidine tracts at the 3’ end of specific introns of *sex-lethal* and *transformer* pre-mRNA. It is thought that binding by SXL inhibits binding of U2AF, a protein which is required for recruitment of the U2 snRNP to the branch site, an essential step in splicing. Thus, SXL binding leads to skipping of the adjacent exon and alternative splicing of the transcript.
Both RRM of SXL are important for specific binding to the polypyrimidine tract. Individually, each RRM recognizes the binding site weakly and with poor specificity, although RRM1 has substantially higher affinity for RNA than RRM2. The combined domains, however, have improved specificity and nanomolar affinity (in low salt) (20). A crystal structure of the two domains bound to the RNA (Figure 3a), shows that the β-sheet surfaces of the two RRM form an extended binding platform, with the β-sheets facing each other to form a V-shaped cleft. The electropositive binding cleft formed by the β sheets accommodates a 9-nucleotide stretch of RNA, which binds across the two β-sheet surfaces (the sequence is UGUUUUUUUU). RRM2 recognizes the first three nucleotides. The RNA then makes a sharp turn, and the remaining nucleotides are recognized by RRM1. The turn requires a uridine stretch and will not accommodate a nucleotide insertion between the RRM2-bound and RRM1-bound nucleotides. Stacking interactions between RNA nucleobases and aromatic residues of the RNPs occur in both RRM. Multiple hydrogen bonds and salt bridges are formed between the RNA backbone and RRM2 (through residues in β2 and RNP1). Interactions between RRM1 and the backbone are also important, although there are fewer backbone contacts than with RRM2, which uses residues in both RNP2 and RNP1 to interact with the backbone. Residues within the RNP motifs of RRM2 are responsible for most of the specific contacts made between the RNA nucleobases and the domain. In contrast, residues within Loop 1, Loop 3, and Loop 5 of RRM1, as well as residues within the RNP motifs, make contacts with nucleobases. The interdomain linker is also utilized for RNA base recognition (21).
Figure 3. (a) RRM of SXL bound to UGUUUUUUUU sequence of RNA. RRM2 is the topmost RRM (PDB ID 1B7F). (b) RRMs 1 and 2 of U2AF65 binding to a U8 sequence (PDB ID 2YH1). (c) RRMs 3 and 4 of PTB binding to CUCUCU RNAs (PDB ID 2ADC).
SXL is an example of one protein that uses tandem RRM to bind RNA. However, the orientation of the RRM relative to each other can be highly variable. For instance, RRM 1-2 of U2AF65, which also binds polypyrimidine tracts but which has somewhat different sequence specificity than SXL, form an almost continuous, flat β-sheet upon which the RNA can bind (Figure 3b) (22). RRM 3-4 of the polypyrimidine tract binding protein (PTB) also are used together to increase binding affinity for RNA. However, in this case the domains’ α-helices are used to create a substantial protein-protein interface between the RRM. The β-sheets that constitute the RNA binding surfaces are oriented away from each other, making it possible to bind discontinuous segments of RNA (Figure 3c).
**Figure 4.** C-terminal RRM of p65 protein bound to stem IV of telomerase RNA (pdb ID 4ERD). The left panel shows the RRM β-sheet surface as oriented in Figure 1. The right panel emphasizes the bulged nucleotides.

**p65.** New structures of RRM-protein interactions continue to uncover novel RRM binding modes (reviewed most recently in (23)). One example is the second RRM of the *Tetrahymena* telomerase protein, p65 (Figure 4). This protein has a fifth β strand, non-canonical RNP motifs, and a C-terminal tail which is largely disordered when not bound to RNA but which folds into a kinked α-helix upon binding. The protein binds with high affinity to a two-nucleotide GA bulge that protrudes from a longer stem structure. The stacking platform is moved from the middle two strands of the β sheet to β2, which contains a tyrosine that stacks with the bulge guanine. This nucleotide is sandwiched between β2 and α3, and multiple hydrogen bonds and hydrophobic interactions characterize the binding of the nucleotide at this site. The bulge Adenine contacts both β2, β3, and Loop 4, but not α3. The long α3 helix binds across the major groove of the stem. Apart from participating in recognition of the bulge G, aromatic residues from α3 stack with duplex bases flanking the bulge. Lysines and arginines within α3 also contact the phosphate backbone of the double-stranded RNA. Recognition of double-stranded RNA is reminiscent of that seen in the splicing protein U1 70K, whose RRM is followed by a long, C-terminal α-helix that tracks down a long stem in the U1 snRNP (shown in Figure 2b).

Importantly, protein binding causes a large changes to the RNA structure. Binding opens the bulge and widens the major groove to accommodate α3. This results in a large bend between RNA stems flanking the bulge (24). While p65 is an unusual RRM, it appears to be a member of a subclass of RRMs that utilize similar RNA binding strategies. These domains have conserved
but atypical RNP1 and RNP2 motifs and include an additional consensus sequence on β2, ‘RNP3’. Aromatic residues in both RNP3 and α3 form the RNA stacking platform, and α3 contains charged residues that can contribute to double-stranded RNA binding. They seem to exist within larger proteins that contain more traditional RRMs.

These examples show that RRMs are very diverse in how they recognize their RNA targets. However, most RRMs use solvent-exposed aromatic sidechains (generally found within the RNP consensus motifs, although exceptions exist) to stack with single-stranded RNA bases. Specific contacts with the RNA nucleobases or backbone are often made with the more variable regions of the domains, including loops. RRMs recognize single-stranded RNA specifically, but the geometric constraints imposed by neighboring elements of secondary structure can be critical for RNA binding, as seen with both U1A and p65. Protein binding frequently results in changes in RNA structure (true of all the examples presented). Changes to protein structure often occur as well. These can be quite subtle (like the loop rigidification seen in U1A) or more dramatic.

Many RRMs are intrinsically relatively weak, non-specific binders of RNA (like the individual domains of SXL). However, specificity and added affinity can be built in through the interaction of multiple RRMs, which are often found in tandem. RRMs can also have extensions that serve to bind double-stranded RNA, such as in U1-70K and p65 (where part of the extension is also used to form the single-stranded binding pocket). To date, 23 solution structures and 26 crystal structures of RRM-RNA interactions have been solved, which have provided a wealth of information about this protein family. Lacking in the structural descriptions of most of these complexes, however, has been a detailed thermodynamic understanding of how different RRM sequences can be accommodated, how RRMs cooperate with each other to modulate binding
affinity and specificity, and how conformational changes in both the RNA and RRM (which can be as dramatic as domain folding) are linked to binding.

**Protein-protein interactions and Protein-RNA binding**

An additional mechanism of manipulating RNA binding affinity (that has more recently gained interest) is the role of protein-protein interactions acting in *trans* (25-27). For RRM-protein interactions that are intrinsically weak and relatively non-specific, the ability to modulate binding mechanisms and outcomes through protein-protein (or RNA-RNA) interactions may provide a broad framework for complex regulation of cellular processes like micro-RNA biogenesis and activity and post-transcriptional mRNA regulation. That examples of this kind of regulation have been discovered is very exciting.

One of the first RNA binding proteins for which such an effect appeared important was the U2B″ protein. In humans, this is a U2 snRNP-specific protein that binds to SLIV of the U2 snRNA. However, unlike U1A, it did not seem to possess intrinsically high affinity and specificity for its RNA target. Rather, early reports suggested that binding by another U2 snRNP-specific protein, U2A’, seemed to be required for the protein to bind with high affinity and specificity. The interaction between U2A’ and U2B″ was determined to occur between the Leucine-rich repeats of U2A’ (N-terminal domain) and RRM1 of U2B″ (28). U2A’ could not be reconstituted into U2 snRNPs in the absence of U2B″, and U2A’ binding to the U2 snRNA could not be detected, suggesting that U2A’ did not bind directly to the snRNA but rather localized to the U2 snRNP through interactions with U2B″ (29, 30).
Figure 5. U2A′-U2B″-SLIV co-crystal structure (PDBID 1A9N). (a) shows the U2A′ protein (lighter blue) bound to α1 and Loop 5 of the RRM. (b) The RNA and RRM are shown without U2A′, in a similar orientation to U1A-SLII in Figure 2.

A co-crystal structure of U2B″ RRM1 in complex with SLIV of the U2 snRNA and the structured N-terminus of U2A′ (Figure 5) was solved (31). This structure showed that U2A′ consists of leucine-rich repeats (LRRs) which form an 8-stranded β-sheet (the six central strands
are oriented in a parallel fashion). This β-sheet surrounds α1 of the RRM. The C-terminal β-hairpin, which caps the LRRs, also interacts with Loop 5 of the RRM. The protein-protein interface is ~2/3 hydrophobic and 1/3 polar.

RNA binding to U2B” RRM1 shares many similarities to binding of U1A to SLII (Figure 2). Overlays of the two RRM structures show almost no deviations in the peptide backbone. Loop 3 of U2B” RRM1 opens up the RNA loop, unstacking the bases. The stacking interactions between RNA nucleobases and the aromatic residues of the RNP motifs are conserved, as are many of the specific interactions with much of the RNA loop. However, the hydrogen bonding network that is important for recognition of the loop-closing basepair is substantially altered to recognize the U-U basepair that characterizes U1 SLIV, as well as the 5’ A of the loop. Additionally, sidechains from β2 and α1 contact the 3’ end of the loop, which is not recognized in the U1A-SLII interaction. Given the similar structures and binding features of U1A and U2B”, one question that remains unanswered is how U2A’ binding to the RRM changes RNA recognition. Further, if the two binding reactions are indeed linked, this suggests a significant change in the U2A’ binding surface. Such a change is not obvious when comparing protein structures.

Interactions between U2A’ and Drosophila SNF have also been observed. As in humans, U2A’ is a U2 snRNP-specific protein. U2A’ interacts with SNF in yeast-2-hybrid assays, which suggests that the proteins interact directly. As with U2B” and human U2A’, the interaction appears to be mediated by the conserved LRRs of U2A’ and RRM1 of SNF (32). Given that SNF is not U2-specific, it is unclear why U2A’ is found only in U2 snRNP. Interestingly, co-immunoprecipitation of the proteins from nuclear extracts is RNA-dependent (33). In a subsequent chapter, we find that these results are consistent with positive linkage between U2A’-
SNF and SNF-U2 snRNA binding. The linkage effect is RNA-dependent, which explains the failure of U2A′ to localize to the U1 snRNP. However, these results also suggest that the purpose of the linkage effect is to localize U2A′ and not to localize SNF/U2B″.

**U1A, U2B″, and SNF in splicing?**

Pre-mRNA splicing was a feature that emerged in early eukaryotes (34). The splicing reaction removes intronic regions from the pre-mRNA and joins the exons. The splicing machinery can be manipulated to change the splicing pattern of transcripts, resulting in the translation of alternative protein isoforms. Thus, it is an important eukaryotic source of molecular diversity. The splicing machinery is complex and dynamic. The core elements are five snRNPs, or small nuclear ribonucleoprotein particles, which each contain a single snRNA and a variable number of associated proteins. Many additional proteins are associated with the spliceosome, increasing its complexity. The classical model for splicing has the snRNPs follow an ordered, sequential assembly on the pre-mRNA, leading first to a transesterification between the branch point and the 5′ splice site, resulting in an intron lariat intermediate. Subsequently, a second transesterification joins the 5′ and 3′ splice sites, releasing the intron lariat.

The splicing machinery is essentially conserved in all modern eukaryotes, although there is variability in the extent to which and how organisms use pre-mRNA splicing. Between *Drosophila* and humans, a major distinguishing splicing feature is the intron length. *Drosophila* introns are too short to be spliced efficiently by human spliceosomes (35). Larger differences are seen in more divergent organisms. While >95% of human genes are spliced, fewer than 5% of genes in *Saccharomyces cerevisiae* contain introns. The 5′ splice site and branch point consensus sequences are also much stronger in *S. cerevisiae* than they are in other organisms. *S.
*pombe* splices a much larger percent of its pre-mRNAs, and consensus sequences are much more variable than in *S. cerevisiae*. Spliced-leader *trans* splicing joins exons from different pre-mRNA transcripts. This splicing phenomenon utilizes the major spliceosome but has been found in only a subset of organisms, including some trypanosomes, platyhelminthes, and nematodes. Thus, while the major spliceosome is an ancient piece of cellular machinery that has in large part been conserved, there is considerable diversity in how it is used.

While U1A/U2B″/SNF-family proteins are extremely well conserved, their cellular function(s) remain elusive. U1A is notorious as an autoantigen in rheumatic diseases, including systemic lupus erythematosus and mixed connective tissue disorder. *In vitro* splicing can be reconstituted in the absence of U1A (36), and the locations of the U1A and U2B″ binding sites on the U1 and U2 snRNAs also present a puzzle for understanding the function of both the proteins and their RNA partners. The U1 snRNP recognizes the 5′ splice site through basepairing of the 5′ end of the snRNA with the pre-mRNA. SLII is a long stem that is directed away from the 5′ splice site and the core of the snRNP (Figure 2b).

U2 snRNP function is substantially more complicated. The U2 snRNA initially engages the branch site of the pre-mRNA by basepairing (this occurs 5′ to SLII of the U2 snRNA). The U2 snRNA also base-pairs at its 5′ end with the U6 snRNA, which is important for pre-catalytic complex formation. Interestingly, pre-mRNA basepairing with snRNAs is generally weak and is facilitated by hosts of proteins. Both the U2 and U6 snRNAs are extensively remodeled to form the catalytically active splicing complex. This includes the melting of SLI of the U2 snRNP to basepair more extensively with the U6 snRNA. However, like U1A in the U1 snRNP, the U2B″ binding site is located far from the ‘business end’ of the U2 snRNA, on the 3′ terminal stemloop. No active roles for U1A and U2B″ in splicing have been found. Given our understanding of the
splicing cycle, the proteins’ locations within the snRNPs call to mind the Enlightenment notion of God as a master clock-maker: always present but never actively involved.

How U1A and U2B″ participate in splicing remains elusive. But they are very highly conserved throughout eukaryotes, and unlike other proteins, remain associated with their respective snRNPs throughout the different stages of splicing (37). In C. elegans, the absence of either U1A or U2B″ results in no apparent phenotypic abnormalities to the animal, but knockout of both is lethal to the worms. The proteins therefore appear to be redundant, but their functions are essential (38). U2A′ is also an essential protein in both Drosophila and C. elegans (32, 38).

SNF was first discovered in Drosophila not because of an interest in splicing but because of its role in sex determination. As a result, its role in Drosophila biology has been studied more extensively than homologous proteins in other organisms. SNF is an essential protein that is incorporated into both the U1 and U2 snRNPs. In the U2 snRNP, it also binds to U2A′. It is known that SNF interacts at a genetic level with SXL (another protein with a critical role in sex determination), and it is thought that this activity defines its role in sex determination. However, the nature of this interaction remains confusing. SNF appears to be important in sex determination through its role in the U1 snRNP (39). Other work suggests a direct protein-protein interaction between SNF and SXL: purified proteins co-immunoprecipitate in the absence of RNA, and further results suggested that SXL binds RRM1 of SNF (40). However, in cells, SNF-SXL co-immunoprecipitation is RNA-dependent, SXL co-immunoprecipitates with the larger U1 snRNP complex, and it continues to co-immunoprecipitate with these complexes in cells with SNF mutant proteins that are deficient in their ability to incorporate into the U1 snRNP (39). Further, some SNF mutations that affect sex determination have no obvious effect on protein snRNP incorporation or SXL association with the U1 snRNP (39). How SNF affects sex
determination and whether it interacts directly with SXL remain ambiguous. As with U1A and U2B”, its essential functions are even less clear.

Regardless of what this family of proteins does in the cell, members exhibit strikingly distinct snRNP partitioning behavior. The goal of this thesis is to understanding the basis of this partitioning behavior. Previous work suggested that differences in the proteins’ binding specificities and differences in interactions with a second protein, U2A’ were important in determining protein partitioning. The opening chapters establish the nature of the differences in RNA binding specificity between human U1A, U2B”, and Drosophila SNF. The following chapters treat the evolution of this protein family in metazoans. The work concludes with an analysis of linkage effects between U2A’-RRM and RRM-RNA interactions. This provides an understanding of protein partitioning behavior that is much more complete than any done to date and explains the unique partitioning of U1A, U2B”, and SNF in cells.
References.


Chapter 2.

*Drosophila* SNF protein binds two RNA hairpins

Sandra G. Williams and Kathleen B. Hall
Abstract.

SNF is a protein that is found in the U1 and U2 snRNPs (small nuclear ribonucleoproteins) of Drosophila. Its mammalian counterparts are two highly homologous proteins, U1A and U2B”. In vivo, these proteins segregate to the U1 and U2 snRNPs, respectively, where they bind distinct RNA hairpins. The RNA binding properties and mechanism of U1A have been studied extensively, but much less is known about SNF and U2B” binding to their RNA targets. By comparing thermodynamic aspects of SNF:RNA interactions with U1A:RNA interactions, we find that SNF binds its RNA targets in a manner that is distinct from that of U1A. In vitro, SNF is able to bind both Drosophila U1 stem-loop II and U2 stem-loop IV with high affinity, although it binds stem-loop II more tightly than it binds stem-loop IV.

Introduction.

RNA recognition motifs (RRMs) are the most abundant RNA binding element in eukaryotic proteins. Currently, more than 6,000 distinct RRMs have been identified, and RRMs are found in a large number of human genes (1). RRMs are involved in processes that include splicing, translation, and RNA trafficking. They are often part of proteins containing multiple functional domains, and in this context, they play key roles in linking RNA binding with RNA regulation. In spite of the high degree of structural similarity between RRMs, they are highly variable with respect to the RNA sequences and structures that they bind, the affinities they have for RNA, and the mechanisms by which they interact with their RNA targets. It is not possible to predict ab initio what RNA sequence an RRM will bind, how tightly it will bind it, or the mechanism by which it will recognize its RNA.
SNF as a member of the snrpA/snrpB2 family

SNF (Sans fille) is a member of the snrpA/snrpB2 protein family. Members of this family are found in all eukaryotes, and in general, proteins in this family contain two RRMs. In some eukaryotes, these proteins are found in pairs; U1A segregates to the U1 snRNP, and U2B" segregates to the U2 snRNP, where they bind highly related but distinct RNA targets within the U1 and U2 snRNA, respectively. In Drosophila, neither U1A nor U2B" is present. Instead, a single protein, SNF, seems to replace both U1A and U2B" and is found in both the U1 and U2 snRNPs.

The function of these proteins in splicing remains unclear, although U1A and U2B" proteins are extremely conserved amongst eukaryotes. U1A binds with high affinity and specificity to stem-loop II (SL II) of U1 snRNA, while U2B" binding is less well characterized to its RNA target—SL IV in U2 snRNA. The U1 snRNP can be functionally reconstituted without the U1A protein (2); analogous experiments with U2B" are not available, although the protein appears to remain in the spliceosome with the U2 snRNP throughout the splicing cycle (3). In C. elegans, the U1A and U2B" proteins are redundant: in the absence of either one, the remaining protein is incorporated into both snRNPs (4), suggesting that each protein can bind both RNA targets. In Drosophila, SNF is found in both snRNPs, where it is assumed to bind to either SL II or SL IV depending on its snRNP context.

A snf knockout is embryonic lethal to the fly, but the reason for this phenotype remains unclear. Mutational experiments have shown that in Drosophila, SNF is not required for recognition of the 5′ splice site by the U1 snRNP (5) and neither is it required for U2 snRNP activity (6). Rather than playing a primary role in splicing events, SNF may interact with
modulators in order to modify splice site selections, thus controlling U1 (and possibly U2) snRNPs at critical times in response to cues from other proteins (6). If this is SNF’s true role in the spliceosome, it indicates that inherent to SNF’s function is a flexibility of interactions. Part of this flexibility includes an ability to bind different RNA targets.

**SNF and its RRM s**

SNF is a 216-amino acid protein with two RRMs connected by a flexible linker of approximately 35 amino acids. As with other RRMs, the two RRMs of SNF adopt an α/β sandwich global fold with two α-helices and an antiparallel, four-stranded β-sheet. The RRM domains of SNF are homologous to those of U1A and U2B″ (Figures 1a, 1b), and although much work has been done to characterize human U1A, much less has been done to understand how human U2B″ binds RNA.

By analogy to human U1A, the N-terminal RRM (RRM1) is likely to be the work-horse of SNF. The canonical site of RNA binding in RRMs is on the surface of the β-sheet, where an RRM recognizes between two and eight nucleotides of single stranded RNA (7). SNF RRM1 contains several highly conserved motifs that are critical for RNA binding in canonical RRMs. These are the ‘RNP motifs’ located on β₁ and β₃, which contain solvent-exposed aromatic amino acids that stack with RNA bases. In SNF, both the RNP motifs of RRM1 are identical to those of both U2B″ and U1A, with the exception of the last residue of RNP2, which is shared with U1A but not U2B″ (Figure 1b). The loop that connects β₂ and β₃ (Loop 3) is critical for RNA recognition, and in SNF, Loop 3 was initially described as a chimera of its human counterparts
A noncanonical feature of the snrpA/snrpB2 family of proteins is the C-terminal tail of RRM1, which contains a third α helix. This tail is important for RNA recognition.

Figure 1. (a) Cartoon of a crystal structure of U1A RRM1 (PDB ID 1URN) (23). Residues are colored such that purple residues are identical in SNF, U1A, and U2B”; red residues are shared by U1A and SNF; blue residues are shared by U2B” and SNF; and black residues are unique to SNF. The C-terminal tail, Loop 3, and β strands are labeled. RNP1 is located on β3, and RNP2 is located on β1. (b) Sequence alignment of human U1A, U2B”, and Drosophila snf proteins. Human U1A residues are colored red, and U2B” residues are colored blue. For the SNF sequence, residues are colored as in 1a. Sequences highlighted in grey correspond to the RNP motifs. The first arrow indicates the end of the RRM1 construct, and the second arrow indicates the beginning of the RRM2 construct. (c). RNAs recognized by human U1A, U2B”, and Drosophila snf proteins. SL II of the U1 snRNA and SLIV of the U2 snRNA are recognized by U1A and U2B”, respectively. SNF recognizes both stem-loops.
The RNA binding partners of U1A, U2B”, and SNF are U1 snRNA stem-loop II (SLII) and U2 snRNA stem-loop IV (SLIV) (Figure 1c). Although these RNAs are quite similar to each other, human U1A RRM1 binds to SLII with subnanomolar affinity but is unable to bind SLIV. As will be described in Chapter 3, U2B” binds SLII and SLIV with equal affinity and binds SLII more weakly than U1A does. Drosophila U1 SLII and U2 SLIV are also quite similar, and SNF must be able to recognize both RNAs.

In the U1A:SLII interface, aromatic residues in the RNP motifs are involved in nonspecific stacking interactions with RNA nucleobases. Specific RNA binding is achieved in part through a hydrogen bonding network between the RNP motif and the 5’ half of the RNA loop. Loop 3 protrudes through the RNA loop, disrupting base stacking as well as forming specific contacts with SLII. Loop 3 contacts the RNA bases as well as other parts of the protein, and the hydrogen bonding network presumably holds both Loop 3 and the RNA in place, opening up the RNA loop so that it is presented as a single strand. Additionally, the conserved TDS sequence in Loop 6 mediates interactions between the C-terminal tail of RRM1 (from the end of $\beta_4$ through the beginning of the third helix) and the CAC nucleotides at the top of the RNA loop. Conspicuously absent are interactions between U1A and the last three nucleotides (UUC) at the 3’ end of the RNA loop. While nucleotides can be inserted in the 3’ end of the loop, deletions are deleterious to binding.

From a cocrystal complex of human U2B” RRM1:SLIV (8), it appears that most protein:RNA interactions are shared with those observed between U1A and SLII. However, this cocrystal structure also demonstrated interactions between Loop 3 of the protein and the UACC nucleotides at the 3’ end of the SLIV loop. The complex was crystallized in the presence of
human U2A′, a member of the leucine-rich repeat protein family. There has been debate over whether this protein is required for U2B″ to bind SLIV. *Drosophila* also contains a U2A′ protein, which associates with SNF in the U2 snRNP (9). Whether this protein is required for SLIV binding by SNF has also been unclear. The effects of U2A′ on RNA binding by this family of proteins are described in Chapters 6 and 7.

For fifteen years, it has been known that the *snrpA/snrpB2* system functions differently in *Drosophila* than it does in humans. Two fundamental questions emerge regarding how the protein homologs bind their biological targets. First, how do the binding mechanisms across the *snrpA/snrpB2* family of proteins differ? More specifically, how do the RRMs achieve distinct binding specificities (the presumed method of protein partitioning)? The second question is whether the different binding mechanisms result in similar biological consequences, regardless of the organism. Given the high degree of evolutionary conservation in this system, the presumption is that the final biological results of binding are similar; that is, Nature has solved one problem in multiple ways.

In this chapter, we compare the thermodynamics of SNF:RNA interactions. The dual ability of SNF to bind both U1 SLII and U2 SLIV allows us to probe whether the protein binds the two RNAs differently, and if so, what constitutes the basis of these differences. We show that the RNA binding properties of SNF are distinct from those of U1A. Furthermore, the binding mechanism changes depending on the RNA target of SNF.

**Materials and Methods.**
**RNA synthesis.** RNA stem-loops for nitrocellulose filter binding assays were enzymatically synthesized with T7 RNA polymerase from DNA oligonucleotides, as described (10)(11). DNA was obtained from IDT (Integrated DNA Technologies). The RNA was internally labeled with [α-32P]UTP and [α-32P]CTP. Labeled RNAs were gel-purified before use in binding assays. The RNA product for dU1 SLII was: 5′—GGCUUGGCCAUUGCACCUCGGCUGAGCC. The RNA product for dU2 SLIV was: 5′—GGCCGGUAUUGCAGUACCGCCGGGUCC. Regions corresponding to the RNA loops are underlined. Experiments for nonspecific binding were done using a 25 nucleotide RNA of random sequence.

For fluorescence binding experiments, 6-carboxyfluorescein (6-FAM) was used to label dSLIV. The labeled RNA was obtained from IDT. The RNA sequence was: 5′—6-FAM-GGGCCCGGUAUUGCAGUACCGCCGGGUCC.

**Protein purification.** All protein constructs were isolated from *E. coli* BL 21 Codon-Plus RP cells (Stratagene) transformed with a plasmid carrying the protein of interest under control of a TAC promoter. The *E. coli* strain contained extra copies of the *argU* and *proL* genes, which encode tRNAs that recognize codons that are more prevalent in the *Drosophila* genome as compared with the *E. coli* genome. The SNF RRM1 construct is composed of residues 1-102. The SNF RRM2 construct is composed of residues 134-216, as well as an additional N-terminal methionine. Cells were grown in LB medium at 37°C and induced in mid-log phase with 1mM IPTG. FL SNF and SNF RRM1 were grown an additional 6 hours at 30°C, whereas SNF RRM2 was grown an additional 4 hours at 37°C. The cells were pelleted and stored at -70°C until lysis.
Containers used in the purification were acid-washed before use, and all solutions were filtered through a 0.45µm cellulose nitrate (CN) filter (Nalgene) to remove RNases.

Two different methods of purification were used, depending on the protein construct. For FL SNF and SNF RRM1, cells were resuspended on ice in 30mM sodium acetate pH 5.3, 200mM NaCl, 2mM EDTA, 8.5% sucrose. A protease inhibitor cocktail (Sigma), PMSF, and DNase II were added prior to lysis. The suspension was French pressed four times. The lysate was spun down at 4°C, 45,000xg in an ultracentrifuge. The supernatant was passed through a 0.22µm cellulose acetate filter and loaded directly onto an SP Sepharose column pre-equilibrated in 50mM Tris pH 7.5. FL SNF was eluted using a gradient running from 275 to 360mM NaCl. RRM1 was eluted with a gradient running from 100-350mM NaCl. Immediately after elution, EDTA and PMSF were added to protein-containing fractions to a final concentration of 5mM and 20µg/mL, respectively. This was done to minimize protein degradation by trace proteases, which was otherwise substantial.

The estimated pI for SNF RRM2 was 6.1, whereas FL SNF and SNF RRM1 are basic. Because purification was done by ion exchange chromatography, SNF RRM2 had to be purified differently. Cells were resuspended in 20mM TrisCl pH 7.5, 20mM NaCl, 2mM EDTA. A protease inhibitor cocktail (Sigma), PMSF, and DNase II were added prior to lysis. The suspension was French pressed four times. The lysate was centrifuged at 25,000xg for 40 minutes. The supernatant was fractionated with 30%, followed by 65% ammonium sulfate. The 65% fractionation was centrifuged at 25,000xg for 45 minutes, and the pellet was resuspended in 50 mM TrisCl pH 7.5. This was dialyzed overnight against 50 mM TrisCl. After centrifugation, the supernatant was loaded onto a CM column pre-equilibrated with 50 mM TrisCl pH 7.5. The column was run to baseline, and SNF RRM2 was eluted (no gradient was used). Fractions
containing SNF RRM2 were collected and loaded onto a Q Sepharose column pre-equilibrated with 20 mM TrisCl pH 7.5. Fractions were eluted with a gradient run from 0 to 300 mM NaCl.

After purification, fractions were concentrated using Vivaspin concentrators, and the protein was exchanged into 10 mM sodium cacodylate, 50 mM KCl pH 7 for storage. The concentration of all proteins was calculated spectrophotometrically. For FL SNF and SNF RRM1, $\varepsilon_{280}=5120\text{M}^{-1}\text{cm}^{-1}$ (both proteins contain four tyrosine residues). For RRM2, the protein concentration was calculated using $\varepsilon_{260}=1152\text{M}^{-1}\text{cm}^{-1}$ as RRM2 contains 8 phenylalanine residues and $\varepsilon_{260,\text{Phe}}=144\text{M}^{-1}\text{cm}^{-1}$. For all three constructs, the protein absorption spectra from 220 to 360nm was identical regardless of whether the spectra were obtained in denaturing conditions (8M urea) or not (0M urea).

**Circular dichroism spectra and unfolding experiments.** All CD spectra were taken using a Jasco J715 instrument. Recordings were made at room temperature. The buffer contained 50 mM KCl, 10 mM sodium cacodylate at pH 7, and a protein concentration of 20 µM. For denaturation studies, the concentration of urea was varied and the mean residue ellipticity (MRE) at 221nm was followed as a function of the urea concentration. Chemical denaturation using guanidine chloride was also performed, but for SNF RRM1, this resulted in a folded baseline that was too small to be used for successful fitting of the data. Data were fit in Scientist (Micromath) using the linear extrapolation method (12) for which the equation:

$$y = \frac{(y_0 + m_D C) \cdot \exp[(m \cdot C - \Delta G_{D,\text{H}_2\text{O}})/RT] + (y_N + m_N C)}{1 + \exp[(m_C - \Delta G_{D,\text{H}_2\text{O}})/RT]}$$
was used. $y$ is the observed MRE; $C$ is the concentration of urea; and $m$ is the slope of the unfolding transition. $y_N$ and $y_D$ are the intercepts of the native and denatured baselines, respectively, and $m_N$ and $m_D$ are the slopes of the native and denatured baselines, respectively. $\Delta G^o_{D,H_2O}$ is the standard unfolding free energy in the absence of denaturant, and $\Delta G_D = \Delta G^o_{D,H_2O} - m \cdot C$.

For refolding experiments, a solution of >10M urea was added to proteins such that the final urea concentration was 8M (9M was used for SNF RRM2). The proteins were allowed to equilibrate for at least 6 hours. The unfolded protein samples were then diluted into lower urea concentrations and allowed to equilibrate overnight, before CD spectra were taken. For plotting the refolding data, MRE values for SNF RRM1 were normalized by a factor of 1.05, and MRE values for FL SNF were normalized by a factor 1.1 to account for differences in protein concentration between the stocks used for unfolding and refolding experiments.

**Filter binding assays.** Nitrocellulose filter binding assays were used to determine standard binding free energies for binding of RNA to different SNF constructs, as described (13, 14). A constant, picomolar concentration of RNA and variable protein concentrations were used. BSA (Roche) was added to a final concentration of 40 µg/mL. All experiments described in the text were done at pH 7. However, additional experiments varying the pH between 6 and 8 show that binding is pH-independent within this range. Solution conditions were otherwise variable and are indicated in the text and figures. All experiments were performed in duplicate, and binding curves were fit to a standard Langmuir isotherm using Scientist (Micromath).
RNA:Protein binding by fluorescence. Fluorescence experiments were done using an SLM 8000 fluorimeter. A circulating water bath was used to control the cuvette temperature. Reaction buffers contained 20 μg/mL of BSA and 10 mM potassium phosphate pH 8. Variable amounts of KCl and MgCl₂ were used, as described in the text. A fixed RNA concentration of 10 nM 6-FAM-dSLIV was used in all experiments.

Acid-washed cuvettes were blocked for one hour with buffer. The excitation and emission wavelengths were set to 490 and 520nm, respectively, and polarizers were set at the magic angle for fluorescence intensity measurements. For protein titrations, fluorescence anisotropy and fluorescence intensity were recorded as functions of the total protein concentration. While both measurements yielded similar binding curves, fluorescence intensity measurements provided data sets that were less noisy and were therefore used for further analysis. The fluorescence intensity of the dye was significantly enhanced when protein bound the RNA. Binding curves were fit to a standard Langmuir isotherm using Scientist.

Results

Protein structure and stability.

Structure. Like U1A, SNF is composed of two RRMs, each of which is approximately 90 amino acids long. To determine whether the secondary structures of U1A and SNF were similar, we measured CD spectra of different SNF constructs. Far-UV CD spectra for the full-length (FL), first domain (RRM1), and second domain (RRM2) constructs of SNF are compared in Figure 2a. These spectra are qualitatively similar to those recorded for U1A protein constructs (16, 17). All constructs yield spectra consistent with folded proteins that contain significant β-sheet character. As with U1A, RRM1 displays substantially greater helical character than RRM2,
as indicated by a more negative mean residue ellipticity at 220nm. This is consistent with the presence of an additional helix as well as with the longer helix lengths found in RRM1. Although the magnitude of the MRE for RRM2 at 221nm is small, this domain is certainly folded, as indicated by the well-dispersed NMR spectrum (Figure 3) and the clear transition observed in CD spectra upon titration with urea. The molar ellipticities of the RRM1 and RRM2 constructs add to yield approximately the molar ellipticity of the full protein (not shown). We attribute the difference to the disordered linker, which is present only in the full-length construct.

**Figure 2** (a). Far UV CD spectra of RRM1, RRM2, and Full-length SNF plotted as mean residue ellipticity (MRE) as a function of wavelength (nm). Figure 2(b). Representative unfolding curves for the chemical denaturation of FL SNF (□), RRM1 (△), and RRM2 (○). The mean residue ellipticity (deg cm² dmol⁻¹ residue⁻¹) at 221nm is plotted as a function of the urea concentration (M). Fits to a two-state unfolding model are shown, and data from refolding experiments are indicated (×). All spectra were recorded in 50 mM KCl, 10 mM sodium cacodylate, pH 7.
**Figure 3.** $^1$H/$^{15}$N HSQC spectra of protein constructs. Overlays of RRM1 (purple) and FL SNF (black) HSQCs are shown on the left. Overlays of RRM2 (purple) and FL SNF (black) are shown on the right.

$^1$H/$^{15}$N HSQC (heteronuclear single quantum coherence) spectra correlate the chemical shifts of the amide $^1$H and $^{15}$N, yielding a single crosspeak in the spectrum for each non-proline residue. HSQCs of the three protein constructs are shown in Figure 3. The spectra are well-dispersed. This is consistent with the domains being folded. Crosspeaks of the spectra from the isolated domains overlay well with the spectrum of the full-length protein, which is consistent with the domains having largely the same structure in the full-length protein as in the truncated constructs. Additional residues in the full-length spectrum that do not have corresponding peaks in the isolated domains correspond to residues from the interdomain linker. These residues are clustered in the $^1$H dimension with chemical shifts between 8 and 8.5 ppm, consistent with the linker being unstructured in solution.

After much of this work had been completed, solution structures of SNF RRM1 and RRM2 were solved and showed that the RRM s adopt largely identical structures to their U1A counterparts (27).
Stability. Chemical denaturation of RRM1, RRM2, and FL SNF with urea is fully reversible and can be fit to a two-state model of unfolding (Figure 2b). RRM 1 has an unfolding free energy ($\Delta G^\circ_{D,H_2O}$) of 3.5 ($\pm$0.3) kcal/mol. RRM2 has a $\Delta G^\circ_{D,H_2O}$ of 4.8 ($\pm$0.6) kcal/mol (Table 1). Both of these values are significantly reduced from those of U1A. In U1A, RRM1 has a $\Delta G^\circ_{D,H_2O}$ of 9.4 ($\pm$0.5) kcal/mol (18) and RRM2 has a $\Delta G^\circ_{D,H_2O}$ of 8.3 ($\pm$0.8) kcal/mol (17). Clearly, both the RNA recognition motifs of SNF are significantly destabilized as compared with their U1A counterparts. While it was possible to fit the unfolding of FL SNF according to a 2-state model, the unfolding free energies of RRM1 and RRM2 are similar enough that it is not possible to distinguish whether or not the domains unfold independently in the full-length protein.

<table>
<thead>
<tr>
<th></th>
<th>$\Delta G^\circ_{D,H_2O}$ (kcal/mol)</th>
<th>$m_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRM1</td>
<td>3.5 ($\pm$0.3)</td>
<td>0.85 ($\pm$0.11)</td>
</tr>
<tr>
<td>RRM2</td>
<td>4.8 ($\pm$0.6)</td>
<td>0.92 ($\pm$0.07)</td>
</tr>
</tbody>
</table>

Table 1. SNF unfolding free energies. $\Delta G^\circ_{D,H_2O}$ is the standard unfolding free energy in the absence of denaturant as determined by linear extrapolation. $m_D$ is the slope of $\Delta G_D$ vs. [Denaturant] and is one of the terms of the Linear Extrapolation Model. All experiments were done in duplicate, and the estimates from non-linear least squares fits were averaged to give the values shown. Errors for $\Delta G^\circ_{D,H_2O}$ and $m_D$ are estimated from the non-linear least squares fit.

RNA binding.

Human U1A only binds to SLII with high affinity. In vitro, human U2B″ binds to both SLII and SLIV (described in the following chapter). While studies from Drosophila genetics experiments suggested that SNF could bind both U1 SLII and U2 SLIV, it was not clear from the experiments whether SNF required U2A’ to bind SLIV. Certainly, these experiments provided no insight into the affinity SNF has for its RNA targets. Additionally, it is possible that changes to
the RNA, as well as to the protein, contribute to SNF’s ability to recognize two RNA targets. In order to address these questions, we performed nitrocellulose filter binding experiments with full-length SNF and $^{32}$P-labelled RNA hairpins. The hairpins contained *Drosophila* loop sequences and loop-closing base pairs and were synthesized with eight or nine base-pair stems for binding assays. Data from protein titrations were fit assuming a 1:1 stoichiometry, which was confirmed for SNF:SLIV by fluorescence experiments in which the RNA concentration was well above the apparent $K_d$ (not shown). Apparent binding constants ($K_d$) for SNF binding *Drosophila* SLII and SLIV are given in Table 2 at two salt concentrations. Representative titrations are shown in Figure 4. Under both conditions, SNF binds SLII with a small loss of affinity as compared to U1A. With experiments done in 250 mM KCl, SNF bound SLIV with an affinity of $7.7 \pm 3.0 \times 10^{-8}$ M, which represents a loss of affinity compared to SLII but is still fairly tight. Binding to a random 25mer was very weak (micromolar at best; see Figure 4). It is evident that SNF can bind specifically to dSLIV in the absence of U2A', although this occurs with weaker affinity than binding to SLII. Binding to stem-loops with human loop and loop-closing basepair sequences was indistinguishable to binding the *Drosophila* sequences (not shown).

**Figure 4.** Titrations for FL SNF binding different RNAs. Nitrocellulose filter binding experiments were done to determine binding constants for FL SNF binding to SLII (■), SLIV (●), or a random 25-mer sequence (×). Fits to a model of single-site binding are shown for SNF binding to SLII and SLIV. Experiments were done at 22°C in 250 mM KCl, 10 mM cacodylate, 1 mM MgCl$_2$ at pH 7.
Salt dependence of RNA binding. When considering mechanisms for RNA:protein interactions, it is impossible to do so without considering electrostatic contributions to binding. The association of cations with highly charged nucleic acids reduces the unfavorable charge density of the polyphosphate backbone through counterion condensation and screening effects. Upon ligand binding, some of the associated cations are displaced to allow association of the ligand. This results in a salt dependence of the apparent binding constant, $K_{\text{obs}}$. For an oligopeptide binding to a nucleic acid, analysis of the salt dependence of binding can yield the number of cations released from the nucleic acid and thus provide mechanistic information about binding.

For instance, cation release from the RNA is entropically favored, and in some cases it has been shown that a binding reaction can be driven by cation release (19). The salt dependence of these interactions also provides a framework for understanding the extent to which electrostatics contribute to interaction specificity between the protein and RNA target (20). To this end, we wanted to understand whether electrostatic contributions to binding were different for the U1A:SLII interaction as compared with the SNF:SLII interaction. Additionally, we wanted to know whether electrostatics contributed to the difference in binding affinity for SNF binding to SLII as compared with SLIV.

