Published in final edited form as:

J Mol Biol. 2015 September 25; 427(19): 2997–3000. doi:10.1016/j.jmb.2015.06.009.

Dividing and Conquering the RNA Recognition Motif Family: A Representative Case Based on hnRNP L

Sarah Loerch and Clara L. Kielkopf

Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642

The RNA recognition motif (RRM) is a prevalent RNA-binding fold among proteins that are required for RNA biogenesis, processing, transport, and degradation. Compared to approximately 6000 RRM s that were annotated ten years ago¹, the Protein Families (Pfam) database² currently contains >50,000 RRM s, primarily among eukaryotic proteins (more than 30,000 RRM s), but also among bacterial, and in a few cases, viral proteins. These RRM s typically are multiples of one to six copies and can be combined with other classes of RNA binding domains in a given polypeptide. To date, experimental structures are available for <1% of annotated RRM s. In recent years it has become apparent that the RRM family is far more diverse in terms of its modes of ligand recognition and the types of ligand recognized than was assumed initially. A growing need has emerged for the sub-classification of the large RRM family into groups of distinct structure and function. To date, few primary sequence motifs have been defined for the characterization of such specialized RRM variants. In this issue of the *Journal of Molecular Biology*³, Allain and colleagues set a precedent for identifying the consensus motifs of an RRM sub-class, which in combination with the predicted secondary structures of conserved extensions of the core RRM fold can be leveraged to successfully predict RRM structure-function relationships based on primary sequence features.

Historically, the discovery of RRM s has been driven by the identification of characteristic sequence motifs. A conserved eight-residue motif, called ribonucleoprotein (RNP)1, was first noted in the primary sequences of the poly(A) binding protein and hnRNP A1⁴, and later was extended to the hnRNP C protein⁵. Several RNP1 motifs were repeated within a given polypeptide, and each RNP1 was embedded within a conserved region of approximately 90 residues that in retrospect corresponded to one of multiple consecutive RRM s. As sequences for other RNA binding proteins became available, a second, six-residue consensus motif, termed RNP2, was noted approximately 30 residues N-terminal to the RNP1⁶. These RNP1 and RNP2 consensus sequences continue to serve as a major template search pattern for the classification of RRM family members.

Shortly thereafter, the three-dimensional structure of the N-terminal RRM in U1 small nuclear ribonucleoprotein associated (U1A) protein⁷, revealed that the RNP1 and RNP2 reside on two central β -strands of an antiparallel four-stranded β -sheet backed by two α -

helices ($\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ -topology) (Fig. 1A). A subsequent structure of RNA-bound U1A further demonstrated that aromatic, polar and basic residues of the RNP1 and RNP2 motifs form stacking, hydrogen bonds, and electrostatic interactions with the RNA bases and backbone⁸. Additional RNA contacts by a C-terminal α -helix of the U1A protein raised an early clue that regions outside the core RRM fold can contribute to RNA-binding.

In light of conserved residues that were integral to the U1A RRM fold, Keene and colleagues⁹ and in a subsequent comprehensive study, Krainer and colleagues¹⁰ surveyed then-available sequences for putative RRM s that were independent of RNP1/RNP2 conservation. Beyond two N-terminal RRM s of hnRNP L (RRM1 and RRM2) that were identified originally based on weak sequence matches of the RNP1/RNP2 consensus¹¹, the Keene and Krainer groups respectively identified third⁹ and fourth¹⁰ RRM s of the hnRNP L protein through inclusion of the structural context in the search algorithm. These seminal studies set a precedent for defining the RRM in terms of the three-dimensional topology as opposed to key RNA binding residues, yet the degenerate RNP1/RNP2 motifs already hinted that the RRM family would comprise diversified sub-classes with specialized functional motifs.

More recently, a variety of potential RRM sub-classes have emerged from the expanding database of RRM-containing structures (>400 RRM-containing depositions in the PDB). Consequently, the canonical RNP-centric view of the RRM has shifted to an RRM scaffold that can be modified and extended to meet different ligand requirements. In some cases, RRM variants even bind proteins as opposed to RNA (for example,^{12; 13; 14}). The diverse quasi-(q)RRM s¹⁵ and pseudo-(ψ)RRM s¹⁶ respectively bind RNA via loop and α -helical regions of the RRM fold as opposed to the RNP1/RNP2 surface. Other sub-classes, such as the cryptic (x)RRM s in genuine La proteins [reviewed in¹⁷] and occluded (o)RRM s in the *S. cerevisiae* essential splicing factor Prp24¹⁸, utilize α -helical extensions of the core RRM fold to mediate atypical interactions with double-stranded RNA. As such, consensus motifs for the identification of such auxiliary secondary structural features – or “sub-structural” elements such as those defined here for hnRNP L would offer valuable insight into the structures and functions of uncharacterized proteins.

