J Mol Evol (1999) 48:291–302

JOURNAL OF MOLECULAR EVOLUTION

Novel Predicted RNA-Binding Domains Associated with the Translation Machinery

L. Aravind,[1,2] Eugene V. Koonin[2]

[1] Department of Biology, Texas A&M University, College Station, TX 77843, USA
[2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received: 15 May 1998 / Accepted: 20 July 1998

Abstract. Two previously undetected domains were identified in a variety of RNA-binding proteins, particularly RNA-modifying enzymes, using methods for sequence profile analysis. A small domain consisting of 60–65 amino acid residues was detected in the ribosomal protein S4, two families of pseudouridine synthases, a novel family of predicted RNA methylases, a yeast protein containing a pseudouridine synthetase and a deaminase domain, bacterial tyrosyl-tRNA synthetases, and a number of uncharacterized, small proteins that may be involved in translation regulation. Another novel domain, designated PUA domain, after PseudoUridine synthase and Archaeosine transglycosylase, was detected in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, a family of predicted ATPases that may be involved in RNA modification, a family of predicted archaeal and bacterial rRNA methylases. Additionally, the PUA domain was detected in a family of eukaryotic proteins that also contain a domain homologous to the translation initiation factor eIF1/SUI1; these proteins may comprise a novel type of translation factors. Unexpectedly, the PUA domain was detected also in bacterial and yeast glutamate kinases; this is compatible with the demonstrated role of these enzymes in the regulation of the expression of other genes. We propose that the S4 domain and the PUA domain bind RNA molecules with complex folded structures, adding to the growing collection of nucleic acid-binding domains associated with DNA and RNA modification enzymes. The evolution of the translation machinery components containing the S4, PUA, and SUI1 domains must have included several events of lateral gene transfer and gene loss as well as lineage-specific domain fusions.

Key words: RNA-binding domains — Ribosomal protein S4 — Archaeosine transglycosylase — Pseudouridine synthase — Translation machinery

Introduction

Nucleotide modification in rRNAs and tRNAs is a major venue of structural diversification and change of specificity of these molecules (Limbach et al. 1994). These posttranscriptional modifications include various forms of base methylation, in situ base conversion as in pseudouridine (ψ) formation, or insertion of a modified base by transglycosylation as in the case of queuine (Limbach et al. 1994; Lane et al. 1995; Ofengand et al. 1995; Romier et al. 1996). While most common in the noncoding RNAs (Tollervey and Kiss 1997), modifications also occur in mRNA and, in this case, may involve the more dramatic RNA editing (Smith, Sowden 1996). Some of the RNA modifications, e.g., ψ formation, are universal in all cells, whereas others are restricted to one domain of life such as archaeosine formation in the Archaea (Gregson et al. 1993; Watanabe et al. 1997).

A broad spectrum of RNA modification enzymes has been identified, including various RNA-specific methylases, ψ synthases, tRNA-guanine transglycosylases involved in queuine (in Bacteria and eukaryotes) and archaeosine (in Archaea) formation, cytosine deaminases

Correspondence to: E.V. Koonin; *e-mail:* koonin@ncbi.nlm.nih.gov

involved in mRNA editing, and others (Rottman et al. 1994; Navarathan et al. 1995; Koonin 1996; Romier et al. 1997). Evidently, for each of these enzymes to perform the requisite reactions on RNA molecules, an RNA-binding moiety is required, either as a built-in domain of the enzyme or as a stand-alone protein or RNA. DNA and RNA-binding domains are central to genome replication, expression, and stability maintenance. There is no single structural equivalent of a nucleic acid-binding domain but the repertoire is relatively small, with the same domains recurring in different systems and frequently embedded in multidomain proteins with different architectures. Some of the DNA- and RNA-binding domains tend to combine with enzymatic domains, for which they serve as a vehicle delivering the enzymes to the sites of their action on nucleic acids. Such mobile nucleic acid-binding modules associated with enzymatic domains include the helix–hairpin–helix (HhH) domain identified in a plethora of proteins involved in DNA replication and repair as well as ribosomal proteins (Doherty et al. 1996), the double-stranded RNA-binding domain present in RNAase III (St Johnston et al. 1992), and the S1 domain found in ribosomal proteins and polynucleotide phosphorylase (Bycroft et al. 1997). Other nucleic acid-binding domain do not have this tendency of combining with enzymatic domains and occur on their own as single or multiple copies or with other nonenzymatic domains. Examples of such ''loners'' include the RRM (RNA recognition motif) domain, which is common in a number of RNA-binding proteins (Shamoo et al. 1995), the so-called KOW (Kyrpides–Ouzounis–Woese) domain found in NusG and ribosomal proteins (Kyrpides et al. 1996), and the cold shock domains of the OB (oligonucleotide/oligosaccharide binding) fold class (Schnuchel et al. 1993).

In the course of comparative analysis of the protein sequences encoded in bacterial, archaeal, and eukaryotic genomes, we discovered two conserved and previously undetected domains. The first of these is an ancient domain found in ribosomal protein S4 (and designated the S4 domain after this well-studied protein), tyrosyl-tRNA synthetase, the RluA and RsuA families of ψ synthases, several predicted RNA methylases, a putative novel RNA editing enzyme, and a number of uncharacterized proteins. The second domain (designated PUA, after *p*seudo*u*ridine synthase and *a*rchaeosine transglycosylase) is seen primarily in eukaryotes and Archaea and is associated with ψ synthases of the TruB family, tRNA-guanine transglycosylases, putative novel archaeal ATPases involved in RNA modification, a predicted archaeal rRNA methylase, and a conserved family of putative novel eukaryotic translation factors. In addition and unexpectedly, PUA domain has been detected in bacterial and yeast glutamate kinases. Both S4 and PUA domains are predicted to deliver nucleotide-modifying enzymes to their target sites in RNA and, on other occasions, to act as translation regulators, via structure-specific RNA binding. A combined analysis of the phylogenetic distribution and domain architectures of the proteins containing these newly described domains provides for plausible evolutionary scenarios that are based on the parsimony principle and include vertical inheritance, postulated gene transfer events, and domain fusion and shuffling.

