Letter

Prokaryotic Homologs of the Eukaryotic DNA-End-Binding Protein Ku, Novel Domains in the Ku Protein and Prediction of a Prokaryotic Double-Strand Break Repair System

L. Aravind¹ and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Homologs of the eukaryotic DNA-end-binding protein Ku were identified in several bacterial and one archeal genome using iterative database searches with sequence profiles. Identification of prokaryotic Ku homologs allowed the dissection of the Ku protein sequences into three distinct domains, the Ku core that is conserved in eukaryotes and prokaryotes, a derived von Willebrand A domain that is fused to the amino terminus of the core in eukaryotic Ku proteins, and the newly recognized helix-extension-helix (HEH) domain that is fused to the carboxyl terminus of the core in eukaryotes and in one of the Ku homologs from the Actinomycete Streptomyces coelicolor. The version of the HEH domain present in eukaryotic Ku proteins represents the previously described DNA-binding domain called SAP. The Ku homolog from S. coelicolor contains a distinct version of the HEH domain that belongs to a previously unnoticed family of nucleic-acid-binding domains, which also includes HEH domains from the bacterial transcription termination factor Rho, bacterial and eukaryotic lysyl-tRNA synthetases, bacteriophage T4 endonuclease VII, and several uncharacterized proteins. The distribution of the Ku homologs in bacteria coincides with that of the archeal-eukaryotic-type DNA primase and genes for prokaryotic Ku homologs form predicted operons with genes coding for an ATP-dependent DNA ligase and/or archeal-eukaryotic-type DNA primase. Some of these operons additionally encode an uncharacterized protein that may function as nuclease or an Slxlp-like predicted nuclease containing a URI domain. A hypothesis is proposed that the Ku homolog, together with the associated gene products, comprise a previously unrecognized prokaryotic system for repair of double-strand breaks in DNA.

The multifunctional eukaryotic protein Ku binds to discontinuities in double-stranded (ds) DNA such as double-strand breaks, single-strand gaps, and noncomplementary segments. The repair of double-strand breaks in eukaryotes occurs via a pathway of nonhomologous end-joining, or illegitimate recombination, that depends on the Ku protein (Critchlow and Jackson 1998; Featherstone and Jackson 1999). The Ku protein consists of two tightly associated subunits, Ku70 and Ku80, which bind DNA ends and transiently bring them together (Blier et al. 1993; Ramsden and Gellert 1998). In vertebrates, Ku has been shown to recruit the catalytic subunits of the DNA-dependent protein kinase to initiate a phosphorylation and protein-protein interaction cascade that, in turn, leads to the recruitment of repair enzymes including DNA ligase IV, whose activity the Ku protein stimulates in vitro (Gottlieb and Jackson 1993; Teo and Jackson 1997, 2000; Ramsden and Gellert 1998). Ku is also a part of the telomerebinding complex and is required for the perinuclear localization of the telomeres (Hsu et al. 1999, 2000; Mishra and Shore 1999; Galy et al. 2000). In addition, Ku forms complexes with numerous other chromosomal proteins such as HP1a, Werner syndrome helicase and poly(ADP-ribose)-polymerase, along with which it binds to chromosomal matrix-attachment re-

¹Corresponding author.

E-MAIL aravind@ncbi.nlm.nih.gov; FAX (301) 480-9241.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/ gr.181001. gions (MARs) (Galande and Kohwi-Shigematsu 1999; Li and Comai 2000, 2001; Song et al. 2000).

Ku70 and Ku80 are paralogs (Gell and Jackson 1999) and are both conserved throughout the eukaryotic crown group as well as in early-branching eukaryotes such as trypanosomes. This suggests that Ku is an ancient component of the DNA repair and chromatin integrity system, with the duplication that gave rise to Ku70 and Ku80 probably predating the divergence of most, if not all, extant eukaryotes. Prokaryotic counterparts of Ku and the entire illegitimate-recombinationdependent double-strand break repair system, of which Ku is a central component, have not been identified.

We have reported previously the presence of a homolog of the small, catalytic subunit of the eukaryotic-archeal DNA primase (EP) in several bacteria including *Bacillus, Mycobacterium*, and *Streptomyces* (Koonin et al. 2000). The gene for this predicted primase is fused to or juxtaposed with a gene for a eukaryotic-archeal ATP-dependent DNA ligase (ADDL), which suggests a functional association between the two enzymes and the presence of a previously undetected, eukaryotic-type DNA repair mechanism in bacteria. Here, we report the first prokaryotic homologs of the DNA-binding protein Ku and discuss evidence that they are part of the same DNA repair system with EP and ADDL. This analysis also reveals the modular architecture of the Ku proteins and allows us to define ancient protein modules involved in DNA repair and other aspects of nucleic acid metabolism.

RESULTS AND DISCUSSION

Bacterial and Archeal Ku Homologs

To gain further insight into the functions of the eukaryotetype DNA ligases and primases in bacterial DNA repair, we searched the gene neighborhood of the genes encoding these proteins for other conserved genes. The Bacillus subtilis gene ykoV is adjacent to the ykoU gene, which encodes a twodomain protein with fused EP and ADDL domains; the YkoV protein is highly conserved in all bacteria that encode an EP, but is not detectable in any other bacterial species. Furthermore, the juxtaposition of the ykoV orthologs and the genes coding for EP or ADDL is maintained in phylogenetically diverse bacteria, including Mycobacterium tuberculosis, Streptomyces coelicolor, Mesorhizobium loti, and Bordatella pertussis, and the archeon Archaeoglobus fulgidus. This strong preservation of gene neighborhood of the EP, ADDL, and YkoV orthologs suggests that these genes belong to the same operon, although the exact gene arrangement is variable (Fig. 1). Gene neighborhood or operonic cooccurrence of genes is conserved between multiple, distantly related prokaryotic genomes only when the products of the corresponding genes interact functionally, and often, also physically (Dandekar et al. 1998; Wolf et al. 2001). Hence, it appears most likely that YkoV and its orthologs form a functional complex with EP, ADDL, and another predicted conserved protein (SC9H11.09c and its orthologs) that is also associated with these predicted operons (Fig. 1).

To identify potential distant homologs of the YkoV protein and thus possibly predict its function, we performed iterative PSI-BLAST database searches (Altschul et al. 1997) using the sequences of YkoV and its orthologs as queries; the searches were run to convergence, with a profile inclusion threshold of expect (*E*) value of 0.01. Most of these searches detected eukaryotic Ku proteins in the second iteration. For example, the search with the sequence of AF1726, the A. fulgidus ortholog of YkoV, detects the central region of the fission yeast Ku70 subunit with $E = 10^{-4}$ in iteration two and, at convergence, retrieves only the eukaryotic Ku70 and Ku80 proteins. Similarly, reverse searches with the corresponding regions of Ku70 sequences retrieve first the eukaryotic orthologs, then the paralogous Ku80 sequences and, finally, the prokaryotic YkoV-like proteins (e.g., in a search initiated with the sequence of the central region of human Ku70, B. subtilis YkoV is detected in iteration two with $E = 10^{-3}$). This shows that the YkoV-like proteins are the prokaryotic homologs of the Ku70 and Ku80 proteins. The region of similarity shared by these proteins covers almost the entire length of most prokaryotic proteins. In contrast, the eukaryotic Ku proteins are much larger and contain a conserved amino-terminal extension and, in the case of Ku70, also a conserved carboxyterminal extension. Thus, the prokaryotic Ku homologs described here appear to define a distinct, previously unnoticed domain that forms the ancient core of these proteins. The conserved blocks, identified previously in the eukaryotic Ku proteins and termed 'primary homology regions' (PHR) 3-5 (Gell and Jackson 1999), map entirely within this domain shared by the eukaryotic and prokaryotic Ku homologs. In contrast, PHR 2 and 3 (Gell and Jackson 1999) map to the amino-terminal region exclusively shared by the eukaryotic Ku proteins.

This core domain shared by the prokaryotic YkoV-like proteins and the eukaryotic Ku70 and Ku80 (hereinafter Ku core; Fig. 2) is ~234–280 amino acids long and is larger than most common globular domains. The multiple-alignment-based secondary structure prediction using the PHD program (Rost and Sander 1993) shows that the Ku-core domain is likely to form two distinct substructures. The amino-terminal region (~85 residues) is poorly conserved and is predicted to form a β -strand-rich subdomain. The remaining portion is



Figure 1 Gene organization in the predicted operons encoding components of the postulated novel double-strand-break repair system. The direction of an arrow indicates the direction of transcription. Distinct regions of each gene encoding separate domains in the protein, such as primase and ligase, are indicated in different shades.

ERERGEGEGE, ERE	••••••••••••••••••••••••••••••••••••	e alignment of the Ku-core domains. The secondary structure predicted using the PHD program is shown <i>above</i> the alignment. E indicates a β-strand and H e uppercase used to denote the most confident prediction (>82% accuracy). The 90% consensus shown <i>below</i> the alignment was derived using the following KRHEDQNST) colored blue; hydrophotic (In: ALICVMYFW) and the alignmetic subset of these are (I: ALIVMC) all shaded yellow; small (s: ACDGNPSTV) colored one diverted to the <i>right</i> of the alignment. The second park, big (b: Q.E,R,K,Y,M,F,W,L,I) shaded gray. The limits of the domains are indicated by the position numbers on each side of the alignment. The set are directed by the rotention numbers on each side of the alignment. The set are denoted by their gene names followed by the species abbreviations and GenBank identifiers. on of this manuscript, the prokaryotic KU homologs were identified in the SMART database (Schultz et al. 1.998). The species abbreviations are: At. <i>Arabidopsis</i> Mm, <i>Mus musculus</i> , Dm, <i>Drosophila melanogaster</i> ; Sc. Soccharomyces cereviside; Sp. Schizosaccharomyces pombe; Af, Archaeoglobus fulgidus; Pa, <i>Pseudomonas</i> is; Scoe, Streptomyces coelicolor; Bpe, Bordatella pertussis; ML, <i>Mycobacterium tuberculosis</i> ; MI, <i>Mesohizobium</i> loti.
 LE LELENCY LE LELENCY LE LELENCY LITISENI LIT	KEVAALGE REVIAALGE REVIAALGE RUKLAALGE SURTOTAAL SURTOTAAL LERFAVAE LLERFAVAE LLERFAVAE LLERFAVAE LLERFAVAE RUKLAVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN TGKVALAN	nce alignir he upperci p: KRHEDC) colored p ains are inc ion of this 5, Mm, <i>Mu</i>
ture: 92415 2 92415 2 666 2 15 229270 2 2478 2 548736 2 6448736 2 1479721 1942 2 1479721 1942 8 1468 8 8469 8 8469	tture: 992415 992415 66 66 53157 53157 53157 53157 53157 53157 53157 53157 1942 1942 1942 1942 1942 1942 1942 3869 8469	tiple sequel lelix, with tl ses: polar (r (c: DEHKR) (c: DEHKR) (c
Secondary struc KU70_Dm.6226892 Yku70_Dm.6226892 F611.25 02C_Sp_74 F611.25 4 9627 Ku80_Hs_1086394 Ku80_Ls_1086394 Ku80_Ls_1086394 F11347 Ku80p_Sc_52537 T2115.4_At_1064 F11347 T2115.4_At_1064 F11347 T2115.4_At_1064 F11347 T2115.4_At_1347 m112074_D1_3481 m114607_At_13481 m114607_	Secondary struc KU70_Dm.6226892 CU70_Dm.6226892 FG170P_SC.622892 FG112_At_98027 FG112_At_98027 FG112_At_98024 FG112_At_1063 FG123_03_5299 BC543_03_529 BC543_03_52626 At1754_At_1064 At1754_525_525_526 At1754_54_11499 At1757_4At_11493 At1757_4At_11493 At177275_255_526 At177271215_4At_11493 At1772715_4At_11493 At1772715_4At_11493 At1772715_24At_11347 mi19607_M1_1347 mi19607_M1_1348 mi14607_M1_1348 C0056053441_1348 C0156053441_1348 C0156053441_1348 C0156053441_1348 C0156053441_1348 C0156056667 M119607_M1_1348 C015605666666666666666666666666666666666	Figure 2 Mult indicates an α -h amino acid class green, charged i subclasses of Ku. subclasses of Ku. subsequent to the thaliana; Hs, Ho, aeruginosa; Bs, B

Prokaryotic Ku Homologs

Aravind and Koonin

more strongly conserved and is predicted to form an α/β structure ending in a strongly predicted bihelical hairpin (Fig. 2). This complex fold is consistent with the functions associated with this region as demonstrated by experimental studies on the eukaryotic Ku70 and Ku80. The principal determinants of heterodimerization (Osipovich et al. 1997; Cary et al. 1998; Koike et al. 1998; Gell and Jackson 1999) and DNA-binding (Wu and Lieber 1996; Wang et al. 1998; Osipovich et al. 1999) of these proteins map to the Ku-core domain as defined by the present sequence comparisons with the prokaryotic Ku homologs. This region also mediates the interactions of the eukaryotic Ku proteins with other chromosomal proteins (Song et al. 2000). Thus, the prokaryotic Ku homologs are predicted to form a homodimer that binds DNA and also associates with other proteins via the conserved Ku core.

