Volume 324, number 3, 361–366 FEBS 12587 June 1993
© 1993 Federation of European Biochemical Societies 00145793/93/$6.00

The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid

Toby J. Gibson, Julie D. Thompson and Jaap Heringa

European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, W-6900 Heidelberg, Germany

Received 21 April 1993

New findings are presented for the ~ 50 residue KH motif, a domain recently discovered in RNA-binding proteins The conserved sequence is ~ 10 residues larger than previously reported. Profile searches have revealed new members of this family, including two, *E. coli* NusA and human GAP-associated p62 phosphoprotein, for which RNA-binding data exists. A *nusA* homolog was detected in the RNA polymerase gene complex of six archaebacterial species and may encode an antiterminator. All KH-containing proteins are linked with RNA and the KH motif most probably functions as a nucleic acid binding domain.

hnRNP; Vigilin; Ribosomal protein; RNA polymerase; 3′-5′ Exonuclease, Profile search

1. INTRODUCTION

The hnRNP K protein is one of many associated with heterogeneous nuclear RNA [1]. As with most hnRNP proteins, the function of hnRNP K is not understood. The cDNA sequence revealed a ~ 40 amino acid repeat within the translated sequence that was found to be homologous to four proteins of surprisingly diverse occurrence [1]: *Escherichia coli* polynucleotide phosphorylase, a 3′–5′ exonuclease involved in RNA turnover [2]; yeast MER1, involved in regulation of meiosis-specific RNA splicing [3]; yeast HX, of unknown function [4]; and ribosomal protein S3, found in all taxonomic kingdoms and shown in *E. coli* ribosomes to be in close proximity to mRNA downstream of the decoding site [5]. The obvious property these proteins share is a close association with nucleic acid, leading to the suggestion that this domain is likely to be RNA-binding [1]. The domain was designated KH (for K-homology). A second recently determined sequence, independently reported to be homologous with the HX protein, is that of vigilin. Vigilin cDNA was isolated by comparative screening with mRNA from differentiated and dedifferentiated chicken chondrocytes [6]. Vigilin is of unknown function and consists almost exclusively of fourteen repeats of the KH motif. In this article we report the identification of new members of the KH family, present a multiple alignment of the known family members, indicating significant features, and discuss implications for the role of the KH domain.

2. MATERIALS AND METHODS

2.1 *Computation*

Profiles spanning the 50-residue KH domain were prepared by the program PROFILEWEIGHT [7] using the BLOSUM62 substitution matrix [8], sequence weighting and excision of gapped positions with more than 80% padding characters. PROFILESEARCH [9] provided in the GCG package [10] was used for searching SWISS-PROT [11] and PIR [12] protein databases with the profiles. Default normalisations for amino acid composition and sequence length were turned off. As with all sequence searches, in profile searches there is no clear demarcation of positive scores from noise The scores are dependent on factors such as the length of the aligned sequences, the residue composition, gap penalties, and choice of substitution matrix. In each run, scores for candidate hits were compared against the standards of previously detected KH domains. Scores given in the results may be compared to a maximum score of 9.71 for vigilin down to 6.00 for *E. coli* S3 Discrimination of true versus false ended below about 4.80 after which the score curve flattened and was obviously noisy.

MPsrch (J. Collins and S. Sturrock, Edinburgh), running on a MASPAR parallel computer and FASTA [13] were used for searching with single sequences. Given the short probe sequence, MPsrch scores < 10^{-1} were followed up and < 10^{-2} considered highly significant. For FASTA, optimised scores > 90 were followed up and > 100 were considered highly significant. TFASTA [13] was used for searching a six-phase translation of the EMBL [14] and GENBANK [15] DNA sequence databases. To check for repeats, profiles were slid along a single sequence according to [16]. The GDE colour alignment editor (S. Smith, Harvard University) and the program COLORMASK (J Thompson, unpublished) were used to prepare the coloured alignment in Fig. 1.

Correspondence address: T.J. Gibson, European Molecular Biology Laboratory Postfach 102209, Meyerhofstrasse 1, W-6900 Heidelberg, Germany. Fax: (49) (6221) 387517.

