MOLECULAR AND CELLULAR BIOLOGY, Jan. 1993, p. 174–183
0270-7306/93/010174-10$02.00/0

Vol. 13, No. 1

Isolation of RRM-Type RNA-Binding Protein Genes and the Analysis of Their Relatedness by Using a Numerical Approach

YOUNG-JOON KIM AND BRUCE S. BAKER*

Department of Biological Sciences, Stanford University, Stanford, California 94305

Received 31 July 1992/Returned for modification 9 October 1992/Accepted 28 October 1992

Proteins with RNA recognition motifs (RRMs) have important roles in a great many aspects of RNA metabolism. However, this family has yet to be systematically studied in any single organism. In order to investigate the size of the RRM gene family in *Drosophila melanogaster* and to clone members of this family, we used a polymerase chain reaction (PCR) with highly degenerate oligonucleotides to amplify DNA fragments between the RNP-1 and RNP-2 consensus sequences of the RRM proteins. Cloning and sequencing of 124 PCR products revealed 12 different RRM sequences (RRM1 to RRM12). When PCR products were used as probes in genomic Southern and Northern (RNA) analyses, 16 restriction fragments and 25 transcripts, respectively, were detected. Since the combinations of nucleotide sequences represented in the PCR primers correspond to only 4% of the RRM sequences inferred to be possible from known RRM sequences, we estimate the size of the RRM gene family in the order of three hundred genes in flies. In order to gain insight into the possible functions of the genes encoding the RRMs, we analyzed the sequence similarities between the 12 RRMs and 62 RRM sequences of known proteins. This analysis showed that the RRMs of functionally related proteins have similar sequences and are clustered together in the RRM gene tree. On the basis of this observation, the RRMs can be divided into three groups: a heterogeneous nuclear ribonucleoprotein type, a splicing regulator type, and a development-specific factor type. This result suggests that we have isolated good candidates for both housekeeping and developmentally important genes involved in RNA metabolism.

In eukaryotes, there is substantial posttranscriptional metabolism of RNA. The major events of RNA metabolism include capping, polyadenylation, splicing (and alternative splicing), transport, localization, translation, and turnover (for reviews, see references 6, 7, 10, and 20). In all of these processes, *trans*-acting factors must recognize particular parts of RNA molecules and, in many cases, particular species of RNA molecules. Work during the last few years has shown that many, but by no means all, of the proteins that bind to RNA share a loosely conserved domain of about 80 to 90 amino acids referred to as an RNA recognition motif (RRM) (also referred to as RNP-CS type RNA-binding domain, RNP-80, and RNP motif; for reviews, see references 3, 14, 23, and 34).

RRM is a loosely conserved RNA binding domain which has 21 conserved amino acid residues spread across an 80- to 90-amino-acid region, with the most conserved sequences being the RNP-1 octapeptide and RNP-2 hexapeptide (14). Recently, the three-dimensional structure has been worked out for the RRM of U1A protein (21, 37), and this structure seems to be conserved in other members of the RRM gene family (23).

Studies of several RRM-containing proteins have shown that this motif can confer the ability to bind single-stranded nucleic acids, although for most proteins in this family, experimental evidence for RNA binding is currently lacking. A proteolytic fragment of the heterogeneous nuclear ribonucleoprotein (hnRNP) A1 protein containing only two RRM repeats retained RNA binding activity (29, 40). The region that is responsible for RNA binding in members of this family has been also defined by protein truncation experiments. For example, only one motif of the four repeated RRMs present in the poly(A)-binding protein of *Saccharomyces cerevisiae* is enough to retain poly(A) binding activity (46). Similarly, studies of the U1 70K protein (42) and the U1A protein (51) have shown that a single RRM in these proteins with minimal flanking sequences retains the ability to specifically bind U1 RNA. Within the RRM, the RNP-1 consensus sequence seems to be involved in RNA binding. UV cross-linking (35) and filter binding assays (52) have shown that RNP-1 is directly involved in single-stranded nucleic acid binding. In addition, Scherly et al. (50) showed that in exchanges of RNP-1 between the U1A and U2B″ proteins, the binding specificity of these proteins follows the sequences around RNP-1. These data suggest that the binding specificities of at least some proteins in this family reside in RRM, rather than in unique flanking sequences. However, studies of the RNA binding affinities of the La-encoded (9) and Ro-60K (12) proteins show that sequences outside the RRM are also required for RNA binding in some members of this family.

While many members of this family, such as hnRNP proteins (11, 30, 56), small nuclear ribonucleoprotein (snRNP) proteins (18, 54, 57), and poly(A)-binding proteins (45), carry out functions in basic housekeeping aspects of RNA metabolism, other RRM-containing genes have developmentally important regulatory roles. For example, the *elav* gene of *Drosophila melanogaster* contains RRMs (43) and has been shown to encode a nervous system-specific function necessary for the differentiation of neural cells (44). In addition, two of the regulatory genes that control sexual differentiation in *D. melanogaster*, Sex-lethal (*Sxl*) and transformer-2 (*tra-2*), also contain RRMs (1, 4, 16). Both the *Sxl* and *tra-2* genes control the splicing patterns of the pre-mRNAs produced by other genes in the regulatory

* Corresponding author.

hierarchy governing sexual differentiation (2, 38). Since RRM-containing genes carry out a diverse array of housekeeping and regulatory functions, the consensus sequence that defines this family probably represents a general RNA binding domain.

The fact that *Sxl*, *tra-2*, and the recently identified human splicing factor ASF (15) or SF2 (27) regulate alternative splicing events and encode proteins with sequences that place them in this family of RRM-containing proteins is particularly intriguing, since few *trans*-acting factors that regulate alternative splicing decisions have been identified. In contrast to the small number of genes known to regulate alternative splicing decisions, many genes that are regulated by alternative splicing are known (for reviews, see references 8 and 55). Thus, there are probably many, as yet unidentified, *trans*-acting factors that control these alternative splicing decisions.

Indeed, for many aspects of RNA metabolism, only a small fraction of the gene products that are involved in these processes is currently known. Since many of the known proteins that interact with RNA contain RRM sequences, we reasoned that the isolation and characterization of new members of the RRM family would increase our understanding of RNA metabolism. To estimate the size of this gene family in *D. melanogaster* and to isolate new sequences encoding proteins with RRMs, we used the polymerase chain reaction (PCR) (47) to amplify RRM sequences. Twelve new sequences encoding RRMs have been isolated. This brings the known number of RRM-containing proteins in *D. melanogaster* to nearly 20. A sequence comparison and a numerical analysis of the RRM sequences suggest that functionally related RRM-containing proteins have significant sequence similarities in their RRMs, and thus it may be possible to obtain clues as to the functions of uncharacterized RRM proteins by their similarity to RRM proteins of known function.

MATERIALS AND METHODS

RNA preparation and cDNA synthesis. Adult flies (2 g) were homogenized in a Dounce homogenizer in 10-ml of homogenization buffer (0.15 M NaCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl [pH 8.0], 0.5% sodium dodecyl sulfate). The supernatant was extracted twice with phenol and then ethanol precipitated. The pellet was resuspended in water and frozen at −70°C. Poly(A)$^+$ RNA (20 μg) was purified by using oligo(dT)-cellulose column chromatography. cDNA was synthesized by adding 5 μg of poly(A)$^+$ RNA, 5 μg of oligo(dT), 20 U of Moloney murine leukemia virus reverse transcriptase, and 20 U of RNAsin to 50 μl of a solution consisting of 50 mM Tris-Cl (pH 8.0), 5 mM $MgCl_2$, 5 mM dithiothreitol, and 50 mM KCl.

PCR amplification of RNA recognition motifs. cDNA (100 ng) was amplified via PCR in 100-μl volumes of a solution consisting of 50 mM KCl, 10 mM Tris-Cl (pH 8.3), 1.5 mM $MgCl_2$, 0.01% (wt/vol) gelatin, 0.45 mM (each) of the four deoxynucleoside triphosphates, 1 nmol (each) of the degenerate primers, and 2.5 U of *Taq* polymerase. For the first 5 cycles, the reaction was incubated for 1 min (each) at 95, 48, and 74°C; the annealing temperature was shifted from 48 to 55°C for the remaining 25 cycles. After amplification, 10 μl of the PCR was analyzed by 2% agarose gel electrophoresis. The rest of the PCR was digested with *Eco*RI and *Hin*dIII and then cloned into a pSK vector (Stratagene). Single-stranded DNAs were prepared from the clones and sequenced by the method of Sanger et al. (49).

Northern and genomic Southern analysis. Samples (2 μg) of poly(A)$^+$ RNA of adult flies or of restriction enzyme-digested *D. melanogaster* genomic DNA was electrophoresed and transferred to Nytran membranes (Schleicher and Schuell). Hybridizations were done using the conditions described by Nagoshi et al. (38).