The common ancestor of the prokaryotic and eukaryotic Ku proteins might have resembled the extant prokaryotic version, with the essential functions of dimeric DNA-endbinding and interactions with other components of the DNA repair complex. The conserved structure of the predicted operons that encode Ku homologs (Koonin et al. 2000) strongly suggests that these proteins function together as subunits of a protein complex with a possible role in DNA repair or replication. The Ku homologs, EP, and the associated ADDL show sporadic distribution in prokaryotes, as is typically the case with DNA repair systems (Aravind et al. 1999). This is in sharp contrast to the DNA replication components such as, for example, bacterial-type DnaG-primase or NAD-dependent DNA ligase, which show a practically universal distribution among bacteria, including those that possess the Ku-EP-ADDL operons. These observations support a function for these proteins in a DNA repair system, most likely one involved in correction of double-strand breaks in DNA, similar to their eukaryotic counterparts.

In addition to the EP and ADDL, other potential components of the predicted prokaryotic, Ku-associated DNA repair system are revealed by examination of the operons encoding these proteins. Rv0938 contains a conserved domain between its ADDL and EP domains (Fig. 1) that occurs as a stand-alone protein (SC9H11.09c) in S. coelicolor. A homologous domain is also present amino-terminal of the ADDL and EP domains in the ligase-primase proteins from Pseudomonas aeruginosa, B. pertussis, and M. loti, which in the latter two organisms cooccur in a predicted operon with the genes coding for Ku homologs. Thus, this uncharacterized domain is only found in those organisms that also encode Ku and EP, and, given the predicted operonic organization, probably interacts with them functionally. A multiple alignment of this domain reveals conserved histidine and aspartate residues that could form a metal-coordinating cluster within an all β-strand fold (Fig. 3A). This strongly suggests a catalytic function, most probably that of a DNAse, for this conserved domain.

One of the predicted Ku-encoding operons from *M. loti* includes a small gene (Msl2076) between the genes coding for the EP–ADDL fusion protein and the Ku homolog (Fig. 1). This gene shows the same direction of transcription as the two other genes and probably is a part of the operon. Sequence profile searches with the Msl2076 showed that it belongs to a distinct family of UvrC-Intron-type (URI) endonucleases (Aravind et al. 1999) typified by the *Escherichia coli* YhbQ, *B. sub-tilis* YazA, and yeast Slx1p. This family of URI nucleases (Fig. 3B) is represented widely in single or duplicate copies in bacteria, eukaryotes, DNA viruses, and, so far, in a single archeon, *Halobacterium salinarium*. The prokaryotic members of this

family are characterized by their distinct, small size (typically, <100 amino acids); thus, they represent stand-alone forms of the URI endonuclease domain. The eukaryotic members typified by the yeast DNA repair protein Slx1p (Mullen et al. 2001) contain an additional, carboxy-terminal PHD-finger domain, whereas one of the paralogs in *Arabidopsis* is fused to the MutS DNA-repair ATPase (Fig. 3B). Yeast Slx1p functionally interacts with the yeast RecQ-like helicase Sgs1p and is likely to function in resolution of recombination intermediates in DNA repair (Mullen et al. 2001), which is consistent with its predicted nuclease activity. Thus the Slx1p-YhbQ family of proteins is likely to define a highly conserved repair-recombination pathway present in both eukaryotes and bacteria. This hypothetical repair pathway might interact with the predicted Ku–EP–ADDL-dependent pathway.

The Helix–Extension–Helix Fold and its Association with the Carboxyl Terminus of the Ku-Core Domain in Bacteria and Eukaryotes

In an attempt to glean more details of the functional interactions and evolution of the Ku proteins, we analyzed the domains that are associated with the Ku core in eukaryotes and prokaryotes. All prokaryotic Ku homologs, with the exception of SCF55.25c from S. coelicolor, consist of the Ku-core domain alone. SCF55.25c contains a carboxy-terminal extension of ~40 amino acid residues that show significant sequence similarity to several small, uncharacterized proteins from bacteria, bacteriophages, and Arabidopsis thaliana. Iterative PSI-BLAST searches resulted in the detection of the same region of similarity in the bacterial transcription terminator Rho, where it occurs at the extreme amino terminus, immediately upstream of the OB-fold domain. The conserved region precisely corresponds to the amino-terminal α -helical domain of Rho (hereinafter Rho-N) as defined by its X-ray and NMR structures (Allison et al. 1998; Bogden et al. 1999). To further investigate the distribution of this small domain, we searched the protein structure database using the DALI search tool (Holm and Sander 1998). This resulted in the detection of two structures with high similarity to Rho-N, namely the carboxy-terminal domain of Endonuclease VII (Raaijmakers et al. 1999) (a nuclease and Holliday junction resolvase from bacteriophage T4) and the small α -helical domain inserted into the catalytic domain of bacterial and eukaryotic lysyl-tRNA synthetases (KTRS) (Onesti et al. 2000). In PSI-BLAST searches initiated with the Rho-N domain sequence, these proteins were detected with borderline *E*-values. A structure-based sequence alignment of the Rho-N domain with the α -helical domains of KTRS and Endonuclease VII shows that, in addition to the structural similarity, they contain the conserved residues characteristic of Rho-N and its homologs that were detected in sequence searches (Fig. 4A). Thus, we conclude that these α -helical domains have a common evolutionary origin and define a novel superfamily of ancient mobile domains that are found in various contexts related to nucleic acid metabolism.

Examination of the multiple alignment and the three structural prototypes of this superfamily (Rho-N, Endonuclease VII carboxy-terminal domain, and KTRS-insert domain) shows that these domains share a novel fold (Fig. 4B) that is distinct from other α -helical folds found in small, primarily nucleic-acid-binding domains such as Helix–turn–Helix, Helix–loop–Helix, and Helix–hairpin–Helix (Doherty et al. 1996; Aravind and Koonin 1999; Massari and Murre 2000). In this newly detected fold, the first short, mobile helix almost im-

 EEEEEEE JHTJ 1943 TGUVYNTASQKRGTIT KIGVTNDLGR NP BETAGGGSNFTSRYGVQRLWYEEHPIDITDA LQBEKSEKRWP-RLWKIELJEKANPEMFELFR JAT 1943 TGUVYNTASQKRGTIT KIGVTNDLGR NP BETAGGGSNFTSRYGVQRLWYEEHPIDITDA LQBEKSEKRWP-RQWKIDLIERANPEMFELFR JAT 1943 TGUVYNTASQKRGTIT KUTGVTNDLGR NP BETAGGGSNFTSRYGVQRLWYEEHPIDITDA LQBEKSEKRWP-RQWKIDLIERDNPEMPELFR JAT 1943 TGUVYNTASQKRGTIT KUTGVTNDLGR NP BETAGGGSNFTSRYGVQRLWYEEHMLMTEAL QBETSEKRWP-RQWKIDLIERDNPRMPELFR MITAVMMADHYRGTILT KVTGVTNDLGR NP BETAGGGSNFTSRYGLTRLWWYEEHMLMTEAL QBETSEKRWP-RQWKIDLIERDNPRMPDLYA MITAVMMADHYRGTILT KVTGVTNDVRRYCHARGGGSNFTSRYGLTRLWYEEHMLMTEAL QBETSEKRWP-RQWKIDLIERDNPRMPDLYA SILAVTMARPYGTILT KVTGVTNDVRRYCHARGGGSNFTSRYGLTRLWYEEHMLMTEAL QBETSEKRWP-RQWKIDLIERDNPRMPDLYA MITAVMMARPYGTILT KVTGVGR NGRLAGGGARFFHSSP ANYNILASQRNGTILT KVTGARATTKYCH ANYNILASQRNGTILT KVTGARATTKYCH ANYNILASQRNNKLYNG FTAVDVRRYCHARGANATTKRCHARALLYVERCKMP-RAKKON-RAKKIDVYEWEDDLYA ANYNILASQRN	 Jadaszek Jadaszek<	351068 36 RYCW 3473868 34 IFVIR 3471944 26 RFVVR 3488470 34 RFVVR 7 SFVVI 00% .hli	VE DASRI VER HKI AATRI VELK HKI HATAD VELK HKI HATAD VELK QRI DATRI VELK QRI QARRI VELK H.A HADHR	LECDGVLVSW LELDGTLKSW LELDGTLKSW LELGGVLVSW LELGGVLVSW LELDGVLVSW	AVPRNLPDNT AVPRGPCLDP AVTRGPSLDP AVTRGPSLDP AVPRGPSLDP AVPRGPSLDP AVTRGPSLDP	FSVNHLATH PAVKRLAVQ PHEKRLAVH PADKRLAVE PHDKRLAVE PHDKRLAVE PSVKRLAVE PSVKRLAVI	TBDH PLEYATPOCA VBDH PLDYADFDCS VBDH PLDYAPFDC1 VBDH PLDYAPFDC1 VBDH PLEYLDFDCV VBDH PLEYLDFDCV VBDH PLEYCGFSGD SEDHP1-Y. FDG.	ALPSGEYGAG SIPOGHYCAG TIPKGEYGGG TIPKGEYGGG TIPKGQYCAG TP.G.YGUG	KVIIW SGT DVIVW EGT SVIVW EGT PMIVW FIGT TVMIVW FIGT TVMIM SGV HVDIW SGV	YDTEKF - HDI WTPLDDP-REO WMPETDP-AKU WAPMDDV-EK WEPEGKN2EQ WIPRGDP-RQC	DPHTGEVIN GLEKGHLSF AMKKGHLSF SLRTGAFKF SLRTGAFKF ALAKGDFKF GLRRGKLHF	PALDGEKLS FELNGHKLF PRLAGEKLN TLEGERLA ALRGRRLA	SGRWHLIRTNI HGLWHLVRLKF HGGWMLTRLKF HGSFVLVRMAN NGDWVLVRMAN G.a.L.R		HRLKN 4 VKAKD 1 LKSDD 1 FKERD 1 FKERD 1 IKHRD 1 RKEDD ?
 	 														