Analysis of protein:nucleic acid interactions is complicated by a number of factors, including preferential interactions of cations, anions, protons, and water with the protein. The resulting salt dependence of the interaction is therefore described by:

\[
Stem-Loop & 100\text{mM KCl} & 250\text{mM KCl} \\
\hline
II & 1.6 \pm 0.8 \times 10^{-10} & 3.3 \pm 0.2 \times 10^{-9} \\
IV & 5.6 \pm 1.6 \times 10^{-9} & 7.7 \pm 3.0 \times 10^{-8} \\
\hline
\]

Table 2. Apparent binding constants ($K_d$) for SNF:RNA interactions.
\[ \frac{\partial \log K_{obs}}{\partial \log [MX]} = -(\Delta c + \Delta a) + 2[X] \cdot \Delta w/[H_2O] \quad \text{(Eqn. 1)} \]

where \( \Delta c \), \( \Delta a \), and \( \Delta w \) represent the net gain or loss of cations (\( \Delta c \)), anions (\( \Delta a \)), and water (\( \Delta w \)) (21). Unless the salt concentration is quite high (> 0.5M) preferential hydration effects are expected to be small compared with the ion terms, simplifying the analysis, and the salt dependence of binding can be understood in terms of net cation and anion release from the interacting species. In plotting log(\( K_{obs} \)) as a function of the log of the salt concentration, the slope yields the net release (slope < 0) or uptake (slope > 0) of ions upon protein:RNA association.

We measured the association constants (\( K_{obs} \)) of SNF:RNA interactions at multiple concentrations of KCl (Figure 5). When U1A RRM1 binds SLII, there is a net release of 6.7 (± 1.1) ions. These data were obtained from titrations done in NaCl. For U1A RRM1, binding is equivalent in NaCl or KCl; the binding affinity is independent of the nature of the monovalent cation. SNF binding to SLII results in a net release of 5.7 (± 0.2) ions; these results are the same within error to those observed for U1A:SLII. In contrast, SLIV binds to SNF with a net release of 4.0 (± 0.2) ions, 1.7 ions fewer than were released by SLII binding. This suggests that the mechanism of binding SLII and SLIV by SNF is different.
If the assumption is made that SLIV and SLII bind to the same site on the protein, it follows that the anion effects can be considered equivalent for the binding of SNF to both SLII and SLIV. If the specificity of the protein for SLII over SLIV is independent of salt, this would indicate that electrostatics do not contribute to binding specificity (20). However, our analysis shows that for SNF, the salt dependence of binding its two partners is indeed different; electrostatic interactions affect the specificity of binding.

Temperature dependence of RNA binding. The temperature dependence of the observed binding constants provides information about the standard enthalpy and entropy change associated with the binding reaction. In addition, a non-linear van’t Hoff plot indicates an apparent heat capacity change that can be described by the following equation:

\[
\ln(K_{obs}) = (\Delta C_{P,obs}/R) \cdot [(T_H/T) - \ln(T_S/T) - 1] \quad (19) \quad \text{(Eqn. 2)}
\]

where the enthalpy and entropy are considered to be temperature-dependent. \(T_H\) and \(T_S\) are the temperatures at which the enthalpy and entropy of complex formation, \(\Delta H^o\) and \(\Delta S^o\), are 0.
\( \Delta C_{P,\text{obs}} \) is the apparent heat capacity change that characterizes the reaction. There are a number of reasons why an association may indicate an apparent heat capacity change, but in general an apparent \( \Delta C_P \) accompanying a protein:ligand interaction is interpreted as resulting from hydrophobic surface burial or a coupled conformational change, either of the ligand, protein, or both.

The van’t Hoff plot of U1A binding to SLII shows significant deviation from linearity, indicative of a large apparent \( \Delta C_P \). The binding event occurs with significant conformational changes in both the RNA and the protein: base-stacking is disrupted in the RNA, and Loop 3 of the protein is rigidified as it protrudes through the RNA loop and contacts RNA nucleobases. These changes likely contribute to the apparent \( \Delta C_P \). Given the similarities of protein and RNA structures, we thought that binding would occur with similar changes to the RNA and protein, resulting in an apparent \( \Delta C_P \). However, we wished to know whether the apparent heat capacity change was similar in magnitude for both RNA targets, and whether this was similar to the \( \Delta C_P \) observed for the U1A:SLII interaction. Along with comparing the \( \Delta C_P \)s for these different interactions, comparing the characteristic temperatures \( T_H \) and \( T_S \) could provide information about mechanistic differences of binding.

We measured the temperature dependence of \( K_{\text{obs}} \) for SNF:RNA interactions at a variety of salt concentrations (Figure 6, tables 3,4). We varied the salt concentration between experiments because of the difficulties in accurately measuring \( K_{\text{obs}} \) for both RNAs over the assessed temperature range (4 to 37\(^\circ\) C). In particular, binding of SLII to SNF was too tight to measure the \( K_{\text{obs}} \) accurately at low temperatures, especially when the salt concentration was below 300mM KCl.
Figure 6. Van’t Hoff plots for SNF and U1A binding to target RNAs. (a) Human U1A binding to SLII (●) and SNF binding to SLII (■). (b) SNF binding to SLIV in 100 mM KCl (▲) and in 250 mM KCl (◆). Binding experiments for U1A were done in 200 mM NaCl, 10 mM cacodylate, and 1 mM MgCl₂. Binding experiments for SNF were done in 10 mM cacodylate, 1 mM MgCl₂ pH 7, at the described concentrations of potassium chloride. Errors for K_{obs} in SNF experiments are estimated from the standard deviation of at least two separate results.

<table>
<thead>
<tr>
<th>Protein:RNA</th>
<th>[KCl] (mM)</th>
<th>∆Hº (kcal/mol)</th>
<th>∆Sº (cal/mol·K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNF:SLII</td>
<td>300</td>
<td>-31.1 (±2.0)</td>
<td>-67.3 (±6.9)</td>
</tr>
<tr>
<td>SNF:SLIV</td>
<td>250</td>
<td>-23.4 (±2.2)</td>
<td>-47.6 (±7.4)</td>
</tr>
<tr>
<td>SNF:SLIV†</td>
<td>250</td>
<td>-23.4 (±3.5)</td>
<td>-48.9 (±11.7)</td>
</tr>
<tr>
<td>SNF:SLIV</td>
<td>300</td>
<td>-24.4 (±1.9)</td>
<td>-53.2 (±6.5)</td>
</tr>
</tbody>
</table>

Table 3. Enthalpy and entropy of binding for data fit to the van’t Hoff equation. †Determined from fluorescence experiments rather than filter binding experiments.

<table>
<thead>
<tr>
<th>Protein:RNA</th>
<th>[KCl]</th>
<th>∆C_{P,obs} (kcal/mol)</th>
<th>T_H(K)</th>
<th>T_S(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1A:SLII</td>
<td>200</td>
<td>-3.1 (±0.4)</td>
<td>284 (±5)</td>
<td>288 (±2)</td>
</tr>
<tr>
<td>SNF:SLIV</td>
<td>100</td>
<td>-1.37 (±0.42)</td>
<td>281.7 (±3.6)</td>
<td>290.2 (±1.6)</td>
</tr>
</tbody>
</table>

Table 4. Fits of non-linear van’t Hoff plots. Apparent heat capacity change (∆C_{P,obs}), T_H, and T_S values for interactions that showed curvature in the van’t Hoff plots. All results for U1A are from Williams and Hall, 1996.
The results of these experiments were surprising. Unlike U1A binding to SLII, the van’t Hoff plot for SNF binding to SLII was linear and could thus be analyzed without invoking an apparent $\Delta C_p$. More surprising was that for SLIV, the presence of a heat capacity change was found to be salt-dependent. Whereas in 100 mM KCl, the van’t Hoff plot was nonlinear and could be fit to Eqn. 2 to yield a $\Delta C_{P,\text{obs}}$ of -1.4 kcal/mol, at 250 mM KCl and 300 mM KCl, the apparent heat capacity change vanished. While the $\Delta H^\circ$ of complex formation was the same at these higher salt concentrations (~24 kcal/mol), $\Delta S^\circ$ was more variable. The data also show that for SNF binding to both RNAs at higher salt concentrations, the reactions are enthalpy-driven. For SNF binding to SLIV at 100 mM KCl, the reactions are enthalpy-driven above 17°C but entropy-driven below 9°C. We conclude that SNF uses different binding mechanisms to bind these two RNAs.

**Comparison of FL SNF and RRM1 binding.** In U1A, the first RRM is sufficient for full RNA binding. The second RRM does not bind RNA and does not contribute to the binding affinity of the protein:RNA interaction. We did binding experiments with SNF RRM1 to see if this domain was sufficient for full RNA binding. Like U1A, SNF RRM2 does not bind either SLII or SLIV on its own (data not shown). Unlike U1A, loss of the linker and RRM2 of SNF is significantly detrimental to SNF’s ability to bind both its targets. Table 5 gives binding affinities ($K_{\text{obs}}$) and free energies ($\Delta G^\circ$) for full-length and RRM1 of SNF and U1A binding to their respective RNAs. In 250 mM KCl, SNF truncation results in a 30-fold decrease in binding affinity for SLII and a 3-fold decrease in binding affinity for SLIV.

Interestingly, titrating FL SNF into SLIV labeled with 6-carboxyfluorescein (6-FAM) at the 5’ end of the stem resulted in fluorescence enhancement upon binding, which could be
monitored to measure binding affinities. The results of such measurements yielded equivalent
binding constants to those obtained from nitrocellulose filter binding experiments under similar
conditions. The observation that binding results in fluorescence enhancement also may provide
insights into the increased affinity of FL SNF for RNA.

<table>
<thead>
<tr>
<th>Protein:RNA</th>
<th>$K_{\text{obs, FL}}$ (M)</th>
<th>$\Delta G_{\text{FL}}^{\circ}$ (kcal/mol)</th>
<th>$K_{\text{obs, RRM1}}$ (M)</th>
<th>$\Delta G_{\text{RRM1}}^{\circ}$ (kcal/mol)</th>
<th>$K_{\text{obs, RRM1}}/K_{\text{obs, FL}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1A:SLII</td>
<td>4 (±3) · 10^{-10}</td>
<td>-13 (±0.4)</td>
<td>4 (±3) · 10^{-10}</td>
<td>-13 (±0.4)</td>
<td>1</td>
</tr>
<tr>
<td>SNF:SL II</td>
<td>7 (±6) · 10^{-10}</td>
<td>-12 (±0.5)</td>
<td>2.18 (±0.16) · 10^{-8}</td>
<td>-10.3 (±0.1)</td>
<td>30</td>
</tr>
<tr>
<td>SNF:SL IV</td>
<td>7.7 (±3.0) · 10^{-8}</td>
<td>-9.6 (±0.2)</td>
<td>2.33 (±0.66) · 10^{-7}</td>
<td>-8.95 (±0.17)</td>
<td>3</td>
</tr>
<tr>
<td>SNF:SL IV 100mM KCl</td>
<td>5.6 (±1.6) · 10^{-9}</td>
<td>-11 (±0.2)</td>
<td>3.33 (±0.35) · 10^{-8}†</td>
<td>-10.1 (±0.1)</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Table 5. Comparison of FL SNF vs RRM 1 binding free energies. Experiments were done in
250mM KCl (unless noted), 10mM cacodylate, 1mM MgCl$_2$ pH 7. †Determined from
fluorescence experiments rather than filter binding experiments. Unless noted, errors were
determined as the standard deviation of at least two data sets. This generally gave a higher
estimation of uncertainty than was achieved with error propagation analysis. ‡Error was
determined from propagation of uncertainty.

Fluorescence enhancement upon binding is both salt-dependent and dependent on the
protein construct length. A salt titration was performed to assess the effect of salt upon
fluorescence enhancement (Figure 7). Experiments were performed at constant protein and
RNA concentrations of 10µM protein and 10nM RNA. At this protein concentration, FL SNF is fully bound to SLIV at all salt concentrations tested, and RRM1 is fully bound (>95%) to SLIV up to at least 300mM KCl. The fluorescence enhancement of the RNA upon protein binding is shifted to lower salt concentrations for RRM1 as compared to FLSNF, but the shape of the salt dependence is similar. Given that the dye is attached to the 5′ end of the RNA stem, these data suggest that RRM2 or the linker of the full-length protein is interacting with the RNA stem; this interaction is highly salt-dependent. Alternatively, the linker or RRM2 may be interacting with RRM1 to facilitate specific interactions between RRM1 and the RNA stem. The difference in RNA binding affinities for FL SNF as compared with RRM1 underscores a major mechanistic difference in binding when SNF is compared with U1A.

\(^{1}\)H/\(^{15}\)N HSQC spectra of full-length SNF in the presence and absence of excess SLII RNA show dramatic changes to the RRM1 crosspeaks (most of them disappear upon the addition of RNA, consistent with binding and slower tumbling of the entire domain upon association with the RNA), while changes to the RRM2 crosspeaks are essentially absent (Figure 8). Some of the linker peaks also appear to be affected by the presence of the RNA. This supports the view that RRM2 does not contribute to RNA binding.

RNA constructs of both SLII and SLIV were made that varied in stem length between 6 and 12 basepairs. The binding affinity for these RNAs was identical to that of binding to the 9 basepair stem RNAs used in the rest of the study. Shorter stems were not assessed, as these were likely to have compromised hairpin stability. Fluorescence enhancement of fluorescein-labeled SLIV RNAs with 6 and 12 basepair stems was also assessed. While RNA fluorescence enhancement was assessed at a single salt concentration, these RNAs showed identical enhancement properties to that of the 9 basepair hairpin used in other studies. These results
suggest that the interaction between the full-length protein and the RNA that causes the fluorescein fluorescence enhancement (and improved binding compared to RRM1) is nonspecific.

Figure 8. $^1$H/$^15$N HSQC spectra of $^{15}$N-labeled FL SNF. In black is free SNF; in pink is a spectrum taken in the presence of unlabeled SLII RNA.

A comparison between the salt-dependence of SNF RRM1 and FL SNF is provided in Chapter 4. The results show a substantial difference in salt dependence between the two protein constructs (RRM1 has a substantially weaker salt-dependence of binding for the two RNAs), which is consistent with nonspecific electrostatic contributions to binding between the RNA stem and the positively charged linker. Although the binding affinities of FL SNF and U1A for SLII are quite similar, SNF RRM1 is not able to achieve a high binding affinity for both of its targets without compensating interactions from other parts of the protein. Further, while the salt
dependence of U1A RRM1 and FL SNF for SLII were very similar, the salt dependence of U1A RRM1 and SNF RRM1 for SLII is quite different, again suggesting different electrostatic contributions to RNA recognition by the two RRMs.

Substituting Phenylalanine 98 with Tryptophan results in a protein with identical RNA binding characteristics to SNF (when substituted into either RRM1 or FL SNF). This site is at the beginning of the interdomain linker, C-terminal to the third α helix. The fluorescence emission spectrum is both quenched and red-shifted upon the addition of RNA (Figure 9). While this site is most likely too far from the end of the RNA stem to contribute to the observed fluorescein fluorescence enhancement, its sensitivity to RNA shows that residues within the linker either change conformation upon RNA binding or can interact with the RNA.

![Figure 9](image.png)

**Figure 9.** Emission spectrum of SNF F98W in the absence (blue) or presence (pink) of 1.1x SLII. The emission maxima are indicated. The site of the mutation is also indicated on the RRM structure. Emission spectra using FL SNF F98W or SLIV are identical.

**Discussion**

*Structural and Thermodynamic Properties of SNF*
The far UV CD spectra of the two RRM domains of SNF are almost identical to those of U1A, indicating that both proteins have similar secondary structure. Furthermore, the unfolding curves of both domains fit well to a two-state model, characteristic of many globular proteins. In spite of these similarities, the thermodynamic stability of the first and second domains is dramatically reduced in SNF; RRM2 is destabilized by 3.5 kcal/mol and RRM1 is destabilized by 6 kcal/mol when compared with U1A. This suggests there was the evolutionary capacity for SNF to be more stable than it is. The marginal stability of SNF RRM1 is therefore surprising, but it is not unique. The folded state of globular proteins is generally only slightly stabilized compared to the unfolded state (folding free energies typically range from 5-10 kcal/mol) (22). Much more stable proteins have been engineered, suggesting that there is little evolutionary pressure for highly stable proteins and that perhaps the converse is true; that marginal thermodynamic stability is in fact a biologically important characteristic of proteins. However, the relationship between protein instability and protein function has not been well-defined.

An unresolved structural question is the presence or absence of a third α helix at the C-terminus of SNF RRM1. In U1A, the C-terminal tail and its helix contribute directly and indirectly to substrate discrimination and specificity (18, 23). This helix is not a feature of canonical RRMs, and although the U1A sequence is similar to that of SNF, it is not identical. The sensitivity of far-UV CD to helical content renders CD suitable for addressing whether the third helix is present. The similarity between the CD spectra of RRM1 of U1A and SNF suggests that this helix does exist in SNF.

SNF:RNA interactions
A comparison of the binding partners of SNF and U1A reveals that the differences between *Drosophila* and human SLII are minimal. Both loops close with a C:G base pair, and the single-stranded AUUGCAC loop RNA sequence recognized in the U1A:SLII complex (13) (23)(24) is identical between the human and *Drosophila* RNAs. In the U1A:SLII complex, the remaining 3 residues at the 3’ end of the RNA loop extend away from the protein surface and do not make any contacts with the protein (23). They are important in that they act as a spacer and therefore must not form base-pairs, but otherwise the sequence of these residues is unimportant (25). Individual mutations of the two discrepant nucleotides to their *Drosophila* counterparts result in no change in binding affinity for U1A (24).

Because SLII binding to SNF was very tight, the van’t Hoff analysis was done at a higher salt concentration (300mM KCl) than had been used in the original U1A van’t Hoff analysis (200mM NaCl). The comparison between SNF and U1A binding to SLII is made with this caveat. At 200 mM NaCl, the van’t Hoff plot of U1A binding SLII clearly deviates from linearity and shows an apparent $\Delta C_P$. Despite the similarity of both proteins and SLII RNAs, SNF binding to SLII is linear at 300 mM KCl: there is no $\Delta C_{P,\text{obs}}$.

In the U1A:SLII interaction, conformational transitions occur in both the RNA and protein. In the RNA, the loop must be splayed open by part of the protein. Along with the disruption of stacking interactions throughout the loop, this is likely the basis of changes in the RNA loop structure that occur upon protein binding and contribute to the $\Delta C_{P,\text{obs}}$ (13, 26). Given that dSLII and hSLII are essentially identical, the disparity in $\Delta C_{P,\text{obs}}$ is unlikely to result from differences intrinsic to the RNA. Additionally, the U1A:hSLII interaction results in changes to the protein structure. Notably, Loop 3 becomes confined through its interactions with the RNA. While the exact nature of the SNF:RNA interface and hydrogen bonding network is unknown, it
is likely that binding to either RNA target results in conformational changes of both the RNA and protein. It seems unlikely that the RNA:protein interaction would result in significant hydrophobic surface burial of the human protein but not of SNF. It also seems unlikely that the SNF:dSLII interaction occurs through a ligand docking mechanism, but such a mechanism is consistent with the results. It is possible that changes in the protein and RNA compensate for each other so that there is no $\Delta C_{P, \text{obs}}$.

A large $\Delta C_{P, \text{obs}}$ characterized the SNF:SLIV interaction at lower salt concentrations. The large salt dependence of $\Delta C_{P, \text{obs}}$ for SLIV was surprising. This apparent salt dependence could result from a shift of the van’t Hoff plot such that over the temperature range measured, the curve appears linear when in fact it is not. This could also account for the apparent linearity of the van’t Hoff plot observed for SNF binding to SLII. Alternatively, the salt dependence of the $\Delta C_{P, \text{obs}}$ for the SNF:SLIV interaction may reflect a large electrostatic contribution to the observed $\Delta C_P$. It is possible that in this system, the linker between RRM1 and RRM2 interacts with the RNA. The linker contains a high density of positive charges that could contribute to a large electrostatic interaction, and we find that removing the linker and RRM2 results in significant loss of binding affinity. For SLIV, we can compare the loss of binding affinity at 100 mM KCl and 250 mM KCl, and we find that at the lower salt concentration, there is a 6-fold loss of binding affinity upon truncation, whereas in 250 mM KCl, there is only a 3-fold loss of affinity. Truncation also leads to a loss of discrimination between RNAs: $\Delta \Delta G^\circ = \Delta G^\circ_{(\text{dSLII})} - \Delta G^\circ_{(\text{dSLIV})} = -2 \text{ kcal/mol}$ for SNF, but only $-0.6 \text{ kcal/mol}$ for RRM1. Fluorescence enhancement upon binding is also diminished with protein truncation. Fluorescence enhancement upon binding is highly salt-dependent for both FL SNF and SNF RRM1, but fluorescence enhancement decreases at much lower salt concentrations for RRM1. We suggest that the highly charged linker binds the
phosphate backbone of the RNA stem, contributing to an overall increase in binding affinity. Furthermore, we propose that these additional interactions are necessary to compensate for the compromised binding that SNF RRM1 exhibits for dSLII as compared with the affinity that U1A RRM1 has for SLII.

One of the most striking differences between U1A and SNF is that while U1A does not detectably bind SLIV, SNF binds SLIV with nanomolar affinity. At 300 mM KCl, van’t Hoff plots for SNF binding both SLII and SLIV are linear, which makes it possible to compare the entropic and enthalpic contributions to binding. Binding of the two RNAs is distinguished by \( \Delta \Delta S^o = S^o_{\text{SNF:SLII}} - S^o_{\text{SNF:SLIV}} = -14.1 (\pm 9.5) \text{ cal/mol-K} \) (\( T \Delta \Delta S^o \) is -4.2 kcal/mol at 298K), whereas \( \Delta \Delta H^o = H^o_{\text{SNF:SLII}} - H^o_{\text{SNF:SLIV}} = -6.7 (\pm 2.8) \text{ kcal/mol} \). Binding of both RNA targets is enthalpically driven, and the additional binding affinity of SNF for SLII vs SLIV appears to be completely enthalpic in origin. While there is a considerable entropic penalty for binding both RNAs, at 22°C the penalty associated with binding SLII is larger than that of binding SLIV by +4 kcal/mol. More ions are released upon SLII binding than upon SLIV binding, so the difference in this entropic penalty cannot be attributed to differential ion release.

Enthalpic contributions to binding, including hydrogen bonding, salt bridges, and stacking interactions between bases and amino acid side chains, appear to be more favorable for the SNF:SLII interaction. But of more interest, enthalpic and entropic contributions to binding appear to favor different RNAs. This presents intriguing possibilities for tweaking the binding system. In biological contexts, SNF binds at least one other protein partner, U2A′. Although it is unclear how U2A′ alters SNF’s interactions with RNA, studies of the mammalian U2A′:U2B″ interaction suggest that U2A′ alters the specificity of U2B″ for its RNA target. Shifting the
balance between entropic and enthalpic effects may well provide a mechanism for driving specificity, and this may occur in the context of protein modulators. The effects of U2A’ binding on SNF function will be discussed further in Chapter 6.
References


Chapter 3.

Human U2B” Protein Binding to snRNA Stemloops
Human U2B⁺ protein binding to snRNA stemloops

Sandra G. Williams, Kathleen B. Hall *

Department of Biochemistry and Molecular Biophysics, Washington University Medical School, St Louis, MO 63110, United States

**A R T I C L E   I N F O**

Article history:
Received 3 April 2011
Received in revised form 9 May 2011
Accepted 9 May 2011
Available online 16 May 2011

Keywords:
RRM
RNA binding protein
U2B⁺
Protein:RNA thermodynamics

**A B S T R A C T**

The human U2B⁺ protein is one of the unique proteins that comprise the U2 snRNP, but it is also a representative of the U1A/U2B⁺ protein family. In the U2 snRNP, it is bound to Stem-Loop IV (SLIV) of the U2 snRNA. We find that in vitro it binds not only to human SLIV, but also to Stem-Loop II (SLII) from human U1 snRNA and to Drosophila U2 snRNA SLIV. The thermodynamics of these binding interactions show a striking similarity, leading to the conclusion that U2B⁺ has a relaxed specificity for its RNA targets. The binding properties of U2B⁺ are distinct from those of human U1A and of Drosophila SNF, despite its high homology to those proteins, and so provide important new information on how this protein family has modulated its target preferences.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

The spliceosomal proteins U1A and U2B⁺ are highly conserved in eukaryotes, where they are components of the U1 and U2 snRNPs, respectively. In the snRNPs, U1A protein is thought to bind exclusively to stem-loop II (SLII) of the U1 snRNA and U2B⁺ to stem-loop IV (SLIV) of the U2 snRNA. These RNA hairpins are highly conserved, and their loop sequences are very similar to each other. Although they are components of snRNPs, the roles of U1A and U2B⁺ in splicing remain unclear, and in fact, in vitro snRNP reconstitution in the absence of U1A has no effect on splicing [1]. Other experiments using mutations of the fly homologue, SNF, that exclude the protein from either the U1 or U2 snRNP resulted in relatively mild phenotypic consequences [2,3]. In contrast, knocking out both U1A and U2B⁺ in C. elegans is embryonic-lethal, as is the SNF knockout in Drosophila [4,5]. There is certainly a possibility that these proteins have alternative functions outside of the snRNPs.

U1A, U2B⁺, and SNF consist of two RNA recognition motifs (RRMs) connected by a variable, flexible linker. The RRM is the most commonly used RNA binding domain in eukaryotes (Maris et al., 2005, FEBS J, 272, 2118-31), and can be identified by two amino acid sequences (RNP1 and RNP2) that are located in two of the four β-strands on its β-sheet (Ghetti et al., 1989, FEBS Lett, 257, 373-6). The canonical view of RNA–RRM interactions is that single-stranded RNA binds to the β-sheet surface. Favorable electrostatic interactions, hydrogen bonding, and stacking between RNA bases and aromatic residues located in the RNP motifs are regarded as the predominant determinants of RNA binding (Clery et al., 2008, Curr Opin Struct Biol, 18, 290-8). U1A uses its N-terminal RRM to bind its in vivo U1 snRNA target, Stem-Loop II (SLII), with very high affinity and specificity [6,7]. There are crystal and solution structures of U1A RRM1 [8,9], and a co-crystal structure of U1A RRM1 bound to SLII [10] shows how the loop of the RNA hairpin is positioned on the surface of the β-sheet. A co-crystal of U2B⁺ RRM1 bound to both SLIV and U2A⁺ (an auxiliary protein) positions the RRM in the middle of the complex, with the RNA bound on the β-sheet surface and U2A⁺ wrapping around the opposite face of the RRM, predominantly making contacts with RRM α₁ [11]. There is a solution structure of SNF RRM1 [12]. Not surprisingly, the structures of these three RRMs are similar to each other and for U1A, the structures of the free and bound proteins are also similar (Nagai et al., 1990, Nature, 348, 515-20).

1.1. U2A⁺ in U1 snRNP

The U1A/U2B⁺ family of proteins provides a valuable opportunity to understand determinants of RNA:protein affinity and specificity. RRM1 of U1A is ~75% identical to the N-terminal RRMs of U2B⁺ and the Drosophila homologue, SNF. While SNF binds to both U1 snRNA SLII and U2 snRNA SLIV [13], there has been significant debate regarding how (and how well) U2B⁺ binds to its target, U2 snRNA SLIV. Some data report that U2B⁺ binds both SLII and SLIV [14], and there are conflicting studies on whether the U2A⁺ protein, which is present in the U2 snRNP and which binds U2B⁺, is required for U2B⁺ to bind SLIV [15]. Regardless, the studies do suggest that RNA binding by U2B⁺ is much weaker than binding of SLII by U1A.

The co-crystal structure of U2B⁺:SLIV showed that many of the interactions between U2B⁺ and SLIV were also present in the earlier U1A:SLII co-crystal structure [11]. Since both the RNA loops contain identical 5′ sequences (human SLII: A1UUGCA6CUCC and human SLIV A1UUGCA6GUAGCC) and the RNP sequences of the proteins are identical, preservation of these contacts is not unexpected. Stacking of the nucleobases with Tyr10 in RNP2 (U1) and Phe53 in RNP1 (U2B⁺ numbering) occurs in both complexes. Despite the
phylogenetic conservation of A9 on the 3’ side of the SLIV RNA loop (and its absence from stem-loop II sequences), the U2B’ co-crystal showed no interactions between the protein and A9. Within the 3’ U1A1CC sequence, U8 and C11 do pack against the VALKT amino acids of U2B’ (i.e., the protein’s biological functions is that human U2B’ has different RNA binding preferences in vitro which could not be anticipated by the apparent similarities in their respective cocrystal structures.

The U1A:RNA interaction is characterized by complicated thermodynamics. More specifically, the U1A:SLII interaction has a large apparent heat capacity (ΔCp,obs) of ~3 kcal/mol, and its enthalpy and entropy are both temperature-dependent [7,16]. Interpretation of ΔCp,obs is made more difficult by the conformational transitions of both RNA and protein upon complex formation. It is reasonable to anticipate that the U2B’-RNA interaction will also involve conformational changes of RNA and protein.

Measurements of thermodynamic pairwise coupling in the U1A protein and in the complex helped to identify a network of amino acid side chains that span the RNA binding surface, including the conserved Tyr13, Phe56, and Gln54 (numbered as per U1A) on the surface of the 3-sheet, residues in Loop 3, and in the C-terminal tail of RRM1 [Kranz and Hall, 1998, J Mol Biol, 275, 465-81] [17]. While the Loop 3 sequences in the two proteins are quite different, the other amino acids that have been implicated in this network are conserved between U1A and U2B’, leading to our expectation that the RNA binding surface of U2B’ will also span the entire face of RRM1. A comprehensive thermodynamic study of U2B’-RNA interactions is essential to describe this protein’s binding mechanism and elucidate how and why it differs from the other members of this protein family. Most significantly, this analysis will help us understand how RRMNs with such similar sequences and structures have such different RNA binding properties.

In this study, we use full-length wild-type human U2B’ to assess the binding affinities of this protein for several RNA stemloops (hairpins). We also assess the salt and temperature dependence of these interactions. An important result that has implications for the protein’s biological functions is that human U2B’ binds U1 snRNA SLII and U2 snRNA SLIV with almost equal affinity. In contrast, human U1A protein effectively binds only SLII. Within this protein family, binding and U2 snRNA SLIV with almost equal affinity.

2. Results

The U2B’ protein has two RNA recognition motifs, separated by a 40 amino acid linker. This linker is much shorter than the corresponding linker of U1A. Sequences of the N-terminal RRMs of human U1A, U2B’, and Drosophila SNF are compared in Fig. 1, including some of the linker sequences. A structuraldepiction is shown that highlights the residues on the RNA binding surface that differ between U1A and U2B’. Differences in i2 are extensive, but the most significant difference with respect to RNA binding is in Loop 3, which contacts the RNA.

Loop 3 of U1A, (SRSLMKMRG) is a site that contributes to RNA:protein specificity; in both U1A:SLII and U2B’: SLIV co-crystals, the protein Loop 3 protrudes into the RNA loop where it is juxtaposed with the nucleotides on the bottom of the RNA loop, and so spays the RNA open. The sequence of this important protein loop is different in U2B’ (LKTMKMRG) and Drosophila SNF (LKTLMKMRG). For U1A, Loop 3 makes contacts with the RNA, including a hydrogen bond between the R47 amide and the backbone at the RNA loop-closing G. In the U2B’: SLIV co-crystal, Loop 3 contacts with the U–U appear to be minimal.

The RNA also needs to be considered in a discussion of U1A/U2B’ binding. In the cocrystal of U2B’ RRM1, Stemloop IV of U2 snRNA, and the U2A’ auxiliary protein, the RRM contacts the 5’ AUUGCA sequence of the RNA loop. These interactions appear similar to those formed in the U1A:SLII complex, as expected based on the RNA sequence (AUUGCA in SLII). One significant difference between SLII and SLIV is the loop-closing base pair, which in SLIV is a C-G but in SLIV is a noncanonical U-U pair. In Drosophila, the U-U sequence has become U–G, and this difference appears to be important for recognition of the RNA by SNF [Williams and Hall, 2010, Biochemistry, 49, 4571-82]. The adjacent base pair in the stem is a G:C in human and fly SLIV (Fig. 2), and it may be that this is the effective loop-closing base pair. Here, we refer to the U–U as either the loop-closing base pair or as inserted nucleotides at the bottom of the RNA loop. The difference in the loop-closing basepair may well be important for discrimination between the RNAs by the proteins.

2.1. RNA Binding to U2B’

Stemloop IV, the binding site for U2B’, is located at the 3’ end of U2 snRNA. In vertebrates, the stem of this hairpin has eleven base pairs and an asymmetric internal A bulge in the middle of the stem (Fig. 2). In most vertebrates, the loop sequence is conserved: 5′gUA1UUG-C5AGUAC10CUc, where the loop-closing base pair is indicated in lower case. Much of this sequence is shared with U1 snRNA SLII: cA1UUGC5ACUCC10g. Notable differences include the ‘inserted’ U/U pair at the bottom of the SLIV loop, the C7-to-G7 substitution, and the longer SLIV loop. These phylogenetic differences could identify unique contacts with the corresponding protein, or indicate sites that are insensitive to mutation.

Binding of U2B’ to several RNA sequences suggests some features recognized by the protein. As summarized in Table 1, the human SLII and SLIV RNAs are bound with equal affinity by the protein at room temperature in 100 mM KCl, 1 mM MgCl2, 10 mM sodium cacodylate pH 7.4. This rather surprising result indicates that the presence of the inserted U-U nucleotides does not limit protein binding. Indeed, the protein binds with equal affinity to Drosophila SLIV RNA with its U–G pair. In contrast, the human U1A protein does not detectably bind to the human SLIV RNA in nitrocellulose filter binding assays, but it is able to weakly bind to Drosophila SLIV (5 ± 3 × 10−7 M in 100 mM KCl, 12 mM MgCl2, 10 mM sodium cacodylate pH 7, 22°C). The U–U pair in SLIV appears to function as a discriminator for U1A, but not for U2B’, pointing to a fundamental difference in their mechanisms of RNA recognition. Additionally, a C7-to-G7 substitution in SLII leads to a 10-fold weaker affinity of the U1A:RNA interaction (ΔAG = 1.4 kcal/mol) [6]. In contrast, while the binding affinity of U2B’ for SLII is significantly lower than the affinity of U1A for SLII, the protein tolerates the G7 in SLIV without a further loss of affinity (ΔAG = 0), indicating different binding mechanisms.

Although the stem of SLIV in U2 snRNA contains an internal asymmetric bulge, here we use perfect duplexes for the hairpins. We find that stems of length 6 and 9 base pairs are bound with equal affinity by the protein. In the co-crystal, the RNA stem was a perfect 6 base pair duplex that appeared to make contacts with α5 of U2B’ [11]. We posit that this contact arose through crystal packing that led to a buckling of the base pairs that unwind the stem, together with a twist of the RNA loop.

A significant difference between SLII and SLIV is the insertion of A9 on the 3’ side of the loop. While interactions with the protein were not observed in the cocrystal structure, this A9 could be used as a point of specific contact for U2B’ recognition. To observe the structural
consequence of U2B\(^*\) binding to SLIV, A9 was replaced with 2-aminopurine (2AP), and its fluorescence was monitored with and without bound protein. The introduction of 2AP could potentially disrupt an RNA:protein contact, so U2B\(^*\) affinity for 5'-32P-2AP-SLIV RNA was measured using nitrocellulose filter binding. The binding affinities for human and Drosophila SLIV RNAs were identical to affinities for the wild-type RNAs, giving us confidence that the fluorescence data report on normal binding events.

The steady state 2AP fluorescence intensity of the human SLIV RNA free and bound to U2B\(^*\) is shown in Fig. 3a. There is a 50% increase in the 2AP fluorescence intensity upon binding, indicating that stacking of the 2AP with neighboring bases (which significantly quenches 2AP fluorescence) has been disrupted. U2B\(^*\) binding to Drosophila 2AP-SLIV results in a very similar increase in the fluorescence intensity, shown in Fig. 3b. A feasible interpretation of these results, consistent with the binding data and with the U2B\(^*\):hSLIV crystal structure, is that the protein does not interact with the 2AP (or A9) base in the RNA but needs to open the loop to make specific contacts with the AUUAGCAG sequence. Certainly this result indicates that binding of U2B\(^*\) results in a large change in environment at the 2AP, and by extension, a large conformational change of the RNA.

2.2. Salt dependence of binding

The salt concentration can affect the binding affinity of the protein for the RNA, and indeed, given the highly negative charge of RNA, one

Table 1: U2B\(^*\) binding affinities to RNA hairpins.

<table>
<thead>
<tr>
<th>RNA</th>
<th>[KCl] mM</th>
<th>K(_D) (M)</th>
<th>(\Delta G^*) (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hSLIV 9 base pair stem</td>
<td>100</td>
<td>(1 \pm 2) \times 10(^{-9})</td>
<td>-12.2 (±0.3)</td>
</tr>
<tr>
<td>6 base pair stem</td>
<td>100</td>
<td>(2 \pm 2) \times 10(^{-9})</td>
<td>-12 (±0.2)</td>
</tr>
<tr>
<td>2AP9</td>
<td>100</td>
<td>(2 \pm 2) \times 10(^{-9})</td>
<td>-12 (±0.2)</td>
</tr>
<tr>
<td>250</td>
<td>(2.6 \pm 0.6) \times 10(^{-8})</td>
<td>-10.3 (±0.1)</td>
<td></td>
</tr>
<tr>
<td>hSLII</td>
<td>100</td>
<td>(2.1 \pm 0.2) \times 10(^{-9})</td>
<td>-11.7 (±0.2)</td>
</tr>
<tr>
<td>100/2*</td>
<td>1.9 \pm 0.7 \times 10(^{-8})</td>
<td>-10.5 (±0.2)</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>1 \pm 2 \times 10(^{-8})</td>
<td>-10.1 (±0.2)</td>
<td></td>
</tr>
<tr>
<td>dSLIV</td>
<td>100</td>
<td>(2 \pm 2) \times 10(^{-9})</td>
<td>-12 (±0.2)</td>
</tr>
<tr>
<td>250</td>
<td>3.8 \pm 1.5 \times 10(^{-8})</td>
<td>-10.0 (±0.2)</td>
<td></td>
</tr>
</tbody>
</table>

Binding measured by nitrocellulose filter binding/fluorescence intensity at 22 °C in 100 or 250 mM KCl, 1 mM MgCl\(_2\) (except for hSLII in 100/2 indicating 2 mM MgCl\(_2\)), 10 mM sodium cacodylate pH 7.4 or 10 mM potassium phosphate pH 8. Free energies calculated from the K\(_D\).
expects a large salt dependence for the interaction. On the basis of our experiments to date, we suspect electrostatic interactions do occur between the protein and the phosphates of the stem. U2B$^+$ binding to a dSLIV construct with a fluorescent probe at its 5’ end results in a significant enhancement of fluorescence intensity, as well as an increase in the anisotropy of the fluorophore (Fig. 4). Protein binding results in increases to both the steady-state anisotropy and fluorescence intensity. When both anisotropy and fluorescence intensity were measured, fits yielded equivalent values for $K_{\text{obs}}$. However, data were less noisy for fluorescence enhancement. Similar fluorescence enhancements upon U2B$^+$ binding are seen in SLII and hSLIV constructs with 5’ fluorescentin labels (data not shown).

As with U2B$^+$, SNF binding to the same RNAs results in a large enhancement of fluorescence intensity. With SNF, this enhancement is highly salt-dependent and is greatly decreased when RRM1 alone is used in titrations [13]. Such changes to the fluorescence intensity are hard to explain absent interactions between the protein and the RNA stem. A recent crystal structure of the U1 snRNP suggests that interactions between U1A and SLII occur well beyond the putative end of the RRM domain [18]. It therefore seems likely that the charged linker between the two RRMs interacts with the backbone of the RNA stem to increase the affinity of the interaction. An examination of the sequence of the U2B$^+$ linker shows that there are many lysines that follow an$^3$ and might contribute to electrostatic association with the RNA. If the RRM1 tail is flexible and disordered, those lysines could reach to the stem of the RNA hairpin and interact with the negatively charged backbone. This is probably an important contributor to the salt dependence observed in the interactions.

As a first measure of the salt dependence, the KCl concentration was varied from 150 to 400 mM. The salt dependence of U2B$^+$ binding to hSLII and SLIV, and dSLIV was determined by titrating 5’-FAM-RNAs (fluorescently labeled RNAs) with U2B$^+$ under various salt concentrations. To adequately probe the cation/anion response of the protein to RNA binding, these experiments should be repeated with a larger anion to remove the bias from the RNA; assessing anion effects to the protein:RNA interaction would also allow us to discriminate between the relative contributions of anions and cations to the salt dependence of these interactions [19]. As shown in Fig. 5 and Table 2, the slope of the plot of $\ln(K_{\text{obs}}) \text{ vs } \ln[\text{KCl}]$ for U2B$^+$ binding to human SLII and SLIV indicates a net release of ions for all interactions studied. U2B$^+$ binding to SLII, hIV, and dIV occurs with a net release of 5.2 (± 0.3), 4.6 (± 0.3), and 4.8 (± 0.2) ions, respectively.

This analysis of the salt dependence and the interpretation of the data follows from studies undertaken by the Lohman lab. In a series of experiments to measure the interactions between oligopeptides and DNA and RNA strands, they defined a simple relationship between the change in the observed equilibrium constant with added monovalent salts:

$$\frac{\Delta \log K_{\text{obs}}}{\log [\text{MX}]} = -\Delta \alpha + 2[\text{MX}] \Delta \omega \div [\text{H}_2\text{O}]$$

where $\Delta \alpha$ is the change in cations bound, $\Delta \omega$ is the change in anions bound, and $\Delta \omega \div [\text{H}_2\text{O}]$ indicates net uptake of anions and cations upon binding, whereas a
negative slope indicates ion release. For the U2B’ interaction with its RNAs, we find a net release of approximately 5 ions. By comparison, the value for SNF:dSLIV is \(-4\), for SNF:SLII is \(-6\), and U1A:SLII is \(-7\) [6,13]. This indicates that the electrostatic contributions to binding are distinct in the various RNA:protein interactions and implies that electrostatics play an important role in distinguishing the different RNA binding mechanisms within this protein family. We found that SNF discriminates between the RNAs, and that electrostatics contribute to this discrimination. However, we find that U2B’ binds with similar affinities to the three RNAs, and the overall electrostatic contributions to binding these RNAs appear to be similar.