Computer analysis. The sequences of 47 genes with RRMs were selected from the NBRF/PIR protein database (release 20) by using the PROFILE SEARCH program of the University of Wisconsin Genetics Computer Group (13) and saved as a local data file named RRMdata. Each RRM sequence was compared with the sequences in the RRMdata file with the FASTA program of the University of Wisconsin Genetics Computer Group looking for similar sequences. The means and the standard deviations of the similarity scores for each comparison were calculated to show the significance of the similarity.

In order to construct a RRM gene tree, a sequence distance matrix was calculated by using the DNA mutational distance between different amino acids (26). Insertions and deletions at each amino acid position were treated as three mutational steps. The topology of the RRM gene tree was calculated from the distance matrix using the neighbor-joining algorithm (48).

RESULTS

Probes for RNA recognition motifs. The approach we have taken to estimate the size of the RRM gene family and to clone members of this family is based on PCR. RRM is a region of 80 to 90 amino acids, across which there are about 21 conserved residues in the 19 family members identified at the time we initiated this work (Fig. 1). The number of RRMs in a given protein ranges from one to four, and thus 32 RRM sequences were available to us. Clusters of conserved residues occur at only two positions in RRM (Fig. 1) and represent the only portions of RRM where there is enough conservation to design probes for oligonucleotide hybridization screens and/or primers to amplify sequences from members of this family via PCR. These two relatively conserved subdomains are RNP-1, which is an octapeptide, and RNP-2, which is a hexapeptide (14). However, there is little, if any, constraint on the amino acid in the seventh position in RNP-1. Thus, in designing degenerate oligonucleotide pools corresponding to this region, we have focused on the first six residues in RNP-1, the consensus sequence of which is depicted in Fig. 2a. This consensus amino acid sequence represents between 53 and 95% of the residues found at these six positions in the RNP-1 sequences depicted in Fig. 1. Assuming that all combinations of known amino acid residues at each position in RNP-1 are possible, this consensus sequence encompasses 29% of possible RNP-1 sequences. (Statistical calculations show that among the 32 RRMs available at the time we started this study, the numbers of RNP-1 and RNP-2 with 0, 1, 2, etc. deviations from the consensus sequence fit random expectations.) RNP-2 is a sequence of six amino acids at the 5′ end of the RRM and is less conserved than RNP-1. The most common amino acids found at each position of the RNP-2 (Fig. 2a) represent between 58 and 84% of the known amino acid residues at these six positions. Again, assuming all combinations of known residues are possible, this consensus sequence represents about 14% of the theoretically possible RNP-2 sequences.

The DNA sequences for RNP-1 and RNP-2 obtained by reverse translation of the amino acid consensus sequences

```
HUMAN    U1snRNP-70K      TLFVARV-NYDTTESK---LRREFEVYGPIKRIHMVYSKRS-GKPRGYAFIEYEHERDMHSAY-KHADGKKIDGRR-VLVDVERGRT
         U1snRNP-A#1      TIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVS----RSLKMRGQAFVIFKEVSSATNAL-RSMQGFPFYDKP-MRIQYAKTDS
                -A#2      ILFLTNL-PEETNELM---LSMLFNQFPGFKEVRLVP----G--RHDIAFVEFDNEVQAGAAR-DALQGFKITQNNAMKISFAKK
         U2snRNP-B"#1     TIYINNMNDKIKKEELKRSLYALFSQFGHVVDIVAL----KTMKMRGQAFVIFKELGSSTNAL-RQLQGFPFYGKP-MRIQYAKTDS
                -B"#2     ILFLNNL-PEETNEMM---LSMLFNQFPGFKEVRLVP----G--RHDIAFVEFENDGQAGAAR-DALQGFKITPSHAMKITYAKK
         RoRNP-60K        MLFAKAI-CSQCSDIST--KQAAFKAVSEVCRIP----------THLFTFIQFKKDLK-ESMK-CGMNGRALRKAI-AD-WYNEKGG
         UP2              KIFVGGL-SPDTPEEK---IREYFGGFGEVESIELPMDNKTN-KRRGFCFITFNQEEP
         hnRNP C1/C2      RVFIGNL-NTLVVKKSD--VEAIFSKYGKIVGCSV---------HKGFAFVQYVNERNARAAV-AGEDGRMIAGQV-LDINLAAEPK
         La               SVYIKGFPTDATLDD----IKEWLEDKGQVLNIQMRRTL--HKAFKGSIFVVFD---SIESAK-KFVETPGQKYKE-TDLLILFKDD
HAMSTER  C23              TLFVKGLSEDTTEET----LKGSF--EGSV-RARIVTDRET-GSSKGFGFVDFNSEEDAKAAK-EAMEDGEIDGNK-VTLDYAKPKG
RAT      HDP #1           KLFIGGL-SFETTDES---LRSHFEQWGTLTDCVVMRDPNT-KRSRGFGFVTYATVEEVDAAM-NAR-PHKVDGRV-VEPKRAVSRE
             #2           KIFVGGI-KED-TEEHH--LRDYFEQYGKIEVIEIMTDRGS-GKKRGFAFVTFDDHDSVDKIV-IQK-YHTVNGHN-CEVRKALSKQ
MOUSE    NUCLEOLIN #1     NIFIGNLNPNKSVAELKVAISEPFAKN-DLAVVDV----RT-GTNRKFGYVDFESAEDLEKAL--ELTGLKVFGNE-IKLEKPKPKG
                   #2     TLLAKNL-SFNITEDE---LKEVFEDALEIRLVS-----QD-GKSKGIAYIEFKSEADAEKNL-EEKQGAEIDGRS-VSLYYTGEKG
                   #3     TLVLSNL-SYSATEET---LQEVFEKATFIKVPQ----NQQ-GKSKGYAFIEFASFEDAKEAL-NSCNKMEIEGRT-IRLELQGPRG
                   #4     TLFVKGL-SEDTTEET---LKESFE--GSVRARIV-TDRET-GSSKGFGFVDFNSEEDAKAAK-EAMEDGEIDGNK-VTLDWAKPKG
FLY      P9 #1            KLFIGGL-DYRTTDEN---LKAHFEKWGNIVDVVVMKDPRT-KRSRGFGFITYSHSSMIDEAQ-KSRP-HKIDGRV-VEPKRAVPRQ
            #2            KLFVGAL-KDDHDEQS---IRDYFQHFGNIVDINIVIDKET-GKKRGFAFVEFDDYDPVDKVV-LQKQ-HQLNGKM-VDVKKALPKQ
         ELAV #1          SLFSSVG-EIESVK-L---IRDKSQVYIDPLNPQAPSK----GQSLGYGFVNYVRPQDAEQAV-NVLNGLRLQNKT-IKVSFARPSS
              #2          NLYVSGL-PKTMTQQE---LEAIFAPFGAIITSRILQNAGNDTQTKGVGFIRFDKREEATRAIIALNGTTPSSCTDPIVVKFSNTPG
              #3          PIFIYNL-APETEEAA---LWQLFGPFGAVQSVKIVKDPTT-NQCKGYGFVSMTNYDEAAMAI-RALNGYTM-GNRVLQVSFKTNKA
         SXL #1           NLIVNYL-PQDMTDRE---LYALFRAIGPINTCRIMRDYKT-GYSFGYAFVDFTSEMDSQRAI-KVLNGITVRNKR-LKVSYARPGG
             #2           NLYVTNL-PRTITDDQ---LDTIFGKYGSIVQKNILRDKLT-GRPRGVAFVRYNKREEAQEAI-SALNNVIPEGGSQPLSVRLAEEH
         tra-2            GVFGLNT-NTSQHK-----VRELFNKYGPIERIQMVIAQTR-QRSRGFCFIYFEKLSDARAAK-DSCSGIEVDGRR-IRVDFSITQR
         U1 70K snRNP     TLFIARI-NYDTSESK---LRREFEFYGPIKKIVLIHDQ-ESGKPKGYAFIEYEHERDMHAAY-KHADGKKIDSKR-VLVDVERART
YEAST    SSB              AEFFGTD-ADSISLPM---RKMRDQHTGRIF--------TSDSANRGMAFVTFSGENVDIEAKAEEFKGK-VFGDRELTVDVAVIRP
         PABP #1          SLYVGDL-EPSVSEAH---LYDIFSPIGSV-SSIRVCRDAITKTSLGYAYVNFNDHEAGRKAI-EQLNYTPIKGRL-CRIMWSQRDP
              #2          NIFIKNL-HPDIDNKA---LYDTFSVFGDILSSKIATDEN--GKSKGFGFVHFEEEGAAKEAI-DALNGMLLNGQE-IYVAPHLSRK
              #3          NLYVKNI-NSETTDEQ---FQELFAKFGPIVSASLEKDAD--GKLKGFGFVNYEKHEDAVKAV-EALNDSELNGEK-LYVGRAQKKN
              #4          NLFVKNL-DDSVDDEK---LEEEFAPYGTITSAKVMRTEN--GKSKGFGFVCFSTPEEATKAI-TEKNQQIVAGKP-LYVAIAQRKD
E.coli   Rho N1           NMGLENL-ARMRKQDI---IFAILKQHAKSGEDIFGDGVLE-ILQDGFGFLRSADSSYL--AGPDDIYVSPSQIRR-FNLRTGDTIS
T4       GP32             CQYISKNDLYNTDNKEYSLVKRKTSYWANILVVKDPAAPEN-EG-KVFKYRFGKKIWDKINAMIAVDVEMGETPVD-VTCPWEGANF

CONSENSUS                  LFVGNL            L   F  FG V               RGFGFV.F        A         I G   V V
                           IYIKG             I      Y  I               K  A I Y                  V     I I
                           RNPII                                         RNP I
```

