© 1993 Oxford University Press Nucleic Acids Research, 1993, Vol. 21, No. 25 5803–5816

Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors

Ewan Birney+, Sanjay Kumar§ and Adrian R.Krainer*
Cold Spring Harbor Laboratory, PO Box 100, 1 Bungtown Road, Cold Spring Harbor,
NY 11724-2208, USA

Received July 30, 1993; Revised and Accepted October 27, 1993

ABSTRACT

We present a systematic analysis of sequence motifs found in metazoan protein factors involved in constitutive pre-mRNA splicing and in alternative splicing regulation. Using profile analysis we constructed a database enriched in protein sequences containing one or more presumptive copies of the RNA-recognition motif (RRM). We provide an accurate alignment of RRMs and structure-based criteria for identifying new RRMs, including many that lack the prototype RNP-1 submotif. We present a comprehensive table of 125 sequences containing 252 RRMs, including 22 previously unreported RRMs in 17 proteins. The presence of a putative RRM in these proteins, which are implicated in a variety of cellular processes, strongly suggests that their function involves binding to RNA. Unreported homologies in the RRM-enriched database to the metazoan SR family of splicing factors are described for an Arg-rich human nuclear protein and two yeast proteins (*S. pombe* mei2 and *S. cerevisiae* Npl3). We have rigorously tested the phylogenetic relationships of a large sample of RRMs. This analysis indicates that the RRM is an ancient conserved region (ACR) that has diversified by duplication of genes and intragenic domains. Statistical analyses and classification of repeated Arg–Ser (RS) and RGG domains in various protein splicing factors are presented.

INTRODUCTION

Several metazoan protein factors have been identified that are required for pre-mRNA splicing *in vitro*, or that regulate the selection of alternative splice sites *in vivo* and/or *in vitro* (reviewed in 1,2). In addition, several polypeptides associated with all, or with specific snRNPs have been described, as have a number of RNA-binding proteins, termed hnRNPs, which are associated with nuclear pre-mRNAs (reviewed in 1–4). These proteins, and many others, are present in spliceosomes, although perhaps not in all spliceosomes. Some may participate directly in splicing whereas others may be involved for example in snRNP assembly and transport to the nucleus, or in mRNA export to the cytoplasm.

The generic or essential metazoan protein splicing factors for which complete amino acid sequences are available to date include human SF2/ASF (also known as SRp30a) (5,6), human and mouse U2AF^{65} (7,8), human and chicken SC35 (also known as PR264 or SRp30b) (9,10), human PSF (11), *Drosophila* SRp55 (a variant of which is known as B52) (12,13), mouse and human X16 (also known as SRp20) (14,15) and its probable *Drosophila* homolog, RBP1 (16). Partial peptide sequences are also available for human SRp40, SRp55, and SRp75 (15). With the exception of U2AF^{65} and PSF, all of the above proteins belong to the same protein family, and many of them, perhaps all, have similar functions in constitutive and alternative splicing *in vitro* (15–21). Functional roles for constitutive and/or alternative splicing *in vitro* have also been reported for hnRNPA1 and hnRNP C1/C2, for which sequences are available from several species (22–24). In addition, the complete sequences of several other hnRNP and snRNP polypeptides have been known for some time (reviewed in 3,4). Finally, several regulators of alternative splicing in *Drosophila*, first identified genetically, have been cloned and sequenced, including Transformer (Tra), Transformer-2 (Tra-2), Sex-lethal (Sxl), and Suppressor of white apricot (Su(w^a)) (reviewed in 25).

Several of the above proteins are closely related in sequence, and many share structural features among themselves and with other RNA-binding proteins. Other sequence features are unique to a small subset of these proteins. The most notable features present in many of these proteins include one or more RNP-type RNA-recognition motifs (RRM or CS-RBD or RNP-80), and clusters of Arg and Ser residues (RS domains). Gly-rich clusters, sometimes referred to as hinge regions, are also common, as are domains with many Gly–Gly dipeptides interspersed with aromatic and Arg residues (RGG or GAR domains). The RRMs of various proteins have been extensively studied (reviewed in 26–30). As more members of the RRM superfamily have been

* To whom correspondence should be addressed

Present addresses: +Balliol College, Oxford, OX1 3BJ, UK and §New England Biolabs, Beverly, MA 01915, USA

Table 1. Sequences in the comprehensive RRM database

```
SR proteins
    humSF2=gp:m69040(17:96)(122:197!)
    droSRP55=sw:sr55_drome[gp:x62599](5:79)(116:193!)
    araSR=gp:m98340(8:89)(122:197!)
    humSC35=gp:x62447[gp:x62446](15:98)
    murX16=sw:x16_mouse[gp:l10838](11:88)
        droRBP1=gp:l04929(12:91)
Other RS domain proteins
    droTRA2=sw:tra2_drome(98:180)
    humU2AF=sw:ua2f_human[gp:x64587](151:237)(261:343)(378:472!)
    sacYCl11c=sw:ycb1_yeast(123:203)(220:301)(350:431!)
#   humARGNP=gp:m74002(34:120!)
snRNP-associated proteins
    humU1A=sw:ru1a_human(11:94)(208:287)
        droU1A=gp:m89775(8:93)(144:216)
        xenU1A=gp:x57953(11:96)(209:284+)
    humU2B=sw:ru2b_human(8:90)(152:229)
    solU2B=gp:m72892(12:103)(209:284+)
    humU170k=sw:ru17_human[gp:x15776][gp:x06815](281:363)
        droU170k=sw:ru17_drome(101:187)
        xenU170k=sw:ru17_xenla(103:191)
    araU170k=gp:m93439(33:117)
    sacU170k=gp:x59986(108:195)
hnRNP proteins
    humHNRNPA1=sw:roa1_human[gp:m99167](14:95)(105:186)
        droHNRNPA1=roa1_drome(24:107)(115:198)
        graHNRNPA1=roa1_scham(18:101)(109:192)
        droHRP6=gp:m25545(27:110)(118:201)
        droHRP40=gp:x62637(57:136)(138:220)
        droHRP36=gp:x62636(25:108)
        droHRP97=gp:l02106(33:116)(12:207)
        humHNRNPA2/B1=sw:roa2_human(22:104)(113:195)
        xenHNRNPAA=sw:roaa_xenla[sw:roab_xenla](15:98)(106:89)
        droP11=gp:x59691[gp:x58183](28:118)(117:197)
        caeRBP=gp:d10877(21:104)(111:185)
        xenNRP=gp:m34895(21:104)(111:185)
    humHNRNPAB=gp:m65028(69:152)(153:237)
    murCBF=gp:d90151(76:159)(160:243)
    ratHNRNPAB=sw:roc_rat(33:117)
        humE2BP=gp:m94630(77:160)(162:243)
    humHNRNPC=sw:roc_human(17:87)
        xenHNRNPC=sw:roc_xenla(18:88)
    humHNRNPL=sw:rol_human(71:150!)(164:242!)(352:430!)(472:559!)
    humHNRNPI=sw:ptb_human[sw:ptb_mouse](60:142!)(185:263!)(338:416!)(455:533!)
    humHNRNPM=gp:l03532(72:155)(204:287)(653:729)
Other hnRNA-associated proteins
    droELAV=sw:elav_drome[sw:elav_drovi](150:245)(259:332)(403:485)
        droSXL=sw:sxlf_drome(126:208)(212:294)
        humHUD=sw:hud_human(47:129)(133:215)(298:380)
        droCPO=gp:z14974(452:536)
    droBJ6=gp:x55902(303:375)(377:460)
        humPSF=gp:x70944[#sw:cs24_human(298:367)(372:451)
    sacNPL3=gp:m86731(126:200)(#201:280!)
Chloroplast RNA-binding proteins
    nic28kd=sw:ro28_nicsy(98:181)(192:275)
        spi28kd=ro28_spiol(49:133)(143:227)
        zea31kd=gp:m74566(126:209)(220:296)
    plu30kd=gp:x65118[gp:x61113](88:172)(195:279)
        tob31kd=gp:x65117(89:173)(209:293)
    nic31kd=sw:ro31_nicsy[gp:s38122](137:220)(231:315)
    araRNABP=gp:m94554[gp:x65255](108:192)(202:286)
    nic33kd=sw:ro33_nicsy(114:197)(218:301)
        tab33K6=gp:x61115(104:180)(205:289)
Stress-induced plant proteins
    zeaABI=sw:abal_maize(10:86)
        zeaGRP=gp:x61121(5:84)
    carGRP=gp:x58146(8:84)
    napGLY=gp:z14143(8:84)
    araCCRB=gp:l04171[gp:l00649][z14988](2:91)
    vulGLY=gp:x57663[gp:x57662](+1:75)
Pre-rRNA processing factors
    humNUCL=sw:nucl_human(308:388)(394:471)(487:566)(573:649)
        ratNUCL=sw:nucl_rat[sw:nucl_mouse](311:393)(397:476)(489:569)(575:656)
        mesNUCL=sw:nucl_mesau(308:389)(394:467)(486:559)(572:646)
        galNUCL=sw:nucl_chick(280:364)(372:452)(462:542)(554:635)
        xenNUCL=sw:nucl_xenla(109:191)(201:281)(291:370)(379:460)
    sacNSR1=sw:nsr1_yeast(169:251)(268:350)
    sacRNA12+=gp:s92205(196:279!)
    sacSSB1=sw:ssb1_yeast(38:124)(186:278)
Poly(A)-binding proteins
    humPABP=sw:papb_human[gp:x65553](12:94)(100:180)(192:270)(292:372)
        droPABP=sw:papb_drome(4:90)(92:174)(178:266)(286:372)
        xenPABP=sw:pabp_xenla[gp:x57483](12:97)(100:184)(188:276)(295:379)
        sacPABP=sw:pabp_yeast(39:123)(127:210)(220:303)(323:406)
        schPABP=gp:m64603(67:151)(154:238)(248:331)(354:434)
        araPABP=gp:m97657(46:129)(133:216)(226:309)(329:412)
    sacRNA15=sw:rn15_yeast(19:103)
Other proteins
    humTIA-1=gb:m77142=pir:a39293(8:85)(96:174)(206:276)
        humTIAR=gp:m96954(11:89)(99:179)(207:280)
        sacPUB1=gp:l13725[gp:l01797](76:159)(164:247)(342:420)
    humRO=sw:ro6_human(203:300!!)
    humLA=sw:la_human(230:300!!)
        bosLA=gp:x13698(203:330!!)
    humEIF4B=sw:if4b_human(97:179)
#   droMODU=sw:modu_drome(177:258!)(261:339!)(343:419!)(421:508!)
#   sacCDC63=sw:cc63_yeast(80:168)
#   hanPRT1=sw:ptr1_hanpo(40:126)
    humRD=sw:rdp_human[gp:m21332](262:339)
    humSCF=gp:m85085(17:99)
#   schMEI2=sw:mei2_schpo(196:277!)(296:373!)
#   humMSSP-1=gp:x64652(25:109)(109:198)
#   murBF41=sw:bf41_mouse(+1:69)
    plaARP2=sw:arp2_plafa(25:121!)(347:451!)
    sacNAM8=gp:x64763(#56:154!!)(164:249)(314:392)
        sacNGR=gp:z14097(#36:179!!)(193:272)(361:432)
    murP16=gp:x52102(2:77)(79:155)
    sacRNP1=gp:m88608(36:121)(142:230)
    humEWS=gp:x66899(362:446)
    droORB=gp:x64412(577:656!!)(689:837!!)
Possible RRM-containing proteins
#   humXE7A=gp:l03426(160:268!!)
#       humLSPRO=gp:m99578(153:260!!)
#   sacSEN3=gp:l06321(476:556!!)
#   araDDP=gp:m98455(281:371!!)
#   humU2AF35=gp:m96982(60:154!!)
```