2.3. Temperature dependence of binding

Although we anticipate from our previous work with U1A and SNF that only U2B’ RRM1 is specific for RNA recognition, the full length U2B’ protein was used in these experiments. The protein is folded and stable to 35 °C as determined by 2H/1H NMR experiments using 300 μM protein (data not shown). Binding to the human SLIV RNA was measured by nitrocellulose filter binding in 100 mM KCl, 1 mM MgCl₂, 10 mM sodium cacodylate, pH 7.4, or by monitoring binding-induced changes in the fluorescence intensity of 5’-fluorescein-labeled RNA in 250 mM KCl. Where the measurements could be made using both nitrocellulose filter binding and protein titrations into fluorescein-labeled RNA, agreement was excellent, indicating that introduction of the fluorescein did not perturb the binding.

As anticipated, U2B’ binding to the RNAs is both temperature- and salt-dependent. The van’t Hoff plot for binding to hSLIV in 250 mM KCl is linear and best fit by the van’t Hoff equation [\(\Delta H^\circ/T - R \Delta S^\circ/R = \ln (K_{\text{obs}})\)] to give \(\Delta H^\circ = -16.4 (\pm 1.0)\) kcal/mol and \(\Delta S^\circ = -21.2 (\pm 3.5)\) cal/mol-K (Fig. 6). In this salt, binding of hSLIV by U2B’ is enthalpy-driven, possibly due to favorable stable stacking of the nucleobases with the aromatic amino acids on the β-sheet surface. In Fig. 6, these data were also fit using the formalism from Ha et al. (1989), to give a small heat capacity of \(-414 (\pm 190)\) cal/mol, but we consider that the temperature-independent values of enthalpy and entropy give an adequate description of this interaction. Given the narrow temperature range for which we can perform these experiments, it is likely that at this salt concentration, we are sampling temperatures above the critical point, where a non-linear van’t Hoff relationship will nevertheless appear linear.

In contrast, at a lower salt concentration (100 mM KCl), binding of U2B’ to hSLIV is better fit by an expression that includes an apparent heat capacity, \(\ln (K_{\text{obs}}) = (\Delta C_p^{\text{obs}}/R)\Delta T - \ln (T_p/T) - 1\) [20], where \(\Delta C_p^{\text{obs}}\) is the observed heat capacity change, \(R\) is the gas constant, \(T\) is the temperature in Kelvin, and \(T_p\) and \(T\) are values where the enthalpy and entropy of complex formation are zero. The result is that the enthalpy and entropy are temperature-dependent. The data do not fit perfectly, but we estimate \(\Delta C_p^{\text{obs}}\) for complex formation is \(-1.6 (\pm 0.4)\) kcal/mol.

U2B’ binding to Drosophila SLIV was measured by 5’-FAM-RNA fluorescence in 250 mM KCl. The binding data could be fit to the van’t Hoff relation to give \(\Delta H^\circ = -15.3 (\pm 2.5)\) kcal/mol and \(\Delta S^\circ = -18.3 (\pm 8.4)\) cal/mol-K. However, these data can also be fit by the expression \(\ln (K_{\text{obs}}) = (\Delta C_p^{\text{obs}}/R)\Delta T - \ln (T_p/T) - 1\) to give the apparent heat capacity of the association: here, \(\Delta C_p^{\text{obs}} = -1.1 (\pm 0.4)\) kcal/mol. The only difference in the hSLIV and dSLIV loops is a U/U for human SLIV is nearly identical, suggesting that the protein does not make significant contacts with the base of the RNA loop.

U2B’ also binds to human SLII. The data cannot be fit by the van’t Hoff equation, and instead leads to \(\Delta C_p^{\text{obs}} = -2.2 (\pm 0.4)\) kcal/mol, in 100 mM KCl, 1 mM MgCl₂. As Fig. 6 shows, the temperature dependence of U2B’ binding to hSLIV and hSLII is nearly identical, suggesting that either the binding mechanism is the same, or that there are compensating interactions that result in similar thermodynamic signatures.

True heat capacity changes upon ligand binding are in general attributed to hydrophobic surface burial. We have observed large apparent heat capacity changes in U1A, SNF, and now U2B’. However, one difficulty with interpretation of the heat capacity term for these RNA:protein complexes lies in the possibility that there are conformational changes of the components coupled to binding. It has been shown that apparent heat capacity changes can occur in the absence of hydrophobic surface burial, as a result of temperature-dependent conformational changes that are coupled to binding but that occur with a \(\Delta C_p = 0\) [21]. The ZAP fluorescence data show clearly that there is a large conformational change in the RNA upon binding, and it is likely that conformational changes also occur in the protein (particularly in Loop 3). Measurement of the temperature dependence of the calorimetric enthalpy of binding using isothermal titration calorimetry (ITC) would make interpretation clearer, as temperature dependence of this measurement...
would indicate contributions from a true ΔGp. These experiments are planned for U2B*.

3. Conclusions

These data show conclusively 1) that the human U2B* protein does not require an auxiliary protein for binding to RNA; 2) that U2B* binds to both U1 snRNA stemloop II as well as to U2 snRNA stemloop IV; and 3) U2B* shows no preference for SLIV or SLII on the basis of binding affinity. Temperature and salt dependence of these interactions are similar, which is distinct from recognition of the same RNAs by Drosophila SNF, suggesting that U2B* recognizes the different RNAs through a common mechanism. These results have important implications for the biology of snRNP assembly and composition, but also for development of a general model for RNA recognition by the U1A/U2B* family of RRM proteins.

3.1. snRNP composition

The U1 and U2 snRNPs mostly assemble in the cytoplasm, adding the Sm proteins to the snRNA after its transport out of the nucleus. Other protein components are added subsequently. The U1 snRNP is relatively simple, with only three U1-specific proteins in addition to the Sm cluster: U1A binds to SLII, the U1 70 K protein binds to SLI, and the U1C protein binds to the tail of U1 70 K [18,22]. The U2 snRNP is the most complex snRNP, and its composition is functionally regulated. Recent studies suggest that the U2B* protein is present in the particle throughout its life, even though it is located at the extreme 3′ end of the U2 snRNA which, in contrast to much of the rest of the snRNA, does not rearrange during splicing [23].

Both U1A and U2B* are found in the cytoplasm where they can bind to their snRNA targets, which for U2B* includes both SLI and SLIV. U1A would win the binding contest for its SLII target on the basis of its higher affinity. Comparing U1A and U2B* binding to SLII in vitro shows that U1A has (approximately) a 100-fold greater binding affinity. If cytoplasmic concentrations of each protein are equal, then U2B* would be effectively excluded from the U1 snRNP. Under conditions where the U1A concentration is limiting, U2B* could certainly bind to SLII, as indeed occurs in C. elegans when the U1A protein is knocked out [4]. However, in vertebrates, U1A could not replace U2B* since its affinity for the human SLIV is too weak. In this respect, the two vertebrate proteins are not interchangeable, and it seems that U2 SLIV may have evolved to exclude U1A.

3.2. RNA recognition

Based on our in vitro results, it appears that the U2B* protein is quite tolerant of substitutions in the RNA loop, at least on the 3′ side and at the loop-closing base pair. One model is that only the loop AUJUGCAG nucleotides constitute the recognition site for the protein, and all other interactions are nonspecific electrostatic contacts. Certainly the very similar binding affinity, temperature dependence, and salt dependence for hSLIV, hSLII, and dSLIV RNAs are consistent with this scheme. The protein T89D90S91 amino acids (numbered according to U1A) that recognize the G4A6(C7/G7) sequences of SLIV and SLII are conserved in both U1A and U2B*. The T89D90S91 residues are located after the hinge that links [4] with the third α-helix, but it is their peptide backbone amide and carbonyl oxygen atoms that hydrogen bond with the nucleobases. This interaction appears to be modulated by the geometry of the protein that presents the TDS backbone for contacts, all of which are conserved in the two proteins.

The more interesting interactions are those that the thermodynamics show are not critical for RNA binding, but that involve Loop 3 of the protein, one of the phylogenetically conserved sequences in each U1A and U2B* sub-family. The construction of thermodynamic cubes that included RNA binding together with pairwise coupled sites on the protein [17] was inspired by the work of DiCera, which was presented at Gibbs Biothermodynamics meetings. Kranz’s novel application to a bimolecular system provided the first evidence of a network of interactions that together formed the RNA binding site of the protein. Thermodynamic pairwise coupling applied to U1A and the U1A:SLII complex helped to define a network of interactions that linked Loop 3 with the aromatic residues on the surface of the β-sheet.

One intriguing possibility is that in the U2B* protein, the pairwise coupling patterns are altered with the result that Loop 3 is no longer linked to the β-sheet surface. The U1A protein is extremely sensitive to the nucleobases at the bottom of SLII through interactions with the C:G loop-closing base pair mediated by Loop 3. Insertion of a single U between that C:G base pair and the first A of SLII reduces the binding affinity of U1A by several orders of magnitude [7]. We proposed that the insertion causes a shift in the frame of RNA binding on the protein β-sheet through disruption of Loop 3/RNA interactions. Here, however, we observe that the U2B* protein seems insensitive to the nucleotides at the bottom of the loop, which implies a very different role for its Loop 3.

Although it is not apparent from the structures of the proteins alone or in their respective co-crystals, the thermodynamic analysis of U2B* binding to three RNA hairpins shows that its RNA binding surface is quite different from that of U1A. Its specificity is relaxed, and its mechanism of RNA recognition appears to rely on only a subset of the RNA loop nucleotides. Appreciation of the thermodynamic properties of this complex will direct further experiments that probe the atomic details of the interactions.

4. Materials and methods

4.1. U2B* cloning and purification

The human U2B* clone was purchased from the American Type Culture Collection (ATCC). The full length cDNA was amplified using the following primers: Forward primer: GGTGGTCCATGGAATACTCA-GACC; Reverse Primer: GGTGGTAAAGCTTATTATTTCTGGCATAGGand subcloned into the IPTG-inducible pTAC vector using the Ncol and HindIII sites. The plasmid was transformed into BL-21(DE3) cells for protein expression. Cells were grown in LB medium at 37 °C to an optical density of 0.6–0.8 and were induced with 1 mM IPTG for 3 h. After addition of IPTG, the temperature was dropped to 28 °C. Cells were harvested and stored at −70 °C until lysis. Cells were resuspended in 30 mM sodium acetate (pH 5.3), 200 mM NaCl, 2 mM EDTA, and 8.5% sucrose. PMSF, DNase II, and a protease inhibitor cocktail (Sigma) were added prior to French pressing the cells. The lysate was collected and spun down in an ultracentrifuge at 4 °C, 45,000 g. The supernatant was filtered through a 0.22 µm cellulose acetate membrane and loaded onto an SP Sepharose column pre-equilibrated in 50 mM Tris (pH 7.5). U2B* was eluted over 170 min, using a 190 to 350 mM NaCl gradient. EDTA and PMSF were added to the eluted fractions to a final concentration of 5 mM and 20 µg/ml, respectively. All column buffers were sterile-filtered through 0.45 µm cellulose nitrate filters (Nalgene), and containers used in the purification were acid washed to remove RNases. Fractions containing U2B* were concentrated using a Vivaspin concentrator with a molecular weight cutoff of 10 kDa and buffer-exchanged into 100 mM KCl, 1 mM cacodylate pH 7, 5 mM EDTA. Concentration was determined spectrophotometrically using ε280 = 8960 M−1 cm−1.

4.2. RNA synthesis

Nitrocellulose filter binding experiments were performed with RNA transcripts that were enzymatically synthesized with T7 RNA polymerase, as described [24,25]. The DNA oligonucleotides were obtained from IDT (Integrated DNA Technologies). The transcripts
were internally labeled with [α-32P]UTP and [α-32P]CTP. Labelled RNAs were gel-purified before use in binding assays. The RNA product for dU1 SLII was 5′-GGGCGGUAUUGCAUCUCCGGCCGGCC. The RNA product for dU2 SLIV was 5′-GGGCCGGUUAUUGCAUACCCGGCCGGCC. Two hU2 SLIV constructs were: 5′-GAGACCUUAUUGCAUACCCGGCCGGCC and 5′-GGGCCGGUUAUUGCAUACCCGGCCGGCC. The sequences corresponding to the loops are underlined.

4.3. 2-aminopurine fluorescence experiments

RNA products internally labeled with 2AP were obtained from Dharmaco and IBA GmbH. The RNAs were: 5′-GGGCCGUUAUUGCAUACCCGGCCGGCC (hSLIV) and 5′-GGGCCGUUAUUGCAUACCCGGCCGGCC (dSLIV). An SLM 8000 instrument was used to perform fluorescence experiments. The temperature was set to 23 °C and controlled by a circulating water bath. Cuvettes were blocked for at least one hour with 20 μg/mL of BSA. The cuvettes were subsequently rinsed with 1× buffer (20 mM KCL, 10 mM sodium cacodylate pH 7) and snap-cooled on ice for 2 min. 1/10 volume of RNA was diluted to 2.2 μM in water, heated at 95 °C for 5 min, and snap-cooled on ice for 2 min. 1/10 volume of a sample was added and the RNA was folded as described for the 2AP-labelled RNAs with 20 mM potassium phosphate, 1 mM MgCl2 for 1 h with buffer containing appropriate concentrations of KCl and MgCl2. 20 μg/mL of BSA and 10 mM potassium phosphate pH 8 to prevent sticking of U2B′ to the cuvette walls. For initial experiments, steady-state anisotropy and fluorescence intensity were both measured throughout the protein titrations. While fits of the data yielded comparable Ks for both methods, fluorescence intensity proved to be more sensitive. Subsequent experiments were performed by measuring fluorescence intensity alone. Cuvette temperature was controlled using a circulating water bath. The excitation and emission wavelengths were set to 490 and 520 nm, respectively. Temperature dependence experiments were performed using 250 mM KCl, 10 mM potassium phosphate, 1 mM MgCl2, pH 8. Salt dependence experiments were performed using 250 mM KCl, 10 mM potassium phosphate, 2 mM MgCl2, pH 8. Salt dependence experiments were performed using 250 mM KCl, 10 mM potassium phosphate, 2 mM MgCl2, pH 8. Binding curves were fit to a standard single-site binding model using Scientist (Micromath). Experiments were performed at least twice, and the reported values reflect the average of these experiments. Uncertainties are given as the larger value of either the standard deviation of these values or the propagated uncertainty.

4.5. RNA binding assays

Nitrocellulose filter binding assays were used to determine standard binding free energies for binding of RNA to different SNF protein constructs, as described [16]. A constant, picomolar concentration of RNA and variable protein concentrations were used. BSA (Roche) was added to a final concentration of 40 μg/mL. All experiments described in the text were done at pH 7.4. Solution conditions were otherwise variable and are indicated in the text and figures. All experiments were performed in duplicate, and binding curves were fit to a standard Langmuir isotherm using Scientist (Micromath). Kaleidagraph was used to fit van’t Hoff data.

Acknowledgments

This work is supported in part by NIH (SCW and KBH). Thanks to Dr. Kim Delaney for cloning human U2B′. Thanks to many members of the Gibbs community who patiently schooled KBH in thermodynamics and appreciated its application to RNA.

References


Chapter 4.

Introduction to Phylogenies and Ancestral Sequence Reconstruction
The next two chapters build an understanding of the evolution of the U1A/U2B”/SNF protein family. They rely on a phylogenetic analysis of modern metazoan U1A, U2B”, and SNF sequences. As part of this analysis, Maximum Likelihood sequences of ancestral proteins were inferred. Before describing the results, I will introduce the concepts and methods underlying this approach. Most of this introduction is summarized from (1).

Phylogenetics is the study of the evolutionary relationship between different populations, species, or genes. A phylogenetic tree, or phylogeny, is used to depict these relationships. Clearly, we do not have a complete record of organisms and their genes over evolutionary time, so the phylogeny must be inferred from available data. Various kinds of data can be used, as can various methods of inference.

Figure 1a shows a tree depicting the genealogic relationship between five modern sequences in a gene family. The modern sequences are the external nodes, or the tips of the tree, and are labeled A-E. Traditional phylogenies are built from morphological analyses of species; such analyses use multiple morphological traits present in living organisms or in the fossil record of extinct ones to make inferences about how organisms are related to each other (and thus, how they evolved). In contrast, in molecular phylogenetics the tree is inferred from sequence data at the external nodes. In Figure 1a, the internal nodes are shown as circles colored magenta, and these correspond to sequences in ancestral, extinct organisms. The node representing the last common ancestor of all modern sequences (A-E) is called the root of the tree. It is an internal node, which in Figure 1(a) is depicted as larger than the other nodes. The edges connecting the nodes are called branches.
Figure 1b shows three different trees that could represent the relationship of the five sequences to each other; these trees have different topologies, or branching patterns. In all of these, D and E are more closely related to each other than to any other sequence (they are separated by a single internal node). However, the relationship between D and C is the same in the bottom two trees (separated by two internal nodes) but different in the top tree (separated by 3 internal nodes). Similarly, the relationship between A and B is shared in the top trees but different in the bottom tree.

A.             B.

Figure 1. Phylogenetic trees showing how sequences or species can be related.

The above trees are fully resolved; all non-root internal nodes have three branches, and the root has two branches. However, trees can also be partially or fully unresolved, which is
represented in a tree by a multifurcation (Figure 2). A multifurcation is also called a polytomy, and most commonly, such polytomies are the result of insufficient information to resolve the true relationship between sequences, rather than a result of multiple simultaneous divergence events. Tree-building algorithms generally restrict the tree search to fully resolved trees, as this simplifies the search process. However, branch supports on the final tree give information about the robustness of a particular node, which may indicate that the branching is not truly resolved. In reconstructing metazoan U1A/U2B"/SNF protein phylogeny, the deuterostome branch was not well-resolved. However, it was possible to reconstruct the deuterostome branch separately. The inclusion of additional sequence data in this reconstruction improved the resolution of the deuterostome phylogeny.

![Fully resolved tree](image1.png) ![Partially resolved tree](image2.png) ![Star tree](image3.png)

**Figure 2.**

**Tree searching.**

Sequence-based (molecular) tree reconstructions search multiple trees, assign each tree a score, and choose the tree with the best score. The tree search, however, is very computationally
intensive for large numbers of sequences. The number of distinct, unrooted, fully resolved trees is given by:

\[ B(T) = \prod_{i=3}^{T} (2i - 5) \quad (Eqn. 1) \]

(1), where \( T \) is the number of sequences. For the modest number of 50 sequences, there are \( 3 \times 10^{74} \) such trees; exhaustively evaluating all trees is, in general, not feasible. One approach to the tree searching problem is to generate an initial tree and perform iterative branch-swapping steps. At each step, a collection of trees is generated by branch-swapping. These trees are evaluated by some criterion (maximum-likelihood or parsimony), and the best tree is selected as the basis for the next round of branch-swapping steps until there is no further improvement in the criterion used for optimization. This procedure is clearly not exhaustive and can lead to trapping in a local minimum. Often, the branch-swapping procedure is performed on multiple starting trees, in an attempt to find the true minimum. PhyML, the program used for phylogenetic inference in this thesis, utilizes this kind of tree-searching approach (2).

**Gene tree vs. Species tree**

The term ‘species tree’ is used to describe the evolutionary relationship between different organisms. Historically, species trees have been inferred from morphological analyses. In molecular phylogenetics, a ‘gene tree’ or ‘sequence tree’ is instead inferred. These trees describe the relationship between particular genes or sequences and may or may not be identical to the species tree. In some cases (probably most famously, that of ribosomal RNA (3)), the gene tree
has been used as a best approximation of the species tree. Modern methods often use sequences from multiple genes to refine the species tree.

There are several reasons for potential discrepancies between the species tree and a gene tree. These include errors in estimating the tree (the gene tree should resemble the true species tree, but an insufficient sequence pool or deficiencies in the reconstruction methods make the trees appear dissimilar). True deviations between a gene tree and a species tree can be the result of lateral gene transfer, ancestral polymorphisms, or a consequence of gene duplications and losses. Important in the U1A/U2B''/SNF family is the occurrence of gene duplications. Absent other reasons for deviation from the species tree, if a gene duplication occurred at an early stage, the gene tree should yield two species trees, separated by the node at which the gene duplication occurred. Late gene duplications should resemble a species tree, with bifurcations occurring far down a branch.

**Methods of tree inference.**

Several methods have been developed to infer a gene tree topology. These include distance-based methods, parsimony, maximum-likelihood, and Bayesian methods. Parsimony methods are fairly intuitive and will be described first. A description of the maximum-likelihood method, which was used in this work, will follow.

The premise of maximum-parsimony approaches is that the evolutionary trajectory that best explains the relationship between modern sequences is the one requiring the fewest changes (fewest total number of nucleotide or amino acid changes). This method (like maximum
likelihood approaches) relies on an alignment of modern sequences. Each position within the alignment is treated separately, and reconstructed trees are given a score based on the minimum number of character changes that are necessary to obtain the modern set of sequences. These character changes are summed over all sites and all branches of the tree, yielding the tree score. Trees with the lowest score are considered Maximum Parsimony trees. The principle of parsimony was used by Zuckerkandl and Pauling to infer ancestral hemoglobin states (4), although the hemoglobin/myoglobin phylogeny used in the analysis was inferred largely from more global protein characteristics, including assembly state, molecular weight, and some knowledge of genetic linkages (5). Parsimony was also used in one of the first sequence-based phylogenetic reconstructions (of ferredoxin) (6) and to develop amino acid substitution rate matrices (7). Refined versions of the Dayhoff matrix is used in Maximum Likelihood methods to account for differences in amino acid substitution rates.

Unfortunately, particularly with closely related sequences, there are frequently multiple trees that have the same lowest parsimony score. Other problems with the parsimony method include that it ignores branch lengths (the distance between nodes) and variations in rates of change between nucleotides or amino acids. For example, K->R substitutions occur with much higher frequency than K->F substitutions (both because fewer nucleotide mutations must occur and because the more conservative change is more likely to be tolerated), but these transitions contribute equally to a tree’s parsimony score. Also, branches within the true gene tree are of different lengths, reflecting different evolutionary rates on that branch or different lapses of evolutionary time. One expects longer branches to accumulate more changes than shorter
branches, but the parsimony score fails to take this into account. In contrast, likelihood methods can incorporate both differences in substitution rates and differences in branch lengths.

Maximum likelihood methods for phylogenetic reconstruction work by establishing a likelihood function, based on a Markov model of evolution. A fairly early model of nucleotide sequence evolution is the K80 model. The substitution rate for a transition is $\alpha$ and for a transversion is $\beta$. A substitution rate matrix, $Q$, describes the instantaneous rate of change for one nucleotide to another, and thus the relative substitution rates. In the case of the K80 model, $Q$ is given by:

$$Q = \begin{pmatrix}
-(\alpha+2\beta) & \alpha & \beta & \beta \\
\alpha & -(\alpha+2\beta) & \beta & \beta \\
\beta & \beta & -(\alpha+2\beta) & \alpha \\
\beta & \beta & \alpha & -(\alpha+2\beta)
\end{pmatrix} \quad (Eqn. 2)$$

The total substitution rate for any nucleotide is $\alpha+2\beta$ (one transition and two transversions are possible), and the distance $d$ between two sequences separated by time $t$ is $d = t(\alpha+2\beta)$. Evolution is modeled as a continuous Markov process, where $dP/dt = P(t)Q$, with a boundary condition $P(0) = I$. The matrix of transition probabilities is therefore given by:

$$P(t) = e^{Qt} \quad (Eqn. 3)$$

There are three distinct elements of the matrix, corresponding to the probability of a transition, a transversion, or no nucleotide change (the diagonal elements) after time $t$. The probabilities can be expressed in terms of $\alpha$, $\beta$, and $t$, or alternatively they can be expressed in terms of the distance $d$ and the transition:transversion rate ratio, $\kappa = \alpha/\beta$. If one considers two related
nucleotide sequences of the same length, the distance between the sequences is the number of expected changes. One can estimate $d$ and $\kappa$ by counting the number of transitional and transversional changes between the sequences.

Estimators for $d$ and $\kappa$ can also be obtained by Maximum Likelihood methods. In this case, the probabilities given by $P$ can then be used to establish estimates of $\kappa$ and the distance $d$ between sequences in terms of the observed proportion of sites with transitional or transversional differences. The log likelihood of the two sequences being separated by a distance $d$ and the transition:transversion ratio being $\kappa$, given $n$ total sites, $n_s$ transitional differences, and $n_v$ transversional differences, is given by:

$$\ell(d, \kappa | n_s, n_v) = (n-n_s-n_v) \log(p_0/4) + n_s \log(p_1/4) + n_v \log(p_2/4)$$  \hspace{1em} (Eqn. 4)

where $p_0$, $p_1$, and $p_2$ are the probabilities for no change, a transition, and a transversion, respectively; they are functions of $d$ and $\kappa$ as given by $P(t)$. The maximum likelihood estimators of $d$ and $\kappa$ are obtained in terms of $n$, $n_s$, and $n_v$ after setting $\partial \ell / \partial d = 0$, $\partial \ell / \partial \kappa = 0$.

The premise of ML tree reconstruction is to use this statistical framework of how sequences evolve probabilistically to evaluate a tree and assign it a score. The data are the aligned sequence set. The sites are assumed to evolve independently, so the probability of the entire data set is the product of the probabilities at the different sites.

The log likelihood of observing a given sequence alignment is:

$$\ell = \sum_{h=1}^{s} \log[f(x_h | \theta)]$$  \hspace{1em} (Eqn. 5)

where $h$ is the index of sites, $x_h$ are the nucleotides in the sequence alignment, and $\theta$ are the model parameters. An example tree is shown in Figure 3. The external nodes are labelled 1-5,
and at a given site of the DNA sequence alignment, these nodes have the nucleotide indicated. The data, $x_h$, are TCACC. The model parameters $\theta$ include the distances $d1$ to $d8$ and the transition:transversion ratio $\kappa$. The data $x_h$ could result from any set of nucleotides at the internal ancestral nodes. $f(x_h | \theta)$ must therefore sum over all possible nucleotide combinations at nodes 0, 6-8. With this function, ML will estimate the model parameters for the data set and also provide a maximum likelihood score for the tree. This procedure can be repeated over different possible trees, and the tree with the best ML score is selected as the ML tree, with the model parameters (branch lengths $d$ and the transition:transversion ratio) estimated for the tree.

Figure 3.

For protein sequences, the substitution rate matrix $Q$ is of course a 20 x 20 matrix that has been empirically determined based on large protein data sets. As with the K80 model, branch lengths are included. Additionally, parameters can be added to take into account variability in substitution rates across sites. In the case of the phylogenetic reconstruction presented, this is done by considering that the rate variation across sites is described by a gamma distribution.
**Ancestral sequence reconstruction**

Ancestral sequence reconstruction seeks to determine what the probable sequences were at the tree’s internal nodes. For a parsimony approach, the Maximum Parsimony reconstruction is the reconstruction in which the sequences of the internal nodes result in the fewest total sequence changes. For the ML reconstruction, the problem can be treated using Bayes’ theorem to calculate the posterior probability that a node has a particular amino acid. While this can be done using an entirely Bayesian framework, it is more often done using the discrete ML estimates of the model parameters determined in the tree reconstruction. Returning to the K80 nucleotide model and the tree in Figure 3, the posterior probability that node 0 has nucleotide $x_0$ is given by:

$$f(x_0|x_h;\theta) = \frac{f(x_h|x_0;\theta) f(x_0;\theta)}{\sum_{x_0} f(x_h|x_0;\theta) f(x_0;\theta)}$$  \(\text{(Eqn. 6)}\)

where $\pi_{x_h} x_0$ is the joint probability of the states at the external nodes ($x_h$) and the state at Node 0. The denominator sums over all possible nucleotides at Node 0. The prior probability $\pi_{x_0}$ is the equilibrium nucleotide frequency (1/4 for any nucleotide in this model; for amino acids, equilibrium amino acid frequencies have been determined empirically from the same datasets used to generate the substitution rate matrices).

**Some limitations of phylogeny and sequence reconstruction.**

While progress continues to be made in improving and refining ancestral reconstruction methods, there are several caveats and remaining problems in phylogeny and sequence reconstruction methods. ML and parsimony methods treat each site in the sequence as
independent of all other sites. This allows likelihoods at different sites to be calculated separately. Although sites can certainly co-evolve (and this information is in itself interesting), treating sites independently for the purpose of both phylogenetic and ancestral reconstruction appears to work well. Similarly, ML methods of ancestral sequence reconstruction assume the ML phylogeny to be true and ignore uncertainties in the phylogenetic reconstruction. While this is a large assumption, it appears to compromise the accuracy of the ancestral sequence reconstruction only rarely (as compared with a Bayesian integration of ancestral states over alternative trees) (8).

The phylogenetic reconstruction is relatively sensitive to the level of conservation between related sequences. Sites that show very little conservation contain very little information about the evolutionary trajectory leading to the modern sequences. This is also true of sites that are universally (or nearly universally) conserved. Thus, one needs an intermediate level of conservation to obtain high topological accuracy of the reconstructed tree (2). In our study, the conservation of the RRM sequence was in fact higher than ideal for tree reconstruction. The interdomain linker was more divergent and thus provided important information. However, the linker is itself a site that experienced a large number of insertions and deletions.

Importantly, treatment of sequence gaps, insertions, and deletions is problematic in both phylogenetic and ancestral sequence reconstruction. There is no obvious way to deal with this problem, but these changes are common in sequence alignments. It is not uncommon to throw away sites that contain insertions or deletions in both the phylogenetic and sequence
reconstruction. Alternatively, missing amino acids can be treated as a separate character. This is unfortunate, as insertions or deletions can contain substantial phylogenetic information. They can also be important sources of novel or changing gene or protein function.

In the next two chapters, a phylogeny of metazoan U1A/U2B″/SNF proteins shows that the last common ancestor (Urbilaterian) of Drosophila and humans had a single SNF-like protein and that the gene duplication giving rise to separate U1A and U2B″ proteins occurred as recently as in the last common ancestor of jawed vertebrates. This information is also used to begin to understand how separate U1A and U2B″ biochemical functions emerged.
References


Chapter 5.

Resurrection of an Urbilaterian U1A/U2B′′/SNF Protein
Resurrection of an Urbilaterrian U1A/U2B''/SNF Protein

Sandra G. Williams¹, Michael J. Harms² and Kathleen B. Hall¹

¹ - Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO 63108, USA
² - Center for Ecology and Evolution, University of Oregon, Eugene, OR 97403-5289, USA

Correspondence to Kathleen B. Hall: kathleenhal@gmail.com
http://dx.doi.org/10.1016/j.jmb.2013.05.031
Edited by M. F. Summers

Abstract

The U1A/U2B''/SNF family of proteins found in the U1 and U2 spliceosomal small nuclear ribonucleoproteins is highly conserved. In spite of the high degree of sequence and structural conservation, modern members of this protein family have unique RNA binding properties. These differences have necessarily resulted from evolutionary processes, and therefore, we reconstructed the protein phylogeny in order to understand how and when divergence occurred and how protein function has been modulated. Contrary to the conventional understanding of an ancient human U1A/U2B'' gene duplication, we show that the last common ancestor of bilaterians contained a single ancestral protein (URB). The gene for URB was synthesized, the protein was overexpressed and purified, and we assessed RNA binding to modern snRNA sequences. We find that URB binds human and Drosophila U1 snRNA SLII and U2 snRNA SLIV with higher affinity than do modern homologs, suggesting that both Drosophila SNF and human U1A/U2B'' have evolved into weaker binders of one RNA or both RNAs.

Introduction

The eukaryotic spliceosome is a large, complex, and highly dynamic macromolecular machine that splices pre-mRNA. Among its many associated proteins is the U1A/U2B''/SNF family, which are components of the U1 and U2 small nuclear ribonucleoproteins (snRNPs). These three proteins use RNA recognition motifs (RRMs) to recognize their RNA targets, but their N-terminal RRM
Reconstruction of an Ancestral RRM

(RRM1) are distinguished from most RRMs by their extremely high affinity and the exquisite specificity of their RNA binding. The RRMs of these three proteins have high sequence identity, as do the two RNA stem–loops that they recognize, consistent with a shared phylogenetic lineage of the proteins.

One copy of U1A, U2B+, or SNF is present in the U1 and U2 snRNPs, where the protein binds to a specific stem–loop. In spite of their high sequence identity and structural similarity, Drosophila SNF, human U1A, and human U2B+ have distinct RNA binding properties. Drosophila SNF binds both U1 snRNA SLII and U2 snRNA SLIV [1], whereas in humans, U1A binds exclusively to U1 SLII, and in nuclear extracts, U2B+ localizes exclusively to the U2 snRNP, where it binds SLIV [2]. In vitro, these differences are manifested in very high affinity and specificity of U1A for SLII, modest affinity and no specificity between SLII and SLIV for U2B+, and an intermediate specificity for SNF [3–5]. Therefore, comparing the modern RRMs, it is clear that each has unique RNA binding properties, but the molecular basis for these differences has been difficult to explain from structural considerations of the proteins. Placing the observed functional diversity within its evolutionary context promises to provide new insight into modern protein function: residues responsible for altered function can be determined, allowing insights into how these mutations altered the protein to result in changes to RNA binding.

Pauling and Zuckerkandl first explored the possibility of studying extinct proteins 50 years ago [6], but it has only been recently that advances in phylogenetic analysis and the explosion of available sequence data have made these “molecular restoration studies” feasible [7–10]. The power of this approach is that it provides the evolutionary context for studying proteins, thus taking advantage of Nature’s long-running experiments. This obviates many of the difficulties of traditional comparative mutagenesis studies, which include sifting through the large number of functionally irrelevant background mutations, identifying interacting mutations, and contending with lineage dependence of functionally relevant mutations [11].

We used U1A/U2B+/SNF sequences from broadly diverse organisms to reconstruct the metazoan protein phylogeny and resurrect the ancestral protein of the last common ancestor of human U1A, human U2B+, and Drosophila SNF. Our goals were to determine whether the gene duplication responsible for subfunctionalization of human proteins occurred early or late in metazoan evolution and how the RNA binding properties of the ancestral protein compare with its modern descendants. Our new phylogeny revises the current understanding of U1A/U2B+/SNF functional divergence: a single protein family member was present in the last common ancestor of bilaterians, and gene duplications resulting in separate U1A and U2B+ proteins are relatively recent. Like its modern counterparts, the resurrected URB protein (which corresponds to the last common ancestor of bilaterians) has two RRMs separated by a flexible linker. URB binds modern U1 snRNA SLII and U2 snRNA SLIV with high affinity, and its specificities for RNAs most closely resemble those of Drosophila SNF. URB’s dynamic properties in solution also resemble those of SNF and differ from human U1A, suggesting that dynamics have been evolutionarily conserved and may be implicated in protein function.

Results

A new family phylogeny

Because humans, potatoes, and yeast each code for separate U1A and U2B+ proteins, prior characterization of U1A/U2B+ molecular evolution posited on the basis of parsimony that U1A and U2B+ paralogs emerged after a single, ancient gene duplication prior to the divergence of plants, fungi, and metazoans [1]. We tested whether metazoan proteins were consistent with this model using modern phylogenetic methods and sequences from broadly diverse metazoans. A schematic of the ancestral reconstruction approach we employed is shown in Supplementary Fig. 1. Putative U1A, U2B+, and SNF proteins from 160 diverse metazoan organisms were obtained from BLAST searches and subsequently aligned. PhyML [12] was used to reconstruct the protein phylogeny, and a resulting cladogram of the maximum likelihood (ML) tree is presented in Fig. 1a. The striking result of this analysis is that the last common ancestor of all bilaterians had a single U1A/U2B+ protein, as do most modern metazoans.

The reconstructed phylogeny reveals that gene duplications within the bilaterian lineage have occurred at least three times: once in the evolution of jawed vertebrates, once in the lophotrochozoan lineage, and once in the nematode lineage. The implication of our reconstruction is that, if gene duplication results in subfunctionalization of RNA binding and localization to distinct snRNPs, this occurred late in the proteins’ molecular evolution. Reconstruction of the full tree (Fig. 1a) results in poor resolution of the deuterostome phylogeny, prompting a separate analysis of the deuterostome sequences. This reconstruction used an alignment containing more residues from the interdomain linker (see Materials and Methods). ML and maximum parsimony (MP) reconstructions of this deuterostome phylogeny are consistent with a single gene duplication in an ancestor of jawed
vertebrates (Fig. 1b and Supplementary Fig. 2) that resulted in separate U1A and U2B* proteins in these animals.

U1A/U2B*/SNF ancestral protein sequences were subsequently inferred from modern sequence alignments and the reconstructed phylogeny using CodeML [13]. Sequence alignments of the reconstructed Urbilaterian SNF (indicated in Fig. 1a and which we call URB) and modern homologs are shown in Fig. 2a. Also shown in Fig. 2b are the residues that have diverged between URB and the human proteins (left panel) and URB and Drosophila SNF (right panel), plotted on the RRM structure. With few exceptions, amino acids within the RRMs were unambiguously predicted (Supplementary Fig. 3 and Supplementary Table 1), and the alignment illustrates the high sequence conservation of the RRMs. Particularly striking is the conservation of RRM2, for which there is no known biological or biochemical function.

A resurrected Urbilaterian ancestral protein resembles Drosophila SNF

RNA stem–loops and URB binding

In order to study the RNA binding properties of the ancestral RRMs, it was important to determine appropriate RNA sequences for binding studies. Modern U1 stem–loop II and U2 stem–loop IV sequences from diverse metazoans were obtained and aligned. Consensus sequences for the loop and loop-closing base pair, which are known to be important for protein recognition, are shown in Fig. 3 (Supplementary Fig. 4). Features of the RNA stem–loops that have previously been identified as
Reconstruction of an Ancestral RRM

### (a) RRM1

<table>
<thead>
<tr>
<th>Species</th>
<th>Residue Range</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human U1A</td>
<td>human U1A 1</td>
<td>MAVPETRPHYIYINLNEKIKDELKSLSYIALPSQFGQILDIVSRLKMRG</td>
</tr>
<tr>
<td>URB</td>
<td>human U1A 1</td>
<td>N---DIRPHRTIYINLNEKIKDELKSLSYIALPSQFGQILDIVALKTLKMRG</td>
</tr>
<tr>
<td>Dros. SNF</td>
<td>human U1A 1</td>
<td>M---EMLPQZIYINLNEKIKKELKSLSYIALPSQFGQILDIVALKTLKMRG</td>
</tr>
<tr>
<td>Human U2B’</td>
<td>human U1A 1</td>
<td>M---DIRPHRTIYINLNEKIKDELKSLSYIALPSQFGQILDIVALKTLKMRG</td>
</tr>
<tr>
<td>Human U1A</td>
<td>human U2B’ 51</td>
<td>QAFTKFKEVSAATNLRSMQGFYDKPMLRIRYAKTDSEQIAKMKGTFVER</td>
</tr>
<tr>
<td>URB</td>
<td>human U2B’ 51</td>
<td>QAFTKFKEVSAATNLRSMQGFYDKPMLRIRYAKTDSEQIAKMKGTFVER</td>
</tr>
<tr>
<td>Dros. SNF</td>
<td>human U2B’ 51</td>
<td>QAFTKFKEVSAATNLRSMQGFYDKPMLRIRYAKTDSEQIAKMKGTFVER</td>
</tr>
<tr>
<td>Human U2B’</td>
<td>human U2B’ 51</td>
<td>QAFTKFKEVSAATNLRSMQGFYDKPMLRIRYAKTDSEQIAKMKGTFVER</td>
</tr>
</tbody>
</table>

### (b) Linker

<table>
<thead>
<tr>
<th>Species</th>
<th>Residue Range</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human U1A</td>
<td>human U1A 105</td>
<td>DRRREKRPPKTEGTPQMPQSQSSENPPNHILFLLFLPVENELMLSLNFQF</td>
</tr>
<tr>
<td>URB</td>
<td>human U1A 102</td>
<td>EEKKEAAAATKAPAAPAAAPAAPAAPAAPAAPAAPAAPAAPAAPAAPA</td>
</tr>
<tr>
<td>Dros. SNF</td>
<td>human U1A 102</td>
<td>HKKKKPPKAPDEASSADNP-AGTEQPPQILFLWFLVNLEEMSLSHLFQF</td>
</tr>
<tr>
<td>Human U2B’</td>
<td>human U1A 102</td>
<td>EEKKEAAAATKAPAAPAAAPAAPAAPAAPAAPAAPAAPAAPAAPAAPA</td>
</tr>
<tr>
<td>Human U1A</td>
<td>human U2B’ 175</td>
<td>PGFKEVRLVPGRHDIAFVSEPDEVQAGAADAFLQFGKPTQNNMKISFAKK</td>
</tr>
<tr>
<td>URB</td>
<td>human U2B’ 154</td>
<td>PGFKEVRLVPGRHDIAFVSEPDEVQAGAADAFLQFGKPTQNNMKISFAKK</td>
</tr>
<tr>
<td>Snf</td>
<td>human U2B’ 166</td>
<td>PGFKEVRLVPGRHDIAFVSEPDEVQAGAADAFLQFGKPTQNNMKISFAKK</td>
</tr>
</tbody>
</table>

### Fig. 2 (legend on previous page)
important for recognition by human U1A and U2B* [2,14,15] are highly conserved across phyla. In particular, the AUUGCA sequence at the 5’ side of the loops is almost invariant, and loop length is typically 10–11 nucleotides. Due to the RNA conservation across phyla, it is likely that these features were also shared by the Urbilateria counterparts of these snRNA stem–loops.