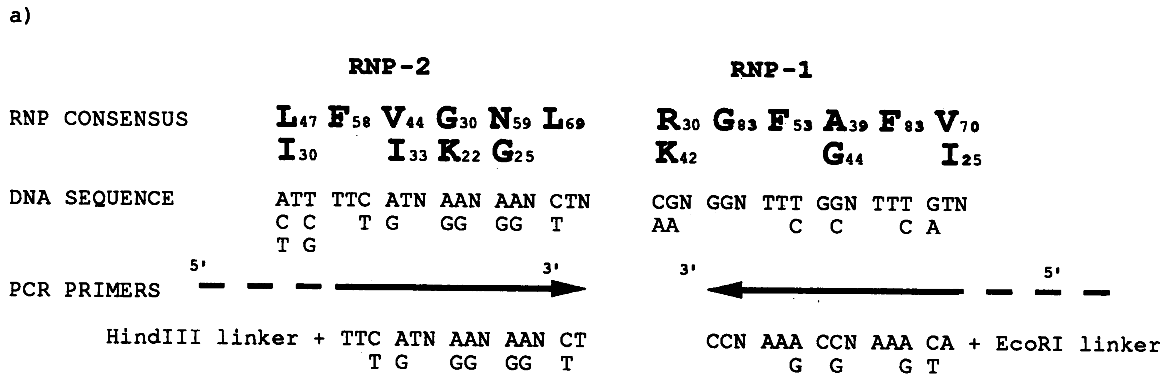
FIG. 1. Amino Acid Sequences of RNA recognition motifs. 32 RNA Recognition Motifs are aligned by the method of Kenan et al. (23). The conserved features are in bold type and include the most highly conserved segments denoted RNP-1 and RNP-2. The consensus sequence is shown at the bottom of the figure. The amino acid sequences were obtained from the following references: U1 snRNP 70K protein (42), U1 snRNP-A (54), U2 snRNP-B″ (18), Ro RNP-60K (12), UP2 (30), hnRNP C1/C2 (56), La (9), helix-destabilizing protein (11), UP1 (59), nucleolin (32), fly p9 (19), fly ELAV (42), SXL (4), tra-2 (1), fly U1 70K snRNP (33), SSB1 (22), polyadenylate-binding protein (45); *Escherichia coli* Rho protein (41); and T4 gp32 (28).

are very degenerate (Fig. 2a). The numbers of sequences in the 18 nucleotides of RNP-1 and RNP-2 are 16,384 and 294,912, respectively. We initially attempted to perform oligonucleotide hybridization screens of libraries using degenerate oligonucleotide probes made against RNP-1. These probes gave poor signal-to-noise ratios and large numbers of false-positives results, so we turned instead to PCR. We expected PCR to provide higher specificity by allowing primers to RNP-1 and RNP-2 to be used simultaneously. In order to keep the degeneracies of the PCR primers relatively low, we used the 14 nucleotides at the 3′ ends of the RNP-1 and RNP-2 to make PCR primers (Fig. 2a). A single pool of degenerate primers for RNP-1 and 4 pools of nonoverlapping degenerate primers for RNP-2 was made (Fig. 2b). Eight nucleotides containing a restriction enzyme site were incorporated at the 5′ end of each primer to facilitate cloning of the PCR-amplified products. These extra nucleotides also increased the stability of the hybrids after the first round of PCR amplification.

Amplification of RNA recognition motifs. In order to amplify the sequences between RNP-1 and RNP-2, we carried out first-strand cDNA synthesis with *D. melanogaster* adult poly(A)$^+$ RNA. We used mRNA as the starting material, since this allowed us to predict the size of the amplified products from RRM-containing genes. In all known RRMs, the RNP-1 and RNP-2 sequences are a relatively fixed distance apart, so the amplified product would be expected to be 130 to 140 nucleotides (Fig. 3). In addition, in genes with more than one RRM, these motifs are also a relatively constant distance apart, so the amplified product for a gene containing two RRMs would be expected to be 370 to 410 nucleotides (Fig. 3).

The products of the cDNA synthesis were amplified with the PCR primers depicted in Fig. 2b in a two-step PCR. The first 5 rounds of the PCR reaction were conducted at 48°C (low stringency), and the subsequent 25 rounds of the PCR reaction were conducted at 55°C (high stringency). The products of the PCR reaction were then separated on agarose gels (Fig. 3). The reactions carried out with RNP-2 primers RNP-2-1, RNP-2-2, and RNP-2-4 yielded products in the size ranges expected from genes containing one and two RRMs. However, PCR with primer RNP-2-3 consistently failed to produce products in the size ranges expected. This primer was subsequently found to contain a 3′ overlap with the RNP-1 primer, which led to the two primers priming off of one another rather than the cDNA.

To determine whether the PCR products in the 140- and 400 nt size classes (hereafter referred to as the S and L size classes, respectively) represented RRMs and to determine whether RRMs from more than one gene were present in these bands, we cloned the PCR products and sequenced 124 clones. Of these clones, 119 had open reading frames with the expected conserved residues between RNP-1 and RNP-2 (Fig. 4), suggesting that they represented RRM sequences.


a)
RNP-2
RNP-1
RNP CONSENSUS
L_{47} F_{58} V_{44} G_{30} N_{59} L_{69}
I_{30} I_{33} K_{22} G_{25}
R_{30} G_{83} F_{53} A_{39} F_{83} V_{70}
K_{42} G_{44} I_{25}
DNA SEQUENCE
ATT TTC ATN AAN AAN CTN
C C T G GG GG T
T G
CGN GGN TTT GGN TTT GTN
AA C C C A
PCR PRIMERS
5'
3'
3'
5'
HindIII linker + TTC ATN AAN AAN CT
T G GG GG T
CCN AAA CCN AAA CA + EcoRI linker
G G G T

b)

1. RNP-1 primer	5'-tcgaattcNAYRAANSCRAANCC-3'
2. RNP-2-1 primer	5'-tcaagcttTTYRTNAARAAYYT-3'
3. RNP-2-2 primer	5'-tcaagcttTTYRTNGGNGGNYT-3'
4. RNP-2-3 primer	5'-tcaagcttTTYRTNAARGGNYT-3'
5. RNP-2-4 primer	5'-tcaagcttTTYRTNGGNAAYYT-3'

FIG. 2. RRM Consensus sequences used to generate PCR primers. (a) The RRM consensus amino acid sequences for RNP-1 and RNP-2 derived from the known RRM sequences from Fig. 1 are shown. The number adjacent to each amino acid indicates the frequency that the amino acid is found at that position among the known RNP consensus sequences. The possible DNA sequences encoding these RNP consensus sequences are indicated. Degenerate PCR primers for the RNP consensus sequences were made using the 14 nucleotides at the 3′ end of each consensus sequence and 8 nucleotides containing a restriction enzyme site at the 5′ end of each primer. (b) Sequences of the PCR primers used to amplify cDNA fragments containing RRMs are listed. Sequences in capital letters are complementary to cDNAs containing the indicated RNP consensus sequences defined above. Sequences in lowercase type are linker sequences that create a *Eco*RI or *Hin*dIII site used to facilitate cloning of PCR-amplified fragments. The nucleotide ambiguities in the primer sequences were shown using the letter codes for nucleotides proposed by the International Union of Biochemistry (39). S means C or G, R means A or G, Y means T or C, and N means A or G or C or T.

Each of the 119 clones had one of 12 different RRM sequences, and these 12 different types of RRM sequences were named RRM1 to RRM12. Of 124 clones sequenced, 5 had one of two sequences which did not appear to be derived from an RRM sequence.