Sequences are grouped by the known or presumed functions of the proteins. These groupings are generally not phylogenetic. Phylogenetic relationships are shown by indentation under a representative sequence. Homologs that only differ in species of origin and/or by only a few amino acids are given in [ ] after the sequence. Sequences are given an abbreviated name in UPPER case (usually the appropiate gene name) preceded by a species code of three lower case letters. SWISS-PROT locus names are given wherever possible, in the form sw:locus_name; otherwise GenPept or GenBank accession numbers are given as gp:accession_number or gb:accession_number, respectively. Finally, the coordinates of the RRMs are given as (res1:res2). An ! inside a bracket indicates an atypical RRM, a !! indicates a very atypical/questionable RRM. A + at either end denotes a truncation within the RRM, usually because the sequence was a cDNA or protein fragment. A # in the margin indicates a sequence in which no RRMs had been previously reported; a # inside a bracket indicates that that particular RRM had not been previously reported. Sequences that were used in the first database search are in *italics*, and those in the second search are in **bold** (see Methods). This table is available on a file-server. Database codes: sw, SwissProt; gp, GenPept; gb, GenBank. Species codes: ara, *Arabidopsis thaliana*; bos, bovine; cae, *Caenorhabditis elegans*; car, *Daucus carota*; dro, *Drosophila melanogaster*; gal, *Gallus gallus*; han, *Hansenula polymorpha*; hum, human; mes, *Mesocricetus auratus*; mur, mouse; nic, *Nicotiana sylvestris*; plu, *Nicotiana plumbaginifolia*; rat, rat; sac, *Saccharomyces cerevisiae*; sch, *Schizosaccharomyces pombe*; sci, *Schistocerca americana*; sol, *Solanum tuberosum*; spi, *Spinacia oleracea*; tab, *Nicotiana tabacum*; vul, *Sorghum vulgare*; xen, *Xenopus laevis*; zea, *Zea mays*.

discovered, the range of sequence variation has greatly increased. Thus, a systematic reevaluation of sequence criteria for identifying bonafide RRMs has become necessary. In contrast to RRMs, much less is known at present about the structure and function of RS (7,31,32) and RGG domains (33,34).

We report a systematic comparison of sequence features present in these splicing factors, in an attempt to understand the structural, functional and evolutionary relationships among these proteins. Conventional global pairwise alignment searches (e.g., FASTA (35)) do not properly identify homologies between proteins with similar domains that are arranged differently, or between proteins that share one domain but differ otherwise. Even programs based on local alignments (e.g., BLAST (36), BLAZE (37), or BLITZ (38)) have difficulty identifying degenerate motifs composed of conserved but dispersed residues. Since most of the proteins of interest are composed of multiple degenerate domains, such searches are of limited value. Our approach was to analyze each identifiable domain separately, and to build a database of all proteins containing one of the domains, the RRM, using profile analysis (39). This database was then searched by conventional pairwise alignments with other domains or full-length proteins of interest. This approach was effective in identifying proteins with unreported homologies to splicing factors.

METHODS

Putative RRM-containing sequences were identified by first aligning 99 known RRMs from 45 sequences (shown in *italics* in Table 1) using the program PILEUP (gap weight = 2.3, gap extension = 0.05) in the GCG suite of sequence analysis programs (40). This alignment was then used to generate a profile with the GCG program PROFILEMAKE (logarithmic weighting, gap ratio = 0.33 and gap length ratio = 0.1 were unchanged from default). The profile was used to search the release of the GenPept database corresponding to GenBank 72 (gap penalty = 3.4; gap extension = 0.05). The scores were not corrected for amino acid composition nor normalized for length. Sequences scoring above 8.3 were analyzed further, and a subset of these sequences (75 sequences shown in **bold** in Table 1, comprising 113 RRMs) were manually aligned. A representative subset of this alignment is shown in Figure 1, which displays 70 RRMs, 67 of which were taken from the larger alignment of 113 RRMs. The remaining three RRMs shown in Figure 1 are novel and were found subsequently. The large alignment was used to generate the RRM profile, as above. The RRM profile was then used to search the release of GenPept corresponding to GenBank 75 (scores not normalized for length or composition; gap penalty = 3.5; gap extension = 0.05), and sequences with scores better than 8.18 were selected as the RRM-enriched database. Each of these sequences was analyzed further, and the putative RRM-containing sequences are given in Table 1.

Statistical analyses of amino acid composition and arrangement (clustering and periodicity) were carried out using the SAPS program (41). Residue distributions were calculated with reference to the SWISS-PROT database, release 20.

Phylogenetic trees were inferred from the aligned RRM sequences using two methods: the neighbor-joining method (42) as implemented in the CLUSTALV sequence analysis package (43), and the method of maximum parsimony using the PROTPARS module of the PHYLIP phylogeny inference package (44,45). Gapped regions were included in the alignments used for each analysis. In the neighbor-joining method, distances were corrected by the method of Kimura (46). In the CLUSTALV implementation of Kimura's distance correction, distances greater than 82% were arbitrarily corrected to 330%. Therefore, accurate branch length estimates were not possible for very distant sequences. Confidence limits on the trees were estimated by bootstrap sampling the aligned data set (47), applying the appropriate inference method to each bootstrap sample, and tallying the occurrence of each monophyletic grouping. 1000 bootstrap samples were used for the neighbor-joining method, 100 for the maximum parsimony procedure. For the latter method, a majority consensus tree was constructed by applying the PHYLIP CONSENSE module to the set of trees generated from the bootstrap samples. The trees generated by either procedure were initially displayed using the PHYLIP DRAWTREE module, and then redrawn on a personal computer to add confidence intervals.

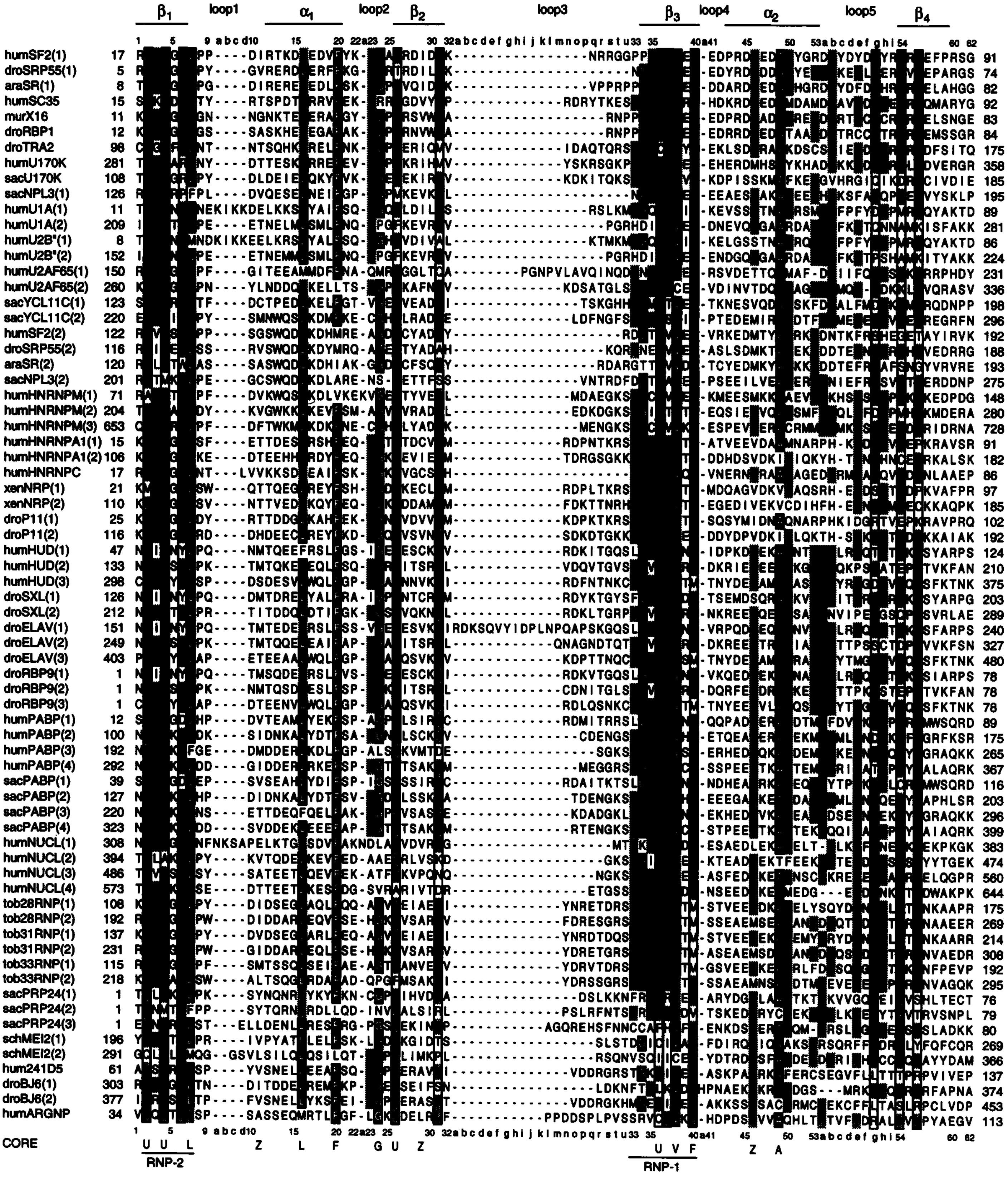
The GenPept database was used for virtually all database searches, but SWISS-PROT locus names are given wherever possible. Entries that appear in one or more figures or tables are given an abbreviated name in uppercase preceded by a three-letter species code in lowercase, using the nomenclature defined in Table 1. This is followed by the equivalent locus name or accession number. Sequences mentioned in the text, but which do not appear in the figures, are given only a locus name or accession number with the format sw:locus_name for SWISS-PROT entries and gp:accession_number for GenPept entries. The accession numbers for GenPept sequences are the same as their parent DNA sequences in GenBank.

Profile searches were performed principally on the Vax cluster at the Oxford Molecular Biology Data Centre, with additional searches on the Vax cluster at the ICRF Bioinformatics group and on the Vax at Cold Spring Harbor Laboratory. The network server at NCBI was used for BLAST searches, the BLAZE(TM) mail server for BLAZE searches, and the BLITZ mail server at Heidelberg for BLITZ searches. Sequence alignment and residue distribution figures were produced with software written by J. Posfai and E.B. using the PostScript(TM) language.

RESULTS AND DISCUSSION

Analysis of RRM domains

The RRM is a region of around eighty amino acids containing several well conserved residues, some of which cluster into two short submotifs, RNP-1 (octamer) and RNP-2 (hexamer) (reviewed in 26–30). One or more RRMs are found in a variety of RNA-binding proteins, including hnRNP proteins, translation factors, snRNP polypeptides, proteins involved in pre-mRNA and pre-rRNA processing, and poly(A)-binding proteins (see Table 1). Each RRM can form a globular domain that in at least some cases is capable of independently binding RNA. However, in other cases regions distinct from the RRM, or synergy between RRMs in a multi-domain protein, are required for either general or sequence-specific RNA binding (7,27,32,48–50). The first of two RRMs in the U1-A polypeptide of U1 snRNP (humU1A=sw:ru1a_human), as an 89 aa fragment, is sufficient for binding to stem-loop IV of U1 snRNA with specificity and affinity comparable to the full-length protein (48,51–55). The three-dimensional structure of this fragment has been solved by X-ray diffraction and NMR (56,58). In addition, a 93 amino acid fragment of hnRNP C (humHNRNPC=sw:roc_human), containing its RRM, has been characterized by NMR in the presence and absence of RNA (59,60). These two RRMs are very similar and consist of a $\beta_1-\alpha_1-\beta_2-\beta_3-\alpha_2-\beta_4$ structure, with the RNP-1 and RNP-2 submotifs lying in the central anti-parallel β_3 and β_1 strands, respectively. The only notable difference between these two RRM structures is the longer α_1 helix in U1-A. The β_1 and β_3 strands are involved in RNA binding, as shown by NMR (60), and have conserved solvent-exposed aromatic residues that are among the residues implicated in contacting RNA, as shown by mutational and UV crosslinking studies (27,32,48,52–55,61,62).


