Characterization of the Major hnRNP Proteins from *Drosophila melanogaster*

Erika L. Matunis, Michael J. Matunis, and Gideon Dreyfuss

Howard Hughes Medical Institute and Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-6148

Abstract. To better understand the role(s) of hnRNP proteins in the process of mRNA formation, we have identified and characterized the major nuclear proteins that interact with hnRNAs in *Drosophila melanogaster*. cDNA clones of several *D. melanogaster* hnRNP proteins have been isolated and sequenced, and the genes encoding these proteins have been mapped cytologically on polytene chromosomes. These include the hnRNP proteins hrp36, hrp40, and hrp48, which together account for the major proteins of hnRNP complexes in *D. melanogaster* (Matunis et al., 1992, accompanying paper). All of the proteins described here contain two amino-terminal RNP consensus sequence RNA-binding domains and a carboxyl-terminal glycine-rich domain. We refer to this configuration, which is also found in the hnRNP A/B proteins of vertebrates, as 2×RBD-Gly. The sequences of the *D. melanogaster* hnRNP proteins help define both highly conserved and variable amino acids within each RBD and support the suggestion that each RBD in multiple RBD-containing proteins has been conserved independently and has a different function. Although 2×RBD-Gly proteins from evolutionarily distant organisms are conserved in their general structure, we find a surprising diversity among the members of this family of proteins. A mAb to the hrp40 proteins crossreacts with the human A/B and G hnRNP proteins and detects immunologically related proteins in divergent organisms from yeast to man. These data establish 2×RBD-Gly as a prevalent hnRNP protein structure across eukaryotes. This information about the composition of hnRNP complexes and about the structure of hnRNA-binding proteins will facilitate studies of the functions of these proteins.

HETEROGENEOUS nuclear RNAs (hnRNAs), from which mRNAs in eukaryotic cells are derived by RNA processing, are associated with specific nuclear proteins and form heterogeneous nuclear RNP (hnRNP) complexes (reviewed in Dreyfuss, 1986; Dreyfuss et al., 1988). Several hnRNP proteins in vertebrate cells have been extensively characterized, and some are likely to participate in various steps along the pathway of gene expression, including pre-mRNA processing (Choi et al., 1986; Sierakowska et al., 1986; Moore et al., 1988; Willusz et al., 1988).

The primary structures of several hnRNP proteins have been determined, and they can be placed into a few general categories. The hnRNP A, B, C, and E proteins (Williams et al., 1985; Cobianchi et al., 1986; Lahiri and Thomas, 1986; Swanson et al., 1987; Buvoli et al., 1988; Preugschat and Wold, 1988; Burd et al., 1989; Kay et al., 1990) are composed of one or two amino-terminal 80–90 amino acid RNA-binding domains (CS-RBDs or RBDs) which contain two highly conserved peptides: an octapeptide termed the RNP consensus sequence (RNP-CS or RNP1) and a hexapeptide designated RNP2. These motifs, as well as several specific amino acids throughout the RBD, are identifying features of a rapidly growing number of RNA-binding proteins (reviewed in Dreyfuss et al., 1988; Bandziulis et al., 1989; Mattaj, 1989; Query et al., 1989). These RNP-CS proteins also have a carboxyl-terminal auxiliary domain which is unique for each type of protein. Other hnRNP proteins such as the L protein (Piñol-Roma et al., 1989), the M protein (M. Swanson and G. Dreyfuss, unpublished results), and the I protein (A. Ghetti, S. Piñol-Roma, M. Michael, C. Morandi, and G. Dreyfuss, manuscript in preparation) have been found to contain two or more repeated domains that are distantly related to the CS-RBD described above. Finally, there are hnRNP proteins such as the K (Matunis et al., 1992*b*) and U proteins (M. Kiledjian and G. Dreyfuss, manuscript in preparation) which have no apparent homology to any previously described hnRNP proteins. The degree of evolutionary conservation of these latter five proteins is not yet known, but the A1 and C proteins have been cloned and sequenced in both mammalians and amphibians (Williams et al., 1985; Cobianchi et al., 1986; Swanson et al., 1987; Buvoli et al., 1988; Preugschat and Wold, 1988; Kay et al., 1990), and their sequences in these vertebrate organisms are very highly conserved.

Considerably less is known about invertebrate hnRNP proteins. Two *D. melanogaster* loci have been described which encode proteins with sequence similarity to the A/B hnRNP proteins (Haynes et al., 1990; Haynes et al., 1991),

suggesting that vertebrate and invertebrate hnRNP complexes could have proteins with similar structures. In the accompanying paper (Matunis et al., 1992*a*) we describe the isolation of *D. melanogaster* hnRNP complexes and the characterization of several of the major constituent proteins using monoclonal antibodies. We found that invertebrate hnRNP complexes, like vertebrate hnRNP complexes (Piñol-Roma et al., 1988), contain a large number of proteins. The hnRNP complexes from *D. melanogaster* are composed of at least 10 major proteins that migrate between 36 and 75 kD on SDS-polyacrylamide gels. In addition, many less abundant proteins are present in these complexes. Interestingly, several RNP-CS proteins involved in sex determination have been found in *D. melanogaster* (reviewed in Baker, 1989). These proteins, initially identified in genetic studies and subsequently described at a molecular level, participate in RNA processing events and have recently been shown to bind specific hnRNAs directly (Inoue et al., 1990; Hedley and Maniatis, 1991; Hoshijima et al., 1991). However, the relationship between these splicing factors and the abundant hnRNP proteins from *D. melanogaster* is not known. Since this organism offers unique advantages for analyzing the function of a protein, including amenability to genetic approaches, we have begun to characterize hnRNP proteins in *D. melanogaster*. Here we describe the isolation of cDNA clones encoding representatives of several major groups of *D. melanogaster* hnRNP proteins, their primary structures, cytological map positions, and relatedness to vertebrate hnRNP proteins.

Materials and Methods

Isolation of cDNA Clones

Mouse polyclonal antisera recognizing several proteins in the *D. melanogaster* embryo 2M ssDNA eluate and the mAb 10D5 (both described in the accompanying paper) were used at a 1:500 dilution to screen a random primed λgt11 cDNA library from 0–16 h *D. melanogaster* embryos (a kind gift of Dr. Bernd Hovemann, University of Heidelberg, Heidelberg) as described previously (Nakagawa et al., 1986). Purified positive plaques obtained with the polyclonal antisera were used to affinity select antibodies (Snyder et al., 1987), and affinity-selected antibodies were used to probe immunoblots of *D. melanogaster* 2M ssDNA eluate as described previously (Choi and Dreyfuss, 1984). The mAbs 5A5 and 4C2 (Matunis et al., 1992*a*) were used to screen immunologically positive plaques obtained with the polyclonal antisera at a 1:500 dilution as described above. When necessary, inserts from immunologically positive clones were used as hybridization probes to rescreen the library in order to obtain cDNA clones encompassing the entire coding sequence of these genes as described by Maniatis et al. (1982).

In Vitro Transcription, In Vitro Translation, and Immunoprecipitation

Phage inserts from the λgt11 positive clones were subcloned into pGEM plasmids (Promega Biotech, Madison, WI) and linearized with appropriate restriction sites to provide templates for in vitro transcription (Melton et al., 1984) using SP6 or T7 polymerase (Promega Biotech). These RNAs were translated in nuclease treated rabbit reticulocyte lysate in the presence of [^{35}S]methionine according to the manufacturer's instructions (Promega Biotech). Immunoprecipitation of the [^{35}S]methionine-labeled proteins was performed as previously described (Adam et al., 1986). Ascites fluid from a BALB/c mouse inoculated with the parent myeloma cell line SP2/0 was used as a nonimmune control.

Gel Electrophoresis and Immunoblotting

Electrophoresis of proteins in SDS-PAGE was performed as described previously (Dreyfuss et al., 1984) using 12.5% acrylamide in the separating gel. Fluorography of [^{35}S]methionine-labeled proteins was performed as described (Laskey and Mills, 1975). Two-dimensional NEPHGE was performed according to O'Farrell et al. (1977). The first dimension was separated using a pH 3–10 ampholine gradient at 475 V for 4 h, and the second dimension was separated by SDS-PAGE (see above). Immunoblotting was done as described previously (Choi and Dreyfuss, 1984), using each ascites fluid at a 1:1,000 dilution.

Cell Culture, Labeling, and ssDNA Chromatography

D. melanogaster Schneider's line 2 (S2) cells were maintained, labeled, and used as a source for the isolation of proteins by single-stranded DNA (ssDNA) chromatography as described in the accompanying paper (Matunis et al., 1992*a*).

DNA Sequence Analysis

A combination of nested deletions (Erase-A-Base kit; Promega Biotech), restriction fragments, and specific oligonucleotide primers were sequenced by the dideoxy chain termination method (Sanger et al., 1977) using double-stranded plasmid DNA and T7 DNA polymerase according to the manufacturer's instructions (Pharmacia LKB, Piscataway, NJ). The University of Wisconsin Genetics Computer Group (UWGCG) Sequence Analysis Programs (Devereux et al., 1984) and the MacVector programs (IBI, New Haven, Connecticut) were used for sequence analysis. Database searches were performed using the universal sequence database searching programs FASTA and TFASTA (Pearson and Lipman, 1988).

Cytological Localization

Salivary glands from third instar Oregon-R larvae grown at 18°C on medium containing 10% dextrose, 10% dry nutritional yeast, and 1.2% agarose were isolated and squashed according to the procedure of Atherton and Gall (1972) as described in Ashburner (1989). Biotinylated probes prepared using biotin-16-dUTP (Boehringer Mannheim, Indianapolis, IN) were hybridized to the squashes, and the resulting signals detected by peroxidase staining using the Detek-I-HRP kit (Enzo Biochemicals, New York, NY) as described in Ashburner (1989). Slides were photographed with a Zeiss Axiophot using Kodak Tmax 400 film.

