Multiple RNA binding domains (RBDs) just don't add up

Yousif Shamoo*, Norzehan Abdul-Manan and Kenneth R. Williams[1]

Department of Molecular Biophysics and Biochemistry and [1]Howard Hughes Medical Institute, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510-8024, USA

Received September 30, 1994; Revised and Accepted January 5, 1995

ABSTRACT

One of the most common motifs for binding RNA in eukaryotes is the RNA binding domain (RBD) or RNA Recognition Motif (RRM). One of the more intriguing aspects of these proteins is their modular nature. Proteins have been found containing from one to four RRMs. In most instances, these domains have some basal level of non-sequence specific RNA binding affinity. In addition, many also have a higher affinity for a specific structure or sequence of RNA. In the cases of heterogenous nuclear ribonucleoprotein A1 (hnRNP A1), yeast poly-A binding protein and splicing factor U2AF^{65}, the individual free energy of binding of the RBDs for RNA are not strictly additive. By invoking a model in which the amino acids connecting adjoining RBDs are considered to be flexible linkers with an interresidue spacing of about 3.5 Å, it is possible to predict the apparent association constants for at least some multi-RBD proteins to single-stranded RNA. We have surveyed the literature and found that individual RBDs are separated by 'linker' sequences of highly variable length. These linkers provide a critical determinant of binding affinity and may modulate *cis* versus *trans* binding. A clearer understanding of multi-RBD binding is essential to critically evaluating the role of these proteins in RNA splicing, packaging and transport.

How do proteins having more than one RNA binding domain (RBD) (1) bind ssRNA? The RBD or RRM (RNA Recognition Motif) (2) is a ~90 residue domain that contains two conserved 'RNP' consensus sequences that are usually separated by 25–35 residues (1,3). More than 100 RBDs have been identified in over 30 proteins involved in pre-mRNA splicing, rRNA processing, RNA packaging and transport (1,4). Although most RRMs have been assumed to have a basal level of non-specific binding affinity for RNA (5), only in a relatively few instances has this been proven. Hence, in this report the term RBD will generally be restricted to those RRM domains that have been shown to bind RNA.

Many RBDs also have high affinity for specific RNA sequences such as yeast poly-A binding protein (6), splicing factor U2AF for the poly-pyrimidine tract of pre-mRNA (7), and snRNP U1A binding for U1 RNA (8). Each RBD structure that has been solved has the same β_1-α_1-β_2-β_3-α_2-β_4 folding pattern in which a four stranded β-sheet is backed by a pair of helices (9–12). Since proteins contain from one to four RBDs, an intriguing question is what is the function of multiple RBDs (see Fig. 1)? In many cases, individual RBDs within the same protein have different binding specificities which suggests they may allow a single protein to bridge multiple RNAs (*trans*) (8), whereas in others, such as the hnRNP A1 protein, multiple RBDs interact with 'non-specific' RNA lattice to increase binding affinity (*cis*) (5).

While the binding affinity of a multiple RBD containing protein might be expected to be the product of the affinities of its isolated RBDs (which assumes the free energy of binding of RBDs is additive), this was not the case with hnRNP A1. The affinity of an A1 construct containing both its RBDs was nearly 1000-fold less than the product of the affinities of its isolated RBDs (5). This result arises from the flexible 17-residue linker joining the two A1 RBDs that allows a 'free' A1 RBD (linked to a bound RBD) to sweep out a sphere with a radius as large as 90 Å. *In extremis*, the affinity of a protein containing two identical RBDs connected by an infinitely long and flexible linker for ssRNA is simply twice the individual affinity of one domain, whereas if the two domains were superimposed, the overall association constant would be the product of the individual RBDs. This phenomena is illustrated in Table 1 where the dimensions of the U1A RBD have been assumed to be representative of a typical RBD (9). For this example we have used an affinity of 10^5 M^{-1} which is in between the ~10^4 M^{-1} and 10^6 M^{-1} affinities observed for the A1 (5) and type C (13) RBDs and assumed the linkers are extended to give an inter-residue distance of about 3.5 Å. NMR studies have clearly shown that RBDs from U1A (14), hnRNP C (10) and hnRNP A1 (11) do not aggregate, and therefore the model does not consider RBD dimerization. These assumptions produce a hypothetical set of affinities for a two RBD protein connected by varying length linkers. As seen in Table 1, even when the linker is as short as 10 amino acids, the predicted affinities for the two domain protein are nearly 350-fold less than the product of the individual RBD affinities. In order to fully segregate the two RBD affinities from one another, the two domains must be separated by a rather long linker, somewhere in the order of 60 residues. Under these conditions, the overall affinity of the hypothetical two RBD