Size of the RRM gene family. The frequencies with which the individual RRM sequences were represented among the 124 clones sequenced varied substantially (Fig. 4). The different levels at which these genes were represented could be due either to the use of RNA as the starting material for the amplification or to different efficiencies of amplification for those RRMs. This result suggests that the 124 clones we sequenced may not have been enough to isolate RRM sequences amplified at low levels and led us to ask the next two questions. Are there unidentified RRM sequences amplified at low levels? Is there any way we can increase the level of amplification for weakly amplified RRM sequences?

In order to address the first question and to find out what portion of the amplified RRM sequences were identified by the sequence analysis, we used genomic Southern analysis to determine how many genes were represented in each of the PCR L and PCR S products. We probed Southern blots of restriction digests of wild-type (Canton S) DNA with probes prepared from each of the PCR L and PCR S fragments (Fig. 5). The three PCR L fragments each hybridized with a single major band. The PCR S1, S2, and S4 probes hybridized with seven, three, and three prominent bands, respectively, suggesting that they contained sequences from at least that many genes. The fact that the *Eco*RI digestions and the *Bam*HI digestions gave comparable numbers of bands and that these numbers are close to the number of RRM sequences plus the two nonspecific sequences identified by sequencing the PCR products suggests that 12 to 14 RRM sequences were amplified in the PCR. Therefore, the genomic Southern analysis showed that we may have identified essentially all of the RRMs present in the PCR products.

Probing of Northern blots of poly(A)$^+$ RNA from wild-type (Canton S) adults with probes prepared from each of the PCR S and PCR L products showed about 25 transcripts hybridized to these probes (Fig. 5). The fact that these probes in aggregate detected 16 bands in genomic Southern analysis and about 25 bands in Northern analysis suggests that some of these transcripts may come from the same gene and be due to alternative promoter usage or alternative processing events.

In order to test whether we could overcome the problem of biased amplification of RRM sequences, part of which resulted from the different concentrations of RRM gene transcripts in the mRNA substrate used for PCR, fly genomic DNA instead of cDNA was used as the template with the three pairs of PCR primers described above. The amplification of genomic DNA with these primers was less efficient and produced more size classes of products than when cDNA was used as the template (Fig. 3). Since any introns present in RRMs will be present in genomic DNA,

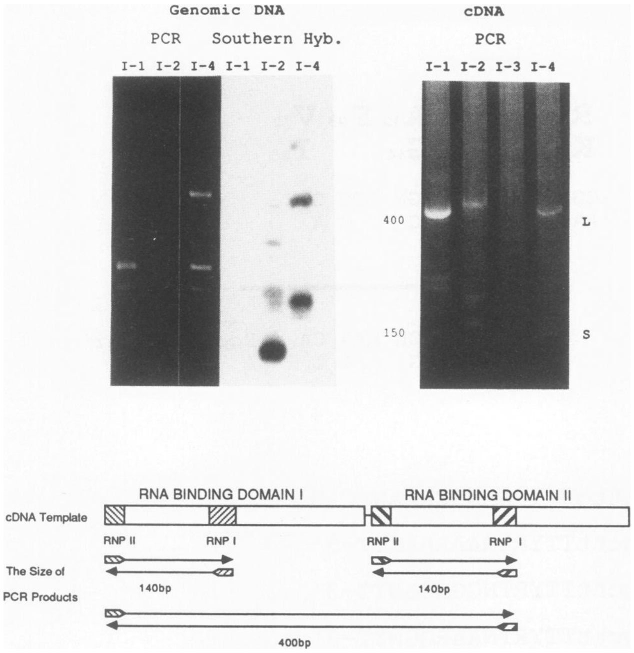


FIG. 3. Agarose gel electrophoresis of PCR products. PCR products amplified from both cDNA and genomic DNA using the degenerate primers shown in Fig. 2 were electrophoresed through a 2% agarose gel and visualized by ethidium bromide staining. The numbers above each lane indicate which RNP-2 primer was used in conjunction with the RNP-1 primer to generate the corresponding reaction products. The diagram at the bottom of the figure illustrates how amplified products of 140 and 400 bp could be generated from cDNA. Products of these sizes are denoted with the letters S for the small 140-bp products and L for the large 400-bp products. PCR products amplified from genomic DNA were analyzed by Southern hybridization (Southern Hyb.) The probe was made from the gel-purified cDNA PCR products L and S.

we could not predict the size of products amplified from RRMs. To determine which fragments in the genomic amplification come from the RRMs represented in the cDNA PCR fragments, a probe containing all the PCR L and PCR S products of the cDNA amplification was used to probe a Southern blot of the genomic PCR products. Less than half of the amplified bands from the PCR-amplified genomic DNA hybridized with the probe (Fig. 3). Even though we cannot rule out nonspecific amplifications from the noncoding sequences of the genomic DNA, considering that the same PCR primers amplified RRM sequences very specifically from cDNAs, the larger number of amplified fragments from the genomic DNA suggest that some of the fragments represent RRM genes that are either not expressed or expressed at lower levels in the adult stage. Therefore, the number of RRM sequences which can be amplified from *D. melanogaster* (not just from adult mRNA) using the degenerate PCR primers is greater than the 12 which were identified from adult cDNAs.

Numerical analysis of RRM sequences. As a first step in gaining insight into the functions of these RRMs, we analyzed the sequence similarities between the 12 RRMs reported here and the 62 RRMs found in 47 RRM-type RNA-binding proteins reported in the literature. The sequence comparison of these RRMs were done in two ways.