The gene for URB was synthesized (GenScript) and the full-length (FL) protein and each RRM were overexpressed and purified. Representative modern U1 SLII and U2 SLIV sequences were used to characterize URB binding (Fig. 4 and Table 1). The RNA binding affinities and specificities of URB are unique, although its specificity is most similar to that of Drosophila SNF (Table 2). URB and SNF share a marked preference for SLII over SLIV, but URB binds with higher affinity to both RNAs. Human U2B* also binds to both SLII and SLIV but does not discriminate between the two RNAs, and its affinity is substantially weaker than that of URB or SNF [4]. URB and U1A bind with equally high affinity to SLII (human U1A does not detectably bind SLIV). We conclude that human U1A and U2B* have evolved away from URB to subfunctionalize RNA binding through radical changes in their RNA specificity while many of URB’s RNA binding preferences are retained by Drosophila SNF.

It is important to evaluate whether the experimental results obtained from the resurrected URB protein are robust to uncertainties in the reconstruction. Although the ML sequence is the sequence with the highest posterior probability for a given node, position 84 could be plausibly reconstructed as either Ala or Ser (Supplementary Fig. 3). This was the only site in RRM1 that had a significant alternative reconstruction. We introduced the A84S substitution into URB and found that binding to the RNA targets tested was identical with that of the ML RRM1 (Table 1).

The comparison of RNA binding affinity of U1A, U2B*, SNF, and URB illustrates an important evolutionary adaptation. U1A RRM1 binds only SLII with subnanomolar affinity in physiological solution, while the other proteins bind both SLII and SLIV. Our original hypothesis, based on our measured affinity of SNF for RNAs, was that RNA binding affinity would be compromised when an RRM bound two (slightly) different RNA targets. However, a comparison of URB binding with U1A shows that the RRM’s ability to bind both SLII and SLIV does not need to compromise the protein’s intrinsic high affinity for SLII. Rather, it is clear that, in both Drosophila and humans, the RRMs have evolved into weaker binders of one RNA or both RNAs.

Only the N-terminal RRM1 of URB binds RNA; URB C-terminal RRM2 alone does not detectably bind to SLII, SLIV, or a 25-nucleotide random pool RNA at concentrations as high as 10 μM (data not shown). The interdomain linker sequence and length is poorly conserved in the protein family, but it does contain regions with high positive charge density, typically from multiple lysine residues that could interact with the RNA backbone contributing to binding electrostatics. For URB, the difference in RNA binding affinity between the FL protein and RRM1 is modest (Table 3), indicating that RRM1 is the predominant source of RNA binding affinity. In contrast, FL SNF protein has a higher affinity for SLII RNA than does SNF RRM1 alone. In 250 mM KCl and 1 mM MgCl₂ (22 °C), the RNA binding affinity of SNF RRM1 alone for SLII is weaker than the affinity of the FL protein by ΔΔG°-binding = −3.3 ± 0.4 kcal/mol (Table 3). At lower salt concentrations, FL SNF also binds with higher affinity to SLIV than does RRM1 alone. Since SNF RRM2 does not detectably bind RNA, its linker must contribute to RNA binding affinity [5].

Because the salt dependence of RNA binding indicates the contribution of electrostatics to the association, we measured the binding of URB RRM1 and FL protein to SLII and SLIV as a function of KCl concentration and compared its properties to those of SNF (Fig. 5). A comparison of the net ions released upon RNA binding (Table 4) shows that URB/SNF RRM1 binding to SLII releases 3.2/3.1 ions while binding to SLIV releases 3.9/2.7 net ions. In addition, SNF’s linker does contribute to binding affinity, most likely through nonspecific interactions between lysines and the RNA stem. While URB’s linker has seven lysines near RRM1 (Fig. 2), they are apparently not involved in binding the stem–loops.

**Protein structure and dynamics**

Given the different RNA binding properties of URB, Drosophila SNF, human U1A, and human U2B*, we were interested in further investigating differences between the RRMs that could explain differences in binding. Homology models for URB RRM1 (modeled on existing structures of Drosophila SNF, human U1A, and human U2B*) are shown in Fig. 6a. These models predict a structure that is very similar to that of the three modern proteins. A similar alignment of homology models for URB RRM2 is also shown.

---

**Fig. 3.** RNA sequences for U1 SLII and U2 SLIV are highly conserved in metazoans, as shown in the consensus sequences for these RNAs. Input sequences and accessions (Supplementary Table 6), as well as consensus sequences broken up by clade (Supplementary Fig. 4), are available in the supplementary material.

**Table 1.** The comparison of RNA binding affinity of U1A, U2B*, SNF, and URB illustrates an important evolutionary adaptation. U1A RRM1 binds only SLII with subnanomolar affinity in physiological solution, while the other proteins bind both SLII and SLIV. Our original hypothesis, based on our measured affinity of SNF for RNAs, was that RNA binding affinity would be compromised when an RRM bound two (slightly) different RNA targets. However, a comparison of URB binding with U1A shows that the RRM’s ability to bind both SLII and SLIV does not need to compromise the protein’s intrinsic high affinity for SLII. Rather, it is clear that, in both Drosophila and humans, the RRMs have evolved into weaker binders of one RNA or both RNAs.

---

**Table 2.** The gene for URB was synthesized (GenScript) and the full-length (FL) protein and each RRM were overexpressed and purified. Representative modern U1 SLII and U2 SLIV sequences were used to characterize URB binding (Fig. 4 and Table 1). The RNA binding affinities and specificities of URB are unique, although its specificity is most similar to that of Drosophila SNF (Table 2). URB and SNF share a marked preference for SLII over SLIV, but URB binds with higher affinity to both RNAs. Human U2B* also binds to both SLII and SLIV but does not discriminate between the two RNAs, and its affinity is substantially weaker than that of URB or SNF [4]. URB and U1A bind with equally high affinity to SLII (human U1A does not detectably bind SLIV). We conclude that human U1A and U2B* have evolved away from URB to subfunctionalize RNA binding through radical changes in their RNA specificity while many of URB’s RNA binding preferences are retained by Drosophila SNF.

It is important to evaluate whether the experimental results obtained from the resurrected URB protein are robust to uncertainties in the reconstruction. Although the ML sequence is the sequence with the highest posterior probability for a given node, position 84 could be plausibly reconstructed as either Ala or Ser (Supplementary Fig. 3). This was the only site in RRM1 that had a significant alternative reconstruction. We introduced the A84S substitution into URB and found that binding to the RNA targets tested was identical with that of the ML RRM1 (Table 1).

The comparison of RNA binding affinity of U1A, U2B*, SNF, and URB illustrates an important evolutionary adaptation. U1A RRM1 binds only SLII with subnanomolar affinity in physiological solution, while the other proteins bind both SLII and SLIV. Our original hypothesis, based on our measured affinity of SNF for RNAs, was that RNA binding affinity would be compromised when an RRM bound two (slightly) different RNA targets. However, a comparison of URB binding with U1A shows that the RRM’s ability to bind both SLII and SLIV does not need to compromise the protein’s intrinsic high affinity for SLII. Rather, it is clear that, in both Drosophila and humans, the RRMs have evolved into weaker binders of one RNA or both RNAs.

Only the N-terminal RRM1 of URB binds RNA; URB C-terminal RRM2 alone does not detectably bind to SLII, SLIV, or a 25-nucleotide random pool RNA at concentrations as high as 10 μM (data not shown). The interdomain linker sequence and length is poorly conserved in the protein family, but it does contain regions with high positive charge density, typically from multiple lysine residues that could interact with the RNA backbone contributing to binding electrostatics. For URB, the difference in RNA binding affinity between the FL protein and RRM1 is modest (Table 3), indicating that RRM1 is the predominant source of RNA binding affinity. In contrast, FL SNF protein has a higher affinity for SLII RNA than does SNF RRM1 alone. In 250 mM KCl and 1 mM MgCl₂ (22 °C), the RNA binding affinity of SNF RRM1 alone for SLII is weaker than the affinity of the FL protein by ΔΔG°-binding = −3.3 ± 0.4 kcal/mol (Table 3). At lower salt concentrations, FL SNF also binds with higher affinity to SLIV than does RRM1 alone. Since SNF RRM2 does not detectably bind RNA, its linker must contribute to RNA binding affinity [5].

Because the salt dependence of RNA binding indicates the contribution of electrostatics to the association, we measured the binding of URB RRM1 and FL protein to SLII and SLIV as a function of KCl concentration and compared its properties to those of SNF (Fig. 5). A comparison of the net ions released upon RNA binding (Table 4) shows that URB/SNF RRM1 binding to SLII releases 3.2/3.1 ions while binding to SLIV releases 3.9/2.7 net ions. In addition, SNF’s linker does contribute to binding affinity, most likely through nonspecific interactions between lysines and the RNA stem. While URB’s linker has seven lysines near RRM1 (Fig. 2), they are apparently not involved in binding the stem–loops.

**Protein structure and dynamics**

Given the different RNA binding properties of URB, Drosophila SNF, human U1A, and human U2B*, we were interested in further investigating differences between the RRMs that could explain differences in binding. Homology models for URB RRM1 (modeled on existing structures of Drosophila SNF, human U1A, and human U2B*) are shown in Fig. 6a. These models predict a structure that is very similar to that of the three modern proteins. A similar alignment of homology models for URB RRM2 is also shown.
These models also predict a typical RRM whose structures are similar, with the exception of the loops, which are likely to sample multiple conformations and are difficult to model correctly.

Chemical denaturation monitored by circular dichroism shows that URB RRM1 has a folding free energy of $-5.1$ kcal/mol (Fig. 6c), which is intermediate in stability between that of U1A RRM1 \[ \Delta G^\circ_{\text{folding}} = -9.4 \text{ kcal/mol} \] \cite{16} and that of SNF RRM1 \[ \Delta G^\circ_{\text{folding}} = -3.5 \text{ kcal/mol} \] \cite{5}. The notion that stability is an evolutionarily neutral trait as long as proper folding can be maintained \cite{17} seems plausible for this protein family.

Binding was assessed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl$_2$ (pH 7 at 22 °C).

Table 1. RNA binding of FL URB and URB RRM1

<table>
<thead>
<tr>
<th></th>
<th>$K_{\text{obs,FL}}$ (M)</th>
<th>$\Delta G_{\text{FL}}$ (kcal/mol)</th>
<th>$K_{\text{obs,RRM1}}$ (M)</th>
<th>$\Delta G_{\text{RRM1}}$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URB-SLII</td>
<td>4.2 ± 0.4 × 10$^{-10}$</td>
<td>$-12.7$ ± 0.1</td>
<td>1.2 ± 0.2 × 10$^{-9}$</td>
<td>$-12.1$ ± 0.1</td>
</tr>
<tr>
<td>URB-SLIV</td>
<td>6.9 ± 2.9 × 10$^{-9}$</td>
<td>$-11.0$ ± 0.3</td>
<td>1.5 ± 0.2 × 10$^{-9}$</td>
<td>$-10.6$ ± 0.1</td>
</tr>
<tr>
<td>URB-N25$^a$</td>
<td>$-1$ × 10$^{-6}$</td>
<td>$-8$</td>
<td>$&gt;1$ × 10$^{-8}$</td>
<td>$&gt;-8$</td>
</tr>
<tr>
<td>URB A84S$^b$ SLII</td>
<td>7.9 ± 2.0 × 10$^{-10}$</td>
<td>$-12.3$ ± 0.1</td>
<td>1.8 ± 0.5 × 10$^{-8}$</td>
<td>$-10.5$ ± 0.1</td>
</tr>
</tbody>
</table>

Binding free energies of binding of RRM1 and the FL proteins are shown. Data for SNF and U1A binding were previously reported \cite{2}. All binding experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl$_2$ (pH 7) at room temperature.

$^a$ N25 is a control for nonspecific RNA binding.

$^b$ URB A84S is the second most probable reconstruction of the RRM1 sequence.
Table 2. Protein family specificity for SLII/SNV

<table>
<thead>
<tr>
<th>Protein family</th>
<th>ΔΔG°(kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URB</td>
<td>1.6 ± 0.3</td>
</tr>
<tr>
<td>SNF</td>
<td>2.4 ± 0.5</td>
</tr>
<tr>
<td>U1A</td>
<td>-6</td>
</tr>
<tr>
<td>U2B</td>
<td>0.1 ± 0.3</td>
</tr>
</tbody>
</table>

The difference in the binding free energy [ΔΔG° = ΔG°(SLII) - ΔG°(SLIV)] of the FL proteins binding SLII and human SLIV was assessed with nitrocellulose filter binding experiments in 250 mM KCl, 10 mM sodium cacodylate (pH 7.0), and 1 mM MgCl₂ at room temperature.

Table 3. Protein family RNA affinity of RRM1 versus FL protein

<table>
<thead>
<tr>
<th>Protein family</th>
<th>ΔΔG°(kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URB-SLII</td>
<td>-0.6 ± 0.1</td>
</tr>
<tr>
<td>URB-SLIV</td>
<td>-0.4 ± 0.3</td>
</tr>
<tr>
<td>U1A-SLII</td>
<td>-0 ± 0.6</td>
</tr>
<tr>
<td>SNF-SLII</td>
<td>-3.3 ± 0.4</td>
</tr>
<tr>
<td>SNF-SLIV</td>
<td>-0.7 ± 0.3</td>
</tr>
</tbody>
</table>

Difference in free energies of binding of RRM1 and FL proteins is shown. Data for SNF and U1A are from [1]. All experiments were performed in 250 mM KCl, 10 mM sodium cacodylate (pH 7), and 1 mM MgCl₂, room temperature.

Reconstruction of an Ancestral RRM

CMPG (Carr-Purcell-Meiboom-Gill) experiments can be used to quantify exchange on the millisecond-to-microsecond timescale. Figure 8 shows the difference in the effective transverse relaxation rates (ΔR₂,eff) of the amide nitrogen when pulsed at high and low CPMG field strengths (1000 and 50 Hz) for each residue. Residues showing CPMG dispersion (nonzero ΔR₂,eff) are experiencing exchange on this timescale. While residues in α₁ and Loop 5 of U1A exhibit exchange (regions distant from the RNA binding surface), resonances from the rest of the domain do not show dispersion. In contrast, both SNF and URB show evidence of fast exchange throughout much of the protein. While ΔR₂,eff is largest in α₁ and Loop 5, the RNA binding surface of both proteins exhibits significant dispersion, indicating that much of the domain, including the RNA binding surface, is exchanging between at least two conformations on the millisecond-to-microsecond timescale.

Jawed vertebrates experienced a unique evolution of U1A and U2B“

Experimental evidence has established that the RNA binding properties of URB family proteins changed significantly subsequent to the gene duplication in an ancestor of jawed vertebrates. However, the RNA binding properties of most metazoan U1A/U2B/SNF proteins have not been determined. While the sequence similarity of proteins in this family is high, we as yet have a poor grasp of the extent of functional conservation or divergence of proteins within this family. Given the results of the phylogenetic reconstruction, it is tempting to hypothesize that protein function is highly conserved in organisms with a single U1A/U2B/SNF family protein.

Functional divergence can be assessed through statistical comparisons of evolutionary rates between clusters of a phylogeny. The premise of this analysis is that functional divergence is highly correlated with changes in evolutionary rates, and
the analysis tests for differences in evolutionary rates between clusters of proteins [21]. Using functional divergence analysis on the U1A/U2B” SNF protein family shows that the functional distance between most bilaterian clusters (particularly those with a single SNF protein) is small. With the exception of nematodes and jawed vertebrates, coefficients of functional divergence are low (0.15) between other bilaterian clusters (Supplementary Table 2). In contrast, coefficients of functional divergence are much higher between gnathostome clusters and other bilaterian clusters (0.65; Supplementary Table 2 and Supplementary Fig. 6a), represented schematically in the functional divergence map in Supplementary Fig. 6b. Given the protein phylogeny, the results of the functional divergence analysis, and the experimentally determined functional similarities between Drosophila SNF and URB, it is likely that most SNF paralogs from organisms containing a single protein share very similar RNA binding properties. Our results with the resurrected URB protein indicate that these properties have been conserved since prior to the Cambrian radiation.

Functional divergence analysis can be extended to look at individual sites within the sequence and test what parts of the protein are likely to be contributing to the functional differences. Artificially engineered U1A/U2B” chimeras established β2 and Loop 3 as a region of the proteins that determined their specificity (their “specificity motif”) for either U1 SLII or U2 SLIV [2]. The RNA specificity motif of Drosophila SNF includes sequences that appear in either U1A or U2B”.

Table 4. Comparison of salt dependencies for different RNAs and proteins

<table>
<thead>
<tr>
<th></th>
<th>FL Protein</th>
<th>RRM1</th>
<th>∆∆ (Net ions released)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1A-SLII</td>
<td>–6.7 ± 1.1</td>
<td>–3.1 ± 0.5</td>
<td>2.5 ± 0.6</td>
</tr>
<tr>
<td>SNF-SLII</td>
<td>–5.7 ± 0.2</td>
<td>–3.1 ± 0.5</td>
<td>1.4 ± 0.4</td>
</tr>
<tr>
<td>SNF-SLIV</td>
<td>–4.0 ± 0.2</td>
<td>–3.2 ± 0.3</td>
<td>1.0 ± 0.4</td>
</tr>
<tr>
<td>URB-SLII</td>
<td>–4.2 ± 0.2</td>
<td>–3.9 ± 0.1</td>
<td>–0.04 ± 0.23</td>
</tr>
<tr>
<td>U2B”-SLII</td>
<td>–5.2 ± 0.3</td>
<td>–3.9 ± 0.1</td>
<td>–0.04 ± 0.23</td>
</tr>
<tr>
<td>U2B”-SLIV</td>
<td>–4.6 ± 0.3</td>
<td>–3.9 ± 0.1</td>
<td>–0.04 ± 0.23</td>
</tr>
</tbody>
</table>

The slope of the In(KA,app) versus ln([KCl]) indicates the net ions absorbed (positive) or released (negative) upon binding. Data for U1A, FL SNF, and U2B” were previously reported [2–4]. ∆∆ is the difference between the slope of the salt dependence for the FL protein and RRM1, indicating a difference in net ions released between RRM1 and the FL protein.
of an ancient origin of protein subfunctionalization, this led to the identification of SNF as a chimeric protein. However, our reconstruction of U1A/U2B″ phylogeny indicates the opposite: Drosophila SNF’s RNA specificity motif is unchanged from that of its Urbilatian ancestor. In organisms with a single protein, the RNA specificity motif has remained highly conserved (Supplementary Fig. 7). However, the RNA specificity motif evolved away from the original sequence following the gene duplication in

Fig. 6. URB is predicted to be very similar in structure to its modern descendents Drosophila SNF and human U1A and U2B″. Structural models for URB RRM1 (a) were templated from 1FHT (a U1A RRM1 solution structure) [44], 1A9N (a cocrystal structure that includes U2B″ RRM1) [52], and 2K3K (a solution structure of SNF RRM1) [18]. These models are aligned and colored by secondary structure. The RNA binding surface is indicated, as well as Loop 3. Similar models for URB RRM2 are shown (b) and were templated from 2U1A [53] and 2AYM [18]. (c) Chemical denaturation of URB RRM1. Mean residue ellipticity (deg cm$^2$ dmol$^{-1}$ residue$^{-1}$) at 221 nm is plotted as a function of urea concentration for URB RRM1 in 50 mM KCl and 10 mM sodium cacodylate (pH 7). Linear extrapolation of the data [43] gives a folding free energy of $-5.1$ kcal/mol.
an ancestor of jawed vertebrates, and the sites implicated in functional divergence are segregated to different parts of the RNA specificity motif for U1A and U2B″ proteins (Supplementary Table 3). This result is consistent with a model of subfunctionalization of SNF RNA binding properties in U1A and U2B″ proteins of jawed vertebrates.

Lophotrochozoans and nematodes

The lophotrochozoan lineage also contains a gene duplication event, and while there is evidence of functional divergence following this gene duplication (Supplementary Table 4), the functional distance from other lineages is much smaller than that of the gnathostome proteins. In particular, the RNA specificity motifs of these lophotrochozoan clusters show some indication of functional divergence, but this is much less pervasive than in jawed vertebrates. The functional distance between the paralogous lophotrochozoan clusters is smaller, indicating a substantial degree of functional similarity between these proteins. This raises the possibility that these proteins do not function similar to gnathostome U1A and U2B″. It thus appears that the extent to which the gnathostome proteins have diverged from their counterparts in other bilaterians to subfunctionalize RNA binding through adaptations of the RNA specificity motif is unique. If gene duplication in lineages other than gnathostomes result in subfunctionalization of RNA binding, it is likely that this is accomplished through different evolutionary and biochemical mechanisms.

The nematode lineage is notable for its unusual snRNA stem–loops and its U1A/U2B″/SNF proteins. Nematode sequences are distinguished in the protein phylogeny by their long branch lengths (Supplementary Fig. 2a); thus, it is not entirely surprising that the RNA loop sequences have also diverged from those of other bilaterians. snRNA stem–loops from nematodes have shorter loop sequences that sometimes lack the almost universally conserved adenine at the 5′ end of the loop. While an adenine is still the most common 5′ loop position of nematode SLII and SLIV, the decrease in conservation of this RNA feature suggests that different RNA binding mechanisms may have evolved in this lineage, preceding the gene duplication in ancestors of Caenorhabditis. Nematode RRM1 sequences have correspondingly unusual features in the RNA specificity motif. Not surprisingly, Drosophila SNF and human U1A bind to Caenorhabditis elegans U1 SLII and U2 SLIV with much weaker affinities than they bind to their natural counterparts (Table 5), which is likely to have resulted in selection against such sequences in non-nematode lineages.

Discussion

Phylogenetic analysis of the U1A/U2B″/SNF protein family has allowed us to determine that metazoans have a shared history of a single protein
whose sequence and function are highly conserved. URB is our resurrected ancestral protein, which we show to be a thermodynamically stable and soluble protein with unique RNA binding properties.

Implications of conservation of SLII and SLIV sequences for RNA–protein coevolution

The RNA loop sequences important for protein binding are highly conserved, regardless of whether an organism has one or two U1A/U2B″/SNF proteins. One of the distinguishing features between SLII and SLIV is the loop-closing base pair, which is almost universally conserved as a Watson–Crick C-G pair in SLII. The loop-closing base pair of U2 SLIV is much more variable. Most commonly, U-U or U-G is found, but G-U and C-G base pairs also exist in this position. If a gene duplication relaxes the evolutionary constraints of the protein, it is plausible that this could be accompanied by a decrease in conservation of the RNA binding sequences from the ancestral state as the protein–RNA interactions coevolve away from a single protein state. However, there are no major differences in RNA sequence conservation between organisms with separate U1A/U2B″ proteins and those with a single SNF protein. It is

---

Fig. 8. Backbone resonances showing $^{15}$N CPMG dispersion. (a) The difference in the effective $^{15}$N transverse relaxation rate between $\nu_{\text{CPMG}} = 50 \text{ Hz}$ and $\nu_{\text{CPMG}} = 1000 \text{ Hz}$ is shown for U1A (top), SNF (middle), and URB (bottom) RRM1, indicating regions experiencing millisecond-to-microsecond exchange. In (b), regions of significant CPMG dispersion are plotted onto the protein structure of each RRM.
possible that the RNA loop sequence conservation reflects the relatively recent origin of gene duplications and that, with time, it will change, too. However, it may also reflect additional evolutionary constraints on the RNA beyond U1A/U2B″ binding that are currently unknown, including pleiotropic effects.

Mechanistic implications for RNA binding

URB’s properties persist in select modern paralogs. In particular, the slow backbone dynamics of URB RRM1 are also found in SNF. Curiously or suggestively, these two proteins also bind two RNA targets with similar relative specificities. The RNA specificity motif of URB RRM1 is conserved in SNF; however, while this motif is important in determining RNA binding specificity, it is not the sole determinant of RNA binding properties: SNF and URB have identical specificity motifs but nevertheless have distinct RNA binding affinities. Finally, in none of the proteins does RRM2 contribute to RNA binding nor does it interact with RRM1. NMR data show that its backbone amides are in fast exchange both alone and in the context of the FL protein.

The distinct functional properties of human U1A, human U2B″, and Drosophila SNF have evolved in the presence of extensive structural identity; the differences at the level of primary, secondary, and tertiary structure among RRM1 of URB, U1A, U2B″, and SNF are minor. Within the common tertiary structure, the determinants of their unique RNA binding properties remain unclear. A vital contribution to RNA binding comes from the surface hydrogen bonding networks that couple the specificity motif to the RRM’s characteristic RNP motifs. These sequences are coupled to other sites on the RRM, as well as to the RNA [22,23]. We have shown that Loop 3, as part of the RNA specificity motif, is a region of extensive functional divergence in jawed vertebrates. Modulation of this region is therefore likely to alter the surface hydrogen bonding networks of the RRM. We can now use information about protein evolution to understand which mutations were responsible for altering RNA binding and the mechanisms by which these mutations modulate protein function. We predict that changes in protein dynamics and exchange properties that result from alterations to the hydrogen bonding networks are likely to be important determinants of RRM function. While the
importance of protein structure in determining function has long been appreciated, measuring protein motions and, more importantly, establishing their functional significance is an ongoing endeavor [24–32]. Establishing how these motions have been conserved and modulated is therefore critical to our understanding of these molecules.

The RNA binding properties of human U1A and U2B* have long been understood to be characteristic of the U1A/U2B*/SNF protein family. However, it is now clear that the gnathostome proteins are not at all characteristic of this family but functionally quite distant. While gene duplication in this lineage has led to distinct changes in specificity for target RNAs, it is possible that alternative functions have also emerged for these gnathostome proteins. Given our functional divergence analysis and the separate origin of gene duplications in other lineages, it is also likely that gene duplications in different lineages do not result in protein evolution toward identical functional endpoints. Indeed, experiments have shown that C. elegans U1A and U2B* are functionally redundant [33]. Transgenic expression of both human U1A and U2B* is unable to rescue the embryonic lethal phenotype of SNF knockout in flies [34]. Given these results and the small functional distance between lophotrochozoan SNF family paralogs, an intriguing question is whether functional redundancy has been retained for a specific purpose or whether these paralogs in nematodes and lophotrochozoans are still in the process of diverging.

Conclusions

Subfunctionalization following gene duplication is an important source of molecular diversity. While our work shows that the characteristics distinguishing the U1A/U2B*/SNF family’s RRM1 from most RRM—their extremely high affinity and specificity for SLII/SLIV type RNA sequences—were well established in URB, subsequent subfunctionalization of RNA binding as seen in humans is much more recent and is restricted to jawed vertebrates. We can now use our phylogenetic tree to construct intermediates in the evolution from URB to U1A/U2B* to trace the progress of their distinctive properties.

The eukaryotic spliceosome seems to have evolved its modern, complex architecture very early [35], but the presence of a single SNF protein suggests a simpler early architecture preceding the division of the eukaryotic kingdoms. That a single protein was historically found to bind both the U1 and U2 snRNAs raises the question that, in the primitive spliceosome, a single snRNP may have recognized both the 5′ splice site and the branch point, tasks that subsequently were subfunctionalized by the U1 and U2 snRNPs, respectively. Understanding the relationship between modern snRNP proteins and modern snRNAs may further define how the spliceosome evolved to its present state and elucidate fundamental aspects of the splicing reaction. While functional roles for U1A/U2B*/SNF proteins in pre-mRNA splicing have not yet been determined, it will be intriguing to understand the consequences for splicing of whether an organism has one or two SNF family proteins and, in the two-protein case, whether the consequences are lineage dependent.

Materials and Methods

Phylogenetic analysis

The protein phylogeny was inferred from extant protein sequences using ML methods. Protein sequences from broadly diverse metazoans and closely related choanozoa were obtained after BLASTing (TBLASTN) annotated/confirmed U1A, U2B*, and SNF sequences against multiple databases in National Center for Biotechnology Information. Sequences that had greater similarity to other known, RRM-containing proteins (PTB, ELAV-2, etc.) were discarded. Accession numbers of sequences used in further analysis are given in Supplementary Table 10. RRM regions were aligned using MUSCLE and had very high alignment scores when processed in T-Coffee (Core [36]) and GBLOCKS [37,38]. The aligned linker sequences were manually refined, after which they produced high alignment scores in T-Coffee and GBLOCKS. The aligned sequences were used as inputs for ProtTest [39] to identify the best model for tree reconstruction. The sequence alignments were then used for ML tree reconstruction using PhyML [12] with SPR tree searches. Because of results from the ProtTest analysis, the LG model of amino acid substitution was used, along with equilibrium amino

<table>
<thead>
<tr>
<th>Table 5. U1A and SNF binding to C. elegans (CE) RNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K_{obs,CE} (M)</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>U1A-CE II</td>
</tr>
<tr>
<td>SNF-CE II</td>
</tr>
<tr>
<td>SNF-CE IV</td>
</tr>
</tbody>
</table>

CE SLII: AUUGCACUUUU GGUGCC.
CE SLIV: CGUGGC GUUGCCGU GCCGGG.
Putative loop sequences are in boldface. \( \Delta ΔG° = ΔG°(CE) − ΔG°(hs) \); hs, Homo sapiens.
acids from the LG model. Substitution rate variation was described with a discrete gamma model, and the \( \alpha \) parameter was optimized for the data. Branch supports are given as the approximate likelihood ratio test (aLRT) statistic [40]. aLRT scores of 4.6 and 9.2 correspond to being a 10-fold or a 100-fold (respectively) more likely to observe the node than not. In Fig. 1a, the branches in the deuterostomes with aLRT scores less than 4.6 were collapsed into a polytomy, as the phylogeny was poorly resolved.

The sequence identity of the RRM2s (particularly RRM2) is much higher than ideal for phylogenetic reconstruction, as the phylogenetic signal is weak with such high sequence identity, resulting in diminished topological accuracy [12]. Inclusion of a linker alignment was therefore critical for reasonable tree topologies. Tree reconstruction is performed on gapless alignments, and because the interdomain linker is a site of many insertions and deletions, the alignable linker sequence for all proteins was relatively short. For the subgroup of deuterostome sequences, it was possible to include a longer interdomain linker alignment for tree reconstruction. This was used to resolve the topology of the deuterostome tree. Separate ML and MP trees were reconstructed for deuterostomes. The MP tree was reconstructed with ProtPars in the PHYLIP package [41].

The ML trees were rooted using Chaoanozan proteins as an outgroup for the rest of the metazoa tree or C. elegans protein sequences for the deuterostome tree. Proteins were labeled SNF if the organism contained a single U1A/U2B\(^{\prime}\)/SNF protein. In cases outside the jawed vertebrates where organisms contained two proteins, these were arbitrarily numbered 1 and 2. Trees were visualized in Archaeopteryx.

### Ancestral U1A/U2B\(^{\prime}\) synthesis

CodeML in PAML [13,42] was used for inference of the ancestral amino acid sequences of the U1A/U2B\(^{\prime}\)/SNF protein family. The ancestral sequences are inferred from the phylogeny and the modern sequences using ML methods. A marginal reconstruction was performed using similar evolutionary model parameters as those used for tree reconstruction. The ML tree obtained with PhyML was used as the input tree. The highest posterior probability sequence at the node corresponding to the last common ancestor of modern bilaterians was obtained, and this sequence was sent to GenScript for synthesis of the URB gene. The protein sequence was back-translated, and the DNA sequence was optimized for Escherichia coli codon use. NCO (5\( \prime \)) and HindIII (3\( \prime \)) restriction sites were added to the sequence for subsequent subcloning into the Ptac expression vector.

### Protein characterization

Ptac plasmids with the URB gene under an IPTG-inducible promoter were transformed into BL-21 cells for subsequent overexpression of URB proteins. Purification was performed in a manner similar to purification of Drosophila SNF proteins [5]. E. coli BL-21 cells were grown at 37 °C to an optical density of 0.8. Cells were then induced with 1 mM IPTG for 4 h at 25 °C, spun down, and kept at –80 °C. Cells were then thawed and resuspended at 4 °C in 30 mM sodium acetate (pH 5.3), 200 mM NaCl, 2 mM ethylenediaminetetraacetic acid (EDTA), and 8.5% sucrose. Sigma protease inhibitor cocktail, PMSF, and DNase II were added. The cells were French-pressed and spun down. For FL URB, the supernatant was passed over an SP Sepharose FPLC column (GE), washed with 50 mM Tris (pH 7.5) and 175 mM NaCl, and eluted over 275 mM NaCl gradient (4 h, 1.5 mL/min). Fractions with URB were collected, concentrated, and buffer-exchanged into 10 mM sodium cacodylate and 10 mM KCl (pH 7.0). Purification of URB RRM1 was similar, except that the column was pre-equilibrated in 0 M NaCl and 50 mM Tris (pH 7.5); the supernatant was washed with 0 M and 100 mM NaCl, and the salt gradient was adjusted to 100–350 mM NaCl (3 h).

### Purification of RRM2

Cells with the URB RRM2 gene were grown, induced, resuspended, and lysed in a similar manner to the other cells. Following centrifugation, the supernatant was dialyzed against 25 mM sodium acetate (pH 5.3) for 2 h. The dialysate was filtered and loaded onto an SP Sepharose FPLC column (GE) pre-equilibrated with 25 mM NaOAc (pH 5.3). The column was washed in the buffer and then eluted in a 0–250 mM NaCl gradient (3 h, 1.5 mL/min). Fractions with URB RRM2 were collected, concentrated, and buffer-exchanged into 10 mM sodium cacodylate and 10 mM KCl (pH 7.0).

### Protein circular dichroism/denaturation

A Jasco J715 spectropolarimeter was used to record CD spectra. All spectra were recorded at room temperature. The sample buffer contained 50 mM KCl and 10 mM cacodylate (pH 7), and spectra were recorded for samples with a protein concentration of 20 \( \mu \)M. CD spectra of each purified RRM (in 0 M urea) are consistent with a canonical \( \alpha/\beta \) composition.

For chemical denaturation, mean residue ellipticity at 221 nm was monitored as a function of urea concentration. The melts were fit in Scientist (Micromath) to a two-state folding model using the linear extrapolation method of Santoro and Bolen [43]. The unfolding free energy was determined from the average of two separate experiments. Uncertainty was propagated from the two experiments. This resulted in an unfolding free energy of 5.1 ± 0.4 kcal/mol.

NMR samples were prepared following E. coli growth in minimal media supplemented with either \( ^{1}H\text{NH}_{2}\text{Cl} \) or \( ^{15}H\text{NH}_{2}\text{Cl} \) and \( ^{13}C \) glucose. Proteins were purified as described above and buffer-exchanged into 50 mM KCl, 20 mM sodium cacodylate, and 2 mM EDTA at pH 6.5 with 10% D2O. For all samples, the protein concentration was 350 \( \mu \)M. Previously published assignments for SNF and U1A RRM1 were used (Biological Magnetic Resonance Bank ID 6930 and Protein Data Bank ID 1FHT [44]).

All spectra were acquired on Varian Unity Inova 500-MHz or 700-MHz spectrometers. Data were processed in NMRPipe and analyzed using nmrViewJ. URB RRM1 \( ^{1}H\) and \( ^{13}C\) assignments were made based on HNCA spectra acquired at 30 °C and similarities to SNF assignments. With the exception of I30, L46, and K85, all
non-proline amide resonances were assigned. URB 1H/15N HSQC spectra were acquired at 22.5 °C, 700 MHz; spectral widths were 7300 Hz in the 1H dimension and 2000 Hz in 15N.

For 15N-1H NOE measurements, duplicate pairs of NOE spectra were collected for SNF and U1A with and without a 3-s 1H pre-saturation. Spectra were collected at 28 °C on a 500-MHz spectrometer with 80 scans and a 3-s recycle delay. The intensity ratio was used to determine the steady-state NOE. Heteronuclear NOEs determined as SNF, U1A, and URB RRM1 was probed using CPMG delay. For all samples, a reference spectrum was acquired, 700-MHz spectrometer, with 32 scans and a 2.5-s recycle delay. The intensity ratio was used to determine the functional distance analysis, Θ is transformed into a distance.

Different gene clusters in the U1A/SNF/U2B family were analyzed for Type I and Type II functional divergence. The ML tree shown in Fig. 1a was modified such that the deuterostome branch was replaced with the deuterostome tree shown in Fig. 1b. Non-bilaterian sequences were removed, and branch lengths were re-optimized in CodeML. The protein sequence alignment and resulting tree were then used as inputs for DIVERGE 2.0. Gnathostome U1A and U2B clusters were then compared with other clusters of proteins within the tree. Based on the Type I functional divergence analysis, a functional distance map [21] of the clusters was generated.

RNA binding experiments

Nitrocellulose filter binding experiments were performed as previously described [5] to determine binding constants for protein–RNA interactions. Unless otherwise noted, all experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl2 (pH 7.0) at room temperature. Titrations were fit to a Langmuir isotherm in Scientist (Micromath). Experiments were performed in duplicate and repeated at least 2 times. Reported errors are the larger of either the standard deviation from repeat experiments or the propagated error.

RNA stem–loops were transcribed from DNA oligonucleotides (IDT) with T7 RNA polymerase, using [α-32P]UTP and [α-32P]CTP. The transcription products were gel-purified. The different RNAs were as follows:

U1 SLII: 5'-GGAGACCAUUUGAGACUUCCGGGGUUC
U2 SLIV: 5'-GGCCGGAUUUGGAGACUUCCCAGCGACGCC
CE SLII: 5'-GGCCGGAUUUGGAGACUUCCCAGCGACGCC
CE SLIV: 5'-GGCCGGAUUUGGAGACUUCCCAGCGACGCC

The underlined loop sequences for U1 SLII and U2 SLIV correspond to sequences from human (SLII) and Drosophila (SLIV). URB binding to RNAs with the Drosophila loop of SLII (AUUGGACCCUC) and the human loop-closing base pair of SLIV (U-U; human Loop IV is identical with that of Drosophila) was identical under the conditions tested [250 mM KCl, 10 mM cacodylate, and 1 mM MgCl2].
(pH 7.0, room temperature). Experiments for nonspecific binding were conducted using a 25-nucleotide random sequence pool.

**Homology modeling**

Homology modeling of the URB RRM1 structure was performed with SWISS-MODEL [51]. The RRM1 sequence was aligned with RRM1 sequences from *Drosophila* SNF and human U1A and U2B*. Structures for URB were templated from 1FHT (a U1A RRM1 solution structure) [44], 1A9N (a cocrystal structure that includes U2B* RRM1) [52], and 2K3K (a solution structure of SNF RRM1) [18]. Backbone RMSD values from the three resulting structures (excluding Helix 3) were minimized in VMD to align the structures. Models for URB RRM2 were templated from 2U1A [53] and 2AYM [18].

**Acknowledgements**

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01 GM096444 to K.B.H., F31 GM089576 to S.G.W., and F32 GM090650 to M.J.H. We thank Dr. Kim Delaney for assistance with cloning URB constructs and Dr. Gregory DeKoster for assistance with NMR spectroscopy.

**Conflict of Interest.** The authors declare no conflicts of interest.

**Supplementary Data**

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2013.05.031

Received 16 March 2013;
Received in revised form 6 May 2013;
Accepted 8 May 2013
Available online 22 June 2013

**Keywords:**
- RRM;
- ancestral reconstruction;
- RNA binding;
- protein phylogeny;
- U1A protein

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.


‡ http://rfam.sanger.ac.uk/family/RF00003
§ http://rfam.sanger.ac.uk/family/RF00004

**Abbreviations used:**
- snRNP, small nuclear ribonucleoprotein;
- RRM, RNA recognition motif;
- ML, maximum likelihood;
- MP, maximum parsimony;
- FL, full length;
- HSQC, heteronuclear single quantum coherence;
- NOE, nuclear Overhauser enhancement;
- aLRT, approximate likelihood ratio test;
- EDTA, ethylenediaminetetraacetic acid.

**References**


null
Supplementary Material.

Supplementary figure 1.

Supplementary figure 2.

Supplementary figure 3.

Supplementary figure 4.

Supplementary figure 5.

Supplementary figure 6.

Supplementary figure 7.

Supplementary table I.

Supplementary table II.

Supplementary table III.

Supplementary table IV.

Supplementary table V.

Supplementary table VI.
Supplementary Figure 1. Ancestral reconstruction process.
Supplementary Figure S2A. Maximum likelihood tree for U1A/U2B*/SNF family. The tree was reconstructed using PhyML, as described in the Methods and shows SH-aLRT branch supports (aLRT with the non-parametric SH correction). A marker of the branch lengths is shown. Accessions for the input sequences are shown in Supplementary Table V.
Supplementary figure S2B. ML tree (PhyML) (PROTPARS in PHYLIP) for deuterostome U1A/U2B^*/SNF protein sequences. The ML tree shows SH-aLRT branch supports.
Supplementary Figure 2C. Maximum parsimony tree (PROTPARS in PHYLIP) for deuterostome U1A/U2B*/SNF protein sequences.
Supplementary figure 3. Highest posterior probabilities mapped onto a model of the protein structure. The legend shows posterior probability of the most likely residue at each site in the sequence, color-coded by the posterior probability of observing the amino acid. Dark blue indicates the posterior probability of observing the most likely amino acid is close to 1. The linker has the greatest number of ambiguously determined residues.
Supplementary Figure 4. Clustered RNA consensus sequences. Consensus sequences for the loop and loop-closing basepair of U1 SLII and U2 SLIV are shown for a variety of organisms. Here they are grouped as indicated by the brackets on the tree at the left. Sequences and accessions for individual species are given in ST VI.
Supplementary figure 5. 1H/15N HSQC of URB RRM1 with assigned crosspeaks labeled.