* To whom correspondence should be addressed

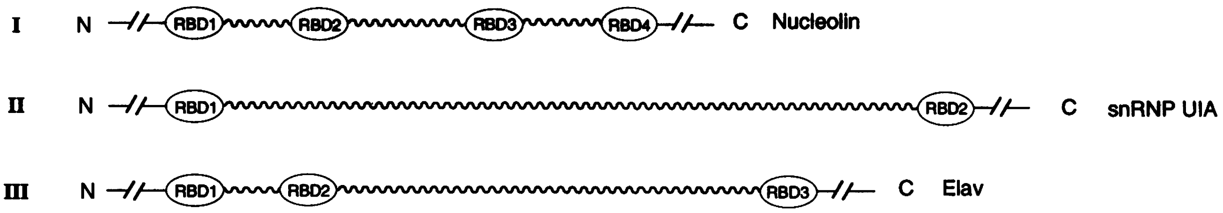

Figure 1. General model for RBD organization. Based on the data in Table 2, there appear to be three general arrangements of multiple RBDs. Nucleolin, yeast poly-A binding protein, hnRNP A1 and similar proteins have closely spaced tandem arrays of two to four RBDs connected by relatively short linkers (4,5). The relatively short linker length in these proteins increases the probability of adjacent, *cis* binding by two or more RBDs which, assuming the target is sufficiently long to accommodate the binding of multiple RBDs, would be expected to increase the overall affinity by at least 5-fold (eg., assuming the protein contains two RBDs with individual affinities of 10^4 M^{-1} and that they are separated by a linker of 30 amino acids). In contrast, proteins like snRNP U1A and U2B″ have a single RBD with very high affinity for a specific RNA target and contain two or more RBDs connected by a long stretch of more than 60 residues (24,25). Individual RBDs in these proteins might be expected to be more likely to either span different RNA species (ie., *trans* binding) or to interact with well separated *cis* targets. U2AF^{65} and ELAV, appear to represent a hybrid of these two linker types since they contain two closely spaced RBDs separated from a third and/or fourth RBD by linkers of variable lengths.

protein (7.7×10^5 M^{-1}) described in Table 1 is less than 10-fold greater than that of the corresponding single RBD containing protein (10^5 M^{-1}). As a consequence of the 17 amino acid linker in A1, after the first RBD is bound, the concentration of the 'second or free' A1 RBD in the sphere swept out by the second RBD is only about 550 μM, nearly 2000-fold less than the 1 M concentration required for the free energies of binding of the individual RBDs to be additive. Assuming the linker is a fully extended random coil, the 'effective' affinity resulting from binding of the second A1 RBD (K_2') can be predicted by the following relationship (15):

$$\text{(I)}\quad K_2' = 3V(K_2)/4\pi(r^3)N$$

K_2 is the affinity that would be observed for the second A1 RBD had it bound first, 'r' is the mean free radius linking the two RBDs and N is the number of particles per volume (V) in the standard state (ie., 1 M) defined for the binding of the first RBD (K_1). The overriding importance of the flexible linker was demonstrated by studies (5) on an A1 hnRNP construct that contained two RBDs with individual affinities of 1.5×10^4 M^{-1} and 4.5×10^4 M^{-1}. In this instance, the second A1 RBD to bind had a calculated K_2' of only 8.2 M^{-1}. Using this value, the calculated K_{app} of this A1 construct was 8.3×10^5 M^{-1}, which was close to the measured affinity of 5.4×10^5 M^{-1}.