Results

Isolation of cDNA Clones for the Major D. Melanogaster hnRNP Proteins

Proteins isolated from *D. melanogaster* embryo lysate by ssDNA-chromatography were used as antigens to generate antibodies to hrp proteins (see Matunis et al., 1992, accompanying paper, for the designation of the proteins), as described in detail in Matunis et al. (1992*a*). A λgt11 *D. melanogaster* embryo cDNA library was screened with the resultant polyclonal antisera, and a large collection of clones was obtained. We were interested in identifying clones encoding homologues of the well-characterized mammalian A, B, and C hnRNP proteins that migrate between 30–40 kD on SDS-polyacrylamide gels (Beyer et al., 1977; Choi and Dreyfuss, 1984; Leser et al., 1984; Wilk et al., 1985; Matunis et al., 1992*a*). Therefore, clones encoding lacZ-fusion proteins which affinity-selected antibodies to proteins with an apparent molecular mass in this range were further characterized. Among these were clones designated pHRP40.2 and pHRP38.1. Additional hnRNP cDNA clones were obtained by screening the collection of clones isolated using the polyclonal antisera with mAbs that became available later. Several clones reacted with the mAb 4C2 (which recognizes the hrp40 proteins) and these were classified into two different groups, pHRP40.1 and pHRP40.2, which differed at the 3′ end both by restriction mapping and sequence analysis. The mAb 5A5 (specific for hrp36) identified the clone designated

Table I. cDNA Clones, mAbs, and Cytological Positions for the Major D. Melanogaster hnRNP Proteins

Protein	Antibody	Clone	Cytological position
hrp48	10D5	pHRP48.1	27C
hrp40	8G6, 4C2, 8D2	pHRP40.1	87F
		pHRP40.2	87F
hrp38	5A5	pHRP38.1	98DE
hrp36	5A5	pHRP36.1	87F

All of the antibodies react with the indicated proteins both in immunoprecipitations and on immunoblots after SDS-PAGE except for hrp38 which does not react with 5A5 on immunoblots.

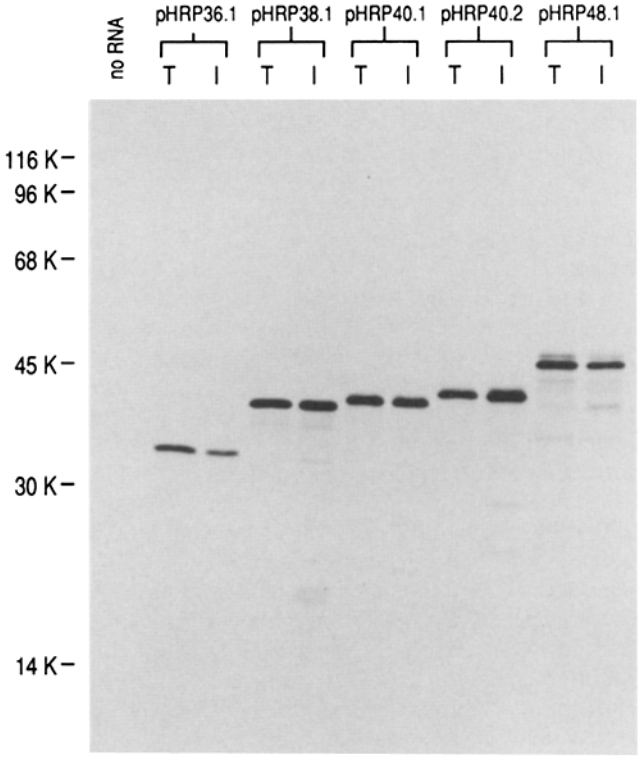


Figure 1. In vitro transcription/translation and immunoprecipitation of pHRP36.1, pHRP38.1, pHRP40.1, pHRP40.2, and pHRP48.1 cDNA clones. Each clone was linearized at a restriction site located downstream of the coding region and transcribed with SP6 or T7 polymerase. The RNAs were translated in reticulocyte lysate, and aliquots of the total translation reactions (lanes *T*) or the proteins immunoprecipitated with mAbs (lanes *I*) were analyzed by SDS-PAGE and fluorography. No DNA template was added to the transcription reaction used in the translation shown in the lane "no RNA." Sizes are given in kD.

pHRP36.1. The mAb 10D5 (which is specific for hrp48) did not recognize any of the clones isolated using the polyclonal antisera, and was used to isolate the cDNA clone designated pHRP48.1 from the λgt11 expression library. When necessary, partial cDNA clones were used as hybridization probes to obtain full-length clones. Several cDNA clones differing slightly by sequence and restriction maps were obtained with each antibody, as exemplified by the clones encoding hrp40 proteins (pHRP40.1 and pHRP40.2). For the remaining hrp proteins characterized here, a single representative member of each family was chosen for further analysis. These full-length clones are listed in Table I.

The plasmids containing full-length cDNAs were analyzed by transcription and translation in vitro. The translation products from the coding strand of each cDNA are shown in Fig. 1 (lanes *T*). No proteins were produced from the opposite strand of each clone (data not shown). To verify the identity of the proteins encoded by these cDNA clones, the translation products of each cDNA were immunoprecipitated with specific monoclonal antibodies as shown in Fig. 1 (lanes *I*). Each of the cDNA clones encoded a protein of the expected size range based on the reactivity of the antibody used for its isolation, and the proteins were immunoprecipitated by the corresponding mAb (see Table I), but not with nonimmune control ascites fluid from mice inoculated with the parent myeloma cell line SP2/0 (data not shown).

To further confirm the identity of the proteins encoded by the individual clones, the translation products were analyzed by two-dimensional gel electrophoresis (Fig. 2, *left panel*). [^{35}S]methionine-labeled proteins isolated by ssDNA chromatography from *D. melanogaster* Schneider's line 2 (S2) cells were analyzed simultaneously for comparison (Fig. 2, *right panel*). The proteins produced in vitro were identified by superimposing the two autoradiograms in Fig. 2. All of the proteins within a dashed box are immunologically related to each other (Matunis et al., 1992*a*). The protein encoded by pHRP48.1 (recognized by the mAb 10D5) co-migrates with the most basic 48-kD protein within the hrp48 cluster. Proteins encoded by the pHRP40.1 and pHRP40.2 clones co-migrate precisely with two of the hrp40 proteins. The highest molecular weight most basic protein in the cluster, a minor constituent of the hrp40 group of proteins which is visible on a longer exposure of this autoradiogram (see Matunis et al., 1992*a*), corresponds to pHRP40.2. The abundant, more acidic protein with a molecular weight slightly less than the protein encoded by pHRP40.2 is encoded by pHRP40.1. The protein produced from the clone pHRP38.1 is of the same relative molecular mass as the most prominent hrp38 protein, and is slightly more acidic. pHRP36.1 encodes a protein with an apparent molecular mass of ~34 kD which migrates at the acidic region of the hrp36 cluster. In S2 cells, this protein is visible on a longer exposure of the autoradiogram in the left panel of Fig. 2.

The cytological localization of the genes encoding these proteins was determined by in situ hybridization of the *D. melanogaster* hnRNP cDNA clones to polytene chromosomes isolated from third instar Oregon-R larvae (Fig. 3). A single band of hybridization was obtained on the polytene chromosomes with each cDNA, suggesting that these proteins are encoded by single-copy genes. pHRP38.1 hybridizes to 98DE on the right arm of chromosome 3 (Fig. 3 *A*). The hybridization signal from pHRP48.1 is a 27C on the left arm of chromosome 2 (Fig. 3 *B*). The remaining clones, pHRP40.1, pHRP40.2, and pHRP36.1 all localize to the right arm of chromosome 3 at 87F. These three cDNAs were used as hybridization probes both individually and in pairwise combinations, and the signals completely overlapped (data not shown). The hybridization signal from pHRP40.2 is shown in Fig. 3 *C*. The results of these and the preceding experiments are summarized in Table I.

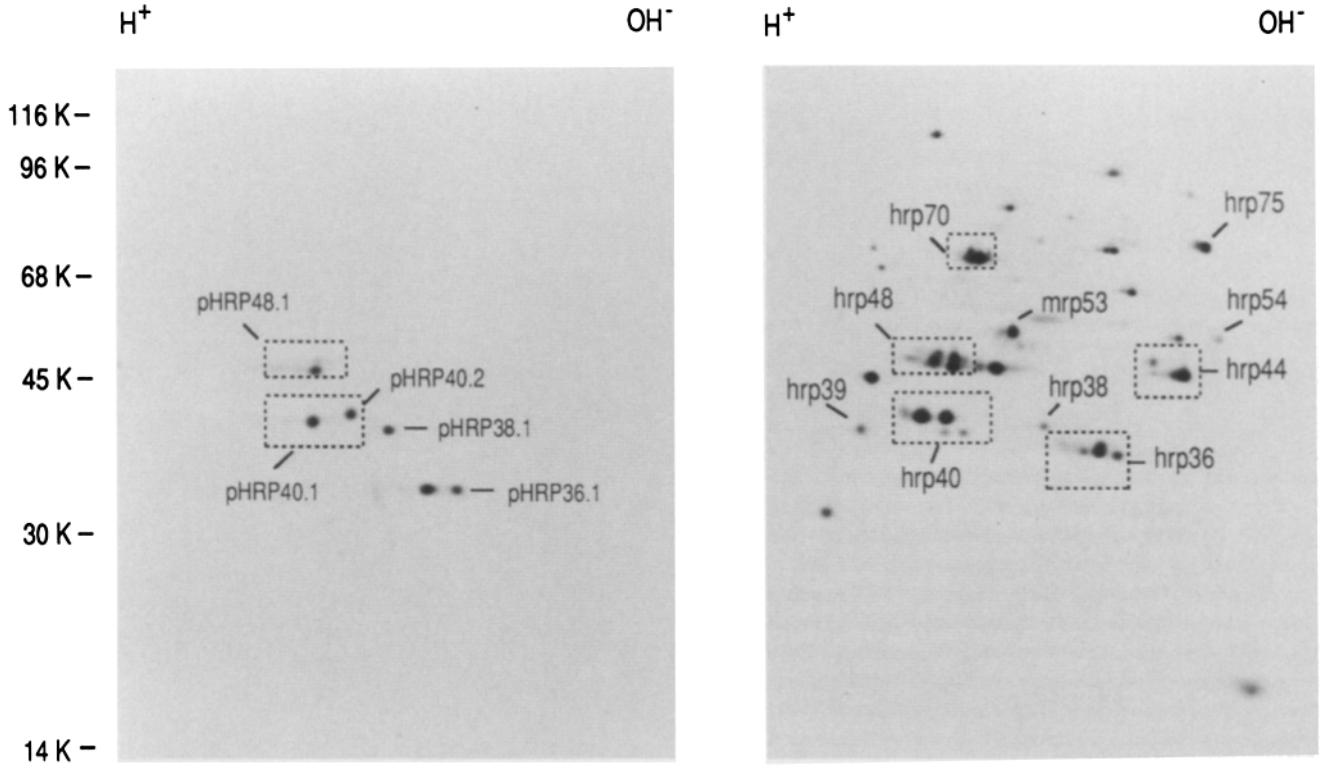


Figure 2. Two dimensional electrophoresis of translation products produced in vitro from *D. melanogaster* hnRNP cDNA clones. Aliquots of the translation products shown in Fig. 1 were run on two-dimensional gels (NEPHGE in the first dimension and SDS-PAGE in the second dimension) and compared to the position of ssDNA binding proteins purified from [^{35}S]methionine-labeled S2 cells (*right panel*). No polypeptides were visible from an equal amount of reticulocyte lysate when exogenous RNA was omitted (not shown).

