S. Adinolfi¹
C. Bagni²
M. A. Castiglione Morelli³
F. Fraternali¹
G. Musco¹
A. Pastore¹

¹ The National Institute for
Medical Research,
The Ridgeway, Mill Hill,
London NW7 1AA, U.K.

² Dipartimento di Biologia,
Università Tor Vergata,
Roma, Italy

³ Dipartimento di Chimica
Organica,
Università della Basilicata,
Potenza, Italy

Novel RNA-Binding Motif: The KH Module

Abstract: The KH motif has recently been identified in single or multiple copies in a number of RNA associated proteins. Here we review the current knowledge accumulated about the sequence, structure, and functions of the KH. The multidomain architecture of most of the KH-containing proteins inspired an approach based on the production of peptides spanning the sequence of an isolated KH motif. Correct identification of the minimal length necessary for producing a folded peptide has had a number of important consequences for interpreting functional data. The presence of the KH motifs in *fmr1*, the protein responsible for the fragile X syndrome, and their possible role in the *fmr1* functions are also discussed. © 1999 John Wiley & Sons, Inc. Biopoly 51: 153–164, 1999

Keywords: modular proteins; RNA-binding; fragile X; trinucleotide expansion

INTRODUCTION

In eukaryotic cells, the primary transcripts of protein coding genes, pre-mRNA or heterogeneous nuclear RNA, undergo a complex and highly regulated series of events that include excision of introns, generation of the 5'- and 3'-terminal structures, and polyadenylation before being exported as processed mRNA to the cyto-

plasm.¹ Specific RNA/protein interactions occupy a central position in the post-transcriptional regulation of gene expression. In particular, the RNA-binding proteins associated with nuclear pre-mRNAs—also known as heterogeneous ribonuclear proteins (hnRNP)—play a primary role in the control of post-transcriptional events.² The cloning and sequencing of cDNA for most of the hnRNP proteins have revealed the multidomain structure

Correspondence to: A. Pastore. Tel. no.: +44-181-959 3666.
Fax no.: +44-181-906 4477
Biopolymers (Peptide Science), Vol. 51, 153–164 (1999)
© 1999 John Wiley & Sons, Inc.

of many of the RNA-binding proteins and have led to the identification of several different motifs. One of the latest described is the K homology (KH) module, an evolutionary conserved sequence motif that has been originally identified as three repeats in the human hnRNP K protein from nuclear ribonuclear-protein complexes.³ Because of a suggested implication in RNA regulation that is increasingly supported by experimental evidence, the KH motif has been the subject of intense studies since its first identification.⁴ In the present paper, we review some of the knowledge accumulated so far about the sequence, structure, and functions of this module and the theoretical and experimental work done at the level of peptides spanning the KH sequence.

SEQUENCE ANALYSIS

The KH motif was first suggested to have a consensus of ≈ 30 amino acids,^{3,4} which was then extended to comprise ≈ 50 amino acids on the basis of sequence conservation.⁵ From sequence similarity, it was inferred that the consensus represented an autonomously folded unit. The motif expands around a conserved VIGxxGxxI core sequence (where x is any amino acid, with a preference for positive residues). KH modules are present in a wide variety of quite diverse proteins, from bacteria to human, indicating an early origin in cellular evolution. The only common property among members of the KH family is that they all function in close association with RNA.

The KH motif appears in most of the proteins of the family in multiple copies. Vigilin and its yeast homolog (Scp160) contain up to 15 repeats.^{6,7} Only relatively few members (although the number is rapidly increasing) contain one KH, usually flanked by other conserved regions.⁸ A new subfamily of KH-motif-containing proteins is encoded by genes that are conserved from yeast to humans. Mutations with interesting developmental phenotypes have been identified in *Ceanorhabditis elegans*,⁹ *Drosophila*,¹⁰ and mouse.¹¹ Members of this growing family, designated with the alternative acronyms STAR or GSG, include GRP33,¹² ZNF162,¹³ sam68,¹⁴ gld-1,⁹ Who/How,¹⁵ Xqua,¹⁶ and Qk1.¹¹ These proteins seem to homodimerize and be bifunctionally involved both in signal transduction and in RNA metabolism.^{8,17} Recently, two new proteins involved in splicing have been demonstrated to belong to the KH family and to bind RNA: the mammalian splicing factors SF1, which contains both a KH motif and a Zn-knuckles domain (another RNA-binding sequence)¹⁸ and the regulatory protein KSRP (KH type splicing regulatory) which contains four KH domains.¹⁹ Several

members of the KH family have been found to contain both a nuclear export (NES) and a nuclear localization signal (NLS) suggesting that the proteins may work as nucleocytoplasmic shuttles.^{7,20,21}

In the aligned sequences there are no absolutely conserved residues.^{22,23} This should exclude a catalytic role of the KH domain. A preference for positive charged residues on and between the proposed helices suggests a possible surface of interaction with a negatively charged partner, such as nucleic acids. Because the only common sequence motif in the whole family is the KH module, it was suggested that the KH domain is directly involved in RNA binding.⁵

INVOLVEMENT OF THE KH IN THE FRAGILE X SYNDROME

One of the most interesting members of the KH family is *fmr1*, the protein responsible for the fragile X syndrome.⁴ Fragile X is the most common cause of inherited mental retardation with an occurrence of one case over 4000 males and one over 7000 females.^{24,25} Because of its medical importance, it is noteworthy to spend a few words about this protein and the presence in it of the KH domain. The disease is characterized by mental retardation of variable severity, facial deformities, and macroorchidism in postpubertal males. The gene involved in the syndrome is the *fragile X mental retardation 1* gene (*fmr1*).²⁵ Appearance of the syndrome has been correlated with the presence of an anomalous expansion (up to thousand copies) of a trinucleotide triplet (CGG) in the 5'-untranslated region of the *fmr1* gene. As a consequence of this amplification, both the CGG repeat region and the CpG island within the promoter region become methylated,²⁶ thus causing transcriptional silencing of the *fmr1* gene which encodes in normal individuals the RNA-binding protein *fmr1*.³