Materials and Methods

The nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda, MD) was searched using the gapped BLASTP program (Altschul et al. 1997). The nucleotide database of expressed sequence tags (dbEST) was searched using the gapped version of the TBLASTN program, which uses a protein sequence as a query and searches the nucleotide database translated in all six frames. Further iterative searches were carried out using the PSI-BLAST program, which constructs a position-specific weight matrix based on the multiple alignment generated in the first pass of the database screening and subsequently iterates the search using this profile (Altschul et al. 1997). Additionally, separately constructed multiple alignments were used to derive a weighted profile and seed the PSI-BLAST iterations (A. Schaffer, L.A., and E.V.K., unpublished observations). Further database searches were carried out using the MoST program, which uses ungapped alignment blocks as an input to construct a position-specific weight matrix and iteratively searches the database until convergence using the ratio of the number of observed matches to the expected number as a cutoff (Tatusov et al. 1994). Multiple alignments were constructed by searching for similarity blocks using Gibbs sampling as implemented in the MACAW program (Schuler et al. 1991; Neuwald et al. 1995); the alignments were subsequently optimized globally using the ClustalW program (Thompson et al. 1994) and further modified using the outputs of the PSI-BLAST searches as a guide. The proteins were filtered for low-complexity regions and decomposed into their predicted globular domains using the SEG program with the parameters of window size 45, trigger complexity of 3.4, and extension complexity of 3.75 (Wootton and Federhen 1996). Secondary structure predictions were performed with the PHD program using the entire family alignment or subsets that produced most reliable alignments (Rost and Sander 1994). Structural database threading was performed on the basis of the secondary structure predictions (Rost et al. 1997).

Results and Discussion

The S4 Domain

A number of ψ synthases, such as SFHB and YPUL of the RSUA family and YCEC and YA32 of the RLUA family (Koonin, 1996), contain a predicted globular domain, which is located between the N terminus and the ψ synthase catalytic domain. When the NR database was searched with the sequence of this domain using the PSI-BLAST program, ribosomal protein S4 sequences were retrieved at random expectation (*e*) values in the range of 10^{-3} to 10^{-4} in the first or second iteration. When these searches were run to convergence, we also

detected several additional proteins at *e* values below 10^{-3}. These include bacterial tyrosyl-tRNA synthetases and a conserved group of bacterial proteins with additional methyltransferase domains (Figs. 1 and 2), which have been annotated in sequence databases as hemolysins on the basis of the properties of a single protein from the spirochaete *Serpulina hyodysenteriae* (ter Huurne et al. 1994). In order to assess the statistical significance of these sequence similarities more robustly, we carried out reverse searches using the corresponding regions from these database hits. In each of these cases, it was possible to reproducibly retrieve essentially the same set of proteins at *e* values below 10^{-3} within four iterations, suggesting a coherent evolutionary and structural relationship between these proteins. Similar results were obtained in motif searches using the MoST program with *r* values of 0.005–0.01.

Thus, we defined a common globular domain in all these proteins, which we call S4 domain after the thoroughly studied ribosomal protein that appears to have an important conserved role in ribosomal structure and function (Tang and Draper 1989; Baker and Draper 1994; Heilek et al. 1995). The S4 domain consists of 60–65 amino acid residues and typically occurs in a single copy at various positions in different proteins. The boundaries of the domain were determined on the basis of both sequence conservation and the domain organization of the respective proteins. More specifically, in tyrosyl-tRNA synthetases the S4 domain is in the extreme C terminus, defining the C-terminal boundary, and in the predicted methyltransferases, it is at the extreme N terminus, indicating the N-terminal limit (Fig. 2). Notably, it is precisely the S4 domain in the boundaries defined here that is duplicated in the *Chlamydomonas* chloroplast S4 ribosomal protein, providing additional support for its identification as an independent, structurally distinct unit (Figs. 1 and 2).

Inspection of the multiple alignment of the S4 domain superfamily shows a characteristic pattern of conserved polar, hydrophobic, and small residues in its N-terminal and central parts; these residues are likely to define the conserved structural features of the domain. The C-terminal part does not show a high degree of conservation between different families within the superfamily and may be important for interactions specific to each family (Fig. 1). Structural prediction using the PHD program suggests that this domain has an α/β structure with the central highly conserved region corresponding to the β elements (Fig. 1).

The evidence of the RNA-binding function of the S4 domain comes from several independent studies on ribosomal protein S4 itself and on the C-terminal domain of bacterial tyrosyl-tRNA synthetase. The *E. coli* S4 is a multifunctional protein, which binds to a region of 16S rRNA that comprises most of the 5′ domain and spans about 460 nucleotides (Heilek et al. 1995). In addition, S4 binds a complex pseudoknot in its own transcript and represses its translation; the protein also shows nonspecific RNA binding (Tang and Draper 1989). There appears to be no similarity between the two independent RNA structures recognized by this protein. Early studies yielded conflicting results regarding the parts of the protein needed for the interaction with 16S RNA. However, a more recent detailed analysis of deletions in the S4 protein, along with circular dichroism (CD) studies, have suggested that the region spanning residues 48 to 177 reproduces the binding specificity typical of the full length protein, indicating that the main RNA binding determinants lie in this region (Baker and Draper 1995). The CD spectra suggest that this fragment of the protein is structured similarly to the intact protein but any further deletions moving into the S4 domain defined above cause the loss of structure (Baker and Draper 1995). Furthermore, toeprint assays imply a bimodal RNA contact by S4 protein, with one of these contacts specifically dependent on the region encompassing the S4 domain and the other one mediated by a predicted α helix-rich globular domain which is located N-terminally of the S4 domain (Fig. 2).

Independent evidence of the RNA-binding properties of the S4 domain come from the studies on the C-terminal domain of the *Bacillus stearothermophilus* tyrosyl-tRNA synthetase, which is required for tRNA binding by this enzyme (Carter et al. 1986). Though this domain appears disordered in the crystal structure (Brick et al. 1989), CD spectra and fluorescence spectroscopy suggest that it is structured in solution (Guez-Ivanier and Bedouelle 1996). It may be speculated that tRNA binding is required for the S4 domain in the tyrosyl-tRNA synthetase to assume its proper conformation. These experimental results, together with our observations on the presence of the S4 domain in RNA-modifying enzymes, suggest that this domain is capable of recognizing complex, unique 3D features in highly folded RNA molecules such as rRNAs, tRNAs, and untranslated regions of mRNAs. The presence of the S4 domain in ψ synthases such as RsuA, which pseudouridylates position 516 located in a highly structured portion of *E. coli* 16S RNA (Wrzesinski et al. 1995a), is compatible with these observations. Our database searches show that the *E. coli* RluA ψ synthase, which acts on both tRNA and 23S RNA (Wrzesinski et al. 1995b), lacks a S4 domain but homologous, highly conserved proteins from *E. coli* and a number of other bacterial species (e.g., the *E. coli* proteins YceC and SfhB) contain it, suggesting mechanistic similarities between substrate recognition by many ψ synthases, tRNA binding by tyrosyl-tRNA synthetases, and rRNA binding by S4. The differences in the S4 domains of these enzymes, particularly in the variable C-terminal portions, may contribute to the unique target