Abbreviations: ORF, open reading frame; indel, site of insertion or deletion in aligned sequences; hnRNA, heterogeneous nuclear RNA; hnRNP, protein associated with hnRNA; KH, domain with hnRNP K similarity; 5-APAS-UTP, 5-((4-azido-phenacyl)thio)-uridine-5′ triphosphate.

Published by Elsevier Science Publishers B.V.

2.2. *Database entries*

Sequence entries extracted from SWISS-PROT are: HX_YEAST, *S cerevisiae* HX; PNP_ECOLI, *E. coli* polynucleotide phosphorylase; MER1_YEAST, *S cerevisiae* MER1, GRP3_ARTSA, *Artemia salina* grp33; Y14K_HALMO, *Halococcus morrhuae* ORF139; Y14K_HALHA, *Halobacterium halobium* ORF139; YRP7_METVA, *Methanococcus vannielii* ORF2; YRP3_SULAC, *Sulfolobus acidocaldarius* ORF130, YRPL_THECE, *Thermococcus celer* ORF-X; RS3_ECOLI, *E coli* S3, RS3_HALHA, *Halobacterium halobium* S3, RS3_HUMAN, human S3.

A sequence entry extracted from the PIR database is: S23464, human vigilin.

Sequence entries extracted from the EMBL database are: BSORF1T7A, *Bacillus subtilis* NusA; HSP62, human GAP-associated P62; TARPOG, *Thermoplasma acidophilum* ORF-X.

A sequence entry extracted from GENBANK is: S74678, human hnRNP K.

3. RESULTS

3.1. *Detection of KH repeats*

Database searches with profiles constructed from multiple alignments are more sensitive than searching with single sequences but alignment accuracy is critical for maximum sensitivity [9]. Before preparing profiles, the hnRNP K, HX and vigilin sequences were reexamined. Inspection of sequences adjacent to the KH repeats revealed that the domain is some ten residues larger than originally reported [1], with a C-terminal conserved element that is separated from the core of the motif by a segment which is unconserved and highly variable in length. Full-length alignments were used for subsequent searches. Profiles were constructed from the aligned repeats of human hnRNP K, HX and vigilin and first used to check the number of repeats in these proteins by sliding the sequence past the profile [16]. Fig. 1A,B shows that the number of repeats detectable by this method in hnRNP K and HX differs from that reported [1,6], showing 3 and 6 peaks, respectively. The additional repeats were straightforward to align and added to the profile dataset.

The KH profile was used to search the SWISS-PROT and PIR protein databases. Scores comparable to the known homologs (see section 2 for the score range) were obtained for three types of protein: 6.33 for *Artemia salina* hnRNP protein grp33 [17]; a range of 5.59–7.49 for five ORFs (in part known to be homologs), that