Before extending these findings to other multiple RBD-containing proteins, we surveyed the literature and found that RRM linkers vary from four to as many as 122 residues (Table 1). While the A1 linker is highly flexible (19), the degree of flexibility of other RRM linker sequences is not known. However, since the sequences in Table 1 contain only about half the average content of several hydrophobic amino acids (eg., isoleucine, leucine, phenylalanine, tryptophan and valine) and 2.4-fold the average content of proline (based on the composition of an 'average' protein in the PIR Database), linker regions are probably less structured and more solvent exposed than an 'average' sequence. In fact, using the method of Karplus and Schultz (20) the average predicted flexibility of the sequences in Table 1 is about equal to that of the A1 linker. Although flexibility predictions suggest many of the shorter linker regions in Table 1 are entirely flexible, some of the longer linkers are predicted to include some relatively rigid domains. In fact, some of the linkers such as those found in nucleolin RBDs 1 to 2 and YPAB 1 to 2 are very rich in lysines and arginines and may directly participate in RNA binding. If so, one could well imagine the linker acting to 'thread' the RNA to the adjoining RBD. This, however, would not alter the conclusion that RBD linker length is an important determinant of binding affinity. To illustrate this, Table 1 demonstrates that even a 10-residue linker decreases the predicted affinity for a multi-site ligand by >300-fold. Based on these data, any study directed at quantitating the contribution of individual RBDs to the overall free energy of binding must take into account the effect of the linker.

Table 1. Effect of linker length on the predicted binding affinity of a protein with two RBDs with individual affinities of 10^5 M^{-1}

Linker length	Distance[a] (Å)	K_2'[b] (M^{-1})	K_{app} (M^{-1})
(#Res.)			(Overall Affinity[c])
0	30	1500	2.9×10^8
10[d]	65	150	2.9×10^7
20	100	40	8.2×10^6
30	135	16	3.4×10^6
40	170	8.1	1.8×10^6
60	240	2.9	7.7×10^5
80	310	1.3	4.7×10^5
100	380	0.72	3.4×10^5
120	450	0.44	2.9×10^5
Infinite	Infinite	0	2.0×10^5

[a] Approximate distance between the centers of two neighboring RBDs. This calculation assumes the linker sequence is a random coil with a residue distance of 3.5 Å and that each RBD has a radius of ~15 Å.

[b] Calculated affinity for the subsequent binding of the second RBD of a hypothetical two RBD protein to a nucleic acid ligand that is already bound by the first RBD of the same protein. The affinities were calculated based on equation (I) and by assuming that each isolated RBD has an affinity of 10^5 M^{-1} for this nucleic acid ligand.

[c] The overall apparent affinity was calculated based on the following equation which was derived by Crothers and Metzger (15): $K_{app} = 2[K_1(1 + K_2')]$. In this equation, the factor of two is for degeneracy arising from interchange of the two RBDs and K_1 and K_2 correspond to the effective affinities of the first and second RBD that bind the nucleic acid lattice.

[d] Clearly, as the linker decreases below the diameter of the RBDs, steric effects between the adjoining RBDs become a critical determinant in binding affinity.