β1
loop1
α1
loop2
β2
loop3
β3
loop4
α2
loop5
β4
CORE
U U L
Z L F G U Z
U V F
Z A
RNP-2
RNP-1

An alignment of RRMs. An alignment of 70 RRMs (Figure 1) was constructed with special emphasis on tertiary structural requirements, modeled on the two available structures. The alignment includes sequences of proteins thought to be involved in pre-mRNA splicing, spliceosome-associated factors, and some hnRNP proteins, as well as some additional putative RNA processing factors. Human and yeast poly(A)-binding proteins were included, together with nucleolin, to give some sampling of other RRM sequences, and to illustrate phylogenetic relationships (see below). All the RRMs present in these proteins were included, except for RRM3 of U2AF^{65} and RRM3 of a homologous protein in yeast, YCL11c, which are highly atypical (7,63). Alignments with different sets of RRMs have been constructed and all the features described below are consistent with all the alignments (data not shown).

The alignment shown in Figure 1 differs from previous alignments of RRMs (27) in the positioning of conserved residues in α_1 and loop2. The alignment shown here benefits from the recent availability of the hnRNP C RRM structure (59), in addition to that of the N-terminal U1-A RRM (56,58), which allowed us to look at the relative positions of residues within both structures. The most significant change implicates the conserved Gly in the tight turn at the end of loop2 leading into β_2, rather than in the last turn of the α_1 helix (27). This alignment also shows that the U1-A sequence is atypical in its longer α_1 helix, which gives rise to the first gapped region. The other major gap in the alignment corresponds to loop3 (also known as the variable loop), which can be easily accommodated into the model tertiary structure. Other small alignment gaps occur between α_1 and β_2, β_3 and α_2, and α_2 and β_4, all of which are plausible sites for insertions or deletions within the conserved tertiary structure. In some RRMs (not shown in Figure 1), there is also a gap at position 28 within the β-bulge in β_2 (positions 27–29). The RNP-1 and RNP-2 submotifs lie in β_3 and β_1 strands, respectively. The conserved aromatic residues at positions 3, 35, and 37 protrude from the β sheet to interact with the RNA, as shown by crosslinking and mutational studies (32,48,52–55, 61,62). Ring-stacking interactions between these solvent-exposed residues and single-stranded bases have been postulated (27).

A consensus RRM structural core sequence. The most conserved positions in the alignment (Leu7, Leu16, Phe20, Val38, Phe40, and Ala49) (Figure 1) correspond precisely to residues that form the hydrophobic core of the U1-A tertiary structure (56,58), as previously noted (27). In addition, Gly24 seems to be required for the turn into β_2. Based on the analysis of the RRM alignment and the model three-dimensional RRM structure, we propose the following consensus structural core sequence for RRMs:

UxUxxLxxx[x$_{0-6}$]Z[x]xxxLxxxFxxx[x]GxUx[x]Zxxxxxx
[x$_{0-21+}$]UxVxF[x]xxxxxxZxxA

(x = any residue; U = uncharged residues: L,I,V,A,G,F, W,Y,C,M; Z = U + S,T; + indicates that loop3 may be expanded further)

We note that this is a degenerate consensus, i.e., no single position is absolutely invariant. Although position 34 has a highly conserved Gly residue, we did not include it in the consensus because its role appears to be to connect loop3 to β_3. Since loop3 does not appear to be involved in the structure of the domain (see below), only a subset of RRMs might require a Gly at this position. There are many additional positions in the alignment that show conservation, and in general, additional conserved residues must be required to form the two α-helices and four β-strands, and to fold them into a correct RRM structure. However, it is unlikely that an RRM can exist without at least conservative substitutions in the above consensus sequence. Although this consensus is too permissive to be used as the sole criterion to identify RRMs, it will identify most, if not all RRMs, including those with atypical RNP-1 submotifs. The RNP-1 submotif remains the most obvious signature for typical RRMs, but many RRMs contain atypical RNP-1 submotifs (see Table 1) and sequences matching the RNP-1 submotif consensus are found in proteins that lack an RRM (see below).

RRM positions with potential to contact RNA. All the solvent-exposed positions in and near the β-sheet have the potential to contribute to RNA binding. The three solvent-exposed aromatic positions (3 in RNP-2; 35 and 37 in RNP-1), which have been implicated in RNA binding, are predominantly Phe or Tyr (73%, 60%, and 74%, respectively; taken from a weighted dataset of 179 sequences; data not shown) but they can tolerate substitutions (Figure 1). Thus, these conserved aromatic residues are not always required for RNA binding, as they are absent from many putative RRMs. For example, hnRNP L (sw:rol_human), which lacks these conserved aromatic residues in all but one position of one of its four RRMs, can bind RNA (64). The variability seen at these usually aromatic positions could reflect sequence-specific contacts. For example, Gln54 of the U1-A polypeptide, which is in the usual aromatic position 35, has been implicated in sequence-specific hydrogen bonding to stem-loop II of U1

Figure 1. Alignment of 70 selected RRMs. Selected RRMs were aligned manually with special emphasis on tertiary structural requirements modeled after two available structures, as described in the text. Alignment gaps are indicated by dashes. Sequence names are as in Table 1. For sequences with more than one RRM, each one is numbered from the N-terminus, with the number given in parentheses after the sequence name. Beginning and end residue positions within the parent protein are given for each RRM at left and right of the alignment. The sacPRP24 and droRBP9 sequences are from reference (27); the accession numbers for the remaining sequences are given in Table 1. The alignment positions are numbered above and below, following the nomenclature of Kenan *et al.* (27), in which conserved positions are given sequential numbers, whereas positions that are not present in all RRMs are alphabetized. However, as the alignment in reference (27) differs from the one above, the numbers are not interchangeable between the two. The positions of conserved residues are highlighted by vertical shading. Black shading indicates positions in which a single residue occurs in at least 75% of the sequences. Grey shading at the same positions represent conservative substitutions. Elsewhere, grey shading indicates positions in which residues belonging to a single conservative grouping are present in at least 50% of the sequences. When a single conservative grouping represents at least 75% of the sequences, this is denoted by grey shading in boxed columns. Acceptable conservative groupings were I=V=L,F=Y=W,Q=N,R=K,D=E,S=T. Three additional positions at which the consensus is split between two residues were also shaded: positions 6 and 53g (G=N) and position 36 (A=G). The consensus structural core residues are shown below the alignment (U = uncharged residues: L,I,V,A,G,F,W,Y,C,M; Z = U + S,T), along with the position of the RNP-1 and RNP-2 submotifs. Secondary structure, modeled primarily after the humU1A(1) tertiary structure (56,58) except for α_1, which was based on the humHNRNPC secondary structure (59), is given above the alignment. In humU1A(1) α_1 extends another 2 residues towards the N-terminus. Most positions of the RRM lack a firm consensus. The range of sequence similarity between two otherwise unrelated RRMs is 10–20% identity.

snRNA, and mutating this residue to Phe drastically reduces binding (52). In some RRMs, solvent-exposed aromatic residues in β_2 (humSC35, murX16, sacNPL3) or β_4 (sacPABP, humPABP) might have similar functions to those of the aromatic residues in β_1 and β_3.

Since residues within loop3 are not involved in the overall structure of the RRM, substitutions, insertions and deletions in this loop may be easily tolerated. Thus, alignments between direct homologs, such as human *Xenopus*, chick and rat nucleolin, or human, *Xenopus*, *Drosophila* and yeast U1-70K, show greatest conservation in the β strands, and greatest variability in loop3 of the corresponding RRMs, suggesting that accumulated mutations in these regions do not affect RRM function (data not shown; see Table 1 for accession numbers). Variability is also seen in the α-helices, as expected, since they are on the opposite side of the RNA-binding surface. Furthermore, in the yeast U1-70K homolog (sacU170K=gp:x59986), whose RRM can be functionally replaced by the human U1-70K RRM (humU170K=sw:ru17__human), loop3 is one of the most variable regions compared to the human RRM (65,66). However, in other protein families, such as the hnRNP A1-like proteins, loop3 sequences do show conservation.

Replacement of loop3 of RRM1 of the U1-A polypeptide (humU1A=sw:ru1a__human) by the analogous region of the U2-B″ polypeptide (humU2B=sw:ru2b__human) abolishes its ability to distinguish between U1 and U2 snRNAs (55). Further replacement of residues in part of β_2, in addition to loop3, reverses the RNA-binding specificity of this RRM (54). However, this is one of the few regions of divergence between the N-terminal RRMs of these two highly homologous proteins, and one of the only divergent solvent-exposed regions near the β-sheet. The issue of binding specificity is further complicated in this case by the fact that the U2-B″ protein requires the U2-A′ protein for specific binding to U2 snRNA (55,56). Interestingly, the *Drosophila* protein dro25 (droU1A=gp:m89775), whose first RRM has a loop3 region that is almost identical to that of human U2-B″, in fact binds to *Drosophila* U1 RNA *in vivo* (57). In short, although in general loop3 shows the greatest variability in sequence, this need not mean that it is the major determinant of sequence-specific binding.

Construction of an RRM-enriched database

The RRM has only a few well-conserved residues, mostly in the RNP-1 and RNP-2 submotifs (27; Figure 1). Pairwise alignments of unrelated RRM sequences are often incorrect, due to the additional residues within the RRM. To overcome these limitations, we employed profile analysis (39). A profile is a position-dependent scoring table that summarizes the preferences for amino acid residues and the acceptability of gaps at each position in a set of aligned sequences. By constructing a profile from a set of aligned RRM sequences, flexibility for residue type and gaps is promoted in some regions whereas conserved residues are strongly enforced. In addition subtle preferences for broad residue type (e.g., uncharged, small residues) become apparent and are represented in the profile. By using a profile generated from an alignment of 75 RRM-containing sequences (113 RRMs in total), essentially a larger version of the alignment in Figure 1, we could identify many, if not all sequences with potential RRMs in a large database. The resulting limited set of sequences, which represents only 0.66% of the entries in the GenPept database, can then be used to search for other domains and motifs. The reduced database size increases the statistical significance of otherwise weak similarities found among these proteins outside their RRMs.

As expected, known RRMs (both present in and absent from the profile alignment) consistently produced high scores (except for *E. coli* rho protein, *Drosophila* bicoid, *Drosophila* Suppressor of sable and bacteriophage ϕ29gp10, see below). Atypical RRMs, whether present in the alignment, such as hnRNP L (sw:rol__human), or absent from it, such as La protein (sw:la__human), gave lower scores. The first 434 scores (cutoff score of 8.18) were arbitrarily selected to produce the RRM-enriched database. This database was over twice the size necessary to include the last known positive in the search (La protein) (67,68). This RRM-enriched database included sequences that lack well-defined RNP-1 and RNP-2 submotifs, which are the hallmarks of RRMs, but that contain the virtually invariant hydrophobic residues located in the RRM core. The cutoff was chosen liberally to include most, if not all sequences with potential RRMs. As a result, only 29% of the sequences in this RRM-enriched database appear to contain an RRM. However, this contrasts with 0.0019% putative RRM-containing sequences in the GenPept database.

