Resurrection of an Urbilaterian 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...
(RRM1) are distinguished from most RRM s by their extremely high affinity and the exquisite specificity of their RNA binding. The RRM s 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\textsuperscript{*}, 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\textsuperscript{*} have distinct RNA binding properties. Drosophila SNF binds both U1 snRNA SLII and U2 snRNA SLIV \cite{1}, whereas in humans, U1A binds exclusively to U1 SLII, and in nuclear extracts, U2B\textsuperscript{*} localizes exclusively to the U2 snRNP, where it binds SLIV \cite{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\textsuperscript{*}, and an intermediate specificity for SNF \cite{3–5}. Therefore, comparing the modern RRM s, 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 \cite{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 \cite{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 \cite{11}.

We used U1A/U2B\textsuperscript{*}/SNF sequences from broadly diverse metazoans to reconstruct the metazoan protein phylogeny and resurrect the ancestral protein of the last common ancestor of human U1A, human U2B\textsuperscript{*}, 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\textsuperscript{*}/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\textsuperscript{*} proteins are relatively recent. Like its modern counterparts, the resurrected URB protein (which corresponds to the last common ancestor of bilaterians) has two RRM s 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\textsuperscript{*} proteins, prior characterization of U1A/U2B\textsuperscript{*} molecular evolution postulated that U1A and U2B\textsuperscript{*} paralogs emerged after a single, ancient gene duplication prior to the divergence of plants, fungi, and metazoans \cite{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\textsuperscript{*}, and SNF proteins from 160 diverse metazoan organisms were obtained from BLAST searches and subsequently aligned. PhyML \cite{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\textsuperscript{*} 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\textsuperscript{*} 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 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 Urbilaterian 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...
Reconstruction of an Ancestral RRM

Fig. 2 (legend on next page)

Please cite this article as: Williams, S. G. et al, Resurrection of an Urbilaterian U1A/U2B″/SNF Protein, J. Mol. Biol. (2013), http://dx.doi.org/10.1016/j.jmb.2013.05.031
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 SLI 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 ∆∆G°(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. 2.** The RRMs of human U1A, human U2B*, and Drosophila SNF have remained highly conserved with those of their Urbilaterian ancestor. (a) Alignments of the ML ancestral protein (URB) and SNF with both human U1A and U2B* are shown. Residues are highlighted in red if shared with human U1A only and in blue if shared with U2B* only. RNP motifs are italicized, and secondary structure elements are indicated above the alignment. (b) Regions of sequence divergence are plotted on the RRM structure. On the left, residues that have diverged between URB and U1A but not U2B* are colored blue, residues that have diverged between URB and U2B* but not U1A are colored red, and residues that have diverged in both U1A and U2B* are colored black. On the right, residues that have diverged in Drosophila SNF are colored black. Residues conserved between all proteins are shown in purple.

---

**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 3), as well as consensus sequences broken up by clade (Supplementary Fig. 4), are available in the supplementary material.

---

**Fig. 4.** The RRMs of human U1A, human U2B*, and Drosophila SNF have remained highly conserved with those of their Urbilaterian ancestor. (a) Alignments of the ML ancestral protein (URB) and SNF with both human U1A and U2B* are shown. Residues are highlighted in red if shared with human U1A only and in blue if shared with U2B* only. RNP motifs are italicized, and secondary structure elements are indicated above the alignment. (b) Regions of sequence divergence are plotted on the RRM structure. On the left, residues that have diverged between URB and U1A but not U2B* are colored blue, residues that have diverged between URB and U2B* but not U1A are colored red, and residues that have diverged in both U1A and U2B* are colored black. On the right, residues that have diverged in Drosophila SNF are colored black. Residues conserved between all proteins are shown in purple.
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 [$\Delta G^{\circ}(\text{folding}) = -9.4$ kcal/mol] [16] and that of SNF RRM1 [$\Delta G^{\circ}(\text{folding}) = -3.5$ kcal/mol] [5]. The notion that stability is an evolutionarily neutral trait as long as proper folding can be maintained [17] seems plausible for this protein family.

$^1$H/$^15$N heteronuclear single quantum coherence (HSQC) NMR spectra of FL URB, RRM1, and RRM2 are overlaid in Fig. 7a. The $^1$H/$^15$N HSQC amide spectra of the two independent RRMs can clearly be identified in the context of the FL protein. Assignments of the RRM1 cross-peaks are indicated on the spectrum of Supplementary Fig. 5. Amide resonances for the individual RRMs are well dispersed, indicating structured, folded domains.