Table 1. Selected linker regions from various RRM containing proteins

Protein	RRMs	Linker Length	Sequence
Bj6	1-2	4	RFAPNATILR
SXL	1-2	10	SYARPGGESIKDTNLY
ELAV	1-2	10	SFARPSSDAIKGANLY
HuD	1-2	10	SYARPSSASIRDANLY
NUCL	1-2	12	EKPKGRDSKKVRAARTLL
YPAB	1-2	12	NVDQRDPSLRKKGSGNIF
NUCL	3-4	14	ELQGPRGSPNARSQPSKTLF
hnRNP A2/B1	1-2	14	AVAREESGKPGAHVTVKKLF
hnRNP A1	1-2	17	PKRAVSREDSQRPGAHLTVKKIF
YPAB	2-3	18	APNLSRKERDSQLEETKAHYTNLY
hnRNP L	1-2	21	NYGTSQKISRPGDSDDSRSVNSVLLFT
NUCL	2-3	21	YYTGEKGQRQERTGKNSTWSGESKTLV
NSR1	1-2	23	ENSTSKPAGNNDRAKKFGDTPSEPSDTLF
chplst 33Kd	1-2	25	PEVPRGGEREVMSAKIRSTYQGFVDSPHKLY
YPAB	3-4	28	GRAQKKNERMHVLKKQYEAYRLEKMAKYQGVNLF
U2AF^{65}	1-2	31	NRPHDYQPLPGMSENPSVYVPGVVSTVVPDSAHKLFI
ASF/SF2	1-2	32	EFPRSGRGTGRGGGGGGGGGAPRGRYGPPSRRSENRVV
SRp55	1-2	43	EPARGSARGRNRDRYDDRYGGRRGGGGGRYNEKSSSRYGPPLRTEYRLI
hnRNP I	3-4	46	TLSKHQNVQLPREGQEDQGLTKDYGNSPLHRFKKPGSKNFQNIFPPSATLHL
hnRNP L	3-4	47	CVSKQPAIMPGQSYGLEDGSCSYKDFSESRNNRFSTPEQAAKNRIQHPSNVLH
hnRNP I	1-2	50	FSNHKELKTDSSPNQARAQAALQAVNSVQSGNLALAASAAAVDAGMAMAGQSPVLR
pPTB	1-2	53	QFSNHKELKTDSSPNQARAQAALQAVNSVQSGNLALAASAAAVDAGMAMAGQSPVLRII
U2AF^{65}	2-3	60	LVQRASVGAKNATLVSPPSTINQTPVTLQVPGLMSSQVQMGGHPTEVLCLMNMVLPEELLDDEEYE
snRNP U2B"	1-2	67	QYAKTDSDIISKMRGTFADKEKKKEKKKAKTVEQTATTTNKKPGQGTPNSANTQGNSTPNPQVPDYPPNYILF
ELAV	2-3	75	KFSNTPGSTSKIIQPQLPAFLNPQLVRRIGGAMHTPVNKGLARFSPMAGDMLDVMLPNGLGAAAAAATTLASGPGGAYPIF
pPTB	2-3	79	DFSKLTSLNVKYNNDKSRDYTRPDLPSGDSQPSLDQTMAAAFGLSVPNVHGALAPLAIPSAAAAAAAAGRIAIPGLAGAGNSVLL
HuD	2-3	87	KFANNPSQKSSQALLSQLYQSPNRRYPGPLHHQAQRFRLDNLLNMAYGVKRLMSGPVPPSACSPRFSPITIDGMTSLVGMNIPGHTGTGWCIF
hnRNP I	2-3	105	DFSKLTSLNVKYNNDKSRDYTRPDLPSGDSQPSLDQTMAAAFGAPGIISASPYAGAGFPPTFAIPQAAGLSVPNVHGALAPLAIPSAAAAAAAAGRIAIPGLAGAGNSVLL
hnRNP L	2-3	114	EYAKPTRLNVFKNDQDTWDYTNPNLSGQGDPGSNPNKRQRQPPLLGDHPAEYGGPHGGYHSHYHDEGYGPPPPHYEGRRMGPPVGGHRRGPSRYGPQYGHPPPPPPPPEYGPHADSPVLM
snRNP U1A	1-2	122	IQYAKTDSDIIAKMKGTFVERDRKREKRKPKSQETPATKKAVQGGGATPVVGAVQGPVPGMPPMTQAPRIMHHMPGQPPYMPPPGMIPPPGLAPGQIPPGAMPPQQLMPGQMPPAQPLSENPPNHILF