In none of the searches to identify RRM-containing sequences were either the bacterial rho protein (sw:rho__ecoli) (69), the ϕ29gp10 protein (sw:vg10__bpph2) (70), the bicoid protein (sw:hmbc__drome) (71) or Suppressor of sable (gp:m57889) (72) identified, even though they were all previously reported to have RRMs (69–72). They scored 7.00, 7.83, 6.87 and 7.46, respectively, against the RRM profile using identical search conditions as above. In each case, randomized sequences that maintained the composition but altered the order of residues gave significantly lower scores than the original sequence. The above core consensus can be made to fit the ϕ29gp10 sequence, whereas for rho, bicoid, and Suppressor of sable, unprecedented gaps and substitutions must be accommodated. Mutations of the two solvent-exposed Phe residues in the putative RNP-1 sequence of rho protein reduce RNA binding (69). It is unclear if the criteria proposed here would support the presence of an RRM in ϕ29gp10, but neither bicoid, nor rho, nor Suppressor of sable proteins satisfy these criteria. This may indicate that the consensus for an RRM is broader than suggested here, that these regions have diverged from an ancestral RRM, or that the sequence homology found is spurious, despite the requirement for two Phe residues for RNA binding by rho.

The presence of RNP-1 submotifs in prokaryotic cold shock proteins and eukaryotic Y box transcription factors has been noted (73). However, in these proteins the submotif differs in the final position, which is predominantly Phe in RRM RNP-1 submotifs and Arg in cold shock proteins and Y box factors. Neither the cold shock proteins nor the Y box factors were identified in our RRM searches. The crystal and solution structures of the *B. subtilis* major cold shock protein have been recently reported (74,75) and show that the RNP-1 submotif lies in the second β strand, but with an entirely different overall topology from that of the prototype RRM. Interestingly, the position of the residues in this β_2 strand match very closely with those in the U1-A RNP-1, and β_3 of the cold shock protein has residues in analogous positions to those of the RNP-2 submotif in U1-A (74,75), although in a very different position relative to the RNP-1 along the primary sequence. The conservation of these RNP-1 submotifs could be a case of convergent evolution of nucleic acid-binding domains. This illustrates one limitation of using the RNP-1 submotif as the sole signature for an RRM. For example,

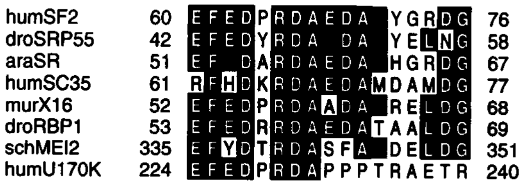

humSF2 60 E F E D P R D A E D A Y G R D G 76
droSRP55 42 E F E D Y R D A D A Y E L N G 58
araSR 51 E F D A R D A E D A H G R D G 67
humSC35 61 R F H D K R D A E D A M D A M D G 77
murX16 52 E F E D P R D A A D A R E L D G 68
droRBP1 53 E F E D R R D A E D A T A A L D G 69
schMEI2 335 E F Y D T R D A S F A D E L D G 351
humU170K 224 E F E D P R D A P P P T R A E T R 240

Figure 2. A conserved octapeptide motif. The alignment shows a partially conserved octapeptide motif found in all SR proteins in the current databases (human SF2/ASF, *Drosophila* SRp55, an *Arabidopsis* SR protein, human SC35, murine X16 and *Drosophila* RBP1), as well as in *S. pombe* mei2 and human U1-70K polypeptide. In all but the U1-70K polypeptide this motif is found within an RRM, overlapping the last two amino acids of the RNP-1 submotif. The alignment was extended by nine residues towards the C-terminus to illustrate the further homology found among the first seven sequences, which is higher than between unrelated RRMs. Sequences upstream of the octapeptide are not shown because they comprise the rest of the RNP-1 submotif (except in U1-70K) and the homology is high, as expected. Black boxes indicate identities, whereas grey boxes denote conservative groupings, as in Figure 1. Positions were shaded if identities or conservative groupings were present in four or more sequences.

bacteriophage T4gp32, which has been noted to have a sequence with similarity to the RNP-1 submotif (26,76), does not satisfy the structural criteria for an RRM, and binds RNA by a mechanism that does not involve an RRM-like structure (26,77).

A comprehensive RRM database

Although the entire RRM-enriched database was employed for subsequent searches, we attempted to identify all sequences containing bonafide RRMs by manually examining each entry for the presence of the motif using the structure-based RRM consensus given above. A comprehensive list of the resulting RRM-containing sequences is given in Table 1. This RRM database contains 22 previously unreported RRMs, strongly suggesting that the function of the corresponding proteins, which are implicated in a variety of cellular processes, involves binding to RNA. For example, the *S. cerevisiae* gene RNA12+ (gp:s92205), which is involved in rRNA maturation, contains a previously unreported RRM, suggesting that the gene product directly binds rRNA. This analysis also revealed previously unreported RRMs in *Drosophila* modulo (sw:modu_drome) and human MPSS-1 (gp:x64652), both of which have been shown to bind DNA. Given the presence of multiple RRMs in these proteins, their relative affinities for dsDNA, ssDNA and RNA should be measured. A fourth example is a partial human cDNA that was originally identified as a myoblast cell surface protein (hum241D5 = sw:cs24_human), and has a good fit to the RRM consensus (Figure 1), suggesting that it was cloned fortuitously. Recently, it was shown that this cDNA represents a fragment of a recently cloned human splicing factor, PSF (gp:x70944) (11).

Unreported homologies between sequences in the RRM-enriched database and known splicing factors

The RRM-enriched database was searched using FASTA (35), with the sequences of known splicing factors and spliceosomal proteins, or domains thereof. Unexpected sequences identified through these searches were then manually examined for the presence of one or more RRMs using the consensus derived above, and for other sequence features.

Human Arg-rich nuclear protein. When the RRM-enriched database was searched with the RS domains from various proteins, an Arg-rich human nuclear protein of 54 kDa (humARGNP = gp:m74002) (78) gave consistently high scores. This sequence has both an extensive RS domain and a slightly atypical, previously unreported RRM upstream of the RS domain (residues 34-113; Figure 1; see also Figure 5, below). This protein localizes to the nucleoplasmic speckled region of the mammalian nucleus (78), a region enriched in many splicing components (reviewed in 79). This subnuclear localization is consistent with the presence of an RS domain (31), and together with the presence of an RRM, suggests a possible role in pre-mRNA splicing. The humARGNP RRM does not resemble the RRMs of SR proteins, nor was any additional homology observed between humARGNP and other SR proteins, besides the RS domain and the expected similarity between otherwise unrelated RRMs (data not shown).

S. pombe mei2. Another sequence we identified as having homology to splicing factors was mei2, the product of a gene involved in meiosis in *S. pombe* (schMEI2 = sw:mei2_schpo) (80). It showed a partial match to a region previously shown to be homologous between SF2/ASF (humSF2 = gp:m69040) and U1-70K (humU170K = sw:ru17_human), the octapeptide EFEDPRDA (5). Exact or almost exact fits to this octapeptide were also found in the six other sequenced SR proteins, all just after the RNP-1 submotif, extending from the end of β_3 into α_2 (Figure 2). The octapeptide in mei2 is in the identical position in a putative RRM (residues 296–373) as in the SR proteins. Therefore, the mei2 octapeptide is present in the same structural context, although the sequence of the putative RRM is very atypical. The octapeptide in U1-70K is located upstream of the single RRM, and is strongly conserved in human, *Xenopus*, and *Drosophila* proteins (5), but not in the more divergent *S. cerevisiae* homolog (65,66) (see Table 1 for accession numbers). An ungapped alignment of the homologous segments of yeast mei2, the six SR proteins, and human U1-70K is shown in Figure 2. The homology between U1-70K and the other proteins does not extend beyond either side of the octapeptide (Figure 2 and data not shown). No other exact matches to the octapeptide were found either in the RRM-enriched database, in SWISS-PROT, or in six-frame translations of GenBank. The function of the octapeptide motif in any of the above proteins is presently unknown.

S. pombe mei2 has at least two previously unreported RRMs: the one containing the octapeptide, and another immediately preceding it (Figure 1). Neither of these RRMs contains a good fit to the RNP-1 consensus (DGICIVAF and VSQIICEF). However, the correct spacing of the hydrophobic core residues strongly suggests that these putative RRMs can fold into the correct prototype structure, which in turn suggests that the ability to bind RNA has also been conserved. A good fit to the RNP-1 submotif (VGYAFINF) is found towards the C-terminus of mei2. It is unclear whether this is part of a third RRM, as the requisite Ala residue in α_2 is replaced by a Phe residue, which would not fit in the model RRM tertiary structure.

Recent work showed that wild type mei2 is required for efficient splicing of the mes1 intron during meiosis in *S. pombe* (C. Shimoda, personal communication). Alternative splicing of the mes1 intron is of the intron retention type (2). The presence of two RRMs (Figure 1) and the additional homology to SR proteins (Figure 2) are consistent with a direct role of mei2 in pre-mRNA splicing in *S. pombe*. Unlike SR proteins, mei2 lacks an RS domain, but extensive RS domains have not yet been found in fission or budding yeasts.

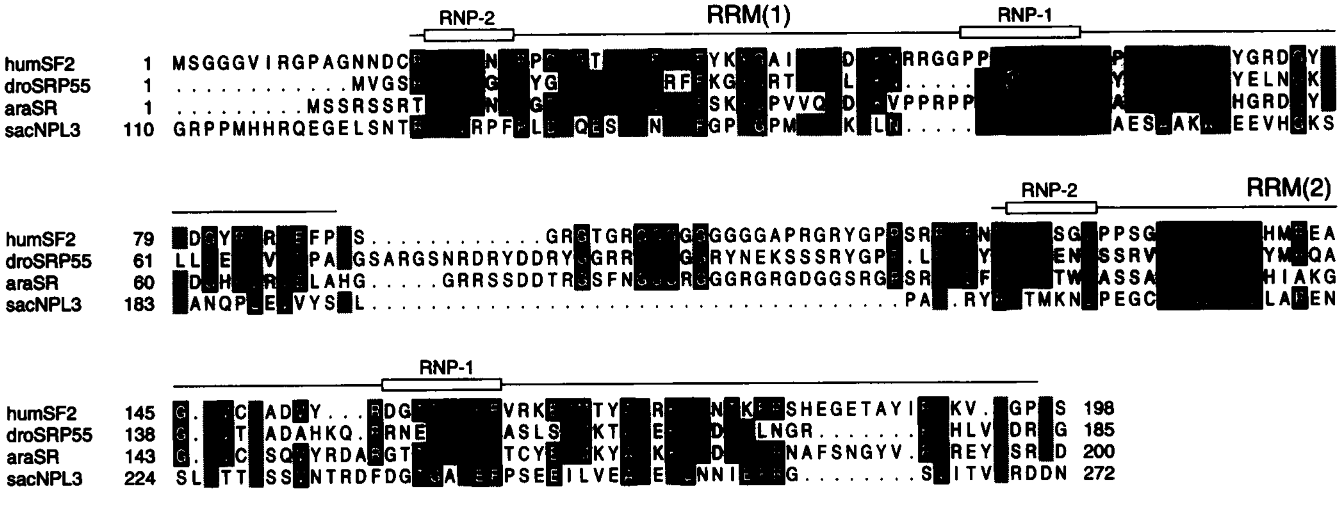


Figure 3. Sequence similarity between three metazoan SR proteins and yeast Npl3. The alignment shows N-terminal regions of human SF2/ASF, *Drosophila* SRp55, an *Arabidopsis* SR protein, and a central portion of *S. cerevisiae* Npl3. Residue positions for each protein are shown at left and at the end of the alignment. Black boxes indicate identities, whereas grey boxes show conservative groupings, as in Figure 1. Positions were shaded if identities or conservative groupings were present in three or more sequences. The positions of the two RRMs are shown by horizontal lines above the alignment, and the RNP-2 and RNP-1 submotifs of each RRM are boxed.