 3471943 2 TGY UTM TASÖKRGTLYT GVTWDLGR WP BELKSGE GSRFT SQYGVÖRL WY EEYFDITDAT OF BKSLKRRL-RQWKIDLIEKANPDWFELFR 91 3425248 15 WIA YYTMADHTRGTLYT GVTRAOFT TTD BEGRGSAFTRRYGLTRL WY EEHMLMTEAT OR BT STKRWP - RQWKINLIERDNPRWDDDYD 96 3425248 15 MIA YYTMADHTRGTLYT GVTRAOFT TTD BEGRGSAFTRRYGLTRL WY EEHMLMTEAT OR BT STKRWP - RQWKINLIERDNPRWDDDYD 96 3425269 2 OPA YTLASORNGTLYT GVTSOR LKGT OR HGUTML WY EEHMLMTEAT OR BT STKRWP - RQWKINLIERDNPRWDDDYD 96 3425264 14 CWS YTLASORNGTLYT GVTSOR LY GRART FRSDNTML WY EEHMLMTEAT OR BT RAKLOLIERDNPYLIREND 90 357 14 CWS YTLYRADGTLYT GVTSOR LY GSGKGARETRS FORNALVY FSA PYGGDAL REBRAIKALS - KRAKERLIVGNVLLT GT TTDV FSA PYGGDAL REBRAIKALS - KRAKERLIVGNVLLT GT TTDV FSA PYGGDAL REBRAIKALS - KRAKERLIVGNVLLT GT TTDV FSA PYGGGAL RAGA AND PACKINT - RAKLOLIERENNOV RAKLOV 93 379016 14 CWS YTLYRADNKLT GT TTDV FRED GGKGARKTRNALVY FSA PYGGDAL RABATRALS - KRAKERLIVGAGRAELLS 91 37015 31 2004 11 2 FYCCTLICSTIKRSFLYVERARA AND PKGARKTRPRUNNUCUVHGF PSKLAALOF FRAMALNERAKKGGR 110 3129014 31 3 FFGYLLINCLARPE - YRGRYT RAKGGARKTRRAMMUUNH FSLHT PASKLATRAKKGGR 110 3129014 31 3 FFGYLLINCLARPESKYVERAAA FROM MY CUVHGF PSKLAALOF FRAMALNERAGGARTRRAKGGARTR	 3471943 2 TGYVTMTASQKRGTITICOTEGENERGESETTERYGVQRLWYEEYPITIDLICR_SKLKRRL-RQWKIDLIEKANPDWFELFR 9 3422692 1 GYTVTMADHYYGTLYTGVTREPRESEGSETTERYGLTRLWYEEHMLMTERIGRESEWPFROMKINLIERDNPRMDDLYA 10 3422692 1 SIAYTMANDHYYGTLYTGVTREPRESEGSETTERYGLTRLWYEEHMLMTERIGRESEWPFROMKINLIERDNPRMDDLYA 90 3422692 2 QPAYTLASQRNGTLYTGVTREPRESEGSETTERYGVTMLWYETHADMTYLIHRERTEREWP-ROWKINLIERDNPRMDDLYA 90 3422692 2 QPAYTLASGRNGTLYTGUTSTERERGSERFRAGGENERFERGENERFRAGENERFERGENERFRAGENE	ructure: 3473728 2	. effeffee.	CRGT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	LGREVPE	TRAGOGSNETSE	RYG	-VORL	GehHH	MI OR KSI	LKRWP-R	I	KANPEWFE L	FR 9
3425248 15 WIAVYMMADHYRGTLYAGVTAQFITRITD REGRGSAFTRYGLTRLWWEBHMIMTERIORFTCRWP-RQWKINLIERDNPRWDDLYA 104 3422692 7 SIAVYTMANKPYGTLYIGVTSRFEARLVBREKQLKGFTQRHGLTRLWWETHADMTVALHREKRLKEWP-RQWKINLIERDNPRWDDLYD 96 0056 2 OPAYYTLASQRMGTLYIGVTSRFEARLVBERKQLKGFTQRHGLTRLWWEBATREKCLKWP-RQWKINLIERDNPRWDDLYD 96 14 CWSYTURASONGTLYIGVTSDLVGRYQBRCRIGARFTSRYNVINEWWYELHPTMESATREKCLARGENLERDNPRWDDLYD 95 2016 14 CWSYTURASONGALYCGISDDLRRFDTGSGRGARFTSRYNVINEWWYELHPTMESATREKCLARGENLERDNPRWDDLYD 96 15 PWTXLASQRMGALYCGISDDLRRFDTGSGRGARFTSRYNVRLIFYSAPVGDRSILAREREKLAGGAGFAELLSS 91 9705 13 FYCYLLQSTIKRORLYTGITTDVERRCPMGKGAKYTKVRRPVELIFASAPVGDRSILAREREKLAGGFATLHH 112 870 13 FYCYTLQSTRHPAGGSTWRACOMGCAXFTKDGGAKTFKRCPVELIFASAPVGBRSILAREREKLAGGFATLHH 112 9710 13 FYCYTLQSTRHPAGGSTWRACOMGCAXFTKDGGAKTFKRR	3425248 15 WIAVYMMADHYRGTLYA GVTAQFITRTD REGROSAFTRYGLTRLWYEEHMIATEN LORETS KRWP-RQWKINLIERDNPRWDDYN 90 3422692 7 SIAVYIMTWRYYGTLYI GVTBOLKGFTORHGLITRLWYEEHMIATEN FREKKEWP-RQWKINLIERDNPRWDDYN 90 2 OPAVYITRABNGTLYI GVTBOLKRFDYI GSKGARFLIGGFTORHGVTMPWYEEHPTHADMTVA IRREKKEWP-RQWKINLIERDNPRWDDYN 90 25016 14 CWSYTLYRABNGTLYI GVTBOLKRFDYI GSKGARFLIGGFTORHGVTMPWYELHPTHESATTREKCEKWP-RQWKINLIERDNPRWDDYN 90 257 13 FYCYTLABNGALYCG I SDDELREPDYI GSKGARFLIGGG Y-THYTARAS-FYRAKOLTRRAKERLUVGVPULREVGE 10 3 PWFXYURABNGALYCG I SDDELREPDYI GSKGARFLIGGG Y PRINCHTALS-KRAKERLUVGEVPULREVGE 10 35015 13 FYCCYLLQSINKR	3471943 2	TGYVYMTASQF	CRGT	UNTUDITIO	LGRRMPE	HKSGEGSRFTS	5XC	-VQRLWW	(EEYFDITD	ALQREKSI	LKRRL-R	QWKIDLIE	KANPDWFEL	FR 9
 3422692 7 SIAVYIMTNKPYGTLYIGVTSRFEAR_UVB REKOLKGFTQRHGLTRLWWYETHADMTVATHREKREKEMP-RQWKINLIERDNPRWDDLYD 96 3 2 QPAVYILASQRNGTLYIGVTSBLVQELYCERDLYGERFLIERFYMSEATREKOLKFWN-RAWRLQLIEENNYRWDDLYD 95 2 2 PAVYILASQRNGALYYGTTPDVERKPDHSEATREFTERD 3 PWFLYLIRTADGALYYGTTPDVERKPDHSEATREFTERD 3 PWFLYLIRTADGALYYGTTPDVERKPOHSEALGGGARFHSPP 3 PWFLYLIRTADGALYYGTTPDVERKPOHSEAGARERFINESATREFTERD 3 PWFLYLIRTADGALYZGTSDDLRFFTYDGGARFHSPP 3 PWFLYLIRTADGALYZGTSDDLRFFTYDGGARFHSPF 3 PWFLYLIRTAD	3422692 7 SIAVYIMTUKPYGTLYIGVERFEARUVBIREKOLKGFTORHGLTRLWWETHADMTVALHREKELKEWP-RQMKINLIERDNPRWDDYD 9 3422692 7 SIAVYIMTUKPYGTLYIGVERDLYOETYOERELLIEGFTSRYNLTRLWWELHPTMESALTREKOLKKWN-RMKLOLIEENVSWRDUWF 9 30016 14 GWSVYLVAERNGTLYIGVTSDLYOETYOERELLIEGFTSRYNLARLWYVERGASKGARKHWYELHPTMESALTRELS-KRAKERLJVGVPULREVGE 10 3 PWETXLIRTADGTLYIGVTSDLYOETYOERGAGARFFHSSPBELTUFSREADAOPEHARLEKALS-KRAKERLJVGVPULREVGE 11 3 PWETXLIRTAD	3425248 15 1	WIAVYMMADHN	(RGT	LYAGUTAQ	PITRITU	REGRGSAFTRI	RYG	-LTRLWW	TEEHMLMTE	ALQRETS	IKRWP-R	SQWKINLIE:	RDNPRWDDL	YA 10
 2 QPAVTLASQRNGTLYIGVTSDLVQRIYOHRHLIEGFTSRYNVTMPWYELHPTMESRITREKOLKKWN-RAWKLQLIEENNVSWRDLWF 2 2 QPAVTLASQRNGTLYIGVTSDLVQRIYOHRCHIEGFTSRYNVTMPWYELHPTMESRITREKOLKWN-RAWKLQLIEENNVSWRDLWF 3 2 WETXLIFRADNKLYTGITTDVERREDHESETARSE 4 2 WAYTLASKDFTSEFTVGSTPARYOGAKKTRKRGSRWEMIMITVGFPSKIAMOHSLALRABYKKUTT-RKKELVIEGKVFKKGGRTHH 3 2 FGYKLLYKLLGSINKRGSFTVGSTPARYOGAKKTRDVELLIFSESSTRAMOHSLARMOHSLYKKKGGRTHH 4 112 FYCYTLRSQPRHSLYIGSTPHPFRLSOH-GLAKGGAKKTSGRGPWEMVLVVHGFPSSVALLRFEMMOHSLAFERAEAVKGLGR 3 12 12 FYCYTLIRSQPRHASLYIGSTPHPFRLSOH-GLAKGGAKKTSSRGPWEMVLCVHGFPSSVALLRFEMMOHSLAFERAEAVKGLGR 3 12 2004 31 17 FYCYTLIRSQPRHASLYIGSTPHPFRLSOH-GERKGGAKKTSSRGPWEMVLCVHGFPSSVALLRFEMMOHSL-VSRLQFPRAFOL 3 17 5 FGYYLLYCLINPRYRGRVYGFTVNTRREMOMGGRKKGGAMRTSGRGPWEMVLCVHGFPSSVALLRFEMMOHSL-VSRLQFPRAFOL 3 17 5 FGYYLLYCLINPRYRGRVYGFTVNTRREMENDGAKKTSSRGPWEMVLCVHGFPSSVALLRFEMMOHSL-VSRLQFPARYOL 3 17 5 FGYYLLYCLINPRYRGRVYGFTVNTRREMERDMOHSLOHNYCVHGFPSSVALLRFEMMOHSL-VSRLAFFRAERVKGGAKKTGANTKRKRKREMOHSLOHNYCVHGFPSSVALLRFEMMOHSL	0056 2 QPAYTLIASQRNGTLYTGVTSDLVQBTYOERHLIEGFTSRYNVTWPWYELHPTMESATTREKQLAFWRMACAMARCSRNALS 9 3 PWFXLTASQRNGTLYTGVTSDLVQBTYOERFDFGSGKGARKERSP ARALYYYEACASKGDARRERATKALS-KRAKERLUGGVPVLREVGE 10 3 PWFXLTRTADGALYCGISDDPLREPTOGOSGKGARKLEGSRPWEMIMILYKGFPSKLALS-KRAKERLUGGVPVLREVGE 11 257 3 PWFXLTRTADGALYCGISDDPLREPTOGOSGKGARKLEGSRPWEMIMILYKGFPSKLALS-KRAKERLVGVPVLREVGE 11 27 13 FYCCYLLQSINKRGSFYVGSTRWVRENCGAGKYTKKRSPVELLFASSFSKREAMQGFY-QFFHKGGRTHH 11 3700 16 LYTYTLRSQPHSIXTIGSTRPPRELSCING-CAKKGGAKKTSKLSLRWMWVCUVGFPSKLALRAGFYRKLTRAKKAGRTHH 11 35913 16 LYTYTLRSQPHHSIXTIGSTRPPRELSCING-CAKKGGAKKTSKLSLRWMWVCUVHGFPSKLALRAGFYRKLARGF 11 35913 16 LYTYXILRSQPHHSIXTIGSTRPPRELSCING-CAKKGGAKKTSKLSLRWMWVCUVHGFPSKLALREFMALLERAKKKERLYU 91 35913 16 LYTYXILRSQPHHSIXTIGSTRPPRELSCING-CAKKGGAKKTSKLSLRWMVCUVHGFPSKLREMMOHP 851 27 3129004 172 FFGYTLIYSLRFKRAKKRELYU 91 91 91 3129004 172 FFGYTLIYSLRFKRAKKRELYU 91 91 91 91 91 91 91 91 91 91 <td>3422692 7</td> <td>SIAVYIMTNKE</td> <td>LDXc</td> <td>LYIGUTSR</td> <td>LFEARLVE</td> <td>REKOLKGFTQI</td> <td>RHG</td> <td>-LTRLWW</td> <td>(ETHADMTV)</td> <td>ATHREKR</td> <td>LKEWP-R</td> <td>SQWKINLIE</td> <td>RDNPRWDDL</td> <td>YD 9</td>	3422692 7	SIAVYIMTNKE	LDXc	LYIGUTSR	LFEARLVE	REKOLKGFTQI	RHG	-LTRLWW	(ETHADMTV)	ATHREKR	LKEWP-R	SQWKINLIE	RDNPRWDDL	YD 9