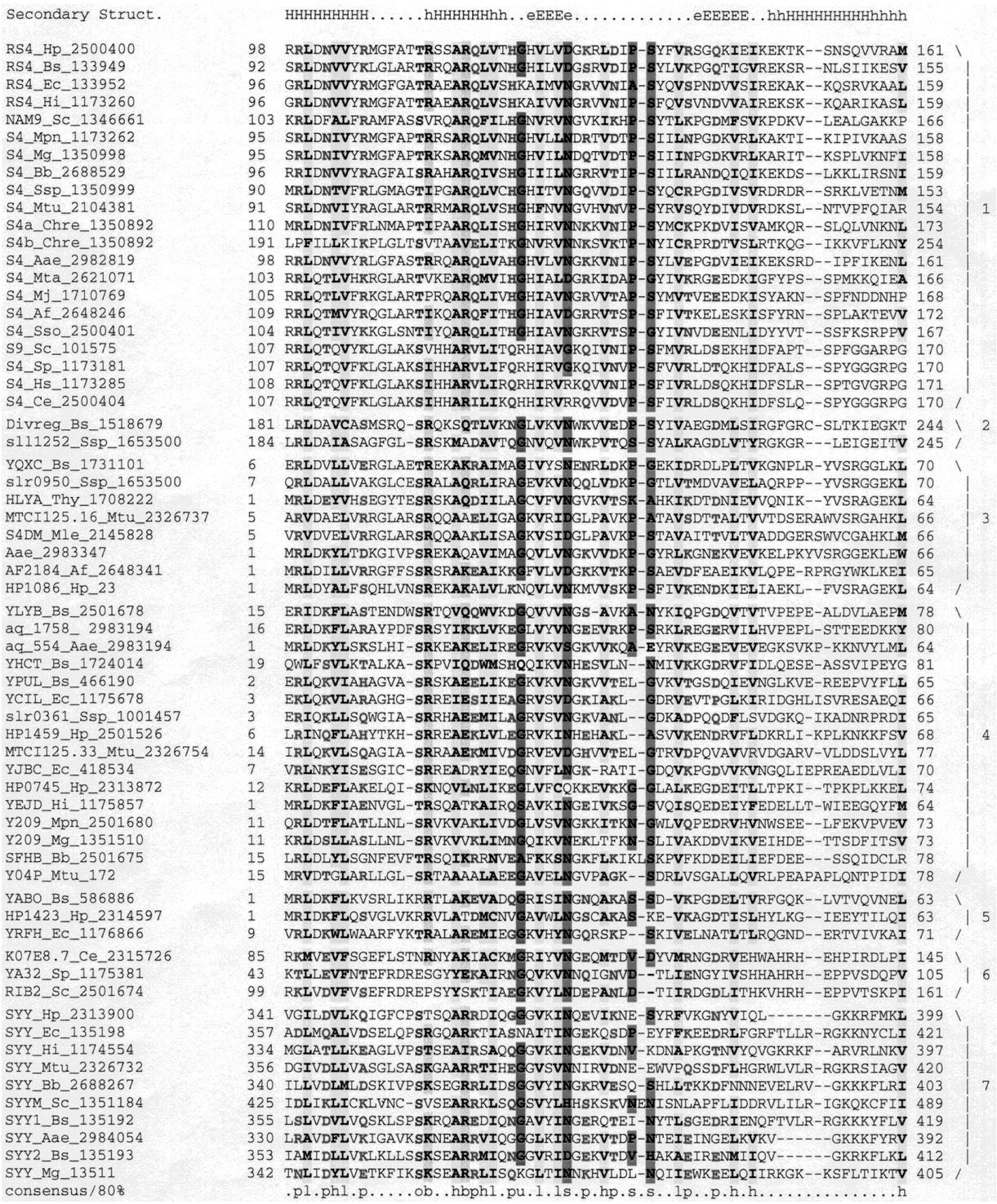

Fig. 1. The S4 domain. The multiple alignment was constructed as described in the text using MACAW and CLUSTALW. Names of the proteins are as assigned in SWISSPROT or as per designations of the individual genome sequencing projects, with the GenBank numbers shown next to them. The numbers flanking each line of the alignment indicate the boundaries of the S4 domain in the polypeptide. The sequences are grouped based on similarity and domain organization. Group 1, sequences of the S4 domains from the ribosomal protein S4; group 2, uncharacterized bacterial proteins; group 3, methyltransferases; group 4, bacterial ψ synthases; group 5, bacterial stand-alone versions of the S4 domain; group 6, the eukaryotic ψ synthases; and group 7, bacterial tyrosyl-tRNA synthetases. The conserved positions are shaded using the 80% consensus rule. The amino acid classes used in building the consensus were hydrophobic residues (A, C, F, I, L, M, V, W, Y; *yellow background*), small residues (A, C, S, T, D, N, V, G, P; *blue background*), tiny residues (A, G, S; *green background*), big residues (L, F, W, Y, E, R, K, I; *gray background*), polar residues (S, T, E, D, R, K, H, N, Q; *brown coloring*), hydroxylic residues (S, T; *blue coloring*), and charged residues (R, K, H, D, E; *purple coloring*). The multiple alignment-based secondary structure prediction is indicated on *top* of the alignment; E(e) indicates extended conformation (β-strands), and H(h) indicates α helices; *capital letters* indicate the most confident predictions. Species abbreviations: Aae, *Aquifex aeolicus;* Af, *Archaeoglobus fulgidus;* Bb, *Borrelia burgdorferi;* Bs, *Bacillus subtilis;* Ce, *Caenorhabditis elegans;* Chre, *Chlamydomonas rheinhardtii;* Ec, *Escherischia coli;* Hi, *Haemophilus influenzae;* Hp, *Helicobacter pylori;* Hs, *Homo sapiens;* Mj, *Methanococcus jannaschii;* Mpn, *Mycoplasma pneumoniae;* Mg, *Mycoplasma genetalium;* Mta, *Methanobacterium thermoautotrophicum;* Mtu, *Mycobacterium tuberculosis;* Mle, *Mycobacterium leprae;* Sc, *Saccharomyces cerevisiae;* Sp, *Schizosaccharomyces pombe;* Thy, *Treponema hyodysenteriae.*