S. cerevisiae Npl3. A third striking homology was detected between *S. cerevisiae* Npl3 (sacNPL3A=gp:m86731), also known as NOP3 (gp:x66019), and three SR proteins: human SF2/ASF (humSF2=gp:m69040), *Drosophila* SRp55 (droSRP55 =sw:sr55_drome), and an *Arabidopsis* protein (araSR= gp:m98340). Npl3 was isolated in a genetic screen for factors involved in nuclear protein localization (81). However, this function may be indirect, as gene disruption did not affect protein localization, although mutant forms of the protein blocked transport. Npl3 has a good fit to the RRM consensus (81) followed by a second unreported, atypical RRM, similar to the above SR proteins (Figure 1). However, the proteins differ in that Npl3 has an N-terminal domain rich in Pro (22%) and Gln (16%), with imperfect Ala–Pro–Gln–Glu repeats unique in the database, and a C-terminal RGG domain (see below), whereas the SR proteins have characteristic C-terminal RS domains (5,6,12).

The alignment of the central region of *S. cerevisiae* Npl3 with the first 185–200 residues of human SF2/ASF, *Drosophila* SRp55, and the *Arabidopsis* SR protein is shown in Figure 3. Within this region, Npl3 is 30%, 29%, and 24% identical to araSR, SRp55 and SF2/ASF, respectively. This level of homology is far greater than expected for unrelated RRMs, and is particularly striking in the case of the second atypical RRM. Npl3 lacks the Gly hinge region between the RRMs, resulting in a large gap in the alignment. The two other significant gaps lie in the presumptive loop3 in the RRM structure. None of these gaps would therefore disrupt the expected RRM tertiary structure.

To date, proteins containing extensive RS domains have not been identified in either fission or budding yeast. The N-terminus of *S. cerevisiae* YCL11c (sacYCL11C=sw:ycb1_yeast) is Arg-rich (26% Arg in the first 52 amino acids), with four RS dipeptides (63). The observed homology between the RRMs of metazoan SR proteins and yeast Npl3 suggests that these proteins have a common ancestor. In addition to its expected ability to bind RNA, Npl3 may be involved in some aspect of mRNA processing, although its auxiliary domains suggest that its function is different from those of SR proteins. Recent experiments have implicated Npl3 in mRNA nuclear–cytoplasmic export, in addition to nuclear protein import, with several temperature-sensitive alleles mapping to Gly241 and Ala254 (P. Silver, personal communication). These residues are located within the second RRM of Npl3 (Figure 3), and are two of the positions of the structural core consensus defined above (Figure 1). The temperature-sensitive phenotype of these mutations is consistent with the requirement for the structural core consensus for the integrity of the tertiary structure of the RRM.

A conserved heptapeptide in RRMs of certain SR proteins, Npl3, YCL11c, and hnRNP M. The second RRMs of SF2/ASF (humSF2=gp:m69040), SRp55 (droSRP55=sw:sr55_drome), the *Arabidopsis* SR protein (araSR=gp:m98340) and yeast Npl3 (sacNPL3A=gp:m86731) includes the heptapeptide SWQDLKD, which is completely conserved in location and sequence (Figure 3). Exact matches to this heptapeptide were not found in any other protein in the RRM-enriched database. However, partial matches were found in YCL11c (sacYCL11C=sw:ycb1_yeast), an open reading frame in chromosome III of *S. cerevisiae*, which shows similarity to human $U2AF^{65}$ (63), and in all three RRMs of hnRNP M (humHNRNPM=gp:l03532) (see under α_1 region in Fig. 1). Although the sequence similarity between the above heptapeptides is low, the motif lies in precisely the same location in each of the RRMs, i.e., in α_1 on the opposite side of the β-sheet where the RNA is thought to lie. Only partial fits to the heptapeptide were found in six-frame translations of GenBank. The function of the heptapeptide remains unknown, but it is unlikely to be involved in directly contacting RNA, given its position within the tertiary structure of these proteins.

Human $U2AF^{35}$. The sequence $U2AF^{35}$ (humU2AF35=gp: m96983) yielded a high score against the RRM profile. $U2AF^{35}$ is a subunit of U2AF not required for biochemical complementation of splicing extracts depleted of strong poly(U)-binding proteins, and has an extensive RS domain (82). Further analysis did not identify an RRM, although we note that the region

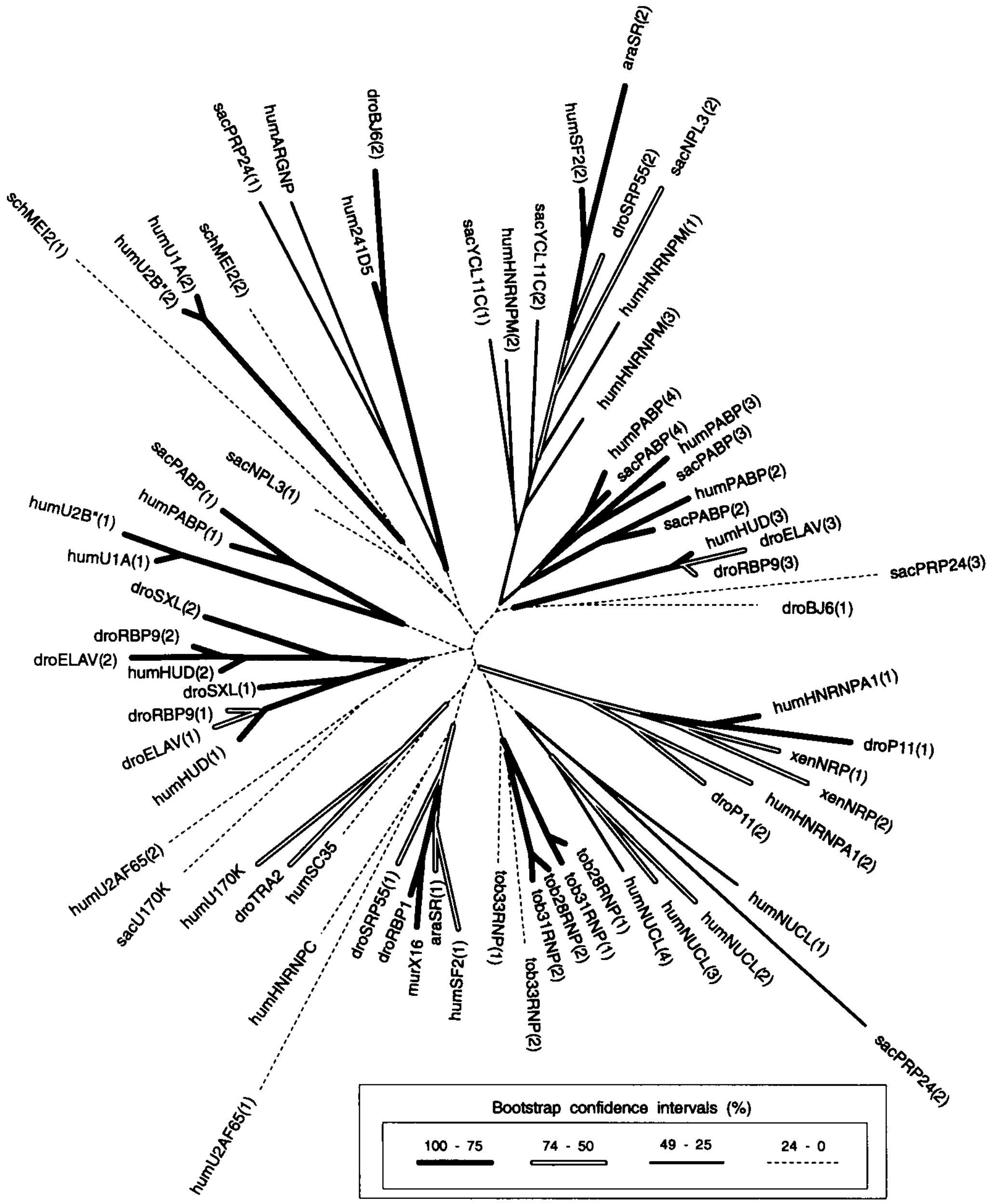


Figure 4. Phylogenetic tree of 70 selected RRMs. The phylogeny was derived by the neighbor-joining method from the alignment of all the RRMs shown in Fig. 1. Sequence names are as in Fig. 1. Confidence limits on the phylogeny were obtained by the bootstrap method, as described in Methods, and are represented by lines of different thickness, as shown at the bottom of the tree. A low bootstrap confidence interval for a particular grouping indicates that the homology seen is not consistent across the alignment. The central node for this unrooted tree was chosen arbitrarily. The groupings that have a 50% or greater confidence interval are in broad agreement with the phylogenies derived by the maximum parsimony method (see Methods). The overall similarity of the alignment used to generate this phylogeny is insufficient to derive accurate lengths for all branches (see Methods).

between residues 110–144 is very reminiscent of the β_3–loop4–α_2 region of the model RRM (not shown). This similarity could indicate a very atypical RRM or perhaps a structure that has evolved from an RRM to fulfill a different role. It will be interesting to see whether or not U2AF^{35} directly binds RNA.

Phylogenetic analysis of RRMs

The modular nature of RRMs has led to the proposal that these domains have evolved by duplication and diversification from an ancestral RNA-binding protein (reviewed in 26). The availability of a large set of RRM-containing proteins afforded us the opportunity to examine this model and the evolutionary relationships among RRMs. Phylogenetic analysis of the RRM alignment of Figure 1 by both the neighbor-joining method (42) and by the method of maximum parsimony (44,45) was carried out. The two trees were in broad agreement and the neighbor-joining tree is shown in Figure 4. Confidence intervals on the phylogeny were estimated by the bootstrap method (47), and are represented in the figure.

Table 2. Sequence conservation among SR proteins

	droSRP55(349)	araSR(303)	humSF2(248)	humSC35(221)	murX16(164)	droRBP1(135)
droSRP55		141 47%	123 50%	107 48%	86 52%	66 49%
araSR			145 58%	109 49%	80 49%	70 52%
humSF2				86 38%	79 48%	71 52%
humSC35					77 47%	64 47%
murX16						84 62%

The length in amino acids for each protein is given in parentheses. For each pairwise score the upper number shows the number of identical amino acids, and the lower number expresses this as a percentage of the smaller sequence. Alignments were constructed with full length sequences using the global alignment algorithm of Needleman and Wunsch, as implemented in the GCG GAP program with the Dayhoff scoring matrix. For the sake of consistency each alignment was made with a gap penalty of 1.7 and gap extension of 0.03. These parameters were found to give in nearly all cases alignments consistent with the domain structure of the proteins. These numbers do not accurately reflect homology and cannot be used to infer phylogenetic relationships. See Fig. 1 and Fig. 4 for the alignment and phylogeny of the first RRM for each of these sequences.