Here, Allain and colleagues³ rigorously analyze three sub-structural elements that they find to be present in different combinations in each of the four hnRNP L RRM s (representative features shown in Fig. 1A). The authors confirm that all four linked RRM s (numbered 1–4 from N- to C-terminus) and an adjoining N-terminal region are required for hnRNP L to recognize CA-rich RNAs, as well as to regulate alternative splicing *in vitro*. The authors complete a structural portrayal of the hnRNP L RRM s by adding a new NMR structure of RRM2, an NMR structure of RRM1, NMR and crystal structures of the linked RRM3-RRM4 to prior crystal structures of the hnRNP L RRM1 and RRM34¹⁹. The authors note distinctive structural variations of the hnRNP L RRM s. Based on sequence and structural comparisons, they defined a family of five proteins (hnRNP L, hnRNP LL, Polypyrimidine Tract Binding protein (PTB), nPTB, and Regulator of Differentiation (Rod)1) that can be assumed to share these structural features. The authors leveraged the hnRNP L/PTB family to compile consensus motifs for key sub-structure elements and utilized these in searches for previously unidentified matches.

All of the hnRNP L RRM s share C-terminal “interacting coils”, a C-terminal extension that lacks regular secondary structure and interacts in a consistent perpendicular orientation across the β -sheet surface. The corresponding region of the PTB structure contacts the bound RNA, suggesting a broader role for “interacting coils” in RNA binding by other family members. A search of the Protein Data Bank (PDB) identified five additional RRM structures that exhibit “interacting coil” extensions of the core RRM fold (single RRM s from hnRNP H2, RNPS1, and ABH8, and two RRM s from Matrin-3). Based on alignment of these five RRM s plus 20 RRM s from the hnRNP L and PTB family members (which each contain four RRM s), Allain and colleagues define an interacting C-terminal coil (ICC) consensus motif.

Further searches of the PDB with the refined ICC successfully identified a sixth matching RRM outside the hnRNP L/PTB family from Rna15 (Fig. 1B), a factor that recognizes a 3'-end mRNA processing element in yeast. Based on the original structure of Rna15 in the absence of RNA²⁰, it was anticipated that this C-terminal coil region would interfere with RNA binding. Nevertheless, the RNA-bound structure of Rna15 revealed that the ICC traverses the β -sheet surface and contacts the bound RNA²¹ in an analogous manner as PTB²². This proof-of-principle finding could have been predicted in the absence of an Rna15 structure had the ICC been defined at the time. A few “false” hits were obtained in searches based on the consensus ICC sequence alone. These off-target matches corresponded to C-terminal α -helices that often antagonize, rather than assist, RNA binding by RRM s. However, the implementation of secondary structure predictions to distinguish sequences with α -helical propensities from extended coils such as the ICC strengthened such searches.

Elongations of the ICC added fifth anti-parallel β -strands to the typical four β -strands of the core RRM fold in the second and third RRM s of hnRNP L, which Allain and colleagues term “vRRM” s. Analogous fifth β -strands are apparent in the structures of the second and third RRM s of PTB. Alignments of these hnRNP L and PTB vRRM s alongside the corresponding regions of the hnRNP LL, nPTB, and Rod1 members of the family was used to generate a five-stranded RRM consensus motif (VRC), which adds one hydrophobic residue that interacts with the first α -helix of the RRM to the minimal ICC. Sequence searches recovered one additional vRRM from the Q4DY42 protein. However clustering analysis indicated that the Q4DY42 protein was closely related to hnRNP L and PTB, and as such not considered a bona fide “new” family member.

In a final defining feature, the third and fourth hnRNP L RRM s pack back-to-back *via* a buried hydrophobic surface as observed previously for the corresponding RRM s of PTB, which further had been shown to assist binding to distally separated RNA sites^{19; 22}. Based on comparisons among hnRNP L/PTB family members, Allain and colleagues define an “interacting RRM consensus” (IRC) underlying the back-to-back inter-RRM arrangement. However, no new IRC-containing proteins were identified in sequence searches, suggesting that this back-to-back mode of tandem RRM binding to looped RNAs, as well as the five-stranded β -sheet, is limited to the hnRNP L and PTB paralogues.

Based on consensus motif analysis combined with phylogenetic clustering, Allain and colleagues reason that an ancestral RRM might have acquired the ICC motif first and then diverged into two descendant RRM – one that acquired the additional VRC, and one that did not. Whether the newly identified ICC-containing proteins, some of which do not cluster with the hnRNP L/PTB family, diverged from the ICC of an ancestral RRM or are a case of convergent evolution remains an open question. As noted by the authors, the VRC of hnRNP L is regulated by acetylation²³, which highlights the possibility that such RRM extensions evolved as handles for post-translational modifications that regulate RNA binding. Altogether, this study represents a prototype for the systematic derivation of “sub-structures” from a growing body of RRM structures that in the longer term, could offer an approach to “divide and conquer” the oversized RRM family into practical sub-classes related by modes of ligand binding.