 350016 14 CWYTLYRAENGALYCGISDDPLRFDTGCSGKGARFFHSSPARALYYEACASKGDMLRRERAIKALS-KRAKERLLVGVPVLREVGE 101 257 3 PWFATLIFTADURLYCGISDDPLRFYOEMSKLOGGASKTRKE 2705 3 PWFATLIFTADUSFYVGSTPWPVRRLOGGASKTRKE 2705 3 PWFATLICSINKRGSFYVGSTPWPVRRLOGGASKTKKDGSRPWEMIMIVKGFPSKLALABEYKVCUT-KKOERLVKKGGGGSFYVGGGASKTKH 2705 2870 2016 2013 2014 2015 2014 2014 2015 2014 2014 2014 2015 2014 2014 2015 2014 2014 2015 2014 2015 2014 2015 2014 2016 2014 2014 2015 2014 2014 2014 2014 2014 2014 2014 2014 2014 2015 2014 2014 2014 2014 2015 2014 2014 2014 2014 2014 2014 2014 2014 2015 2015 2014 2014 2014 2014 2015 2014 2015 2015 2014 2014 2015 2014 2014 2015 2014 2015 2014 2014 2014 2014 2014 2014 2014 2015	350016 14 CWYTURAENGALYCGISDPLREPDICSCKGARFHSSPARALYYPEACASKGDALRRERAIKALS-KRAKELLVGVPVLREVGE 10 257 3 PWFYTUIRTADGALYCGISDPLREPTOGACATTRYCUT	0056 2	QPAVYILASQF	TDNS	LYIGUTSD.	LVQRIYQ	REHLIEGFTSI	RYN	KWNAMTV-	(ELHPTMES.	ALTREKO	LKKWN-R	NAWKLQLIE	ENNUSWRDL	WF 9
 257 3 PWFLYLIRTADNKLYTGITTDVERPYQQBQSKGAKALRGKGELTLVFSAPVGDRSLALRAPKVKQLT-KRQKERLVAEGAGFAELLSS 91 9705 13 FYCCYLLQSINKRQSFYVGSTPNPVRELRCBMGKLAVGGAYRTKRDGSRPMEMIMIVRGFPSKIANLOFEHAMQHGY-QTHYLAEKDRVVKHKAGGRTLHH 91 92 93 94694 92 93 94 92 92	257 3 PWFINITRADNKLYTGITTDVERBYQCBGSKALRGKGELTLVFSAPVGDRSIALABEYRVKQLT-KRQKERLVAEGAGFAELLSS 9 3 PYCCELLGSINKRQSFYVGSTPWPVRELRCBMCKLAVGGAYRTKRDGSRPMEMIMIVRGFPSKIAALQFEHAMQHGY-QTHYIAEKDRVVKHKAGGRTHH 11 870 5 NHFFYLVKCKDOSFYVGSTPWPVRELRCBMCKLAVGGAYRTKRDGSRPMEMIMIVRGFPSKIAALQFEHAMQHGY-QTHYIAEKDRVVKHKAGGRTHH 11 870 5 NHFFYLVKCKDOSFYVGSTPWDFHRQNDGKGAXFTKRDGSRPMEMIMIVRGFPSKIAALQFEHAMQHGY-QTHYIAEKDRVVKHKAGGRTHH 11 870 5 NHFFYLVKCKDOSFYVGSTPWDFHRQNDGKGAXFTSKDGGAYRTSKIDGSRPMEMIMIVRGFPSSYIAALQFEHAMQHGY-QTHYIAEKDRVVKHKAGGRTHH 11 870 5 NHFFYLVKCKDSSFYVGSTPWDFMCOGMGGAXRTSCAMRTSGROMEMVLVVHGFPSYVALKFBMALMPFKALLPFRKAGLRFRAEAVRGGGR 11 870 15 FYGVELISSPRPCYKRERAGGASTFRAGGANRTSGROMEMVLVVHGFPSYVALKFBMAMQHPH-SSRLAHVGPREAVRGGGA 12 8713 17 FYGVELISSPRPCYKREY 9500563MRTSGROMEMVLVVHGFPSYVALKFBMAQDHP-SSRLAHVGPR	350016 14	CWSVYLVRAEN	VGA	TACGISDD	PLRRFDT	CSGKGARFFH	SSP	-ARALVYV	/EACASKGD	ALRRERA	IKALS-K	RAKERLLV	3VPVLREV	GE 10
 9705 13 FYCCYLLQSINKRQSFYVGSTPNPVRLKOHMGKLAVGGAYRTKRDGSRPMEMIMIVRGFPSKIAALOFBHAMQHGY-QTHYIAEKDRVVKHKAGGRTLHH 9705 13 FYCCYLLQSINKROSFYVGSTPNDLHKRVKIHMDGKGAKYTKVRRPVELIFAESFSTKREAMQAEYYFKLT-RKKKELYIEEKRNSKEAYY 93 59 423 16 LYTYVTIRSOPHHASINIGSTPHPPREASOHM-GLASCARTSTSIRPMANDVCLVSGFPSMIAAKFBMAQAEYYFKLT-RKKKELYIEEKRNSKEAYY 93 59 423 15 LYTYVTIRSOPHHASINIGSTPHPPREASOHM-GLASCARTSTSIRPMANDVCLVSGFPSMIAAKFBMAQAEYYFKLT-RKKKELYIEEKRNSKEAYY 93 59 423 15 LYTYVTIRSOPHHASINIGSTPHPPREASOHM-GLASCARTSTSIRPMANDVCLVSGFPSMIAAKFBMAQAPH-SKSILAHVGFRBANGGGRK 94 59 13 FFGYYLLYCLNPRYRGRYYGFTWYBRINGMGGRKKGGAKKTDSRGPMBNVCVHGFPNHVAALHFPMAQDNPL-VSKSLKEKQFRAFAF 94 59 13 FFGYYLLTSLSPRHKGRTYIGSTPDPFRLINGM-GETNGGRKKTDSRGPMDNVCVHGFPNHVAALHFPMAQDNPL-VSKSLKEKGFRFAFAF 94 59 13 172 FYGYYLLTSLSPRHKGRTRAFGAGAKKTCNLRLVYHSASAYDNVCVHGFPNHVAALHFPMAQDNPL-VSSLKEKQFRRAFGFAFAF 94 59 13 172 FYGYYLLTSLSPRHKGRTRAFGAGAKKTCNLRLVYHSASAYDNVTAAAMSPHPRASSAAA-FKSFSGVAIQV 94 59 27 FFGXTLTSLSPRHKGRTRAFRERRENGM-GETVGGASKTKHGRPMSISCLVYGFPNKVSAAKFBMNWONLG-ISRYTRCDFR-SKKQYTTMYCL 92 2 SSCYYWRRPDKKLYTGGTFDLEGRTRAFERRENG-CELVGGASKTKFLLVHVYSAAKFBMNWONLG-ISRYTRCDFR-SKKQYTTMYCL 92 2 SSCYYWRRPDKKLYTGQTFDLEGRTRAFERRENGESSFLLVHNQGKSMACOLETLINGL- 92 2 SSCYYWRRPDKKLYTGQTFDLEGRTRAFERRENGESSFLLVHWQGKSMACOLETLINGL- 92 38CYYWRRPDKKLYTGQFFLRGENGSSFLLVHWQGKSMACOLETLINGL- 92 38CYYWRRPDKRLYTGQFFLRGENGSSFLLVHWQGKSMACOLETLINGL- 94 694 91 922 922 9100000000000000000000000000	 PYCCYLLQSINKRQSFYVGSTPNPVRBLRC, MGKLAVGGAYRTKRDGSRPMMI MIVRGFPSKIAALQFBHAQHGY-CTHYI AEKDRVVKHKAGGRTLHH NHFFYVKTKDOSFYVGSTPNPPRASQEN-GGAKYTKRDGSRPMMI MIVRGFPSKIAALQFBHAQHGY-CTHYI AEKDRVVKHKAGGRTLHH NHFFYVKTKDNSIVAGYTND-HKRVTBMOGKGAKYTVRK SNHFFYVKTGISIVAGYTND-HKRVTBMOGKGAKYTVRK E LYTYTLRSOPRHSIVIGSTPPPRASCEN-GGAKYTYRKI SLAPWINNCLVSGFPSKIAAMQABYYFKKIKKKRELYI EEKRINSKEAVY E LYTYTLRSOPRHSIVIGSTPPPRASGEN-GAKGGANRTSPUBLIFBESFSYKREBAMQABYYFKKIKKKRELYI EEKRINSKEAVY E LYTYTLRSOPRH-VCKDSIVIGSTPPRASCEN-GRAGANRTSGRQWEMVLVVHGFPSSVALKPBMALNPH-KSKLHIPAEKRAEAVRGCR I Z FYGVCLISSDRPCYKNEYUGFTWURRENOGBMGGANKTGSRGMEMVLVVHGFPSSVALKPBMAQNPI-VSKSLERKOLRKERETPFAFOL I Z FYGVCLISSDRPCYKNEYUGSTPPPRASCEN-GEITSGAMRTKKKRPWEMVLCI YGFPTNVSALGPBMAMONPI-VSKSLERKOLRKERETPFAFOL I Z FYGVCLISSDRPCYKNEYUGSTPPPRARENOGBMGGAKKTTDSRGPMEMVLCI YGFPTNVSALGPBMAMONPI-VSKSLERKOLRKERETPFAFOL I Z FYGVCLISSDRPCYKNEYUGSTPPPRARENOGBMGGAKKTTDSRGPMEMVLCI YGFPTNVSALGPBMAMOND-SKSLEKKOLRKERETPFAFOL I Z FYGVCLISSDRPCYKNEYUGSTPPPRRLRQBN-GEITSGANRTKNILLIVYSASAYDYNTAALHFBMAMOND-SKSVFKERETPFAFOL I Z FYGVCLISSDRPCYKNEYCGISTRANDRRAGGAKKTDSRGANRTKKKRPWENVLCI YGFPTNVSALGPBMAMOND-SKSVFKERETPFAFOL I V WCVYLLRSDNKRLYGTSTBRVAGAKCIRNATNILLIVYSASAYDYNTAALHFFBMAMOND-SKVFKERETPFAFOL V J SSCVYMRRPDKRLYGGASKKTROGNETARA	257 3	PWFLYL IRTAL	NC	TTTOLLD	WERRYQO	HOSGRGAKALR	GKG	-ELTLUFS	SAPVGDRSL	ALRADYR	VKQLT-K	CROKERLVA	EGAGFAELL	SS 9
