The pre-mRNA binding K protein contains a novel evolutionarily conserved motif

Haruhiko Siomi, Michael J.Matunis, W.Matthew Michael and Gideon Dreyfuss*
Howard Hughes Medical Institute and the Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6148, USA

Received December 2, 1992; Accepted January 23, 1993

ABSTRACT

The K protein is among the major pre-mRNA-binding proteins (hnRNPs) in vertebrate cell nuclei. It binds tenaciously to cytidine-rich sequences and is the major oligo(rC/dC)-binding protein in vertebrate cells. We have cloned a cDNA of the *Xenopus laevis* hnRNP K and determined its sequence. The *X.laevis* hnRNP K is a 47 kD protein that is remarkably similar to its human 66 kD counterpart except for two large internal deletions. The sequence of hnRNP K contains a 45 amino acid repeated motif which is almost completely conserved between the *X.laevis* and human proteins. We found that this repeated motif, the KH motif (for K homology), shows significant homology to several proteins some of which are known nucleic acids binding proteins. The homology is particularly strong with the archeabacterial ribosomal protein S3 and with the *saccharomyces cerevisiae* protein MER1 which is required for meiosis-specific splicing of the MER 2 transcript. As several of the proteins that contain the KH motif are known to bind RNA, this domain may be involved in RNA binding.

INTRODUCTION

In eukaryotic cells, nascent RNA polymerase II transcripts are associated with at least 20 major proteins in heterogeneous nuclear ribonucleoprotein complexes (hnRNPs) (1, 2, 3). The hnRNP proteins bind pre-mRNAs directly and appear to participate in pre-mRNA processing and possibly also in mRNA transport from the nucleus to the cytoplasm (1, 4, 5, 6). Binding experiments with ribohomopolymers have established that several hnRNP proteins have distinct RNA-binding characteristics (7). Among these, the hnRNP K and J proteins were found to be the major poly(rC)-binding proteins in HeLa cells. We have recently described the cloning, sequence and characterization of the human hnRNP K (8). The human K protein sequence did not reveal previously identified structural motifs known to be involved in RNA binding, such as the ribonucleoprotein consensus sequence (RNP-CS) found in many of the previously characterized hnRNP proteins (5, 9, 10, 11). Thus, hnRNP K is a novel type of RNA-binding protein. However, human hnRNP K contains several interesting sequence motifs: gly-arg-gly-gly (GRGG) sequences which occur frequently in many RNA- and single-stranded DNA-binding proteins (12), and two sets of internal repeats—a short 6 amino acid repeat and a long 45 amino acid repeat. It was not known which, if any, of these motifs represent the RNA-binding domain(s) of human hnRNP K.

One of the monoclonal antibodies to the human K and J proteins reacted with a *X.laevis* protein of molecular weight 47 kD (8), which also has strong poly(rC)-binding activity (see below), suggesting that it is likely to be the *X.laevis* homolog of the human hnRNP K. The cloning and sequencing of a cDNA encoding this 47 kD protein described here confirmed that it is in fact the *X.laevis* hnRNP K. As in human cells, *X.laevis* hnRNP K is an abundant nuclear protein and it is bound to poly(A)$^+$ RNA *in vivo*. The deduced amino acid sequence of the *X.laevis* K protein is very similar to the human K protein, including a near complete conservation of the internal repeats. Although *X.laevis* hnRNP K contains two large deletions, this repeated motif is almost completely conserved, suggesting that it plays an important role in the function of the protein possibly in RNA binding. Here we report that the long internal repeats in hnRNP K have a significant sequence homology (which extends over 39 amino acids) with several eukaryotic, eubacterial and archeabacterial proteins all of which are putative nucleic acids binding proteins.

MATERIALS AND METHODS

Cell culture and labeling

X.laevis kidney epithelial cells were grown at 27°C in 80% Dulbecco's modified Eagle's medium (DMED) supplemented with penicillin, streptomycin and 10% fetal calf serum (FCS). Cells were labeled for 20 hours with 20 μCi per ml [^{35}S]methionine in culture medium containing one-tenth the normal concentration of methionine and 5% FCS.

Cell fractionation, ssDNA chromatography and immunopurification

X.laevis nucleoplasm was prepared essentially as previously described for human HeLa cells (2). Cells were lysed in RSB-100 containing 0.5% Triton X-100 and the protease inhibitors leupeptin, pepstatin A, and aprotinin (Sigma Chemical Co., St.

* To whom correspondence should be addressed

louis, MO). Nuclei were pelleted, resuspended in RSB-100 with protease inhibitors and sonicated. The resultant nuclear lysate was layered over a 30% sucrose cushion and spun at 5,000×g for 10 minutes to remove chromatin and nucleoli. After centrifugation, the layer on top of the sucrose cushion (the nucleoplasm) was removed and used for further fractionation. Nucleoplasm was digested with 100 units per ml micrococcal nuclease (Pharmacia LKB Biotechnology, Piscataway, NJ) for 10 minutes at 30°C in the presence of 1 mM $CaCl_2$. Reactions were stopped by the addition of EGTA to a final concentration of 5mM, on ice. The digested nucleoplasm was then loaded onto a ssDNA-cellulose column (United States Biochemical Corp., Cleveland, Ohio) at 0.1 M NaCl. The column was washed with 0.1 M NaCl in 10 mM Na-phosphate buffer (pH 7.4) and, after equilibration, with 1 mg per ml heparin in the same buffer. Bound proteins were then eluted with 2 M NaCl in 10 mM Na-phosphate buffer (pH 7.4).