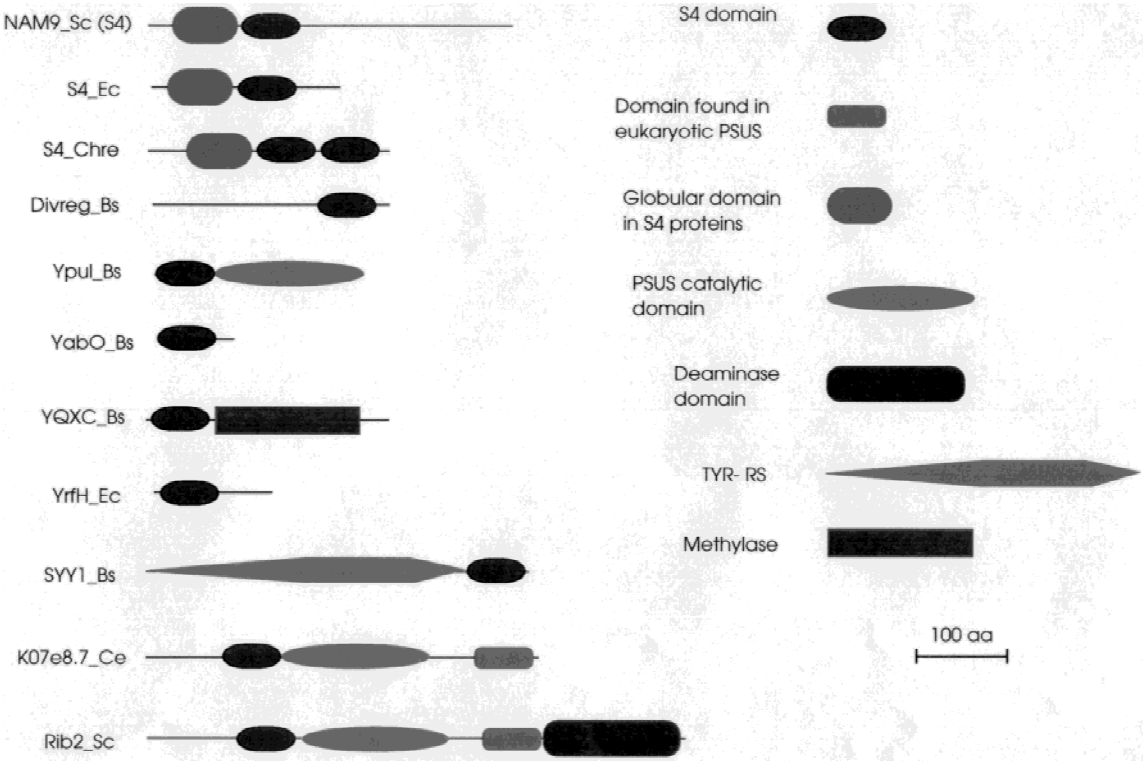

Fig. 2. Representative domain architectures of S4 domain-containing proteins. The proteins are drawn to scale as indicated. PSUS stands for the catalytic domain of the ψ synthases, in this case, of the RluA and RsuA families. The additional globular domain in the eukaryotic PSUS is the predicted metal-binding domain mentioned in the text. The additional α helix-rich globular domain in S4 that has been shown to contact RNA independently is also shown.

specificities, given the high sequence conservation in the catalytic domains. Furthermore, the presence of the S4 domain in bacterial tyrosyl-tRNA synthetases but not in their archaeal or eukaryotic orthologs suggests major differences in the modes of $tRNA^{Tyr}$ binding.

The detection of the S4 domain in several other proteins seems to be of considerable interest. A group of eukaryotic ψ synthases of the RluA family contains the S4 domain as well as additional, uncharacterized, conserved domains, with the distinctive domain organization conserved at least in the crown group of eukaryotes (Fig. 2). Remarkably, one of the four yeast ψ synthases—Rib2, contains, in addition to the S4 domain, a C-terminal domain similar to a wide range of deaminases, including RNA editing cytidine deaminases (Fig. 2). While it has been initially proposed that Rib2 has a function in riboflavin biosynthesis (Tzermia et al. 1997), the observed domain organization suggests that it may be a novel RNA-editing enzyme that produces ψ from cytosine through deamination with subsequent pseudouridylation.

Another previously undetected family includes proteins that, in addition to the S4 domain, contain the conserved motifs typical of SAM-dependent methyltransferases (Kagan and Clarke 1994; Koonin et al. 1995), and are encoded by a variety of taxonomically diverse bacteria, including *Helicobacter pylori, Bacillus subtilis, Spirochaetae, Mycobacteria, Synechocystis, Aquifex aeolicus,* and the archaeon *Archaeoglobus fulgidus* (Figs. 1 and 2). The member of this family from *Treponema hyodysenteriae* has been described as a hemolysin (ter Huurne et al. 1994), but given the broad distribution of the family in both parasitic and free-living species and their domain organization, this is unlikely to define the common function of these proteins. The combination of the S4 domain, predicted to bind RNA, and the methyltransferase domain, suggests that this is a novel family of rRNA methylases. Given their patchy phylogenetic distribution, these enzymes may be responsible for taxon-specific rRNA methylation.

Several small bacterial proteins (e.g., YabO from *B. subtilis* and YrfH from *E. coli*) consist almost entirely of the S4 domain (Fig. 2). An interesting possibility is that these putative RNA-binding proteins may be involved in translational regulation similarly to S4. The sporadic occurrence of these proteins in bacteria suggests that they may have specific functions only in the context of certain operons and their untranslated leaders.

The PUA Domain

Using the same iterative strategy for database searching that is described above for the S4 domain, significant sequence similarity (*e* value below 10^{-4} at the second PSI-BLAST iteration) was detected between the C-terminal regions of the archaeal tRNA guanine-

transglycosylase involved in archaeosine synthesis and the eukaryotic CBF5 proteins, which belong to the TruB family of ψ synthetases (Koonin 1996). Both of these proteins are involved in RNA base modification but their catalytic domains are completely unrelated, leading to the suggestion that the conserved region defines a novel RNA-binding domain. Comprehensive iterative searches resulted in the identification of this domain (hereafter PUA domain, after *p*seudo*u*ridine synthases and *a*rchaeosine-specific transglycosylases) in other archaeal and eukaryotic proteins with additional domains and also as highly conserved stand-alone versions. The PUA domain clearly is less common in bacteria, but quite unexpectedly, in addition to predicted RNA methylases found only in two species, it was detected at the C termini of bacterial and fungal glutamate kinases (Figs. 3 and 4).

Like the S4 domain, PUA is a small, compact domain that consists of 78–83 amino acid residues. The N-terminal boundary was defined by the extreme N-terminal location of the PUA domain in a family of predicted RNA methylases, whereas the C-terminal boundary was easily determined given the C-terminal location of the PUA domain in CBF5-like ψ synthases and tRNA transglycosylases (Fig. 4). Inspection of the multiple alignment of the PUA domain reveals highly conserved motifs that center around stretches of hydrophobic residues (Fig. 3). The domain also contains two highly conserved positions occupied by small amino acids (primarily glycines), which may define a functionally important turn. Secondary structure prediction using the PHD program suggests that PUA domain has a β-strand-rich structure, with the predicted strands corresponding to the conserved hydrophobic patches (Fig. 3). This pattern of secondary structure elements resembles that in some other RNA-binding domains, particularly the OB fold (Murzin 1993). Given the prevalence of the OB fold among RNA binding domains, it cannot be ruled out that PUA is a distinct version of this fold.