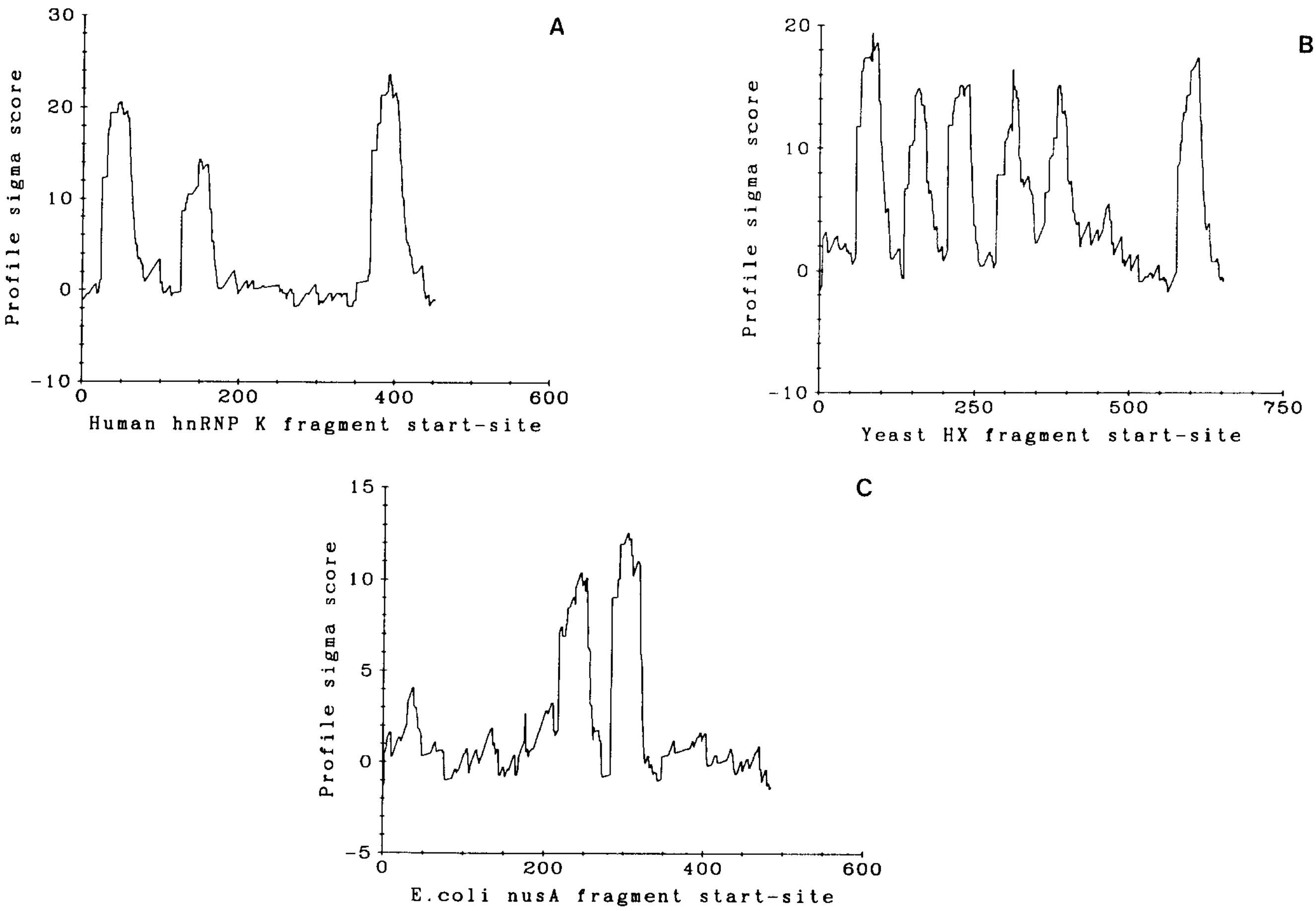

Fig. 1. Plots of scores from the KH profile sliding past a sequence, sampling a window of 50 residue fragments. The score is plotted at the first residue in the window. Scores have been converted to standard deviations relative to scores generated with randomised sequence. (A) hnRNP K against the KH profile. There are three significant peaks (B) HX against the KH profile. There are six significant peaks. (C) NusA against the KH profile. There are two significant peaks

were always 3′ of genes encoding RNA polymerase subunits, and that were isolated from a divergent set of archaebacterial species [18–21]; and 5.77 for the *E. coli* NusA protein which binds within the RNA polymerase complex where it is involved in the mechanism of transcriptional termination [22]. Further examination of the sequences by sliding the profiles against the sequences [16] showed one copy of the KH domain in grp33 and two copies in the other proteins, although no repeated sequences had previously been reported in these proteins. Figure 1C shows the two statistically significant matches to the KH profile in NusA. A profile prepared from aligning the archaebacterial ORFs scored 8.06 on the NusA sequence, suggesting a close similarity, and inspection reveals that they align throughout to a central portion of the larger NusA sequence: they are therefore direct NusA homologs.

The profile searches were complemented by sequence searches with single domains of the protein databases using MPsrch and FASTA. These confirmed the NusA subfamily (FASTA scores > 100; > 30% sequence identity) and that the additions were new members of the KH domain. Finally, TFASTA was used to search the DNA databases for gene sequences not yet in the protein databases. Matches were found for several more proteins: a sixth archaebacterial RNA polymerase operon ORF [23]; *Bacillus subtilis* NusA (K. Shazand, unpublished); and human p62 GAP-associated phosphoprotein [24]. The latter was known to be a homolog of *Artemia* grp33 with a reported similarity extending beyond the KH domain.