Proteins were immunoprecipitated from cell lysate or *in vitro* transcription/translations in phosphate-buffered saline (PBS, pH 7.4) containing 1% Empigen BB (Calbiochem Corp., San Diego, CA) 1 mM EDTA, and 0.1 mM dithiothreitol as previously described (2). Ascites fluid from a BALB/c mouse inoculated with the parent myeloma cell line SP2/0 was used and a non-immune control for precipitations.

Probing protein blots with labeled RNAs

Poly(rC) was 5′ end labeled with T4 polynucleotide kinase. Protein blots were treated for 1 h at room temperature in binding buffer (10 mM Tris−HCl [pH 7.4], 50mM NaCl, 1mM EDTA, 1×Denhardt's solution). The blots were then probed at room temperature for 1 h with labeled RNA (100,000 cpm per lane) in binding buffer containing 20 mg of *Escherichia coli* tRNA (Boehringer Mannheim Biochemicals, Indianapolis, Ind.) per ml. Blots were washed three times for 15 minutes each with binding buffer, air dried, and exposed to X-ray film for autoradiography.

RNA-protein cross-linking in intact cells

Photochemical RNA−protein cross-linking by UV-light irradiation of cells and isolation and analysis of RNPs was carried out essentially as previously described for human HeLa cells (13).

Gel electrophoresis and immunoblotting

Proteins were separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and fluorographed as previously described (13). Two-dimensional nonequilibrium pH gradient polyacrylamide gel electrophoresis (NEPHGE) was performed as described by O'Farrell et al. (14) using an ampholine gradient of pH 3−10 separated for 4 hours at 400 volts in the first dimension. Proteins were separated by SDS-PAGE in the second dimension. Proteins were transferred to nitrocellulose membrane (Schleicher and Schuell, Keene, NH) and the immunoblots were probed with antibodies, as previously described (15).

Immunofluorescence microscopy

X.laevis cells were grown on coverslips and fixed and permeabilized as previously described for human HeLa cells (15). Monoclonal antibody 3C2 ascites fluid was used at a dilution of 1:1,000 and detected with fluorescein isothiocyanate-conjugated goat anti-mouse $F(ab')_2$ (Cappel Laboratories, Malvern, PA).

Isolation of cDNA clones and DNA sequence determination

Monoclonal antibody 3C2 (1:500 dilution of mouse ascites) was used to directly screen a λgt11 *X.laevis* cell cDNA library (a gift from Joe Gall, Carnegie Institute, Baltimore, MD). Positive plaques were purified, and one, pXK 1, was selected for further characterization. The *Eco*RI fragment of pXK1 was sequenced entirely on both strands by a combination of restriction fragment subcloning and oligonucleotide primer synthesis. All sequencing reactions were performed on plasmid DNA with the T7 Sequencing Kit (Pharmacia LKB Biotechnology) according to the manufacturer's instructions.

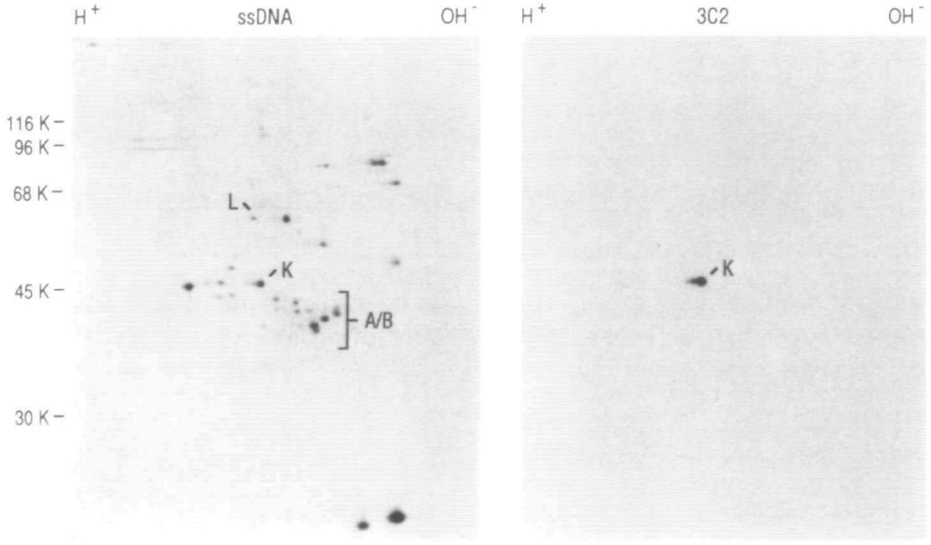


Figure 1. Two-dimensional gel electrophoresis of *X.laevis* proteins purified by ssDNA chromatography and immunoblot analysis with monoclonal antibody 3C2. [^{35}S]methionine-labeled *X.laevis* nucleoplasm was bound to ssDNA-cellulose at 0.1 M NaCl, and heparin-resistant proteins were eluted with 2 M NaCl (panel *ssDNA*). Proteins from the same fraction were also transferred to nitrocellulose membrane and probed with the monoclonal antibody 3C2 (panel *3C2*). Proteins were resolved by two-dimensional gel electrophoresis (NEPHGE in the first dimension, SDS−PAGE in the second dimension). The hnRNP L protein was identified by probing with monoclonal antibody 4D11 (17), the A/B proteins with monoclonal antibody 4C2 (31) (data not shown). Molecular mass standards are on the left.