First, each of the 12 RRM sequences was compared with the 74 RRM fragment sequences using the FASTA homology searching program (13). Because the RRM itself contains a loosely conserved consensus sequence, we looked for similarities between the RRM sequences both in the RRM consensus sequence and in regions where there are normally no known constraints on the use of amino acids (23). Considering just the amino acids that make up the RRM consensus sequence, 10 of 12 RRM sequences had levels of similarity to previously identified RRM sequences that were substantially greater than those found among random pairs of RRMs. Moreover, these 10 RRM sequences also have similarities to those RRM sequences in the nonconserved

```
RRM1    TTTGTCGGCAATTTGGGCTCCTCGGCGTCCAAGCACGAG---ATAGAAGGCGCATTTGCCAAATATGGACCCCTGCGAAACGTGTGGGTGGCC------------------------CGCAATCCACCAGGTTTCGCCTTTGTC
        F  V  G  N  L  G  S  S  A  S  K  H  E     I  E  G  A  F  A  K  Y  G  P  L  R  N  V  W  V  A                          R  N  P  P  G  F  A  F  V

RRM2    TTTGTCGGGAATCTGCCGCAAGGCCTTGTGCAGGGCGAT---GTGATCAAAATATTCCAGGACTTTGAGGTGAAGTACGTGCGGCTGGTGAAGGACCGGGAAACGGAT---------CAG---TTCAAAGGCTTCGGCTTCATC
        F  V  G  N  L  P  Q  G  L  V  Q  G  D     V  I  K  I  F  Q  D  F  E  V  K  Y  V  R  L  V  K  D  R  E  T  D           Q     F  K  G  F  G  F  I

RRM3    TTTGTCGGTGGCTTGAGCTGGGAAACGACTGAGAAGGAA---CTCCGCGATCACTTCGGCAAATATGGCGAGATCGAGAACATCAATGTCAAGACAGATCCCCAGACCGGT------CGG---TCCCGAGGCTTCGCGTTCATC
        F  V  G  G  L  S  W  E  T  T  E  K  E     L  R  D  H  F  G  K  Y  G  E  I  E  N  I  N  V  K  T  D  P  Q  T  G        R     S  R  G  F  A  F  I

RRM4    TTCGTGGGAGGCCTATCCACTCAGACGACCGTGGAAACG---CTGCGCGGATTCTTTAGTCAGTTCGGTATCGTGGCCGATGCGGTGGTCTTGCGGGATCCGGTGAGCAAC------CAT---TCTAGGGGCTTCGGCTTCATG
        F  V  G  G  L  S  T  Q  T  T  V  E  T     L  R  G  F  F  S  Q  F  G  I  V  A  D  A  V  V  L  R  D  P  V  S  N        H     S  R  G  F  G  F  M

RRM5    TTTGTTGGCGGACTAAGTAGTTGTTGGCTTCGGCGTGGTTGTCGTAGTGGTGGTAGTGGTCGTTGGCTTCTTGGTTGTCGTTGTGGTAGTGGGCTTTGGCGTGGTTGTAGTAGTTGTCGTCTTGGTAGTGGGTTCGGTTTCGTC
        F  V  G  G  L  S  S  C  W  L  R  R  G  C  R  S  G  G  S  G  R  W  L  L  G  C  R  C  G  S  G  L  W  R  G  C  S  S  C  R  L  G  S  G  F  G  F  V

RRM6    TTTGTGGGAGGCCTCAGTTGGCAGACAAGTCCAGAGAGC---TTACGCGATTACTTCGGACGTTACGGTGATATCTCAGAGGCTATGGTCATGAAGGATCCCACGACGCGC------AGA---TCCAGAGGTTTTGCGTTTGTC
        F  V  G  G  L  S  W  Q  T  S  P  E  S     L  R  D  Y  F  G  R  Y  G  D  I  S  E  A  M  V  M  K  D  P  T  T  R        R     S  R  G  F  A  F  V

RRM7    TTCGTTGGCGGCCTATCCTGGGAGACGACGCAGGAGAAC---CTGTCGCGCTACTTCTGCCGCTTCGGGGACATCATTGACTGTGTGGTGATGAAGAACAACGAGAGCGGC------AGG---TCGCGCGGCTTTGGCTTCGTT
        F  V  G  G  L  S  W  E  T  T  Q  E  N     L  S  R  Y  F  C  R  F  G  D  I  I  D  C  V  V  M  K  N  N  E  S  G        R     S  R  G  F  G  F  V

RRM8    TTTGTGGGAGGTCTGCCCTACGGAGTGCGCGCAGCGGAT---TTGGAGCGCTTTTTCAAAGGCTACGGCCGCACACGCGACATCCTCATC---------------------------AAA------AATGGCTACGCCTTCATG
        F  V  G  G  L  P  Y  G  V  R  A  A  D     L  E  R  F  F  K  G  Y  G  R  T  R  D  I  L  I                             K        N  G  Y  A  F  M

RRM9    TTCGTCTACAACCTGGCGCCCGAGACCGAGGAGAACGTG---CTGTGGCAACTGTTTGGCCCCTTCGGAGCAGTGCAATCTGTTAAGGAGATTCGTGATCTGCAGAGCAAC------AAG---TGCAAGGGCTTTGGCTTCGTC
        F  V  Y  N  L  A  P  E  T  E  E  N  V     L  W  Q  L  F  G  P  F  G  A  V  Q  S  V  K  E  I  R  D  L  Q  S  N        K     C  K  G  F  G  F  V

RRM10   TTTGTGAACTACTTGCCACAGACGATGTCGCAGGACGAG---ATCCGTTCGTTGTTCGTCAGTTTTGGCGAGGTGGAGAGCTGCAAGTTGATACGCGACAAGGTGACAGGA------CAAAGTCTG---GGCTACGGATTCGTG
        F  V  N  Y  L  P  Q  T  M  S  Q  D  E     I  R  S  L  F  V  S  F  G  E  V  E  S  C  K  L  I  R  D  K  V  T  G        Q  S  L     G  Y  G  F  V

RRM11   TTCGTGGGGAACCTGGCTCCTCGGCGCTCCAAGCCACGA---GATAGAAGCGCATTTGCCAAATATGGACCCCTGCGAAACGTGTGGGTGGCC------------------------CGCAATCCACCAGGGTTCGCTTTCGTC
        F  V  G  N  L  A  P  R  R  S  K  P  R     D  R  S  A  F  A  K  Y  G  P  L  R  N  V  W  V  A                          R  N  P  P  G  F  A  F  V

RRM12   TTCGTGGACAACCTGGATAGCTCAGTGTCCGAGGACCTG---CTAATCGCCCTCTTCAGCACCATGGGGCCCGTCAAAAGCTGCAAAATCATTCGGGAA------------------CCGGGCAACGATCCATATGCCTTCATC
        F  V  D  N  L  D  S  S  V  S  E  D  L     L  I  A  L  F  S  T  M  G  P  V  K  S  C  K  I  I  R  E                    P  G  N  D  P  Y  A  F  I

Cons.   F  V  G  N  L                             L           F        F  G     V                                          R        R  G  F  G  F  V
           I  K  G                                I                    Y        I                                            K        K     Y  A     I
```

FIG. 4. DNA and amino acid sequences of PCR-amplified RRM fragments. The PCR-amplified RRM fragments were cloned and 124 independent clones were sequenced, and the amino acid sequences were derived from the DNA sequences. Twelve different kind of RRM sequences were identified, and two unrelated sequences were found. The number of times each RRM sequence was found in the 124 sequences is shown at the right side of the sequence. The conserved (Cons.) amino acids are identified at the bottom of figure.

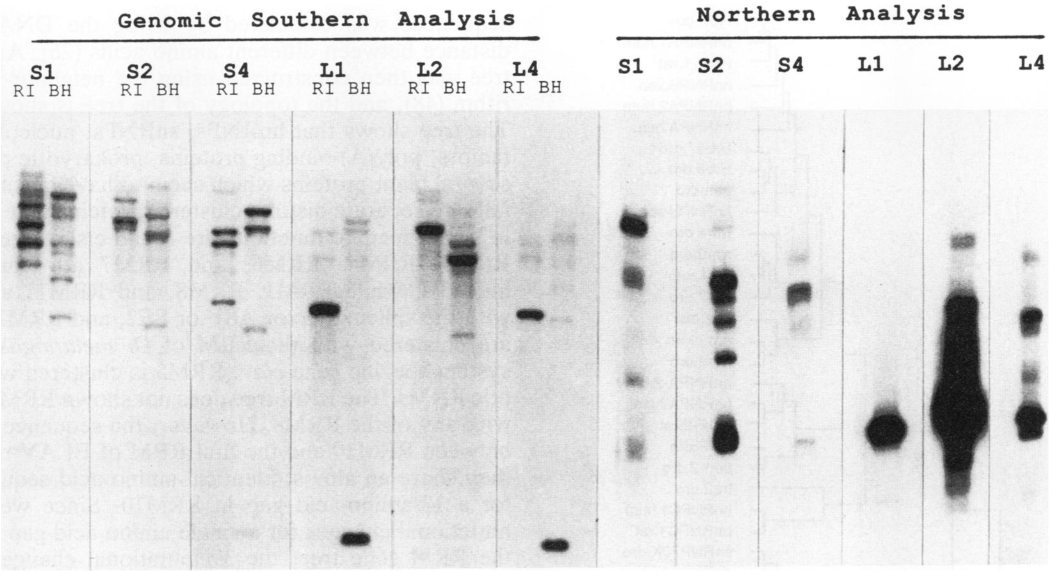


FIG. 5. Northern and genomic Southern hybridization with cDNA PCR products. Two micrograms of poly(A)$^+$ RNA from adult *Drosophila* flies was used in each lane for Northern hybridization, and 1 μg of *Drosophila* genomic DNA digested with *Eco*RI (RI lanes) or *Bam*HI (BH lanes) was used in the Southern hybridizations. The probe for each hybridization was made from one of the six groups of cDNA PCR products (S and L from cDNA PCR reactions 1, 2, and 4, respectively). The probe used in the hybridization is given over each lane.

sites of the RRM (Fig. 6). RRM3, RRM4, RRM6, and RRM7 have similarities to the RRM sequences of hnRNPs, while RRM1, RRM8, and RRM11 have similarities to the RRMs of human ASF or SF2, and RRM2, RRM9, and RRM10 have similarities to the RRMs of *elav*.