Chemical shifts of the amide $^1$H and $^{15}$N are shown in ppm.
Supplementary Figure 6a. Coefficients for Type I ($\Theta_I$) and Type II ($\Theta_{II}$) functional divergence between gnathostome U1A proteins, U2B″ proteins, and other U1A/U2B″/SNF proteins in bilaterians show that vertebrate U1A and U2B″ clusters are functionally divergent from each other and from other bilaterian proteins in the same family. This is indicated by $\Theta \neq 0$. Line color indicates which clusters are compared; blue is used for vertebrate U2B″, red for vertebrate U1A, and black for remaining proteins.
Supplementary Figure 6b. The functional distance\textsuperscript{22} between different bilaterian clusters is based on coefficients of Type I functional divergence (ST II). This map shows vertebrate U1A and U2B\textsuperscript{*} to be functionally distant from each other, as well as from other clusters of U1A/U2B\textsuperscript{*}/SNF proteins. In contrast, proteins from arthropods, lophotrochozoans, and other deuterostomes are functionally very close to each other.
Supplementary Figure 7. RNA specificity motif sequences. Sequences from organisms with a single U1A/U2B*/SNF protein are shown on the right. Sequences from organisms with two U1A/U2B*/SNF proteins are shown on the left. Sites that have diverged from the Urbilaterial state are boldfaced.
Supplementary Table I. Posterior probabilities for different amino acids at the Urbilaterian node.

Residue Number

<table>
<thead>
<tr>
<th>Residue Number</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>I(0.998) V(0.002)</td>
</tr>
<tr>
<td>18</td>
<td>I(0.997) V(0.003)</td>
</tr>
<tr>
<td>21</td>
<td>D(0.998) E(0.001)</td>
</tr>
<tr>
<td>43</td>
<td>L(0.996) M(0.004)</td>
</tr>
<tr>
<td>60</td>
<td>S(0.977) N(0.019) T(0.004)</td>
</tr>
<tr>
<td>84</td>
<td>A(0.783) S(0.216) C(0.001)</td>
</tr>
<tr>
<td>86</td>
<td>T(0.999) S(0.001)</td>
</tr>
<tr>
<td>91</td>
<td>I(0.929) V(0.071)</td>
</tr>
<tr>
<td>98</td>
<td>F(0.951) Y(0.049)</td>
</tr>
<tr>
<td>100</td>
<td>E(0.997) P(0.002) D(0.001)</td>
</tr>
<tr>
<td>102</td>
<td>E(0.992) P(0.005) D(0.002) Q(0.001)</td>
</tr>
<tr>
<td>104</td>
<td>R(0.033) K(0.967)</td>
</tr>
<tr>
<td>105</td>
<td>K(0.984) R(0.014) P(0.001) Q(0.001)</td>
</tr>
<tr>
<td>106</td>
<td>E(0.997) D(0.002) K(0.001)</td>
</tr>
<tr>
<td>107</td>
<td>K(0.997) Q(0.001) R(0.002)</td>
</tr>
<tr>
<td>108</td>
<td>K(0.835) E(0.155) Q(0.005) R(0.002) T(0.001) A(0.001)</td>
</tr>
<tr>
<td>109</td>
<td>A(0.392) K(0.248) N(0.006) D(0.004) G(0.003) E(0.0074) R(0.006) P(0.043) Q(0.023) T(0.019)</td>
</tr>
<tr>
<td>110</td>
<td>K(0.939) N(0.001) E(0.031) Q(0.012) R(0.006) A(0.006) T(0.003) S(0.002)</td>
</tr>
<tr>
<td>112</td>
<td>A(0.940) S(0.041) E(0.012) Q(0.003) T(0.002) D(0.001) K(0.001)</td>
</tr>
<tr>
<td>113</td>
<td>K(0.566) E(0.306) P(0.002) S(0.001)</td>
</tr>
<tr>
<td>114</td>
<td>P(0.787) Q(0.101) E(0.091) A(0.016) S(0.003) D(0.001) K(0.001)</td>
</tr>
<tr>
<td>115</td>
<td>A(0.997) S(0.002) S(0.001)</td>
</tr>
<tr>
<td>116</td>
<td>A(0.966) G(0.027) S(0.005)</td>
</tr>
<tr>
<td>117</td>
<td>P(0.999) Q(0.001)</td>
</tr>
<tr>
<td>119</td>
<td>P(0.600) A(0.371) S(0.025) G(0.003) T(0.001)</td>
</tr>
<tr>
<td>120</td>
<td>A(0.772) G(0.157) S(0.043) P(0.026) T(0.001)</td>
</tr>
<tr>
<td>121</td>
<td>A(0.950) T(0.025) S(0.020) V(0.003) G(0.001)</td>
</tr>
<tr>
<td>122</td>
<td>A(0.993) Q(0.004) S(0.003)</td>
</tr>
<tr>
<td>123</td>
<td>V(0.883) I(0.116) L(0.001)</td>
</tr>
<tr>
<td>130</td>
<td>H(0.504) Q(0.496) N(0.001)</td>
</tr>
<tr>
<td>177</td>
<td>V(0.912) I(0.062) M(0.015) T(0.009) L(0.001) A(0.001)</td>
</tr>
<tr>
<td>179</td>
<td>A(0.963) S(0.037)</td>
</tr>
<tr>
<td>183</td>
<td>K(0.901) R(0.099)</td>
</tr>
<tr>
<td>195</td>
<td>H(0.998) N(0.002)</td>
</tr>
<tr>
<td>200</td>
<td>S(0.993) T(0.007)</td>
</tr>
<tr>
<td>201</td>
<td>F(0.998) Y(0.002)</td>
</tr>
</tbody>
</table>

Posterior probabilities for different amino acids at the Urbilaterian node are generally high, indicating confidence in the reconstructed sequence. Averages for the posterior probabilities of RRM1, RRM2, and the interdomain linker are 0.996, 0.991, and 0.884, respectively. It should be noted that the interdomain linker is a site of frequent insertions and deletions, and the alignments used for the sequence reconstruction were significantly condensed from the full linker sequence. In the table, sites not listed have a single predicted amino acid with a posterior probability of 1. Sites with a second amino acid with a posterior probability >0.2 are boldfaced.
### Supplementary Table II.

(a) Coefficient for functional divergence for gnathostome U1A & other clusters:

<table>
<thead>
<tr>
<th></th>
<th>Other deuterostomes</th>
<th>Arthropods</th>
<th>Lophotrochozoa</th>
<th>U2B’’</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_I$</td>
<td>0.810 ± 0.088</td>
<td>0.683 ± 0.097</td>
<td>0.822 ± 0.114</td>
<td>0.762 ± 0.075</td>
</tr>
<tr>
<td>LR</td>
<td>86</td>
<td>49</td>
<td>52</td>
<td>104</td>
</tr>
<tr>
<td>$\Theta_{II}$</td>
<td>0.204 ± 0.063</td>
<td>0.180 ± 0.078</td>
<td>0.390 ± 0.101</td>
<td>0.223 ± 0.060</td>
</tr>
</tbody>
</table>

(b) Coefficient for functional divergence for gnathostome U2B’’ & other clusters:

<table>
<thead>
<tr>
<th></th>
<th>Other deuterostomes</th>
<th>Arthropods</th>
<th>Lophotrochozoa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_I$</td>
<td>0.653 ± 0.079</td>
<td>0.657 ± 0.068</td>
<td>0.814 ± 0.110</td>
</tr>
<tr>
<td>LR</td>
<td>68</td>
<td>94</td>
<td>55</td>
</tr>
<tr>
<td>$\Theta_{II}$</td>
<td>0.228 ± 0.076</td>
<td>0.206 ± 0.090</td>
<td>0.377 ± 0.118</td>
</tr>
</tbody>
</table>

(c) Coefficient for functional divergence for remaining clusters:

<table>
<thead>
<tr>
<th></th>
<th>Other deuterostomes/arthropods</th>
<th>Other deuterostomes/Lophotrochozoa</th>
<th>Arthropods/Lophotrochozoa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_I$</td>
<td>0.149 ± 0.051</td>
<td>0.024 ± 0.082</td>
<td>0.094 ± 0.048</td>
</tr>
<tr>
<td>LR</td>
<td>8.50</td>
<td>0.088</td>
<td>3.83</td>
</tr>
<tr>
<td>$\Theta_{II}$</td>
<td>0.037 ± 0.101</td>
<td>0.046 ± 0.173</td>
<td>0.086 ± 0.165</td>
</tr>
</tbody>
</table>

(d) Coefficient for functional divergence for nematodes & other clusters:

<table>
<thead>
<tr>
<th></th>
<th>U1A</th>
<th>U2B’’</th>
<th>Other deuterostomes</th>
<th>Arthropods</th>
<th>Lophotrochozoa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_I$</td>
<td>0.999 ± 0.122</td>
<td>0.929 ± 0.100</td>
<td>0.372 ± 0.091</td>
<td>0.492 ± 0.067</td>
<td>0.150 ± 0.071</td>
</tr>
<tr>
<td>LR</td>
<td>67</td>
<td>86</td>
<td>17</td>
<td>54</td>
<td>4.52</td>
</tr>
<tr>
<td>$\Theta_{II}$</td>
<td>0.322 ± 0.087</td>
<td>0.243 ± 0.106</td>
<td>0.082 ± 0.129</td>
<td>0.143 ± 0.129</td>
<td>0.217 ± 0.249</td>
</tr>
</tbody>
</table>

Maximum-likelihood estimates of the coefficients of Type I and Type II functional divergence between different bilaterian clusters show that gnathostome U1A and U2B’’ clusters have experienced significant functional divergence from other branches of the U1A/U2B’’/SNF tree (a and b). The cluster of nematode proteins is also functionally divergent from other groups. However, functional divergence between proteins from arthropods, lophotrochozoans, and deuterostomes with a single SNF protein is small ($\Theta_I$ close to 0) (c). Likelihood ratios (LR) from testing whether $\Theta_I$ is significantly different from 0 are indicated.
## Supplementary table III.

<table>
<thead>
<tr>
<th>URB Residue</th>
<th>Residue #</th>
<th>U2B/U1A</th>
<th>U2B/Other</th>
<th>U1A/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>36</td>
<td>0.79</td>
<td>0.63</td>
<td>0.80</td>
</tr>
<tr>
<td>I</td>
<td>37</td>
<td>0.96</td>
<td>0.98</td>
<td>0.63</td>
</tr>
<tr>
<td>L</td>
<td>38</td>
<td>0.99</td>
<td>0.70</td>
<td>0.58</td>
</tr>
<tr>
<td>D</td>
<td>39</td>
<td>0.94</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>I</td>
<td>40</td>
<td>0.80</td>
<td>0.98</td>
<td>0.58</td>
</tr>
<tr>
<td>V</td>
<td>41</td>
<td>0.72</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>A</td>
<td>42</td>
<td>0.72</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>L</td>
<td>43</td>
<td>0.44</td>
<td>0.37</td>
<td>0.36</td>
</tr>
<tr>
<td>K</td>
<td>44</td>
<td>0.99</td>
<td>0.59</td>
<td>0.99</td>
</tr>
<tr>
<td>T</td>
<td>45</td>
<td>1.00</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>L</td>
<td>46</td>
<td>0.27</td>
<td>0.97</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Site-specific posterior probabilities obtained from functional divergence analysis show that the RNA specificity motif is a region of significant Type I functional divergence when comparing gnathostome U1A, U2B”, and the rest of the bilaterian tree (Other). Sites with posterior probabilities greater than 0.75 are colored red and show that different regions of the RNA specificity motif are functional divergence related when comparing U1A and U2B” with the rest of the tree. Residue numbering corresponds to that of URB.

<table>
<thead>
<tr>
<th>Residue</th>
<th>U2B/U1A</th>
<th>U2B/Other</th>
<th>U1A/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>38</td>
<td>26.75</td>
<td>-0.86</td>
<td>0.00</td>
</tr>
<tr>
<td>39</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>41</td>
<td>1325.40</td>
<td>0.00</td>
<td>-1.51</td>
</tr>
<tr>
<td>42</td>
<td>1325.40</td>
<td>0.00</td>
<td>-3.63</td>
</tr>
<tr>
<td>43</td>
<td>21.72</td>
<td>0.00</td>
<td>-0.30</td>
</tr>
<tr>
<td>44</td>
<td>54.40</td>
<td>0.00</td>
<td>-1.78</td>
</tr>
<tr>
<td>45</td>
<td>37.48</td>
<td>0.00</td>
<td>-0.89</td>
</tr>
<tr>
<td>46</td>
<td>22.29</td>
<td>-1.38</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Site-specific posterior ratios within the RNA specificity motif for Type II functional divergence between gnathostome U1A, U2B”, and the rest of the bilaterian tree (Other). Much of the RNA specificity motif shows strong evidence for Type II functional divergence between U1A and U2B”, indicating that at these sites, while the characters are different, the extent of conservation is similar. Sites with posterior ratios greater than 1 are colored red.
Supplementary Table IV.
Coefficients of Type I and Type II functional divergence.

<table>
<thead>
<tr>
<th></th>
<th>AP1/AP2</th>
<th>AP1/Other</th>
<th>AP2/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Θ_I</td>
<td>0.352 ± 0.196</td>
<td>0.304 ± 0.073</td>
<td>0.474 ± 0.065</td>
</tr>
<tr>
<td>LR</td>
<td>3.241</td>
<td>17.294</td>
<td>53.465</td>
</tr>
<tr>
<td>Θ_{II}</td>
<td>-0.166 ± 0.179</td>
<td>0.000 ± 0.151</td>
<td>0.108 ± 0.169</td>
</tr>
</tbody>
</table>

Analysis of functional divergence between putative annelid/platyhelminth U1A (AP1) or U2B" (AP2) protein clusters and the rest of the bilaterian tree (Other). A gene duplication occurred in the annelid/platyhelminth lineage, providing an opportunity for functional divergence between these groups. While there is evidence of Type I functional divergence, values of Θ_I are smaller than those seen with gnathostome proteins (S7). Likelihood ratios (LR) from tests of Θ_I ≠ 0 are given for coefficients of Type I functional divergence.

<table>
<thead>
<tr>
<th>Residue</th>
<th>AP1/AP2</th>
<th>AP1/Other</th>
<th>AP2/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>0.17</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td>37</td>
<td>0.41</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>38</td>
<td>0.41</td>
<td>0.62</td>
<td>0.32</td>
</tr>
<tr>
<td>39</td>
<td>0.34</td>
<td>0.21</td>
<td>0.38</td>
</tr>
<tr>
<td>40</td>
<td>0.29</td>
<td>0.26</td>
<td>0.41</td>
</tr>
<tr>
<td>41</td>
<td>0.35</td>
<td>0.22</td>
<td>0.37</td>
</tr>
<tr>
<td>42</td>
<td>0.62</td>
<td>0.77</td>
<td>0.41</td>
</tr>
<tr>
<td>43</td>
<td>0.34</td>
<td>0.27</td>
<td>0.72</td>
</tr>
<tr>
<td>44</td>
<td>0.34</td>
<td>0.22</td>
<td>0.37</td>
</tr>
<tr>
<td>45</td>
<td>0.34</td>
<td>0.20</td>
<td>0.49</td>
</tr>
<tr>
<td>46</td>
<td>0.41</td>
<td>0.58</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Site-specific posterior probabilities within the RNA specificity motif for Type I functional divergence show that the RNA specificity motif is much less implicated in functional divergence following the gene duplication in lophotrochozoans than it is in jawed vertebrates. This is further support that the functional consequences of the gene duplication in jawed vertebrates are unique to this lineage. In lophotrochozoans, there is only one site in the RNA specificity motif that is likely to be functionally divergent between the U1A and U2B" proteins. Sites with posterior probabilities greater than 0.6 are colored red.
**Supplementary Table V.** A. Accessions for protein sequences used in phylogenetic analysis.

<table>
<thead>
<tr>
<th>Organism</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acropora millepora</td>
<td>EZ048694</td>
</tr>
<tr>
<td>Acrithosiphon pisum</td>
<td>NP_001156146; ACTA01068414, ACTA01060414;</td>
</tr>
<tr>
<td>Ailuropoda melanoleuca</td>
<td>ACTA01195147</td>
</tr>
<tr>
<td>Alcyonidium diaphanum</td>
<td>GW338016</td>
</tr>
<tr>
<td>Alvinella pompejana</td>
<td>GO120435</td>
</tr>
<tr>
<td>Ambystoma mexicanum</td>
<td>CO778357</td>
</tr>
<tr>
<td>Amphimedon queenslandica</td>
<td>GW154872</td>
</tr>
<tr>
<td>Anas platyrhynchos</td>
<td>DR766439</td>
</tr>
<tr>
<td>Anemonia viridis</td>
<td>FK730358</td>
</tr>
<tr>
<td>Anolis carolinensis</td>
<td>AAWZ02015054; FG757571</td>
</tr>
<tr>
<td>Anopheles darlingi</td>
<td>EFR20150</td>
</tr>
<tr>
<td>Apis florea</td>
<td>AEKZ01009701</td>
</tr>
<tr>
<td>Aplysia californica</td>
<td>AASC02037840</td>
</tr>
<tr>
<td>Artemia franciscana</td>
<td>ES513318</td>
</tr>
<tr>
<td>Asterina pectinifera</td>
<td>DB394317</td>
</tr>
<tr>
<td>Atta cephalotes</td>
<td>ADTU01020803</td>
</tr>
<tr>
<td>Biomphalaria glabrata</td>
<td>EX001526</td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>Q3H40</td>
</tr>
<tr>
<td>Botryllus schlosseri</td>
<td>JG345514</td>
</tr>
<tr>
<td>Branchiostoma floridum</td>
<td>XP_002608892</td>
</tr>
<tr>
<td>Brugia malayi</td>
<td>XM_001898968</td>
</tr>
<tr>
<td>Bursaphelenchus</td>
<td></td>
</tr>
<tr>
<td>C. briggsae</td>
<td>CAP32598; CAP32599</td>
</tr>
<tr>
<td>C. elegans</td>
<td>CCD72714; CCD72715</td>
</tr>
<tr>
<td>Caligus clemensi</td>
<td>ACO15476</td>
</tr>
<tr>
<td>Caligus rogercressei</td>
<td>BT076725</td>
</tr>
<tr>
<td>Camponotus floridanus</td>
<td>AEAB01001334</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>XM_533663, XM_852660</td>
</tr>
<tr>
<td>Capitella teleta</td>
<td>EY555728</td>
</tr>
<tr>
<td>Carcinus maenas</td>
<td>DV111425</td>
</tr>
<tr>
<td>Carteriospongia foliascens</td>
<td>GO083435</td>
</tr>
<tr>
<td>Cavia porcellus</td>
<td>AAKN02054814</td>
</tr>
<tr>
<td>Clytia hemisphaerica</td>
<td>FP962533, FP969551</td>
</tr>
<tr>
<td>Convoluta pulchra</td>
<td>EV601540</td>
</tr>
<tr>
<td>Crassostrea gigas</td>
<td>CU990277</td>
</tr>
<tr>
<td>Culex quinquefasciatus</td>
<td>XP_001869484</td>
</tr>
<tr>
<td>Cynoglossus semilaevis</td>
<td>GH230755</td>
</tr>
<tr>
<td>Cynops pyrrhogaster</td>
<td>FS299517</td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>ACJG01004100</td>
</tr>
<tr>
<td>Dasypus novemcinctus</td>
<td>AAGV020854559, AAGV020030187</td>
</tr>
<tr>
<td>Dermacentor andersoni</td>
<td>EG363225</td>
</tr>
<tr>
<td>Dicentrarchus labrax</td>
<td>CABK01022450; FM010807</td>
</tr>
<tr>
<td>Dirofilaria immitis</td>
<td>BQ482073</td>
</tr>
<tr>
<td>Dissostichus mawsoni</td>
<td>FE195215</td>
</tr>
<tr>
<td>Drosophila</td>
<td>P43332</td>
</tr>
</tbody>
</table>
melanogaster
Drosophila mojavensis XP_002011511
Dugesia japonica BP188316
Echinococcus granulosus CN649625
Eptatretus burgeri BJ653196
Eriocheir sinensis FG358008
Euprymna scolopes DW272106
Fasciola gigantica FN382429
Felis catus AANG02066291
Fundulus heteroclitus EV460328
Gadus morhua FF411832; GW862495
Gallus gallus BU217107
Gekko japonicus EB168095
Haemonchus contortus CB012287
Haliotis asinine GT275982
Halocynthia roretzi DB628174
Helobdella robusta EY362537; EY309972
Herdmannia momus EL733013
Heterodera glycines CB281783, CB299297
Heterorhabditis bacteriophora EG025450
Hippoglossus hippoglossus EB034728, EB036449
Hirudo medicinalis FP608099, FP655814
Homo sapiens P09012, P08579,
Hydra magnipapillata XP_002166796
Hyriopsis cumingii GW694963
Ictalurus punctatus CB938640; FD177029
Ixodes scapularis EW951007
Lepeophtheirus salmonis BT077965
Lernaeocera branchialis GO415835
Leucoraja erinacea AESE012542282, AESE010582384; EE988713
Lipochromis DB868131
Litomosoides sigmodontis DN557960
Loa loa XM_003137831
Lottia gigantia FC716649
Lymnaea stagnalis ES291772
Mayetiola destructor AEGAO1004651
Meleagris gallopavo EX718534
Meloidogyne incognita BQ625325
Mizuhopecten yessoensis GR867202
Molgula tectiformis CJ335932
Monodelphis domestica XM_001371577, XM_001374193
Monosiga brevicollis XM_001750498; XM_001743768
Monosiga ovata EC166548; EC167763
Mus musculus Q9CQ17
Mytilus californianus ES404759
Nematostella vectensis  FC197233
Neobenedenia melleni  GW920266
Notophthalmus viridescens  GO928930
Oikopleura dioica  CBY30800; CBY13172
Oreochromis niloticus  GR619518, GR698177
Ornithorhynchus anatinus  XP_001517530
Oryctolagus cuniculus  Q6TU31
Oscarella lobularis  AM764161
Osmerus mordax  EL527514
Paracentrotus lividus  AM554423
Paragonimus westermani  AT007359
Parasteatoda tepidariorum  FY220251; FY222814
Pediculus humanus  AAZO01007024
Penaeus monodon  GW999592
Perca flavescens  GO660012
Petrolisthes cinctipes  FE828228
Petromyzon marinus  FD707013
Pig  Q06AA4, Q06A96
Pimephales promelas  DT124339, DT315850
Poecilia reticulate  ES374266; ES385823
Pongo abelii  Q5R5E3
Porites astreoides  GE912403
Pratylenchus vulnus  EL887864
Priapulus caudatus  FD659912
Pristionchus pacificus  FE936495
Procavia capensis  ABRQ01088214, ABRQ01088213,
Abbrav288212
Python molurus  AEQU010290590, AEQU010296162
Rana catesbeiana  GO466192, GO472856
Rattus norvegicus  Q5U214
Rhipicephalus microplus  CK176008
Rhodnius prolixus  ACPB02046449
Saccoglossus kowalevskii  XP_002737237
Salmo salar  DW536215, DY711266
Salpingoea  ACSY01000297; ACSY01001408
Schistosoma japonicum  FN314510; Q5DD61
Schmidtea mediterranea  AAWT01069907; AAWT01016515
Scylla paramamosain  FJ774874
Sebastes caurinus  GE806616
Sparus aurata  AM970326
Squalus acanthias  ES606621
Strongylocentrotus purpuratus  XP_001202981
Strongyloides ratti  CACX01001886
Taenia solium  EL761369
Taeniopygia guttata  ABQF01072336, XP_002199945
Tetranychus urticae   GT993501
Tigriopus californicus  FN250640
Torpedo californica  EW690270, EW692457
Tribolium castaneum   XP_968271
Trichinella spiralis   ES566809
Trichoplax adhaerens   XM_002107782
Trichuris muris   FF143725
Tubifex tubifex   EY438538
Tursiops truncatus  ABRN01324180; ABRN02032000
Varroa destructor  ADDG01022410
Wuchereria bancrofti  CK855436
Xenopus levis   P45429; Q5XHF0
Xenopus tropicalis  Q5BL54; NP_001011120

Ensembl identifier
Bos taurus  ENSBTAP000000011963, ENSBTAP00000039319
Ciona intestinalis  ENSCINP00000019335
Cavia porcellus  ENSCPOG00000004514
Ciona saviny  ENSCSAVP00000000709
Echinops telfairi  ENSETEP000000015600, ENSETEP00000000877
Felis catus  ENSFCAP000000015063
Gasterosteus aculeatus  ENSGACP00000013006, ENSGACP00000026011
Gallo gallo  ENSGALP00000031344
Pan troglodytes  ENSPTRP00000018887, ENSPTRP00000022762
Danio rerio  ENSDARP00000057600, ENSDARP00000027914
Macaca mulata  ENSMMUP00000019028
Mus musculus  ENSMUSP00000092248
Ornithorhynchus anatinus  ENSOANP00000001971
Oryctolagus cuniculus  ENSOCUP00000001227
Oryzias latipes  ENSORLP00000008203, ENSORLP00000024176
Rattus norvegicus  ENSRNOP00000006900
Tupaia belangeri  ENSTBEP00000011023, ENSTBEP00000011168
Tetraodon nigroviridis  GSTENP000014781001, GSTENP00003842001
**Supplementary Table VI.** RNA U1 SLII and U2 SLIV sequences and accessions for consensus sequences.

<table>
<thead>
<tr>
<th>U1 snRNA SLII</th>
<th>Stem-loop sequence</th>
<th>Genebank or EMBL Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human, rat</td>
<td>UUAUCCAUGCAGUCUCCGGAGUGU</td>
<td>M14585; X12933</td>
</tr>
<tr>
<td>Frog</td>
<td>UCAAGCCAUUGCACUCGCCGUAGU</td>
<td>X02585</td>
</tr>
<tr>
<td>Anolis carolinesis</td>
<td>UCAUCCAUGCAGUCUCCGGGAGU</td>
<td>AAWZ02017214</td>
</tr>
<tr>
<td>Python molurus</td>
<td>UCAACCUUGCACUCGGUGGU</td>
<td>AEQU010647801</td>
</tr>
<tr>
<td>Taeniopygia guttata</td>
<td>UCAUICCUUGCACUCGGGUGUG</td>
<td>AADN02018455</td>
</tr>
<tr>
<td>Chicken, turkey</td>
<td>UCAUCCUGACACUCGGGGUGUG</td>
<td>AADN02018455; ADDD01057309</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>UUGGCAUGCAGUCUCGCGAGUG</td>
<td>CA590338</td>
</tr>
<tr>
<td>Sparus aurata, Tilapia</td>
<td>UCAGCGCAGCAGCAGCGUUGUG</td>
<td>FR530182; GR652160</td>
</tr>
<tr>
<td>Salmon salar, Takifugu rubripes</td>
<td>UCAGCGCAGCAGCAGCGUUGUG</td>
<td>ET359474; CA590338</td>
</tr>
<tr>
<td>Oryzias latipes</td>
<td>UCAGCGCAGCAGCAGCGUUGUG</td>
<td>DH065676</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
<td>UCGGCAUGCAGUCUCCGGCAGUG</td>
<td>AL244821</td>
</tr>
<tr>
<td>Dissostichus mawsoni</td>
<td>UCGGCAUGCAGUCUCCGGCAGUG</td>
<td>FE211741</td>
</tr>
<tr>
<td>Fundulus heteroclitus</td>
<td>UCGGCAUGCAGUCUCCGGCAGUG</td>
<td>DN951756</td>
</tr>
<tr>
<td>Ambystoma mexicanus</td>
<td>UCAUCCAUGCAGAUGCGUUGUG</td>
<td>M59827</td>
</tr>
<tr>
<td>Chrysemys pic. (turtle)</td>
<td>UCAUCCAUGCAGUCCGGGUGUG</td>
<td>AC239217</td>
</tr>
<tr>
<td>+lamprey</td>
<td>UCAUCCAUGCAGUCCGGGUGUG</td>
<td>In Supplementary Material (S1) of 55</td>
</tr>
<tr>
<td>Branchiostoma floridae</td>
<td>CCGGCAUGCAGUCCGGGUGUG</td>
<td>NW_003101437</td>
</tr>
<tr>
<td>Ciona</td>
<td>CAGGCCAUUGCACAUGCGCUUA</td>
<td>AJ227341</td>
</tr>
<tr>
<td>Botryllus schlosseri</td>
<td>CUGGCCAUUGCACAUGCGCUAG</td>
<td>JG376609</td>
</tr>
<tr>
<td>Molgula tectiformis</td>
<td>CUUCUAUGCAGACUCGCGAGUG</td>
<td>CJ374783</td>
</tr>
<tr>
<td>Halocynthia roretzi</td>
<td>CUGAUCUUGCAGUGCGGUGUG</td>
<td>DB583845</td>
</tr>
<tr>
<td>Oikopleura dioica</td>
<td>CGUAUGCAGACUCCGGGUGGUG</td>
<td>FP869737</td>
</tr>
<tr>
<td>Saccoglossus kowal</td>
<td>CUUCUAUGCAGUCCGGGUGUC</td>
<td>NW_003112883</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>UUGGCCAUUGCACCACGCGUAG</td>
<td>K00787</td>
</tr>
<tr>
<td>Anopheles gamb</td>
<td>UGGGCCAUUGCACCACUUGGGCUGG</td>
<td>ABKQ02022061</td>
</tr>
<tr>
<td>Culex</td>
<td>CUGGCCAUUGCAGUCUCCGGU</td>
<td>NW_001887363</td>
</tr>
<tr>
<td>Mayetiola destructor</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>AEG01027640</td>
</tr>
<tr>
<td>Tribolium castaneum</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>NC_007418</td>
</tr>
<tr>
<td>Acrythosiphon pismus</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>NW_003383501</td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>BP178451</td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>NW_002987798</td>
</tr>
<tr>
<td>Camponatus floridanus; Atta cephalotes; Apis florea</td>
<td>CUGAUCUUGCAGACUUGCGAGUG</td>
<td>AEAB01002246; ADTU01004990; AEKZ01000909; ACPB02019542; NW_002847753; ADDG01080938</td>
</tr>
<tr>
<td>Species</td>
<td>Sequence</td>
<td>Accession</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Paryhale hawaiensis</td>
<td>CUCACCAUUGCACUC CGGUGUUG</td>
<td>FN434129</td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>CUUUCAUUGCACUA AGGAUGGG</td>
<td>FE344712</td>
</tr>
<tr>
<td>Lepeophtheirus salmonis</td>
<td>CUUUCAUUGCACUUCGGUGGCC</td>
<td>GW644660</td>
</tr>
<tr>
<td>Penaeus monodon</td>
<td>CUCACCAUUGCACUC CGGUGGCG</td>
<td>JJ745050</td>
</tr>
<tr>
<td>Artemia franciscana</td>
<td>CCCCCAUUGCACUUCCGGUGGGC</td>
<td>GW644660</td>
</tr>
<tr>
<td>*Proasellus coxalis</td>
<td>CCCCCAUUGCACUUCGGUGG</td>
<td>AY198214</td>
</tr>
<tr>
<td>*Asellus aquaticus</td>
<td>CUCACCAUUGCACUC CGGUGGCG</td>
<td>AY198212</td>
</tr>
<tr>
<td>+Lottia gigantia</td>
<td>UUCUGCAUUGCACCUAGCGGAG</td>
<td>AZZI43324</td>
</tr>
<tr>
<td>Aplysia californica</td>
<td>CUUUCAUUGCACUCCGGUUUG</td>
<td>EB201145</td>
</tr>
<tr>
<td>Mytilus mytilus</td>
<td>UUGGCCAUUGCACUUUUCGGUCCG</td>
<td>AM880502</td>
</tr>
<tr>
<td>Crepidula fornicata</td>
<td>UUGGCCAUUGCACUUUUCGGUCCG</td>
<td>EZ556252</td>
</tr>
<tr>
<td>Ensifera siliqua, Pharus legumen, Elysia timida</td>
<td>CUCUCCAUUGCACUCGGUUUGG</td>
<td>FN908890; FN908891; HP162468</td>
</tr>
<tr>
<td>C. elegans</td>
<td>CACACCAUUGCACUUU UGGUGCG</td>
<td>X51371</td>
</tr>
<tr>
<td>C. briggsae</td>
<td>CUAUCCAUCACUUU AGGAUGG</td>
<td>NC_013488</td>
</tr>
<tr>
<td>Heterodera glycines</td>
<td>UCCACCAUUGCACUUU CGGUGG</td>
<td>ABLA01015619</td>
</tr>
<tr>
<td>Heterorh. bacterioph</td>
<td>UGGUCCACUGCAUU CGGACC</td>
<td>ACKM01001870</td>
</tr>
<tr>
<td>Brugia malayi</td>
<td>UUGGCCAUUGCACUUUUCGGUCCG</td>
<td>AAQA01000393</td>
</tr>
<tr>
<td>Loa loa</td>
<td>UUGGCAUUGCACUUCUGGUCG</td>
<td>ADBU01000479</td>
</tr>
<tr>
<td>Meloidogyne inc</td>
<td>UUCCCAUUGCACUUA AGGGCGAG</td>
<td>CABB01003416</td>
</tr>
<tr>
<td>Pristionchus pac</td>
<td>UUGGCCAUUGCACUUU CGGUGG</td>
<td>ABKE01003605</td>
</tr>
<tr>
<td>Strongyloides ratti</td>
<td>UUCACCAUUGCACUUU AGGUUGAA</td>
<td>CACX01002107</td>
</tr>
<tr>
<td>Wuchereria banc</td>
<td>UUGGCCAUUGCACUUUCAGGUGCGG</td>
<td>ADBV01000954</td>
</tr>
<tr>
<td>Ascaris lumbricoides</td>
<td>UUGGUAUUGCACUUU CGACCAG</td>
<td>L22246</td>
</tr>
<tr>
<td>+Trichinella spir</td>
<td>GUAUUAUUGCACUA CGAUGGC</td>
<td>ABIRO1000393</td>
</tr>
<tr>
<td>Schist mansoni</td>
<td>CUCACCAUUGCACUUGG</td>
<td>L25921</td>
</tr>
<tr>
<td>+Schmidtea mediterranea</td>
<td>UACUUAUUGCACUCUGAAGUG</td>
<td>AAWT01083394</td>
</tr>
<tr>
<td>Echin. Multilocularis, Taenia solium</td>
<td>CUCAGCAUUGCACUCCGCUGUG</td>
<td>M73768; GT892159</td>
</tr>
<tr>
<td>Fasciola hepatica</td>
<td>CUCGCGUGUAUUGCACUGGUGG</td>
<td>GT740508</td>
</tr>
<tr>
<td>Neobenedenia melleni</td>
<td>CUCUCCCAUUGCACACCGGAAGG</td>
<td>GW919958</td>
</tr>
<tr>
<td>Alvinella pompejana</td>
<td>CUUUCAAUGCAUCUCUGGUGG</td>
<td>FP530209</td>
</tr>
<tr>
<td>Hirudo medicinalis</td>
<td>CUCUCAUUGCACUUGGUGG</td>
<td>FP606058.</td>
</tr>
<tr>
<td>Helobdella robusta</td>
<td>CAGCAUUGCACUACCGGUGG</td>
<td>EY389719</td>
</tr>
<tr>
<td>Capitella</td>
<td>CUUUCUAUUGCACUAGGGGG</td>
<td>EY534601</td>
</tr>
<tr>
<td>Arbacia punctulata</td>
<td>CUGUGCAUUGCACACCGGCUCG</td>
<td>X75936</td>
</tr>
<tr>
<td>Asterina pectinifera, Patiria miniata</td>
<td>CUAACCAUUGCACUCCGGUGG</td>
<td>DB422888; HP123087</td>
</tr>
<tr>
<td>Lytechinus variegatus, Strongylocentrotus purpuratus</td>
<td>CUGGCAUUGCACACCCGGCUUG</td>
<td>X04994, X56629</td>
</tr>
</tbody>
</table>
Nematostella vectensis, Porites lobata
Hydra magnapapillata, Acropora millepora
Amphimedon queens.
+Monosiga brevicollis
Trichoplax adhaerens

U2 snRNA SLIV

Human, rat, mouse, frog, Dicentrarchus labrax, Tetraodon nigroviridis:
Salmo salar
Takifugu rubripes, Oryzias latipes
Zebrafish, Ictalurus punctatus
Chick, Pheasant, turkey Branchiostoma floridae
Emu
Anolis carolinensis
Python molurus

+Lamprey

Tribolium castaneum, Pediculus humanus, Bombyx mori; Heliotis virescens; Spodoptera littoralis
Rhipicephalus microplus, Ixodes scapularis, Acaryosiphon pismum
Rhodnius prolixus
Apis florea
Camponotus floridanus
Atta cephalotes
Bombus terrestris
Nasonia vitripennis
Glossina morsitans
Mayetiola destructor

In Supplementary Material (S1) of

Tribolium castaneum
Pediculus humanus
Bombyx mori; Heliotis virescens; Spodoptera littoralis
Rhipicephalus microplus
Ixodes scapularis
Acaryosiphon pismum
Rhodnius prolixus
Apis florea
Camponotus floridanus
Atta cephalotes
Bombus terrestris
Nasonia vitripennis
Glossina morsitans
Mayetiola destructor
Drosophila melanogaster; Anopheles gambiae; Culex quinquefasciatus; Aedes aegypti
Petrolisthes cinctipes, Daphnia pulex, Lepeophtheirus salmonis

C. elegans, C. briggsae, Heterodera glycines, Heterorh. bacterioph
Ciona savigny, Molgula tectiformis, Oikopleura dioica

Saccoglossus kowal. C. elegans

C. elegans

C. briggsae

Heterodera glycines

Heterorh. bacterioph

Brugia malayi

Loa loa

Meloidogyne inc

Pristionchus pac

Strongyloides ratti

Wuchereria banc

Ascaris lumbricoides

Strong. Purpuratus

Euclidaris tribuloides

Lytechinus variegatus

Patricia miniata

Elysis timida

Crassostrea gigas

Biomphalaria glabrata

Aplysia californica

+Lottia gigantea

Alvinella pompejana;
Hirudo medicinalis

Helobdella robusta

Schistosoma japonicum

Schistosoma mansoni

Taenia solium

+Schmidtea

125
<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>+Trichinella spir</td>
<td>CCUGAUUGCAUCAUACAUUCAGG</td>
<td>ABIR01000806</td>
</tr>
<tr>
<td>Nematostella vectensis</td>
<td>CUCGGUAUUGCAUACCUCGAG</td>
<td>ABAV01007607</td>
</tr>
<tr>
<td>Acropora digitifera</td>
<td>CUCGGUAUAGCAUACCUCGAG</td>
<td>BACK01000482</td>
</tr>
<tr>
<td>Hydra magnapapillata</td>
<td>CUCGGCUUUGCAUACUGCUGAG</td>
<td>ABRM010985</td>
</tr>
<tr>
<td>Amphimedon queensl.</td>
<td>CCUGGUUUGCAUACUCCAGG</td>
<td>ACUQ01000983</td>
</tr>
<tr>
<td>+Trichoplax</td>
<td>GCAGGUUUGCAUGUAGCUCUACU</td>
<td>ABGP01000395</td>
</tr>
<tr>
<td>Salpingoea</td>
<td>CCUGGUUUGCAUACUCCUAGG</td>
<td>ACSY01001265</td>
</tr>
<tr>
<td>+Monosiga brevicollis</td>
<td>UGCAUUUCGGACGCUGCAUACU</td>
<td>ABFJ01000830</td>
</tr>
</tbody>
</table>
Chapter 6.

Evolution of Ancestral URB Protein Toward Modern Vertebrate Traits

Sandra G. Williams, Kimberly J. Delaney, Mariah Lawler, Kathleen B. Hall

“Where you come from is gone, where you thought you were going to never was there, and where you are is no good unless you can get away from it.” -Flannery O’Connor, *Wise Blood*
**Introduction**

In the preceding chapter, we reconstructed the phylogeny and ancestral sequences of metazoan U1A/U2B″/SNF family proteins. This showed that U1A and U2B″ proteins emerged following a gene duplication in ancestors of jawed vertebrates (Figure 1). While RNA binding properties were subfunctionalized in humans, there remained questions about whether subfunctionalization occurred gradually or rapidly and what mutations were responsible for the changes in behavior. Additionally, we characterize the evolution of protein exchange properties and dynamics. Our results show that modern RNA binding specificity of U1A and U2B″ RRM1 for RNAs emerged very quickly following the gene duplication in an ancestor of jawed vertebrates. Five sequential residue changes in Loop 3 confer modern U1A RNA binding specificity for SLII vs. SLIV. U2B″ binding specificity is more complicated. While mutations in β2/Loop 3 drastically reduce the binding affinity of Urb for both SLII and SLIV, the reverse mutations in the U2B″ ancestral protein have no effect on binding. Further, the β2/Loop 3 mutations do not affect the RNA binding specificity, suggesting that additional mutations in other parts of the RRM are important for the domain’s final RNA binding properties.

**Results**

The reconstruction of the deuterostome branch of U1A/U2B″/SNF phylogeny is shown in Figure 1A. This is a subtree of the larger U1A/U2B″/SNF phylogeny that was described in the preceding chapter. Proteins at each significant node are designated as follows: Urb-D is the last common ancestral protein for modern deuterostomes; Urb-V for jawed vertebrate proteins (preceding the gene duplication); Urb-Va for the last common ancestral protein of modern U1A
proteins; and Urb-Vb for U2B” proteins. Figure 1B shows a sequence alignment of the first RRM of these proteins, as well as alignments of the corresponding sequences of human U1A and U2B”. Figure 1C shows the location of the evolving residues plotted on the RRM. Urb-Va has an ambiguously reconstructed amino acid in Loop 3 and β2, but many of the ambiguities are at the N- and C-termini of the RRMs (see Supplemental Table 1).