RESULTS AND DISCUSSION

The *X.laevis* 47 kD protein is an abundant nuclear pre-mRNA-binding protein

Previous studies have shown that the monoclonal antibody 3C2, which was generated against the human hnRNP K and J proteins, crossreacted on immunoblots with a protein of ca. 47 kD in *X.laevis* (8). We have shown previously that the human hnRNP K binds ssDNA and binds tenaciously to poly(rC) (3,8). Two-dimensional gel electrophoresis of ssDNA-binding proteins showed that the 47 kD protein recognized by 3C2 is one of the major single-stranded DNA-binding proteins in *X.laevis* cells (Figure 1). Furthermore, a 47 kD poly(rC)-binding protein was identified among the ssDNA-binding proteins in *X.laevis* by Northwestern blotting (Figure 2, lane Xenopus, 2M), in contrast to the 66 kD hnRNP K protein of human HeLa cells (8), (Figure 2, lane HeLa, 2M). We, therefore, considered this *X.laevis* protein to be the likely candidate for the human hnRNP K homolog, and reasoned that since it is of a much smaller size its sequence would provide insight into the regions of hnRNP K necessary for RNA binding. In addition, this protein can be cross-linked to poly(A)$^+$RNA *in vivo* by UV-light irradiation, demonstrating its direct association with RNA in these cells (Figure 3). Immunofluorescence microscopy, shown in Figure 4, demonstrated that the *X.laevis* protein, like the human K protein, is localized to the nucleus. The antibody shows general nucleoplasmic staining as seen for many of the other hnRNP proteins (15, 16) but, also reveals intense staining of 2−3 unidentified loci in each nucleus, reminiscent of the staining of these cells with antibody to human hnRNP L (17). Based on all these considerations we conclude that the antibody 3C2 reacts specifically with the *X.laevis* hnRNP K protein.

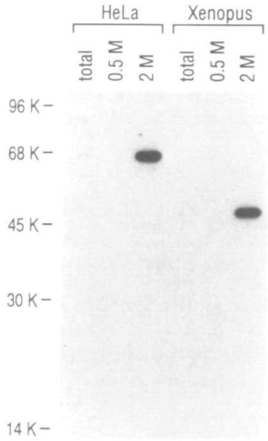


Figure 2. A 47 kD single stranded DNA binding protein of *X.laevis* binds to poly(rC). HeLa and *X.laevis* cells were fractionated by ssDNA-cellulose chromatography. The 0.5 M and 2 M ssDNA fractions were resolved by SDS-PAGE, transferred to nitrocellulose membrane, and probed with ^{32}P-end-labeled poly(rC). Lanes *total*, total cell proteins; lanes *0.5 M*, proteins eluted from ssDNA-cellulose at 0.5 M NaCl; lanes *2 M*, proteins eluted from ssDNA-cellulose at 2 M NaCl.

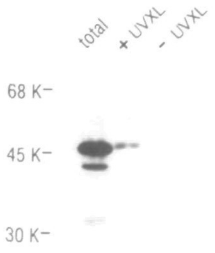


Figure 3. Immunoblot analysis of hnRNP K crosslinked *in vivo* to poly(A)$^+$ RNA. *X.laevis* cells were irradiated with UV-light and poly(A)$^+$ RNA was isolated and digested with RNases. Released proteins were resolved by SDS-PAGE, transferred to nitrocellulose membrane, and probed with the mAb 3C2. Lane *total*, total cell proteins; lane *+UVXL*, proteins crosslinked to poly(A)$^+$ RNA; lane *-UVXL*, proteins from samples treated as in lane *+UVXL* except that cells were not UV-irradiated with UV-light.

Isolation of a cDNA clone for *X.laevis* hnRNP K

To isolate candidate cDNAs for *X.laevis* hnRNP K, the monoclonal antibody 3C2 was used for immunological screening of a *X.laevis* λgt11 library. Several positive clones were isolated and characterized by restriction mapping and by *in vitro* transcription and translation. One of these clones, pXK1 (1.5kb), when translated *in vitro*, produced a polypeptide of 47 kD. This polypeptide comigrated with authentic *X.laevis* hnRNP K, and it was immunoprecipitated with 3C2 but not with control antibody SP2/0 (Figure 5). Based on these criteria, as well as on the binding of this protein to poly(rC) and its migration on two dimensional gels (data not shown), we conclude that it is the cDNA for the *X.laevis* hnRNP K protein. pXK1 was sequenced completely on both strands. The cDNA clone contains an open reading frame encoding a protein of 397 amino acids with a predicted molecular weight of 43,726 (compared with the human protein of 463 amino acids) and a predicted pI of 7.28 (compared