Unlike the situation with the S4 domain, experimental details that could shed light on the specific functions of the PUA domain are sketchy. The conservation of CBF5 and the associated PUA domain in archaea and eukarya suggests that this enzyme performs a fundamental pseudouridylation function, for which PUA domain is required. Interestingly, the second yeast ψ synthase of the TruB family lacks a PUA domain and has been shown to catalyze ψ formation in both cytoplasmic and mitochondrial tRNAs (Becker et al. 1997); the absence of the PUA domain is compatible with the transfer of the gene coding for this enzyme from bacteria to the eukaryotic nuclear genome. In contrast, the PUA domain-containing protein CBF5 is essential for rRNA maturation but does not affect tRNA, suggesting that PUA domain may contribute to the specific rRNA binding. Interestingly, the human orthologue of CBF5, a protein named dyskerin, is the product of the gene mutated in a bone marrow failure disorder, X-linked recessive dyskeratosis congenita, suggesting that a defect in rRNA maturation may be involved in the pathogenesis (Heiss et al., 1998).

The Archaeosine biosynthesis enzymes have catalytic domains that are distantly related to the family of transglycosylases involved in queuine (Q) biosynthesis in bacteria and eukaryotes (Gregson et al. 1993; Watanabe et al. 1997). These Q-specific transglycosylases (TGT) lack the PUA domain and are present only in bacteria and eukarya, with the exception of the archaeon *A. fulgidus*, which appears to have acquired a single gene encoding a Q-specific TGT from bacteria. Conversely, the archaeal-type archaeosine-specific TGTs were not detected in eukaryotes (including ESTs) or in bacterial genomes sampled so far. The four sequenced archaeal genomes each encode two distinct forms of the TGT with a C-terminal PUA domain (Fig. 4), which may be specifically responsible for archaeosine incorporation at different sites in tRNA molecules.

The uncharacterized proteins containing the PUA domain possess interesting features that make them candidates for further studies on RNA modification. One of these is a small protein highly conserved in archaea and eukarya (e.g., yeast YER007). This is a stand-alone version of the PUA domain (Fig. 4), and its extraordinary conservation suggests that it may function as a RNA-binding cofactor for RNA modifying enzymes.

Another protein family, which includes mammalian ligatin, a protein that, apparently erroneously, has been described as a trafficking receptor for phosphoglycoproteins (Jakoi et al. 1989), is characterized by a N-terminal PUA domain and at least one additional C-terminal globular domain (Fig. 4). This family is highly conserved at least in the crown group of eukarya. A database search with the C-terminal domains of the ligatin family proteins unexpectedly showed a marginally significant similarity (*e* value, ~0.07) with the sequences of the eukaryotic translation initiation factor eIF-1/SUI1 and its orthologues from archaea and bacteria (Kasperaitis et al. 1995; Tatusov et al. 1997; Kyrpides and Woese 1998) as well as a conserved family of uncharacterized eukaryotic proteins. Iterative reverse searches with the proteins of the SUI1 family and the new conserved family showed statistically significant similarity to the ligatins (e.g., *e* value $<10^{-4}$ at the third iteration in a search initiated with the yeast protein YJY4). Multiple alignment of the SUI1 orthologues and newly detected similar domains contains a number of conserved positions, with two prominent motifs located in the middle and near the C terminus of the domain, and is predicted to possess a compact β/α structure (Fig. 5). SUI1 is involved in the recognition of the initiation codon by the eIF2–GTP–Met–$tRNA_i$ ternary complex (Naranda et al. 1996; Yoon and Donahue 1992) and, accordingly, is likely to possess RNA-binding activity. So far, the members of this family have been identified only as standalone small proteins (Kyrpides and Woese 1998) but the findings presented here