 870 5 NHFPYVKCKDNSWYAGYTNDLHKRVKTANDGKGAKYTKVRRPVELIFAESFSTKREAMQAEYYFKKLT-RKKKELYIEEKRNSKEAVYV 9359423 16 LYTVYILRSQPRHASLYIGSTPHPPRELSOIN-GLAKGGAKRTSFVELIFAESFSTKREAMQAEYYFKKLT-RKKKELYIEEKRNSKEAVYV 9359423 16 LYTVYILRSQPRHASLYIGSTPHPPRELSOIN-GLAKGGAKRTSGROPMENULVVHGPPSSVAALRFBAMQHPH-ASRLAHVGPREAVKGLGR 103 172 FYGYYCLISRSDRPCYKOFTWARRANOHNGGRKKGGAMRTSGROPMENULVVHGPPSSVAALRFBAMQHPH-ASRLAHVGPREAVKGLGR 103 172 FYGYYCLISRSDRPCYKORTRYRCHOOHNGGRKKGGAMRTSRKRPWENULCIYGFPNHVALHFPAMQOHD-VSKSLREKGLREKGFPFAFQL 273 277 FFACXLLTSLSPRHKGTTYIGFTUNPRRINOHNGGRKKGGAMRTKKKRPWENULCIYGFPNHVALHFPAMQOHD-VSKSLREKGLREKGFPFAFQL 273 277 FFACXLLTSLSPRHKGTTYIGFTUNPRRINOHNGGRKKGGAMRTKKKRPWENULCIYGFPNHVALHFPAMQOHD-VSKSLREKGLREKGFPFAFQL 273 277 FFACXLLTSLSPRHKGTTYIGFTUNPRRINOHN-GETTSGAMRTKKKRPWENULCIYGFPNHVALLHFPAMQOHD-SSKSURANGKREKGGAWRTV- 27106110 6 FYCCYLLKSURPCSSKGYTIGSTPDPPRELROHN-GETTSGAWRTKNLRLVYHSASAYDYNTAARMEYDLG-ISRYTKDCDFP-SKKGYTTMYCL 201100 922 SSCVYVMRRPDKKLYTIGGTPDLEGRIRAHRAKGIGSSFLYLMVQGKKNACGLETLINQL-HEGSYSLANLAKPLLLHKFLL 202 202<	870 5 NHFYYVKCKDNSWYAGYTNDLHKRVKTENDGKGAKYTKVRRPUELTFAESFSTKREAMOAEVYFKKUT-RKKKELYIEEKRNSKEAVYV 9 35943 16 LYTYYLIRSOPHASLYIGSTEPPRELSOEN-GAKGAXYTSKISLRPMNHVCLVSGFPSMIAALKFBMAAINNH-KSLHIRAEKRAEAVKGAGK 11 312904 13 FFGYYLLYCUPRVGRVYVGFTVWTARVOOINGGRKGGAMRTS-GRGAWRYCVVGFPSMIAALKFBMAMOHPI-SSKILHIRAEKRAEAVKGAGK 11 204363 17 FYGYYCLISRSDRPCYKNCYJGYTVDPNRELNOENGGRKGGAMRTS-SGGAWRYCVVGFPNHVJALKFBMAMOHPI-VSKSLKEKQLRKEKRETPFAPOL 21 2341198 27 FFGYLLISSDRPCYKNCYJGYTVDPNRELNOEN-GETYSGAMKTDSRGPWDMVCVVHGFPNHVJALKFBMAMOHPI-VSKSLKEKQLRKEKRETPFAPOL 21 234116110 6 FYCCYLLISSDRPCYKNCYJGFTWNRELNOEN-GETYSGAMRTKHGRPWSISCLVYGFPNHVSALKFBMAMOHDI-VSKSLKEKQLRKEKRETPFAPOL 21 2341165 2 SCVYYMRRPOSSGAVTIGFTWNRELNOEN-GETYGGAKKTNLELLYKSALKFBMAMOHDI-STRYTKDDFR-SKKQYTIMYCL 10 94694 92 5 SSCVYYMRRPDKRLYTGTTSBILNRELKBENERGEN-GETXGAGSSFLILHKESASYDYNTAAANEYVIKKCSKYFKLRLIKAKPLILHKFLL 10 946694 92 5 SSCVYYMRRPDKRLYTGTTSBILNRELKBENERGELOGSSFLILHVESASYDYNTAAANEYVIKKCSKYFKLRLIKAKPLILHKFLL 10 946694 92 5 SSCVYYMRRPDKRLYTGTTSBILNRELKBENERGELOGSSRLILWVGGKRMACOLETILINOL-HEQGYSLIKNGCKRLYTGTTSBILNRELK 10 00 08	9705 13	FYCCYLLQSIN	JKRQS	FYVGSTPN	IPVRRLRO	NGKLAVGGAYI	RTKRDGSR	PWEMIMIV	RGFPSKIA	ALQFEHAN	MQHGY-Q	THYIAEKD	RVVKHKAGGRTL	HH 11
359423 16 LYTVYILRSQPRHASLYIGSTPHPPRUSOHN-GLAKGGAKRTSKISLRPWNMVCLVSGFPSMIAALKFPWALNNPH-KSLHIPAEKRAEAVKGLGR 110 3129004 13 FFGVYLLYCIAPP-YRGRYVGFYVATARPVOGIMGGRKKGGAMRTSGRGPMEMVLVVHGFPNHVAALHFPWAWONPL-VSKSLKEKQLRKEKFPFAFQL 102 504363 172 FYGVYCLISRSDRPCYRNRRTGFYNPRREIROHN-GERKRGGAMRTSKRPPMEMVLCIYGFPNHVAALHFPWAWONPL-VSKSLKEKQLRKEKFPFAFQL 273 504363 172 FYGVYCLISRSDRPCYRNRRTGFYNPRREIROHN-GERYGGAMRTKKKRPWEMVLCIYGFPNHVAALHFPWAWONPL-VSKSLKEKQLRKEKFPFAFQL 273 504363 172 FYGVYCLISRSDRPCYRNRRTGFYNPRREIROHN-GERYGGAMRTKKKRPWEMVLCIYGFPNHVAALHFPWAWONPL-SKSGKATQV 124 504363 172 FYGVYCLISRSDRPCYRNRRTGFYNPRREIROHN-GERYGGAMRTKKKRPWEMVLCIYGFPNHVAALHFPWAWONPL-SKSGKATQV 124 507106110 6 FYCCYLLKSNRTG-SSGAVYIGSTPPPPFRLRQH-GEVGGASKTKHGRPWISCLVYGFPNKVSLQFPMAWONPLG-ISSYTKRDCDFR-SKKGKTIMYCL 103 502 10 VWCVYILRQDNGKLYTGIFFNLNRFIRQHSNRQGASKTKHGRPWISZCLVYGFPNKVSALGFPMAWONLG-ISSYTKDCDFR-SKKGKTIMYCL 100 504 527 52 55CVYWRRPDKRLYIGGYPDLEGGIRRAHRAKEGLQGSSFLYLWVQGKSMACQLPTLLINQL-HEGGYSLANDFGT 1003 504 505hVhhhCG.PSpRh.Hp.bhhp.bhhp.l.hhr.hlkFLL 100 504 505hvhhhflogregerrahrakeGLQGSSFLYLMVQGKSMACQLPTLLINQL-HEGGYSLANDFGT 100 504 505hvhhhvh.hhp.lb.hhr.hlkFLL	359423 16 LYTVYILRSQPRHASLYIGSTPHPPRLSOH-GLAKGGAYRTSKISLRPWNWCLVSGFPSMIAALKEBWALNPH-KSLHIPAEKRAEAVKGLGR 11 3129004 13 FFGVYLLYCLNPRYRGRVYGFTVWTARPVQOHNGRKKGGAWRTSGRGPWEMVLVVHGFPSSVAALREBWAWQHPH-ASRILAHVGPRLRGETAFAF 10 504363 172 FYGVYLISRSDRPCYKURCYIGYTVDPNRRIMOHNGGAKKTOSRGPWEMVLVVHGFPSSVAALREBWAWQHPH-ASRILAHVGPRLRGETAFAF 10 172 FYGVYLISRSDRPCYKURCYIGYTVDPNRRIMOHNGGAKKTOSRGPWEMVLVVHGFPSSVAALREBWAWQHPH-ASRILAHVGPRLRGETAFAF 10 2043163 172 FYGVYLISRSPR-FKGOTMGYTGANTGAKKTOSRGPWEMVLVVHGFPSSVAALREBWAWQHPH-SSRILAHVGPRLRGETAFAF 10 204610 6 FYCCYLLKSNRTCYIGYTUGATUGM-GEITSGAMRTKKKRPWEMVLCIYGFPNHVAALREBWAWQHPH-SSRILAHVGPRFLILHKFLL 10 94694 922 SSCVYNRRPDGKLYTGGTSNLNRTKOHSNRGAKKTONLLLVYHSASAYDYNTRARMSVNLKRKCSKYFKLRLIKAKDILLLHKFLL 10 946694 922 SSCVYWRRPDKLYTGGTSNLNRTKOHSNRGAGKCLRNATRLLLVYHSASAYDYNTRARMSVNLRRKCSKYFKLLINQL-HEGGYSILANLADGKHRNFGT 100 00 00 00 00 00 00 00 00 00 00 00 00	870 5 1	NHFFYVVKCKL	SNC	WYAGYTND	LHKEVKT	HNDGKGAKYTKI	VRR	-PVELIFA	LESFSTKRE	AMQABYYY	FKKLT-R	KKKELYIE	EKRNSKEAV	9 VY
3129004 13 FFGVTLIYCLNPRYRGRVYGFTVNTARRVQOBNGGRKKGGAMRTSGRGPWEMVLVVHGFPSSVALRFEMAWQHPH-ASRRLAHVGPRLRGBTAFAF 109 504363 172 FYGVTCLISRSDRPCYKINCYIGYTVDPNREIMOBMGGRKKGGAMRTKSRGPWEMVLCVHGFPSNHVAALHFEMAWQHPH-SSRLKEKGLKERKETPAFQL 273 2347198 27 FFACXLLTSLSPRHKGGTYIGFTVNPRREIMOBM-GETTSGAMRTKKRPWEMVLC1YGFPNHVAALHFEMAWQHPF-SKKKGTMYCU 124 2106110 6 FYCCXLLKSDRPCSSGAVTIGFTVNPRREIMOBM-GETYGGASKTKHGRPWISCLYGFPNHVSALGFEMAWQHPG-SKKKTTMYCL 100 904694 922 SSCVYWRRPDKRLYTGGTEDALGGATRABRAKGISNRQFNLHUVYHSSANDYNTSAAAMBYNLKRKCSKYFTALRLIKAKPGLLHHFLL 100 94694 9.2 SSCVYWRRPDKRLYTGQTDDLEGRIRABRAKGISSSFLHUVTHSSANDYNTAAAMBYNLKRKCSKYFTALLAFADGKHRWFGT 100 94694hYhY.G.PSPRh.HP.Dh.HP.Dh.H.LVLMOGKSMACQLFTLINQL-HEGCYSLANLADGKHRWFGT 100 93604694hYhY.G.PSPRh.HP.Dh.HP.Dh.H.LVLMOGKSMACQLFTLINQL-HEGCYSLANLADGKHRWFGT 100 94694hYhY.G.PSPRh.HP.DhP.Dh.HR.VLMOGKSMACQLFTLINQL-HEGCYSLANLADGKHRWFGT 100 94694hYhY.G.PSPRh.HP.D	3129004 13 FFGYTLLYCLNPR-YRGRVYGFTWTTRRVQOINGGRKKGGAMRTSGRGPWEMVLVVHGFPSYALLRPMAWQHPH-ASRILAHVGPRLRGETATAF 10 504363 172 FYGYYCLISRSDRPCYKNRCYIGYTVDPNRTMOINGGRDKGGAKKTDSRGPWDMVCVVHGFPNHVALLHPMAWQNPL-VSKSLKEKQLRKERFFFAFQL 27 2341198 27 FFACXLITSLSPRHKGOTYIGFTWDPRRINCIA-GETYSGAMRTKKKRPWEMVLCIYGFPNHVSALLFPMAWQNPL-VSKSLKEKQLRKERFFFAFQL 27 2365782 10 VWCYTLRSDRPC-SGGAYTIGFTWDPRRINCIA-GETYGGASKTTKHGRPWSISCLYYGFPNKVSALLFPMAWQNLG-ISRYTKDCDFR-SKKKTTIMYCL 10 PV_9627822 10 VWCYTLRQDNGKLYTGITSNLNRRIKOISNKGGAKCLRNATNLRLVYHSASAYDYNTTAAMQNLG-ISRYTKDCDFR-SKKKTTIMYCL 10 994694 922 SSCYYWRRPDKRLYIGGTBDLIGGRRAB,RAKEGLQGSSFLYLMVGGKSMAGQLIGTLLNUL-HEGGYSLIALLADGKHRNFGL 00 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	359423 16	LYTVYILRSQE	PRHAS	HATSDIVL	IPPRRLSQ	HN-GLAKGGAY	RTSKISLR	PWINMUCLV	SGFPSMIA	ALKFEWAI	I-HUNNH-K	SLHIPAEK	RAEAVKGL	GR 11