The full-length *fmr1* protein (614 amino acids long) is predominantly cytoplasmic^{27,28} and mainly expressed in the brain and in testes, the major organs affected in the fragile X syndrome.²⁹ Several shorter forms have been observed in vivo as a result of alternative splicing of several of the 17 exons present in the gene.²⁵ Sequence analysis shows in all members of the *fmr1* family the presence of three potential RNA-binding domains: an RGG motif and two KH modules.^{3,30} Although no specific function has yet been associated with the *fmr1* protein, it appears to bind to RNA,⁴ to associate with ribosomes,^{31,32} and may affect protein translation.³³ An Ile304 $\rightarrow$ Asn missense mutation was identified in the second KH module of *fmr1* in a particularly severe fragile X

syndrome patient.³⁴ Fxr1 and fxr2, the two homologs of fmr1,^{30,35} share $\approx 60\%$ sequence homology with fmr1, mostly localized in the N-terminus up to the second KH. In accordance with the suggestion that the vertebrate genome underwent 4-fold amplification in a common ancestor,³⁶ these proteins are likely to show partial overlap in their function(s). In vitro, fmr1 interacts with both fxr proteins.³⁷

DOMAIN BOUNDARY IDENTIFICATION OF THE KH MODULE

A detailed structural investigation of the KH module is crucially important for understanding its function in such a diversified protein family. Until the last 3 years very little was known about the tertiary structure of the KH domain. Secondary structure prediction based on multiple sequence alignment⁵ had suggested that the KH motif might fold into an α - β structure with a three-stranded sheet interrupted by two contiguous helices.

The first step in structural and biochemical studies of modular proteins is the correct identification of the boundaries of each module. Conservation of a region ≈ 50 amino acids long suggested that this block should be necessary and sufficient to the KH to fold as an independent domain.²² The consensus sequence seemed to end at a very conserved glycine C-terminal to the third predicted β -strand, which coincides with the archaeobacterial nusA-like protein C-terminus. The secondary structure was predicted from the multiple alignment as a $\beta\alpha\alpha\beta\beta$ fold.⁵ On the basis of these indications, three synthetic peptides were produced spanning the regions suggested in ref. 5 and their folds analyzed by NMR: the seventh KH repeat from chicken vigilin,⁶ the second KH of the hypothetical protein 139 (ORF-139) from *Halobacterium halobium*,³⁸ and the fifth KH from the yeast protein Scp160.³⁹ Unfortunately all peptides resulted to have no tertiary structure as observed by both circular dichroism (CD) and nuclear magnetic resonance (NMR) (G.M. and A.P., unpublished data). Only the NMR spectrum of ORF-139 showed some residual secondary structure (as shown by sequential amide–amide correlations). These results suggested that the domain boundaries had not been correctly assigned and therefore the necessity of further inspecting the KH-containing sequences. A survey of the sequence database showed that most, but not all, KH motifs are preceded and followed by a stretch of about 18 amino acids with high helical propensity, suggesting a helix–KH–helix pattern. The pattern is particularly clear in

vigilin and the yeast Scp160 protein (from *Saccharomyces cerevisiae*)⁴⁰ (Fig. 1).

Repeat 6 of vigilin was selected, and four peptides spanning this sequence but with different boundary choices were expressed and purified: one with both flanking KH linkers (from residue 414 to 502 of the full-length protein), one with only the N-terminus linker (414–485), one which included only the C-terminal linker but prolonged the N-terminus 2 residues before the first residue of the KH consensus (433–502), and finally one with the KH domain alone (433–485).⁴¹ The four peptides had different behaviors: only those containing both the KH motif and C-terminal extension were expressed solubly and could be shown to be stable in time and folded. The C-terminal stretch should therefore be considered as part of the KH domain. This conclusion is consistent also with the alignments published by Deigaard et al.⁴² and Leffers et al.⁴³

STRUCTURAL ANALYSIS

Two solution structures of the KH domains from vigilin and from fmr1 have been solved so far by NMR spectroscopy^{23,44} (Fig. 2). The first is that of the 6th KH module of vigilin, also used for the identification of the domain boundaries. The second report concerns the structure of the first KH of fmr1 (fmr1-KH1). These two sequences were selected as representative of the KH fold family and because of the high medical interest associated with the fmr1 protein. To date no crystal structures of KH modules have been reported. The similarity between the two structures is within what is expected on the basis of their sequence homology. The tertiary fold of the KH consists of a $\beta\alpha\alpha\beta\beta$ structure with a three-stranded antiparallel β -sheet interrupted by spatially contiguous helices.⁵ A short additional helix (one turn) located between the first β -strand and $\alpha 2$ and T-shaped with the longitudinal axis of the β -sheet is clearly observable in vig-KH6 while in fmr1-KH1 the corresponding region is more disordered. The long loop connecting $\beta 1$ and $\alpha 2$ must be of functional relevance as it contains the conserved VIGxxGxxI motif that is the marker of the whole KH family. The C-terminal $\alpha 3$ helix packs back into the β -sheet and against $\alpha 2$ contributing to the formation of the hydrophobic core. The structure explains the necessity that the C-terminal helix is present to obtain a folded domain: without this stretch the hydrophobic core would remain exposed. Helices $\alpha 2$ and $\alpha 3$ are almost parallel in fmr1-KH while they form a more pronounced V-orientation in the vig-KH6 structure mostly caused by

FIGURE 1 Multiple alignment of the sequences of the subfamily of a subset of the KH family where the C-terminal addition was first identified. The two proteins where the extension of the KH motif was clearer were vigilin and its yeast Scp160 homolog. The repeats of these proteins are aligned with the sequences of *fmr1* and its homologs. It should be noticed that when the work was done the sequence of Scp160 was still partial. The alignment is color coded according to the amino acid properties. In the last lines we report the positions of the mutations for which alterations of the phenotype has been described. The secondary structure predicted from the pattern of conserved hydrophobic residues (indicated in blue in the alignment) is compared with the helical and β -sheet regions observed experimentally in the solved structures. Conserved or partially conserved hydrophobic are indicated with the special characters. The limits of the mini and maxi KH domains as defined are also reported (see the Redefinition of the Domain Boundaries section).

a different length and conformation of the loops preceding the helices.