The Predicted Amino Acid Sequences of the Major hnRNP Proteins

The nucleotide and predicted amino acid sequences of the *D. melanogaster* hnRNP protein cDNA clones pHRP48.1, pHRP40.1, pHRP40.2, and pHRP36.1 are shown in Fig. 4. The start of translation for each protein encoded by these clones is proposed to be the first ATG, which is preceded by an in-frame stop codon in each case, and each long open reading frame is predicted to encode a protein of a molecular mass and pI that is in agreement with values observed for both the in vivo and in vitro produced proteins shown in Fig. 2. Surprisingly, the predicted amino acid sequences of all of these proteins have a similar modular primary structure (Bandziulis et al., 1989); two amino-terminal RNP-CS RNA binding domains (CS-RBDs), and a carboxyl-terminal glycine-rich auxiliary domain, which we abbreviate as 2×RBD-Gly.

Fig. 4 *A* shows the nucleotide and predicted amino acid sequence of pHRP48.1. The proposed start site for translation is at nucleotide 87, and is in agreement with the consensus for *D. melanogaster* translation start sites [$^{C}/_{A}AA^{A}/_{C}ATG$] (Cavener, 1987) in three of the four positions preceding the ATG. This starting position predicts a protein composed of 386 amino acids with a molecular mass of 41,027 D and a pI of 6.7. This protein shares ~35% identity to all of the known 2×RBD-Gly proteins. Interestingly, when compared to all the predicted protein sequences in the Genbank and EMBL databases, the *D. melanogaster* hrp48 protein most closely resembles Nrpl, a protein related to the hnRNP A/B proteins which is expressed in the nervous system of developing frogs (Richter et al., 1990). Nrpl and hrp48 differ from other 2×RBD-Gly proteins in their auxiliary domains. This domain in hrp48 is composed primarily of glycine (22%), serine (12%), proline (10%), and alanine (9%). The four most abundant amino acids in the carboxyl-terminal domain of Nrpl are also glycine (14%), alanine (14%), proline (10%), and serine (9%). The relatively high alanine and proline content and lower glycine content of these two proteins distinguishes hrp48 and Nrpl from the other members of the 2×RBD-Gly protein family, since alanine and proline are rarely found in the glycine-rich domains of other 2×RBD-Gly proteins (see below).

The nucleotide and predicted amino acid sequences of pHRP40.1 and pHRP40.2 are shown in Fig. 4 (*B* and *C*). The proposed start site for translation in each clone is at nucleotide 95, and conforms exactly with the consensus for *D. melanogaster* translation start sites (Cavener, 1987). This starting position predicts a protein composed of 322 amino acids with a molecular mass of 35,000 D and a pI of 6.1 for pHRP40.1, and a 346 amino acid 36,238 D protein with a pI of 7.2 for pHRP40.2. The two predicted protein sequences are identical up to amino acid 285 (nt 951) as indicated by the arrow in Fig. 4 (*B* and *C*), which is near the middle of the carboxyl-terminal glycine-rich domain. Thus, these proteins differ in their auxiliary domain which is composed primarily of glycine (30%), tyrosine (18%), and asparagine (15%) in pHRP40.1, and glycine (49%), tyrosine (14%), and

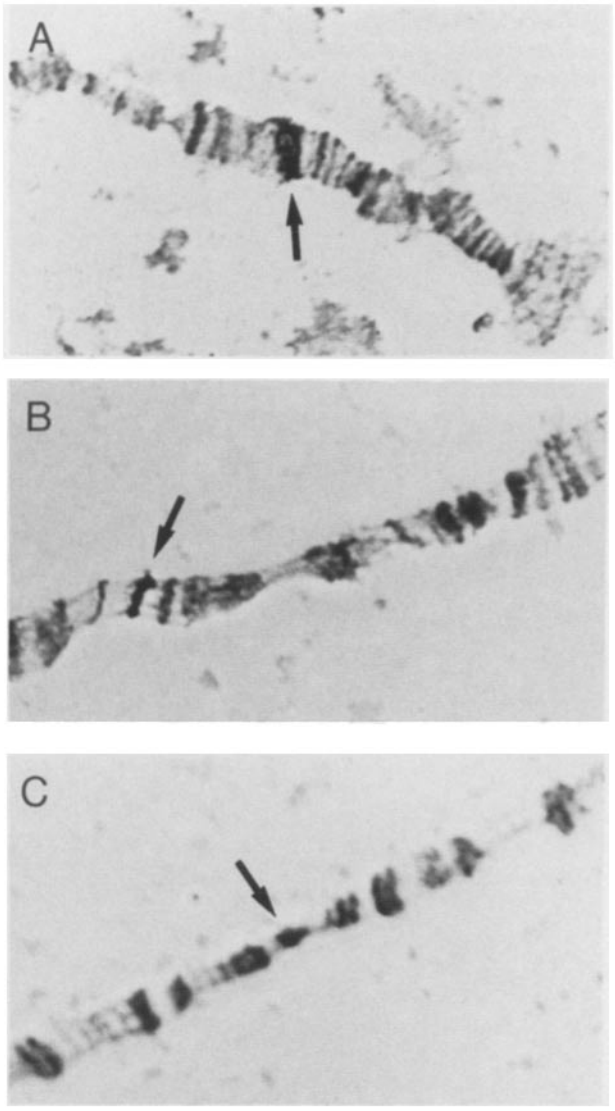


Figure 3. Cytological localization of the *D. melanogaster* hnRNP cDNA clones. Hybridization of biotinylated cDNA probes to polytene chromosomes was detected using peroxidase staining and is indicated by an arrow. The hybridization signal is readily detectable on the original slide due to a color difference between the chromosomes and the reaction product which does not reproduce well in black and white. The signal resulting from pHRP38.1 is localized to 98DE (*A*), pHRP48.1 maps to 27C (*B*), and the signal from pHRP40.1 is at 87F (*C*).

arginine (8%) in pHRP40.2. Three copies of the repeat GRGGX (X = Y or P) are found in both of these proteins. This repeat occurs frequently in many RNA-binding proteins (Matunis et al., 1992*b*) and is a potential site for the dimethylation of arginine in certain nuclear RNA-binding proteins (Christensen and Fuxa, 1988). Interestingly, from the point of divergence at amino acid 285, the auxiliary domain of pHRP40.2 remains similar in overall composition to that of the entire glycine-rich domain, while pHRP40.1 is asparagine and serine-rich rather than glycine-rich. There are three other differences between the two sequences at nucleotides 79, 655, and 667. These G to A changes do not affect the sequences of the predicted proteins, and probably represent polymorphisms. These two clones are most likely alternatively spliced forms of the same gene, as discussed below. The proteins encoded by pHRP40.1 and pHRP40.2 are ~37% identical to all of the known members of the 2×RBD-Gly family.

The nucleotide and predicted amino acid sequence of pHRP36.1 is shown in Fig. 4 *D*. The proposed start site for translation is at nucleotide 27, which is in a reasonable context for *D. melanogaster* translation initiation (Cavener, 1987). This starting position predicts a protein composed of 327 amino acids with a molecular mass of 33,725 D and a pI of 8.7. The carboxyl-terminal domain of the protein predicted by pHRP36.1 is composed primarily of glycine (56%), asparagine (11%), and glutamine (8%), and has two GRGGX (X = P or Q) repeats. This glycine content is the highest found among the auxiliary domains of the 2×RBD-Gly proteins. Unlike the other predicted amino acid sequences presented here, the protein encoded by pHRP36.1 has a high degree of amino acid identity (66%) with the 2×RBD-Gly proteins which are encoded by the *Hrb98DE* locus (see Discussion). The hrp36 proteins are ~37% identical to the remaining members of the 2×RBD-Gly protein family.

Restriction mapping and nucleotide sequencing of pHRP38.1 indicates that the encoded protein, although highly related to that encoded by pHRP36.1, is derived from a separate gene (data not shown). pHRP38.1 corresponds to the previously described *D. melanogaster* cDNA clone p9, which was isolated by hybridization screening with sequences containing a GGN or *pen* repeat (Haynes et al., 1987). Both pHRP38.1 and p9 map to 98DE on polytene chromosomes in situ, and are derived from the *Hrb98DE* locus, which is predicted to encode four protein isoforms varying slightly in molecular mass and charge (Haynes et al., 1990). Therefore, hrp38 most likely corresponds to an isoform of *Hrb98DE*. Additional cDNA clones encoding hrp proteins are being characterized. Partial sequencing indicates that a clone immunologically related to hrp44 encodes a glycine-rich protein with at least one CS-RBD.

2×RBD-Gly Proteins Are Major hnRNP Proteins in Divergent Eukaryotes

To further analyze the primary structures of the *D. melanogaster* hnRNP proteins presented in Fig. 4, the portions of each predicted amino acid sequence corresponding to the RBDs were aligned with the RBDs of 2×RBD-Gly proteins from highly divergent organisms, including *D. melanogaster*, grasshopper, frog, and man (Fig. 5). In addition to previously identified conserved positions within the RBDs of many RNP-CS proteins (Bandziulis et al., 1989), this alignment reveals several positions which are highly conserved specifically in the RBDs of 2×RBD-Gly proteins (Fig. 5, lines "RBD consensus"). Also, comparison of the two lines "RBD1 consensus" and "RBD2 consensus" demonstrates that, in addition to the conserved amino acids shared by the two domains (shown within the gray boxes), each RBD is conserved individually.