3.2. *KH Domain alignment*

An alignment of known KH domains is shown in Fig. 2. Only a representative set (one from each kingdom) of the > 20 ribosomal S3 sequences is included [25–27]. The alignment was made by a combination of automatic matching to the profiles and visual inspection of sequences and of profile-to-sequence dotplots [7]. The aligned sequences have four matching blocks interrupted by indels. These are expected to be loops in the protein structure. In some of the most diverged plastid S3 sequences (not shown), the second block also splits at the conserved GxxG motif at positions 19–22 with loss of the G signature. Hence this is also likely to be a loop region. The five sequence segments that are never split show clear amphipathicity, as expected if they represent elements which pack in a folded globular structure and are partly buried and partly solvent accessible. Segments 1, 4 and 5 have an alternating pattern of hydrophobic and hydrophilic residues. These segments are compatible with β-strands, with packing interactions to one side of the β-sheet, but are inexplicable by α-helical periodicity. By contrast, segments 2 and 3 have conserved hydrophobicity 3/4 residues apart. These fit the behaviour of a-helices with packing interactions to one side. The observed prolines in segment 2 (positions 12–14) would be in the first turn of the helix and are allowed. The glycine preference at position 14 is unusual for a helical position and weakens the inference from amphipathic periodicity. These sequence characteristics may indicate that the KH domain has secondary structure β-α-α-β-β. Given these elements, the folded structure would be a 3-stranded sheet, solvent exposed to one side and on the other packing against the two helices which could themselves range between antiparallel or orthogonal orientations but could not be parallel.

There are no absolutely conserved residues in the aligned sequences, which probably excludes a catalytic role for the KH domain. There is a preference for positively charged residues at positions 12, 13, 15, 19, 20, 25, 26, 28 on and between the proposed helices. These may be an indication of a surface zone that can bind with a negatively charged partner, such as nucleic acid.

4. DISCUSSION

4.1. *KH-Containing protein subfamilies*

Although the KH domains must share a common ancestor, the genetic shuffling processes which characterise multidomain proteins mean that the proteins are not necessarily homologous in their entirety. The known KH proteins fall into subfamilies, some of which have multiple members. In some cases the groupings may allow reasonable guesses to be made about functions of the less well characterised members. These are discussed below.

4.1.1. The NusA family. The *E. coli* NusA protein has been the subject of considerable study, recently reviewed in [28]. NusA modulates the ability of transcription to stop at RNA hairpins which cause polymerase to first pause and then, sometimes, terminate. It forms part of a complex of proteins associated with RNA polymerase during transcriptional extension. It interacts directly with RNA polymerase and the rho terminator protein. N protein from bacteriophage λ, which promotes antitermination, also acts by binding to NusA. However, both the direct mechanism of action of NusA and its mode of interaction with RNA remain obscure. Filter-binding studies imply direct NusA/RNA interaction [29] but no specific recognition sequence has been delineated. Although the presence of NusA in a transcription complex can stabilise RNA at a pause site against nuclease probes, this could be an indirect effect [30]. NusA could not be photo-crosslinked to 5-APAS-U-containing RNA hairpins, indicating that it may not sample the bases at pause sites [31]. The multiple domain structure in NusA, which is revealed here, suggests that functional sites are localised to discrete regions of the sequence, which may be useful in experimental design.