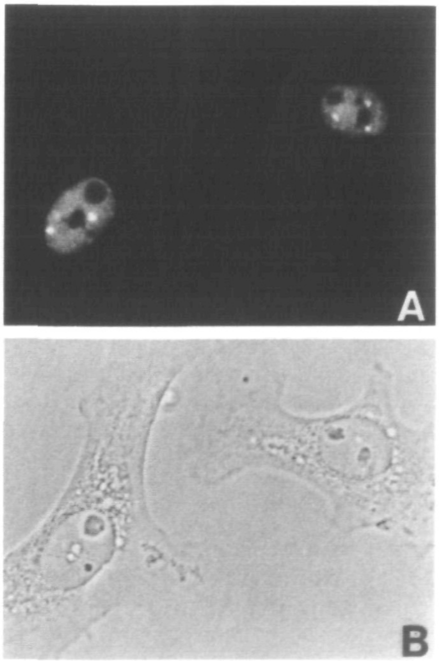


Figure 4. Immunofluorescence microscopy with monoclonal antibody 3C2. (A) *X.laevis* kidney epithelial cells were stained with the anti-hnRNP K monoclonal antibody 3C2. (B) Corresponding phase image.

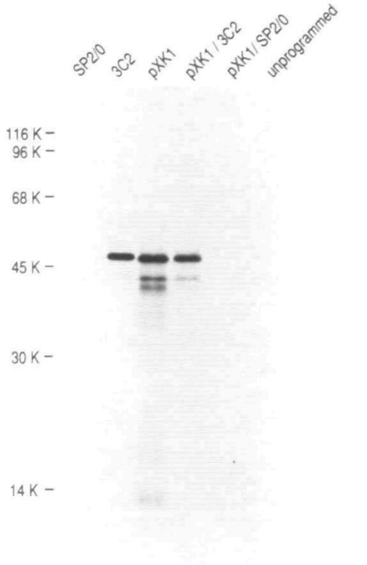
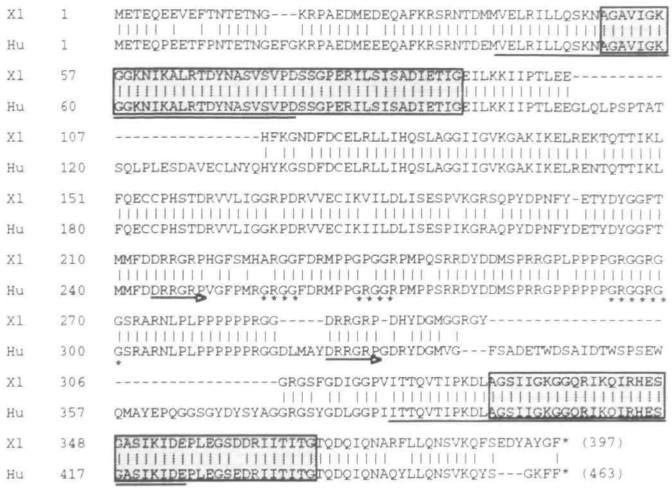


Figure 5. Immunoprecipitation of the pXK1 *in vitro* transcription–translation product and comparison to authentic *X.laevis* hnRNP K. pXK1 was linearized with *HindIII* and transcribed *in vitro* with SP6 polymerase. The RNA was translated in reticulocyte lysate in the presence of [^{35}S]methionine, and the translation product was immunoprecipitated with the monoclonal antibody 3C2. hnRNP K was also immunopurified from [^{35}S]methionine labeled *X.laevis* cell lysate with 3C2. Proteins were analyzed by SDS-PAGE, molecular mass standards are on the left. Lanes: SP2/0, control immunoprecipitation from [^{35}S]methionine labeled *X.laevis* cells with SP2/0 ascites fluid; 3C2, immunoprecipitation from [^{35}S]methionine labeled *X.laevis* cells with 3C2; pXK1, pXK1 *in vitro* transcription–translation product; pXK1/3C2, pXK1 *in vitro* transcription–translation product immunoprecipitated with 3C2; pXK1/SP2/0, control immunoprecipitation of pXK1 *in vitro* transcription-translation product with SP2/0; unprogrammed, product of *in vitro* transcription-translation reaction lacking RNA template.

with 5.17 for the human proteins). Database searches showed that the predicted amino acid sequence (Figure 6) does not have extensive similarity to any known protein on file in the SWISS-PROT (University of Geneva, Switzerland) or Protein Identification Resource (National Biomedical Research Foundation, Washington, D.C., database) other than to the human hnRNP K (8). The overall amino-acid sequence homology of the open reading frame to human hnRNP K is 77.5%, not including conservative changes. These data, taken together with immunological definition of the *X.laevis* clone, indicate that pXK1 encodes the *X.laevis* hnRNP K.