Further evidence of the relatedness of 2×RBD-Gly proteins is derived from the reactivity of the anti-hrp40 mAb 4C2 with the human A, B, and G hnRNP proteins (Matunis et al., 1992*a*). In addition, several immunologically related

A

CS-RBD I

CS-RBD II

```
                                                                CGCTTTT     7
TTATTAACACGGCAAGATTACGTCAATTATTGAGAACCTACCCGGCGCCATAGCTCACAACCACAACCAAAACAACATC    86

ATG GAG GAA GAC GAG AGG GGC AAA CTT TTT GTG GGC GGT CTA TCC TGG GAG ACG ACG CAG   146
Met Glu Glu Asp Glu Arg Gly Lys Leu Phe Val Gly Gly Leu Ser Trp Glu Thr Thr Gln    20

GAG AAC CTG TCG CGC TAC TTC TGC CGC TTC GGG GAC ATC ATT GAC TGT GTG GTG ATG AAG   206
Glu Asn Leu Ser Arg Tyr Phe Cys Arg Phe Gly Asp Ile Ile Asp Cys Val Val Met Lys    40

AAC AAC GAG AGC GGC AGG TCG CGC GGC TTT GGC TTC GTT ACC TTC GCC GAT CCC ACC AAC   266
Asn Asn Glu Ser Gly Arg Ser Arg Gly Phe Gly Phe Val Thr Phe Ala Asp Pro Thr Asn    60

GTC AAC CAC GTG CTG CAG AAC GGA CCG CAC ACG CTC GAC GGT CGC ACC ATC GAC CCC AAG   326
Val Asn His Val Leu Gln Asn Gly Pro His Thr Leu Asp Gly Arg Thr Ile Asp Pro Lys    80

CCG TGC AAT CCG CGC ACT CTG CAG AAA CCG AAG AAG GGC GGC GGC TAC AAG GTC TTC CTG   386
Pro Cys Asn Pro Arg Thr Leu Gln Lys Pro Lys Lys Gly Gly Gly Tyr Lys Val Phe Leu   100

GGT GGC CTG CCC TCG AAC GTC ACC GAG ACC GAT CTG CGG ACC TTC TTT AAC CGC TAC GGC   446
Gly Gly Leu Pro Ser Asn Val Thr Glu Thr Asp Leu Arg Thr Phe Phe Asn Arg Tyr Gly   120

AAG GTC ACC GAG GTG GTT ATC ATG TAC GAC CAG GAG AAG AAG AAG TCC CGC GGC TTC GGC   506
Lys Val Thr Glu Val Val Ile Met Tyr Asp Gln Glu Lys Lys Lys Ser Arg Gly Phe Gly   140

TTT CTC TCC TTC GAG GAG GAG TCC TCC GTT GAG CAC GTG ACC AAC GAG CGG TAC ATC AAT   566
Phe Leu Ser Phe Glu Glu Glu Ser Ser Val Glu His Val Thr Asn Glu Arg Tyr Ile Asn   160

CTG AAT GGC AAG CAG GTC GAA ATC AAG AAG GCC GAG CCT CGT GAT GGA TCT GGC GGC CAA   626
Leu Asn Gly Lys Gln Val Glu Ile Lys Lys Ala Glu Pro Arg Asp Gly Ser Gly Gly Gln   180

AAC TCC AAC AAC AGT ACC GTG GGA GGC GCC TAT GGC AAG CTT GGT AAC GAG TGC AGC CAC   686
Asn Ser Asn Asn Ser Thr Val Gly Gly Ala Tyr Gly Lys Leu Gly Asn Glu Cys Ser His   200

TGG GGA CCG CAC CAT GCT CCC ATC AAC ATG ATG CAG GGC CAG AAT GGC CAG ATG GGT GGA   746
Trp Gly Pro His His Ala Pro Ile Asn Met Met Gln Gly Gln Asn Gly Gln Met Gly Gly   220

CCG CCG CTG AAT ATG CCC ATT GGA GCG CCG AAT ATG ATG CCT GGC TAT CAG GGT TGG GGC   806
Pro Pro Leu Asn Met Pro Ile Gly Ala Pro Asn Met Met Pro Gly Tyr Gln Gly Trp Gly   240

ACC TCG CCG CAG CAG CAA CAA TAC GGC TAC GGC AAC AGT GGC CCA GGA TCG TAC CAG GGA   866
Thr Ser Pro Gln Gln Gln Gln Tyr Gly Tyr Gly Asn Ser Gly Pro Gly Ser Tyr Gln Gly   260

TGG GGA GCT CCT TCG GGA CCA TCG GGC GGC GGC AGC TGG AAC TCG TGG AAC ATG CCA CCT   926
Trp Gly Ala Pro Ser Gly Pro Ser Gly Gly Gly Ser Trp Asn Ser Trp Asn Met Pro Pro   280

AAC TCT GCC GGA CCC ACT GGG GCA CCA GGA GCC GGA GCG GGC ACC GCC ACT GAC ATG TAC   986
Asn Ser Ala Gly Pro Thr Gly Ala Pro Gly Ala Gly Ala Gly Thr Ala Thr Asp Met Tyr   300

TCG CGT GCT CAA GCC TGG GCG ACG GGC GGT CCC TCG ACC ACT GGA CCA GTG GGC GGC ATG  1046
Ser Arg Ala Gln Ala Trp Ala Thr Gly Gly Pro Ser Thr Thr Gly Pro Val Gly Gly Met   320

CCC CGG ACC GGA CCC GGT AAT TCG GCC TCC AAG TCT GGC TCT GAG TAC GAC TAC GGC GGC  1106
Pro Arg Thr Gly Pro Gly Asn Ser Ala Ser Lys Ser Gly Ser Glu Tyr Asp Tyr Gly Gly   340

TAT GGA TCC GGG TAC GAC TAC GAC TAC AGC AAC TAT GTG AAG CAG GAG GGC GCC TCG AAC  1166
Tyr Gly Ser Gly Tyr Asp Tyr Asp Tyr Ser Asn Tyr Val Lys Gln Glu Gly Ala Ser Asn   360

TAC GGA GCA GGG CCG CGA TCA GCG TAC GGC AAC GAC AGC TCC ACG CAG CCA CCC TAT GCA  1226
Tyr Gly Ala Gly Pro Arg Ser Ala Tyr Gly Asn Asp Ser Ser Thr Gln Pro Pro Tyr Ala   380

ACC TCG CAG GCT GTC TAA AGAGAGAAGGATCGTGTGGAGGAGAAGTAGGAGTAGCGGAGCAGAGAGGAGCGGA  1299
Thr Ser Gln Ala Val  *                                                          385

GTAGCATGTCGTCCTCGGCGAATGATGAAGAAGACAGAAAAGATGGTGGTAATTGTAAGCTAAAGCGTCGTCTGTTCGT  1378
TTTTATTTCAGATCAGATCAGTTGTTGAAATTAATATTTAAAACTCTTGTACAACTATTATTTAAAGCGAAAAAAAAAT  1457
TTAACATGAGAATAACTAGCGACAAGTCAGATTCATAGATATACATGTATGTTCTATAGCAGCGAGTATGTAATTCATA  1536
GGCGAAACATCAGCACTCTCACCCATTTAGACAAGAGAACATTATTTCTTTCATGCGCCTCATAAACACCACACTCAAC  1615
AAGCATTCTCCTGTACCCCACACTCACCAACAAACCAACCAACCAGCCAGGATCAACTAACAAGATCTTCCCAACCAAC  1694
TAAAACCCGTAGAC                                                                  1708
```