As a second approach to identify the possible functions of the proteins encoding the RRMs, we analyzed the relationships of the RRM sequences by a numerical approach which has been used to compare phylogenical relationships of proteins from various species. A distance matrix of the RRM

```
a)
RRM8            FVGGLPYGVRAADLER-FFKGYGRTRDILIKN-----GYAFM
                 :: :|   |  | |  : | ::  |||  ||       ::
ASF.hum         YVGNLPPDIRTKDIEDVFYK-YGAIRDIDLKNRRGGPPFAFV
                 ::::        :|   : | ::  |       |  | ::::
RRM1            FVGNLGSSASKHEIEGAFAK-YGPLRNVWVA--RNPPGFAFV
                :::::    ||      |:|| ::||||||||  ||||:::::
RRM11           FVGNLAPRRSKPRDRSAFAK-YGPLRNVWVA--RNPPGFAFV

b)
hnRNPA.hum      FVGGIKEDTEEHHLRDYFEQYGKIEVIEIMTDRGSGKKRGFAFV
                ::::    | |  :|| :  :: || |   ||   | ||::::
RRM3            FVGGLSWETTEKELRDHFDKYGEIENINVKTDPQTGRSRGFAFI
                :::::|| |  | : | : | :|      | || |||||:: ::
ArG.mou         FVGGLSWDTSKKDLKDYFTKFGEVVDCTIKMDPNTGRSRGFGFI
                :::::|| ||   : ||:   :         || | |||::::
RRM6            FVGGLSWQTSPESLRDYFGRYGDISEAMVMKDPTTRRSRGFAFV
                : :::||||  | :| |:   :|  |  || || | |||:: ::
nrp.xen         FIGGLSWQTTQEGLREYFSHFGDVKECLVMRDPLTKRSRGFGFV
                :::::||||||| :  |:  ::||  | ||      |||:::::
RRM7            FVGGLSWETTQENLSRYFCRFGDIIDCVVMKNNESGRSRGFGFV
                :::::|  || | :   :  :: | | ||     |  ||::::
RRM4            FVGGLSTQTTVETLRGFFSQFGIVADAVVLRDPVSNHSRGFGFM
                :::::|  | ||  |  :  :: ||      |   || |::::
hnRNPA.rat      FVGGLSPDTPEEKIREYFGGFGEVESIELPMDNKTNKRRGFGFI

c)
RRM9            FVYNLAPETEENVLWQLFGPFGAVQSVKEIRDLQSNKCKGFGFV
                : :::||||||  :|||:||::||||||   |   | ||: :::
elav-3.dro      FIYNLAPETEEAALWQLFGPFGAVQSVKIVKDPTTNQCKGYGFV
                :  ::               :  |  |  |||  | | |: ::
RRM2            FVGNLPQGLVQGDVIKIFQDFE-VKYVRLVKDRETDQFKGFGFI

RRM10           FVNYLPQTMSQDEIRSLFVSFGEVESCKLIRDKVT-------------GQSLGYGFV
                 ::::||||  ||:|||: | :| || ||||||               ||||:::::
elav-1.dro      IVNYLPQTMTEDEIRSLFSSVGEIESVKLIRDKSQVYIDPLNPQAPSKGQSLGYGFV
```

FIG. 6. Alignments of similar RRM sequences. Only identical amino acids are marked. Vertical bars (|) indicate the identical amino acids at the residues which do not belong to the RRM consensus sequence, while colons (:) indicate the identical amino acids at residues that are part of the consensus sequence of the RRM. (a) splicing component-like RRMs; (b) hnRNP-type RRMs; (c) nerve-specific protein-type RRMs. Abbreviations: hum, human; mou, mouse; xen, *Xenopus laevis*; dro, *D. melanogaster*.

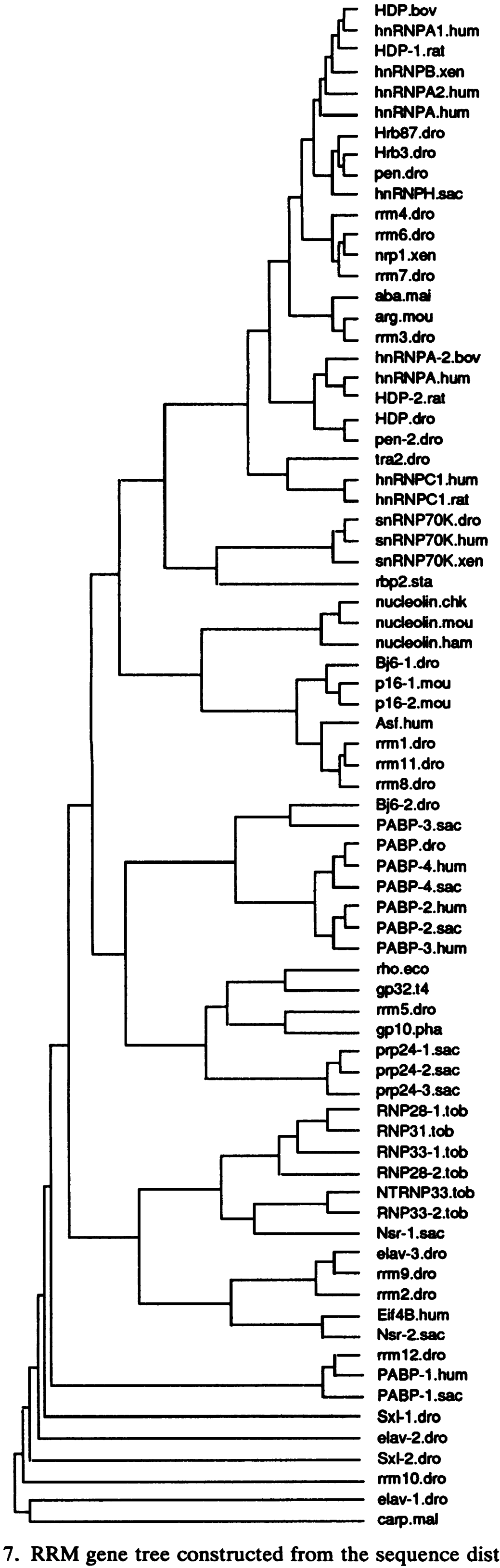


FIG. 7. RRM gene tree constructed from the sequence distance matrix of the 74 RRM sequences, using the neighbor-joining algorithm. Only the topology of the tree is shown here. When there was more than one RRM in a gene, each RRM was given a number after a hyphen. The references for most sequences may be found in Kenan et al. (23), others are shown in the legend to Fig. 1.

sequences was calculated by using the DNA mutational distance between different amino acids (26). An RRM gene tree was then constructed using the neighbor-joining algorithm (48), and the topology of the tree is shown in Fig. 7. The tree shows that hnRNPs, snRNPs, nucleolins, splicing factors, poly(A)-binding proteins, prokaryotic proteins, and several plant proteins which seem to have a similar function fall into separate distinct clusters. Proteins that are involved in tissue-specific functions are found elsewhere in the tree. RRM3, RRM4, RRM6, and RRM7 are clustered with hnRNPs, while RRM1, RRM8, and RRM11 are clustered with the splicing factor ASF or SF2, and RRM2 and RRM9 are clustered with the RRM of *D. melanogaster* nervous system-specific gene *elav*. RRM5 is clustered with prokaryotic RRMs. The RRM tree does not shown RRM10 clustered with any of the RRMs. However, the sequence comparison between RRM10 and the first RRM of ELAV revealed that they share an almost identical amino acid sequence except for a 13-amino-acid gap in RRM10. Since we used three mutational changes for a single amino acid gap to construct the RRM gene tree, the 39 mutational changes calculated from the 13-amino-acid gap in the alignment of RRM10 and ELAV's RRM overshadowed their similarity in the construction of the gene tree. Overall, 11 of the 12 RRMs that we isolated have sequence similarity to three groups of RRMs of what are likely to be functionally distinct classes. These classes are an hnRNP type, a splicing regulator type, and a nerve-specific protein type. The remaining RRM did not show sequence similarity with known RRMs, nor did it cluster with known RRMs in the RRM gene tree.

DISCUSSION

The availability of data bases has led to the identification of a number of consensus sequences found in families of proteins with related functions or properties (for examples, see references 17, 31, 36, and 53). These consensus sequences range from substantial blocks of highly conserved amino acids to rather loosely conserved blocks of amino acids. The RNA recognition motif (RRM) that characterizes the family of RNA-binding proteins that we have focused on here is representative of loosely conserved domains. At the protein level, RRM extends over a region of approximately 90 amino acids, only 21 of which are conserved between family members. Moreover, for a number of these conserved residues, it is the amino acid type (i.e., aliphatic, basic, aromatic) rather than the specific amino acid that is conserved. At the DNA level, the conservation of the RRM is not recognizable, making it very difficult to use hybridizations to isolate other members of this family. Low-stringency or oligonucleotide hybridizations, which are the classic approaches to isolating homologous genes, often give too many false-positives results. PCR allows these problems to be overcome and provides an efficient procedure for isolating genes containing weakly conserved consensus sequences. First, using two primers at a time increases the stringency of the screens. Second, when cDNA is used as a template, the size of the amplified products can be predicted, providing another criteria for distinguishing the real products

Abbreviations: bov, bovine; hum, human; xen, *Xenopus laevis*; dro, *D. melanogaster*; sac, *Saccharomyces cerevisiae*, mai, maize; chk, chicken; sta, *Staphylococcus aureus*; mou, mouse; ham, hamster; eco, *E. coli*; pha, phage; mal, malaria.

from those produced by nonspecific amplification. Last, the amplified products can be sequenced directly to confirm that they have the expected sequence. Here, even though each of the PCR primers we used was only 14 bases in length and had degeneracies ranging from 250- to 1,000-fold, we were able to isolate RRM-containing genes with a high degree of specificity.