Wherever possible, linker lengths have been estimated by aligning the sequences of the RRMs according to Kenan *et al.* (4) where a multiple alignment algorithm was used that tends to align sequences based on secondary structure, placing gaps in loops. References for sequence alignments not found in Kenan *et al.* are indicated below: A1 RBD (5), U2AF^{65} (7), Bj6 (27), HuD (16) and SRp55 (17,18). Shaded residues denote the last three residues of β_4 and the first three residues of β_1 which thus establishes the boundaries of the RBD linkers.

Although there are only a few instances where the affinity of a multiple-RBD protein has been compared to that of the product of the affinities of its component RBDs, the U2AF^{65} splicing factor provides one example (7). In this instance, the affinities of deletion mutants containing RBD-1 and RBD-2,3 were 3.3×10^5 M^{-1} and 2.8×10^4 M^{-1} respectively for the MINX pre-mRNA (as estimated from the data in Fig. 4 of reference 7). Using a linker length of 31 residues (Table 1), the addition of RBD-1 to a bound RBD-2,3 should increase the affinity by 49-fold. This would lead to a predicted affinity for the RBD-1,2,3 construct of 2.8×10^6 M^{-1}, which is close to the observed value of 3.3×10^6 M^{-1} and which is about 3000-fold less than the affinity that might have been predicted from the product of the affinities of the U2AF^{65} RBD-1 and RBD2,3 constructs (i.e., $\sim 10^{10}$ M^{-1}).

Some caution however, needs to be used when applying equation I to non-equilibrium binding data. For example, although filter binding assays readily detect sequence specific binding proteins (which usually have $K_{app} > 10^9$ M^{-1}), they may not detect the short-lived complexes formed when some proteins bind non-specifically to nucleic acids (21). In the case of A1 RBD-1 binding to poly r(εA) (5), the K_{app} is only 3×10^4 M^{-1}. If this complex has a diffusion-limited association rate constant of about 10^8 M^{-1}, the corresponding half-life would be only 2×10^{-4} s. Such a complex would be expected to completely dissociate during the several seconds required to wash a nitrocellulose filter. Hence, if filter binding studies were used to determine the contribution of RBD-1 to the overall affinity of the intact A1 hnRNP [K_{app} for poly r(εA) $>10^8$ M^{-1} (5)] it would be erroneously concluded that RBD-1 makes no contribution to the overall A1 affinity for poly r(εA). The important point is that in a non-equilibrium binding assay, 'no observed binding' may actually mean a K_{app} of 10^4 M^{-1} or more.

One of the interesting ramifications of RBD linkers is their potential impact on *trans* versus *cis* binding. Using A1 as an example, once the first RBD is bound, the effective concentration of the adjacent RBD would be about 550 μM. If there is only one *cis* RNA site in the volume swept out by RBD-2 then A1 would have an equal probability of binding two identical RNA molecules in a *trans* as opposed to a *cis* fashion, when the free RNA (binding site) is 550 μM. Below this concentration, *cis* binding would be favored. Obviously, extending the linker to the 122 residues that are in U1A would promote *trans* binding. RBD-1 of U1A has, in fact, been shown to bind specifically to hairpin II in U1 snRNA (8), while Lutz and Alwine (23) have suggested that RBD-2 binds within the upstream efficiency element of the polyadenylation signal. Based on the available binding data, the 'breakpoint' between *cis/trans* binding might be near 60 residues. Hence both RBDs in ASF have been shown to bind in a *cis* manner (25,26) (linker length of 32), while the specific binding of U2B'' (linker length of 67) to U2 snRNA appears to be due entirely to RBD-1 (22,24). Finally, it should be mentioned that since many other proteins, such as some HMG proteins and zinc finger proteins, also contain multiple nucleic acid binding domains, the analysis described in this work is likely to be applicable to many other proteins beyond just those containing conserved RBDs.