504363 172 FYGVYCLISRSDRPCYKNRCYIGYTVDPNRRIMOHMGENGKGAKKTDSRGPWDMVCVHGFPNHVALHFPMAWONPL-VSKSLKEKQLRKERFFAFACU 273 2347198 27 FFACYLLTSLSPR-HKGQTYIGFTVNPRRIROHM-GEITSGAMRTKKKRPWEMVLCIYGFPTNVSRLQFPMAWOHPR-ESVAVREAAA-FKSFSGVALQV 124 p_7106110 6 FYCCYLLKSNRTQSSGAVYIGSTPDPPRLROHM-GEITSGAKTKHGRPWSISCLVYGFPNKVSRLKFPMNWONLG-ISRYTKDCDFR-SKKQKTIMYCL 103 p_7105110 1 FYCCYLLKSNRTQSSGAVYIGSTPDPPRLROHM-GEITSGAKTKHGRPWSISCLVYGFPNKVSRLKFPMNWONLG-ISRYTKDCDFR-SKKQFTIMYCL 103 PV_9627822 10 VWCYYTLRQDNGKLYTGITSNLNRRIKOHSNLRRAHTANLRLVYHSASAYDYNTRARMEVNLKRKCSKYFKLRLIKAKPLLLHKFLL 100 994694 922 SSCYYWRRPDKKLYIGQTDDLEGRIRAHRAKEGLQGSSFLLVYMVGGKSMAGOLETLLINQL-HEGGYSLANLADGKHRNFGT 1003 03 *********************************	504363 172 FYGVYCLISRSDRPCYKNRCYIGYTUDPNRAIMO, MGGRDKGGAKKTDSRGPWDW/CVVHGFPNHVALHERMAWONPL-VSKSLKEKQLRKERFFAFQL 27 2347198 27 FFACXLLTSLSPRHKGQTYIGFTWNPRRAIRO, M-GEITSGAMRTKKKRPWEMVLCIYGFPTNVSALQFFWAWOHPR-ESVAVREAAA-FKSFSGVALQV 12, p_7166110 6 FYCCXLLKSNRPQSSGATYIGSTUDPPRRAICO, M-GEITSGAMRTKHGRPWSISCLYYGFPTNVSALQFFWAWOHA-ISKTRKCDFR-SKKGYTIMYCL 10, PV_9627822 10 VWCYTIRQDNSSGATYIGSTUDPPRRAICO, M-GEITSGAKTKHGRPWSISCLYYGFPTNVSALQFFWAWOHA-ISKTRKCSKYFKLRLLHKGL 10, PV_9627822 10 VWCYTIRQDNGKLYTGITSNLNRRIKOGN-GEITSGAKCLRNATFLYILWYHSASYDYNTAANONUA-ISKTRKCSKYFKLRLHKKL 10, PV_964694 922 SSCYYWRRPDKRLYIGQTDDLGGAIRAIRARAEGLGGSSSFLYILWVGGKSMACQLETINQL-HEQGYSLANLADGKHRNFGI 03 03 04 00 00 00 00 00 00 00 00 00 00 00 00 0	3129004 13	FFGVYLLYCLA	VPRYRGR	NALSOAR	TTAREVOO	INGGRKKGGAW	RTSGRG	PWEMVLVV	HGFPSSVA	ALRFBWAV	WQHPH-A	SRRLAHVG	PRLRGETAF	AF 10
2347198 27 FFACYLLITSLSPRHKGQTYIGFTWDFRRIRQHN-GEITSGAMRTKKKRPWEMVLCIYGFPTNVSALQFEMAWQHPR-ESVAVREAAAA-FKSFSGVALQV 124 p_7106110 6 FYCCYLLKSNRTQSSGAVYIGSTEDPPRRLRQHN-GEIVGGASKTKHGRPWSISCLVYGFPNKVSRLKFEMNWQNLG-ISRYTKDCDFR-SKKQKTIMYCL 103 pv_9627822 10 VWCYYILRQDNGKLYTGITSMLNRRIKQHSNKRGAKCLRNATNLRLVYHSASAYDYNTRARMEVNLKRKCSKYFKLRLIKAKPLLLHKFLL 100 994694 922 SSCYYVMRRPDKKLYIGQTDDLEGRIRAHRAKEGLQGSSFLYLWYGGKSMAGQLETLINQL-HEQGYSLANLADGKHNRFGT 1003 03 ···hYhY.G.pspRh.Hp.bh	2347198 27 FFACKLLFSLSPRHKGQTVIGFTWDFRRAIRQIN-GEITSGAMRTKKKRPWEMVLCIYGFPTNVShLQPEWAWQHPR-ESVAVREAAA-FKSFSGVAIQV 12 D_7106110 6 FYCCKLLKSNRTQSSGAVYIGSTPDPPRALRQIN-GEIVGGASKTKHGRPWSISELVYGFPKVShLKPEMAWQNLG-ISRYTKDCDFR-SKKQKTINYCL 10 PV_9627822 10 VWCYTLRQDNGKLYTGITSWNRRIKQIENKGGAKCLRNATNLKLVYHSASAYDYNTAARMENVLKRKCSKYFKLRLIKAKPLLLHKFLL 10 994694 922 SSCVYVMRRPDKRLYTGQTDDLEGRIRALRAKEGLQGSSFLYLLVYGGKSMACULETLINQL-HEQGYSLANLADGKHRNFGT 100 03hYhY.G.pspRh.Hp.bhhAhp.E.hHp.U.NhhhAhp.E.h.HEVGYSLANLADGKHRNFGT 100 10 besquence alignment of predicted nucleases associated with the hypothetical FP-ADL-Ku regin system. The novel potential nuclease (A) and The URI-domain-containin potential nuclease (A) and The URI-domain-containin containin potential nuclease (A) and The URI-domain-containin containin potential nuclease (A) and The URI-domain-containin potential nuclease (A) and The URI-domain-containin containin potential nuclease (A) and The URI-domain-containin potential nuclease (A) and The URI-domain-containin potential nuclease (A) and The URI-domain-containin potential nuclease (A) and The ORI-domain-containin potential nuclease (A) and The ORI-domain-containin-containin-containin potential nuclease (A) and The ORI-domain-contain	504363 172	FYGVYCLISRS	SDRPCYKNR	CVICYICYTVD	PNRRIMO	HNGGRDKGGAKI	KTDSRG.	PWDMVCVV	HGFPNHVA	ALHFEWAV	WONPL-V	'SKSLKEKQ	LRKERKETPFAF	QL 27
<pre>p_7106110 6 FYCCYLLKSNRTQ-SSGAVYIGSTPDPFRLKOHN-GEIVGGASKTKHGRPWSISCLVYGFPNKVSRLKFBMNWONLG-ISRYTKDCDFR-SKKQKTIMYCL 103 IPV_9627822 10 VWCYYILRQDNGKLYTGITSMLNRRIKOHSNKAGAKCLRNATNLRLVYHSASAYDYNTRARMBYNLKRKCSKYFKLRLIKAKPLLLHKFLL 100 994694 922 SSCYYVMRRPDKRLYIGQTDDLEGRIRABRAKEGLQGSSSFLYLWYGGKSMAGQLETLINQL-HEQGYSLANLADGKHRNFGT 1003 03</pre>	pp_7106110 6 FYCCYLLKSNRTQSSGAVYIGSTPDPPRALRQEN-GEIVGGASKTKHGRPWSISCLVYGFPNKVSALKPEWNØONLG-ISRYTKDCDFR-SKKQKTIMYCL 10 PV_9627822 10 VWCYYTRQNNGRLYTGITSWINRELKQESKKGRAKGLKNATNIKLVYHSASAYDYNTAARMEYNIKRKSKYFKLHLIKAKPILILHKFLL 101 994694 922 SSCVYMRRPDKRLYTGTTSWINRELKGEGGSSSFLYILMVQGKSMACOLETLINGL-HEQGYSLANLADGKHRNFGT 100 994694 922 SSCVYMRRPDKRLYTGTTSWINRELKGELQGSSSFLYILMVQGKSMACOLETLINGL-HEQGYSLANLADGKHRNFGT 100 08 hYh	2347198 27	FFACYLLTSLS	SPRHKGQ	NUTEFUN	IPRREIR	HN-GEITSGAM	RTKKKR	PWEMVLCI	YGFPTNVS.	ALQFEWAN	WQHPR-E	SVAVREAA	AA-FKSFSGVAI	QV 12
PV_9627822 10 VWCVYILRQDNGKLYTGITSMLNRRIKOHSNKRAKGGAKCLRNATNLRLVYHSASAYDYNTRARMEVNIKRKCSKYFKLRLIKAKPLLLHKFLL 100 994694 922 SSCVYVMRRPDKRLYIGQTDDLEGRIRAREALGSSFLYLMVQGKSMAGQLETLINQL-HEQGYSLANLADGKHRNFGT 1003 08	IPV_9627822 10 VWCVYILIRQDNGKLYTGITSNLMRAIKQESNKCLRNATNLRLVYHSASAYDYNTAARMEVNEKKCSKYFKLRLIKAKPLLLHKFLL 10 994694 922 SSCVYVMRRPDKRLYIGQTDDLEGAIRAMARAKEGLQGSSFLYLMVQGKSMACQLATLINQL-HEQGYSLANLADGKHRNFGT 100 08 hYh	p_7106110 6	FYCCYLLKSNF	RTQSSGA	UTIGSTPD	PPRRLRQ	HN-GEIVGGASH	KTKHGR.	PWSISCLV	'YGFPNKVS	ALKFBWN	NONLG-I	SRYTKDCD	FR-SKKQKTIMY	CL 10
994694 922 SSC YYYM RRPDKRL YIG QTDDLEGGIRAERAKEGLQGSSFLYLMVQGKSMACQLGETLINDL-HEQGYSLANLADGKHRNFGT 1003 08	994694 922 SSCVYYVMRRPDKRLVIQOTDDLEGATRAERAKEGLQGSSFLYLMVQGKSMACQLATLINDL-HEQGYSLANLADGKHRNFGT 100: 108 108 iple sequence alignment of predicted nucleases associated with the hypothetical EP-ADDL-Ku repair system. The novel potential nuclease (A) and The URI-domain-containin	IPV_9627822 10	VWCVYILRODN	VGK	TTGITSN.	ILNRRIKQ	HSNKQGAKCLRI	NAT	NLRLVYHS	ASAYDYNT	AARMEYNI	LKRKCSK	YFKLRLIK	AKPLLLHKF	LL 10
0%	0%	994694 922	SSCVYVMRRPL	0KR	LYIGOTDD	LEGRERA	RAKEGLQGSS		-FLYLMVC	MSXDG	ACQLETLI	TINGL-H	IEQGYSLAN	LADGKHRNF	GT 100
	ple sequence alignment of predicted nucleases associated with the hypothetical EP-ADDL-Ku repair system. The novel potential nuclease (A) and The URI-domain-containin- 	08	hYh		.Y.G.psp	·Rh	Нрb	••••••	h		Ahp.E1	hh		dh	·