REDEFINITION OF THE DOMAIN BOUNDARIES

After the importance of the extra C-terminal helix for the stability of the fold was established, the alignment of members of the KH family was re-examined using an extended KH domain profile.²³ In all the KH motifs selected in this way, the consensus can be extended further to match the required helical periodicity (Fig. 1). There is, however, a subset of sequences of the KH family, mainly consisting of nusA and ribosomal S3 proteins,^{45,46} in which no last helix exists, thus the misleading hypothesis of the domain boundaries. The second KH domain of the archaeobacterial nusA homologs follows immediately after $\beta 3$ of

the first KH, while the end of its own $\beta 3$ coincides with the protein C-terminus. It had therefore to be assumed that in such sequences the role of the C-terminal helix in burying the inner surface of the β -sheet is taken by other regions of the protein. In the S3 protein, the KH domain seems likely to have inserted within a loop of another globular fold. This is quite a common event in modular proteins, but whether it occurred in this case will have to be verified by solving the S3 structure. The family containing the additional helix has been referred to as maxi-KH (3 β -strands and 3 helices), and the other, with boundaries according to the original definition,⁵ as mini-KH (3 β -strands and 2 helices).²³

Redefinition of the KH consensus has a number of important consequences. Some of the motifs that had been overlooked (such as an N-terminal copy of the KH motif in vigilin, previously reported as having only 14 KH repeats) could be identified by the ex-

FIGURE 2 Comparison between the two structures solved.^{23,44} On the left is the average structure of the vig-KH6 repeat (vih.brk), and on the right, that of fmr1-KH1 domain (fmr1.brk). The two structures have been first structurally superposed and shown from the same orientation. The elements of secondary structure are named according to their sequential appearance in the KH fold (latin letters are used instead of the greek ones). Magenta and blue have been used to color the regions in β -sheet and α -helix, respectively. In fmr1-KH1, the region corresponding to α 1 in vigilin is possibly less structured and a helical turn does not appear in all the NMR structures. The presence of a short helix cannot be excluded because of spectral overlap.

tended profile.²³ The C-terminal boundaries of the second KH domain of fmr1 had to be revised. Assuming the extra stretch, the two KH repeats result sequentially contiguous. The second KH should span a much longer stretch than previously identified with an insertion (≈ 70 residues long) between two putative loops (between β -2 and β -3). The exact position of the insertion is clear when comparing the fmr1 sequence with that of fxr1 which contains a much shorter loop.³⁰ Unfortunately, all the attempts made so far to produce a folded second KH module of fmr1 or fxr1 have failed (S.A., unpublished results). Recombinant peptides spanning the regions 281–344 and 281–422 of the full-length protein (corresponding to the domain boundaries predicted in refs. 5 and 23) both led to unfolded structures. These negative results could be explained either assuming that other regions of the protein, such as the region 424–450 in exon 14 (which has a strong helical tendency) or a conserved stretch at the protein C-terminus, are essential for the fold of the domain or that the stability of the isolated peptide is much lower than that of the tandem one.

Insertions in the β 2/ β 3 loop are also observed in the STAR/GSG family.⁸ Several residues of the insertion are conserved among all the members with a stretch of charged residues. Since there are also almost no conserved hydrophobic side chains,⁸ the loop must have a functional rather than a struc-

tural role. The STAR/GSG subfamily is also characterized by a longer consensus that extends on both sides of the KH, forming what have been called Qua1 and Qua2 regions. Two stretches in Qua1 and Qua2 have been recently predicted to adopt a coiled coil structure in Qk1⁴⁷ and be directly involved in homodimerization and RNA-recognition, respectively. However, both regions are too short for the expected critical length of about 28 residues (i.e. 4 heptads) generally assumed necessary for coiled coil stability.⁴⁸ Consistently, the stretch in Qua1 is predicted only using a window of 14 amino acids and with low confidence (48%). The prediction confidence for Qua2 is stronger (up to 99%) but drops when a window of 28 residues is used. The presence of very disruptive residues, that is, a Glu in a “d” position of the coiled coil in Qua1 (Glu 48) and a Gln in an “a” position in Qua2 (Gln 161), also strongly speaks against the prediction (similar considerations apply to a coiled coil prediction reported for fmr1³⁷). It is much more likely that the two stretches form helices according to their strong helical periodicity but not coiled coils and pack against other regions within the same protein or in the homodimer. This suggestion would be supported also by the conserved hydrophobic character of the helix in Qua2.

FIGURE 3 Comparison of the structures of the three most common RNA-binding motifs: the U1A snRNP protein⁵⁰ (left), the fmr1-KH,⁴¹ (middle) and the dsRNA motif^{51,52} (right). On the top, a simplified representation of the β -sheet topology is shown.

COMPARISON WITH OTHER STRUCTURES

The structure of the KH module shows a mixed $\alpha\beta$ fold with a topology directly comparable with the topology of the MuA transposase structure.⁴⁹ However, the mutual orientation of the secondary structure elements as well as the length of the strands in the β -sheet are different. Intriguingly, among the structures picked up when searching the Brookhaven Protein Bank for fold similarity is the U1A snRNP protein⁵⁰ (with an RMSD of 3.1 Å on 59 atoms) (Fig. 3). This structure is formed by two intercalated β - α - β units. Comparison of the KH domain to RNP shows that the former lacks the last strand of the second β - α - β unit. However, the KH domain does not share any of the aromatic residues typical for the functional activity of RNP, therefore suggesting an independent evolution. An $\alpha\beta$ -fold is also present in the dsRBD,^{51,52} although with nonequivalent topology (Fig. 3). As in the RNP motif and the dsRBD (as well as in several other protein modules), the N- and C-

termini of the KH domain point to opposite ends of the structure. Common to the most favored folds in multidomain proteins is the positioning of N- and C-terminus at distal ends in the long axis of the module (in contrast with the proximity of N- and C-terminus in the majority of globular proteins), thus allowing the organization of the modules in those proteins which contain multiple copies as a consecutive array of beads usually without linkers between single modules. The extent of interactions and/or stabilization between the different domains remain to be established.