Comparison of the *X.laevis* and human hnRNP K reveals that there are two large deletions in the central region of the *X.laevis* sequence. A 90.4% sequence identity is seen between the *X.laevis* and human hnRNP K proteins in the regions outside of the two deletions. Notably, two of four GRGG sequences in human hnRNP K are not conserved in the *X.laevis* sequence. It seems, therefore, unlikely that the GRGG sequences are responsible for the nucleic acid binding activity of hnRNP K although they may contribute to it. However, the proteins share a number of sequence landmarks, including a near complete conservation of the direct repeats, namely the asp-arg-arg-gly-arg-pro (DRRGRP) sequence and the long 45 amino acid direct repeats (Figure 6), implying that they are functionally important.



```
Xl    1   METEQEEVEFTNTETNG---KRPAEDMEDEQAFKRSRNTDMMVELRILLQSKNAGAVIGK
          ||||| |  | ||||||   |||||||| ||||||||||| |||||||||||||||||||
Hu    1   METEQPEETFPNTETNGEFGKRPAEDMEEEQAFKRSRNTDEMVELRILLQSKNAGAVIGK

Xl   57   GGKNIKALRTDYNASVSVPDSSGPERILSISADIETIGEILKKIIPTLEE----------
          ||||||||||||||||||||||||||||||||||||||||||||||||||
Hu   60   GGKNIKALRTDYNASVSVPDSSGPERILSISADIETIGEILKKIIPTLEEGLQLPSPTAT

Xl  107   ----------------HFKGNDFDCELRLLIHQSLAGGIIGVKGAKIKELREKTQTTIKL
                          | || ||||||||||||||||||||||||||||||| |||||||
Hu  120   SQLPLESDAVECLNYQHYKGSDFDCELRLLIHQSLAGGIIGVKGAKIKELRENTQTTIKL

Xl  151   FQECCPHSTDRVVLIGGRPDRVVECIKVILDLISESPVKGRSQPYDPNFY-ETYDYGGFT
          ||||||||||||||||| ||||||||| ||||||||| ||| |||||||| |||||||||
Hu  180   FQECCPHSTDRVVLIGGKPDRVVECIKIILDLISESPIKGRAQPYDPNFYDETYDYGGFT

Xl  210   MMFDDRRGRPHGFSMHARGGFDRMPPGPGGRPMPQSRRDYDDMSPRRGPLPPPPGRGGRG
          |||||||||| || |  |||||||||| |||||| |||||||||||||| ||||||||||
Hu  240   MMFDDRRGRPVGFPMRGRGGFDRMPPGRGGRPMPPSRRDYDDMSPRRGPPPPPPGRGGRG
                          ****      ****                        ******

Xl  270   GSRARNLPLPPPPPPRGG-----DRRGRP-DHYDGMGGRGY-------------------
          ||||||||||||||||||     |||||| | |||| |
Hu  300   GSRARNLPLPPPPPPRGGDLMAYDRRGRPGDRYDGMVG---FSADETWDSAIDTWSPSEW
          *

Xl  306   ------------------GRGSFGDIGGPVITTQVTIPKDLAGSIIGKGGQRIKQIRHES
                            |||| || ||| ||||||||||||||||||||||||||||||
Hu  357   QMAYEPQGGSGYDYSYAGGRGSYGDLGGPIITTQVTIPKDLAGSIIGKGGQRIKQIRHES

Xl  348   GASIKIDEPLEGSDDRIITITGTQDQIQNARFLLQNSVKQFSEDYAYGF* (397)
          ||||||||||||| ||||||||||||||||  |||||||| |      |
Hu  417   GASIKIDEPLEGSEDRIITITGTQDQIQNAQYLLQNSVKQYS---GKFF* (463)
```

Figure 6. Comparison of the predicted amino acid sequences of *X.laevis* and human hnRNP K. The sequence of *X.laevis* is on the top; that of human is on the bottom. Amino acid identity is indicated by a vertical line. Underlined portions of the sequence represent the KH motif (see Text). The long internal repeats are indicated in boxes. The arrows show the short repeats. Asterisks indicate the GRGG sequences.

Table 1. Quantitative analysis of hnRNP K repeats and related sequences

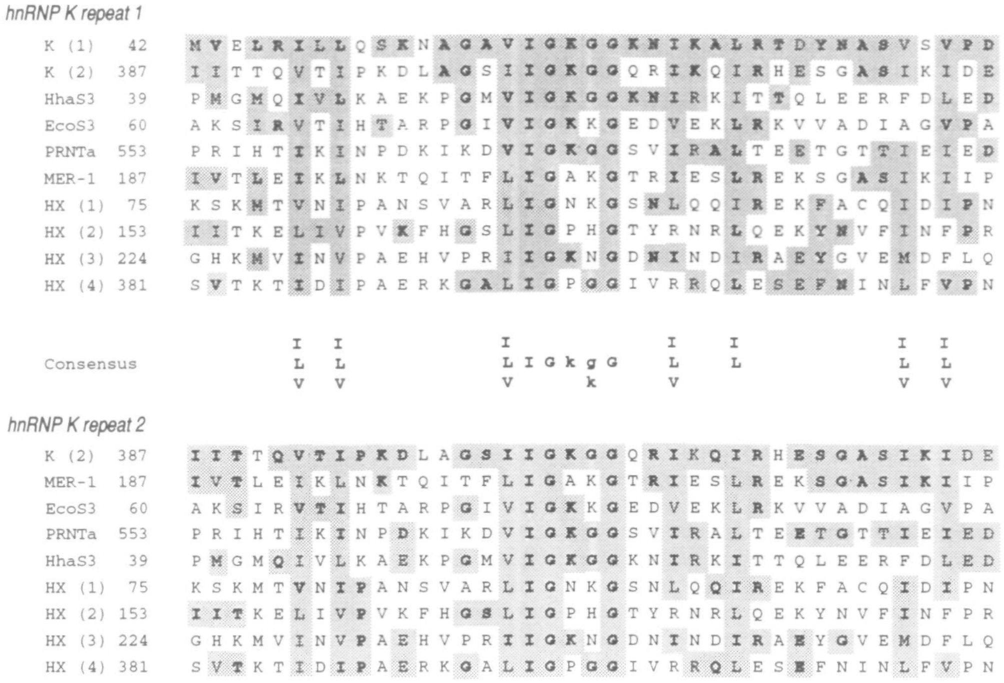
A. Per cent Identity (and Similarity) Between Different Sequences