The diverse array of metabolic processes that RNA molecules undergo in eukaryotes requires a large number of *trans*-acting factors that recognize RNA in many different ways. RRM has been found in a number of proteins involved in diverse aspects of RNA metabolism, suggesting that this gene family may be of substantial size. However, the members of this family identified thus far have largely been identified fortuitously, and thus a clear idea of the actual size of this gene family has been lacking. We tried to estimate the size of the RRM-containing protein gene family by amplifying members of the family using highly degenerate PCR primers which corresponded to known RNP-1 and RNP-2 sequences. The combinations of the nucleotide sequences of PCR primers for RNP-1 correspond to only 29% of possible RNP-1 sequences, whereas the RNP-2-1, RNP-2-2, and RNP-2-4 primers correspond to only 14% of the possible RNP-2 sequences. Thus, assuming the sequences present in RNP-1 and RNP-2 are independent of one another, the PCR products that could be amplified by the combinations of primers we utilized represent only 4.1% (14% × 29%) of possible RRM sequences. Using these PCR products, 12 RRM sequences were identified and 16 restriction fragments were hybridized in the genomic Southern analysis. Theoretically, this result suggests about 300 as the number of RRM sequences in flies. However, there may be a difference between the estimated and actual sizes of the RNA-binding protein gene family for the following reasons. Only 32 RRM sequences were available when we designed the PCR probes. Also, those RNP-1 and RNP-2 sequences that were the basis of the PCR primers came from all known prokaryotic and eukaryotic members of this family, while only a small number of these sequences were from *D. melanogaster*, which may increase the pool size of the consensus sequences larger than that actually found in *D. melanogaster*, resulting in a higher estimation of the RNA-binding protein gene family size. However, the facts that more than 30 different RRM sequences representing about 20 genes, including the 12 RRMs reported here, are now identified from *D. melanogaster* and that the 12 RRMs represent only RRM genes expressed at adult stage suggest that 300 may be a reasonable estimate.

The similarity between RRMs suggests that they may have evolved from a common ancestor. Because functionally important sites on proteins appear to be resistant to evolutionary change, functionally related RRMs may show significant sequence similarity, even when they belong to different species. So far, more than 50 RRM proteins with housekeeping functions or developmentally regulatory roles in RNA metabolism have been identified. As expected from the diverse functions of the proteins with RRMs, they interact with many classes of RNA molecules, with high specificities. The studies on the U1 70K (42), U1A and U2B″ (5, 51), and poly(A)-binding protein (46) showed that at least a portion of RNA binding activities as well as binding specificities of RRM-type RNA-binding proteins reside in the RRM sequences. Therefore, a systematic analysis of the RRM sequences may reveal a connection between the sequence distance or similarity and functional relatedness of the RRMs. A search of the protein data base with the RRM sequences showed that some of the RRM sequences had a high similarity to those of known proteins—hnRNPs, human ASF or SF2, or *D. melanogaster* ELAV. However, some RRMs did not have obviously high similarities to known RRMs, even though their sequences appear to have an RRM consensus sequence. To test whether these moderate similarities indicate functional relatedness beyond that due to the RRMs themselves, we analyzed the RRM sequences by a numerical approach which has been used to find phylogenical relationships between species based on the sequence distances of conserved proteins. The topology of the RRM gene tree showed that functionally related RRMs are clustered together and that RRMs with different functions are separated into different clusters (Fig. 6). The fact that most of the RRMs we have isolated cluster in this tree with RRMs from proteins with known function suggests that these RRMs may be functionally related to these proteins.

It is noteworthy that a large proportion of the RRM sequences that we recovered are hnRNP-like. While this may reflect the real constitution of the RRM gene family in *D. melanogaster*, it is worth keeping in mind that the degenerate PCR primers we used were designed from RRM sequences in which hnRNPs were heavily represented. Hence the high proportion of hnRNP-like sequences among the RRMs may be a reflection of the primers used in the isolation of RRMs. Regardless of the origin of the high proportion of hnRNP-like sequences among the RRMs, the identification of *Drosophila* genes that were clustered with hnRNP proteins and have extensive similarities to these proteins throughout the region spanned by the RRM suggests that these RRMs may function as hnRNPs in RNA metabolism. However, whether hnRNPs share a common biological function is not known yet. Even though their functions are not clearly understood (except for their high nucleic acid affinities), the fact that they are clustered together in the numerical analysis suggests that their functions may be related at least in a biochemical sense.

The construction of the RRM gene tree and the discovery of similarities between the RRMs and known components of RNA metabolism provide clues as to the functions of particular RRMs. It will be interesting to see whether the similarities between these genes extends beyond their RRMs. We are in the process of currently testing this idea. We have thus far found in the two cases (RRM1 and RRM9/RRM10) that we have examined in some detail (24, 25) that the functions of the genes encoding these RRMs are indeed functionally related to those of the genes with which they cluster in this tree (ASF [or SF2] and *elav*, respectively). This finding strengthens the result of the numerical analysis which showed sequence similarities between RRMs and various types of RNA processing factors and suggests that we may have identified genes involved in diverse aspects of RNA metabolism. These clones thus provide a means to study diverse aspects of RNA metabolism.

ADDENDUM

Recently we cloned and sequenced the *rbp12* gene that encodes the RRM12 sequence. RRM12 was clustered with the RRMs of poly(A)-binding proteins and *rbp12* showed a high similarity to a human poly(A)-binding protein (58). Therefore, together with the results from the analysis of *rbp1* and *rbp9* (24, 25), three genes we have analyzed so far show extensive homology to the genes with which their RRM sequences are clustered by the numerical analysis.

ACKNOWLEDGMENTS

Junhyong Kim helped to construct the RRM gene tree and provided instructive comments. We thank Wendy Wall, Lisa Ryner, Claudia Zraly, and Monica Gorman for comments on the manuscript. We also thank Gennet Bohm for preparing media and laboratory supplies.

This work was supported by a grant from the National Institutes of Health.