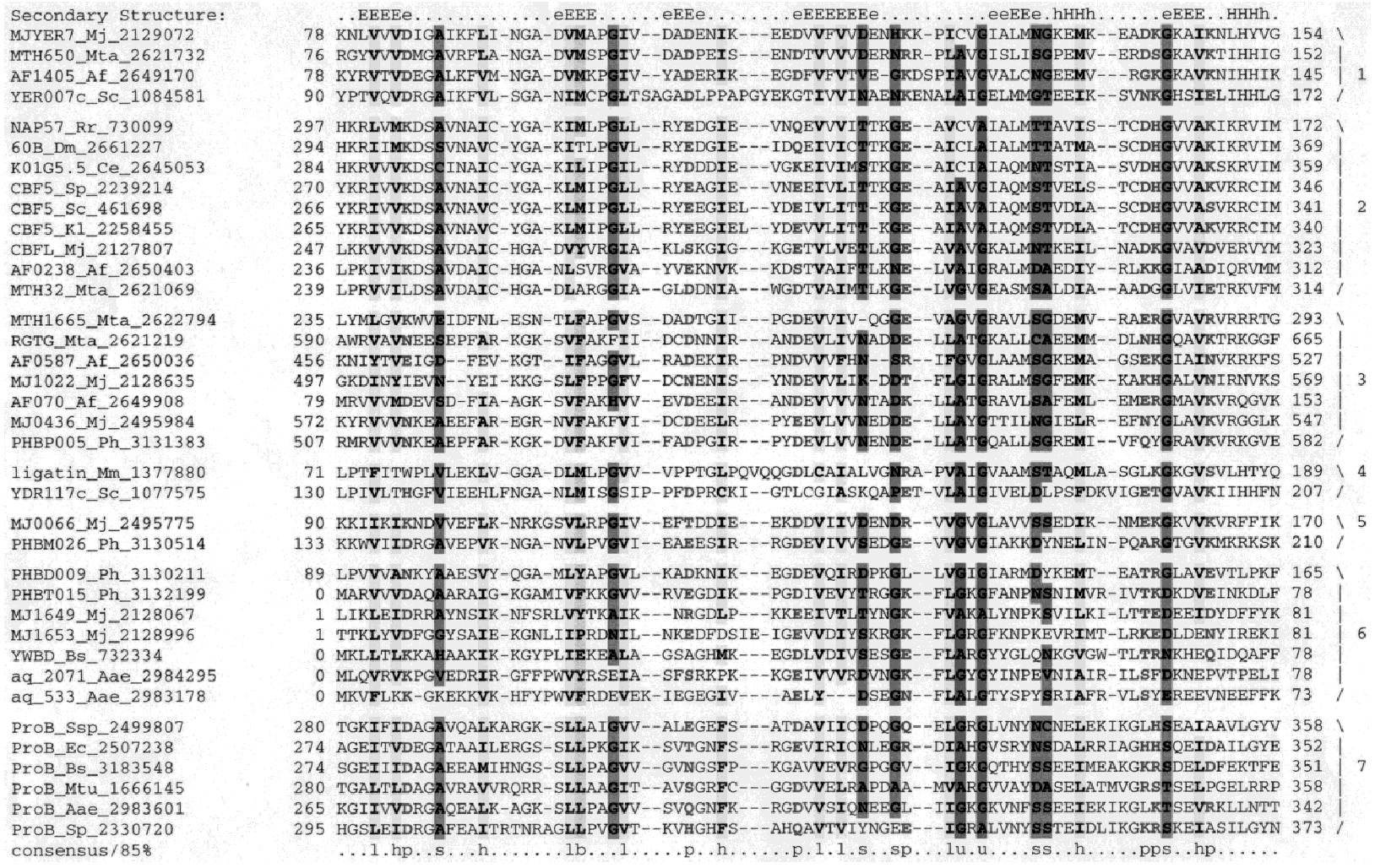

Fig. 3. The PUA domain. The multiple alignment of the PUA domain was constructed as indicated in the legend to Fig. 1. The conserved residues are shaded using the 80% consensus rule. The shading conventions and secondary structure predictions designations are as in Fig. 1. Group 1, stand-alone PUA domain proteins such as yeast YER007 and its homologues; group 2, PUA domains fused to the TRUB family ψ synthase domain; group 3, PUA domains fused to the tRNA transglycosylases; group 4, PUA domains in ligatin-like proteins; group 5, PUA domains fused to PAPS reductase-like domains; group 6, PUA domain fused to methylases; group 7, PUA domain fused to glutamate kinases. The species abbreviations are as in Fig. 1. Additional abbreviations: Dm, *Drosophila melanogaster;* Kl, *Kluyveromyces lactis;* Mm, *Mus musculus;* Rr, *Rattus rattus;* Ph, *P. horikoshii.*

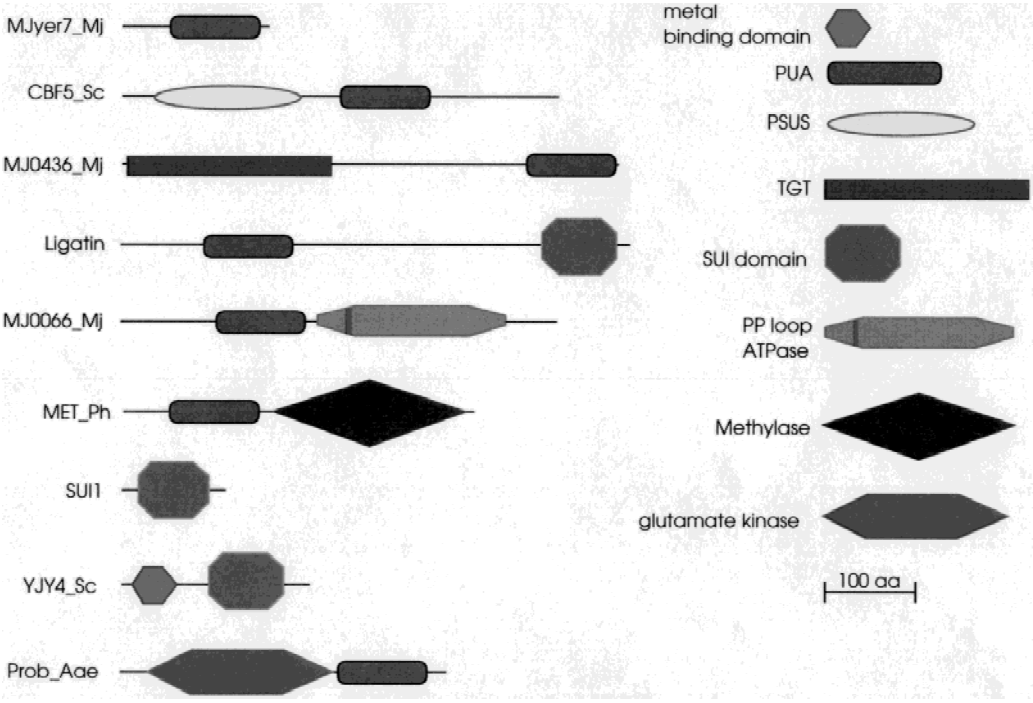

Fig. 4. Representative domain architectures of PUA domain-containing proteins. The proteins are drawn to scale as indicated. The PSUS domain is the catalytic domain of the ψ synthases of the TruB family and the TGT domain is the TIM barrel fold catalytic domain of the tRNA glycosylases. The PP-loop ATPase domain is the domain similar to PAPS reductases. The *red bar* in this domain specifies the intact PP loop, which is typically disrupted in the PAPS reductases (42). The proteins are named as in Fig. 3.

strongly suggest that SUI1 is yet another mobile RNA-binding domain. The ligatin-like protein and the other conserved eukaryotic proteins typified by YJY4 may be a novel class of translation factors with two RNA-binding domains of different specificity.

Two proteins found in *M. jannaschii*, with an orthologue of at least one of them detected in *Pyrococcus horikoshii*, contain a PUA domain preceded by a specific ferredoxin-like domain and succeeded by a domain that is similar to PAPS reductase (Fig. 4) but retains an intact PP-loop motif, suggesting ATPase activity (Bork and Koonin 1994). An interesting possibility is that these proteins are involved in taxa-specific RNA modifications. In a number of archaeal species, tRNAs contain thiolated bases such as thiouridine derivatives and thiothymine, which may play a role in the thermal stability (Edmonds et al. 1991). It seems possible that the PUA domain–PAPS reductase proteins are uncharacterized enzymes involved in the biosynthesis of such bases through sulfate activation and subsequent reduction. Alternatively, they may be involved in sulfate modification of the sugar backbone analogous to the reaction catalyzed by the NodP protein (Schwedock et al. 1994). In a striking parallel with the S4 domain family, a unique fusion of the PUA domain with a predicted rRNA methylase of a distinct family that includes eukaryotic nucleolar proteins (e.g., P120) and their bacterial homologues (Koonin 1994) was observed in *P. horikoshii* (Fig. 4).

Finally, the presence of a bona fide PUA domain in the glutamate kinases from most bacteria and from yeasts was completely unexpected, as the principal function of this enzyme in proline biosynthesis has nothing to do with RNA binding. It has been shown, however, that *Bacillus subtilis* glutamate kinase (the *proB* gene product) down-regulates the expression of sigma D-dependent genes by an as yet unknown mechanism, possibly by inhibiting the translation of sigma D itself (Ogura, Tanaka, 1996). It may be speculated that such a regulatory mechanism is wide spread in bacteria and requires binding of a specific RNA structure by the PUA domain. Experimental demonstration of the role of the PUA domain in glutamate kinases will be of major interest.