	hnRNP K1	hnRNP K2	Hha S3	MER-1	PRNT
hnRNP K1	-	31(56)	36(54)	31(41)	28(44)
hnRNP K2		-	23(46)	41(54)	31(49)
Hha S3			-	15(27)	36(49)
MER-1				-	31(46)
PRNT					-

B. Statistical Significance of Similarity Scores (SD unit)

	hnRNP K1	hnRNP K2	Hha S3	MER-1	PRNT
hnRNP K1	-	11.96	12.54	7.19	7.52
hnRNP K2		-	7.06	9.25	8.06
Hha S3			-	2.09	6.78
MER-1				-	8.29
PRNT					-

Quantification of the homologies in the KH domain of hnRNP K, Hha S3, MER-1 and PRNT. In (A), Percent identity is defined as the ratio of the number of positions containing identical amino acids between any pairwise analysis of sequences aligned as shown in Figure 7 to the total number of amino acids. Percent similarity (parenthesis) was calculated likewise, except that chemically similar amino acids were counted in addition to identities. Conservative amino acid changes are defined as described in the legend of Figure 7. In (B), Statistical analyses between any pairwise sequences were carried out using the RDF2 program for scoring (26). Briefly, This program compares two sequences, calculating initial and optimized scores, and then shuffles the second sequence a specified number of times (in this case, 100 times), again calculating initial and optimized scores. The z value is calculated by subtracting the mean score of the randomly shuffled sequences from the score of the unshuffled sequence and then dividing by the standard deviation of the distribution of shuffled scores. It is suggested that one should be skeptical of conclusions based on sequence similarity scores with z values less than 3, and more confident when the z values are greater than 6.



```
hnRNP K repeat 1

K (1)       42   M V E L R I L L Q S K N A G A V I G K G G K N I K A L R T D Y N A S V S V P D
K (2)      387   I I T T Q V T I P K D L A G S I I G K G G Q R I K Q I R H E S G A S I K I D E
HhaS3       39   P M G M Q I V L K A E K P G M V I G K G G K N I R K I T T Q L E E R F D L E D
EcoS3       60   A K S I R V T I H T A R P G I V I G K K G E D V E K L R K V V A D I A G V P A
PRNTa      553   P R I H T I K I N P D K I K D V I G K G G S V I R A L T E E T G T T I E I E D
MER-1      187   I V T L E I K L N K T Q I T F L I G A K G T R I E S L R E K S G A S I K I I P
HX (1)      75   K S K M T V N I P A N S V A R L I G N K G S N L Q Q I R E K F A C Q I D I P N
HX (2)     153   I I T K E L I V P V K F H G S L I G P H G T Y R N R L Q E K Y N V F I N F P R
HX (3)     224   G H K M V I N V P A E H V P R I I G K N G D N I N D I R A E Y G V E M D F L Q
HX (4)     381   S V T K T I D I P A E R K G A L I G P G G I V R R Q L E S E F M I N L F V P N

                           I   I               I               I     I               I   I
Consensus                  L   L               L I G k g G     L     L               L   L
                           V   V               V       k       V                     V   V

hnRNP K repeat 2

K (2)      387   I I T T Q V T I P K D L A G S I I G K G G Q R I K Q I R H E S G A S I K I D E
MER-1      187   I V T L E I K L N K T Q I T F L I G A K G T R I E S L R E K S G A S I K I I P
EcoS3       60   A K S I R V T I H T A R P G I V I G K K G E D V E K L R K V V A D I A G V P A
PRNTa      553   P R I H T I K I N P D K I K D V I G K G G S V I R A L T E E T G T T I E I E D
HhaS3       39   P M G M Q I V L K A E K P G M V I G K G G K N I R K I T T Q L E E R F D L E D
HX (1)      75   K S K M T V N I P A N S V A R L I G N K G S N L Q Q I R E K F A C Q I D I P N
HX (2)     153   I I T K E L I V P V K F H G S L I G P H G T Y R N R L Q E K Y N V F I N F P R
HX (3)     224   G H K M V I N V P A E H V P R I I G K N G D N I N D I R A E Y G V E M D F L Q
HX (4)     381   S V T K T I D I P A E R K G A L I G P G G I V R R Q L E S E F M I N L F V P N
```

Figure 7. Sequence alignment of hnRNP K with putative RNA-binding proteins. Amino-acid sequences of the repeated motif of hnRNP K deduced from the nucleotide sequences are aligned with the sequences of HhaS3 (18), EcoS3 (21), PRNT α subunit (22), MER1 (19), and a yeast protein HX (23). Sequences are indicated using the single-letter code. The position in the primary sequence of the first amino acid on every line is shown. In the upper panel, amino acids identical to those in hnRNP K repeat 1 are in bold and conserved amino acids are shaded. Conservative amino acid changes are defined by the following groups: (I, L, M, V), (K, R), (D, E), (F, Y, W), and (S, T). In the lower panel, amino acids identical or similar to those in hnRNP K repeat 2 are indicated in the same manner. The consensus (capital letters) was defined as residues present in at least seven of the individual sequences.