REFERENCES

1. **Amrein, H., M. Gorman, and R. Nothiger.** 1988. The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein. Cell **55:**1025–1035. (Erratum, **58:**420, 1989.)
2. **Baker, B. S.** 1989. Sex in flies: the splice of life. Nature (London) **340:**521–524.
3. **Bandziulis, R. J., M. S. Swanson, and G. Dreyfuss.** 1989. RNA-binding proteins as developmental regulators. Genes Dev **3:**431–437.
4. **Bell, L. R., E. M. Maine, P. Schedl, and T. W. Cline.** 1988. Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell **55:**1037–1046.
5. **Bentley, R. C., and J. D. Keene.** 1991. Recognition of U1 and U2 small nuclear RNAs can be altered by a 5-amino-acid segment in the U2 small nuclear ribonucleoprotein particle (snRNP) B″ protein and through interactions with U2 snRNP-A′ protein. Mol. Cell. Biol. **11:**1829–1839.
6. **Bernstein, P., and J. Ross.** 1989. Poly(A), poly(A) binding protein and the regulation of mRNA stability. Trends Biochem. Sci. **14:**373–377.
7. **Bingham, P. M., T. B. Chou, I. Mims, and Z. Zachar.** 1988. On/off regulation of gene expression at the level of splicing. Trends Genet. **4:**134–138.
8. **Breitbart, R. E., A. Andreadis, and B. Nadal-Ginard.** 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. **56:**467–495.
9. **Chambers, J. C., and J. D. Keene.** 1985. Isolation and analysis of cDNA clones expressing human lupus *La* antigen. Proc. Natl. Acad. Sci. USA **82:**2115–2119.
10. **Clawson, G. A., C. M. Feldherr, and E. A. Smuckler.** 1985. Nucleocytoplasmic RNA transport. Mol. Cell. Biochem. **67:**87–99.
11. **Cobianchi, F., D. N. SenGupta, B. Z. Zmudzka, and S. H. Wilson.** 1986. Structure of rodent helix-destabilizing protein revealed by cDNA cloning. J. Biol. Chem. **261:**3536–3543.
12. **Deutscher, S. L., and J. D. Keene.** 1988. A sequence-specific conformational epitope on U1 RNA is recognized by a unique autoantibody. Proc. Natl. Acad. Sci. USA **85:**3299–3303.
13. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:**387–395.
14. **Dreyfuss, G., M. S. Swanson, and R. S. Pinol.** 1988. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. Trends Biochem. Sci. **13:**86–91.
15. **Ge, H., P. Zuo, and J. L. Manley.** 1991. Primary structure of the human splicing factor ASF reveals similarities with Drosophila regulators. Cell **66:**373–382.
16. **Goralski, T. J., J. E. Edstrom, and B. S. Baker.** 1989. The sex determination locus transformer-2 of Drosophila encodes a polypeptide with similarity to RNA binding proteins. Cell **56:**1011–1018.
17. **Gorbalenya, A. E., E. V. Koonin, A. P. Donchenko, and V. M. Blinov.** 1988. A conserved NTP-motif in putative helicases. Nature (London) **333:**22. (Letter.)
18. **Habets, W. J., M. H. Hoet, P. Bringmann, R. Luhrmann, and W. J. van Venrooij.** 1985. Autoantibodies to ribonucleoprotein particles containing U2 small nuclear RNA. EMBO J. **4:**1545–1550.
19. **Haynes, S. R., M. L. Rebbert, B. A. Moxer, R. Forquignon, and I. B. Dawdid.** 1987. *pen* repeat sequences are GGN clusters and encode a glycine-rich domain in a Drosophila cDNA homologous to the rat helix destabilizing protein. Proc. Natl. Acad. Sci. USA **84:**1819–1823.
20. **Hinnebusch, A. G.** 1988. Mechanisms of gene regulation in the general control of amino acid biosynthesis in *Saccharomyces cerevisiae*. Microbiol. Rev. **52:**248–273.
21. **Hoffman, D. W., C. C. Query, B. L. Golden, S. W. White, and J. D. Keene.** 1991. RNA-binding domain of the A protein component of the U1 small nuclear ribonucleoprotein analyzed by NMR spectroscopy is structurally similar to ribosomal proteins. Proc. Natl. Acad. Sci. USA **88:**2495–2499.
22. **Jong, A. Y.-S., M. W. Clark, M. Gilbert, A. Oehm, and J. L. Campbell.** 1987. *Saccharomyces cerevisiae* SSB1 protein and its relationship to nucleolar RNA-binding proteins. Mol. Cell. Biol. **7:**2947–2955.
23. **Kenan, D. J., C. C. Query, and J. D. Keene.** 1991. RNA recognition: towards identifying determinants of specificity. Trends Biochem. Sci. **16:**214–220.
24. **Kim, Y.-J., and B. S. Baker.** The Drosophila gene *rbp9* encodes a protein that is a member of a conserved group of putative RNA binding proteins that are nervous system-specific in both flies and humans. J. Neurosci., in press.
25. **Kim, Y.-J., P. Zuo, J. L. Manley, and B. S. Baker.** A Drosophila RNA binding protein of *rbp1* is localized to transcriptionally active sites of chromosomes and shows a functional similarity to human splicing factor ASF/SF2. Genes Dev., in press.
26. **Kimura, M.** 1983. The neutral theory of molecular evolution. Cambridge University Press, New York.
27. **Krainer, A. R., A. Mayeda, D. Kozak, and G. Binns.** 1991. Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and Drosophila splicing regulators. Cell **66:**383–394.
28. **Krisch, H. M., and B. Allet.** 1982. Nucleotide sequences involved in bacteriophage T4 gene 32 translational self-regulation. Proc. Natl. Acad. Sci. USA **79:**4937–4941.
29. **Kumar, A., K. R. Williams, and W. Szer.** 1986. Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-stranded DNA binding protein. J. Biol. Chem. **261:**11266–11273.
30. **Lahiri, D. K., and J. O. Thomas.** 1986. A cDNA clone of the hnRNP C proteins and its homology with the single-stranded DNA binding protein UP2. Nucleic Acids Res. **14:**4077–4094.
31. **Landschulz, W. H., P. F. Johnson, and S. L. McKnight.** 1988. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science **240:**1759–1764.
32. **Lapeyre, B., H. Bourbon, and F. Amalric.** 1987. Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence. Proc. Natl. Acad. Sci. USA **84:**1472–1476.
33. **Mancebo, R., P. C. H. Lo, and S. M. Mount.** 1990. Structure and expression of the *Drosophila melanogaster* gene for the U1 small nuclear ribonucleoprotein particle 70K protein. Mol. Cell. Biol. **10:**2492–2502.
34. **Mattaj, I. W.** 1989. A binding consensus: RNA-protein interactions in splicing, snRNPs, and sex. Cell **57:**1–3.
35. **Merrill, B. M., K. L. Stone, F. Cobianchi, S. H. Wilson, and K. R. Williams.** 1988. Phenylalanines that are conserved among several RNA-binding proteins form part of a nucleic acid-binding pocket in the A1 heterogeneous nuclear ribonucleoprotein. J. Biol. Chem. **263:**3307–3313.
36. **Murre, C., P. S. McCaw, and D. Baltimore.** 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell **56:**777–783.
37. **Nagai, K., C. Oubridge, T. H. Jessen, J. Li, and P. R. Evans.** 1990. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. Nature (London) **348:**515–520.
38. **Nagoshi, R. N., M. McKeown, K. C. Burtis, J. M. Belote, and B. S. Baker.** 1988. The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster. Cell **53:**229–236.
39. **Nomenclature Committee of the International Union of Biochemistry (NC-IUB).** 1985. Nomenclature for incompletely specified bases in nucleic acid sequences. Eur. J. Biochem. **150:**1–5.

40. **Pandolfo, M., D. Valentini, G. Biomont, C. Morandi, and S. Riva.** 1985. Single stranded DNA binding proteins derive from hnRNP proteins by proteolysis in mammalian cells. Nucleic Acids Res. **13:**6577–6590.
41. **Pinkham, J. L., and T. Platt.** 1983. The nucleotide sequence of the *rho* gene of *E. coli* K-12. Nucleic Acids Res. **11:**3531–3545.
42. **Query, C. C., R. C. Bentley, and J. D. Keene.** 1989. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell **57:**89–101.
43. **Robinow, S., A. R. Campos, K. M. Yao, and K. White.** 1988. The elav gene product of Drosophila, required in neurons, has three RNP consensus motifs. Science **242:**1570–1572. (Erratum, **243:** 12, 1989).
44. **Robinow, S., and K. White.** 1988. The locus elav of Drosophila melanogaster is expressed in neurons at all developmental stages. Dev. Biol. **126:**294–303.
45. **Sachs, A. B., M. W. Bond, and R. D. Kornberg.** 1986. A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: domain structure and expression. Cell **45:**827–835.
46. **Sachs, A. B., R. W. Davis, and R. D. Kornberg.** 1987. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Mol. Cell. Biol. **7:**3268–3276.
47. **Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich.** 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:**487–491.
48. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
49. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
50. **Scherly, D., W. Boelens, N. A. Dathan, W. J. van Venrooij, and I. W. Mattaj.** 1990. Major determinants of the specificity of interaction between small nuclear ribonucleoproteins U1A and U2B″ and their cognate RNAs. Nature (London) **345:**502–506.
51. **Scherly, D., W. Boelens, W. J. van Venrooij, N. A. Dathan, J. Hamm, and I. W. Mattaj.** 1989. Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. EMBO J. **8:**4163–4170.
52. **Schwemmle, M., M. Gorlach, M. Bader, T. F. Sarre, and K. Hilse.** 1989. Binding of mRNA by an oligopeptide containing an evolutionarily conserved sequence from RNA binding proteins. FEBS Lett. **251:**117–120.
53. **Scott, M. P., and A. J. Weiner.** 1984. Structural relationships among genes that control development: sequence homology between the *Antennapedia*, *Ultrabithorax*, and *fushi tarazu* loci of Drosophila. Proc. Natl. Acad. Sci. USA **81:**4115–4119.
54. **Sillekens, P. T., R. P. Beijer, W. J. Habets, and W. J. van Venrooij.** 1988. Human U1 snRNP-specific C protein: complete cDNA and protein sequence and identification of a multigene family in mammals. Nucleic Acids Res. **16:**8307–8321.
55. **Smith, C. W., J. G. Patton, and G. B. Nadal.** 1989. Alternative splicing in the control of gene expression. Annu. Rev. Genet. **23:**527–577.
56. **Swanson, M. S., T. Y. Nakagawa, K. LeVan, and G. Dreyfuss.** 1987. Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA and pre-rRNA-binding proteins. Mol. Cell. Biol. **7:**1731–1739.
57. **Theissen, H., M. Etzerodt, R. Reuter, C. Schneider, F. Lottspeich, P. Argos, R. Luhrmann, and L. Philipson.** 1986. Cloning of the human cDNA for the U1 RNA-associated 70K protein. EMBO J. **5:**3209–3217.
58. **Tian, Q., M. Streuli, H. Saito, S. F. Schlossman, and P. Anderson.** 1991. A polyadenylate binding protein localized to the granules of cytolytic lymphocytes induces DNA fragmentation in target cells. Cell **67:**629–639.
59. **Williams, K. R., K. L. Stone, M. B. LoPresti, B. M. Merrill, and S. R. Planck.** 1985. Amino acid sequence of the UP1 calf thymus helix-destabilizing protein and its homology to an analogous protein from mouse myeloma. Proc. Natl. Acad. Sci. USA **82:**5666–5670.