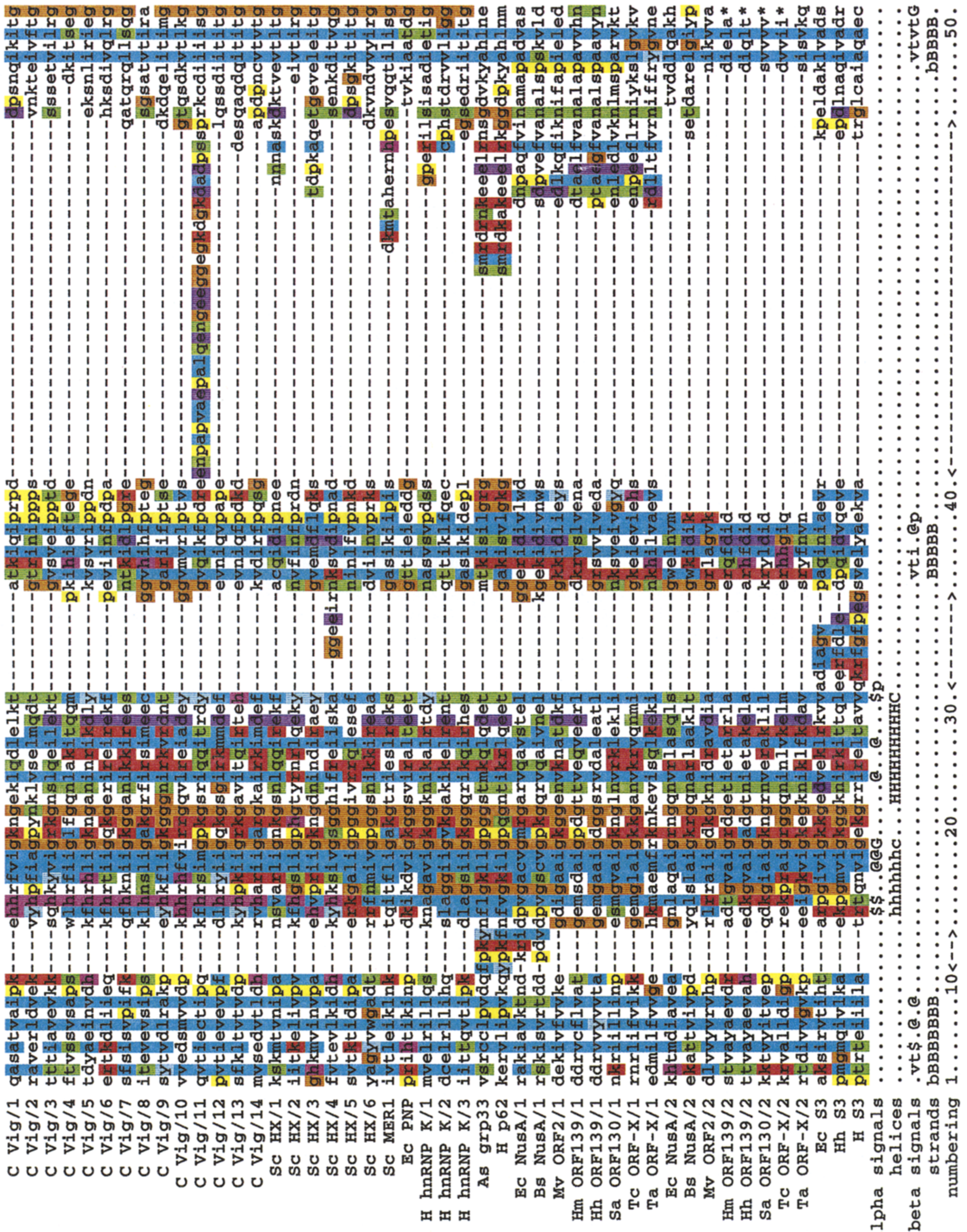

Fig. 2. KH domain alignment colour-coded to highlight significant features All G (orange) and P (yellow) are coloured. Other colouring is by recurring feature: hydrophobic, blue; hydrophobic tendency, lightblue; Q, N or S + T green; D or E, purple; R and K, red; H, pink. Generally, 40 or 50% occurrence of a property results in colouring. S and T are also coloured in a hydrophobic preference. Positive residues (R,K,H) are coloured above 30% occurrence or if they are an alternative to a hydrophobic preference. Symbols used for the structural signals are: @, conserved hydrophobic; $, semiconserved hydrophobic; B or H, strong structure assignment; b or h, weaker structure assignment, C, C-cap of helix.

The archaebacterial homologs of NusA would be expected to have a related function. A close involvement with RNA seems likely since the ORFs are sandwiched between four RNA polymerase subunit genes and three that encode ribosomal proteins. The simplest hypothesis is that they too are anti-terminators. However, as they are shorter, lacking ~ 200 N-terminal and ~ 150 C-terminal residues, it would follow that they associate with fewer factors and that regulation of transcription termination may therefore be simpler in these organisms.

4.1.2. The Vigilin/HX family. Vigilin and HX are linked by two features. First, they consist almost entirely of large numbers of KH repeats, 14 and 6 respectively. Second, in these proteins the KH repeats are preceded by a conserved and obviously helical ~ 15 residue motif [6]. Essentially the entire proteins thus have a common origin, although the domain amplification process has gone further in vigilin. Their functions will be built on the function of the single KH unit itself. High levels of vigilin expression have been shown for chondrocytes and other collagen-producing tissues, but its expression in other tissue types has not been assessed [6]. Therefore it is too early to judge whether a role in tissue-specific RNA metabolism is likely and whether this extends to an involvement with collagen gene transcripts.