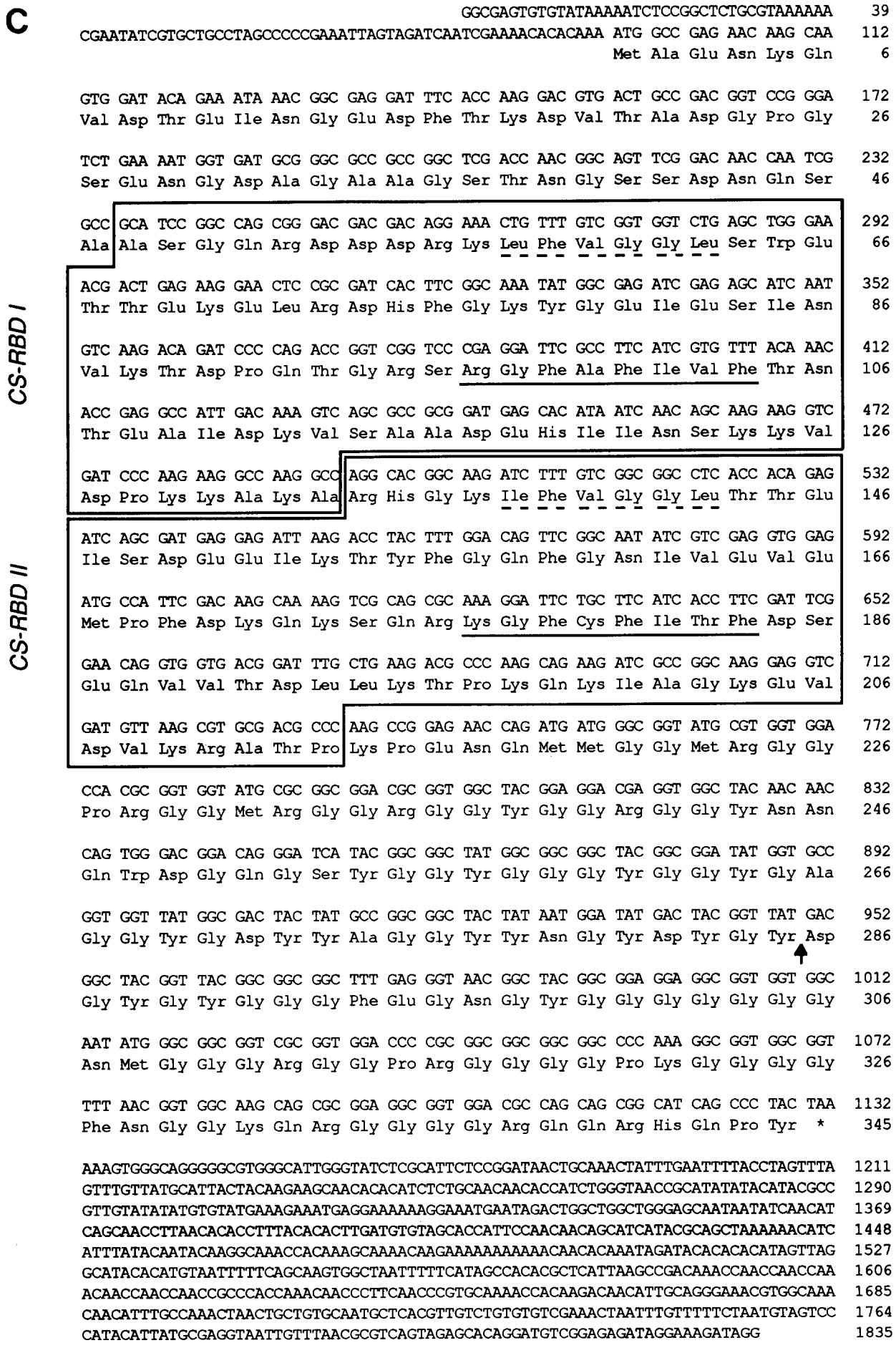
B

```
                                                                    GGCGAGTGTGTATAAAAATCTCCGGCTCTGCGTAAAAAA    39

CGAATATCGTGCTGCCTAGCCCCCGAAATTAGTAGATCAGTCGAAAACACACAAA ATG GCC GAG AAC AAG CAA   112
                                                          Met Ala Glu Asn Lys Gln     6

GTG GAT ACA GAA ATA AAC GGC GAG GAT TTC ACC AAG GAC GTG ACT GCC GAC GGT CCG GGA   172
Val Asp Thr Glu Ile Asn Gly Glu Asp Phe Thr Lys Asp Val Thr Ala Asp Gly Pro Gly    26

TCT GAA AAT GGT GAT GCC GGC GCC GCC GGC TCG ACC AAC GGC AGT TCG GAC AAC CAA TCG   232
Ser Glu Asn Gly Asp Ala Gly Ala Ala Gly Ser Thr Asn Gly Ser Ser Asp Asn Gln Ser    46

GCC GCA TCC GGC CAG CGG GAC GAC GAC AGG AAA CTG TTT GTC GGT GGT CTG AGC TGG GAA   292
Ala Ala Ser Gly Gln Arg Asp Asp Asp Arg Lys Leu Phe Val Gly Gly Leu Ser Trp Glu    66
                                            - - - - - - - - - - - -

ACG ACT GAG AAG GAA CTC CGC GAT CAC TTC GGC AAA TAT GGC GAG ATC GAG AGC ATC AAT   352
Thr Thr Glu Lys Glu Leu Arg Asp His Phe Gly Lys Tyr Gly Glu Ile Glu Ser Ile Asn    86

GTC AAG ACA GAT CCC CAG ACC GGT CGG TCC CGA GGA TTC GCC TTC ATC GTG TTT ACA AAC   412
Val Lys Thr Asp Pro Gln Thr Gly Arg Ser Arg Gly Phe Ala Phe Ile Val Phe Thr Asn   106
                                        ___________________

ACC GAG GCC ATT GAC AAA GTC AGC GCC GCG GAT GAG CAC ATA ATC AAC AGC AAG AAG GTC   472
Thr Glu Ala Ile Asp Lys Val Ser Ala Ala Asp Glu His Ile Ile Asn Ser Lys Lys Val   126

GAT CCC AAG AAG GCC AAG GCC AGG CAC GGC AAG ATC TTT GTC GGC GGC CTC ACC ACA GAG   532
Asp Pro Lys Lys Ala Lys Ala Arg His Gly Lys Ile Phe Val Gly Gly Leu Thr Thr Glu   146
                                            - - - - - - - - - - - -

ATC AGC GAT GAG GAG ATT AAG ACC TAC TTT GGA CAG TTC GGC AAT ATC GTC GAG GTG GAG   592
Ile Ser Asp Glu Glu Ile Lys Thr Tyr Phe Gly Gln Phe Gly Asn Ile Val Glu Val Glu   166

ATG CCA TTC GAC AAG CAA AAG TCG CAG CGC AAA GGA TTC TGC TTC ATC ACC TTC GAT TCG   652
Met Pro Phe Asp Lys Gln Lys Ser Gln Arg Lys Gly Phe Cys Phe Ile Thr Phe Asp Ser   186
                                        ___________________

GAG CAG GTG GTG ACA GAT TTG CTG AAG ACG CCC AAG CAG AAG ATC GCC GGC AAG GAG GTC   712
Glu Gln Val Val Thr Asp Leu Leu Lys Thr Pro Lys Gln Lys Ile Ala Gly Lys Glu Val   206

GAT GTT AAG CGT GCG ACG CCC AAG CCG GAG AAC CAG ATG ATG GGC GGT ATG CGT GGT GGA   772
Asp Val Lys Arg Ala Thr Pro Lys Pro Glu Asn Gln Met Met Gly Gly Met Arg Gly Gly   226

CCA CGC GGT GGT ATG CGC GGC GGA CGC GGT GGC TAC GGA GGA CGA GGT GGC TAC AAC AAC   832
Pro Arg Gly Gly Met Arg Gly Gly Arg Gly Gly Tyr Gly Gly Arg Gly Gly Tyr Asn Asn   246

CAG TGG GAC GGA CAG GGA TCA TAC GGC GGC TAT GGC GGC GGC TAC GGC GGA TAT GGT GCC   892
Gln Trp Asp Gly Gln Gly Ser Tyr Gly Gly Tyr Gly Gly Gly Tyr Gly Gly Tyr Gly Ala   266

GGT GGT TAT GGC GAC TAC TAT GCC GGC GGC TAC TAT AAT GGA TAT GAC TAC GGT TAT GGT   952
Gly Gly Tyr Gly Asp Tyr Tyr Ala Gly Gly Tyr Tyr Asn Gly Tyr Asp Tyr Gly Tyr Gly   286
                                                                          ▲

AAA TAC AAC AAA CAA CAA AGT AGT GCT CAA AAC AAT TAT TAC AAT AAC AAC ACG TCG AGT  1012
Lys Tyr Asn Lys Gln Gln Ser Ser Ala Gln Asn Asn Tyr Tyr Asn Asn Asn Thr Ser Ser   306

AAT TAC CAT CAA AAC AAG AAC AAT AGC AAC AAC TAT CAG CAG TTC TAA GTAGAGCCTCAGCCG  1075
Asn Tyr His Gln Asn Lys Asn Asn Ser Asn Asn Tyr Gln Gln Phe  *                    321

CATAGAGCTGGCCTGGCCAGCAACAAAATCAGAAAGAAATAATCAAACAAAGCAAACAACAACAACCGC          1144
```

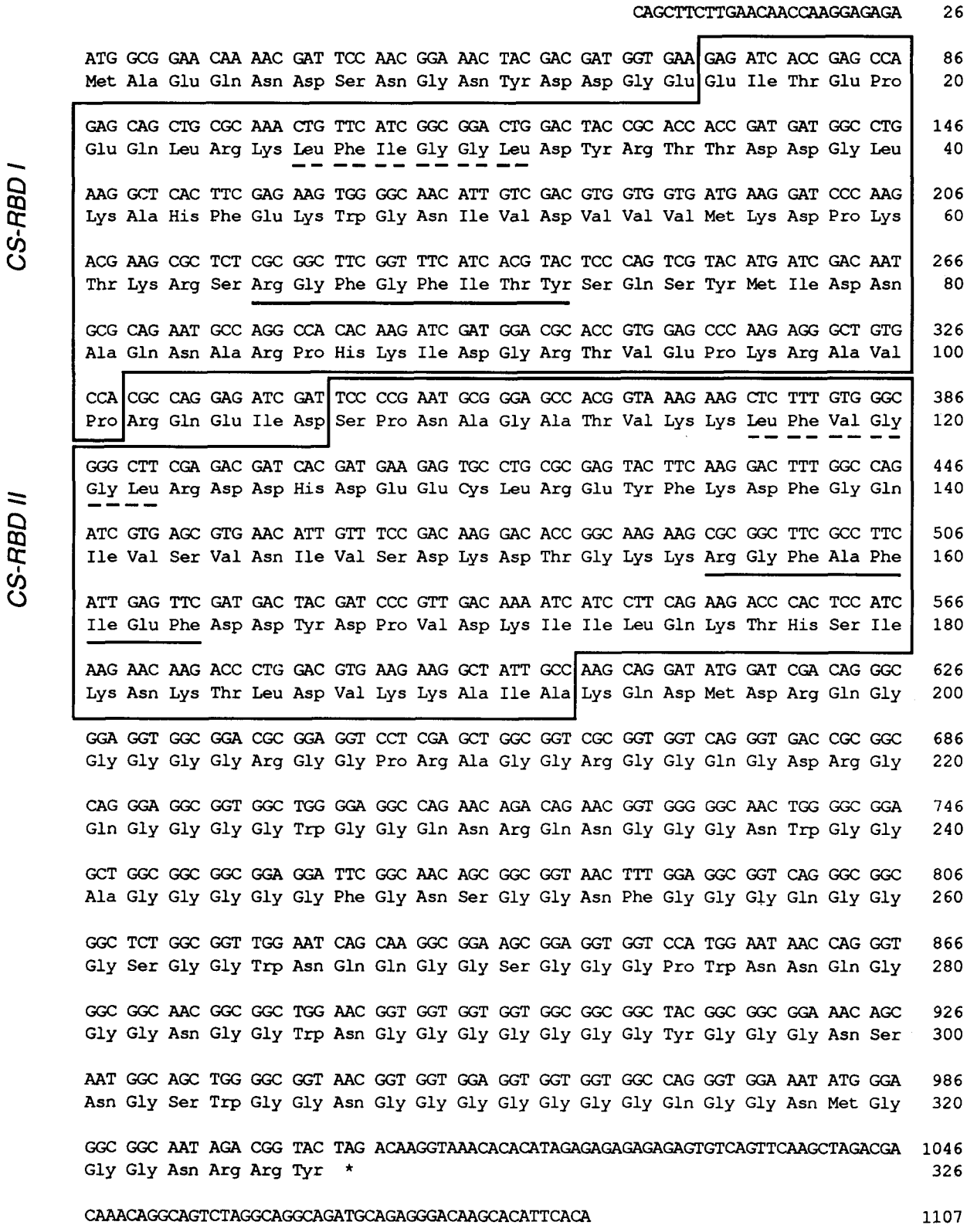
CS-RBD I

CS-RBD II

Figure 4. Nucleotide and predicted amino acid sequence of pHRP48.1 (*A*), pHRP40.1 (*B*), pHRP40.2 (*C*), and pHRP36.1 (*D*) cDNA clones. The consensus sequence-RNA binding domains (CS-RBD I and II) are boxed. Within each CS-RBD, the two most highly conserved sequence elements, RNP-CS (*underlined*) and RNP2 (*dashed line*) are indicated. The arrows in *B* and *C* mark the point from which these two sequences differ. These sequence data are available from EMBL/GenBank/DDBJ under accession numbers X62636 for hrp36.1, X62637 for hrp 40.1, X62638 for hrp 40.2 and X62639 for hrp48.1.