Evolutionary Implications of the Phylogenetic Distribution of S4 and PUA Domains

Examination of the phylogenetic distribution and sequence relationships of S4 and PUA domains, together with those of the RNA modifying enzymes, may help in unraveling the evolutionary history of the respective protein families, which seem to be intimately linked to the evolution of translation (Fig. 6). The resulting evolutionary scenarios are based on the parsimony principle and accordingly are speculative inasmuch as it can never be proved that evolution had followed the most parsimonious scheme. Nevertheless, such scenarios seem to provide the most likely explanation for the observed relationships.

Obviously, ribosomal protein S4 has been encoded by the cenancestor genome. This is compatible with the uni-

Fig. 5. The eIF1/SUI1 domain in translation factors and in PUA domain-containing proteins. The alignment construction and all the designations are as in Figs. 1 and 3. Group 1, eukaryotic and archaeal eIF1/SUI1 factors; group 2, uncharacterized family of eukaryotic proteins containing a C-terminal SUI1 domain and a predicted metal-binding domain at the N terminus; 3, the ligatin-like family of PUA domain-containing proteins. Additional species names: At, *Arabidopsis thaliana;* Zm, *Zea mays.*

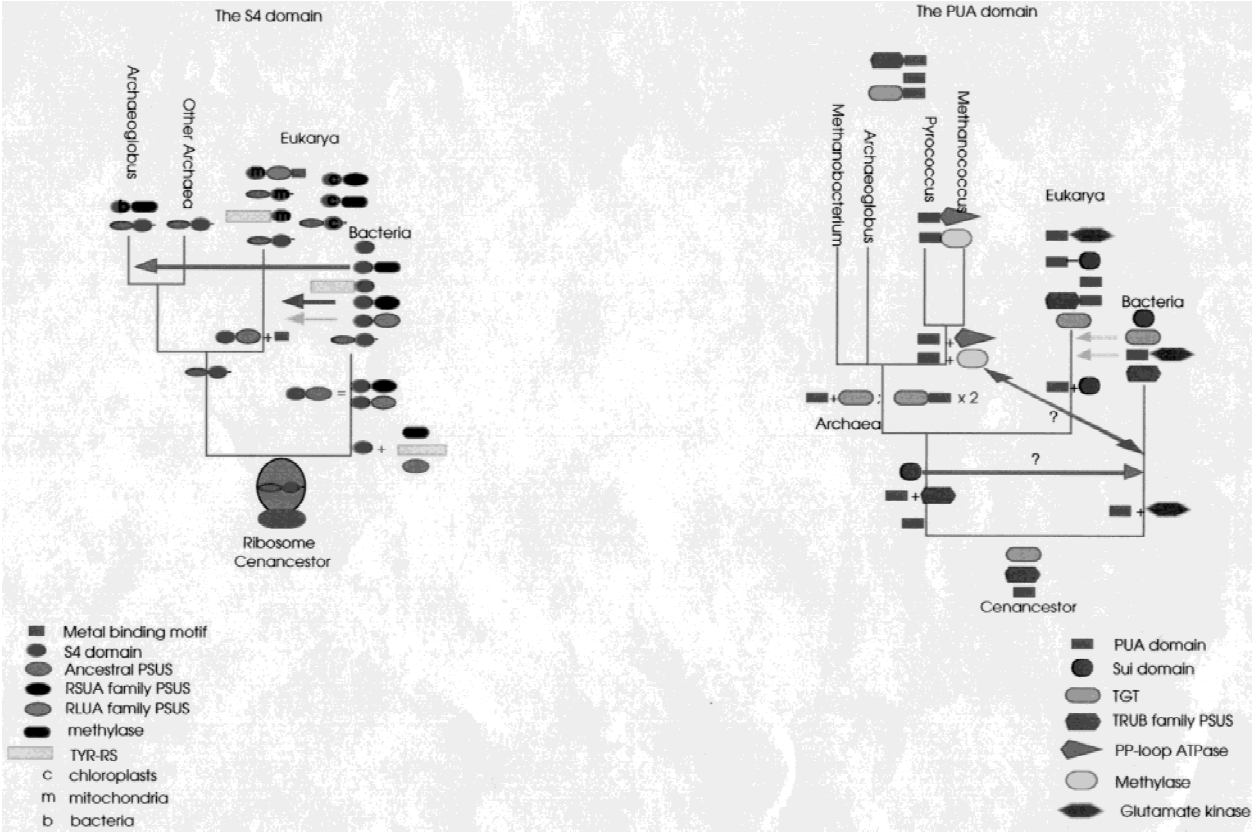

Fig. 6. Hypothetical scenarios for the evolution of the translation machinery components containing the S4 and PUA domains. *Arrows* show postulated horizontal gene transfers; ''×2'' indicates a postulated duplication.

versal role of S4 in the translation accuracy center of the ribosome, suggesting that the interactions between S4 and 16S RNA (Heilek et al. 1995) had already been established in the cenancestor. In contrast, the related RsuA and the RluA families of ψ synthases (Koonin 1996) and the small, stand-alone versions of the S4 domain are largely restricted to Bacteria and eukaryotes, to the exclusion of the Archaea. The bacteria show the greatest diversity in terms of domain organization of the proteins containing the S4 domain (Fig. 2) and also the largest number of ψ synthases of the RsuA and the RluA families. Taken together with the presence of the S4 domain in members of both of these families, this suggests that early in the evolution of the bacteria, the ribosomal S4 domain has been exapted for the related function of rRNA recognition by the common ancestor of the RsuA and RluA families of ψ synthases (Fig. 6a). A similar selective pressure for the recognition of specific

structures in RNA appears to have favored its fusion with the predicted RNA methylases and with tyrosyl-tRNA synthetase. Given the involvement of S4 in translational regulation in bacteria, it is likely that certain versions of this domain have been were recruited for solely regulatory functions as proposed above. The bacterial S4 domain-containing ψ synthases appear to have entered the eukaryotes from their endosymbionts. Specifically, so far only RluA family members have been detected in fungi and animals, whereas plants encode also at least one RsuA family enzyme, which suggests gene transfer from the mitochondrial genome and from the chloroplast genome, respectively (Fig. 6a). Some of these proteins clearly have undergone further evolution in eukaryotes as exemplified by the fusion of enzymatic domains in the putative editing enzyme Rib2 and the acquisition by eukaryotic ψ synthases of an additional predicted metal-binding domain C terminal of the catalytic domain (Figs. 2 and 6a). In the same vein, the S4 domain-containing methylase of *Archaeoglobus fulgidus* most likely has been acquired via horizontal transfer from bacteria. The fusion of the S4 domain to the tyrosyl-tRNA synthetase appears to have been an ancient event, as it predates the separation of all known bacterial lines; interestingly, it is paralleled by a similar fusion of an unrelated RNA-binding domain of the OB fold class to animal tyrosyl-tRNA synthetases (L.A., A.G. Murzin, and E.V.K., unpublished observations).