Table 1. RNA binding of FL URB and URB RRM1

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

Please cite this article as: Williams, S. G. et al, Resurrection of an Urbilaterian U1A/U2B*/SNF Protein, J. Mol. Biol. (2013), http://dx.doi.org/10.1016/j.jmb.2013.05.031
The difference in the binding free energy \[ \Delta \Delta G^0 = \Delta G^0_{\text{SLIV}} - \Delta G^0_{\text{SLII}} \] 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 2. Protein family specificity for SLII/SLIV

<table>
<thead>
<tr>
<th></th>
<th>( \Delta \Delta G^0 ) (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>&gt;6</td>
</tr>
<tr>
<td>U2B</td>
<td>0.1 ± 0.3</td>
</tr>
</tbody>
</table>

The difference in the binding free energy \( \Delta \Delta G^0 = \Delta G^0_{\text{SLIV}} - \Delta G^0_{\text{SLII}} \) 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.

The difference in the binding free energy \( \Delta \Delta G^0 = \Delta G^0_{\text{SLIV}} - \Delta G^0_{\text{SLII}} \) 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></th>
<th>( \Delta \Delta \beta_{\text{RRM1}} ) (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>0.6 ± 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>

The difference in the binding free energy \( \Delta \Delta G^0 = \Delta G^0_{\text{SLIV}} - \Delta G^0_{\text{SLII}} \) 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.

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″. Under the previous framework

![Graphs showing the salt dependence of URB and SNF binding to SLII and SLIV.](image)

**Fig. 5.** Salt dependence of URB and SNF binding to SLII and SLIV. Protein–RNA pairs are indicated in the panels. Black data points and lines are for the FL protein, and blue data points and lines are for experiments performed with RRM1. All experiments were performed in 10 mM cacodylate and 2 mM MgCl2 (pH 7, 22 °C). The salt dependence for FL SNF was previously reported [5] and is shown for comparison. Slopes of the lines are interpreted in terms of net ions released and are tabulated in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>FL Protein</th>
<th>RRM1</th>
<th>ΔΔ (Net ions released)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1A-SLII</td>
<td>−5.7 ± 0.2</td>
<td>−6.7 ± 1.1</td>
<td>2.5 ± 0.6</td>
</tr>
<tr>
<td>SNF-SLII</td>
<td>−4.0 ± 0.2</td>
<td>−3.1 ± 0.5</td>
<td>1.4 ± 0.4</td>
</tr>
<tr>
<td>SNF-SLIV</td>
<td>−4.2 ± 0.3</td>
<td>−2.7 ± 0.4</td>
<td>1.0 ± 0.4</td>
</tr>
<tr>
<td>URB-SLII</td>
<td>−3.8 ± 0.2</td>
<td>−3.2 ± 0.3</td>
<td>−0.04 ± 0.23</td>
</tr>
<tr>
<td>URB-SLIV</td>
<td>−5.2 ± 0.3</td>
<td>−3.9 ± 0.1</td>
<td>1.4 ± 0.4</td>
</tr>
<tr>
<td>U2B″-SLII</td>
<td>−4.6 ± 0.3</td>
<td>−4.4 ± 0.2</td>
<td>2.0 ± 0.2</td>
</tr>
<tr>
<td>U2B″-SLIV</td>
<td>−5.2 ± 0.3</td>
<td>−3.9 ± 0.1</td>
<td>1.3 ± 0.3</td>
</tr>
</tbody>
</table>

The slope of the ln(K_\text{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]. ΔΔ 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.

---

Please cite this article as: Williams, S. G. et al, Resurrection of an Urbilaterian U1A/U2B″/SNF Protein, J. Mol. Biol. (2013), http://dx.doi.org/10.1016/j.jmb.2013.05.031
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 Urbilaterian 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 RNA binding surface Loop 3

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 cm² 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 lophotrochoan 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 lophotrochoan clusters show some indication of functional divergence, but this is much less pervasive than in jawed vertebrates. The functional distance between the paralogous lophotrochoan clusters is smaller, indicating a substantial degree of functional similarity between these proteins. The functional distance between the paralogous lophotrochoan 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 $\nu_{\text{CPMG}} = 50$ Hz and $\nu_{\text{CPMG}} = 1000$ 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 distinct. 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 RRMs—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 possibility that, in a 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 analysis, the LG model of amino acid substitution was used, along with equilibrium amino
induced with 1 mM IPTG for 4 h at 25 °C, spun down, and Drosophila SNF proteins [5].

Ancestral U1A/U2B′ synthesis

CodeML in PAML [13,42] was used for inference of the ancestral amino acid sequences of the U1A/U2B′/SNF protein family. The ancestral sequences are inferred from the phylogenetic and 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 ancestor of modern bilaterians was obtained, and this tree reconstruction. The ML tree obtained with PhyML was similar evolutionary model parameters as those used for methods. A marginal reconstruction was performed using the phylogeny and the modern sequences using ML.

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 μM. CD spectra of each purified RRM (in 0 M urea) are consistent with a canonical αβ composition.

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 15N,NH4Cl or 15NH4Cl and 13C 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 μM. 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 1H/15N and 13Cα 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 \(^{1}H^{15}N\) HSQC spectra were acquired at 22.5 °C, 700 MHz; spectral widths were 7300 Hz in the \(^{1}H\) dimension and 2000 Hz in \(^{15}N\).