**Figure 1.** The deuterostome branch of the U1A/U2B”/SNF phylogenetic tree. (A) The deuterostome subset from the original phylogenetic tree. Red circles indicate nodes that were resurrected for biochemical experiments. (B) Alignments of the maximum-likelihood sequences of each resurrected protein as well as modern U1A and U2B”. Amino acids in gray indicate variation from Urb-V prior to the gene duplication. Amino acids in red indicate variation from Urb-V in the Urb-Va lineage. Amino acids in cyan indicate variation from Urb-V in the Urb-Vb lineage. (C) Amino acid divergence is plotted on the RRM structure. Colored amino acids indicate change from immediate predecessor. (D) Sequences of modern human stem–loop II and stem–loop IV.
RRM1 evolved very little between Urb, Urb-D, and Urb-V; the amino acid sequences of the three proteins are very similar. Following the gene duplication, the Urb-Va/U1A branch accumulated several changes within the N-terminal tail, one mutation in the β4/α3 junction at a relatively nonconserved position, and five simultaneous changes within β2/Loop 3. These five changes in β2/Loop 3 are preserved from reconstructed Urb-Va to modern U1A. Only two additional mutations (one in Loop 4 and one in β4) distinguish Urb-Va from U1A RRM1. Many more mutations were introduced in the Urb-Vb/U2B″ branch between the Urb-V and Urb-Vb nodes than in the U1A branch. These include mutations throughout the RRM (Fig. 1C). On the U2B″ branch, seven additional mutations throughout the domain were accumulated between Urb-Vb and human U2B″ RRM1.

From our deuterostome phylogeny, we selected RRM1s at critical nodes (Urb, Urb-D, Urb-V, Urb-Va, and Urb-Vb) for further functional characterization. We used the modern human U1 SLII and U2 SLIV RNA hairpins to assess the RNA-binding properties of these proteins, as these RNAs have evolved very little and are considered reasonable surrogates for their ancestral counterparts (Fig. 2). Results of the binding experiments are shown in Table 1. Results for human U1A RRM1 and full-length human U2B″ are shown for comparison (these values were previously published (1, 2)). Urb, Urb-D, and Urb-V show almost identical binding to both SLII and SLIV. In contrast, Urb-Va has gained affinity for SLII and lost affinity for SLIV. Urb-Vb has lost affinity for both SLII and SLIV and does not appear to discriminate between the two RNAs. Comparing Urb-Va with U1A RRM1 and Urb-Vb with (FL) U2B″, we find that the ancestral RRM1s and their modern orthologs have very similar RNA-binding specificities for their in vivo RNA targets.
Figure 2. RNA sequence conservation. (A) Vertebrate SLII (left) and SLIV (right) sequence logos-loop and loop-closing base pair. (B) Metazoan SLII (left) and SLIV (right) sequence logos-loop and loop-closing base pair, as reported in (3). (C) Cocrystal of U1A RRM1:SLII (1URN) and (D) U2B″ RRM1 and SLIV from the cocrystal of the ternary complex (1A9N). LVSRS amino acids are shown in U1A, and VALKT in U2B″. Graphics using VMD.

Table 1.

<table>
<thead>
<tr>
<th>Protein</th>
<th>SLII (M)</th>
<th>SLIV (M)</th>
<th>ΔΔG° (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URB</td>
<td>1.2 ± 0.2 × 10^{-9}</td>
<td>1.5 ± 0.2 × 10^{-8}</td>
<td>-1.5 ± 0.1</td>
</tr>
<tr>
<td>Urb-D</td>
<td>1.5 ± 0.2 × 10^{-9}</td>
<td>2.6 ± 0.4 × 10^{-8}</td>
<td>-1.7 ± 0.1</td>
</tr>
<tr>
<td>Urb-V</td>
<td>1.2 ± 0.1 × 10^{-9}</td>
<td>2.5 ± 0.4 × 10^{-8}</td>
<td>-1.8 ± 0.1</td>
</tr>
<tr>
<td>Urb-Va</td>
<td>4.4 ± 0.2 × 10^{-10}</td>
<td>2.75 ± 0.04 × 10^{-7}</td>
<td>-3.8 ± 0.03</td>
</tr>
<tr>
<td>h. U1A RRM1</td>
<td>4 ± 3 × 10^{-10}</td>
<td>&gt; 1 × 10^{-6}</td>
<td>&lt; -4.6</td>
</tr>
<tr>
<td>Urb-Vb</td>
<td>5.0 ± 0.2 × 10^{-7}</td>
<td>3.8 ± 1.8 × 10^{-7}</td>
<td>0.15 ± 0.28</td>
</tr>
<tr>
<td>h. FL U2B″</td>
<td>3.3 ± 1.0 × 10^{-8}</td>
<td>2.6 ± 0.6 × 10^{-8}</td>
<td>0.14 ± 0.22</td>
</tr>
</tbody>
</table>

Apparent dissociation constants for proteins binding to RNAs. Binding constants were measured by nitrocellulose filter binding, and experiments were performed in 250mM KCl 10mM cacodylate, 1mM MgCl₂ pH 7 at 22°C. ΔΔG° indicates the binding specificity of the protein for SLII vs. SLIV.
Source of changes to binding affinity in the U1A branch.

Prior in vitro experiments with U1A/U2B'' chimeric RRMs showed that β2/Loop 3 (LVSRSLKMRG53 in U1A; VALKTMKMRG50 in U2B'') was important for determining RNA-binding specificity (4-6). In our reconstruction, β2/Loop 3 (VALKTMKMRG50) is unchanged in ancestors of U1A and U2B'' prior to the vertebrate (gnathostome) gene duplication. Between the Urb-V and Urb-Va nodes, β2/Loop 3 accumulates five mutations, changing VALKT to LVSRS. In modern proteins, U1A contains the LVSRS sequence, and U2B'' has VALKT; the positions of these amino acids in the structures of the respective RNA:protein complexes are shown in Figure 2. It is important to appreciate that these amino acids do not directly participate in RNA recognition, so any effect they have on RNA binding must be a consequence of changes to the RRM.

While the individual mutations VALKT to LVSRS are relatively conservative, the appearance of the mutations in this branch suggests that they confer new functionality. We have systematically altered VALKT/LVSRS sequences to evaluate the amino acid contributions to RNA binding.

Our most substantial alteration was the reversion of β2/Loop 3 in Urb-Va to the sequence of Urb-V (prior to the gene duplication) to create the Urb-Va-VALKT RRM. This protein differs from Urb-V at sites in the N terminus and in the β4/α3 junction. Urb-Va-VALKT does not discriminate between SLII and SLIV and binds both with an affinity comparable to that of Urb-V (Table 2). The β2/Loop 3 sequence in the Urb-Va protein is, therefore, sufficient to revert the binding specificity of the RRM to that of its immediate ancestor.
Table 2. Apparent dissociation constants for U1A ancestral proteins and related mutants. \( \Delta \Delta G^\circ = \Delta G^\circ \text{mutant protein} - \Delta G^\circ \text{Urb-V} \).

**RNA binding is salt-dependent**

Protein binding to nucleic acids typically has an electrostatic component to the association. The net contribution of the electrostatics to the binding free energy is unique to each complex, but the salt dependence of the interaction can be used to determine if ions are taken up or released and so provides a means to compare binding modes of the interactions. Electrostatic interactions are known to play a key role in the interaction of U1A and SLII (1, 7), and so we examined the dissociation constant of Urb proteins binding to SLII and SLIV as a function of \([K^+]\). These data are plotted as log\(K_{d, \text{app}}\) vs. log [KCl], using the formalism of (8). A positive slope indicates ion uptake; a negative slope indicates net ions released. We measured binding of each protein to SLII and SLIV over a range of KCl concentrations (Fig. 3), and as expected, the negative slope of the salt dependence curve indicates that a net number of ions are released upon binding. However, we find that each protein:RNA interaction is unique: Urb-V binding to SLII/SLIV releases 3.3/4.2 ions; Urb-Va releases 5.4/2.1; and Urb-Vb releases 2.6/5.9 ions. This
diversity of response indicates a significant difference in how each protein interacts with each target RNA.

Figure 3. Salt dependence of binding to wild-type RNA stem–loops. (A) Binding isotherm data and fits to a 1:1 complex; 250 mM KCl, 1 mM MgCl$_2$, 10 mM sodium cacodylate, pH 7. (B–D) Protein–RNA pairs are indicated in the panels. (▪) SLII, (•) SLIV. All experiments were performed in 10 mM cacodylate and 1 mM MgCl$_2$ at room temperature with indicated salt. Slopes of the lines are interpreted in terms of net ions released and are reported adjacent to each line.

RNA mutagenesis.

Stem–loops II and IV are both highly conserved throughout metazoans (Fig. 2) and are remarkably similar in size and structure (Fig. 1D). In the vertebrate lineage, and specifically in humans, there are three main differences between SLII and SLIV: (1) Loop position 7 is cytosine
in SLII but guanosine in SLIV; (2) the loop-closing base pair is a C:G in SLII and a noncanonical U:U in SLIV; and (3) the 3’ loop nucleotides (UCC) in SLII RNA are poorly conserved, but in SLIV the analogous UACC is conserved. SLII has 10 loop nucleotides, while SLIV has 11. The high degree of sequence conservation in these large RNA loops is likely to be predominantly driven by protein recognition. However, the ancestral RRMs clearly show differences in RNA-binding affinity and specificity, which can account for much of the specific protein localization seen in modern snRNPs. A series of RNA mutations were made to probe the source of differences in the RNA-binding specificities of Urb-V, Urb-Va, and Urb-Vb.

Loop size and structure. Extensive work with U1A has previously shown that the secondary structure of its RNA target is important for high-affinity binding: the recognition sequence must occur within the context of a stem–loop structure, and the optimal loop size is at least 10 nucleotides (9, 10). We used RNA variants to address secondary structure requirements and optimal loop size for Urb-V, Urb-Va, and Urb-Vb. The ssLoop RNA construct puts the loop sequence of SLII in a completely single-stranded RNA context (secondary structure predictions via mfold. As shown in Figure 4, all three proteins experience a significant loss of binding affinity in the absence of a stem–loop: binding of both Urb-Va and Urb-Vb is weaker than the sensitivity of the filter binding assay. This is unsurprising as crystal structures of U1A and U2B” RRM1 in complex with their RNA targets indicate interactions between Loop 3 of the protein and the loop-closing base pair of the stem (11, 12). It appears that the ancestral proteins share a similar requirement for the presence of a stem.
We probed the loop size requirement by deleting one (9-nt loop) or two (8-nt loop) nucleotides from the 3’ side of the loop. Previous work with U1A showed that the protein does not specifically recognize these nucleotides in the U1A:SLII complex (9, 13), and in the U1A:SLII cocrystal, the UCC bases have no protein contacts (11). Consistent with the U1A results, all three proteins show a loss of binding affinity to the 9-nt loop RNA and an even greater loss for the 8-nt loop RNA (Fig. 4). This result is supported by cocrystal structures of modern protein-RNA complexes (11, 12), which show that Loop 3 of the protein protrudes through the RNA loop; the loop size must be large enough to accommodate the insertion of the protein and position the central nucleotides near the β4/α3 junction.

![RNA structures](image)

**Figure 4.** Binding to loop size and structure mutants. Dissociation constants for binding to each RNA are shown. All binding experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl2 (pH 7) at room temperature.

**Loop position 1.** Both biochemical and structural data have shown that the adenosine in loop position 1 (A1) is specifically recognized by U1A RRM (11, 14-16). A1 is conserved in both
SLII and SLIV. To investigate the importance of A1 in the ancestral complexes, we replaced it with cytosine (A1C) or deleted it (ΔA1, with a 3’ C insertion to maintain a 10-nt loop) in SLII. All three proteins exhibit decreased affinity for these RNAs (Fig. 5), consistent with A1 conservation in SLII. However, it is clear that perturbation of this nucleobase impedes Urb-Va binding significantly more than it disturbs Urb-Vb binding.

Figure 5. Binding to nucleotide A1 mutants. Dissociation constants for binding to each RNA are shown. ∆∆G° = ∆G° (Mutant RNA) − ∆G° (WT SLII). All binding experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl₂ (pH 7) at room temperature.

Urb-V binding affinities for A1C and ΔA1 SLII indicated some dependence of this loop position (Fig. 5). Binding free energies of both SLII mutant RNAs to Urb-Va were significantly perturbed (∆∆G° = (ΔG°MUT − ΔG°WT) = +2.3–2.4 kcal/mol), but Urb-Vb binding was nearly unchanged (∆∆G° = +0.3–0.6 kcal/mol), indicating that this protein is not very sensitive to this position. In fact, deletion of A1 resulted in less disruption of Urb-Vb:SLII binding than a mutation, suggesting it has no need for this nucleobase. This constitutes a major difference in the
RNA recognition mechanisms of Urb-Va and Urb-Vb. While these results provide a rationale for the phylogenetic conservation of A1 in SLII, they do not explain the conservation of A1 in vertebrate U2 SLIV snRNAs.

**SLIV 3′ side.** Previous mutational analysis and structural work with U1A showed that there is no sequence recognition of the 3′ side of SLII (10, 11, 13, 17). However, the importance of this region of the SLIV hairpin has not been studied. Interestingly, phylogenetic analysis of the RNA sequences indicates that the 3′ side of SLIV shows more sequence conservation than that of SLII and that, in particular, position U8 [AUUGCAGU8ACC] is universally conserved among metazoans (Fig. 2). This level of conservation suggests a strong evolutionary pressure against mutation, which could be consistent with protein interaction with the 3′ side of the loop and the U8 nucleotide in particular. However, Urb-V and Urb-Vb exhibit very little change in binding affinity to the SLIV U8C mutant (Fig. 6). Surprisingly, Urb-Va exhibited the most dramatic response to this mutation with a significant loss of binding affinity (at least one order of magnitude). Previous work has shown that the 3′ UCC does not contact the protein in the U1A:SLII complex; the nucleotides were replaced with a polyethylene glycol linker with no loss of binding affinity (17), so the response of Urb-Va to the U8C substitution is rather mysterious. The SLIV A9G mutation also resulted in drastically reduced binding affinity by Urb-Va but insignificant changes in Urb-V and Urb-Vb affinity. The unexpectedly strong dependence of Urb-Va for this side of the loop reveals a new aspect of the different binding mechanisms of Urb-Va:SLIV and Urb-Va:SLII.
Figure 6. Binding to 3′ loop mutants. Dissociation constants for binding to each RNA are shown. All binding experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl₂ (pH 7) at room temperature.

SLII and SLIV differences.

To further probe the different RNA-binding mechanisms of the proteins and their specific nucleobase recognition, we constructed three RNA mutants: SLII C7G, SLIV G7C, and SLIV LCB in which the U-U loop-closing base pair (LCB) was replaced with the C-G that is found in SLII. In metazoans, C7 appears to be universally conserved in SLII sequences (Fig. 2). In metazoan SLIV sequences, the 7 loop position is considerably more variable; it is most commonly G or C, but U can also be found. However, in vertebrate RNA sequences, it appears to be universally conserved as a G (Fig. 2).

As expected, the SLII C7G mutation resulted in decreased affinity of Urb-Va, consistent with previous U1A binding data (1). Urb-V and Urb-Vb binding to SLII C7G was not perturbed, consistent with a lack of specific recognition of this nucleotide (Fig. 7). Conversely, the SLIV G7C mutation resulted in a significant increase in affinity of Urb-Va for SLIV; again, there was
little or no change in Urb-Vb or Urb-V affinity for these RNAs. These data indicate that prior to the gene duplication, Urb-V did not specifically recognize the nucleobase at position 7 in either SLII or SLIV. Urb-Vb retained this lack of discrimination, but this site is recognized specifically by Urb-Va, and the C at this position is important for high-affinity binding. The free energy associated with this recognition largely accounts for the increased affinity of Urb-Va for SLII, compared with Urb-V. Nucleotide C7 interacts with the peptide backbone in the β4/α3 junction (11), suggesting that the mode of recognition of this base in Urb-Va/U1A will depend on orientation and dynamics of this region of the protein.

The SLIV LCB UU to CG mutation resulted in an increase in affinity for Urb-V and Urb-Va but resulted in no change to the binding affinity of Urb-Vb (Fig. 7). This mutation results in (almost) identical increases to Urb-V and Urb-Va binding affinity: ΔΔG°binding (Mutant − WT) = −1.5/−1.2 kcal/mol, respectively (Table 3). C7 and the loop-closing base pair account for much of the difference in RNA discrimination between Urb-Va and Urb-Vb.

Because SLIV G7C and SLIV LCB mutations account for two of the three differences between SLII and SLIV, we summed the binding free energy differences of these mutations to compare their contributions to protein specificity for SLII over SLIV (Table 3). A simple sum of the ΔΔG° for Urb-V binding to both SLIV mutations results in a −1.8 kcal/mol preference for the mutants over SLIV, indicating that position 7 and the LCB could entirely account for Urb-V’s preference for SLII over SLIV (assuming no cooperativity or contribution from other factors). In 250 mM KCl, Urb-Vb shows no significant difference in binding affinity for WT SLII vs. SLIV, SLIV G7C, or SLIV LCB, indicating that loop position 7 and the LCB do not contribute to Urb-
Vb recognition of the stem–loops. However, Urb-Va presents a different pattern. While Urb-Va has $\Delta \Delta G^{\circ} \text{binding} \ (\text{SLII} - \text{SLIV}) = -3.8 \text{ kcal/mol}$ preference for SLII over SLIV, the additive contributions of the LCB mutation and the mutation at position 7 result in a more modest $-2.4 \text{ kcal/mol}$ preference for the mutants over SLIV. Clearly, position 7 and the LCB are not sufficient to account for Urb-Va's specificity for SLII over SLIV, assuming a model of site independence. Like the modern U1A RRM, the binding mechanism of Urb-Va appears to be quite complex.

![Diagram of RNA structures](image)

**Figure 7.** Binding to SLII and SLIV conversion mutants. Dissociation constants for binding to each RNA are shown. All binding experiments were performed in 250 mM KCl, 10 mM cacodylate, and 1 mM MgCl$_2$ (pH 7) at room temperature.

**Table 3.**

<table>
<thead>
<tr>
<th></th>
<th>Urb-V</th>
<th>Urb-Va</th>
<th>Urb-Vb</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta \Delta G^{\circ} \text{ (SLII-SLIV)}$</td>
<td>$1.2 \pm 0.1 \times 10^{-9}$</td>
<td>$2.5 \pm 0.1 \times 10^{-8}$</td>
<td></td>
</tr>
<tr>
<td>$\Delta \Delta G^{\circ} \text{ (G7C-SLIV)}$</td>
<td>$1.0 \pm 0.1 \times 10^{-9}$</td>
<td>$1.4 \pm 0.1 \times 10^{-8}$</td>
<td></td>
</tr>
<tr>
<td>$\Delta \Delta G^{\circ} \text{ (LCB-SLIV)}$</td>
<td>$5.3 \pm 0.5 \times 10^{-10}$</td>
<td>$-0.4 \pm 0.1$</td>
<td>$1.6 \pm 0.1 \times 10^{-7}$</td>
</tr>
<tr>
<td>Sum of G7C + LCB</td>
<td>$1.7 \pm 0.1 \times 10^{-9}$</td>
<td>$0.3 \pm 0.1$</td>
<td>$1.5 \pm 0.2 \times 10^{-7}$</td>
</tr>
</tbody>
</table>

Difference in free energies of binding of each protein to the indicated RNAs.
Changes to binding affinity. U2B” branch.

The corresponding evolution in the U2B” branch is more complex than that of U1A, and we present below a preliminary analysis of some of the important changes to the protein. The results suggest that mutations throughout the body of the RRM (and not just those restricted to Loop 3) have consequences for RNA binding.

In contrast to the effect of Loop 3 mutational changes in Urb-Va, reverting the β2/Loop 3 mutations to the amino acids of their immediate ancestor (Urb-Vb V38L L46M, Table 4) has little change on RNA binding affinity; increased affinity for both RNAs (and particularly for SLII) was not observed. However, L38V/L46M mutations in the protein prior to the gene duplication (Urb-V L38V/L46M) do have profound consequences for the binding of these RNAs (Table 4). These mutations are not sufficient to establish the RNA binding properties of Urb-Vb, suggesting that other mutations along the Urb-V to Urb-Vb branch are also important for determining the protein’s RNA binding specificity. However, L38V/L46M mutations into Urb-V do drastically reduce the binding affinities for both RNAs and markedly change the binding properties of the protein. This occurs with practically no change to the binding specificity of the protein for SLII vs SLIV. \( \Delta \Delta G^\circ_{\text{SLII-SLIV}} = -1.5 \pm 0.3 \text{ kcal/mol} \), identical to the specificities of predecessor proteins (Urb, Urb-D, and Urb-V; Table 1). V38L/M46L mutations to Urb-Vb have a very small impact on the binding free energy for both SLII and SLIV (> -0.5kcal/mol). Like the opposite mutations in Urb-V, the mutations do not change the binding specificity \( \Delta \Delta G^\circ_{\text{SLII-SLIV}} = 0.2 \pm 0.3 \text{ kcal/mol} \), similar to that of Urb-Vb; Table 1). The overall effect on RNA binding affinity of residues at sites 38 and 46 is highly dependent on the protein background. 

L38V/
L46M mutations in Urb-V result in a protein with weaker affinity for SLIV than the final Urb-Vb protein, suggesting that mutations outside the β2-Loop 3 region are important for changing the protein’s RNA binding specificity (and increasing the protein’s affinity for SLIV).

<table>
<thead>
<tr>
<th>Protein</th>
<th>$K_{d,\text{app}}$ (M)</th>
<th>$\Delta \Delta G^\circ$ (kcal/mol)</th>
<th>Protein</th>
<th>$K_{d,\text{app}}$ (M)</th>
<th>$\Delta \Delta G^\circ$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urb-V</td>
<td>$1.2 \pm 0.1 \times 10^{-9}$</td>
<td></td>
<td>Urb-V</td>
<td>$2.5 \pm 0.4 \times 10^{-8}$</td>
<td></td>
</tr>
<tr>
<td>L38V/L46M</td>
<td>$2.3 \pm 0.5 \times 10^{-7}$</td>
<td>$3.1 \pm 0.1$</td>
<td>Urb-Vb</td>
<td>$5.1 \pm 0.2 \times 10^{-7}$</td>
<td>$3.6 \pm 0.1$</td>
</tr>
<tr>
<td>Urb-Vb</td>
<td>$5.1 \pm 0.2 \times 10^{-7}$</td>
<td>$3.8 \pm 1 \times 10^{-7}$</td>
<td>Urb-Vb</td>
<td>$3.8 \pm 1 \times 10^{-7}$</td>
<td>$1.6 \pm 0.3$</td>
</tr>
<tr>
<td>V38L/M46L</td>
<td>$2.6 \pm 0.3 \times 10^{-7}$</td>
<td>$-0.4 \pm 0.1$</td>
<td>Urb-Vb</td>
<td>$1.7 \pm 0.8 \times 10^{-7}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V38L/M46L</td>
<td></td>
<td>$-0.5 \pm 0.4$</td>
</tr>
</tbody>
</table>

Table 4. Apparent dissociation constants for binding of U2B” ancestral proteins and related mutant proteins to SLII and SLIV. $\Delta \Delta G^\circ$ does not refer to the difference in affinity for the two RNAs but rather the difference in binding free energy for the given protein as compared with the reference (starting) protein (Urb-V is the reference for the top half of the table and Urb-Vb is the reference for the bottom half of the table).

Another way to understand the effects of Loop 3 and the protein background is to use pairwise coupling theory (18) to assess how the 38/46 sites interact with the rest of the protein. To analyze the proteins, we consider the j site to be residues 38/46; the perturbation is from an LL state to a VM state. The i site is considered to be the rest of the protein, which is perturbed from the Urb-V state to the Urb-Vb state (abbreviated in Figure 8 as V and B; the perturbation is characterized by differences at 14 amino acids. Analyzing the binding data shows that the two sets of mutations have a negative coupling free energy of ~ -2.5kcal/mol, which is quite strong.
Evolution of protein dynamics and exchange.

One of the features that distinguishes URB and U1A RRM1 is the difference in their backbone exchange properties. U1A experiences millisecond-microsecond exchange, but this is restricted to α1 and Loop 5. In contrast, millisecond-microsecond exchange is broadly present in URB, although the effects are largest in Loop 1 and Loop 5 (see Chapter 4). To assess the exchange properties of ancestral RRMs and how they changed following the vertebrate gene duplication, we performed similar backbone experiments to those performed on URB, SNF, and U1A RRM1 in Chapter 4. Figure 9 shows results of such experiments for Urb-Va VALKT, which binds SLII and SLIV with similar affinities to those seen in URB RRM1 (Table 2). An overlay of the HSQC of this protein with URB RRM1 is shown in Figure 9a, with the URB protein shown in deeper purple and Urb-Va VALKT shown in lilac. The chemical shift differences between corresponding crosspeaks in the two spectra are plotted in Figure 9b.
Residues that are different between the two spectra are indicated with an asterisk. The spectra overlay very well, with very small differences in the backbone chemical shifts across the body of the protein. Modest chemical shift differences are localized to regions around the mutated residues, shown in the inset of Figure 9b.

**Figure 9.** (a) $^{15}$N-$^1$H HSQC overlay of URB and Urb-Va VALKT proteins. (b) Chemical shift differences of backbone peaks between the two proteins. The residues that are different between the two proteins are identified by asterisks (*). Residues with larger chemical shift differences are plotted on the RRM in pink (>1 standard deviation of the average chemical shift difference) and red (>3 standard deviations of the average chemical shift difference). (c) Heteronuclear NOEs for backbone amides. (d) Amide resonances in fast exchange (ms-us) assessed with $\Delta R_2$, effective measurements from CPMG experiments.

Backbone picosecond-nanosecond motions were assessed with heteronuclear NOE experiments (Figure 9c) which show similar patterns to those previously observed for URB; the
backbone is fairly rigid, with the exception of the far N-terminus and the C-terminus (including Loop 6 and α3). Millisecond-microsecond exchange was probed with CPMG experiments, and the ΔR₂ profiles are similar to those previously seen with URB. The domain shows large ΔR₂ values in Loop 5 and α1, with additional sites experiencing exchange throughout the body of the protein.

VALKTΔLVSRS mutations in Loop 3 of Urb Va-VALKT yield the ancestor of vertebrate U1A proteins (Urb-Va) and are sufficient to increase (very slightly) the RRM’s affinity for SLII while substantially decreasing its affinity for SLIV. Overlays of ¹⁵N/¹H HSQC spectra of Urb-Va and Urb-Va VALKT are shown in Figure 10a, and the chemical shift differences are plotted in Figure 10b. While the spectra still overlay relatively well and the chemical shift differences are small, modest chemical shift differences are observed distant from the mutation site at Loop 3. Fast motions in Urb-Va are similar to those observed in both URB and U1A RRM1. The mutations in Loop 3 have a considerable effect on millisecond-microsecond exchange throughout the domain. Exchange is now restricted to Loop 5 and α1, and even at these sites, the ΔR₂ values are much smaller than in the URB proteins. These exchange properties are identical to those observed in U1A.
Figure 10. (a) $^{15}$N-1H HSQC overlay of Urb V (red) and Urb-Va VALKT (purple) proteins. (b) Chemical shift differences of backbone peaks between the two proteins. The residues that are different between the two proteins are identified by asterisks (*). Below the chemical shift difference plot are RRMs showing residues with larger chemical shift differences (>1 SD of the Urb-Va VALKT average in pink and >2SD in orange). (c) Heteronuclear NOEs for backbone amides. (d) Amide resonances in fast exchange (ms-µs) assessed with $\Delta R_2$, effective measurements from CPMG experiments.

Overlays of HSQCs of URB (purple) and URB-Vb (blue) are shown in Figure 11a. The backbone is generally rigid, but as with other RRMs, it is flexible at its far N-terminus, in Loop 3, and C-terminal to Loop 6 (Figure 11b). The exchange properties of the backbone, however, are markedly changed from the ancestral state and more globally resemble the exchange
properties of URB-Va: residues in α1 and Loop 5 continue to exchange on the millisecond-microsecond timescale, but the \( \Delta R_2 \)s are smaller than those seen in the proteins with URB-like binding. Outside of these regions, much of the exchange on this timescale is lost. The exception (as compared with Urb Va) is β1, the location of the RNP2 motif, which continues to experience exchange.

![Image showing HSQC overlay, NOEs, and \( \Delta R_2 \) measurements](image)

**Figure 11.** (a) \(^{15}\)N-\(^1\)H HSQC overlay of Urb (purple) and Urb-Vb (blue) proteins. (b) Heteronuclear NOEs for backbone amides. (c) Amide resonances in fast exchange (ms-μs) assessed with \( \Delta R_2 \), effective measurements from CPMG experiments.

**Discussion and future directions.**

The conservation of the RRM protein sequence and RNA-binding activity among Urb, Urb-D and Urb-V represents an estimated 150 million years of evolution in which this family remained remarkably stable. However, a gene duplication in an ancestor of jawed vertebrates
resulted in a short period of RRM sequence and functional divergence in both protein paralogs. Urb-Va and Urb-Vb are reconstructions of the last common ancestral RRM1 of modern vertebrate U1A and U2B" proteins, respectively. Given the similarities in RNA-binding specificity between Urb-Va and U1A and between Urb-Vb and U2B", it is likely that the mutations to the proteins (following the gene duplication) rapidly resulted in subfunctionalization of protein binding and localization.

In the Urb-Va/U1A lineage, we are able to identify five amino acids in β2/Loop 3 that are responsible for the protein's specificity. The transition from VALKT (found in Urb, Urb-D, Urb-V, and Urb-Vb) to LVSRS (in Urb-Va), though conservative at each amino acid position, is sufficient to both increase affinity for SLII and decrease affinity for SLIV compared to its predecessor (Urb-V). This is consistent with previous structural and mutagenesis data that show U1A Loop 3 interacting with the 5′ side of the RNA loop and the loop-closing base pair of SLII (4, 11, 19). Some of the single amino acid mutations in the loop sequence result in binding effects that are comparable to that of the full loop substitution. The single mutation of VALKT to LALKT is one of the mutations that has the most dramatic effect on RNA binding. In U1A, the comparable reciprocal substitution, L44V (U1A numbering), confers opposite changes to binding affinity and specificity (20). Our phylogenetic analysis and protein sequence reconstruction cannot determine which of these mutations came first.

In contrast to the Urb-Va branch, where functional changes are localized to a short stretch of residues in β2/Loop 3, the evolution of Urb-Vb RRM1 after the gene duplication was much more complicated. Fourteen amino acid changes accumulated between Urb-V and Urb-Vb.
Possibly, this represents the results of strong negative selective pressure on one copy of the gene to maintain functionality, while the second copy was allowed drift until subfunctionalization of RNA binding was established. The first copy eventually became U1A, and U2B” binding properties emerged in the second copy. The fourteen amino acid mutations between Urb-V and Urb-Vb are not localized to any single region of the RRM, and our preliminary results suggest that mutations throughout the RRM (and not localized to β2/Loop 3) are important for changing the RNA-binding properties of the molecule.

Urb-Vb, the ancestor of modern vertebrate U2B” proteins, binds with modest affinity to both SLII and SLIV but does not substantially discriminate between the two RNAs. This is similar to what is observed for human U2B”. However, the protein binds with ~10-fold weaker affinity to both RNAs than modern U2B”. This potentially results from the absence of the interdomain linker, which may nonspecifically improve the RNA binding affinity of full-length U2B” compared with RRM1.

Compared with its immediate ancestor prior to the gene duplication, URB-Vb has lost affinity for both SLII and SLIV, although the change in binding free energy is greater for SLII. Our results show that mutations in β2/Loop 3 were important to decrease the overall binding affinity of the RRM for SLII and SLIV. However, these mutations did not change the specificity of the protein for the two RNAs. Further, mutating these residues in Urb-Vb to the ancestral amino acids does not impact RNA binding affinity, showing that the effect of one or both of these residues to RNA binding affinity is epistatic. These results leave open two important questions: What were the mutations responsible for changing the specificity of URB? How is the effect of particular amino acids in β2/Loop 3 dependent on other protein mutations? Two
mutations accumulated in Loop 1 between Urb-V and Urb-Vb. This loop packs against Loop 3 in the free protein and partially packs against the RNA in structures of modern homologs complexed with RNA. The mutations are therefore interesting candidates for both affecting RNA binding specificity and the observed coupling between the background protein and β2/Loop 3.

Our structural and biochemical mutational studies of U1A and Drosophila SNF RRMs have identified a network of hydrogen bonding interactions between amino acid side chains that are important for RNA binding (21-23). This network is quite extensive, encompassing nearly the entire RRM surface, but it is different in Drosophila SNF and human U1A. We propose that the network of interactions in each RRM contributes substantially to their different RNA-binding properties. The similarity in slow and fast timescale conformational dynamics between SNF and Urb RRM1 (3) makes it likely that preformed networks similar to SNF were present in Urb-family proteins and were conserved prior to the vertebrate (gnathostome) gene duplication to allow these proteins to bind both SLII and SLIV with high affinities, while still preserving some discrimination between RNAs. However, the network was likely altered in both branches following the gene duplication. In U1A and SNF, engineered mutations of residues on the β-sheet have substantial effects on both the surface hydrogen bonding network and RNA binding (21-23). Given the many changes throughout the body of the RRM in the Urb-Vb branch, it is likely that the hydrogen bonding network has been substantially altered.

**U2A’ interactions**

In addition to binding RNA, U2B″ and SNF also bind the U2A’ protein in the U2 snRNP. U2A’ is conserved in eukaryotes, indicating that it is an ancient protein (24) that is found only in
the U2 snRNP. When analyzing the effects of evolutionary changes between Urb, Urb-Va, and Urb-Vb, interactions with the protein binding partner U2A′ may also have altered.

In the absence of RNA, U2A′ binds tightly to human U2B″ but binds with much weaker affinity to both U1A and Drosophila SNF (data not shown). It is tempting to speculate that prior to the vertebrate gene duplication, Urb proteins bound ancestral U2A′ proteins with weak affinity and that, at some point in the evolution of the Urb-Vb/U2B″ branch, the RRM adapted into a high-affinity binder for U2A′. Whether this is true and whether the subsequent high-affinity U2B″-U2A′ interaction was the result of protein-protein coevolution or adaptations of a single protein remain to be determined. Recent work on the evolution of yeast transcription factors has suggested that in macromolecular assemblies, one of the consequences of gene duplication followed by protein subfunctionalization is a dominant-negative effect between the paralogous proteins with respect to other components of the assembly. Because this can be functionally deleterious, in order for the duplicated proteins to persist, there is a strong evolutionary pressure to minimize paralog interference, thus accounting for differences in binding to other members of the macromolecular complex (25). The difference in the binding affinities of U1A and U2B″ for U2A′ is suggestive of adaptations to minimize paralog interference, but conclusive evidence for such an evolutionary pressure would improve our understanding of the functional role of these proteins in the snRNPs.

**RNA recognition**

The high level of conservation of SLII and SLIV in all family lineages implies significant pressure against change. RNA mutagenesis allowed us to examine the binding requirements of
each of the proteins surrounding the gene duplication to see how these pressures may have changed with protein mutations.

The adenosine residue in loop position 1 (A1) is found in both SLII and SLIV. Mutations to A1 resulted in weakened binding affinity by Urb-V and Urb-Va but, surprisingly, not Urb-Vb. Urb-Va's requirement for A1 provides the evolutionary pressure to maintain this nucleotide in SLII. Prior to the gene duplication, SLIV was bound by Urb-V, which also specifically recognized A1. However, following the duplication, all vertebrate SLIV loops retain A1 despite a seeming lack of recognition by Urb-Vb. It is possible that in the ternary SLIV:RRM:U2A’ complexes of vertebrates, the RRM interacts directly with A1, particularly if Loop 3 undergoes a conformational change upon U2A’ binding. However, if this is not the case, it is possible that this nucleotide may eventually be mutated or eliminated in vertebrate SLIV RNAs.

*Caenorhabditis elegans* evolution has resulted in a different solution to U1A/U2B”/SLII/SLIV recognition that is pertinent to vertebrate SLIV A1. In worms, the U1A and U2B” proteins are redundant, and the worm is viable upon loss of either one (but not both) (26). The snRNAs in *C. elegans* differ from those in vertebrates: in particular, SLIV has lost nucleotide A1 (loop sequence: UUGCACUGC), although SLII retains it. In the absence of biochemical data describing specific RRM:RNA interactions, in vivo data show that in *C. elegans*, U1A and U2B” proteins are able to bind both RNAs with sufficient affinity to be retained in both snRNPs.

The 3’ side of SLIV has four nucleotides (AUUGCAGU8A9C10C11) that are fairly well-conserved in vertebrates (U8 is universally conserved, not only in vertebrates but in all metazoans) (Fig. 2). Urb-V and Urb-Vb, two proteins that likely bound SLIV in vivo, do not
discriminate among U8C and A9G mutations, which indicates that the nucleobases are not recognized specifically by the proteins in the bimolecular complex. Unexpectedly, we find that Urb-Va exhibits significantly weaker binding affinity for U8C and A9G 3' loop mutants of SLIV RNAs. While it is important to remember that nucleotide substitutions within the loop could potentially alter its secondary structure, the unique response of Urb-Va binding shows that it is a protein-specific phenomenon. The result is unexpected since, in SLII, this side of the RNA loop does not make contact with the modern U1A protein. It is possible that the interaction with the 3' side of the loop was subsequently lost as the protein continued to evolve, or that this interaction is only seen in SLIV, an RNA that is not bound by this protein in the cell.

In addition to nucleobases that are specifically recognized by the RRM, there are examples of nucleobases that act as “negative discriminators” (27) to prevent binding by an RRM. The loop G7 in SLIV is one example; while Urb-V and Urb-Vb are insensitive to the specific nucleobase at this position (and it is variable across metazoan SLIV sequences), Urb-Va binds with significantly weaker affinity when it is present. In vertebrate SLIV, this is universally conserved as a G, consistent with evolution of SLIVs to negatively discriminate against binding by Va proteins. The loop-closing base pair serves as a negative discriminator in SLIV and a positive discriminator in SLII RNAs (for Urb-Va/U1A). In vertebrate RNAs, the LCB nucleotides are universally conserved as CG in SLII and UU in SLIV. While both Urb-V and Urb-Va have similar differences in binding free energy for hairpins with CG vs. UU loop-closing base pairs, Urb-V does not discriminate between CG and UG loop-closing base pairs (high affinity for both), while Urb-Va does (data not shown). Loop-closing base pairs in SLIV RNAs
across metazoans are substantially more variable than in vertebrates, reinforcing the distinctiveness of LCB sequence conservation in vertebrates.

Although our experiments are designed to test the affinity and specificity of the RRM during their evolution, they also report on the evolution of the RNA. Notably, while the RRM in both Urb-Va and Urb-Vb lineages have acquired multiple mutations, we have observed that the RNAs have evolved minimally in vertebrates (Fig. 2). Evolutionary pressures are exerted by both the RRM and the RNA stem–loops to maintain their functional relationships, and we have identified several nucleotides that Urb-V and Urb-Va use for discrimination of SLII from SLIV even as the protein sequences vary. From the perspective of the snRNA, it would appear that SLII and SLIV RNA sequences determine what mutations of Urb-Va and Urb-Vb are evolutionary winners.

**Subfunctionalization: different binding modes**

Prior to the gnathostome (vertebrate) duplication of the Urb gene, a single Urb-family protein localized to both the U1 and U2 snRNPs by binding both U1 SLII and U2 SLIV. This protein bound SLII RNAs with very high affinity. It also bound to SLIV RNAs, although binding affinity for SLIV was somewhat weaker. However, after the gene duplication, Urb-Va and Urb-Vb mutated to adopt specialized binding mechanisms suited to their in vivo targets.

Our data show that Urb-Va gained marginal affinity for SLII while losing substantial affinity for SLIV following the gene duplication by adopting a mode of binding that is almost certainly a modification of the RNA recognition employed by Urb-V, effects that were entirely
mediated by mutations to the protein β2/Loop 3. Recognition of SLII by U1A is complex (6, 7, 9-11, 19) and includes interactions between the β4/α3 junction and the top of the RNA loop (C5A6C7), interactions between the RNP motifs and the 5' side of the loop, and interactions between Loop 3 and the CG loop-closing base pair and the adjacent A1. In U1A, discrimination of both the loop-closing base pair and C7 are implicated in the difference in binding affinity for SLII and SLIV. In comparison, Urb-V specificity for SLII over SLIV is almost entirely mediated by differences in recognition of the LCB.

A valine to leucine mutation of amino acid 41, which is at the junction of β2 and Loop 3 (V41ALKT), is sufficient to confer both increased affinity for SLII and decreased affinity for SLIV. This one change is important for the protein to distinguish between the two stem–loops, presumably by altering the recognition of RNA C7. The adjacent A42V substitution has a very similar effect. These two amino acids are nominally located in β2; the V/L41 side chain is solvent-exposed, while the A/V42 side chain is directed toward the core of the protein and is unlikely to bind the RNA directly (see Fig. 2C). The mechanisms by which this amino acid facilitates RNA discrimination could be mediated by propagated effects on Loop 3 structure/dynamics.