Sequence homology between the repeats in hnRNP K and putative nucleic acids binding proteins

Since the most conspicuous structural feature of hnRNP K is a 45 amino acid repeated motif, we searched the GenBank data base to determine whether either repeat might be related to previously described proteins. Significant amino acid sequence similarity was observed between parts of this repeated motif and several putative RNA-binding proteins (Figure 7). We refer to these homology domains within the hnRNP K protein as KH (for *K* protein *H*omology) motifs. The KH domains overlap the long direct repeats of hnRNP K (Figure 6). The sequence similarities were most extensive for repeat 1 in hnRNP K with the *Halobacterium halobium* ribosomal protein S3 (Hha S3) (18), and for repeat 2 in hnRNP K with the yeast *Saccharomyces cerevisiae* MER1 protein (19), a meiosis-specific splicing regulator of MER2 which is essential for meiosis (20) (Figure 7 and Table 1A). HnRNP K repeat 1 (KH 1) and repeat 2 (KH 2) have 31% identity and 56% similarity over 39 residues; the *X.laevis* and human proteins being identical throughout this region. The hnRNP KH 1 and the Hha S3 KH motif have 36% identity and 54% similarity, and hnRNP KH 2 and MER1 have 41% identity and 54% similarity throughout the same regions. KH sequences were identified in the *Escherichia coli* ribosomal protein S3 (EcoS3) (21), polyribonucleotide:orthophosphate nucleotidyltransferase (PRNT) (22), and a yeast (*Saccharomyces cerevisiae)* protein (HX) of unknown function (23). A consensus sequence of the most common amino acid found was derived from all 10 sequences (Fig. 7). The greatest frequency of identities in the sequences are clustered in the central region; I-G-k-g/k-G sequence, which is remeniscent of, but clearly distinct from, the core sequences of the phosphate-binding sites for mono- and dinucleotides (24, 25). Another notable feature is that these sequences have regularly spaced hydrophobic amino acid residues (Leu, Ile, Val). The similar pattern of hydrophobic residues among these sequences may reflect the existence of a similar secondary structure.

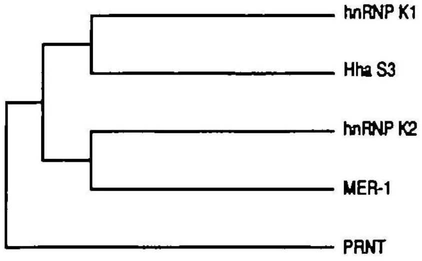


Figure 8. Evolutionary tree based on sequence divergence among the KH motifs. The relatedness tree was constructed using the PILEUP program (27); this program plots a dendrogram that shows the clustering relationships used to determine the order of the pairwise alignments that together create the final multiple sequence alignment. Distance along the vertical axis is proportional to the difference between sequences; distance along the horizontal axis has no significance.

Sequence relationships among these sequences were more carefully examined using the RDF2 program (26) which can test whether high similarity scores are likely to reflect sequence similarity that is due to common ancestry or simply a locally biased amino acid composition, and the results are summarized in Table 1B. The z-value (SD) scores for all pairwise comparisons are highly significant. Interestingly, this statistical analysis also shows that hnRNP K KH 1 is more closely related to the Hha S3 KH motif than to hnRNP K KH 2. Consistent with these findings, comparison of a possible relatedness tree constructed using the PILEUP program (27) shows that the repeats in hnRNP K exhibit considerable divergence, more than those observed within KH 1 of hnRNP K and Hha S3, and KH 2 of hnRNP K and MER-1 (Figure 8). An important evolutionary consequence of the relaxed selection pressures for a precise amino acid sequence among KH motifs is the opportunity for functional differentiation among duplicated KH motifs. Thus, these sequence data suggest that repeats 1 (KH 1) and 2 (KH 2) in hnRNP K may not be functionally equivalent.

A common feature to the proteins with KH motifs is physical or functional association, or both, with RNA molecules. The *E.coli* ribosomal S3 protein is positioned in the 3′ region of the 16S rRNA in the 30S ribosomal subunit (28) and it can be crosslinked to mRNA (29). *In vitro*, PRNT, as a trimeric molecule, catalyzes the polymerization of ribonucleoside diphosphates, and the phosphorolysis of polyribonucleotides in the presence of phosphate. In the yeast *Sacchromyces cerevisiae*, the MER 1 gene product regulates the splicing of the MER 2 transcript during meiosis (30). The existance of the KH domain in several proteins which associate with RNA suggests that this domain may have an important role for RNA-binding. The *S.cerevisiae* HX protein was described as a histone-like protein (23) and it is therefore quite likely that it also is a nucleic acid binding protein. Thus, although the function of the KH motif is not known, at least two of the proteins in which it is found (hnRNP K and ribosomal protein S3) are known to bind RNA and several of the others are likely to have such activity. This invites the speculation that KH motifs may be important for RNA binding. The conservation of this motif in eubacterial and archeabacterial proteins indicates that it arose early in evolution. Future experiments will be required to determine the role of KH domains in RNA binding or in the other functions of these proteins.