For \(^{15}N–^{1}H\) NOE measurements, duplicate pairs of NOE spectra were collected for SNF and U1A with and without a 3-s \(^{1}H\) 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 from duplicate data sets were averaged.

 Backbone \(^{15}N\) millisecond-to-microsecond exchange in SNF, U1A, and URB RRM1 was probed using CPMG experiments [45]. Spectra were recorded at 22.5 °C on a 700-MHz spectrometer, with 32 scans and a 2.5-s recycle delay. For all samples, a reference spectrum was acquired, as well as spectra with CPMG field strengths of 50 Hz and 1000 Hz. The total CPMG block was 40 ms. The \(\Delta R_{2,\text{eff}} = \Delta R_{2,\text{eff},50 \text{ Hz}} - \Delta R_{2,\text{eff},1000 \text{ Hz}}\) was calculated as:

\[
\Delta R_{2,\text{eff}} = \frac{-\ln\left(\frac{I_{\nu=50 \text{ Hz}}}{I_{\text{ref}}}\right)-\ln\left(\frac{I_{\nu=1000 \text{ Hz}}}{I_{\text{ref}}}\right)}{0.04}
\]

where \(I\) indicates the peak intensity from the reference spectrum (\(I_{\text{ref}}\)) or the 50-Hz or 1000-Hz CPMG field strength spectra (\(I_{\nu = 50 \text{ Hz}}, I_{\nu = 1000 \text{ Hz}}\)). \(\Delta R_{2,\text{eff}}\) is a function of the chemical shift difference among the states, the exchange rate, and the population of the states. Uncertainties were estimated from the baseline noise.

Uncertainty in URB reconstruction

The marginal reconstruction calculates posterior probabilities for each amino acid at each site. This allows assessment of the protein sequence for sites that are ambiguously determined. There was a single site in RRM1 for which a second amino acid had a posterior probability (PP) 0.2 < PP < 0.8 (Supplementary Table 3). Site-directed mutagenesis (QuikChange, Agilent) was used to introduce the URB RRM1 A84S mutation in order to assess the functional implication of the uncertainty in the reconstruction. The "mutant" protein was expressed and purified identically with its wild-type counterpart.

Functional divergence analysis

The premise of Gu’s functional divergence analysis is that functional change is highly correlated with changes in evolutionary rate [46,47]. After a major evolutionary event, such as a gene duplication, the evolutionary rate of one of the genes may increase at sites that are responsible for functional divergence. Tests for Type I functional divergence detect whether or not there are significant differences in the evolutionary rates between two clusters or branches of the tree. Type II functional divergence characterizes the situation in which a change in character (amino acid) occurred early after the gene duplication, but subsequent evolutionary rates between the paralogous clusters are similar. This is also called “constant but different” or “cluster specific” divergence [48].

DIVERGE compares monophyletic clusters within the tree to determine if there is a significant difference in evolutionary rates between sites and between clusters. It estimates a coefficient of functional divergence (\(\Theta\)) between two clusters. \(\Theta\) varies between 0 and 1, where 0 indicates no functional divergence, and larger values of \(\Theta\) indicate a larger degree of divergence. The statistical test for functional divergence is to determine whether \(\Theta\) is significantly different from 0. For the functional distance analysis, \(\Theta\) 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 consensus sequences

Curated U1 and U2 snRNA sequences from multiple organisms are available through the former uRNADB† and Rfam‡. Metazoan sequences from these databases were obtained and used as BLAST inputs to obtain additional snRNA sequences. From the snRNA sequences and fragments obtained following BLAST searches, stem–loop II of U1 and stem–loop IV of U2 were aligned. These sequences are shown in Supplementary Table 10b. Consensus sequence figures were made with WebLogo [49,50].

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 [\(\alpha\text{-}^{32}\text{P}\)]UTP and [\(\alpha\text{-}^{32}\text{P}\)]CTP. The transcription products were gel-purified. The different RNAs were as follows:

- U1 SLI: 5′-GGAGACCAUUUGCAGUCGGUUCUC
- U2 SLI: 5′-GGCCGCAUUGGACAGUCGGCAGGCGGUCC
- CE SLI: 5′-GGCGCCGCAUUGCACUUGCGCCGCG
- CE SLIV: 5′-GGCGCCGCAUUGCAGUCGGCAGGCG

The underlined loop sequences for U1 SLII and U2 SLIV correspond to sequences from humans (SLII) and Drosophila (SLIV). URB binding to RNAs with the Drosophila loop of SLII (AUGACCCUC) 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 [52]. Structures for URB were templated from 1FHT (a U1A RRM1 solution structure) [44], 1A9N (a cocrystal structure that includes U2B* RRM1) [53], 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. Award numbers R01 GM096444 to K.B.H., F31 GM089576 to S.G.W., and F32 GM090650 to M.J.H. were 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](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

**Keywords:**

RRM;
ancestral reconstruction;
RNA binding;
protein phylogeny;
U1A protein

‡ 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