Domain duplication. RRM duplications within individual proteins appear to be common, as seen in the following groupings: (i) the last three RRMs of poly(A)-binding proteins (humPABP and sacPABP); (ii) both RRMs of the hnRNP A1-like proteins (humHNRNPA1, droP11, and xenNRP); (iii) the second and third RRMs of nucleolin (humNUCL); (iv) the tobacco chloroplast RNA-binding proteins (tob28RNP, tob31RNP, and tob33RNP), except for tob33RNP(1). The level of amino acid similarity between corresponding RRMs in two homologous proteins that contain multiple RRMs is often greater than the similarity between the multiple RRMs of either protein (83). This is most obvious in the poly(A)-binding proteins, in which the grouping of corresponding yeast and human RRMs implies that RRM duplication preceded the divergence between yeasts and metazoans. The observed phylogenetic pattern is consistent with the idea that each RRM evolved independently after the duplication event.

Gene duplication. RRM relatedness between different proteins, reflecting gene duplication, is also evident for: (i) the SR family of splicing factors as seen in the grouping of their first RRMs (humSF2, droSRP55, murX16, droRBP1, and araSR; humSC35 does not group with this phylogeny, and other groupings of humSC35 are not supported by the bootstrap analysis), and, when present, of their second RRMs; (ii) hnRNP A1-type proteins (humHNRNPA1, droP11, and xenNRP) (84,85); (iii) chloroplast RNA-binding proteins (tob28RNP, tob31RNP, and tob33RNP) (86); (iv) droSXL, droELAV, humHUD, and droRBP9 (87,88). Gene duplication provides an opportunity to evolve restricted developmental or tissue-specific expression patterns for certain RRM-containing proteins that may acquire unique functions or binding specificities. For example, mammalian hnRNP A1 (humHNRNPA1=sw:roa1_human) is expressed in many tissues, whereas its amphibian relative, NRP-1 (xenNRP=gp: m34895), is expressed only in neuronal tissues (89).

When related multi-RRM proteins from different species are aligned (e.g. poly(A)-binding proteins, hnRNP A1-like proteins), not only are the RRMs highly conserved, but also the length and sequence of the linking regions between RRMs (data not shown) (85). This phylogenetic conservation suggests that the linking regions are important and that individual RRMs in these proteins do not act independently but rather require interactions between, and precise positioning of, the RRMs. Indeed, a number of multi-RRM proteins show synergy between two or more of their RRMs for binding to RNA. Examples include yeast poly(A)-binding protein (49), $U2AF^{65}$ (7), and SF2/ASF (32). In the chloroplast RNA-binding protein cp29B, the 37 amino acid Gly-rich linker between the two RRMs is required for the synergistic effect on RNA binding (50).

The phylogenetic data support a model in which at least several of the major groups of RNA-binding proteins arose through a combination of gene duplication and intragenic domain duplication. The presence of RRMs in proteins from a variety of distant phyla, combined with the modular organization of multiple RRMs suggests that the entire domain, rather than simply the RNP-1 submotif, is an ancient conserved region (ACR) (90). After this dataset was assembled, the sequences of RRM-containing proteins from several species of cyanobacteria were submitted to Genbank (gp:120890, gp:120891, gp:120892) (M. Mulligan, personal communication). This is in contrast to our earlier searches, which failed to identify RRMs in any eubacteria (with the possible exception of bacteriophage ϕ29gp10, see above), despite the success of these sensitive searches in identifying a broad range of RRMs in eukaryotes. The presence of RRM-containing proteins in both cyanobacteria and eukaryotes underscores the ancient origin of the RRM.

Phylogeny and function of SR proteins. The SR family of phosphoproteins in metazoans is a striking example of gene duplication and phylogenetic conservation (15) (Table 2). Mouse X16, which is 100% identical at the amino acid level to human SRp20, is 62% identical to RBP1, a probable *Drosophila* SRp20 homolog (14,15,16). Human and avian SC35 (PR264) (9,10) are 98% homologous. An *Arabidopsis* SR protein is 58% identical to human SF2/ASF (59% without RS domain) and 47% (48% without RS domain) identical to *Drosophila* SRp55 (12) (for accession numbers see SR protein group in Table 1). Mouse (gb:x66091) and human SF2/ASF are 100% identical at the amino acid level and 95% identical at the DNA level within the coding region (data not shown). Table 2 gives an indication of the conservation in sequence between SR proteins, and the phylogeny of their RRMs is shown in Figure 4. The evolution of SR proteins probably involved complex events such as domain duplication and subsequent deletion, as well as extension of the RS domain. As a consequence, numerical pairwise homology scores cannot accurately reflect phylogenetic relationships among these proteins.

Partial amino acid sequences of human SRp40 and SRp75 show the presence of at least part of the second atypical RRM (15). However, an analogous RRM is absent from SC35 and X16, despite the fact that human SF2/ASF, *Drosophila* SRp55, and human SC35 have equivalent *in vitro* activities in general and alternative splicing, and bovine SRp40 and SRp75 have equivalent general splicing activity (15,20,21). The single RRM in SC35 and X16 probably substitutes for the two synergistic RRMs in the other SR proteins (32).

All SR proteins studied so far have both constitutive and alternative splicing activities in vitro (15,16,17–21). For example, human SF2/ASF and human SC35, which are only 38% identical have indistinguishable biochemical activities (21). In a comparison of human SF2/ASF and *Drosophila* RBP1 activities in human extracts, some qualitative and quantitative differences were reported, although the possibility that these differences are due to the use of mixed human and *Drosophila* factors has not been ruled out (16). Quantitative differences in splice site selection preferences among several SR proteins *in vitro* were recently reported (91). The *in vivo* expression of individual SR proteins has been studied in a few cases at the level of mRNA or protein (10,14,16,91; A. Hanamura and A.R.K., unpublished data). In each case, a wide range of expression was observed in different tissues or cell lines. In addition, the activities of these proteins may be regulated by phosphorylation or nuclear localization. The selective pressure to maintain such a high degree of sequence conservation between individual members of the SR family from different species is inconsistent with the apparently redundant biochemical activities of less homologous members from the same species. This strongly suggests that individual SR proteins have unique functions *in vivo*.

Analysis of motifs with low compositional complexity (RS and RGG domains)

Domains with repeated Arg−Ser dipeptides are often found in: (i) generic splicing factors, such as SR proteins (see Table 1 for accession numbers) and U2AF^{65} (humU2AF65=sw:ua2f__human); (ii) gene-specific splicing factors, such as Tra2 (droTRA2=sw:tra2__drome), Tra (droTRA=sw:trsf__drome), and Su(w^a) (droSU(WA)=sw:suwa__drome); (iii) U1-70K polypeptide from several species (see Table 1 under snRNP-associated proteins for accession numbers). RS domains so far are exclusive to known or suspected splicing or spliceosome-associated factors, with the possible exception of the E2 protein of some, but not all, isolates of human papillomaviruses (e.g. sw:ve2__bpv4) (data not shown). The *Drosophila* protein Suppressor of sable (gp:m57889), which has been implicated in alternative splicing regulation, also has a highly charged Arg-rich region, although with very few RS dipeptides (72). The RS domain is distinct from the Arg-motif present in several RNA-binding proteins, including bacterial transcription anti-terminators and HIV regulatory proteins (92). RS domains are found in a variety of positions in the above proteins, some of which also have RRMs (Figure 5). Although the Arg−Ser repeats are evident in all cases, some are embedded within other domains, and additional simple imperfect repeats are also common. It is unclear at present what constitutes a minimal RS domain, both from statistical relevance and protein structural standpoints. The RS domains of Su(w^a) and Tra have been shown to be responsible for *in vivo* localization of these proteins to the nucleoplasmic speckled region (31). The RS domains of U2AF^{65} and SF2/ASF have been shown to be required for constitutive splicing *in vitro* (7,32). In the case of SF2/ASF, both Arg and Ser residues are specifically required (32).

At least some of the Ser residues in several RS domains of SR proteins appear to be phosphorylated (15; A. Hanamura, R. Kobayashi, and A.R.K., unpublished data). This modification results in a pronounced change in the charge and probably the conformation of this domain. The RS domain in U1-70K is phosphorylated *in vitro* by a kinase activity associated with U1 snRNP (93). It is presently unclear whether phosphorylated or dephosphorylated RS domains are the active form of the domain in the above proteins. Whereas the highly basic RS domain can be expected to interact with the phosphate backbone of RNA, the state of Ser phosphorylation may influence these electrostatic interactions. In addition, RS domains may be involved in protein−protein interactions.

Another simple imperfect repeated domain found in many RNA-binding proteins is made up of repeated Gly−Gly dipeptides interspersed with Arg and aromatic residues. The RGG domain is ubiquitous in proteins involved in diverse aspects of RNA metabolism, such as pre-rRNA processing and pre-mRNA splicing factors, hnRNP polypeptides, and RNA helicases (33,34). This domain is distinct from segments made up of only consecutive Gly residues, which have been termed Gly hinges because of their flexibility, and are found for example in U1-70K and some SR proteins. In hnRNP U (gp:x65488), which lacks RRMs, the single C-terminal RGG domain as part of a fusion protein is sufficient for RNA binding (34). In nucleolin (humNUCL=sw:nucl__human), which contains four RRMs, a C-terminal fragment containing the RGG domain can also independently bind RNA (94). Likewise, in hnRNP A1 (humHNRNPA1=sw:roa1__human), which contains two RRMs, the C-terminal RGG domain mediates binding to poly(etheno-A) and is responsible for RNA-binding cooperativity (95). Cooperativity is thought to involve both protein−protein and protein−RNA interactions by the RGG domain of hnRNP A1 (95). A proteolytic product of hnRNP A1, known as UP1, which lacks the C-terminal RGG domain, has nucleic acid helix destabilizing activity (96), whereas the intact protein has nucleic acid annealing activity (97−99).

It is likely that in general the RGG domain is involved in RNA binding. However, the primary structure of the domain differs considerably among different proteins. Thus, in nucleolin and other nucleolar pre-RNA processing factors, a common repeat is Arg−Gly−Gly−Phe, whereas in hnRNP A1 only Gly−Gly dipeptides are evident. In addition, up to eight of the Arg residues in the RGG domain of nucleolin are dimethylated (100,101). Whether this modification changes the activity of the protein is unclear, but it would prevent electrostatic interactions with the phosphate backbone of RNA (102).

We performed BLAST (36), BLAZE (37) and BLITZ (38) searches of the GenPept and SWISS-PROT databases with RS and RGG domains derived from several proteins, as well as with idealized repeats. We did not identify any sequences with unreported RS or RGG domains. Although the expected proteins often produced high scores, the known positives and the known negatives did not separate well. This is probably an inherent problem with domains of such low sequence complexity and in addition the searches were hampered by the presence of similar but distinct repeats, such as Gly-rich repeats in keratin. Effective multiple alignments of these domains cannot be generated, thus preventing profile analysis. FASTA searches of the RRM-enriched database were effective in identifying RS and RGG domains (see above), due to the fact that this database is far more limited and enriched in relevant sequences.