4



Prokaryotic Ku Homologs









1370 Genome Research www.genome.org

mediately leads to the second helix that is separated from the parallel third helix by a prominent extended segment (Fig. 4B). We designated this fold the helix–extended-region–helix (HEH) domain after this unique structural pattern. The fusion of the HEH to the Ku-core and Endonuclease VII is suggestive of DNA-binding, whereas the presence of this domain in Rho and the KTRS is more consistent with RNA-binding. In KTRS, the HEH domain undergoes movement on lysine-binding and might facilitate recognition of specific structural features of tRNA^{Lys} (Onesti et al. 2000).

Additional extensive PSI-BLAST searches with the HEH domain sequences detected, with a moderately significant Evalue of 0.09, a previously identified nucleic-acid-binding domain, the SAP domain (Aravind and Koonin 2000). Reciprocal searches with the SAP domain sequences also retrieved from the database some of the HEH domain sequences with significant E-values (e.g., a stand-alone HEH-domain protein from Listeria phage A118 was recovered in a search with the Arabidopsis AP-endonuclease SAP domain in iteration 8 with $E = 4 \times 10^{-4}$), suggesting a potential evolutionary relationship. SAP is a small domain of approximately the same size as the HEH domain, and its core is strongly predicted to contain two helices separated by a relatively long extended region, similar to helix-2 and helix-3 of the HEH domain (Aravind and Koonin 2000). The amino acid residue conservation pattern in these helices and the extended region between them is also similar in the HEH and SAP domains (Fig. 4A). The only noticeable difference in the sequence pattern between these two domains is the presence of an insert of two amino acids at the end of the extended region in the HEH domains. A homology model of the SAP domain from the Acinus protein (Sahara et al. 1999; Aravind and Koonin 2000) built using the HEH from E. coli Rho as the structural template showed that the absence of these two residues is unlikely to disrupt the extended region characteristic of this fold (Fig. 4B). Thus, it appears likely that SAP domain is a derived, eukaryote-specific version of the HEH fold, which interacts with DNA via the charged surfaces of the helices.

Notably, the SAP domain is present in the eukaryotic Ku70 proteins as a conserved carboxy-terminal extension (Aravind and Koonin 2000). Based on the functions of the characterized SAP domains, it has been predicted that this domain binds the MARs and participates in tethering chromosomal proteins to these sites (Aravind and Koonin 2000); the SAP domain is probably responsible for MAR-binding by the Ku protein (Galande and Kohwi-Shigematsu 1999). It appears that HEH-fold domains, namely Rho-N and SAP, have been fused to the carboxyl termini of the Ku-core domain on two independent occasions, in bacteria and eukaryotes, re-

spectively. This may point to a specific cooperation between the HEH and Ku-core domains in binding unusual DNA structures including MARs and their analogs in bacteria.

The Amino-Terminal Region of the Eukaryotic Ku Proteins Contains a Divergent Von Willebrand Factor A Domain

The bacterial Ku homologs contain no counterpart of the conserved amino-terminal extension that is present in the eukaryotic Ku70 and Ku80. To determine the origin of this extension, we performed PSI-BLAST searches with these regions from the eukaryotic Ku proteins. At convergence, these searches retrieved from the database, in addition to the Ku proteins, several Von Willebrand factor A (vWA) domains from various organisms (e.g., the sequence of the Serum opacity factor from Streptococcus pyogenes was retrieved in iteration 4, $E = 10^{-2}$, and the YwmC protein from *B. subtilis* in iteration s, $E = 10^{-3}$, and chicken Collagen $\alpha 2$ in iteration 11, $E = 10^{-3}$). To further assess these observations, we constructed a multiple alignment of all amino-terminal extensions from diverse Ku70s and Ku80s and predicted the secondary structure of this domain using the PHD program (Rost and Sander 1993). The predicted structural elements exactly matched the pattern characteristic of the vWA domain for which experimentally determined structures are available (Lee et al. 1995; Leitinger and Hogg 2000). Furthermore, the two Mg²⁺-binding aspartates located at the ends of strands 1 and 4 of the vWA domains are typically conserved in the Ku proteins (Fig. 5) (Lee et al. 1995). Sequence-structure threading using the hybrid fold recognition method (Fischer 2000) with the human Ku70 protein as a query recovers the vWA domain of integrin (PDB:1ido) as the best hit. Thus, the amino-terminal extension of the eukaryotic Ku proteins appears to be a divergent version of the vWA domain.

The vWA domain, although originally discovered in several animal extracellular adhesion molecules (Lee et al. 1995), has been subsequently detected in intracellular contexts in both prokaryotes and eukaryotes (Ponting et al. 1999). In addition to Ku, at least one other protein with a function in DNA repair and transcription, the TFIIH subunit p44, contains a vWA domain (Ponting et al. 1999). In Ku70 and Ku80, the region encompassing the vWA domain is the second determinant of heterodimerization, which is consistent with the role of vWA in protein-protein interactions (Singleton et al. 1997; Wang et al. 1998). The conservation of the Mg⁺²binding aspartates in most sequences of the vWA domains from the Ku proteins (Fig. 5) suggests that they probably function as cation-dependent interaction modules similar to vWA domains in other contexts. Some of the numerous proteinprotein interactions demonstrated for the eukaryotic Ku proteins, in addition to heterodimerization, probably depend on the vWA domains of Ku70 and Ku80. To our knowledge the experiments to address the Mg+2 dependence of these interactions has not been performed.

Evolutionary Implications and Conclusions

The dissection of the Ku protein from eukaryotes and prokaryotes into individual domains described above suggests an evolutionary scenario for these proteins. The Ku core is an ancient domain that was probably present in bacteria and archea even before the advent of the eukaryotes. There are clear indications that, in these organisms, the Ku homologs are functionally associated with the ATP-dependent DNA ligase

Figure 4 (*A*) Multiple sequence alignment of different classes of HEH domains. Each of the alignments is colored according to a separate consensus using the rules described in the legend to Figure 2. The secondary structure shown above the alignment was derived from the structures of Rho, Endo-VII and K-TRS. For the SAP domains, the structure was predicted using the PHD program. The species abbreviations are the same as in Figure 2; those not present in Figure 2 are : Ec, *Escherichia coli;* Ssp, *Synechocystis sp.;* Tma, *Thermotoga maritima;* Dr, *Deinococcus radiodurans;* Aae, *Aquifex aeolicus;* BPL2, lactococcal Bacteriophage T2; BPA118, *Listeria* bacteriophage A118; T4, Bacteriophage T4; Miclu, *Micrococcus luteus;* Ce, *Caenorhabditis elegans;* Ct, *Chlamydia trachomatis;* Hp, *Helicobacter pylor;* Bst, *Bacillus stearothermophilus.* (*B*) Structures and models of different forms of the HEH domain shown in the alignment. The NH2 (N) and COOH (C) termini of the HEH domains are indicated.



Secondary structure: KU70_Hs_4503841 YKU70_Sc_6323940 CC126.02c_Sp_4008550 KU70_Dm_557036 Ku86_Hs_125731 BC543.03c_Sp_9929270 YKU80_Sc_6323753 consensus/100%	<pre>celese</pre>
SOF_Spy_4995872 YWMC_Bs_8928525 IntegrinA_Hs_6680482 Matrilin3_Hs_3647275	100 DVALLTVEKKNYCENFORAKEDIKKYVYTTT-EKTSGECHAMMERNEVRFITYVRVNEFILTALMVEAKLEEVMEQAKKMMAWJ-DDCGATHKAREITMAKE 36 MVAVLL ASGSAAKRIDGVSKPNSAKKEISKPASELPEGTVVNEVYGSEGUNNNAQSCEATRHVYGFQSPHEQSFINSLATIOPTGMTPIAKALMEAKSSPQL 195 ELAYTLGSGSIGFSDFOKANDFISTMENNYFECCE
Secondary structure: KU70_Hs_4503841 YKU70_Sc_6323940 CC126.02c_Sp_4008550 KU70_Dm_557036 Ku86_Hs_125731 BC543.03c_Sp_9929270 YKU80_Sc_6323753 consensus/1008	hhheEe.e
SOF_Spy_4995872 YWMC_Bs_8928525 IntegrinA_Hs_6680482 Matrilin3_Hs_3647275	KEKKSGK-RONIALFROGEST-FSYDILKEKKDNKN-NTV-KKNRIAEKVTTSNPLFP4FNH-THQKADMLDDVKQLIQLSEKLGIEGLS-GLNDKLRLAGVAS-GFAGALL 384\ DAKGEKVYVLDROGEETCGANFIKT-AKEL-QKONITUNVIGPDY-KEGYKAGQLAATAKVGGGEYFPAYTQKOV-EKIFTQGGLAMLSK

Figure 5 Multiple sequence alignment of the vWA domains of Ku70 and Ku80. The secondary structure shown *above* the figure was based on the solved structures of vWA domains; the same consensus-based coloring scheme as in Figures 2 and 3 is used. The species abbreviation Spy is for *Streptococcus pyogenes*, whereas the rest are the same as in Figures 2 and 3.

and the eukaryotic-type primase, probably as components of a double-strand break repair system. Because ADDL and EP are ubiquitous in archea, but are present only sporadically in bacteria, it seems plausible that this hypothetical repair system, including the Ku-core domain, has originally evolved within the archeal lineage and subsequently has been disseminated among bacteria through multiple horizontal transfers. However, the difficulty with this scheme is that a Ku homolog so far was identified in only one archeon, *A. fulgidus*, whereas the mobile Ku–EP–ADDL operon so far is widely represented only in bacteria. Hence, a bacterial origin for this mobile operon, with a subsequent transfer of the gene coding for Ku to *A. fulgidus* is also equally possible. In *Streptomyces*, there was a duplication of the Ku-core-encoding gene, and one of the paralogs fused with another ancient nucleic-acid-binding domain, the HEH (Fig. 6), whereas, in *M. loti*, the entire Ku–EP–ADDL operon was duplicated (Fig. 1).

Eukaryotes might have vertically inherited the Ku-core protein, along with the primase and the ATP-dependent ligase, from a common ancestor shared with a certain archeal lineage or through horizontal transfer from a bacterial lineage such as the mitochondrial precursor. Under this scenario, early in the evolution of eukaryotes, the Ku-core domain underwent an amino-terminal fusion with the vWA domain,



(Ssp, Scoe, Dr, Pa, BPL2, BPA118)

Figure 6 Protein domain architectures and possible evolutionary trajectories for the constituent domains of the Ku proteins. The domains are indicated by different shapes; the two distinct forms of the HEH are indicated by differential coloring. For each domain architecture, the phyletic distribution is shown in parentheses; the species name abbreviations are as in Figures 2 and 3. The arrows indicate probable evolutionary events such as derivation of a new form of a particular domain (SAP from ancestral HEH) and domain fusion. The connection shown between prokaryotic and eukaryotic forms of the Ku protein does not differentiate between two evolutionary scenarios discussed in the text.