4.1.3. The grp33/p62 family. These proteins share > 65% homology over ~ 100 residues including the KH domain [24]. The remainder of the sequences, which cannot be aligned, are non-globular and it is difficult to assess whether there are residual related elements. Grp33 was purified from hnRNP complexes of brine shrimp but has no known function. P62 was purified as a GAP-binding protein exhibiting tyrosine phosphorylation. Thus it is involved in the $P21^{ras}$ signalling pathway. The homology to grp33, which suggests a link between the signalling pathway and RNA metabolism, was very unexpected and prompted RNA band shift experiments [24]. It was found that both the intact protein and an N-terminal cyanogen bromide fragment, containing the KH domain, could bind and shift RNA. This preliminary experiment provides perhaps the most direct evidence that the KH domain is RNA-binding.

4.2 The probable role of the KH domain

Siomi et al. [1] have suggested that the KH domain might bind RNA on the basis of its presence exclusively in proteins involved in RNA metabolism. The additional KH-containing proteins presented here strongly support this hypothesis. Nevertheless decisive proof of such a function is lacking. If NusA does bind to RNA, it seems neither to recognise the sequence nor to be close to any bases in the terminator hairpin. Thus for NusA, RNA-binding would be independent of sequence. This is also true of polynucleotide phosphorylase which attaches to the 3′ terminus of an RNA molecule and processively degrades it to mononucleotides [32]. The ribosomal S3 protein crosslinks to ribosome-bound mRNA at a position 3′ to the decoding triplet bound by tRNA. This implies a close association, with no suggestion of sequence specificity [5]. Based on these three proteins, it seems reasonable to suggest that the KH domain functions in the non-specific recognition of single-stranded nucleic acid, perhaps with the ability to slide along the molecule without detaching. Although this could always be the function of KH domains, the possibility of sequence specificity in other KH proteins should not yet be ruled out. Indeed in the RNP motif family [33,34] there are both highly and poorly sequence specific members. MER1, which is involved in cell cycle-specific regulation of splicing activity is a candidate for sequence-specific binding, as perhaps is the vigilin protein.

If the KH motif functions in RNA-binding, several parallels can be drawn with the other generic RNA-binding motif, the RNP domain [33,34]. Both domains have a taxonomic distribution ranging from bacteria to eukaryotes, indicating an early origin in cellular evolution. Both occur, singly or in several copies, although there are no examples of massively repeated RNP domains in the manner of vigilin. Both are found in a broad range of proteins eg. RNP family members include *E. coli* transcription terminator rho [35], mRNA poly(A)-binding protein [36] and many snRNPs and spliceosomal proteins [listed in 33]. Differences are that the RNP domain is larger (~ 90 residues) and that there is much evidence for direct base-contact (e.g. [37,38]). Given the present state of knowledge, it would appear that the RNP domain occurs in situations where base sampling is necessary and the KH domain where it is not.

4.3. *Summary*

It has been shown here that the KH domain occurs in a wide variety of proteins whose only common property is that they are associated with RNA. The domain is usually, but not always, repeated two or more times. KH proteins are involved in translation initiation, transcription termination, $P21^{ras}$ signalling pathways, processive 3′–5′ degradation of RNA, meiosis-specific splicing and are frequently found in association with hnRNA. The limited experimental evidence suggests, but does not prove, that KH domains bind to single-stranded RNA in a non-specific manner.

Acknowledgements: We thank Iain Mattaj, Angus Lamond and Bertrand Séraphin for constructively reviewing the manuscript.

REFERENCES

[1] Siomi, H., Matunis, M.J., Michael, W.M. and Dreyfuss, G. (1993) Nucleic Acids Res. 21, 1193–1198.

[2] Régnier, P., Grunberg-Manago, M. and Portier, C. (1987) J. Biol. Chem. 262, 63–68.

[3] Engebrecht, J., Voelkel-Meiman, K. and Roeder, G.S. (1991) Cell 66, 1257–1268.