STABILITY OF THE KH DOMAIN

Stability studies on both the peptides for which the structures were solved have been reported.^{23,44} The melting point of both peptides as followed by circular dichroism was measured around 60°C. Interestingly a pH dependence of the stability has been observed between acidic and neutral pH for vig-KH6 but not for

fmr1-KH.⁵³ This behavior has been correlated to the presence in vig-KH6 and in all the other KH modules in vigilin of a conserved histidine that is otherwise replaced by a hydrophobic residue in most of the family. The histidine is buried in the structure and surrounded by other residues which are part of the hydrophobic core. What seems to stabilize the histidine in the protein interior is the formation of a hydrogen bond with the carboxyl group of a spatially close residue (Pro 46 in vig-KH6).⁵³ The hydrogen bond could either be direct or mediated by two water molecules trapped in the protein interior. In fmr1, replacement of a methionine in the corresponding position with a histidine leads to a folded peptide but with lower stability than the wild-type. The stability of the Ile → Asn mutant (associated with a severe fragile X case) has also been studied. The mutation in a semiconserved position at the beginning of the $\alpha 2$ helix impairs the fold completely leading to an unfolded peptide.

MAPPING NATURALLY OCCURRING MUTATIONS ON THE KH STRUCTURE

Loss of function caused by missense mutations in the KH domain have been demonstrated for members of the family, such as fmr-1, Bic-C, and Gld-1.^{9,10,34} These mutations lead to gain of function (dominant negative) phenotypes, which means that they have a stronger phenotype than null mutants caused by loss of gene expression. Several attempts to reproduce in vitro the effects of the Ile 304 → Asn mutation observed in the second KH of fmr1 have been described.^{34,54} Mutation of equivalent positions results either in lowering the affinity or in completely abolishing binding to RNA and polyribosomes.^{54–57} Ile 304 corresponds to Ile 32 in the vig-6 construct and to Ile 26 in the fmr1 KH1 construct. In both solved structures the Ile is placed in the second helix of the KH, surrounded by hydrophobic residues.^{23,44} On the basis of the mutagenesis results in vig-KH6 and fmr1-KH1 (Ile → Asn), it can be confidently concluded that the residue is essential for the packing. Missense mutation of Ile 304 in fmr1 may therefore be expected to disrupt completely the structure of the second KH domain, rendering its severe effect on the function of the protein unsurprising.

A number of KH domain RNA-binding proteins are known to be required during development of invertebrates. These include *Bicaudal C* in *Drosophila*, which is essential for anterior posterior specification, and *gld-1*, which encodes a protein which is essential for germ line development in *C. elegans*. For both

Bic-C and gld, mutations inside their KH domain have been reported, which can now be mapped and rationalized in the light of the KH structure. In Bic-C, the loss of function point mutation Gly 295 → Arg falls within the third KH repeat.⁵⁸ This mutation corresponds to the conserved Gly 59 in vig-KH6 and to Gly 51 in fmr1-KH1 present at the end of strand β -3 in the constructs. Substitution by arginine would insert a bulky, charged side chain into the KH hydrophobic core and seems to have severe structural consequences.

As with Bic-C, the three point mutations isolated in the gld-1 KH domain are also Gly replacements.⁹ However, in this case, they fall into two very distinct phenotypic classes. The substitution Gly 227 → Asp (Ser) has the full gld-1 loss of function phenotype. Gly 227 is the first glycine of the GkxG motif (corresponding to Gly 26 in the vigilin domain and to Gly 20 in fmr1-KH1). It is unlikely (but cannot be fully ruled out) that the KH fold is damaged, since this region of the structure is quite flexible. Given the evidence from the alignment that this residue is close to the possible RNA-binding site, the mutation seems likely to result in substantially lowered RNA affinity. By contrast, the two mutations Gly 248 → Arg and Gly 250 → Arg both exhibit a change of function phenotype, termed MOG (for masculinization of germline). These mutations map at the beginning of the $\beta 2/\beta 3$ of the KH domain structure and correspond to the exposed residues Pro 47 and Ser 49 in vig-KH6 and to Glu 42 Thr 44 in fmr1-KH1. Whether or not the following insertion was structured as suggested above, it is likely that this GKG motif (highly conserved in the STAR/CSC family) is part of a tight turn preceding a highly charged region. The mutations are unlikely to have a strong destabilization effect on the structure. More likely, they could alter an interaction surface with other proteins or with the rest of the protein. Since the mutations cause only a partial phenotype, they appear to disrupt only one specific function of gld-1, while leaving intact other functions of the protein.

EVIDENCE IN FAVOR OF A DIRECT INVOLVEMENT OF THE KH IN RNA BINDING

RNA-binding properties have been demonstrated for most if not all of the proteins in the KH family at least by in vitro binding to poly(homonucleotide)s.^{42,43,57,59} Different binding preferences are observed. These reports provide increasing evidence that KH-containing proteins play major roles in regulating cellular

RNA metabolism. The presence of the KH motif as the only sequence similarity among these proteins, which are present in diverse cellular locations, strongly suggests that the module itself functions as an RNA-binding domain. Support to this hypothesis is given by mutagenesis studies on the hnRNP K protein^{55,60}: RNA-binding was abolished or significantly reduced, not only by deleting partially or completely any of the KH regions but also in deletion mutants which lack the region C-terminal which would correspond to the additional helix.⁶⁰ Ribosomal protein S3 was shown to cross-link to mRNA in the ribosome.⁴⁶ A sam68 isoform almost entirely lacking the single KH copy but still containing an RGG box binds RNA with a 20% drop in the affinity as compared to the wild-type.⁶¹ The third domain of hnRNP K binds to poly-C, thus suggesting a putative involvement in pre-mRNA metabolism of transcripts containing cytidine-rich sequences.^{42,43,55} Constructs containing isolated KH domains of *fmr1* show strong binding with preference for poly-G (Adinolfi et al., manuscript in preparation).