REFERENCES

1 Burd, C.G. and Dreyfuss, G., (1994) *Science* **265**, 615–621.

2 Query, C.C., Bentley, R.C. and Keene, J.D. (1989) *Cell* **57**, 89–101,.

3 Adam, S.A., Nakagawa, T., Swanson, M.S., Woodruff, T.K., and Dreyfuss, G. (1986) *Mol. Cell. Biol.* **6**, 2932–2943.

4 Kenan, D.J., Query, C.C. and Keene, J.D. (1991) *Trends Biochem. Sci.* **16**, 214–220.

5 Shamoo, Y., Abdul-Manan, N., Patten, A.M., Crawford, J.K., Pellegrini, M.C. and Williams, K.R. (1994) *Biochemistry* **33**, 8272–8281.

6 Sachs, A.B., Davis, R.W. and Kornberg, R.D. (1987) *Mol. Cell. Biol.* **7**, 3268–3276.

7 Zamore, P.D., and Green, M.R. (1992) *Nature* **355**, 609–614.

8 Lutz-Freyermuth, C., Query, C.C. and Keene, J.D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6393–6397.

9 Nagai, K. Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) *Nature* **348**, 515–520.

10 Wittekind, M., Görlach, M., Friedrichs, M., Dreyfuss, G. and Mueller, L. (1992) *Biochemistry* **31**, 6254–6265.

11 Garrett, D.S., Lodi, P., Shamoo, Y., Williams, K.R., Clore, G.M., And Gronenborn, A.M. (1993) *Biochemistry* **33**, 2852–2858.

12 Jessen, T.H., Oubridge, C., Teo, C.H., Pritchard, C. and Nagai, K., (1991) *EMBO J.* **10**, 3447–3456.

13 Amrute, S.B., Abdul-Manan, Z., Pandey, V., Williams, K.R. and Modak, M.J. (1994) *Biochemistry*, **33**, 8282–8291.

14 Hoffman, D.W., Query, C.C., Golden, B.L., White, S.W. and Keene, J.D. (1988) *Proc. Natl. Acad. Sci. USA*, **88**, 2495–2499.

15 Crothers, D.M. and Metzger, H. (1972) *Immunochemistry* **9**, 341–357.

16 Szabo, A., Dalmau, J., Manley, G., Rosenfeld, M., Wong, E., Henson, J., Posner, J.B. and Furneaux, H.M. (1991) *Cell* **67**, 325–333.

17 Roth, M.B., Zahler, A.M. and Stolk, J.A. (1991) *J. Cell Biol.* **115**, 587–596.

18 Mayeda, A., Zahler, A.M., Krainer, A.R. and Roth, M.B. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1301–1304.

19 Casas-Finet, J.R., Karpel, R.L., Maki, A.H., Kumar, A. and Wilson, S.H. (1991) *J. Mol. Biol.* **221**, 693–709.

20 Karplus, P.A. and Schultz, G.E. (1985) *Naturwissenschaften*. **72**, 212–213.

21 Revzin, A. (1990) in *The Biology of Nonspecific DNA-Protein Interactions* (ed. A. Revzin) CRC Press, Boca Raton.

22 Scherly, D., Boelens, W., Dathan, N.A., van Venrooij, W.J. and Mattaj, I.W. (1990) *Nature*, **345**, 502–506.

23 Lutz, C.S. and Alwine, J.C. (1994) *Genes Dev.* **8**, 576–586.

24 Bentley, R.C. and Keene. J.D. (1991) *Mol. Cell. Biol.* **11**, 1829–1839.

25 Cáceres, J.F. and Krainer, A.R. (1993) *EMBO J.* **12**, 4715–4726.

26 Zuo, P. and Manley, L. (1993) *EMBO J.* **12**, 4727–4737.

27 von Besser, H., Schnabel, P., Wieland, C., Fritz, E., Stanewsky, R. and Saumweber, H. (1990) *Chromosoma* **100**, 37–47.