[4] Delahodde, A., Becam, A.M., Perea, J and Jacq, C. (1986) Nucleic Acids Res. 14, 9213–9214.

[5] Rinke-Appel, J., Jünke, N , Stade, K. and Brimacombe, R. (1991) EMBO J. 10, 2195–2202.

[6] Schmidt, C., Henkel, B., Pöschl, E., Zorbas, H., Purschke, W G., Gloe, T.R., Müller, P.K. (1992) Eur. J. Biochem. 206, 625–634

[7] Thompson, J.D., Higgins, D.G and Gibson, T.J (1993) CABIOS, submitted

[8] Hennikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad Sci. USA 89, 10915–10919.

[9] Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Proc. Natl. Acad. Sci. USA 84, 4355–4358.

[10] Genetics Computer Group (1991) Program Manual for the GCG Package, Version 7, April 1991, 575 Science Drive, Madison, Wisconsin, USA, 53711

[11] Bairoch, A. and Boeckmann, B. (1991) Nucleic Acids Res. 19, 2247–2249.

[12] Sideman, K.E., George, D.G., Barker, W.C. and Hunt, L.T. (1988) Nucleic Acids Res. 16, 1869–1870

[13] Pearson, W. R and Lipman, D.J. (1988) Proc. Natl. Acad. Sci USA 85, 2444–2448.

[14] Hamm, G.H. and Cameron, G.N. (1986) Nucleic Acids Res. 14, 5–10.

[15] Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C S. and Bilofsky, H.S (1985) CABIOS 1, 225–233.

[16] Heringa, J. and Argos, P. (1993) J. Mol Biol., submitted.

[17] Cruz-Alvarez, M. and Pellicer, A (1987) J Biol Chem. 262, 13377–13380.

[18] Klenk, H.P., Schwass,V. and Zillig, W. (1991) Nucleic Acids Res. 19, 6047.

[19] Lechner, K., Heller, G. and Boeck, A (1989) J. Mol. Evol. 29, 20–27.

[20] Pühler, G., Lottspeich, F. and Zillig, W. (1989) Nucleic Acids Res. 17, 4517–4534.

[21] Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. and Garrett, R.A. (1989) J. Mol. Biol. 206, 1–17.

[22] Ito, K., Egawa, K. and Nakamura, Y. (1991) J. Bacteriol. 173, 1492–1501.

[23] Klenk, H.P, Renner, O., Schwass, V. and Zillig, W. (1992) Nucleic Acids Res. 20, 5226.

[24] Wong, G., Müller, O , Clark, R., Conroy, L., Moran, M.F., Polakis, P. and McCormick, F. (1992) Cell 69, 551–558

[25] Zhang, X.T., Tan, Y.M. and Tan, Y.H (1990) Nucleic Acids Res. 18, 6689.

[26] Spiridonova, V.A., Akhmanova, A.S., Kagramanova, V.K., Koepke, A.K.E. and Mankin, A.S. (1989) Can. J Microbiol. 35, 153–159.

[27] Zurawski, G. and Zurawski, S.M. (1985) Nucleic Acids Res. 13, 4521–4526.

[28] Friedman, D.I. (1992) Curr Op. Gen. Dev. 2, 727–738.

[29] Tsugawa, A., Kurihara, T, Zuber, M., Court, D.L. and Nakamura, Y. (1985) EMBO J. 4, 2337–2342.

[30] Landick, R. and Yanofsky, C. (1987) J. Mol. Biol. 196, 363–377.

[31] Dissinger, S. and Hanna, M.M (1991) J. Mol. Biol. 219, 11–25.

[32] Guarneros, G and Portier, C. (1991) Biochimie 73, 543–549.

[33] Kenan, D.J., Query, C. C and Keene, J D. (1991) Trends Biochem. Sci 16, 214–220.

[34] Lamm, G.M. and Lamond, A. I (1993) Biochim. Biophys. Acta in press (July issue).

[35] Pinkham, J.L. and Platt, T. (1983) Nucleic Acids Res. 11, 3531–3545.

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

[37] Merrill, B.M., Stone, K.L., Cobianchi, F., Wilson, S.H. and Williams, K R. (1988) J. Biol Chem. 263, 3307–3313.

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