The KH-containing proteins nusA, S3, and PNP operate in contexts where nonspecific RNA recognition is called for. With extrapolation from these proteins, it was originally suggested that the KH domain could recognize ssRNA without sequence specificity. However, accumulating evidence shows that KH-containing proteins exhibit strong preferences for binding to specific RNAs. The phenotypes of *Bic-C*, *gld-1*, and *fmr1* suggest that their gene products influence the cell fate of specific RNAs. Vigilin, almost entirely built by KH repeats, has been found to bind with high affinity the estrogen-inducible vitellogenin mRNA 3' untranslated region and a 3' end of human dystrophin mRNA.^{62,63} In vitro selection techniques performed on the neuronal RNA-binding protein Nova-1 (onconeural ventral antigen) have shown that it recognizes specific RNA targets.^{45,64} Mutagenesis studies defined the third KH domain of Nova-1 as necessary for in vitro binding to the [UCAU(N)(0-2)]₃ RNA sequence.^{45,64} In the protein context, when more than one KH repeat is present, the different KH domains may provide a cooperative effect and thereby modulate the overall affinity and sequence specificity of the individual domains.

HYPOTHESES ABOUT AN RNA-BINDING SURFACE

Analysis of the multiple sequence alignment and comparison with the structure suggests a possible recognition surface for RNA binding. The most strongly

FIGURE 4 Molscript representation of the vigilin structure with the residues involved in the naturally occurring mutations (in cyan). The GxxG loop is shown in green. Asp 37, indicated in yellow, corresponds to the position found to cross-link with RNA in S3. The positions are also marked under the alignment shown in Figure 1.

conserved motif in the KH alignment is the GkxG tetrapeptide between helices $\alpha 1$ and $\alpha 2$ (Figs. 1 and 4). The central positions of this loop are occupied by semiconserved positively charged amino acids. These residues are in an exposed position favorable for possible interactions with nucleic acids and are part of a flexible loop (as demonstrated by ¹⁵N relaxation data⁴⁴). Sequence conservation of the loop cannot therefore be explained by structural criteria and must have a functional role. Flexibility could be an intrinsic feature of the KH motif similar to what has been observed in the RNP motif⁶⁵ and as suggested for the dsRBD^{51,52}: an induced fit mechanism of RNA recognition would be expected to stabilize the loop upon

binding. Such binding surface is supported also by cross-linking of S3 with RNA at position 80 (that corresponds to residue 37 in vigilin and 31 in *fmr1* on helix $\alpha 2$,⁶⁶ Fig. 4). This residue is weakly conserved, with a slight preference for negative charges. Although cross-linking does not necessarily imply direct contact between the residue and the RNA, spatial proximity to the RNA-binding site is certainly expected.

The presence of conserved glycines in the GkxG loop is a feature common to other RNA-binding domains. The importance of glycines in RNA recognition has already been identified in the Tat-TAR peptide-RNA complex.⁶⁷ A conserved GxG loop is also thought to be involved in the helix-hairpin-helix (HhH) DNA-binding motif,^{68,69} a structural motif involved in nonsequence-specific recognition of both single and double-stranded DNA. Glycines have a steric advantage, as their small size allows closer contact necessary for main chain hydrogen bonding which would otherwise be prevented by amino acids with bulkier side chains. More fundamentally, glycines confer a high degree of conformational freedom to the polypeptide, which may facilitate the interaction with the nucleic acids. As observed in the HhH motif of endonuclease III⁶⁹ and in the ribosomal protein L11-C76,⁷⁰ the GkxG loop is the only part of the KH domain which protrudes out in solution, thus creating a flexible anchoring surface potentially available for RNA-binding.

The GkxG loop usually contains a lysine or an arginine either in second or third position. In a few examples, both or either positions are replaced by serines or threonines. The presence of semiconserved positive charges or of exposed hydroxyl groups suggest that the loop might bind to one or more phosphates. This hypothesis is reminiscent of the mechanism of binding in the HhH family^{68,69} the hairpin loop with the consensus G[I/V]G, immediately preceding the second helix. Position +2 after the second glycine has a high propensity for a hydrogen bond donor, usually Lys, Thr, and Ser. Two backbone amides of loop residues and two from the second helix N-terminus make direct contacts to two DNA phosphates, while the side chains of Lys, Thr, and Ser are well suited to make additional nonspecific hydrogen bonds with DNA phosphate oxygens. Although different in the glycine spacing, the KH domain GkxG motif and the following helix are remarkably similar to the HhH motif (with a conserved hydrophilic residue in +2 position). An analogous mechanism could therefore be suggested for nonspecific recognition of the KH with single-stranded RNA. Specificity should however be provided by additional binding sites

which might be different for each KH subfamily. It cannot be excluded, for instance, that additional binding sites could be present on the exposed surface of the two antiparallel helices which together contain most of the positive charges.

Four of the proteins with highly repeated KH domains—vigilin, Scp160, Bic-C, and C06G4.1—possess divergent KH domains in which the GkxG loop has insertions or deletions: in these repeats the site also lacks sequence conservation. The high sequence and length conservation of the GkxG motif in proteins with only a few KH domains points to functionally-imposed restraints, again suggesting that this is part of the RNA-binding site. Since loss of conservation occurs only in highly repeated domains, this might suggest that the RNA-binding affinity of the repeats is varied in multi-KH proteins. This could ensure that the combinatorial affinity of several KH repeats for RNA is not too strong or would allow regulatory processes to modulate the RNA-binding ability of the protein.

CONCLUSIONS

The KH motif seems to be a highly ubiquitous module occurring in proteins with functions ranging from RNA binding to cell transduction. The efforts made in the years since its first identification seem all to indicate that the primary function of the KH itself is the specific recognition of RNA by an induced-fit mechanism in which the GxxG loop, highly flexible and mainly unstructured in the isolated domain, acquires rigidity upon binding to RNA. Such a mechanism, found also in other protein-nucleic acid complexes, seems to be a common theme for RNA recognition. In the future, a deeper description of the functions of the KH module may significantly contribute to a better understanding of the KH-containing protein family and in particular of *fmr1*.