C
CS-RBD I
CS-RBD II
GGCGAGTGTGTATAAAAATCTCCGGCTCTGCGTAAAAAA 39
CGAATATCGTGCTGCCTAGCCCCCGAAATTAGTAGATCAATCGAAAACACACAAA ATG GCC GAG AAC AAG CAA 112
Met Ala Glu Asn Lys Gln 6
GTG GAT ACA GAA ATA AAC GGC GAG GAT TTC ACC AAG GAC GTG ACT GCC GAC GGT CCG GGA 172
Val Asp Thr Glu Ile Asn Gly Glu Asp Phe Thr Lys Asp Val Thr Ala Asp Gly Pro Gly 26
TCT GAA AAT GGT GAT GCG GGC GCC GCC GGC TCG ACC AAC GGC AGT TCG GAC AAC CAA TCG 232
Ser Glu Asn Gly Asp Ala Gly Ala Ala Gly Ser Thr Asn Gly Ser Ser Asp Asn Gln Ser 46
GCC GCA TCC GGC CAG CGG GAC GAC GAC AGG AAA CTG TTT GTC GGT GGT CTG AGC TGG GAA 292
Ala Ala Ser Gly Gln Arg Asp Asp Asp Arg Lys Leu Phe Val Gly Gly Leu Ser Trp Glu 66
ACG ACT GAG AAG GAA CTC CGC GAT CAC TTC GGC AAA TAT GGC GAG ATC GAG AGC ATC AAT 352
Thr Thr Glu Lys Glu Leu Arg Asp His Phe Gly Lys Tyr Gly Glu Ile Glu Ser Ile Asn 86
GTC AAG ACA GAT CCC CAG ACC GGT CGG TCC CGA GGA TTC GCC TTC ATC GTG TTT ACA AAC 412
Val Lys Thr Asp Pro Gln Thr Gly Arg Ser Arg Gly Phe Ala Phe Ile Val Phe Thr Asn 106
ACC GAG GCC ATT GAC AAA GTC AGC GCC GCG GAT GAG CAC ATA ATC AAC AGC AAG AAG GTC 472
Thr Glu Ala Ile Asp Lys Val Ser Ala Ala Asp Glu His Ile Ile Asn Ser Lys Lys Val 126
GAT CCC AAG AAG GCC AAG GCC AGG CAC GGC AAG ATC TTT GTC GGC GGC CTC ACC ACA GAG 532
Asp Pro Lys Lys Ala Lys Ala Arg His Gly Lys Ile Phe Val Gly Gly Leu Thr Thr Glu 146
ATC AGC GAT GAG GAG ATT AAG ACC TAC TTT GGA CAG TTC GGC AAT ATC GTC GAG GTG GAG 592
Ile Ser Asp Glu Glu Ile Lys Thr Tyr Phe Gly Gln Phe Gly Asn Ile Val Glu Val Glu 166
ATG CCA TTC GAC AAG CAA AAG TCG CAG CGC AAA GGA TTC TGC TTC ATC ACC TTC GAT TCG 652
Met Pro Phe Asp Lys Gln Lys Ser Gln Arg Lys Gly Phe Cys Phe Ile Thr Phe Asp Ser 186
GAA CAG GTG GTG ACG GAT TTG CTG AAG ACG CCC AAG CAG AAG ATC GCC GGC AAG GAG GTC 712
Glu Gln Val Val Thr Asp Leu Leu Lys Thr Pro Lys Gln Lys Ile Ala Gly Lys Glu Val 206
GAT GTT AAG CGT GCG ACG CCC AAG CCG GAG AAC CAG ATG ATG GGC GGT ATG CGT GGT GGA 772
Asp Val Lys Arg Ala Thr Pro Lys Pro Glu Asn Gln Met Met Gly Gly Met Arg Gly Gly 226
CCA CGC GGT GGT ATG CGC GGC GGA CGC GGT GGC TAC GGA GGA CGA GGT GGC TAC AAC AAC 832
Pro Arg Gly Gly Met Arg Gly Gly Arg Gly Gly Tyr Gly Gly Arg Gly Gly Tyr Asn Asn 246
CAG TGG GAC GGA CAG GGA TCA TAC GGC GGC TAT GGC GGC GGC TAC GGC GGA TAT GGT GCC 892
Gln Trp Asp Gly Gln Gly Ser Tyr Gly Gly Tyr Gly Gly Gly Tyr Gly Gly Tyr Gly Ala 266
GGT GGT TAT GGC GAC TAC TAT GCC GGC GGC TAC TAT AAT GGA TAT GAC TAC GGT TAT GAC 952
Gly Gly Tyr Gly Asp Tyr Tyr Ala Gly Gly Tyr Tyr Asn Gly Tyr Asp Tyr Gly Tyr Asp 286
GGC TAC GGT TAC GGC GGC GGC TTT GAG GGT AAC GGC TAC GGC GGA GGA GGC GGT GGT GGC 1012
Gly Tyr Gly Tyr Gly Gly Gly Phe Glu Gly Asn Gly Tyr Gly Gly Gly Gly Gly Gly Gly 306
AAT ATG GGC GGC GGT CGC GGT GGA CCC CGC GGC GGC GGC GGC CCC AAA GGC GGT GGC GGT 1072
Asn Met Gly Gly Gly Arg Gly Gly Pro Arg Gly Gly Gly Gly Pro Lys Gly Gly Gly Gly 326
TTT AAC GGT GGC AAG CAG CGC GGA GGC GGT GGA CGC CAG CAG CGG CAT CAG CCC TAC TAA 1132
Phe Asn Gly Gly Lys Gln Arg Gly Gly Gly Gly Arg Gln Gln Arg His Gln Pro Tyr * 345
AAAGTGGGCAGGGGGCGTGGGCATTGGGTATCTCGCATTCTCCGGATAACTGCAAACTATTTGAATTTTACCTAGTTTA 1211
GTTTGTTATGCATTACTACAAGAAGCAACACACATCTCTGCAACAACACCATCTGGGTAACCGCATATATACATACGCC 1290
GTTGTATATATGTGTATGAAAGAAATGAGGAAAAAAAGGAAATGAATAGACTGGCTGGCTGGGAGCAATAATATCAACAT 1369
CAGCAACCTTAACACACCTTTACACACTTGATGTGTAGCACCATTCCAACAACAGCATCATACGCAGCTAAAAAACATC 1448
ATTTATACAATACAAGGCAAACCACAAAGCAAAACAAGAAAAAAAAAAAACAACACAAATAGATACACACACATAGTTAG 1527
GCATACACATGTAATTTTTCAGCAAGTGGCTAATTTTTCATAGCCACACGCTCATTAAGCCGACAAACCAACCAACCAA 1606
ACAACCAACCAACCGCCCACCAAACAACCCTTCAACCCGTGCAAAACCACAAGACAACATTGCAGGGAAACGTGGCAAA 1685
CAACATTTGCCAAACTAACTGCTGTGCAATGCTCACGTTGTCTGTGTGTCGAAACTAATTTGTTTTTCTAATGTAGTCC 1764
CATACATTATGCGAGGTAATTGTTTAACGCGTCAGTAGAGCACAGGATGTCGGAGAGATAGGAAAGATAGG 1835

D



proteins are present in chicken, frog, and yeast cells (Fig. 6). In *X. laevis* and chicken cells, the reactive proteins in the 30–45-kD range have been identified as 2×RBD-Gly proteins (M. Matunis, S. Piñol-Roma, and G. Dreyfuss, manuscript submitted for publication; M. Matunis and G. Dreyfuss, unpublished results). Thus, 2×RBD-Gly proteins are related both by their primary structure and by immunological criteria, and are highly conserved in a wide range of

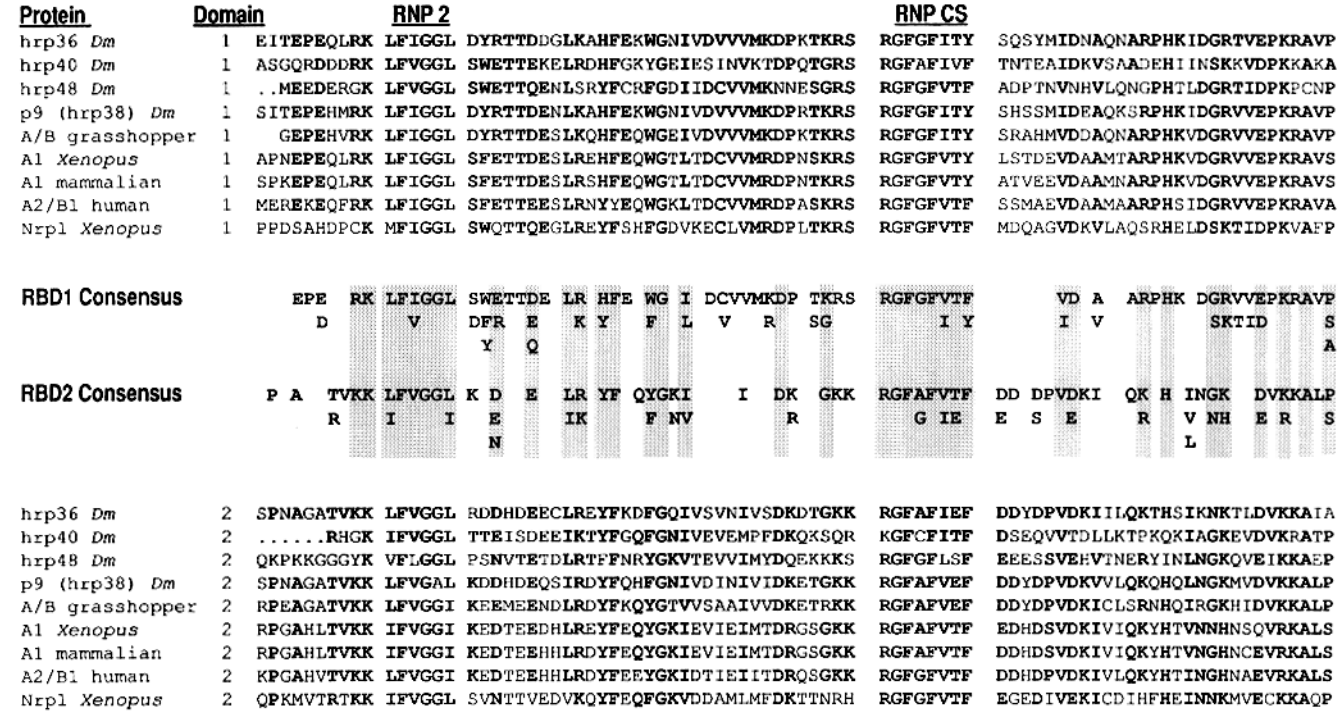


Figure 5. Alignment of the RNA binding domains of 2×RBD-Gly proteins from divergent organisms. The most highly conserved segments of these domains are denoted RNP-CS and RNP2. If an amino acid occurs at a given position in five or more of the nine sequences presented in this figure, it appears in bold type and is listed in the line designated "RBD consensus" either below (*RBD1*) or above (*RBD2*) that position. If, at that same position a second amino acid is present at least twice, it is also in bold type and is listed in the second line of the consensus. Amino acids within the gray shaded boxes in the two consensus lines are conserved in both RBD1 and RBD2. Gaps are indicated with periods (.). Protein sequences other than those presented in this paper were obtained from the following references: *D. melanogaster* clone p9 (Haynes et al., 1987); grasshopper A/B (Ball et al., 1991); *Xenopus* A1 (Kay et al., 1990); mammalian hnRNP A1 from calf thymus (Williams et al., 1985), rat (Cobianchi et al., 1986), and human (Buvoli et al., 1988); human hnRNP A2/B1 (Burd et al., 1989); and *Xenopus* Nrp1 (Richter et al., 1990).

evolutionarily distant organisms, suggesting an essential function.