Urb-V and Urb-Va discriminate a CG vs a UU loop-closing basepair with identical differences in binding free energy, suggesting that SRSLKMRG50 R49, conserved in all Urb-family proteins, interacts with the loop-closing G and also makes contacts with the backbone of K44 and T45. The amide group of K44 can also interact with the RNA backbone. The similarities in RNA recognition are preserved in spite of mutations to K44 and T45 in Urb-Va.
Loop 3 is also important for recognition of A1. Our results show that A1 is recognized by both Urb-V and Urb-Va. However, recognition of this site is somewhat stronger for Urb-Va than it is for Urb-V. This suggests that while the interactions between Loop 3 and A1/LCB are mostly preserved between Urb-V and Urb-Va, the changes to Loop 3 resulted in stronger interactions with A1 by Urb-Va.

Like Urb-V and Urb-Va, optimal binding of Urb-Vb to target RNA sequences requires that the RNA-binding element be present within the context of a loop that is at least 10 nucleotides long. Urb-Vb retains Urb-V's tolerance of either a C or G residue at loop position 7 but has lost recognition of the LCB and A1. The implication is that changes in other parts of the protein impede interactions with the LCB and A1. It is plausible that in Urb-Vb, specific binding to the stem–loops is mediated predominantly through interactions between the RNP motifs and the 5′ side of the loop. Discrimination between SLII and SLIV by Urb-Vb becomes apparent at lower salt concentrations where, in contrast to Urb-V and Urb-Va, Urb-Vb has a higher affinity for SLIV than for SLII.

**Urb Exchange & Dynamics**

Once the residues important for changing the RNA binding specificity between Urb-V and Urb-Vb are determined (as well as whether either or both the β2 and Loop 3 mutations are responsible for decreasing the protein’s affinity for RNA), we will be able to show how these mutations affect the backbone exchange properties. This warrants a more complete treatment of backbone exchange, which can be used to quantify the exchange rates, populations of the major and minor states, and chemical shift differences between and two states. For SNF RRM1,
backbone exchange throughout the body of the protein can be fit to a model of a single two-state exchange process, suggesting that the observed relaxation behavior is part of a correlated process, with the largest structural changes occurring at α1 and Loop 5. If this is also true of URB, how did backbone exchange properties evolve between Urb-V and Urb-Va (or Urb-Vb)? Was there a change in rates, are the populations of the two states different (or the states themselves) different, or is exchange in different regions of the protein no longer correlated? Such questions can only be answered with a more complete description of backbone exchange.
References.


27. Schimmel, P. (1989) Parameters for the molecular recognition of transfer RNAs,

*Biochemistry* 28, 2747-2759.
**Supplementary Table 1.** Ambiguity in deuterostome and vertebrate sequence reconstructions. Sites with amino acid probabilities >0.1 are shown.

<table>
<thead>
<tr>
<th>Site</th>
<th>UrbD</th>
<th>Urb V</th>
<th>Urb Va</th>
<th>Urb Vb</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>R(0.779)</td>
<td>K(0.213)</td>
<td>T(0.189)</td>
<td>L(0.274)</td>
</tr>
<tr>
<td>11</td>
<td>I(0.522)</td>
<td>V(0.478)</td>
<td>C(0.109)</td>
<td>I(0.116)</td>
</tr>
<tr>
<td>21</td>
<td>D(0.836)</td>
<td>E(0.163)</td>
<td>A(0.369)</td>
<td>G(0.113)</td>
</tr>
<tr>
<td>84</td>
<td>S(0.695)</td>
<td>A(0.301)</td>
<td>T(0.170)</td>
<td>I(0.307)</td>
</tr>
<tr>
<td>86</td>
<td>K(0.419)</td>
<td>T(0.344)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>I(0.686)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 7.

Linkage and Allostery in snRNP Protein/RNA Complexes

Sandra G. Williams and Kathleen B. Hall
Linkage and Allostery in snRNP Protein/RNA Complexes

Sandra G. Williams and Kathleen B. Hall

Department of Biochemistry and Molecular Biophysics, Washington University Medical School, St. Louis, Missouri 63110, United States

Supporting Information

ABSTRACT: Drosophila SNF is a member of the U1A/U2B′/SNF protein family that is found in U1 and U2 snRNPs, respectively. SNF also binds to the U2A′ protein, but only in the U2 snRNP. Although previous reports have implicated U2A′ as a necessary auxiliary protein for the binding of SNF to Stemloop IV, there are no mechanisms that explain the partitioning of U2A′ to the U2 snRNP and its absence from the U1 snRNP. Using in vitro RNA binding isotherms and isothermal titration calorimetry, the thermodynamics of SNF/RNA/U2A′ ternary complex formation have now been characterized. There is a very large binding cooperativity unique to Stemloop IV that favors formation of the SLIV/SNF/U2A′ complex. The binding cooperativity, or heterotropic linkage, is interpreted with respect to linked conformational equilibria of both SNF and its RNA ligand and so represents an example of protein–RNA allosteric linkage.

The spliceosome catalyzes eukaryotic pre-mRNA splicing and is one of the most complex and dynamic macromolecular machines in the nucleus. At the core of this machinery are five major snRNPs [small nuclear ribonucleoproteins (U1, U2, and U4–U6 snRNPs)], which each contain a single unique snRNA and multiple associated proteins, some of which are unique to a given snRNP and others of which are shared among snRNPs. In particular, the U1 and U2 snRNPs of many metazoans have a common protein, first identified in Drosophila. This Drosophila SNF protein (for sans fille) binds to U1 snRNA Stemloop II (SLII) and U2 snRNA Stemloop IV (SLIV). To date, there are no data regarding the in vivo function of SNF in the snRNPs, although protein mutations result in defects to Drosophila sex determination, and genetic data show that a SNF deletion is embryonic lethal in the fly.

SNF contains two RNA recognition motifs (RRMs), the first of which is responsible for specific binding to both RNAs. RRM1 and RRM2 are the most abundant RNA binding domains in eukaryotes and are characterized by an α/β sandwich topology. A nuclear magnetic resonance (NMR) solution structure of SNF RRM1 shows its classic RRM fold (Figure 1A), but there are no structures of SNF in bimolecular complexes with either SLII or SLIV. However, SNF is a member of the U1A/U2B′/SNF family of RNA binding proteins, all of which contain an N-terminal RNA binding RRM. The homology between the three RRM1s (~74% identical) allows us to use existing cocrystals of human U1A RRM1 bound to SLII and human U2B′ RRM1 bound to SLIV as models for possibly analogous SNF interactions. In cocrystals of U1A RRM1 bound to SLII, and U2B′ bound to SLIV, the RNA is spread out over the surface of the four-stranded antiparallel β-sheet. In these complexes, two aromatic amino acids stack with nucleobases (Figure 1B); we anticipate that this orientation also describes SNF/RNA complexes.

The sequences of SLII and SLIV are remarkably similar, and SNF binds to each with affinities that are uniquely dependent on salt and temperature, reflecting differences in binding mechanisms for the two RNAs. The RNA sequences are shown in Figure 1C; the conserved nucleobases in the loops (S′AUUGAC/G) are primary contacts for U1A and are likely to be maintained for SNF. Binding of SNF to SLIV is complicated by the association of SNF with U2A′ protein, which is also phylogenetically conserved in metazoans.

In Drosophila, U2A′ is a 265-residue protein that contains an N-terminal leucine-rich repeat (LRR) and a C-terminal domain which is also phylogenetically conserved in metazoans.

Received: February 13, 2014
Revised: May 16, 2014
Published: May 21, 2014

© 2014 American Chemical Society

dx.doi.org/10.1021/bi500192a | Biochemistry 2014, 53, 3529−3539

Open Access on 05/21/2015
U2A′ or with Drosophila nuclear extract, which presumably contained Drosophila U2A′. These results led to the conclusion that protein/protein interactions between U2A′ and SNF enhanced the affinity of SNF for SLIV, promoting the formation of the SLIV/SNF/U2A′ ternary complex. However, those experiments neither explained the apparent absence of a SLII/SNF/U2A′ ternary complex nor provided a mechanism that explained the formation of a SLIV/SNF/U2A′ ternary complex.

We used purified recombinant proteins and RNAs to perform in vitro experiments that compare binding in the ternary complex system (RNA/SNF/U2A′) with bimolecular binding of SNF/U2A′ and SNF/RNA complexes. Our system allows us to analyze SNF binding in terms of binding cooperativity, which we define as the degree to which binding by one ligand (RNA) affects binding of the second ligand (U2A′). Intriguingly, while the cooperativity for the SLIV/SNF/U2A′ complex is large, the cooperativity of SLII/SNF/U2A′ binding is marginal. Of most significance is the fact that the RNA-dependent thermodynamic cooperativity between protein/RNA and protein/protein interactions is sufficient to explain the characteristic partitioning behavior of U2A′ to the U2 snRNP and exclusion from the U1 snRNP. We finally describe protein/protein and protein/RNA binding in terms of allosteric models that include contributions of RNA and protein internal conformational equilibria.

# MATERIALS AND METHODS

## Protein Constructs and Purification

Full-length SNF was purified as previously described.4 A pGEX-2T plasmid containing the gene for Drosophila U2A′ was obtained from H. Salz. The U2A′ gene was subcloned into our Ptc expression vector under an isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible promoter, and the three cysteines in the protein were Quick-changed to their human sequence counterparts (C19V, C38T, and C119S) for the sake of biochemical convenience. [EMSA experiments using both constructs showed no difference in binding properties (data not shown).] The protein construct was truncated at position 180, so what we call U2A′ is the protein LRR domain. Protein expression was induced in Escherichia coli BL 21 cells at an OD of 0.8 in LB medium with 0.1 mM IPTG at 17 °C overnight to reduce the level of inclusion body formation. Cells were spun down and resuspended on ice in 30 mM sodium acetate (pH 5.3), 200 mM NaCl, 2 mM EDTA, and 8.5% sucrose. A protease inhibitor cocktail (Sigma), phenylmethylsulfonyl fluoride, and DNase II were added prior to lysis. Cells were French pressed, spun down in an ultracentrifuge, and filtered through a 0.22 μm filter, and the supernatant was loaded onto a prepacked GE Hi-Trap SP-HP cation exchange column at 4 °C. The column was washed with 50 mM NaCl and 50 mM Tris (pH 7.5). A salt gradient from 50 to 375 mM NaCl was run at a rate of 1.5 mL/min over 2.5 h. Fractions with U2A′ were collected and concentrated into 100 mM arginine, 50 mM KCl, and 10 mM cacodylate (pH 7). The arginine was necessary to maintain protein solubility at high concentrations. The concentrated protein was then run on a Superdex 75 10/300 GL (GE) gel filtration column in the same buffer, with a flow rate of 0.3 mL/min to remove impurities. The protein was eluted as a single, symmetric peak. Clean fractions were collected and concentrated to ∼100 μM for further use.

## Fluorescence Titrations

For fluorescence binding experiments, we used 6-carboxyfluorescein (6-FAM) 5′-end-labeled RNAs (IDT) with sequences of 5′-6-FAM-GGGCCCGGAUUUGACCCUGGUGUC (SLII) and 5′-6-FAM-GGCGCGGUAGUGACCUACCGCGGGGUC (SLIV). Loop nucleotides are underlined. These RNAs were also 3′-end-labeled (using T4 RNA ligase) with [α-32P]PCp (cytidine 3′,5′-bis-phosphate) to assess whether the 5′-fluorescein label affects RNA binding as measured by nuclease filter binding experiments. Filter binding assays with FAM-RNAs and in vitro T7 RNA polymerase SLII and SLIV showed no difference in dissociation constants (data not shown).

Fluorescence experiments were performed using an SLM 8000 fluorimeter. Cuvettes and stir bars were soaked in HCl for 15 min to eliminate RNase contamination, thoroughly rinsed with RNase-free water, and then blocked for 1 h with 250 mM KCl, 10 mM potassium phosphate (pH 8), 1 mM MgCl₂, and 40 μg/mL BSA. RNA stocks were diluted in water, heated to 65 °C for 5 min, and quenched on ice. A 1/10 volume of 10X buffer was added to complete RNA folding.

Fluorescence emission spectra were recorded on samples containing 10 nM RNA and variable protein concentrations (as indicated in the figures). The buffer was the same as that used for blocking. The temperature was held constant with a
circulating water bath at 23 °C. Protein stocks were sufficiently concentrated that the RNA dilution was <1%. The excitation wavelength was 490 nm, and the slit widths were 8 and 2 nm for the excitation and emission monochromators, respectively. The emission wavelength was varied between 495 and 600 nm. Background reference spectra were subtracted from the sample spectra, and the emission intensities were normalized to the maximal intensity of the free RNA.

SNF/U2A′ titrations were performed in 250 mM KCl, 10 mM potassium phosphate (pH 8), 1 mM MgCl₂, 40 μg/mL BSA, 5 mM DTT, and RNasin. Titrations were performed at 23 °C while the mixtures were being constantly stirred. For a single titration of SNF or the SNF/U2A′ complex into FAM-RNA, the cuvette and titrant concentration of FAM-RNA was held constant at 0.1 or 0.5 nM (the lower concentration was used for the highest-affinity interactions). The excitation and emission slit openings were 8 and 16 nm, respectively. SNF or the SNF/U2A′ complex was titrated into the RNA, and the fluorescence emission intensity was recorded for each addition of SNF. The intensity data were converted to fraction fluorescence enhancement and normalized to the maximal fluorescence enhancement. Titrations were collected at multiple concentrations of U2A′, and the data were globally fit in Scientist (Micromath) to eqs 1−4; fractional fluorescence enhancement corresponds to the fraction of RNA bound either by SNF or by the SNF/U2A′ complex. Titration series were repeated at least twice for each RNA. The parameter values listed in Table 1 represent the average of the series fits, with uncertainties that are the larger of either the propagated error or the standard deviation between measurements.

Partitioning surfaces were simulated in Scientist based on binding parameters determined in the fluorescence experiments. For these surfaces, SLII and SLIV were considered competitive ligands for SNF. The partitioning surfaces were plotted in MatLab.

2-Aminopurine Fluorescence Experiments. 2-Aminopurine (2AP) SLIV (Dharmacon) had the sequence 5′-GGCCGUAAUUGCAGU-2AP-CCGCGGCC. The RNA stock was diluted to 300 nM in water, heated to 95 °C for 3 min, and quenched on ice. A concentrated buffer stock was added to bring the salt concentration to 50 mM KCl, with 10 mM cacodylate (pH 7) (the lower salt concentration prevented RNA dimerization).
Cuvettes and stir bars were washed with acid and blocked with BSA as described. The temperature was held constant with a circulating water bath at 23 °C. Protein stocks were sufficiently concentrated such that the RNA dilution was <1% upon addition. The excitation wavelength was 310 nm, and the slit widths were 8 and 2 nm for the excitation and emission monochromators, respectively. A polarizer in the emission path parallel to the monochromator gratings eliminated monochromator artifacts from Wood’s anomaly. The emission wavelength was varied between 340 and 460 nm. Buoyant reference spectra were subtracted from the sample spectra, and the emission intensities were normalized to the maximal intensity of the free RNA.

Circular Dichroism (CD) Spectroscopy. CD spectra were buffer-subtracted and recorded at room temperature on a Jasco J715 instrument. RNA experiments were performed with an RNA concentration of 2 μM in 50 mM KCl and 10 mM cacodylate. Spectra were collected from 375 to 210 nm. For experiments with protein, protein was added to a concentration of 2 μM (SNF or SNF and U2A'). The hairpin RNA sequences were 5′-GGCCGCAUUGCAUCCCGGCCG (SLII) and 5′-GGCCGUAUUUGCAUCCCGGCCG (SLIV).

ITC Experiments. Protein samples were diluted from stock solutions into 100 mM arginine, 50 mM KCl, and 10 mM cacodylate (pH 7) and dialyzed in mini dialyzers (Thermo-Scientific, 2000 molecular weight cutoff) against that buffer. Final samples were prepared by diluting the protein solutions (SNF and U2A') with equal volumes of the final buffer, including 5 mM BME. Samples were degassed prior to being loaded into the ITC injection syringe or cell. Titrations were performed on a NanoITC instrument (TA Instruments) and analyzed using NanoAnalyze.

RESULTS

SNF/RNA/U2A' Ternary Complexes. We have previously determined dissociation constants for binding of SNF to SLII and SLIV.4 In those experiments, we compared the binding of FL SNF, RRM1, and RRM2. We found that RRM2 does not bind to either SLIV or SLII or to a single-stranded random sequence RNA. We also determined that FL SNF and SNF RRM1 bind with a 1:1 stoichiometry to either hairpin. We now consider three-component systems (RNA, SNF, and U2A') to explore possible mechanisms of U2A' localization.

To determine the properties of formation of the RNA/SNF/U2A' complex, the SNF/RNA binding affinity was measured at different U2A' concentrations. Binding was monitored by fluorescence intensity changes of FAM-RNA upon addition of protein [FAM does not alter RNA binding affinity (see Materials and Methods)]. Addition of a saturating amount of SNF results in a 20% enhancement of the FAM-SLII or FAM-SLIV fluorescence intensity at 520 nm. No further change in fluorescence was observed when a large excess of U2A' was added to the RNA alone or to the RNA bound to SNF (Figure 2a). For binding titrations, the fluorescence intensity can therefore be monitored to detect protein binding, and the enhancement is a result of binding of RNA to SNF alone or to the SNF/U2A' complex. Representative binding curves for these experiments are shown in panels b and c of Figure 2. We observed that the presence of U2A' imparts a marginal increase in the affinity of SNF for SLII but a very large increase in the affinity of SNF for SLIV.

A schematic of the thermodynamic cycle for ternary complex formation is shown in Figure 3, with the right panel depicting the macromolecules. On the left, S represents the SNF protein, U represents U2A', and R represents the RNA, either SLII or SLIV. The individual bimolecular binding events have characteristic binding parameters; K_R and K_U represent the bimolecular association constants for the SNF/RNA and SNF/U2A' interactions, respectively. These binding events are also characterized by free energies of binding, ΔG_R and ΔG_U2A', respectively. The ternary complexes can be formed by binding of U2A' to the preformed SNF/RNA complex or by binding of RNA to the preformed SNF/U2A' complex. These are defined by association constants K_RU and K_RU2A', respectively.

Consider that SNF (S) is the macromolecule that can bind two ligands, each of which binds at a single site, in the thermodynamic cycle shown in Figure 3. Conservation of energy requires that ρ = K_RU = αK_R and K_RU2A' = αK_RU2A', where α is the cooperativity parameter and describes the extent to which binding by one ligand affects binding of the second ligand. If α > 1, there is cooperative binding between the binding events (binding by either ligand improves binding of the second ligand). When α = 1, there is no cooperativity; binding by either ligand is independent of the other. If α < 1, there is negative cooperativity in binding of the ligands. In the case of competitive ligand binding, where binding by one ligand completely precludes binding of the second ligand, α = 0. The free energy associated with cooperativity is given by ΔG = −RT ln(α).

![Figure 3](dx.doi.org/10.1021/bi500192a) Biochemistry 2014, 53, 3529−3539
All binding data were globally fit to eqs 1–4 to obtain the two bimolecular association constants \( K_R \) and \( K_U \), as well as \( \alpha \), the cooperativity parameter.

\[
F_{S+US} = F = (1/R_T)[K^0_R S (1 + \alpha K^0_U U)]
\]  

(1)

\[
S = (S_T - FR_T)/(1 + K^0_R U)
\]  

(2)

\[
R = R_T/(1 + K^0_R S + \alpha K^0_U S U)
\]  

(3)

\[
U = U_T/(1 + K^0_S + \alpha K^0_U S R)
\]  

(4)

where \( F_{S+US} \) is the fraction of the total RNA, bound either to SNF (S) or to the U2A'/SNF complex (US); \( R_T \), \( U_T \), and \( S_T \) are the total RNA, U2A', and SNF concentrations, respectively; \( R \), \( U \), and \( S \) are the concentrations of free RNA, U2A', and SNF, respectively; \( \alpha \) is the cooperativity parameter; and \( K_R \) and \( K_U \) are the bimolecular association constants for the SNF/RNA and SNF/U2A' interactions, respectively.

We find that cooperativity of ternary complex formation depends on the RNA species bound (Table 1) [note that binding dissociation constants \( K^0 \) are given; \( K^0_{D(U,R)} = 1/K^0_{R,U} \)]. Cooperativity between U2A' and SLII binding to SNF is only marginal (\( \alpha = 2; \Delta G^\circ = -0.5 \text{ kcal/mol} \)), so it was not possible to reliably determine the bimolecular binding constant for the SNF/U2A' interaction from these titrations. Instead, the protein/protein bimolecular binding constant was fixed to the value determined in the SLIV binding assays. For SNF binding to SLIV, the cooperativity between U2A' and SLIV binding is very large; binding by either molecule increases the binding

---

**Table 1. Thermodynamic Binding Parameters for SNF, RNA, and U2A'**

<table>
<thead>
<tr>
<th></th>
<th>SNF and FAM-SLII</th>
<th>SNF and FAM-SLIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_{D,R,app} ) (M) ( (1/K^0_R) )</td>
<td>( (1.1 \pm 0.5) \times 10^{-9} )</td>
<td>( (8.3 \pm 0.4) \times 10^{-8} )</td>
</tr>
<tr>
<td>( \Delta G^\circ_{D(R,app)} ) (kcal/mol)</td>
<td>( -12.1 \pm 0.3 )</td>
<td>( -9.6 \pm 0.1 )</td>
</tr>
<tr>
<td>( K_{D,U,app} ) (M) ( (1/K^0_U) )</td>
<td>na*</td>
<td>( (1.4 \pm 0.2) \times 10^{-7} )</td>
</tr>
<tr>
<td>( \Delta G^\circ_{D(U,app)} ) (kcal/mol)</td>
<td>na*</td>
<td>( -7.9 \pm 1 )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>( 2.2 \pm 0.4 )</td>
<td>( 361 \pm 51 )</td>
</tr>
<tr>
<td>( \Delta G = -RT \ln(\alpha) ) (kcal/mol)</td>
<td>( -0.47 \pm 0.1 )</td>
<td>( -3.5 \pm 0.1 )</td>
</tr>
</tbody>
</table>

*S is RNA. U is U2A'. SLII and SLIV are labeled with fluorescein (FAM). Binding buffer for all experiments consisted of 250 mM KCl, 10 mM potassium phosphate (pH 8.0), 1 mM MgCl$_2$, 40 \( \mu \)g/mL BSA, 5 mM DTT, and RNasin at 22 °C. Parameter values reflect the average values from at least two separate data series. Uncertainties represent the larger of either the standard deviation of parameter values from different fits or the propagated error. *Data not available from this experiment.

---

**Figure 4.** Modeling protein distributions on snRNPs. (a) Thermodynamic model including both SLII and SLIV RNAs with binding parameters obtained from fluorescence titrations. (b) Fractions of SLII found in a bimolecular complex with SNF (red) and SLIV in a ternary complex (purple). (c) Partitioning surface showing the fraction of both SLII (red) and SLIV (purple) in ternary complexes. SLII is found primarily in the bimolecular complex and SLIV primarily in the ternary complex when \([\text{SNF}] > [\text{U2A}']\).
affinity for the other by a factor of 350 ($\alpha$). Even though the apparent affinity of U2A′ for SNF in the absence of RNA is only $\sim 1.5 \mu M$, the high degree of cooperativity between U2A′ and SLIV binding to SNF means that the affinity of the SLIV/SNF complex for U2A′ is 4 nM. Similarly, the apparent affinity of SNF for SLIV is shifted from 80 to 0.25 nM. Given the large cooperativity, the shift in the SLIV binding curve approaches the U2A′ saturation limit. This result is striking, corresponding to a free energy of cooperativity ($\Delta g^*$) of $\sim 3.5$ kcal/mol. This is a dramatic example of both the degree to which cooperativity can affect binding and of the RNA dependence of this phenomenon.

**In Vivo Partitioning of Proteins in snRNPs.** Using the experimentally determined thermodynamic parameters, we simulated the fraction of cellular U1 and U2 snRNA that would be bound by the various proteins when both proteins and both RNAs are considered simultaneously. Figure 4a shows a schematic of the two-protein, two-RNA system and all relevant binding constants. In this analysis, SLII and SLIV are considered to be competitive ligands for SNF. Panels b and c of Figure 4 and Figure 1 of the Supporting Information show the fraction of SLII and SLIV bound by SNF and U2A′ over a wide range of possible SNF and U2A′ concentrations. These simulations use the approximate cellular concentrations of U1 and U2 snRNAs of 3 and 1.5 $\mu M$, respectively.

Several important observations can be made from the models of protein partitioning. First, there is a significant range of U2A′ and SNF protein concentrations for which most of SLII is found in a bimolecular complex with SNF and most of SLIV is in a ternary complex with both SNF and U2A′ (Figure 4b). Second, U2A′ partitions to the U1 snRNP only when [U2A′] > [SNF] (Figure 4c), which is generally not a condition found in cells. Even though binding of U2A′ and SLII is not negatively cooperative, the difference in the free energy of binding cooperativity ($\Delta \Delta g = 3$ kcal/mol) between U2A′ and the SLII/SLIV complex binding to SNF is sufficient to effectively partition the U2A′ protein away from the U1 snRNP and into the U2 snRNP, when the concentrations of the various components are found at expected cellular levels.

**Protein/Protein Interaction.** Direct measurement of the bimolecular association of U2A′ and RRM has not been done previously. We used ITC to measure the binding thermodynamics. The titration of U2A′ with SNF shows a very large apparent enthalpy of binding (Figure 5a) that is temperature-dependent (Figure 5b), indicating a change in heat capacity ($\Delta C_p$) associated with binding. Given the nonlinearity of the temperature dependence, the data were fitted to a model that takes into account a temperature dependence of $\Delta C_p$: 170
The protein/protein binding mechanism includes burial of hydrophobic surfaces. On the basis of the SLIV/U2B'/U2A'-cocrystal structure, we calculate there is burial of 629 Å² of polar surface area and 1184 Å² of apolar surface at the U2B'/U2A' interface. Applying estimates of binding enthalpy from surface burial yields a predicted binding enthalpy ($\Delta H_p$) of $-1.5$ kcal/mol at 22 °C. The measured apparent heat capacity and enthalpy of binding for SNF/U2A' far exceed this estimate, so unless the binding of SNF to U2A' is very different from the binding of U2B' RRM1 to U2A', there must be other contributions.

Contributions to the observed enthalpy could come from coupling of protonation or ion binding and/or release to complex formation. Cacodylate was used as the buffer in most of the calorimetric titrations in part because the ionization enthalpy of cacodylate is very small ($-0.72$ kcal/mol). To estimate the effect of linked protonation equilibria, experiments were repeated in ACES buffer, which has a much higher ionization enthalpy ($7.17$ kcal/mol) (both experiments conducted at pH 7.0). This analysis showed a net release of approximately eight protons from the solvent on binding. The source of the large linkage between binding and protonation needs to be investigated to improve our understanding of the binding mechanism.

Conformational changes coupled to binding are a common source of an apparent heat capacity. We used CD to assess changes in the secondary structure of the proteins upon binding (Figure 5c). CD spectra of SNF, U2A', and a 1:1 mixture of the proteins show that the spectra are not entirely additive, suggesting some degree of change to the secondary structure upon binding. However, the difference spectrum is small compared to that of other protein–protein interactions with large values of $\Delta C_p$. For the SNF/U2A' interaction, while the changes in overall secondary structure appear to be minor, it is possible that there are significant changes in the tertiary structure of one or both components that are coupled to binding and contribute to the large apparent $\Delta H$ and $\Delta C_p$.

**Protein/RNA Interactions.** Cocrystals first suggested that RNA binding to RRMs results in significant distortion of the loops of U1 SLII and U2 SLIV. Most significantly, the RNA loop must open up upon protein binding, which allows formation of the specific contacts between the protein and RNA. To probe conformational changes to the RNA upon protein binding, we measured CD spectra of SLII and SLIV in the presence and absence of SNF and U2A' (Figure 6a,b).
Between 240 and 300 nm, the contribution of the protein to the CD signal is negligible compared to that of the RNA. Changes in the CD spectrum can therefore be attributed to changes in the RNA structure upon binding.

SNF binding results in an overall increase in the magnitude of the CD signal of the RNA band centered at ~265 nm, consistent with an increased level of base stacking. Further addition of U2A′ (and formation of the ternary complex), however, results in a significant decrease in the intensity of the CD bands, suggesting unstacking of the loop nucleobases.

The 3′-UCC of the SLII loop does not make direct contact with U1A;18 neither does the 3′-ACC of SLIV make contact with U2B′ or U2A′ in the cocrystal.3 We previously replaced the 3′-loop adenine of SLIV with 2-aminopurine (2AP) and showed that it does not affect the RNA binding affinity of SLIV for SNF;19 this nucleotide is stacked with its neighboring bases in the free RNA but becomes flipped out of the stack upon binding to SNF.19 Unexpectedly, when U2A′ is added to the preformed SNF/SLIV complex, the 2AP fluorescence intensity is quenched (Figure 6c). The signal can be recovered by addition of a large excess of SNF (data not shown), which presumably increases the relative population of the bimolecular SNF/RNA complex.

At wavelengths greater than 300 nm, 2-aminopurine can show an induced CD band that is sensitive to the environment of U2 snRNP stability and spliceosome assembly.

The molecular origin of the cooperativity (α) is the predominant unknown that arises from these results. Because the degree of cooperativity determines the form of the SNF complex in vivo, the physical basis of the thermodynamic signature is important to understand. More specifically, we want to understand the origin of α = 350-fold enhancement (positive heterotrophic linkage) of binding of SNF to either SLIV or U2A′ (upon binding by the other) with a corresponding free energy (ΔG°) of ~3.5 kcal/mol. The cooperativity is dependent on the RNA, because linkage effects between binding of SNF to U2A′ and SLII are slightly positive but weak (α = 2.2; ΔG° = ~0.5 kcal/mol). The RNA dependence of the linkage effects is sufficient to explain why U2A′ is effectively partitioned to the U2 snRNP and excluded from the U1 snRNP.

Conformational changes of proteins and RNA coupled to binding are known contributors to observed large and negative apparent binding enthalpies as well as a large apparent ΔCp." Formation of the SLIV/SNF/U2A′ complex certainly requires conformational changes of the RNA, as shown by our spectroscopic data that monitor the SLIV hairpin loop. SNF likely undergoes conformational changes upon RNA binding, as well, much as U1A RRM1 undergoes a conformational change when bound to SLII (loop 3 protrudes through the RNA loop). SNF RRM1 itself is sampling conformational space, as determined by its NMR spectra: the entire RRM1 undergoes conformational averaging on the chemical shift time scale, suggesting that it is best described as an ensemble of structures.23 In addition, the free and bound conformation of the U2A′ LRR domain could be significantly different, or its conformational sampling could be altered. Coupled conformational changes are likely to be a major contributor to the observed cooperativity, the heat capacity, and the large apparent enthalpy of binding.

Conformational changes coupled to binding might imply that a macromolecule alters its conformation only when a ligand is bound, and such is the premise of the concept of induced fit.24,25 However, the free states of SNF and the RNAs are best described as ensembles of structures. Their binding is best described by conformational selection,26,27 in which the structural ensemble is thought to include conformations that are competent to bind ligand. A recent example of this process for RNA/protein binding is seen at the single-molecule level, looking at the conformational ensemble of an RNA five-way junction before and after a protein binds (S4 protein binding to a RNA five-way junction).28 Combining the mechanisms of induced fit and conformational selection29,30 with the thermodynamics that couple conformational changes to binding will be a challenge in the SLIV/SNF/U2A′ system.

What Is α? Implications for Allostery in SNF Interactions. In 1961, Monod and Jacob introduced the term “allosteric,”31 and the first model to explain the allosteric effect was proposed in 1965.32 The model postulated that the protein existed in an equilibrium between at least two states. Since then, additional models for allosterie have emerged (most notably the KNF or sequential model).33 However, the term “allostere” has been used to encompass a much broader range of phenomena; almost any “action at a distance” has been described as allosteric. The feature common to most of what is described as allosteric is the presence of an allosteric binding site. This is a site that is distant from the functional (orthosteric) site; the allosteric site can be a catalytic site or a binding site for a second molecule. When the allosteric site is occupied, the activity of the molecule at the orthosteric site is altered.

If the SLIV/U2B′/U2A′ cocrystal structure is representative of the SLIV/SNF/U2A′ ternary complex, then binding sites for the two ligands (the RNA and U2A′) are distinct. Our data show that binding of RNA to the SNF RRM affects binding of U2A′, and vice versa. Thus, the system meets the two criteria for allosterie.

The system is unusual in terms of descriptions of allosterie because the ligands (and the ligand binding surfaces) are quite large. Using U2B′ as a model, we calculated that 40% of the SNF RRM1 surface is part of an intermolecular interface. More important than the size of the ligands, however, is the fact that
at least one ligand (the RNA) clearly experiences its own conformational heterogeneity, which is modulated by binding. Allostery in larger macromolecular complexes will need to account for conformational heterogeneity of “ligands” as well as conformational changes of the “macromolecule”.

If in a considerable simplification of the system, we consider the RNA/SNF/U2A complex in terms of a two-state SNF equilibrium ensemble, \( \alpha \) is given by (see the Supporting Information)

\[
\alpha = 1 + \left[ K_\alpha (\beta - 1) / (1 + \gamma K_\alpha) \right] / \left[ 1 + \gamma K_\alpha \right]
\]

(6)

where \( K_\alpha \) is the equilibrium constant between the two states of SNF and \( \beta \) and \( \gamma \) are the ratios of the binding constants of each state of SNF for each ligand. As a consequence, \( \alpha \) is limited by \( K_\alpha \), and regardless of \( \beta \) and \( \gamma \), the maximal value of \( \alpha \) is \( \sim 1/K_\alpha \).

For unbound SNF, this means that the free energy difference between the low- and high-affinity states must be at least 3.5 kcal/mol to account for the experimental data \( \Delta G = -RT \ln(\alpha) = -3.5 \text{ kcal/mol} \), but this difference is equal to the SNF RRM1 folding free energy \( \Delta G^\circ (\text{folding}) = -3.5 \pm 0.3 \text{ kcal/mol} \). The observed linkage \( \alpha \) between U2A’ and RNA binding to SNF could occur if the major conformation of free SNF has a weak affinity for the two ligands but a minor conformation has a high affinity for the ligands. This scenario would require that both the RNA and U2A’ binding surfaces of SNF are substantially different in the two conformations.

Assuming two-state exchange is the basis of the allosteric effect, the difference in linkage between SLII and SLIV binding and U2A’ binding could be explained by substantially different affinities of the RNAs for the two states (in eq 6, \( \beta_\text{SLIV} \gg \beta_\text{SLII} \)). However, we know that at the very least, the conformational landscape of the RNAs is best described as an ensemble of states, so we must consider whether the internal equilibria of the ligands can substantially alter the measured linkage parameter and/or allosteric response. If we introduce two-state exchange phenomena in one ligand, we obtain the following dependence of \( \alpha \):

\[
\alpha = 1 + \left[ K_\alpha (\beta - 1) / (1 + \gamma K_\alpha) \right] / \left[ 1 + \gamma K_\alpha + K_\gamma (1 + \mu K_\gamma) \right]
\]

(7)

where \( K_\gamma \) is the equilibrium constant for the ligand exchange process and \( \beta \) and \( \mu \) are the ratios of the binding constants to the two states of the macromolecule for the two states of the ligand (Figure 2 of the Supporting Information). This model requires an allosteric response of the macromolecule (if \( K_\gamma = 0 \), then \( \alpha = 1 \)). Given identical ligand exchange-independent parameters, \( \alpha_\text{(no ligand exchange)} \) can be greater than or less than \( \alpha_\text{(no ligand exchange)} \). The analysis can be extended to include internal equilibrium of both ligands, with similar results.

The ensemble allosteric model (EAM) is a more general model of allosterity that includes both MWC and KNF models as special cases. In the EAM, the two ligand binding sites can be treated formally as separate “domains” that can interact. Each domain can sample distinct conformations. Assuming two-state exchange, the equilibria between states of both domains (in the absence of interactions between them) are given by \( K_\gamma \) and \( K_\gamma \). This is modified by a factor when the two domains interact. If simple two-state ligand internal equilibria are introduced into the EAM, modulation of the linkage parameter \( \alpha \) is also seen. Like the simpler model, an allosteric response (\( \alpha \neq 1 \)) requires that the macromolecule undergo exchange. If \( K_\gamma \) or \( K_\gamma \) is zero or if there is no interaction between the domains, then \( \alpha = 1 \).

While ligand internal equilibria can modify the degree of the allosteric response, this model predicts that allostery requires an energetic change in both domains. It also predicts that the two domains thermodynamically interact when the two binding sites are occupied.

While it is possible that the two RNAs have very different \( \Delta G \) values for the states of SNF (which could account for the difference in the linkage effect), it is also possible that differences in the conformational landscapes of the RNAs (and how they bind protein) are important in the difference between \( \gamma_\text{SLII} \) and \( \gamma_\text{SLIV} \). Ligand internal equilibria can have a dramatic impact on the observed allostery of the system, but determining the thermodynamic origins of allostery in this system and in other systems will be challenging. Attention has recently focused on allosteric effects that are mediated by changes in protein dynamics, as well as changes in protein structure. 37-39 We suggest that such effects are probably ubiquitous and important in the assembly and function of larger macromolecular complexes. This is particularly likely in RNA-protein complexes, where both macromolecules are flexible.

**RNA Recognition by Proteins.** Protein recognition of RNAs is a complex process; while many structural studies have provided insight into the binding of discrete protein domains to particular tracts of RNA, most RNA binding domains are found in the context of larger proteins, which often contribute to RNA binding. Careful studies of multidomain protein recognition of RNA targets have been undertaken; these studies highlight the heterogeneity of mechanisms used to achieve RNA binding specificity.

Large changes in the free energy of binding have been reported for protein/RNA/protein complexes, in which binding by one protein is coupled to a large conformational change in the RNA, which results in a large apparent increase in the affinity for the second protein. One example occurs in 16S rRNA where S15 protein binding to the RNA was found to increase the free energy of binding of the S6/S18 heterodimer to the 16S rRNA by at least 6.5 kcal/mol. 40 Substantial work has shown that protein/protein interactions, coupled with protein/RNA interactions, very significantly impact the catalytic activity of archael RNase P; 41,42 although the thermodynamics and kinetics have not been completely resolved.

Our results show that a protein/RNA interaction can have a very large (350-fold) impact on protein/protein binding; this is an RNA-specific effect, as a highly similar RNA sequence elicits very little change in the protein/protein interaction. The effect has biological consequences, as it is sufficient to explain the protein partitioning behavior of the system and localize U2A’ exclusively to the U2 snRNP. We consider it likely that such phenomena of coupled binding are important in localizing many proteins within RNP.

**ASSOCIATED CONTENT**

Supporting Information

Simulations of protein distributions on snRNAs and the dependence of \( \alpha \) on systems where both components undergo exchange, as well as the formalism for the description of \( \alpha \). This material is available free of charge via the Internet at http://pubs.acs.org.
Biochemistry

■ AUTHOR INFORMATION

Corresponding Author
*Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8231, St. Louis, MO 63110. E-mail: kathleenhal@gmail.com. Phone: (314) 362-4196.

Funding
This work was supported by National Institutes of Health Grants SRO1 GM096444-01 to K.B.H. and F31-GM089576-01 to S.G.W.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS
We thank W. Tom Stump for lab management.

■ ABBREVIATIONS
RRM, RNA recognition motif; LRR, leucine-rich repeat; ITC, isothermal titration calorimetry; snRNP, small nuclear ribonucleoprotein; EMSA, electrophoretic mobility shift assay.

■ REFERENCES


Supplementary figure 1. Fraction of RNA (SLII or SLIV) in different compartments.
Consider a macromolecule M, which is in internal equilibrium between two states, T and R, so

\[ T \rightleftharpoons R \]  

\( K_c \) describes the equilibrium between the two conformations. M also binds ligands A and L; L is in equilibrium between two states, Z and Y; the Z \rightleftharpoons Y equilibrium is defined by the equilibrium constant, \( K_L \). The binding constants for Y and Z binding to either T or R given by \( K_{YT}, K_{ZT}, K_{YR}, \) and \( K_{ZR} \).

The different species are \( t, r, z, y \), and \( a \) (free); \( AT, AR, TZ, TY, RZ \), and \( RY \) (biomolecular complexes); and \( ATZ, ARZ, ATY, \) and \( ARY \) (ternary complex).

\[
K_a,_{app} = \frac{AT + AR}{a \times (T + R)} = \frac{K_{AT}(1 + \gamma K_c)}{1 + K_c}
\]

(where \( \gamma = K_{AR}/K_{AT} \))

\[
a \ K_a,_{app} = \frac{ATY + ATZ + ARY + ARZ}{a \times (TY + TZ + RY + RZ)}
\]

\[
= \frac{K_{AT}(K_{YT} + K_L K_{ZT}) + \gamma K_c K_{AT}(K_{YR} + K_L K_{ZR})}{K_{YT} + K_L K_{ZT} + K_c(K_{YR} + K_L K_{ZR})}
\]

\[ \alpha = \frac{K_{YT} + K_L K_{ZT} + \gamma K_c(K_{YR} + K_L K_{ZR})}{K_{YT} + K_L K_{ZT} + K_c(K_{YR} + K_L K_{ZR})} \cdot \frac{1 + K_c}{1 + \gamma K_c} \]

\[ K_{YR}/K_{YT} = \beta; \ K_{ZR}/K_{ZT} = \mu; \ K_{YR}/K_{YT} = \tau \]

\[
\alpha = \frac{1 + \tau K_L + \gamma K_c \beta + \tau \gamma \mu K_c K_L}{1 + \tau K_L + K_c \beta + \tau \gamma \mu K_c K_L} \cdot \frac{1 + K_c}{1 + \gamma K_c}
\]

\[ = 1 + \frac{K_c((\beta - 1)(\gamma - 1) + \tau K_L(\mu - 1)(\gamma - 1))}{(1 + \gamma K_c)(1 + \beta K_c + \tau K_L(1 + \mu K_c))} \]
In the absence of ligand exchange, $K_L = 0$, and $\alpha$ simplifies to:

$$\alpha = 1 + \frac{K_C((\beta-1)(\gamma-1))}{(1+\gamma K_C)(1+\beta K_C)}.$$ 

Supplementary figure 2. Two different plots of $\alpha$, for a linkage model that includes two-state ligand exchange equilibrium. $K_L$ and $\mu$ are varied, while other parameters are held constant at the values shown in the panels. For both of these plots, in the absence of ligand exchange, $\alpha$ will be constant.
Chapter 8.