ACKNOWLEDGEMENTS

We thank Dr Joseph Gall for the *X.laevis* cDNA library and Eric Sulman for help with isolation of the hnRNP K cDNA clone. We are grateful members of our laboratory for helpful discussions and comments on the manuscript, especially to Matthias Görlach and Douglas Portman for their help with the analysis of the sequences. This work was supported by the Howard Hughes Medical Institute and by a grant from the National Institutes of Health.

REFERENCES

1. Dreyfuss, G. (1986) *Annu. Rev. Cell Biol.* **2**, 459–498.
2. Choi, Y. D. & Dreyfuss, G. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7471-7475.
3. Piñol-Roma, S., Choi, Y. D., Matunis, M. J. & Dreyfuss, G. (1988) *Genes Dev.* **2**, 215–227.
4. Chung, S. Y. & Wooley, J. (1986) *Proteins* 1, 195–210.
5. Dreyfuss, G., Swanson, M.S. & Piñol-Roma, S. (1988) *Trends Biochem.* **13**, 86–91.
6. Piñol-Roma, S. & Dreyfuss, G. (1992) *Nature* **355**, 730–732.
7. Swanson, M. S. & Dreyfuss, G. (1988) *Mol. Cell. Biol.* **8**, 2237–2241.
8. Matunis, M. J., Michael, W. M. & Dreyfuss, G. (1992) *Mol. Cell. Biol.* **12**, 167- 171.
9. Bandziulis, R. J., Swanson, M. S. & Dreyfuss, G. (1989) *Genes Dev.* **3**, 431- 437.
10. Frankel, A. D., Mattaj, I. W. & Rio, D. C. (1991) *Cell* 67, 1041–1046.
11. Kenan, D. J., Query, C. C. & Keene, J. D. (1991) *TIBS* **16**, 214–220.
12. Christensen, M. E. & Fuxa, K. P. (1988) *Biochem. Biophys. Res. Commun.* **155**, 1278–1283.
13. Dreyfuss, G., Adam, S. A. & Choi, Y. D. (1984) *Mol. Cell. Biol.* **4**, 415-432
14. O'Farrell, P. Z., Goodman, H. M. & O'Farrell, P. H. (1977) *Cell* 12, 1133-1142.
15. Choi, Y. D. & Dreyfuss, G. (1984) *J. Cell. Biol.* **99**, 1997–2004.
16. Dreyfuss, G., Choi, Y. D. & Adam, S. A. (1984) *Mol. Cell. Biol.* **4**, 1104–1114.
17. Piñol-Roma, S., Swanson, M. S., Gall, J. G. & Dreyfuss, G. (1989) *J. Cell Biol.* 109, 2575–2587.
18. Spiridonova, V. A., Akhomanova, S. A., Kagramanova, V. K., Köpke, A. K. E. & Mankin, A. S. (1989) *Can. J. Microbiol.* **35**, 153–159.
19. Engebrecht, J. & Roeder, G. S. (1990) *Mol. Cell. Biol.* **10**, 2379–2389.
20. Engebrecht, J., Hirsch, J. & Roeder, G. S. (1990) *Cell* **62**, 927–937.
21. Brauer, D. & Röming, R. (1979) *FEBS Lett.* **106**, 352–357.
22. Régnier, P., Grunberg-Manago, M. & Portier, C. (1987) *J. Biol. Chem.* **262**, 63–68.
23. Delahodde, A., Becam, A. M., Perea, J. & Jacq, C. (1986) *Nucleic Acids Res.* **14**, 9213–9214.
24. Möller, W. & Amons, R. (1985) *FEBS Lett.* **186**, 1–7.
25. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990) *Trends Biochem.* **15**, 430- 434.
26. Pearson, W.R. (1990) *Methods Enzymol.* **183**, 63–98.
27. Feng, D. F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25**, 351–360.
28. Noller, H. F., Moazed, D., Stern, S., Powers, T., Allen, P. N., Robertson, J. M., Weiser, B. & Triman, K. (1990) in *The Ribosome*, ed. Hill, W. E., Dahlberg, A., Garrett, R. A., Moore, P. B., Schlessinger, D. & Warner, J. R. (American Society for Microbiology, Washington, D.C.), pp73–92.
29. Rinke-Appel, J., Jünke, N., Stade, K. & Brimacombe, R. (1991) *EMBO J.* **10**, 2195- 2202.
30. Engebrecht, J., Voelkel-Meiman, K. & Roeder G. S. (1991) *Cell* **66**, 1257–1268.
31. Matunis, M.J., Matunis, E.L. & Dreyfuss, G. *J. Cell. Biol.* **116**, 245–255.