Note added in proof: After acceptance of the manuscript, a crystal structure of the third domain of Nova was published by an American team (Lewis H.A., Chen H., Edo C., Buckanovich R.J., Yang Y.Y., Musunuru K., Zhong R., Darnell R.B., Burley S.K. *Structure* (1999) 7, 191–203). The structure agrees well with the NMR structures previously reported with a higher similarity with that of vigilin-KH6.

We thank T. Gibson for helpful discussions and all those who have contributed, at different times, to the KH project.

REFERENCES

- Siomi, H.; Dreyfuss, G. Rna-binding proteins as regulators of gene expression. *Curr Opin Genet Dev* 1997, 7, 345–353.
- Dreyfuss, G.; Swanson, M. S.; Pinol-Roman, S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mrna formation. *Trends Biochem Sci* 1988, 13, 86–90.
- Siomi, H.; Matunis, M. J.; Michael, W. M.; Dreyfuss, G. The pre-mrna binding k protein contains a novel evolutionary conserved motif. *Nucleic Acids Res* 1993, 21, 1193–1198.
- Ashley, C. T.; Wilkinson, K. D.; Reines, D.; Warren, S. T. Fmr1 protein: conserved rnp family domains and selective rna binding. *Science* 1993, 262, 563–565.
- Gibson, T. J.; Thompson, J. D.; Heringa, J. 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. *Febs Lett* 1993, 3, 361–366.
- Schmidt, C.; Henkel, B.; Poeschl, E.; Zorbas, H.; Puschke, W. E.; Gloe, T. R.; Mueller, P. K. Complete cDNA sequence of chicken vigilin, a novel protein with amplified and evolutionary conserved domains. *Eur J Biochem* 1992, 206, 625–634.
- Weber, V.; Wernitznig, A.; Hager, G.; Harata, M.; Frank, P.; Wintersberger, U. Purification and nucleic-acid-binding properties of a *saccharomyces cerevisiae* protein involved in the control of ploidy. *Eur J Biochem* 1997, 249, 309–317.
- Vernet, C.; Artzt, K. Star, a gene family involved in signal transduction and activation of rna. *TIG* 1998, 13, 479–484.
- Jones, A. R.; Schedl, T. Mutations in *gld-1*, a female germ cell-specific tumor suppressor gene in *ceanorhabditis elegans*, affect a conserved domain also found in src-associated protein sam68. *Genes Dev* 1995, 9, 1491–1504.
- Mahone, M.; Saffman, E. E.; Lasko, P. F. Localised bicaudal-c rna encodes a protein containing a kh domain the rna-binding motif of *fmr1*. *EMBO J* 1995, 14, 2043–2055.
- Ebersole, T. A.; Chen, Q.; Justice, M. J.; Artzt, K. The quaking gene product necessary in embryogenesis and myelination combines features of rna binding and signal transduction proteins. *Nat Genet* 1996, 12, 260–265.
- Cruz-Alvarez, M.; Pellicer, A. Cloning of a full-length complementary cDNA for an *artemia salina* glycine-rich protein. *J Biol Chem* 1987, 262, 13377–13380.
- Caslini, C.; Spinelli, O.; Cazzaniga, G.; Golay, J.; De-Gioia, L.; Rambaldi, A.; et al. Identification of two novel isoforms of the *znf162* gene: a growing family of signal transduction and activator of rna proteins. *Genomics* 1997, 42, 268–277.
- Wong, G.; Muller, O.; Clark, R.; Conroy, L.; Moran, M.; Polakis, P.; McCormick, F. Molecular cloning and nucleic acid binding properties of the gap-associated tyrosine phosphoprotein p62. *Cell* 1992, 69, 551–558.
- Zaffran, S.; Astier, M.; Gratecos, D.; Semeriva, M. The held out wings (how) drosophila gene encodes a putative rna binding protein involved in the control of muscular and cardiac activity. *Development* 1997, 124, 2087–2098.
- Zorn, A. M.; Grow, M.; Patterson, K. D.; Ebersole, T. A.; Chen, Q.; Artzt, K.; Krieg, P. A. Remarkable sequence conservation of transcripts encoding amphibian and mammalian homologues of quaking, a kh domain rna-binding protein. *Gene* 1997, 188, 199–206.
- Chen, T.; Damaj, B. B.; Herrera, C.; Lasko, P.; Richard, S. Self-association of the single-kh-domain family members sam68, *grp33*, *gld-1*, and *qk1*: role of the kh domain. *Mol Cell Biol* 1997, 17, 5707–5718.
- Arning, S.; Grueter, P.; Bilbe, G.; Kraemer, A. Mammalian splicing factor *sf1* is encoded by variant *cdnad* and binds to rna. *Rna* 1996, 2, 794–810.
- Min, H.; Turck, C. W.; Nikolic, J. M.; Black, D. L. A new regulatory protein, *ksrp*, mediates exon exclusion through an intronic splicing enhancer. *Genes Dev* 1997, 11, 1023–1036.
- Kuegler, S.; Gruenweller, A.; Probst, C.; Klinger, M.; Mueller, P.; Kruse, C. Vigilin contains a functional nuclear localisation sequence and is present in both the cytoplasm and the nucleus. *Febs Lett* 1996, 382, 330–334.
- Eberhart, D. E.; Malter, H. E.; Feng, Y.; Warren, S. T. The fragile x retardation protein is a ribonucleoprotein containing both nuclear localization and nuclear export signals. *Mol Genet* 1996, 5, 1083–1091.
- Gibson, T.; Rice, P. M. J. T.; Heringa, J. Kh domains within the *fmr1* sequence suggest that fragile x syndrome stems from a defect in rna metabolism. *Trends Biochem Sci* 1993, 19, 331–333.
- Musco, G.; Stier, G.; Joseph, C.; Morelli, M. A. C.; Nilges, M.; Gibson, T. J.; Pastore, A. Three-dimensional structure and stability of the kh domain: Molecular insights into the fragile x syndrome. *Cell* 1996, 85, 237–245.
- Hagermann, R.; Cronister, A., Eds. Physical and behavioral phenotype. *The Fragile X Syndrome: diagnosis and treatment*; The John Hopkins University Press: Baltimore, MD, 1991.
- Verkerk, A. J.; et al. Identification of a gene (*fmr1*) containing a gcc repeat coincident with a breakpoint cluster region exhibiting length variation in fragile x syndrome. *Cell* 1991, 65, 905–914.
- Pieretti, M.; et al. Absence of expression of the *fmr1* gene in fragile x syndrome. *Cell* 1991, 66, 817–822.
- Devys, D.; Lutz, Y.; Rouyer, N.; Bellocq, J. P.; Mandel, J. L. The *fmr1* protein is cytoplasmic, most abundant in neurons and appears normal in carriers of a fragile x premutation. *Nat Genet* 1993, 4, 335–340.
- Sittler, A.; Devys, D.; Weber, S.; Mandel, J. L. Alternative splicing of exon 14 determines nuclear or cyto-