Binding Affinity and Cooperativity Control U2B"/snRNA/U2A’ RNP Formation

Sandra G. Williams and Kathleen B. Hall
Binding Affinity and Cooperativity Control U2B″/snRNP/U2A′ RNP Formation

Sandra G. Williams and Kathleen B. Hall*

Department of Biochemistry and Molecular Biophysics, Washington University Medical School, St. Louis, Missouri 63110, United States

Supporting Information

ABSTRACT: The U1A and U2B″ proteins are components of the U1 and U2 snRNP s, respectively, where they bind to snRNA stemloops. While localization of U1A and U2B″ to their respective snRNP is a well-known phenomenon, binding of U2B″ to U2 snRNA is typically thought to be accompanied by the U2A′ protein. The molecular mechanisms that lead to formation of the RNA/U2B″/U2A′ complex and its localization to the U2 snRNP are investigated here, using a combination of in vitro RNA–protein and protein–protein fluorescence and isothermal titration calorimetry binding experiments. We find that U2A′ protein binds to U2B″ with nanomolar affinity but binds to U1A with only micromolar affinity. In addition, there is RNA-dependent cooperativity (linkage) between protein–protein and protein–RNA binding. The unique combination of tight binding and cooperativity ensures that the U2A′/U2B″ complex is partitioned only to the U2 snRNP.

Unlike U1A, U2B″ is bound not only to the U2 snRNA but also to the U2A′ protein. During pre-mRNA splicing, the U2 snRNA undergoes significant conformational changes, and many U2 snRNP proteins are exchanged during these rearrangements. However, U2B″ and U2A′ are found in the U2 snRNP throughout its tenure in the spliceosome. Human U2A′ is a modular protein with a 180-amino acid N-terminal leucine-rich repeat (LRR) domain and a C-terminus predicted to be mostly disordered (IUPRED). In a cocrystal of U2 snRNA stemloop IV (SLIV), U2B″ RM1, and the U2A′-LRR domain, the LRR is sandwiched between the RNA on the surface of the β-sheet and the LRR that wraps around α1 on the opposite face of the RRM (Figure 1).

In early studies, several reports concluded that U2B″ was unable to bind to the U2 snRNA specifically in the absence of U2A′. These results also showed a direct protein–protein interaction between U2B″ and U2A′ and suggested that in spite of the similarity between the U1A and U2B″ protein sequences, U1A was compromised in its ability to bind U2A′. More quantitative experiments established that U2A′ did appear to increase the apparent binding affinity of U2B″ for SLIV, supporting the existing hypothesis that U2A′ function was to increase the affinity of U2B″ for SLIV. In more recent in vitro experiments with recombinant human U2B″, we showed that U2B″ does not discriminate between SLII and SLIV but that the binding affinity is still reasonably tight for both RNAs ($K_D$...
values of $\sim 1 \times 10^{-8} \text{ M}$ in 250 mM KCl and $\sim 2 \times 10^{-9} \text{ M}$ in
100 mM KCl).11

Homologues of U1A, U2B”, and U2A’ have been found to
be essential for the viability of Drosophila and Caenorhabditis
elegans, and in both Drosophila and C. elegans, U2A’ has
functions that are independent of snRNP.19,20,22 However, the
cellular functions of U1A, U2B”, and U2A’ remain largely
elusive. The U1 snRNP can be functionally reconstituted
without U1A.13 Deletion of U2 SLIV from Xenopus snRNA and
the resulting loss of U2B”/U2A’ from the U2 snRNP do not
inhibit pre-mRNA splicing, although levels of truncated U2
snRNA and prespliceosomes were low,20,21 suggesting that the
U2B”/U2A’ complex has a function in spliceosome integrity. A
common feature of U1 and U2 snRNPs in organisms as diverse
as humans, Drosophila, C. elegans, and Saccharomyces cerevisiae is
that U2A’ localizes uniquely to the U2 snRNP and is excluded
from the U1 snRNP.17,22,23 The apparent conservation of U2A’
localization therefore appears to be an important
feature of snRNP protein composition.

Here we quantify the interactions between human U1A and
U2B”, U1 snRNA SL II and U2 snRNA SLIV, and human
U2A’ protein with the goal of understanding the mechanism of
protein localization to specific snRNPs. We find that most of
the ternary complexes formed with RRM, RNA, and LRR
domains exhibit positive thermodynamic linkage (coopera-
tivity) that enhances the stability of specific complexes. U2B”
binds U2A’ with nanomolar affinity, and the SLIV/U2B”/U2A’
ternary complex is characterized by a large cooperativity
(linkage) parameter. Surprisingly, SLIV/U1A/U2A’ binding is
also characterized by a large linkage and/or cooperativity
parameter, but the protein–protein interaction is much weaker
(micromolar), effectively preventing the formation of this
ternary complex in vivo. We find that the localization of U2A’ to
the U2 snRNP is a result of its relative binding affinities for
U1A and U2B”, as well as the RNA dependence of thermodynamic linkage between binding of SLIV and U2A’
to U2B”. The linkage between U2A’ and RNA binding also
reinforces the protein partitioning of U1A and U2B” to the U1
and U2 snRNAs, respectively. Given the phylogenetics of this
protein–RNA system and the results of our analysis, we posit
that the protein–protein interactions serve primarily to localize
U2A’ to the U2 snRNP and exclude it from the U1 snRNP,
rather than to enhance RNA binding of U2B”.

Figure 1. RRM structures and sequences. Sequence alignment of human U1A and U2B” RRM1. β-Strands and α-helices are indicated above the
sequences. Structure of human U1A (from the SLI:RRM cocrystal 1URN) indicating Tyr13 that stacks with RNA, and the sites on α1 that
potentially interact with U2A’. Structure of U2B” (from the SLIV/U2B”/U2A’ cocrystal 1A9N), in which Tyr10 stacks with RNA, while residues on
α1 are packed with U2A’. SLIV/U2B” structure from 1A9N and U2B”/U2A’ from 1A9N. To see the U2B” interfaces, the two complexes are
shown separately. Sequences of human SLII and SLIV. Cartoons constructed with VMD.38

Biochemistry

dx.doi.org/10.1021/bi500438e | Biochemistry 2014, 53, 3727–3737
**EXPERIMENTAL PROCEDURES**

**Protein Expression and Purification.** Full-length human U1A and U2B' proteins were purified as described previously. The full-length human U2A' protein was highly prone to aggregation and went entirely into inclusion bodies previously. The cells were grown in LB medium at 37 °C to an optical density of 0.6–0.8 and were induced with 0.1 mM IPTG overnight at 17 °C. Cells were harvested and stored at −70 °C until they were lyzed or processed immediately. Cells were resuspended in 30 mM sodium acetate (pH 5.3), 200 mM NaCl, 2 mM EDTA, 8.5% sucrose, and 10 mM BME. PMSF, DNase II, and a protease inhibitor cocktail (Sigma) were added prior to French pressing the cells. The lysate was collected and spun down in an ultracentrifuge at 4 °C and 45000g. The supernatant was filtered through a 0.22 μm cellulose acetate membrane and loaded onto an SP Sepharose column pre-equilibrated in 50 mM Tris (pH 7.5). U2A' was eluted over 170 ml, using a 50 to 375 mM NaCl gradient. All column buffers were sterile-filtered through 0.45 μm cellulose nitrate filters (Nalgene), and containers used in the purification were acid washed to remove RNases. Fractions containing U2A' were concentrated using a Vivaspin concentrator with a molecular mass cutoff of 10 kDa and buffer-exchanged into 100 mM arginine, 50 mM KCl, 10 mM cacodylate (pH 7), and 5 mM DTT; arginine was necessary to maintain protein solubility at high concentrations. Gel filtration of the protein with a Superdex 75 10/300 GL (GE) column was performed with a flow rate of 0.3 mL/min to remove impurities. The protein was eluted as a single symmetric peak. Clean fractions were collected and concentrated to ~100 μM for further use, and the final protein concentration was determined spectrophotometrically.

**Fluorescently Labeled RNA Hairpins.** For fluorescence binding experiments, we used chemically synthesized RNAs (IDT) with 5′-6-carboxyfluorescein (6-FAM): 5′-6-FAM-GGGCCGCAUUGCACCUCGGCGGGUC (SLII) and g5′-6-FAM-GGGCCGCGUAAUUGCAGUACC GCCGGGGUC (SLIV).

Loop nucleotides are underlined. To assess whether the 5′-fluorescein label affects RNA binding, these RNAs were 3′-end-labeled (using T4 RNA ligase) with [α-32P]pCp (cytidine 3′,5′-bis-phosphate) for use in nitrocellulose filter binding experiments. FAM-RNA and RNAs transcribed with T7 RNA polymerase were bound with equal affinity by U1A and U2B', so the FAM-RNAs were used in fluorescence experiments to measure binding affinity.

**Fluorescence Titrations.** U1A or U2B′/U2A′ titrations were performed in 250 mM KCl, 10 mM potassium phosphate (pH 8), 1 mM MgCl₂, 40 μg/mL BSA, 5 mM DTT, and RNasin. Titrations were performed at 23 °C, with constant stirring. For a single titration of U1A/U2B′ or U1A/U2B′/U2A′ into fluorescein-labeled RNA, the cuvette and titrant concentration of fluorescein-labeled RNA was held constant at 0.1 or 0.5 mM (the lower concentration was used for the highest-affinity interactions). The cuvette and titrant also contained identical concentrations of U2A'. The sample was excited at 490 nm, and the emission intensity at 520 nm was recorded (excitation and emission slit openings of 8 and 16 nm, respectively). U1A or U2B′ with or without U2A′ was titrated into the RNA, and the fluorescence emission intensity was recorded for each addition of protein. The intensity data were converted to fluorescence enhancement and normalized to the maximal fluorescence enhancement to represent the fraction of bound RNA. Titrations were collected at multiple concentrations of U2A′, and the data were globally fit in Scientist (Micromath) to eqs 1–4:

\[
F_{M+UM} = \frac{1}{R_T} [K_R M R (1 + K_U U)]
\]

\[
M = \frac{M_T - F_{M+UM} R_T}{1 + K_U U}
\]

\[
R = \frac{R_T}{1 + K_R M + K_R K_U M U}
\]

\[
U = \frac{U_T}{1 + K_U M + K_R K_U M R}
\]

where \(F_{M+UM}\) is the fraction of the total RNA, bound either to U1A/U2B′ (M) or to U2A′:U1A/U2B′ (UM); \(R_T\), \(U_T\), and \(M_T\) are the total RNA, U2A′, and U1A/U2B′ concentrations, respectively; \(R, U, M\) are the concentrations of free RNA, U2A′, and U1A or U2B′, respectively; \(R_T\) is the cooperativity parameter; and \(K_R\) and \(K_U\) are the bimolecular association constants for the SNF–RNA and SNF–U2A′ interactions.
respectively. The schematic for data analysis in terms of a thermodynamic cycle of protein and RNA binding is illustrated in Figure 2.

Titration series were performed at least twice for each RNA. The parameter values represent the average of the series fits, with uncertainties that are the larger of either the propagated error or the standard deviation between measurements. For SLII, the difference in binding affinity with or without U2A is small, such that U2A binding affinity could not be extracted from these experiments. In fits of the binding data, the U1A–U2A binding constant was fixed to the value obtained from experiments with SLIV. This allowed fitting of the linkage parameter α.

Partitioning surfaces were calculated in Scientist based on the model parameters determined in the fluorescence binding experiments. For these surfaces, SLII and SLIV were considered competitive ligands. The partitioning surfaces were plotted in MatLab.

**ITC Experiments.** RNAs for ITC experiments were transcribed in vitro using T7 RNA polymerase, purified via polyacrylamide gel electrophoresis, and reconstituted in water following ethanol precipitation. Concentrations were determined spectrophotometrically. The RNA was refolded by being heated to 95 °C for 3 min and then quenched on ice. Buffer was added to a total volume of 50 μL and this was placed in mini dialyzers [ThermoScientific, 2000 molecular weight cutoff (MWCO)] to dialyze against the final buffer [250 mM KCl, 10 mM potassium phosphate, and 1 mM MgCl₂ (pH 8)]. The RNA sequences used in ITC experiments were 5′-GGGCAUUGCCACUUGGCCAAGCUUCGGUC (SLII) and 5′-GGGAUUUUCAGGACGUAAUUGCAGUACCCUGCUUC−GG (SLIV). Loop nucleotides are underlined. The longer constructs were necessary to reduce the level of RNA dimerization at concentrations up to 1 mM: SLII includes a UUCG tetraloop at its 5′-end, and SLIV includes a poly-U tail at its 5′-end. Shorter constructs (like those used in the fluorescence binding experiments) were found to dimerize in a concentration-dependent manner, beginning at ~10 μM (data not shown). The longer RNA constructs bind to the RRM proteins with affinities that are identical within error to those of shorter RNA constructs, when assayed by nitrocellulose filter binding (not shown).

Protein samples were diluted to 2 times their final concentration from stock solutions into 250 mM KCl, 10 mM potassium phosphate, and 1 mM MgCl₂ (pH 8) (buffer used in experiments) and dialyzed in mini dialyzers (ThermoScientific, 2000 MWCO) against the experimental buffer. Final samples were prepared by diluting the RNA and/or protein samples (U1A, U2B, U2A, or a mixture of these) with equal volumes of the final buffer supplemented with a final BME concentration of 5 mM. Samples were degassed prior to being loaded into the ITC injection syringe or cell. Titrations were performed on a NanoITC instrument (TA Instruments) and analyzed using the Triple Complex model in SedPhat.²⁷

**RESULTS**

**RRMs, RNA, and U2A.** To determine the thermodynamic parameters of ternary complex formation, we performed titrations of full-length U1A and U2B into RNAs, with and without U2A (we use only the LRR domain of human U2A). For these experiments, the RNAs were labeled at the 5′-end with fluorescein (FAM). Protein binding by either U1A or U2B results in a 20% increase in FAM fluorescence upon saturation by either protein (not shown). Addition of excess U2B to the RNA did not change the RNA fluorescence, and the fluorescence of RNA/U1A/U2B complexes was not altered by the presence of U2A.

Binding of U2B and U1A to RNA was measured directly in the fluorescence experiments with or without U2A. Representative data are shown in Figure 3, and data series were fit to a binding model that takes into account cooperativity between the protein–protein and protein–RNA interaction (see schematic and Figure 2). This provides estimates for the protein–protein and protein–RNA bimolecular binding constants, as well as the concentration-independent linkage parameter α (Tables 1 and 2).

U1A binds with subnanomolar affinity to SLII, and as Figure 3A shows, addition of U2A modestly increases the affinity. However, these titrations show that addition of U2A significantly increases the affinity of U1A for SLIV, and fitting these data to our binding model allowed us to estimate the affinity of U2A for U1A [K_{U2A,app} = (1.4 ± 0.2) × 10⁻³ M⁻¹]. We used this value for fitting the cooperativity parameter α for the U2A/U1A/SLII titrations; all of the binding parameters associated with U1A are summarized in Table 1. Rather surprisingly given the in vivo snRNA partitioning of U1A, the linkage parameter (α) for SLIV/U1A/U2A binding is 89 ± 11, corresponding to a substantial increase in the apparent U1A/SLIV binding affinity when U2A is present.

As a short aside, full length (FL) U1A bound to SLIV with an affinity surprisingly high compared to values reported previously.²⁴ However, most studies to date have used RRM1 constructs to study RNA binding by U1A, and indeed, U1A RRM1 and FL U1A bind SLII with very similar affinities. In contrast, SLIV binding is substantially influenced by U1A construct length, with FL U1A binding more tightly to SLIV than U1A RRM1 (Figure 1 of the Supporting Information), accounting for the discrepancies in SLIV binding between this study and other studies. The U1A interdomain linker is highly positively charged, in particular at its N-terminus, which could contribute to the increased affinity of FL U1A for SLIV (the affinity of RRM1 for SLII is so tight that an effect would not be easily measured). Linker effects on binding affinity have been seen in other members of this protein family,²⁵ and we suspect that contributions of the interdomain linker to RNA binding may be a fairly general method of increasing the binding affinity of this family of proteins for RNA.

The most dramatic U2A-dependent enhancement in RNA binding affinity was seen in the U2B−/SLIV titrations (Figure 3B and Table 2). Formation of the ternary complex is facilitated by thermodynamic linkage (cooperativity α of 140). The affinity of U2B for SLII is also enhanced by the presence of U2A; however, the cooperativity parameter is smaller than that for the SLIV interaction by a factor of ~10 (α = 15). The free energy associated with this cooperativity (Δg) is ~−2.9 ± 0.2 kcal/mol for SLIV/U2B/U2A and ~−1.6 ± 0.1 kcal/mol for SLII/U2B/U2A. Linkage in the U2A/U2B/SLIV ternary complex leads to preferential stabilization of this species over the bimolecular species.

**Localization of the Protein to the snRNPs.** Partitioning of protein to the U1 and U2 snRNAs was modeled using the experimentally determined binding parameters. This is illustrated in simulations of the fractions of SLII and SLIV found in a bimolecular complex with U1A and ternary complex with U2B and U2A (Figure 4; the populations of other species on SLII and SLIV are shown in Figures 2 and 3 of the
10 mM sodium phosphate, and 1 mM MgCl₂ concentration was kept constant for any given titration and is indicated in the figure. All fluorescence experiments were conducted in 250 mM KCl, 10 mM sodium phosphate, and 1 mM MgCl₂ (pH 8) at 22 °C. The linkage parameter for complex formation (α) is calculated to be 7.3 for U1A/SLII/2A; 89 for U1A/SLIV/2A; 16 for U2B'/SLII/2A; and 140 for U2B'/SLIV/2A.

Table 1. Binding Parameters from Global Fits of Titrations of U1A/U2A′ into Fluorescein-Labeled SLII and SLIV

<table>
<thead>
<tr>
<th></th>
<th>SLII</th>
<th>SLIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>K_D,RNA,app (M)</td>
<td>(3.3 ± 0.3) × 10⁻¹⁰</td>
<td>(5.8 ± 0.4) × 10⁻⁶</td>
</tr>
<tr>
<td>ΔG°,RNA (kcal/mol)</td>
<td>-12.8 ± 0.1</td>
<td>-9.8 ± 0.1</td>
</tr>
<tr>
<td>K_D,U2A,app (M)</td>
<td>-</td>
<td>(1.4 ± 0.2) × 10⁻⁶</td>
</tr>
<tr>
<td>ΔG°,U2A (kcal/mol)</td>
<td>-</td>
<td>-7.9 ± 0.1</td>
</tr>
<tr>
<td>α</td>
<td>7.3 ± 1.3</td>
<td>89 ± 11</td>
</tr>
<tr>
<td>Δg (kcal/mol)</td>
<td>-1.2 ± 0.1</td>
<td>-2.6 ± 0.1</td>
</tr>
</tbody>
</table>

“Dissociation constants are reported here. The affinity of U1A for U2A′ was too weak to measure accurately in these experiments. Parameter values reflect the average values from at least two separate data series. The uncertainty represents the larger of either the standard deviation of the parameter values from different fits or the propagated error.

Table 2. Binding Parameters from Global Fits of Titrations of U2B″/U2A′ into Fluorescein-Labeled SLII and SLIV

<table>
<thead>
<tr>
<th></th>
<th>SLII</th>
<th>SLIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>K_D,RNA,app (M)</td>
<td>(1.5 ± 0.1) × 10⁻⁶</td>
<td>(1.7 ± 0.3) × 10⁻⁴</td>
</tr>
<tr>
<td>ΔG°,RNA (kcal/mol)</td>
<td>-10.6 ± 0.1</td>
<td>-10.5 ± 0.1</td>
</tr>
<tr>
<td>K_D,U2A,app (M)</td>
<td>(5.1 ± 1.1) × 10⁻⁹</td>
<td>(4.4 ± 2.4) × 10⁻⁷</td>
</tr>
<tr>
<td>ΔG°,U2A (kcal/mol)</td>
<td>-11.2 ± 0.1</td>
<td>-11.3 ± 0.3</td>
</tr>
<tr>
<td>α</td>
<td>15.8 ± 1.6</td>
<td>139 ± 49</td>
</tr>
<tr>
<td>Δg (kcal/mol)</td>
<td>-1.6 ± 0.1</td>
<td>-2.9 ± 0.2</td>
</tr>
</tbody>
</table>

“Dissociation constants are used here. Parameter values reflect the average values from at least two separate data series. The uncertainty represents the larger of either the standard deviation of the parameter values from different fits or the propagated error.

for U1 SLII and 1.5 μM for U2 SLIV. It is also known that U1A is found at levels in the cell higher than those of U2B′, so simulations were performed assuming [U1A] = 2[U2B′]. Results from additional simulations conducted at various U1A/U2B′ ratios of ≥1 showed overall results similar to those found with a 2:1 ratio (Figure 4 of the Supporting Information).

Figure 4 shows that unless U2A′ concentrations are in excess of U2B′, then U2B′ and U2A′ are effectively excluded from binding U1 SLII, and the ternary complex with U2B′ readily forms on U2 SLIV (U1A is also effectively excluded from binding). The protein concentration ranges over which a bimolecular complex is formed on the U1 snRNA and a ternary complex is formed on the U2 snRNA indicate that the thermodynamics of the systems effectively partition U2A′ to the U2 snRNP and prevent incorporation into the U1 snRNP.

Protein–Protein Interactions. The LRR domain of U2A′ surrounds α1 of U2B′ as illustrated in Figure 1. At this interface, several charged residues from U2B′ (Arg25 and Glu22) form a polar patch that makes contact with U2A′. Despite this conservation of charge, the difference in the binding affinities of U1A and U2B′ for U2A′ is quite large (nearly 3 orders of magnitude). To further probe the thermodynamics of the interactions, we measured the protein–protein interactions directly by ITC. Calorimetric titrations of binding of U1A and U2B′ to U2A′ are shown in Figure 5. Under our experimental solution conditions, the apparent enthalpy of U1A/U2A′ binding is slightly unfavorable (ΔH° = -2 ± 1 kcal/mol), so this association is entropically driven. In contrast, U2B′ binds to U2A′ with a favorable apparent enthalpy of binding (ΔH° = -85 ± 2 kcal/mol).
Using the cocrystal structure \(^{14}\) of SLIV/U2B"/U2A (Figure 1) to estimate the surface areas, the binding enthalpy (at 22 °C) that can be expected from the buried surface area of the two proteins \((\Delta H^\circ)\) is \(-15\) kcal/mol.\(^{27}\) Our experimental U2B"/U2A binding enthalpy is 6-fold larger and favorable. It could originate from conformational changes, linked protonation, or both, which are linked to binding. U2B" uses several charged residues to contact the U2A' surface, which suggests electrostatic interactions also play a role at this interface.

**Calorimetric Analysis of Ternary Complex Formation.** Having determined the calorimetric thermodynamic parameters for the binary protein—protein interactions, we determined the calorimetric parameters for ternary complex formation. Figure 6 shows results from ITC experiments based on titrations of U1A into U2B', RNA into U1A, and RNA into an equimolar mix of U1A and U2A'. Similar results are shown for U2B" in Figure 7. These experiments were performed under the same solution conditions as in the fluorescence binding assays. Fitting parameters from global fits of the data are listed in Tables 3 and 4.

The calorimetry data were fit in SEDPHAT to a binding model that used the injection heats as input.\(^{28,29}\) This provides estimates for the binding enthalpies associated with both of the bimolecular interactions \((\Delta H^\circ_{U2A} \text{ and } \Delta H^\circ_{RNA})\) as well as the enthalpy associated with linkage \((\Delta \Delta H^\circ)\). The calorimetric measurements are more limited in their sensitivity to tight binding than the fluorescence measurements. In particular, the apparent binding of U2B" to SLIV in the presence of U2A' is very tight, producing an extremely steep transition in the titration (Figure 7, bottom right), making it impossible to estimate \(\alpha_{U2A'/U2B''/SLIV}\) with any reliability. Similarly, binding of U1A to SLII is very tight (in the presence and absence of U2A'), making estimates of \(K_{\text{app}U1A/SLII}\) and \(\alpha_{U2A'/U1A/SLII}\) less reliable than those obtained from the fluorescence-based titrations. Within the limitations of the calorimetric data, agreement between spectroscopic and calorimetric data is reasonable, while the binding enthalpies (including the enthalpic contribution to linkage) are accurate and robust.

**Figure 4.** Protein partitioning simulations. (a) Thermodynamic model including all components (U1A, U2B", U2A', SLII, and SLIV) with all binding parameters obtained from fluorescence titrations. Protein partitioning surfaces were calculated over a wide range of U1A, U2B", and U2A' concentrations, given the thermodynamic parameters in panel a. (b) Fractions of SLII found in a bimolecular complex with U1A (red) and of SLIV in a ternary complex with U2B" and U2A' (blue) are shown over a range of U1A, U2B", and U2A' concentrations. (C) Similar partitioning surface showing the fraction of SLII in a ternary complex with U1A and U2A' (red) and SLIV in a ternary complex with U2B" and U2A' (blue). These simulations were conducted assuming [U1A] = 2[U2B"].
accompanied by a larger entropic penalty under these solution conditions.

We have already noted that under the conditions studied, there is significant positive linkage ($\alpha$) between U2A$^{\prime}$ and SLII binding to U1A or U2B$^{\prime\prime}$. The calorimetric titrations show, however, that this is not the result of a net increase in the apparent enthalpy of binding. The enthalpy $\Delta h$ associated with the linked equilibria is approximately twice as large for binding of protein to SLII ($\Delta h = 4.5$ kcal/mol) compared to binding to SLIV ($\Delta h = 2$ kcal/mol); both are unfavorable. Instead, the origin of the positive linkage is entropic. For these complexes, possible entropic contributions could come from water or ion release or the increased flexibility of a structural element.

Binding of U1A and U2B$^{\prime\prime}$ to SLII and SLIV is characterized by salt dependence, enthalpy−entropy compensation, and a negative heat capacity ($\Delta C_{\text{obs}}^{\text{p}} = 0.03$). Previous van’t Hoff determinations of U1A/SLII binding thermodynamics gave the following values: $\Delta H^\circ = -34$ kcal/mol and $\Delta S^\circ = -74$ eu at 22 °C in 200 mM NaCl, 1 mM MgCl$_2$, and 10 mM sodium cacodylate (pH 8), with a heat capacity $\Delta C_{\text{obs}}^{\text{p}}$ of $-3.1 \pm 0.4$ kcal mol$^{-1}$ K$^{-1}$. Binding of SLIV to U2B$^{\prime\prime}$ in 250 mM KCl could be fit to a linear van’t Hoff equation to give the following binding thermodynamics: $\Delta H^\circ = -16 \pm 1$ kcal/mol and $\Delta S^\circ = -21.2 \pm 3.5$ cal mol$^{-1}$ K$^{-1}$. These van’t Hoff values for the enthalpy are in excellent agreement with our new calorimetric determinations at the same temperature at nearly identical salt concentrations. The effect of U2A$^{\prime}$ on the temperature dependence of RNA binding thermodynamics is still to be determined.

**DISCUSSION**

Human U1A and U2B$^{\prime\prime}$ proteins have very different affinities and specificities for U1 snRNA SLII and U2 snRNA SLIV$^{11,32}$ and so segregate to the U1 and U2 snRNP.$^{15,12}$ We propose, however, that the localization of U2A$^{\prime}$ to the U2 snRNP is the
primary raison d’être for the combination of binding affinities and cooperativity among these RNAs and proteins.

In humans, U2A’ localization is largely accomplished by different intrinsic affinities of the protein for U1A and U2B’ and by stronger linkage among SLIV, U2B’, and U2A’ than among SLII, U2B’, and U2A’. The strong intrinsic affinity of U1A for SLII and its relative abundance are also important, as the binding affinity of U1A for SLII is still approximately 5-fold tighter than the affinity of U2B’/U2A’ for SLII. These factors, together with the relative paucity of U2A’, are sufficient to maintain U1A localization to the U1 snRNP and restrict U2B’/U2A’ binding.

The molecular mechanism of the cooperativity we observe in the interaction of U2B’ with SLIV and U2A’ remains unknown, and we can only speculate about its origins. Ternary complex formation that includes SLIV is facilitated by cooperativity, so the SLIV sequence certainly contributes to the binding mechanism. SLII and SLIV differ in their loop-closing base pairs, in the identity of the seventh loop nucleotide (C in SLII and G in SLIV), and in an A inserted on the 3’-side

Table 3. Parameters for Global Fits of Calorimetric Titrations for U1A-Related Thermodynamic Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SLII</th>
<th>SLIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{1A,2A}$ (M)</td>
<td>$(1.0 \pm 0.9) \times 10^{-6}$</td>
<td>$(8.3 \pm 2.2) \times 10^{-7}$</td>
</tr>
<tr>
<td>$\Delta G_{1A,2A}$ (kcal/mol)</td>
<td>$-8.1 \pm 1.5$</td>
<td>$-8.2 \pm 0.9$</td>
</tr>
<tr>
<td>$\Delta H_{1A,2A}$ (kcal/mol)</td>
<td>$2.9 \pm 1.2$</td>
<td>$2.8 \pm 0.6$</td>
</tr>
<tr>
<td>$K_{D,RNA}$ (M)</td>
<td>$&lt;7 \times 10^{-9}$</td>
<td>$(1.0 \pm 1.2) \times 10^{-7}$</td>
</tr>
<tr>
<td>$\Delta G_{D,RNA}$ (kcal/mol)</td>
<td>$&lt;-11$</td>
<td>$-9.1 \pm 0.4$</td>
</tr>
<tr>
<td>$\Delta H_{D,RNA}$ (kcal/mol)</td>
<td>$-23.5 \pm 0.7$</td>
<td>$-14.8 \pm 1.7$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$1.4 \pm 7.9$</td>
<td>$34.0 \pm 17.0$</td>
</tr>
<tr>
<td>$\Delta g$ (kcal/mol)</td>
<td>$-0.2 \pm 1.4$</td>
<td>$-2.1 \pm 0.6$</td>
</tr>
<tr>
<td>$\Delta h$ (kcal/mol)</td>
<td>$4.6 \pm 2.5$</td>
<td>$1.5 \pm 0.6$</td>
</tr>
</tbody>
</table>

*Dissociation constants are used here. U1A binds too tightly to SLII to accurately measure affinity by ITC. Parameter values reflect the average values from at least two separate data series. The uncertainty represents the larger of either the standard deviation of the parameter values from different fits or the propagated error.

Figure 7. Calorimetric titrations and ternary complex formation for U2B’. A titration of U2B’ into U2A’ is shown on the left (red squares), along with fits from global analyses of either SLII (top right) or SLIV (bottom right). RNA was titrated into U2B’ (purple squares) or into an equimolar mix of U2B’ and U2A’ (green diamonds). The results of globally fitting the data for each set of experiments (lines) are also shown. Titrations were conducted in 250 mM KCl, 10 mM potassium phosphate, and 1 mM MgCl₂ (pH 8) at 22.5 °C, and parameters from the fits are listed in Table 4.

Table 4. Parameters for Global Fits of Calorimetric Titrations for U2B’-Related Thermodynamic Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SLII</th>
<th>SLIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{D,UB}$ (M)</td>
<td>$(5.9 \pm 1.1) \times 10^{-9}$</td>
<td>$(6.0 \pm 1.1) \times 10^{-9}$</td>
</tr>
<tr>
<td>$\Delta G_{D,UB}$ (kcal/mol)</td>
<td>$-11.1 \pm 0.1$</td>
<td>$-11.1 \pm 0.1$</td>
</tr>
<tr>
<td>$\Delta H_{D,UB}$ (kcal/mol)</td>
<td>$-73.2 \pm 1.4$</td>
<td>$-73.2 \pm 1.3$</td>
</tr>
<tr>
<td>$K_{D,RNA}$ (M)</td>
<td>$(9.2 \pm 3.9) \times 10^{-8}$</td>
<td>$(2.3 \pm 0.5) \times 10^{-9}$</td>
</tr>
<tr>
<td>$\Delta G_{D,RNA}$ (kcal/mol)</td>
<td>$-9.5 \pm 0.5$</td>
<td>$-10.3 \pm 0.2$</td>
</tr>
<tr>
<td>$\Delta H_{D,RNA}$ (kcal/mol)</td>
<td>$-13.3 \pm 1.0$</td>
<td>$-14.9 \pm 0.2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$7.4 \pm 6.8$</td>
<td>$&gt;30$</td>
</tr>
<tr>
<td>$\Delta g$ (kcal/mol)</td>
<td>$-1.2 \pm 0.5$</td>
<td>$&lt;-2$</td>
</tr>
<tr>
<td>$\Delta h$ (kcal/mol)</td>
<td>$4.5 \pm 0.8$</td>
<td>$2.1 \pm 0.5$</td>
</tr>
</tbody>
</table>

*Dissociation constants are reported here. Values of cooperativity ($\alpha$) could not be accurately determined in ITC experiments with SLIV. Parameter values reflect the average values from at least two experiments. The uncertainty represents the larger of either the standard deviation of the parameter values from different fits or the propagated error.
of the SLIV loop. How U2B′ interacts with these sites on SLIV could determine how it responds to U2A′ binding, resulting in the cooperativity we observe. However, in the reciprocal pathway for ternary complex formation, U2B′ first interacts with U2A′ and also leads to cooperative binding by SLIV. We suspect that loop 3 of U2B′ interacts with SLII and SLIV very differently, specifically near the loop-closing base pair, and that its interactions and flexibility in either binary complex enhance its subsequent interactions in the ternary complex. Loop 3 is a notable site of amino acid variation in this family of RRM's, particularly at its N-terminus, and the unique combination of U2B′ loop 3 amino acids, SLIV, and (uncharacterized) U2A′ amino acids could lead to cooperativity in forming the SLIV/ U2B′/U2A′ ternary complex.

**Linkage in the U1A/U2B′/SNF Family.** The binding parameters that we have determined for this system are sufficient to explain the *in vivo* localization of the different protein components to the U1 and U2 snRNPs. Their reliance on intrinsic differences in binding affinity for SLII and SLIV as a dominant mechanism of localization is strikingly different from that of SNF, the single *Drosophila* protein that binds to both snRNAs. In the protein phylogeny, U1A and U2B′ emerged after a relatively recent gene duplication, while SNF is more closely related to the single ancestral protein. U1A, U2B′, and SNF share many properties, and among these is strong linkage with SLIV binding. All these proteins use an RNA-dependent cooperative binding mechanism to guide their snRNP localization, but while these properties of linked equilibria are fundamental to SNF segregation and contribute significantly to U2B′ ternary complex formation, they would appear to be vestigial for U1A function.

The localization of U2A′ to the U2 snRNP is accomplished through distinct mechanisms in humans and *Drosophila*, and it is of interest to compare the two systems. SNF binds to both U1 snRNA SLII and U2 snRNA SLIV, but only when it is in the U2 snRNP does it form the ternary complex with RNA and dmU2A′ (dm is *Drosophila*). SNF's affinity for dmU2A′ is modest, but binding is very strongly coupled to SLIV binding, increasing the apparent affinity for the RNA by a factor of 350. This effect is a reciprocal one; so SLIV binding increases the apparent binding affinity of the SNF—dmU2A′ interaction by 350-fold. In contrast, dmU2A′ binding has almost no effect on SNF–SLII interactions. The result is that the very large difference in cooperativity is effective in partitioning U2A′ to the U2 snRNP.

Following the gene duplication in an ancestor of jawed vertebrates, one protein evolved to bind U1 SLII with very high affinity and specificity, eventually becoming human U1A. In contrast, the second protein evolved to lose specificity for SLII. The system also evolved a large difference in the intrinsic binding affinities of the two RRM paralogs for U2A′. Given the single protein origin of the phylogeny, it is instructive to think about how the proteins would be expected to partition if a single protein with the characteristics of either U1A or U2B′ were present in humans. Results from simulations similar to those shown in Figure 4 (maintaining the U1A and U2B′ binding parameters but considering only U1A or U2B′ to be present) are shown in Figure 5 of the Supporting Information. If U2B′ were to be lost from human cells, the partitioning of U2A′ to the U2 snRNP would be significantly compromised over a substantial range of protein concentrations. If U1A were lost, the population of U2A′ on the U2 snRNP would be much more significant.

Reports of U2A′ function suggest that its principal biochemical role is to increase the binding affinity of U2B′ for SLIV. While linkage analysis shows that this effect undoubtedly occurs, it seems unlikely that this is the purpose of the protein. Most metazoans function with a single SLII/SLIV binding protein (as in flies), for which a duplicate high-affinity RNA target on the U2 snRNA could easily circumvent the need for an auxiliary protein. In systems with separate U1A and U2B′ proteins, the system would again be more parsimonious with two high-affinity, high-specificity RNAs determining protein partitioning. Therefore, it seems more likely that both systems evolved to execute U2A′ from the U1 snRNP and localize it to the U2 snRNP.

**Biological Necessity of U2B′/U2A′.** In experiments with *Xenopus* oocytes in which endogenous U2 snRNA was inactivated, pre-mRNA splicing could be rescued when exogenous U2 snRNA was expressed. This system was used to investigate the regions of U2 snRNA that are required for spliceosome formation and splicing. In independent experiments, these investigators deleted SLIV in exogenous U2 snRNA and observed that splicing was impaired but not inactivated. Examination of spliceosome assembly revealed that complex A (containing U1 snRNP and U2AF protein) was present in large amounts but complex B (where U2 snRNP is added) and complex C were not detected. Because splicing was observed (complex C is the active spliceosome), Hamm et al. concluded that with the truncated U2 snRNA and the consequent absence of U2B′/U2A′, these complexes were unstable and did not survive purification. Pan and Privé came to a similar conclusion, as they monitored splicing of endogenously transcribed SV40 pre-mRNA in *Xenopus* oocytes. They concluded that U2 snRNA lacking SLIV (and therefore also U2B′/U2A′) was unstable.

*Yeast (S. cerevisiae)* contain both U2B′ (Yib9p or YU2B′) and U2A′ (Lea1p). Cells lacking Lea1p, Yip9p, or both spliced at greatly reduced levels had slow growth, and levels of U2 snRNA were low. Those yeast cells accumulated Commitment Complex 2 (lacking U2 snRNP), and no prespliceosome was detected. In other studies, the yeast U2 snRNA SLIV was deleted; the result was that pre-mRNA splicing was inhibited but not abolished. As Caspary and Seraphin concluded, both Yip9p and Lea1p are essential for efficient prespliceosome formation. Yeast U2B′ has only one RRM, and unlike U2A′, the C-terminal domain of Lea1p is not predicted to be disordered. These different features of the proteins did not seem to preclude their binding to human U2 snRNA when it replaced the yeast snRNA in vivo. In another report, the association of a GST-YU2B′ with human SLIV by an electrophoretic mobility shift assay could not be detected, although the protein did bind to human U1 SLII. The necessity for sequestering U2B′/U2A′ to the U2 snRNP may have its origin in the (unknown) function of the C-terminal tail of U2A′. This region of U2A′ is predicted to be mostly disordered, although there is a putative helical region in human and *Drosophila* proteins. There are many proteins associated with the U2 snRNP, and a large proportion of them are transiently bound. On the basis of the *Xenopus* and yeast results, it has been suggested that U2B′/U2A′ may be essential for stable spliceosome formation. Our investigations have provided a mechanism that explains how the SLIV/U2B′/U2A′ ternary complex is localized to the U2 snRNP. We propose that U2A′ provides protein–protein interactions that stabilize the
U2 snRNP and the prespliceosome and that U2B′ is the scaffold that anchors U2A′ to the snRNP.

**ASSOCIATED CONTENT**

3 Supporting Information

Binding isotherms and simulations of protein distributions on snRNAs. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

Corresponding Author

*Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8231, St. Louis, MO 63110. E-mail: kathleenhal@gmail.com. Phone: (314) 362-4196.

Funding

This work was funded by National Institutes of Health Grants 5 R01 GM096444 to K.B.H. and F31-GM089576 to S.G.W.

Notes

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We thank Michael Rau for his help finding RNA constructs that were not prone to dimerization. He also transcribed and purified some of the RNAs used in this work. We thank Dr. Kim Delaney for cloning the human U2A′ protein.

**ABBREVIATIONS**

RNP, ribonucleoprotein; RRN, RNA recognition motif; LRR, leucine-rich repeat; ITC, isothermal titration calorimetry; snRNP, small nuclear ribonucleoprotein.

**REFERENCES**


Supplementary figure 1.

Binding curves and fits for FL U1A (☐, ■) and U1A RRM1 (△, ▲) binding to SLIV (closed symbols) or N25 (open symbols) RNA. N25 is an RNA with a 25 randommer sequence, used as a control for nonspecific binding. Experiments were done with nitrocellulose filter binding. The solution conditions were 100mM KCl 10mM cacodylate 1mM MgCl2 pH 7 at 22°C.
Supplementary Figure 2.

SLII protein partitioning in humans.
Supplementary figure 3.
Protein partitioning on SLIV in humans.
Supplementary Figure 4.
Ternary complex formation on SLII and SLIV, at variable ratios of U1A:U2B."
Supplementary figure 5.

(A.) SNF only.

(B.) U1A only.

U2B only.