Discussion

The molecular cloning and DNA sequence analysis of several of the major constituents of hnRNP complexes of *D. melanogaster* presented here reveals that all of these proteins share a similar overall primary structure: two CS-RBDs and a glycine-rich carboxyl-terminal domain, or 2×RBD-Gly. These proteins are thus related to the hnRNP A/B proteins of vertebrates which are also 2×RBD-Gly proteins (Williams et al., 1985; Cobianchi et al., 1986; Buvoli et al., 1988; Burd et al., 1989; Kay et al., 1990). Additional previously described cDNA clones predicted to encode *D. melanogaster* hnRNP proteins also have a similar overall structure (Haynes et al., 1987, 1990, 1991). This work, together with the accompanying paper (Matunis et al., 1992*a*), provides the framework for placing proteins, both characterized and uncharacterized, into the picture of *D. melanogaster* hnRNP complexes. Curiously, although the primary structures of several *D. melanogaster* hnRNP proteins are now available, none of these 2×RBD-Gly proteins display a high enough degree of sequence similarity to any particular vertebrate 2×RBD-Gly protein to allow the unambiguous identification of the vertebrate homologues of the *D. melanogaster* hnRNP proteins.

The alignment of the RNA-binding domains from evolutionarily diverse 2×RBD-Gly proteins (Fig. 5) reveals the high degree of conservation of these domains among the members of this protein family, and establishes a specific consensus sequence for 2×RBD-Gly proteins. It is likely that the general three-dimensional structure of these RBDs is similar to that of the first RBD of U1 snRNP A protein (Nagai et al., 1990; Hoffman et al., 1991). Since the two RBDs of 2×RBD-Gly proteins are immediately adjacent to each other, it is also likely that their spatial positioning is important for their function. We note that the size of each RBD in 2×RBD-Gly proteins is conserved and the spacing between RNP-CS and RNP2 is invariant. These features are characteristic of this particular type of RBD, when compared to other CS-RBDs (Bandziulis et al., 1989; E. Matunis and G. Dreyfuss, unpublished observations). Furthermore, comparison of the RBD1 and RBD2 consensus sequences in Fig. 5 extends the observation that the RBDs in multiple RBD-proteins, such as the poly(A)-binding protein and A1 proteins, are conserved individually; that is, each individual RBD shares a higher percent amino acid identity with the corresponding domain from another species than to any of the other RBDs in the same protein (Dreyfuss et al., 1988). These data suggest that the individual RBDs of these proteins may not be functionally equivalent. Indeed, the four CS-RBDs of the mRNA poly(A)-binding protein have recently been shown to have different RNA-binding activities (Niet-

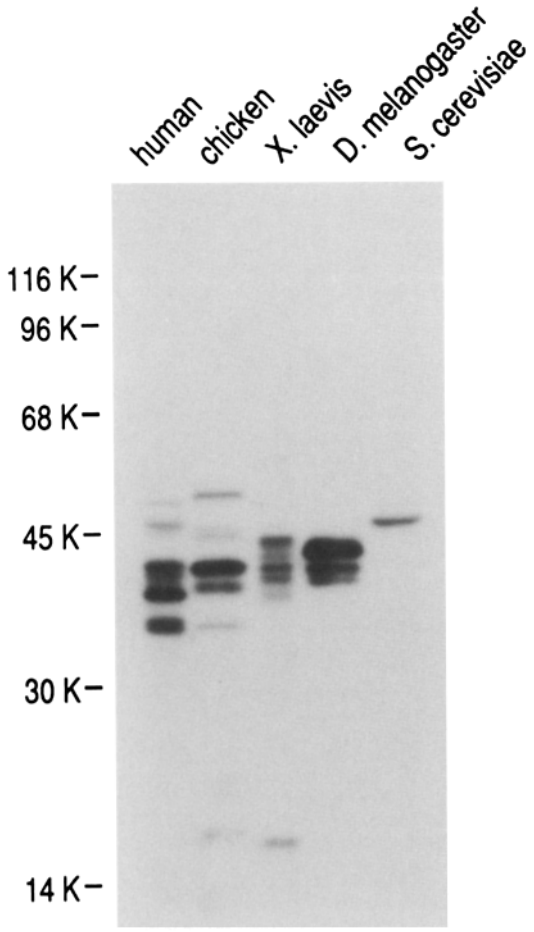


Figure 6. Conservation of 2×RBD-Gly proteins in divergent organisms. Total cellular proteins from the indicated species were resolved by SDS-PAGE, transferred to nitrocellulose, and probed with the mAb 4C2.

feld et al., 1990; Burd et al., 1991). The high degree of conservation of this type of hnRNP protein is also evident from the immunological reactivity of proteins from divergent organisms with the mAb 4C2 (Fig. 6). The immunologically related proteins in vertebrates are known A/B hnRNP proteins. Therefore, the protein reactive with 4C2 in the yeast *Saccharomyces cerevisiae*, from which hnRNP proteins have not yet been described, is a strong candidate hnRNP protein.

Another general theme that emerges about the structure of the hrp proteins is that there are several different isoforms for many of these proteins. The hrp36 proteins are examples of this case, and the different isoforms of these proteins are most likely generated by alternate splicing. During the preparation of this manuscript, Haynes et al. (1991) reported the sequences of cDNA clones from the *Hrb87F* locus, obtained by low-stringency hybridization with sequences derived from a cDNA clone corresponding to the *Hrb98DE* locus (Haynes et al., 1990). Comparison of the sequences of our *D. melanogaster* hnRNP cDNA clone pHRP36.1 with the *Hrb87F*-derived cDNAs indicated that these clones differ by a 180-nt region which is present in the latter but missing in the former clone. The site of this deletion in pHRP36.1 is nucleotide 957, which is flanked by nucleotides that match the *D. melanogaster* splice junction consensus (Keller and Noon, 1985). Also, pHRP36.1 and clones from the *Hrb87F* locus both map cytologically to 87F, supporting the hypothesis that these cDNAs represent variants of a common primary transcript. As a result, the carboxyl-terminal glycine-rich domain of the protein encoded by pHRP36.1 is 60 amino acids shorter than that of *Hrb87F*. Similarly, the cDNA clones pHRP40.1 and pHRP40.2, which are probably derived from a common primary transcript, encode proteins that differ in their carboxyl-terminal domains. Generation of diversity by alternative splicing has been found for several hnRNP 2×RBD-Gly proteins (Buvoli et al., 1988, 1990; Burd et al., 1989; Haynes et al., 1990; Kay et al., 1990; Richter et al., 1990). The functions of this structural diversity are not understood.

A carboxyl-terminal glycine-rich domain is a common feature of many of the major hnRNP proteins. The glycine-rich domain of the mammalian hnRNP A1 has been reported to bind directly to single-stranded nucleic acids (Cobianchi et al., 1988; Kumar et al., 1990) and to have RNA-RNA strand annealing-promoting activity (Kumar and Wilson, 1990). Therefore, a possible explanation for the variation seen in the glycine-rich domains of the hrp proteins is that these differences confer specialized RNA-binding capabilities. In addition, the glycine-rich domain of A1 has been shown to be involved in cooperative binding of A1 to single-stranded nucleic acids (Cobianchi et al., 1988; Kumar et al., 1990), and portions of the glycine-rich domains of the hrp36 and hrp40 proteins are similar to the glycine-rich regions in keratin which are implicated in the assembly of these proteins into filaments (Steinert et al., 1983). This suggests that the glycine-rich domains of the 2×RBD-Gly proteins may be involved in protein-protein interactions. Hence, the variation in the glycine-rich domains of the hrp proteins may alter their ability to interact with other proteins. The differential association of hrp proteins with RNA and/or proteins could give rise to a variety of hnRNP complexes containing subsets of, or varying amounts of, hnRNA-binding proteins. For example, proteins involved in *D. melanogaster* sex determination or other alternative splicing pathways, which are known to bind directly to hnRNA and which do not appear to be stable components of snRNP particles, could be present in a subset of these complexes (reviewed in Baker, 1989). It is also possible that there are tissue- and/or stage-specific hnRNP complexes that contain only specific hrp proteins. For example, the Nrpl protein, detected in the developing *Xenopus laevis* nervous system (Richter et al., 1990), is a 2×RBD-Gly–related protein which is expressed in a tissue- and stage-specific manner. The *D. melanogaster* hrp48 protein most closely resembles Nrpl in overall primary structure. Determining whether the *D. melanogaster* hnRNP proteins, particularly hrp48, are expressed in a developmentally regulated or tissue-specific manner may clarify the relationship between these proteins and Nrpl. We note that the relative abundance of some of the hrp proteins differs between embryo and S2 cell proteins and between different *D. melanogaster* cell lines (M. Matunis and G. Dreyfuss, unpublished results).

It is interesting that all of the hnRNP proteins of *D. melanogaster* characterized thus far are 2×RBD-Gly proteins. This finding is surprising since the predicted primary structures of several mammalian hnRNP proteins represent di-

verse families of proteins: some but not all have CS-RBDs and of these only a subset are 2×RBD-Gly proteins. There are a few abundant *D. melanogaster* hnRNP proteins that have not been characterized yet and we anticipate that they represent other types of hnRNP proteins that are not members of the 2×RBD-Gly family. This includes hrp34, which is a candidate homologue for the hnRNP C proteins, which have one CS-RBD and an acidic auxiliary domain (Swanson et al., 1987; Preugschat and Wold, 1988). The detailed knowledge of the composition of *D. melanogaster* hnRNP complexes and of the structure of the major hnRNP proteins will make it possible to pursue functional studies of these pre-mRNA-binding proteins.

We thank members of our laboratory for comments on this manuscript; Rosemary Hopkins, Mikiko Shiomi, and Miriam Huizinga for their excellent technical assistance; and Dr. Robert Holmgren and members of his lab for their guidance with the *Drosophila* polytene squash preparations. We are grateful to Dr. Bernd Hovemann for providing the *Drosophila* λgt11 library.

This work was supported by grants from the National Institutes of Health and the Howard Hughes Medical Institute.