Acknowledgements

C.L.K. is supported by grant R01 GM070503 from the National Institutes of Health.

References

1. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* 2005; 272:2118–31. [PubMed: 15853797]
2. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. *Nucleic Acids Res.* 2014; 42:D222–30. [PubMed: 24288371]
3. Blatter M, Dunin-Horkawicz S, Grishina I, Maris C, Thore S, Maier T, Bindereif A, Bujnicki JM, Allain FH. The signature of the five-stranded vRRM fold defined by functional, structural and computational analysis of the hnRNP L protein. *J Mol Biol.* 2015 this issue.
4. Adam SA, Nakagawa T, Swanson MS, Woodruff TK, Dreyfuss G. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol Cell Biol.* 1986; 6:2932–43. [PubMed: 3537727]
5. Swanson MS, Nakagawa TY, LeVan K, Dreyfuss G. Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. *Mol Cell Biol.* 1987; 7:1731–9. [PubMed: 3110598]
6. Dreyfuss G, Swanson MS, Pinol-Roma S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci.* 1988; 13:86–91. [PubMed: 3072706]
7. Nagai K, Oubridge C, Jessen TH, Li J, Evans PR. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature.* 1990; 348:515–20. [PubMed: 2147232]
8. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature.* 1994; 372:432–8. [PubMed: 7984237]
9. Kenan DJ, Query CC, Keene JD. RNA recognition: towards identifying determinants of specificity. *Trends Biochem Sci.* 1991; 16:214–20. [PubMed: 1716386]
10. Birney E, Kumar S, Krainer AR. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.* 1993; 21:5803–16. [PubMed: 8290338]
11. Pinol-Roma S, Swanson MS, Gall JG, Dreyfuss G. A novel heterogeneous nuclear RNP protein with a unique distribution on nascent transcripts. *J Cell Biol.* 1989; 109:2575–87. [PubMed: 2687284]
12. Kielkopf CL, Lucke S, Green MR. U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.* 2004; 18:1513–26. [PubMed: 15231733]

13. Kadlec J, Izaurralde E, Cusack S. The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nat Struct Mol Biol.* 2004; 11:330–7. [PubMed: 15004547]
14. Bono F, Ebert J, Lorentzen E, Conti E. The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA. *Cell.* 2006; 126:713–25. [PubMed: 16923391]
15. Dominguez C, Fiset JF, Chabot B, Allain FH. Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nat Struct Mol Biol.* 2010; 17:853–61. [PubMed: 20526337]
16. Clery A, Sinha R, Anczukow O, Corrionero A, Moursy A, Daubner GM, Valcarcel J, Krainer AR, Allain FH. Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. *Proc Natl Acad Sci U S A.* 2013; 110:E2802–11. [PubMed: 23836656]
17. Singh M, Choi CP, Feigon J. xRRM: a new class of RRM found in the telomerase La family protein p65. *RNA Biol.* 2013; 10:353–9. [PubMed: 23328630]
18. Montemayor EJ, Curran EC, Liao HH, Andrews KL, Treba CN, Butcher SE, Brow DA. Core structure of the U6 small nuclear ribonucleoprotein at 1.7-Å resolution. *Nat Struct Mol Biol.* 2014; 21:544–51. [PubMed: 24837192]
19. Zhang W, Zeng F, Liu Y, Zhao Y, Lv H, Niu L, Teng M, Li X. Crystal structures and RNA-binding properties of the RNA recognition motifs of heterogeneous nuclear ribonucleoprotein L: insights into its roles in alternative splicing regulation. *J Biol Chem.* 2013; 288:22636–49. [PubMed: 23782695]
20. Pancevac C, Goldstone DC, Ramos A, Taylor IA. Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors. *Nucleic Acids Res.* 2010; 38:3119–32. [PubMed: 20097654]
21. Leeper TC, Qu X, Lu C, Moore C, Varani G. Novel protein-protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and Hrp1. *J Mol Biol.* 2010; 401:334–49. [PubMed: 20600122]
22. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, Allain FH. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science.* 2005; 309:2054–7. [PubMed: 16179478]
23. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science.* 2009; 325:834–40. [PubMed: 19608861]

Fig. 1. Structures of (a) a representative “vRRM” (third RRM, blue) of hnRNP L determined in the issue and (b) RNA-bound (violet) RRM of Rna15 (blue) with a C-terminal “interacting coil” that was identified based on sequence matches with the “Interacting C-terminal Coil” consensus motif. The fifth β -strand (magenta) and “interacting coil” (red) extensions as well as the canonical RNP2/RNP1 motifs (yellow) are labeled.