Statistical analysis of RS domains. To characterize the RS domains further we used the program SAPS, which calculates the significance of various compositional values of proteins, including repeated peptides (41). As expected, all the proteins of the SR family had significant numbers of Arg−Ser or Ser−Arg dipeptides, and although there were statistically significant higher order repeats in these proteins (e.g., RxxxRx in X16), only the above repeats were common to all SR protein sequences (for accession numbers see SR protein group in Table 1). This repeated structure was not mirrored at the DNA level (data not shown). In contrast, U2AF^{65} (hum-U2AF65=sw:ua2f__human), Su(w^a) (droSU(WA)=sw:suwa__drome), Tra

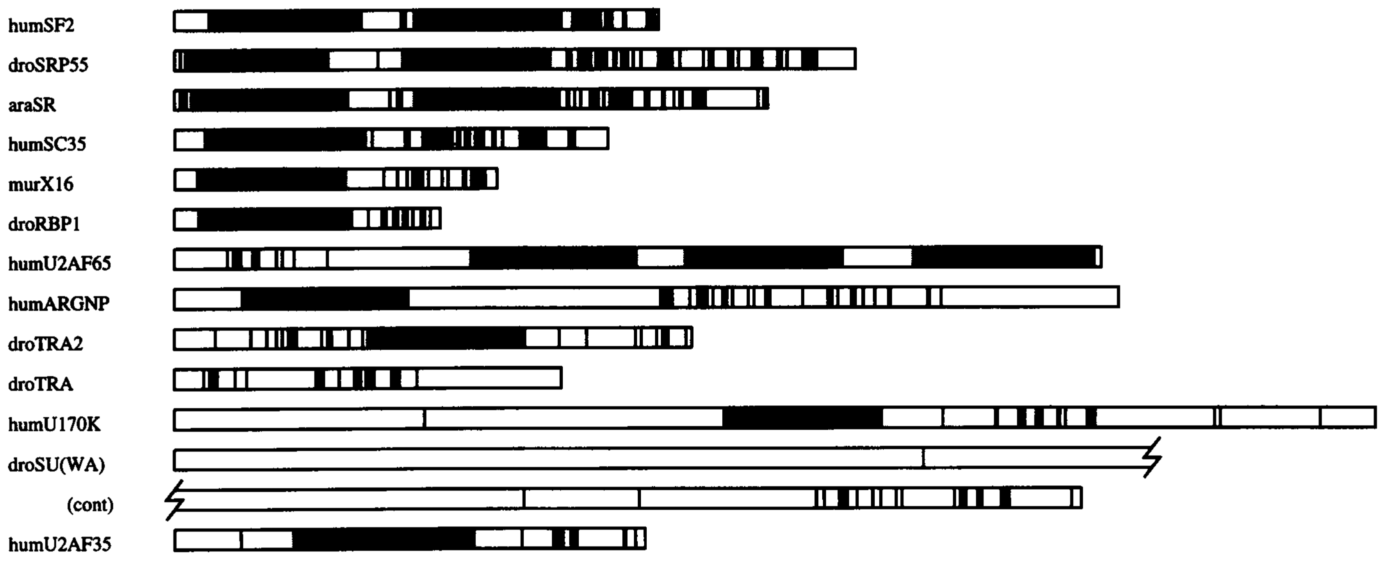


Figure 5. Distribution of RS or SR dipeptides and RRMs in proteins with RS domains. Each protein is represented by a horizontal box drawn to scale. Sequence names are as in Table 1. The *Drosophila* Su(w^{a}) protein (droSU(WA)) was split into two boxes due to its size. Thin vertical black lines represent a single RS or SR dipeptide. The thickness of the black lines is proportional to the number of consecutive RS or SR dipeptides. The peptides can overlap, so that XRSRX is represented by two lines. Grey shading represents RRMs, or in the case of U2AF^{35}, a region reminiscent of an RRM (see text). The RS domain of U1-70K is nested within a region high in RD and RE repeats, which are not shown. The RS domains in some of these proteins include regions with two or more consecutive Arg or two or more consecutive Ser residues, which are not represented in this diagram.

(droTRA=sw:trsf_drome) and Tra2 (droTRA2=sw:tra2_drome) lacked statistically significant repeats of Arg–Ser, but did contain significant repeats of either Rx or Sx. In all these proteins the RS domain is near one of the termini (Figure 5), and given its extremely charged nature, one would expect it to be solvent-exposed.

In U1-70K (humU170K=sw:ru17_human) the RS domain is embedded within an RD/E domain (103), and although Rx is a very significant repeat throughout the RD and RS domains, both of the repeats DR and RE were significant in the same region (data not shown). Without other knowledge, one would expect the function of this domain to be mediated by the RD repeats. Given that the Ser residues in this domain of U1-70K are heavily phosphorylated (93), the RS domain mimics the alternating charge structure of the surrounding RD domain. The RD and RS domains may have separate functions, in which case it is unclear if they have to be nested. Perhaps this RS domain acts like the surrounding RD domain, but in a manner that is subject to regulation by reversible Ser phosphorylation. Another possibility, since the RS domain of U1-70K is uncharacteristically far from the protein termini (Figure 5), is that the RD domain serves to ensure that the RS domain is solvent-exposed and flexible. Interestingly, RS and RD domains are lacking in the *S. cerevisiae* homolog of U1-70K (sacU170K=gp:x59986) (65,66).

The *S. cerevisiae* YCL11c protein (sacYCL11C=sw:ycb1_yeast), for which we previously noted an architectural similarity to human U2AF^{65} (63), did not have significant repeats of either Rx or Sx. In addition, Suppressor of sable (gp:m57889) did not contain significant repeats of either Rx or Sx or (R/K)x. The highly charged region in Suppressor of sable is a very small region in a large protein with many other highly distinctive regions of low sequence complexity.

Sequence criteria for SR proteins. Given that several proteins, in addition to SR proteins, contain both RS domains and RRMs (Figure 5), what structural features distinguish SR proteins? Currently sequenced SR proteins are characterized by an N-terminal RRM and an extensive C-terminal RS domain. The RS domains of these proteins are rich in consecutive RS dipeptides, in contrast to other proteins, in which Arg and Ser residues are dispersed and show less periodicity. The RRM is characterized by a partially conserved octapeptide (EFEDxRDA) that overlaps the RNP-1 submotif (see above). Several, though not all, SR proteins contain a distinctive atypical central RRM, which includes a conserved heptapeptide (SWQDLKD) (see above).

While this manuscript was in preparation, the full sequences of human SRp75 (104) and HRS (105) were published. We note that HRS appears to be identical to human SRp40, based on the reported partial amino acid sequence of the latter (9). These sequences were not retrieved in our searches because they were not available in the databases at the time. Both sequences fully satisfy the above criteria for SR proteins (data not shown). The conserved SWQDLKD heptapeptide appears to be an invariant signature for all SR proteins that contain the central atypical RRM, including SRp75 and HRS, in addition to the proteins shown in Figure 3.

Statistical analysis of RGG domains. A similar SAPS analysis (41) of RGG domains showed limited significant similarities between different proteins. Often GG or GGx, or other spacings of Gly were found to be significant. Gly repeats that also involved Arg or aromatic residues were seldom found to be statistically significant. Within the Gly-rich domain found at the C-terminus of hnRNP A1 from several species, only Gly repeats were significant, and no other repeats were common to all these proteins (data not shown).

The statistical importance of the Gly residues in these repeats is consistent with structural data for nucleolin (humNUCL=sw:nucl_human) (94). Circular dichroism and Fourier transform infrared spectroscopy studies of the RGG domain of nucleolin are consistent with a secondary structure of repeated β-turns stacked together to form a β-spiral (94). Computer-modeling studies of β-spirals indicate that these structures are very flexible (94,106). The fact that each Gly–Gly repeat constitutes an independent unit in the β-spiral model is consistent with our finding that no alignment is possible among the sequences of RGG domains of nucleolin, fibrillarin, and hnRNP A1-type proteins, although all contain aromatic and Arg residues interspersed with the Gly–Gly dipeptides. Since the β-spiral model tolerates other residues, including Gly, outside the Gly–Gly repeats, it is difficult to derive statistically significant consensus repeats. In summary, the common features of all these domains consist of their position near one terminus of the protein, Gly-richness, the presence of few acidic residues, and usually a repeated pattern of Gly residues.

A comprehensive RRM database on a file-server

A comprehensive and frequently updated table of RRM sequences is available on a file-server by sending an e-mail message to RRM@molbiol.ox.ac.uk with the words SEND RRM in the body of the message. Every effort has been made to ensure that this table is comprehensive and non-redundant. Comments, queries, and new or missing sequences are very welcome; send e-mail to birney@molbiol.ox.ac.uk

ACKNOWLEDGEMENTS

We are grateful to R.J. Roberts, J. Pflugrath, and D.S. Horowitz for helpful comments on the manuscript; P.R. Evans for providing the U1-A structural coordinates; J. Felsenstein for advice on phylogenetic analysis; C. Shimoda, P. Silver, and M. Mulligan for communicating results prior to publication; D. Higgins for providing modified Clustal V software prior to its release, and for advice; S. Goodbourn, C. Rawlings, J. Rees, C. Edge, and J. Pflugrath for sharing computing facilities and for advice; L. Cowe and B. O'Connor for assistance with GCG software; and D. Frendewey for helpful discussions. E.B. was supported in part by a fellowship from Baring Brothers and by the Cold Spring Harbor Undergraduate Research Program. S.K. was supported by postdoctoral fellowship grant GM15262 from the N.I.H. A.R.K. is a Pew Scholar in the Biomedical Sciences. This work was supported by grant GM42699 from the N.I.H.