- plasmic localization of *fmr1* protein isoforms. *Hum Mol Genet* 1996, 5, 95–102.
29. Abitbol, M.; Menini, C.; Delezoide, A. L.; Rhyner, T.; Vekemans, M.; Mallet, J. Nucleus basalis magnocellularis and hippocampus are the major sites of *fmr1* expression in the brain. *Nat Genet* 1993, 66, 147–152.
30. Siomi, M. C.; Siomi, H.; Sauer, W. H.; Srinivis, S.; Nussbaum, R. L.; Dreyfuss, G. *Fxr1*, an autosomal homolog of the fragile x mental retardation gene. *EMBO J* 1995, 14, 2401–2408.
31. Khandjian, E. W.; Corbin, F.; Woerly, S.; Rousseau, F. The fragile x mental retardation protein is associated with ribosomes. *Nat Genet* 1996, 12, 91–93.
32. Tamanini, F.; Meijer, N.; Verheij, C.; Willems, P. J.; Galijaard, H.; Oostra, B. A.; Hoogeveen, A. T. *Fmrp* is associated to the ribosomes via rna. *Hum Mol Genet* 1996, 5, 809–813.
33. Feng, Y.; Gutekunst, C. A.; Eberhart, D. E.; Yi, H.; Warren, S. T.; Hersch, S. M. Fragile x mental retardation protein: nucleocytoplasmic shuttling and association with somatodendritic ribosomes. *J Neurosci* 1997, 17, 1538–1547.
34. DeBouille, K.; Verkerk, A. J.; Reyniers, E.; Vits, L.; Heyers, J.; Roy, B. V.; VanDenBos, F.; de Graaff, E.; Oostra, B. A.; Willems, P. J. A. Point mutation in the *fmr1* gene associated with fragile x mental retardation. *Nat Gen* 1993, 3, 31–35.
35. Coy, J. F.; et al. Highly conserved 3' utr and expression pattern of *fxr1* points to a divergent gene regulation of *fxr1* and *fmr1*. *Hum Mol Genet* 1995, 4, 2209–2218.
36. Spring, J. Vertebrate evolution by interspecific hybridization—are we polyploid? *Febs Lett* 1997, 400, 2–8.
37. Siomi, M. C.; Zhang, Y.; Siomi, H.; Dreyfuss, G. Specific sequences in the fragile x syndrome protein *fmr1* and *fxr* proteins mediate their binding to 60s ribosomal subunits and the interactions among them. *Mol Cell Biol* 1996, 16, 3825–3832.
38. Leffers, H.; Gropp, F.; Lottspeich, F.; Zillig, W.; Garrett, R. Sequence, organization, transcription and evolution of rna polymerase subunit genes from the archaeobacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhuae*. *J Mol Biol* 1989, 206, 1–17.
39. Delahodde, A.; Becam, A. M.; Perea, J.; and Jacq, C. A yeast protein *hx* has homologies with the histone *h2af* expressed in chicken embryo. *Nucleic Acids Res* 1996, 14, 9213–9214.
40. Wintersberger, U.; Kuhne, C. A. A. K. *Scp160p*, a new yeast protein associated with the nuclear membrane and the endoplasmic reticulum, is necessary for maintenance of exact ploidy. *Yeast* 1995, 11, 929–944.
41. Morelli, M. A. C.; Musco, G.; Stier, G.; Joseph, C.; Gibson, T.; Travé, G.; Pastore, A. The kh module has an $\alpha\beta$ fold. *Febs Lett* 1996, 358, 193–198.
42. Dejgaard, K.; Leffers, H.; Rasmussen, H. H.; Madsen, P.; Kruse, T. A.; Gesser, B.; Nielsen, H.; Celis, J. E. Identification, molecular cloning, expression and chromosome mapping of a family of transformation up-regulated *hnnp-k* proteins derived by alternative splicing. *J Mol Biol* 1994, 236, 33–48.
43. Leffers, H.; Dejgaard, K.; Celis, J. E. Characterization of two major cellular poly(rc)-binding human proteins each containing three k-homologous (kh) domains. *Eur J Biochem* 1995, 230, 447–453.
44. Musco, G.; Kharrat, A.; Stier, G.; Fraternali, F.; Nilges, M.; Gibson, T. J.; Pastore, A. Solution structure of the first kh domain of *fmr1*, the protein responsible for the fragile x syndrome. *Nat Struct Biol* 1997, 4, 712–716.
45. Liu, K.; Hanna, M. M. *Nusa* contacts nascent rna in *Escherichia coli* transcription complexes. *J Mol Biol* 1995, 247, 547–558.
46. Rinke-Appel, J.; Juenke, N.; Stade, K.; Brimacombe, R. The path of mrna through the *Escherichia coli* ribosome; site-directed cross-linking of mrna analogues rying a photo-reactive label at various points 3' to the decoding site. *EMBO J* 1991, 10, 2195–2202.
47. Chen, T.; Richard, S. Structure-function analysis of *qkl*: a lethal point mutation in mouse quaking prevents homodimerization. *Mol Cell Biol* 1998, 18, 4863–4871.
48. Lau, S. Y. M.; Taneja, A. K.; Hodges, R. S. Synthesis of a model protein of defined secondary and quaternary structure. Effect of chain length on the stabilization and formation of two-stranded alpha-helical coiled-coils. *J Biol Chem* 1984, 259, 13253–13261.
49. Clubb, R. T.; Omichinski, J. G.; Savilahti, H.; Mizuuchi, K.; Gronenborn, A. G.; Clore, G. A novel class of winged helix-turn-helix protein: the dna-binding domain of *mu* transposase. *Structure* 1994, 2, 1041–1048.
50. Nagai, K.; Oubridge, C.; Jessen, T. H.; Li, J.; Evans, P. R. Crystal structure of the u1 small nuclear ribonuclear-protein a. *Nature* 1990, 348, 515–520.
51. Bycroft, M.; Gruenert, S.; Murzin, A. G.; Proctor, M.; St. Johnston, D. Nmr solution structure of a dsrna binding domain from *Drosophila* *stau* protein reveals homology to the n-terminal domain of ribosomal protein s5. *EMBO J* 1995, 14, 3563–3571.
52. Kharrat, A.; Macias, M. J.; Gibson, T. J.; Nilges, M.; Pastore, A. Structure of the dsrna binding domain of *E. coli* *rnase iii*. *EMBO J* 1995, 14, 3572–3584.
53. Fraternali, F.; Amodeo, P.; Musco, G.; Nilges, M.; Pastore, A. Exploring protein interiors: The rôle of a buried histidine in the kh module fold. *Proteins* 1999, 34, 484–496.
54. Lin, Q.; Taylor, S. J.; Shalloway, D. Specificity and determinants of *sam68* rna binding. *J Biol Chem* 1997, 272, 27274–27280.
55. Siomi, H.; Choi, M. C.; Siomi, M. C.; Nussbaum, R. L.; Dreyfuss, G. Essential role for kh domains in rna binding: impaired rna binding by a mutation in the kh domain of *fmr1* that causes fragile x syndrome. *Cell* 1994, 77, 33–39.
56. Feng, Y.; Ansher, D.; Eberhart, D. E.; Brown, V.; Malter, H. E.; Warren, S. T. *Fmrp* associates with polyribosomes as an *mrnp* and the *i304n* mutation of