Received for publication 25 July 1991 and in revised form 23 September 1991.

References

Adam, S. A., T. Nakagawa, M. S. Swanson, T. K. Woodruff, and G. Dreyfuss. 1986. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol. Cell Biol.* 6:2932–2943.

Ashburner, M. 1989. *In Drosophila* A Laboratory Manual. Biotinylated DNA in situ hybridization. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. 434 pp.

Atherton, D., and J. Gall. 1972. Salivary gland squashes for in situ nucleic acid hybridization studies. *Drosophila Inf. Service.* 49:131–133.

Baker, B. S. 1989. Sex in flies: the splice of life. *Nature (Lond.).* 340:521–524.

Ball, E. E., E. J. Rehm, and C. S. Goodman. 1991. Cloning of a grasshopper cDNA coding for a protein homologous to the A1, A2/B1 proteins of mammalian hnRNP. *Nucleic Acids Res.* 19:397.

Bandziulis, R. J., M. S. Swanson, and G. Dreyfuss. 1989. RNA-binding proteins as developmental regulators. *Genes & Dev.* 3:431–437.

Beyer, A. L., M. E. Christensen, B. W. Walker, and W. M. LeStourgeon. 1977. Identification and characterization of the packaging proteins of core 40S hnRNP particles. *Cell.* 11:127–138.

Burd, C. G., M. S. Swanson, M. Görlach, and G. Dreyfuss. 1989. Primary structures of the heterogeneous nuclear ribonucleoprotein A2, B1, and C2 proteins: A diversity of RNA binding proteins is generated by small peptide inserts. *Proc. Natl. Acad. Sci. USA.* 86:9788–9792.

Burd, C. G., E. L. Matunis, and G. Dreyfuss. 1991. The multiple RNA-binding domains of the mRNA poly(A)-binding protein have different RNA-binding activities. *Mol. Cell Biol.* 7:3419–3424.

Buvoli, M., G. Biamonti, P. Tsoulfas, M. T. Bassi, A. Ghetti, S. Riva, and C. Morandi. 1988. cDNA cloning of human hnRNP protein A1 reveals the existence of multiple mRNA isoforms. *Nucleic Acids Res.* 16:3751–3770.

Buvoli, M. G., F. Cobianchi, M. G. Bestagno, A. Mangiarotti, M. T. Bassi, G. Biamonti, and S. Riva. 1990. Alternative splicing in the human gene for the core protein A1 generates another hnRNP protein. *EMBO (Eur. Mol. Biol. Organ.) J.* 9:1229–1235.

Cavener, D. R. 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* 15:1353–1361.

Choi, Y. D., and G. Dreyfuss. 1984. Monoclonal antibody characterization of the C proteins of heterogeneous nuclear ribonucleoprotein complexes in vertebrate cells. *J. Cell Biol.* 99:1997–2004.

Choi, Y. D., P. J. Grabowski, P. A. Sharp, and G. Dreyfuss. 1986. Heterogeneous nuclear ribonucleoproteins: Role in RNA splicing. *Science (Wash. DC).* 231:1534–1539.

Christensen, M. E., and K. P. Fuxa. 1988. The nucleolar protein B-36 contains a glycine and dimethylarginine-rich sequence conserved in several other nuclear RNA-binding proteins. *Biochem. Biophys. Res. Commun.* 155:1278–1283.

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.

Cobianchi, F., R. L. Karpel, K. R. Williams, V. Notario, and S. H. Wilson. 1988. Mammalian heterogeneous nuclear ribonucleoprotein complex protein A1: Large-scale overproduction in *Escherichia coli* and cooperative binding to single-stranded nucleic acids. *J. Biol. Chem.* 263:1063–1071.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12:387–395.

Dreyfuss, G. 1986. Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. *Annu. Rev. Cell Biol.* 2:459–498.

Dreyfuss, G., Y. D. Choi, and S. A. Adam. 1984. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. *Mol. Cell. Biol.* 4:1104–1114.

Dreyfuss, G., M. S. Swanson, and S. Piñol-Roma. 1988. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem. Sci.* 13:86–91.

Haynes, S. R., M. L. Rebbert, B. A. Mozer, F. Forquignon, and I. B. Dawid. 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.

Haynes, S. R., G. Raychaudhuri, and A. L. Beyer. 1990. The *Drosophila* Hrb98DE locus encodes four protein isoforms homologous to the A1 protein of mammalian heterogeneous nuclear ribonucleoprotein complexes. *Mol. Cell Biol.* 10:316–323.

Haynes, S. R., D. Johnson, G. Raychaudhuri, and A. L. Beyer. 1991. The *Drosophila* Hrb87F gene encodes a new member of the A and B hnRNP proteins group. *Nucleic Acids Res.* 19:25–31.

Hedley, M. L., and T. Maniatis. 1991. Sex-specific splicing and polyadenylation of *dsx* pre-mRNA requires a sequence that binds specifically to *tra-2* protein *in vitro*. *Cell.* 65:579–586.

Hoffmann, 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.

Hoshijima, K., K. Inoue, I. Higuchi, H. Sakamoto, and Y. Shimura. 1991. Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in *Drosophila*. *Science (Wash. DC).* 252:833–836.

Inoue, K., K. Hoshijima, H. Sakamoto, and Y. Shimura. 1990. Binding of the Drosophila Sex-lethal gene product to the alternative splice site of transformer primary transcript. *Nature (Lond.).* 344:461–463.

Kay, B. K., R. K. Sawhney, and S. H. Wilson. 1990. Potential for two isoforms of the A1 ribonucleoprotein in *Xenopus laevis*. *Proc. Natl. Acad. Sci. USA.* 87:1367–1371.

Keller, E. B., and W. A. Noon. 1985. Intron splicing: a conserved internal signal in introns of Drosophila pre-mRNAs. *Nucleic Acids Res.* 13:4971–4981.

Kumar, A., and S. H. Wilson. 1990. Studies of the strand-annealing activity of mammalian hnRNP complex protein A1. *Biochemistry.* 29:10717–10722.

Kumar, A., J. R. Casas-Finet, C. J. Luneau, R. L. Karpel, B. M. Merrill, K. R. Williams, and S. H. Wilson. 1990. Mammalian heterogeneous nuclear ribonucleoprotein A1: Nucleic acid binding properties of the COOH-terminal domain. *J. Biol. Chem.* 265:17094–17100.

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.

Laskey, R. A., and A. D. Mills. 1975. Quantitative film detection of ^{3}H and ^{14}C in polyacrylamide gels by fluorography. *Eur. J. Biochem.* 56:335–341.

Leser, G. P., J. Escara-Wilke, and T. E. Martin. 1984. Monoclonal antibodies to heterogeneous nuclear RNA-protein complexes. *J. Biol. Chem.* 259: 1827–1833.

Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. *In* Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 545 pp.

Mattaj, I. W. 1989. A binding consensus: RNA-protein interactions in splicing, snRNPs, and sex. *Cell.* 57:1–3.

Matunis, M. J., E. M. Matunis, and G. Dreyfuss. 1992*a*. Isolation of hnRNP complexes from *Drosophila melanogaster*. *J. Cell Biol.* 116:245–255.

Matunis, M., W. Michael, and G. Dreyfuss. 1992*b*. Characterization and primary structure of the poly(c)-binding hnRNP K protein. *Mol. Cell Biol.* In press.

Melton, D. A., P. A. Krieg, M. R. Rebagliati, T. Maniatis, K. Zinn, and M. R. Green. 1984. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.* 12:7035–7055.

Moore, C. L., J. Chen, and J. Whoriskey. 1988. Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding. *EMBO (Eur. Mol. Biol. Organ.) J.* 7:3159–3169.

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 (Lond.).* 346:515–520.

Nakagawa, T., M. S. Swanson, B. J. Wold, and G. Dreyfuss. 1986. Molecular cloning of cDNA for the nuclear ribonucleoprotein particle C proteins: a conserved gene family. *Proc. Natl. Acad. Sci. USA.* 83:2007–2011.

Nietfeld, W., H. Mentzel, and T. Pieler. 1990. The *Xenopus laevis* poly(A) binding protein is composed of multiple functionally independent RNA binding domains. *EMBO (Eur. Mol. Biol. Organ.) J.* 9:3699–3705.

O'Farrell, P. Z., H. M. Goodman, and P. H. O'Farrell. 1977. High resolution

two-dimensional electrophoresis of basic as well as acidic proteins. *Cell.* 12:1133–1142.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 85:2444–2448.

Piñol-Roma, S., Y. D. Choi, M. J. Matunis, and G. Dreyfuss. 1988. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes & Dev.* 2:215–227.

Piñol-Roma, S., M. S. Swanson, J. G. Gall, and G. Dreyfuss. 1989. A novel heterogeneous nuclear RNP protein with a unique distribution on nascent transcripts. *J. Cell Biol.* 109:2575–2587.

Preugschat, F., and B. Wold. 1988. Isolation and characterization of a *Xenopus laevis* C protein cDNA: Structure and expression of a heterogeneous nuclear ribonucleoprotein core protein. *Proc. Natl. Acad. Sci. USA.* 85:9669–9673.

Query, C. C., R. C. Bently, 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.

Richter, K., P. J. Good, and I. B. Dawid. 1990. A developmentally regulated, nervous system-specific gene in *Xenopus* encodes a putative RNA-binding Protein. *New Biol.* 6:556–565.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74:5463–5467.

Sierakowska, H., W. Szer, P. J. Furdon, and R. Kole. 1986. Antibodies to hnRNP core proteins inhibit in vitro splicing of human B globin pre-mRNA. *Nucleic Acids Res.* 14:5241–5254.

Snyder, M., S. Elledge, D. Sweeter, R. A. Young, and R. W. Davis. 1987. λgt11: gene isolation with antibody probes and applications. *Methods Enzymol.* 154:107–128.

Steinert, P. M., R. H. Rice, D. R. Roop, B. L. Trus, and A. C. Steven. 1983. Complete amino acid sequence of a mouse epidermal keratin subunit and implications for the structure of intermediate filaments. *Nature (Lond.).* 302: 794–800.

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.

Wilk, H.-E., H. Werr, D. Friedrich, H. H. Kiltz, and K. P. Schäfer. 1985. The core proteins of 35S hnRNP complexes: Characterization of nine different species. *Eur. J. Biochem.* 146:71–81.

Williams, K. R., K. L. Stone, M. B. LoPresiti, 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.

Wilusz, J., D. I. Feig, and T. Shenk. 1988. The C proteins of heterogeneous nuclear ribonucleoprotein complexes interact with RNA sequences downstream of polyadenylation cleavage sites. *Mol. Cell. Biol.* 8:4477–4483.