REFERENCES

1. Moore,M.J., Query,C.C. and Sharp,P.A. (1993) In Gesteland,R.F. and Atkins,J.F. (eds.), The RNA World. Cold Spring Harbor Laboratory Press, New York, pp. 303–357.
2. Rio,D.C. (1992) Gene Expr. **2**, 1–5.
3. Dreyfuss,G., Matunis,M.J., Piñol-Roma,S. and Burd,C.G. (1993) Annu. Rev. Biochem. **62**, 289–321.
4. Lührmann,R., Kastner,B. and Bach,M. (1990) Biochim. Biophys. Acta **1087**, 265–292.
5. Krainer,A.R., Mayeda,A., Kozak,D. and Binns,G. (1991) Cell **66**, 383–394.
6. Ge,H., Zuo,P. and Manley,J.L. (1991) Cell **66**, 373–382.
7. Zamore,P.D., Patton,J.G. and Green,M.R. (1992) Nature **355**, 609–614.
8. Sailer,A., MacDonald,N.J. and Weissmann,C. (1992) Nucleic Acids Res. **20**, 2374.
9. Vellard,M., Sureau,A., Soret,J., Martinerie,C. and Perbal,B. (1992) Proc. Natl. Acad. Sci. USA **89**, 2511–2515.
10. Fu,X.-D. and Maniatis,T. (1992) Science **256**, 535–538.
11. Patton,J.G., Porro,E.B., Galceran,J., Tempst,P. and Nadal-Ginard,B. (1993) Genes Dev. **7**, 393–406.
12. Roth,M.B., Zahler,A.M. and Stolk,J.A. (1991) J. Cell Biol. **115**, 587–596.
13. Champlin,D.T., Frasch,M., Saumweber,H. and Lis,J.T. (1991) Genes Dev. **5**, 1611–1621.
14. Ayane,M., Preuss,U., Kohler,G. and Nielsen,P.J. (1991) Nucleic Acids Res. **19**, 1273–1278.
15. Zahler,A.M., Lane,W.S., Stolk,J.A. and Roth,M.B. (1992) Genes Dev. **6**, 837–847.
16. Kim,Y.-J., Zuo,P., Manley,J.L. and Baker,B.S. (1992) Genes Dev. **6**, 2569–2579.
17. Krainer,A.R., Conway,G.C. and Kozak,D. (1990) Genes Dev. **4**, 1158–1171.
18. Krainer,A.R., Conway,G.C. and Kozak,D. (1990) Cell **62**, 35–42.
19. Ge,H. and Manley,J.L. (1990) Cell **62**, 25–34.
20. Mayeda,A., Zahler,A.M., Krainer,A.R. and Roth,M.B. (1992) Proc. Natl. Acad. Sci. USA **89**, 1301–1304.
21. Fu,X.-D., Mayeda,A., Maniatis,T. and Krainer,A.R. (1992) Proc. Natl. Acad. Sci. USA **89**, 11224–11228.
22. Choi,Y.D., Grabowski,P.J., Sharp,P.A. and Dreyfuss,G. (1986) Science **231**, 1534–1539.
23. Sierakowska,H., Szer,W., Furdon,P.J. and Kole,R. (1986) Nucleic Acids Res. **14**, 5241–5254.
24. Mayeda,A. and Krainer,A.R. (1992) Cell **68**, 365–375.
25. Mattox,W., Ryner,L. and Baker,B.S. (1992) J. Biol. Chem. **267**, 19023–19026.
26. Bandziulis,R.J., Swanson,M.S. and Dreyfuss,G. (1989) Genes Dev. **3**, 431–437.
27. Kenan,D.J., Query,C.C. and Keene,J.D. (1991) Trends Biochem. Sci. **16**, 214–220.
28. Nagai,K. (1992) Current Biol. **2**, 131–137.
29. Haynes,S.R. (1992) New Biol. **4**, 421–429.
30. Mattaj,I.W. (1993) Cell **73**, 837–840.
31. Li,H. and Bingham,P.M. (1991) Cell **67**, 335–342.
32. Cáceres,J. and Krainer,A.R. (1993) EMBO J., in press.
33. Steinert,P.M., Mack,J.W., Korge,B.P., Gan,S.Q., Haynes,S.R. and Steven,A.C. (1991) Int. J. Biol. Macromol. **13**, 130–139.
34. Kiledjian,M. and Dreyfuss,G. (1992) EMBO J. **11**, 2655–2664.
35. Pearson,W.R. and Lipman,D.J. (1988) Proc. Natl. Acad. Sci. USA **85**, 2444–2448.
36. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) J. Mol. Biol. **215**, 403–410.
37. Brutlag,D.L., Dautricourt,J.P., Diaz,R., Fier,J., Moxon,B. and Stamm,R. (1993) Computers Chem., in press.
38. Sturroc,S.S. and Collins,J.F. (1993) BLITZ MPsren version 1.3 Biocomputing Research unit, University of Edinburgh, UK.
39. Gribskov,M.. McLachlan,A.D. and Eisenberg,D. (1987) Proc. Natl. Acad. Sci. USA **84**, 4355–4358.
40. Devereux,J., Haeberli,P. and Smithies,O. (1984) Nucleic Acids Res. **12**, 387–395.
41. Brendel,V., Bucher,P., Nourbakhsh,I.R., Blaisdell,B.E. and Karlin,S. (1992) Proc. Natl. Acad. Sci. USA **89**, 2002–2006.
42. Saitou,N. and Nei,M. (1987) Mol. Biol. Evol. **4**, 406–425.
43. Higgins,D.G. and Sharp,P.M. (1988) Gene **73**, 237–244.
44. Felsenstein,J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Dept of Genetics, University of Washington, Seattle.
45. Felsenstein,J. (1989) Cladistics **5**, 164–166.
46. Kimura,M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, U.K.
47. Felsenstein,J. (1985) Evolution **39**, 783–791.
48. Scherly,D., Kambach,C., Boelens,W., van Venrooij,W.J. and Mattaj,I.W. (1991) J. Mol. Biol. **219**, 577–584.
49. Burd,C.G., Matunis,E.L. and Dreyfuss,G. (1991) Mol. Cell. Biol. **11**, 3419–3424.
50. Ye,L. and Sugiura,M. (1992) Nucleic Acids Res. **20**, 6275–6279.
51. Lutz-Freyermuth,C., Query,C.C. and Keene,J.D. (1990) Proc. Natl. Acad. Sci. USA **87**, 6393–6397.
52. Jessen,T.H., Oubridge,C., Teo,C.H., Pritchard,C. and Nagai,K. (1991) EMBO J. **10**, 3447–3456.
53. Boelens,W., Scherly,D., Jansen,E.J., Kolen,K., Mattaj,I.W. and van Venrooij,W.J. (1991) Nucleic Acids Res. **19**, 4611–4618.
54. Scherly,D., Boelens,W., Dathan,N.A., van Venrooij,W.J. and Mattaj,I.W. (1990) Nature **345**, 502–506.
55. Bentley,R.C. and Keene,J.D. (1991) Mol. Cell. Biol. **11**, 1829–1839.
56. Nagai,K., Oubridge,C., Jessen,T.H., Li,J. and Evans,P.R. (1990) Nature **348**, 515–520.

57. Harper,D.S., Fresco,L.D. and Keene,J.D. (1992) Nucleic Acids Res. **20**, 3645–3650.
58. Hoffman,D.W., Query,C.C., Golden,B.L., White,S.W. and Keene,J.D. (1991) Proc. Natl. Acad. Sci. USA **88**, 2495–2499.
59. Wittekind,M., Gorlach,M., Friedrichs,M., Dreyfuss,G. and Mueller,L. (1992) Biochemistry **31**, 6254–6265.
60. Gorlach,M., Wittekind,M., Beckman,R.A., Mueller,L. and Dreyfuss,G. (1992) EMBO J. **11**, 3289–3295.
61. Merrill,B.M., Stone,K.L., Cobianchi,F., Wilson,S.H. and Williams,K.R. (1988) J. Biol. Chem. **263**, 3307–3313.
62. Query,C.C., Bentley,R.C. and Keene,J.D. (1989) Cell **57**, 89–101.
63. Birney,E., Kumar,S. and Krainer,A.R. (1992) Nucleic Acids Res. **20**, 4663.
64. Piñol-Roma,S., Swanson,M.S., Gall,J.G. and Dreyfuss,G. (1989) J. Cell Biol. **109**, 2575–2587.
65. Smith,V., Barrell,B.G. (1991) EMBO J. **10**, 2627–2634.
66. Kao,H.Y. and Siliciano,P.G. (1992) Nucleic Acids Res. **20**, 4009–4013.
67. Chambers,J.C., Kenan,D., Martin,B.J. and Keene,J.D. (1988) J. Biol. Chem. **263**, 18043–18051.
68. Chan,E.K., Sullivan,K.F. and Tan,E.M. (1989) Nucleic Acids Res. **17**, 2233–2244.
69. Brennan,C.A. and Platt,T. (1991) J. Biol. Chem. **266**, 17296–17305.
70. Grimes,S. and Anderson,D. (1990) J. Mol. Biol. **215**, 559–566.
71. Rebagliati,M. (1989) Cell **58**, 231–232.
72. Voelker,R.A., Gibson,W., Graves,J.P., Sterling,J.F. and Eisenberg,M.T. (1991) Mol. Cell. Biol. **11**, 894–905.
73. Landsman,D. (1992) Nucleic Acids Res. **20**, 2861–2864.
74. Schindelin,H., Marahiel,M. and Heinemann,U. (1993) Nature **364**, 164–168.
75. Schnuchel,A., Wiltscheck,R., Czisch,M., Herrier,M., Willimsky,G., Graumann,P., Marahiel,M. and Holak,T. (1993) Nature **364**, 169–171.
76. Kim,Y.-J. and Baker,B.S. (1993) Mol. Cell. Biol. **13**, 174–183.
77. Shamoo,Y., Ghosaini,L.R., Keating,K.M., Williams,K.R., Sturtevant,J.M. and Konigsberg,W.H. (1989) Biochem. **28**, 7409–7417.
78. Chaudhary,N., McMahon,C. and Blobel,G. (1991) Proc. Natl. Acad. Sci. USA **88**, 8189–8193.
79. Huang,S. and Spector,D.L. (1992) Curr. Biol. **2**, 188–190.
80. Shimoda,C., Uehira,M., Kishida,M., Fujioka,H., Iino,Y., Watanabe,Y. and Yamamoto,M. (1987) J. Bacteriol. **169**, 93–96.
81. Bossie,M.A., DeHoratius,C., Barcelo,G. and Silver,P. (1992) Mol. Biol. Cell **3**, 875–893.
82. Zhang,M., Zamore,P.D., Carmo-Fonseca,M., Lamond,A.I. and Green,M.R. (1992) Proc. Natl. Acad. Sci. USA **89**, 8769–8773.
83. Dreyfuss,G., Swanson,M.S., Piñol-Roma,S. (1988) Trends Biochem. Sci. **13**, 86–91.
84. Good,P.J., Rebbert,M.L. and Dawid,I.B. (1993) Nucleic Acids Res. **21**, 999–1006.
85. Matunis,E.L., Matunis,M.J. and Dreyfuss,G. (1992) J. Cell Biol. **116**, 257–269.
86. Ye,L., Li,Y., Fukami-Kobayashi,K., Go,M., Konishi,T., Watanabe,A. and Sugiura,M. (1991) Nucleic Acids Res. **19**, 6485–6490.
87. Szabo,A., Dalmau,J., Manley,G., Rosenfeld,M., Wong,E., Henson,J., Posner,J.B. and Furneaux,H.M. (1991) Cell **67**, 325–333.
88. Kim,Y.-J. and Baker,B.S. (1993) J. Neurosci. **13**, 1045–1056.
89. Richter,K., Good,P.J. and Dawid,I.B. (1990) New Biol. **2**, 556–565.
90. Green,P., Lipman,D., Hillier,L., Waterston,R., States,D. and Claverie,J.M. (1993) Science **259**, 1711–1716.
91. Zahler,A.M., Neugebauer,K.M., Lane,W.S. and Roth,M.B. (1993) Science **260**, 219–222.
92. Lazinski,D., Grzadzielska,E. and Das,A. (1989) Cell **59**, 207–218.
93. Woppmann,A., Patschinsky,T., Bringmann,P., Godt,F. and Lührmann,R. (1990) Nucleic Acids Res. **18**, 4427–4438.
94. Ghisolfi,L., Joseph,G., Amalric,F. and Erard,M. (1992) J. Biol. Chem. **267**, 2955–2959.
95. Kumar,A., Casas-Finet,J.R., Luneau,C.J., Karpel,R.L., Merrill,B.M., Williams,K.R. and Wilson,S.H. (1990) J. Biol. Chem. **265**, 17094–17100.
96. Karpel,R.L. and Burchard,A.C. (1980) Biochem. **19**, 4674–4682.
97. Pontius,B.W. and Berg,P. (1990) Proc. Natl. Acad. Sci. USA **87**, 8403–8407.
98. Kumar,A. and Wilson,S.H. (1990) Biochem. **29**, 10717–10722.
99. Munroe,S.H. and Dong,X.F. (1992) Proc. Natl. Acad. Sci. USA **89**, 895–899.
100. Lischwe,M.A., Cook,R.G., Ahn,Y.S., Yeoman,L.C. and Busch,H. (1985) Biochem. **24**, 6025–6028.
101. Lapeyre,B., Amalric,F., Ghaffari,S.H., Rao,S.V., Dumbar,T.S. and Olson,M.O. (1986) J. Biol. Chem. **261**, 9167–9173.
102. Calnan,B.J., Tidor,B., Biancalana,S., Hudson,D. and Frankel,A.D. (1991) Science **252**, 1167–1171.
103. Mancebo,R., Lo,P.C. and Mount,S.M. (1990) Mol. Cell. Biol. **10**, 2492–2502.
104. Zahler,A.M., Neugebauer,K.M., Stolk,J.A. and Roth,M.B. (1993) Mol. Cell. Biol. **13**, 4023–4028.
105. Diamond,R.H., Du,K., Lee,V.M., Mohn,K.L., Haber,B.A., Tewari,D.S. and Taub,R. (1993) J. Biol. Chem. **268**, 15185–15192.
106. Matsushima,N., Creutz,C.E. and Kretsinger,R.H. (1990) Proteins **7**, 125–155.