Prokaryotic Ku Homologs

followed by a duplication giving rise to the paralogous Ku70 and Ku80 proteins (Fig. 6). Subsequently, but prior to the radiation of the major crown-group lineages, the Ku70 protein fused with the eukaryote-specific version of the HEH fold, the DNA-binding SAP domain (Fig. 6). Ku80 evolved its own unique distinct carboxy-terminal extension resulting in the acquisition of distinct functions by the eukaryotic Ku paralogs. These fusions conferred several new interactive abilities on Ku70 and Ku80 that allowed them to associate with various eukaryote-specific protein complexes involved in DNA repair, telomere formation, and chromatin remodeling. This scenario seems plausible because prokaryotic Ku homologs that contain the Ku-core domain alone seem to be the best candidates for the role of the primitive form of this protein. An alternative scenario would hold that the Ku-core domain evolved at an early stage of eukaryotic evolution and was horizontally acquired by a bacterium or an archeon, probably prior to the fusion with the vWA domain, followed by horizontal dissemination among the prokaryotes. Sequencing of additional archeal genomes and those of early-branching eukaryotes help in resolving these alternative hypotheses.

Regardless of the exact evolutionary scenario, the detection of Ku homologs in prokaryotes and dissection of the Ku protein into previously undetected, distinct domains will allow experimental exploration of simpler model systems to understand the essential functions of these important proteins.

METHODS

The archeal and bacterial genome sequences were retrieved from the Genomes division of the Entrez system (Tatusova et al. 1999). The nonredundant database of protein sequences at the National Center for Biotechnology Information (NIH, Bethesda) was iteratively searched using the PSI-BLAST program (Altschul et al. 1997). The cut-off of E < 0.01 was typically employed for inclusion of sequences in the positionspecific weight matrices. Nucleotide sequences of unfinished bacterial and archeal genomes translated in all six reading frames were searched using the TBLASTN program (Altschul et al. 1997). Multiple alignments of protein sequences were constructed using the ClustalW (Thompson et al. 1994) program and corrected on the basis of PSI-BLAST results. Protein secondary structure was predicted using the PHD program, with a multiple alignment submitted as the query (Rost and Sander 1993; Rost et al. 1997). Sequence-structure threading was performed using the hybrid fold recognition method that incorporates both structural and evolutionary information in sequence comparisons into a single algorithm (Fischer 2000). Homology modeling of protein structures was performed by using the SWISS-MODEL server (Guex and Peitsch 1997). The target was threaded through the template using the SWISS-PDBviewer software and the alignment with the template was manually adjusted to minimize the clashes of the protein backbones. The energy minimization was carried out using the GROMOS program that employs a Sippl-like force field (Guex and Peitsch 1997). The ribbon diagrams of the structures were generated using the MOLSCRIPT program (Kraulis 1991).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Allison, T.J., Wood, T.C., Briercheck, D.M., Rastinejad, F., Richardson, J.P., and Rule, G.S. 1998. Crystal structure of the RNA-binding domain from transcription termination factor rho [letter]. Nat. Struct. Biol. 5: 352-356.

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.
- Aravind, L. and Koonin, E.V. 1999. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* 27: 4658–70.
- 2000. SAP a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem. Sci.* 25: 112–114.
- Aravind, L., Walker, D.R., and Koonin, E.V. 1999. Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.* 27: 1223–1242.

Blier, P.R., Griffith, A.J., Craft, J., and Hardin, J.A. 1993. Binding of Ku protein to DNA. Measurement of affinity for ends and demonstration of binding to nicks. *J. Biol. Chem.* 268: 7594–7601.

- Bogden, C.E., Fass, D., Bergman, N., Nichols, M.D., and Berger, J.M. 1999. The structural basis for terminator recognition by the Rho transcription termination factor. *Mol. Cell* **3**: 487–493.
- Cary, R.B., Chen, F., Shen, Z., and Chen, D.J. 1998. A central region of Ku80 mediates interaction with Ku70 in vivo. *Nucleic Acids Res.* 26: 974–979.
- Critchlow, S.E. and Jackson, S.P. 1998. DNA end-joining: From yeast to man. *Trends Biochem. Sci.* 23: 394–398.
- Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23: 324–328.
- Doherty, A.J., Serpell, L.C., and Ponting, C.P. 1996. The helix-hairpin-helix DNA-binding motif: A structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res.* 24: 2488–2497.
- Featherstone, C. and Jackson, S.P. 1999. Ku, a DNA repair protein with multiple cellular functions? *Mutat. Res.* **434**: 3–15.
- Fischer, D. 2000. Hybrid fold recognition: Combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.* 119–130.
- Galande, S. and Kohwi-Shigematsu, T. 1999. Poly(ADP-ribose) polymerase and Ku autoantigen form a complex and synergistically bind to matrix attachment sequences. *J. Biol. Chem.* **274:** 20521–20528.
- Galy, V., Olivo-Marin, J.C., Scherthan, H., Doye, V., Rascalou, N., and Nehrbass, U. 2000. Nuclear pore complexes in the organization of silent telomeric chromatin. *Nature* **403**: 108–112.
- Gell, D. and Jackson, S.P. 1999. Mapping of protein–protein interactions within the DNA-dependent protein kinase complex. *Nucleic Acids Res.* **27:** 3494–3502.
- Gottlieb, T.M. and Jackson, S.P. 1993. The DNA-dependent protein kinase: Requirement for DNA ends and association with Ku antigen. *Cell* **72**: 131–142.
- Guex, N. and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18: 2714–2723.
- Holm, L. and Sander, C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26:** 316–319.
- Hsu, H.L., Gilley, D., Blackburn, E.H., and Chen, D.J. 1999. Ku is associated with the telomere in mammals. *Proc. Natl. Acad. Sci.* 96: 12454–12458.
- Hsu, H.L., Gilley, D., Galande, S.A., Hande, M.P., Allen, B., Kim, S.H., Li, G.C., Campisi, J., Kohwi-Shigematsu, T., and Chen, D.J. 2000. Ku acts in a unique way at the mammalian telomere to prevent end joining. *Genes & Dev.* 14: 2807–2812.
- Koike, M., Miyasaka, T., Mimori, T., and Shiomi, T. 1998. Subcellular localization and protein–protein interaction regions of Ku proteins. *Biochem. Biophys. Res. Commun.* 252: 679–685.
- Koonin, E.V., Wolf, Y.I., Kondrashov, A.S., and Aravind, L. 2000. Bacterial homologs of the small subunit of eukaryotic DNA primase. J. Mol. Microbiol. Biotechnol. 2: 509–512.
- Kraulis, P.J. 1991. Molscript. J. Appl. Cryst. 24: 946-950.
- Lee, J.O., Rieu, P., Arnaout, M.A., and Liddington, R. 1995. Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18). *Cell* 80: 631–638.
- Leitinger, B. and Hogy, N. 2000. From crystal clear ligand binding to designer I domains. *Nat. Struct. Biol.* 7: 614–616.
- Li, B. and Comai, L. 2000. Functional interaction between Ku and the Werner syndrome protein in DNA end processing. *J. Biol. Chem.* **275**: 28349–28352.
- 2001. Requirements for the nucleolytic processing of DNA ends by the Werner syndrome protein-ku70/80 complex. J. Biol. Chem. 276: 9896–9902.

Aravind and Koonin

Massari, M.E. and Murre, C. 2000. Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* 20: 429–440.

- Mishra, K. and Shore, D. 1999. Yeast Ku protein plays a direct role in telomeric silencing and counteracts inhibition by rif proteins. *Curr. Biol.* 9: 1123–1126.
- Mullen, J.R., Kaliraman, V., Ibrahim, S.S., and Brill, S.J. 2001. Requirement for three novel protein complexes in the absence of the Sgs1 DNA helicase in *Saccharomyces cerevisiae*. *Genetics* **157**: 103–118.
- Onesti, S., Desogus, G., Brevet, A., Chen, J., Plateau, P., Blanquet, S., and Brick, P. 2000. Structural studies of lysyl-tRNA synthetase: Conformational changes induced by substrate binding. *Biochemistry* **39**: 12853–12861.
- Osipovich, O., Durum, S.K., and Muegge, K. 1997. Defining the minimal domain of Ku80 for interaction with Ku70. *J. Biol. Chem.* **272**: 27259–27265.
- Osipovich, O., Duhe, R.J., Hasty, P., Durum, S.K., and Muegge, K. 1999. Defining functional domains of Ku80: DNA end binding and survival after radiation. *Biochem. Biophys. Res. Commun.* 261: 802–807.
- Ponting, C.P., Aravind, L., Schultz, J., Bork, P., and Koonin, E.V. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* 289: 729–745.
- Raaijmakers, H., Vix, O., Toro, I., Golz, S., Kemper, B., and Suck, D. 1999. X-ray structure of T4 endonuclease VII: A DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. *EMBO J.* 18: 1447–1458.
- Ramsden, D.A. and Gellert, M. 1998. Ku protein stimulates DNA end joining by mammalian DNA ligases: A direct role for Ku in repair of DNA double-strand breaks. *EMBO J.* **17**: 609–614.
- Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. **232**: 584–595
- structure at better than 70% accuracy. J. Mol. Biol. **232**: 584–599. Rost, B., Schneider, R., and Sander, C. 1997. Protein fold recognition by prediction-based threading. J. Mol. Biol. **270**: 471–480.
- Sahara, S., Aoto, M., Eguchi, Y., Imamoto, N., Yoneda, Y., and Tsujimoto, Y. 1999. Acinus is a caspase-3-activated protein required for apoptotic chromatin condensation. *Nature*

401: 168–173.

- Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci.* **95**: 5857–5864.
- Singleton, B.K., Priestley, A., Steingrimsdottir, H., Gell, D., Blunt, T., Jackson, S.P., Lehmann, A.R., and Jeggo, P.A. 1997. Molecular and biochemical characterization of xrs mutants defective in Ku80. *Mol. Cell. Biol.* **17**: 1264–1273.
- Song, K., Jung, Y., Jung, D., and Lee, I. 2000. Human Ku70 interacts with HP1alpha. J. Biol. Chem. 276: 8321–8327.
- Tatusova, T.A., Karsch-Mizrachi, I., and Ostell, J.A. 1999. Complete genomes in WWW Entrez: Data representation and analysis. *Bioinformatics* **15**: 536–543.
- Teo, S.H., and Jackson, S.P. 1997. Identification of Saccharomyces cerevisiae DNA ligase IV: Involvement in DNA double-strand break repair. EMBO J. 16: 4788–4795.
- 2000. Lif1p targets the DNA ligase Lig4p to sites of DNA double-strand breaks. *Curr. Biol.* 10: 165–168.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.
- Wang, J., Dong, X., Myung, K., Hendrickson, E.A., and Reeves, W.H. 1998. Identification of two domains of the p70 Ku protein mediating dimerization with p80 and DNA binding. *J. Biol. Chem.* 273: 842–848.
- Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. 2001. Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res.* 11: 356–372.
- Wu, X. and Lieber, M.R. 1996. Protein–protein and protein–DNA interaction regions within the DNA end-binding protein Ku70-Ku86. *Mol. Cell. Biol.* **16**: 5186–5193.

Received January 18, 2001; accepted in revised form May 14, 2001.



Prokaryotic Homologs of the Eukaryotic DNA-End-Binding Protein Ku, Novel Domains in the Ku Protein and Prediction of a Prokaryotic Double-Strand Break Repair System

L. Aravind and Eugene V. Koonin

Genome Res. 2001 11: 1365-1374 Access the most recent version at doi:10.1101/gr.181001

> **References** This article cites 48 articles, 18 of which can be accessed free at: http://genome.cshlp.org/content/11/8/1365.full.html#ref-list-1

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.



To subscribe to Genome Research go to: https://genome.cshlp.org/subscriptions