- severe fragile x syndrome abolishes this association. *Mol Cell* 1997, 1, 109–118.
57. Brown, V.; Small, K.; Lakkis, L.; Feng, Y.; Gunter, C.; Wilkinson, K. D.; Warren, S. Purified recombinant fmpr exhibits selective rna binding as an intrinsic property of the fragile x mental retardation protein. *J Biol Chem* 1998, 273, 15521–15527.
 58. Mahone, M.; Saffman, E. E.; Lasko, P. F. Localised bicaudal-c rna encodes a protein containing a kh domain, the rna-binding motif of fmrl. *EMBO J* 1995, 14, 2043–2055.
 59. Matunis, M. J.; Michael, W. M.; Dreyfuss, G. Characterization and primary structure of the poly(c)-binding heterogeneous nuclear ribonucleoprotein complex k protein. *Mol Cell Biol* 1992, 12, 164–171.
 60. Ito, K.; Sato, K.; Endo, H. Cloning and characterization of a single-stranded dna binding protein that specifically recognizes deoxycytidine stretch. *Nucleic Acids Res* 1994, 22, 53–58.
 61. Barlat, I.; Maurier, F.; Duchenne, M.; Guillard, E.; Tocque, B.; Schweghoffer, F. A role for sam68 in cell cycle progression antagonized by a sliced variant within the kh domain. *J Biol Chem* 1997, 272, 3129–3132.
 62. Dodson, R. E.; Shapiro, D. J. In vitro genetic analysis of the rna binding site of vigilin a multi-kh domain protein. *Mol Cell Biol* 1998, 18, 391–4003.
 63. Kanamori, H.; Dodson, R. E.; Shapiro, D. J. In vitro genetic analysis of the rna binding site of vigilin, a multi-kh domain protein. *Mol Cell Biol* 1998, 18, 3991–4003.
 64. Buckanovich, R. J.; Darnell, R. B. The neuronal rna binding protein nova-1 recognizes specific rna targets in vitro and in vivo. *Mol Cell Biol* 1997, 17, 3194–201.
 65. Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. Crystal structure at 1.92 a resolution of the rna-binding domain of the ula spliceosomal protein complexed with an rna hairpin. *Nat Struct Biol* 1994, 372, 432–438.
 66. Urlaub, H.; Kruf, V.; Bischof, O.; Mueller, E. C.; Wittmann-Liebold, B. Protein-rna binding features and their structural and functional implications in ribosomes as determined by cross-linking studies. *EMBO J* 1995, 14, 4578–4588.
 67. Puglisi, J. D.; Chen, L.; Blanchard, S.; Frankel, A. D. Solution structure of a bovine immunodeficiency virus tat-tar peptide/rna complex. *Science* 1995, 270, 1200–1203.
 68. Doherty, A. J.; Serpell, L. C.; Ponting, C. P. The helix-hairpin-helix dna-binding motif: a structural basis for non-specific recognition of dna. *Nucleic Acids Res* 1996, 24, 2488–2497.
 69. Thayer, M. M.; Ahern, H.; Xingand, D.; Cunningham, R. P.; Tainer, J. Novel dna binding motifs in the dna repair enzyme endonuclease iii crystal structure. *EMBO J* 1995, 14, 4108–4120.
 70. Markus, M. A.; Hinck, A. P.; Huang, S.; Draper, D. E.; Torchia, D. A. High resolution solution structure of ribosomal protein l11-c76, a helical protein with a flexible loop that becomes structured upon binding to rna. *Nature Struct Biol* 1997, 4, 70–77.