Similarly to the evolutionary history of the S4 domain-containing proteins, the evolutionary scenario of the PUA-containing proteins should have included several gene fusion events as well as lateral gene transfers (Fig. 6b). It appears most likely that the cenancestor genome encoded a stand-alone version of the PUA domain, whereas two of the extant PUA-containing protein families, namely, the TruB family of ψ synthases and the TGT family, were also represented, but lacked the PUA domain. The formative events in the evolution of the PUA-containing proteins should have been the early fusion to the TruB domain in the common ancestor of Archaea and Eukarya and a parallel, independent fusion to glutamate kinase in an ancestral bacterium. Along the archaean lineage, a variety of further fusions has occurred, resulting in the transfer of the PUA domain to archaea-specific enzymes at different points in evolution. Similarly, in the eukaryotic lineage, there has been a unique fusion with the SUI1 domain, producing the ligatin-related family of putative novel translation factors. Finally, in some bacterial lineages, e.g., the spirochaetes, the PUA domain has been deleted from the glutamate kinase.

Conclusions

Computer-aided analyses of complete proteomes enabled us to identify two previously undetected but highly conserved families of likely RNA-binding domains, namely, the S4 and PUA domains, to identify novel domain architectures in a variety of proteins associated with the translation machinery and to predict new translation regulators and RNA modification enzymes. We further demonstrated that SUI1, previously identified only as a stand-alone translation factor, is yet another mobile RNA-binding domain.

The translation system is highly conserved in evolution and is generally considered to be the paragon of evolutionary stability (Dennis 1997; Tatusov et al. 1997; Kyrpides and Woese 1998). Recent phylogenetic studies, however, have revealed considerable complexity in the evolution of aminoacyl-tRNA synthetases, apparently including multiple events of horizontal gene transfer and lineage-specific gene loss (Ibba et al. 1997a, b; Koonin and Aravind 1998). These studies are complemented by complex evolutionary scenarios for RNA-binding domains suggested by the present analysis.

Note Added at Proof

While this manuscript was being processed for publication, the NMR and crystal structures of the ribosomal protein S4 from *Bacillus stearothermophilus* have been published [Markus MA, Gerstner RB, Draper DE, Torchia DA (1998)]. The solution structure of ribosomal protein S4 δ41 reveals two subdomains and a positively charged surface that may interact with RNA [EMBO J 17:4559–4571; Davies C, Gerstner RB, Draper DE, Ramakrishnan V, White SW (1998)]. The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: one domain shows structural homology to the ETS DNA-binding motif [EMBO J 17:545–558]. The N-terminal boundary of the conserved S4 domain shown in our Fig. 1 precisely corresponds to the beginning of domain 2 in both structures. The C-terminal boundary of the conserved domain, however, maps between α6 and β4 in domain which shows that the C-terminal β-hairpin of this domain is not required for the formation of a compact and functional structure. The conserved domain consists of 3 α-helices (α4, α5 and α6) and 3 β-strands (β1, β2 and β3). All the secondary structure elements were identified correctly and almost precisely by the alignment-based prediction (Fig. 1), with the exception of β2 that consists of only 2 residues and was not predicted. Markus and co-workers notice the sequence similarity between domain 2, the C-terminal region of bacterial tyrosyl-tRNA synthetases and the *E. coli* YrfH protein. Additionally, however, they claim a similarity to a region of viral reverse transcriptases that could not be corroborated by our analysis and is, in our opinion, spurious.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new gen-

eration of protein database search programs. Nucleic Acids Res 25:3389–3402

Baker A, Draper D (1995) Messenger RNA recognition by fragments of ribosomal protein S4. J Biol Chem 270:22939–22945

Becker HF, Motorin Y, Planta RJ, Grosjean H (1997) The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of psi55 in both mitochondrial and cytoplasmic tRNAs. Nucleic Acids Res 25:4493–4499

Bork P, Koonin EV (1994) A P-loop-like motif in a widespread ATP pyrophosphatase domain: Implications for the evolution of sequence motifs and enzyme activity. Proteins 20:347–355

Brick P, Bhat TN, Blow DM (1989) Structure of tyrosyl-tRNA synthetase refined at 2.3 A resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. J Mol Biol 208:83–98

Carter P, Bedouelle H, Winter G (1986) Construction of heterodimer tyrosyl-tRNA synthetase shows $tRNA^{Tyr}$ interacts with both subunits. Proc Natl Acad Sci USA 83:1189–1192

Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG (1997) The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 88:235–242

Dennis PP (1997) Ancient ciphers: translation in Archaea. Cell 89: 1007–1010

Doherty AJ, Serpell LC, Ponting CP (1996) The helix-hairpin-helix DNA-binding motif: A structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res 24:2488–2497

Edmonds CG, Crain PF, Gupta R, Hashizume T, Hocart CH, Kowalak JA, Pomerantz SC, Stetter KO, McCloskey JA (1991) Posttranscriptional modification of tRNA in thermophilic archaea (Archaebacteria). J Bacteriol 173:3138–3148

Gregson JM, Crain PF, Edmonds CG, Gupta R, Hashizume T, Phillipson DW, McCloskey JA (1993) Structure of the archaeal transfer RNA nucleoside G*-15 (2-amino-4,7-dihydro-4-oxo-7-beta-D-ribofuranosyl-1H-pyrrolo[2,3-d]pyrimidine-5-carboximidamide (archaeosine)). J Biol Chem 268:10076–10086

Guez-Ivanier V, Bedouelle H. (1996) Disordered C-terminal domain of tyrosyl transfer-RNA synthetase: Evidence for a folded state. J Mol Biol 255:110–120

Heilek GM, Marusak R, Meares CF, Noller HF (1995) Directed hydroxyl radical probing of 16S rRNA using Fe(II) tethered to ribosomal protein S4. Proc Natl Acad Sci USA 92:1113–1116

Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, Poustka A, Dokal I (1998) X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet 19:32–38

Ibba M, Morgan S, Curnow AW, Pridmore DR, Vothknecht UC, Gardner W, Lin W, Woese CR, Söll D (1997a) A euryarchaeal lysyl-tRNA synthetase: Resemblance to class I synthetases. Science 278: 1119–1122

Ibba M, Bono JL, Rosa PA, Söll D (1997) Archaeal-type lysyl-tRNA synthetase in the lyme disease spirochete *Borrelia burgdorferi*. Proc Natl Acad Sci USA 94:14383–14388

Jakoi ER, Brown AL, Ho YS, Snyderman R (1989) Molecular cloning of the cDNA for ligatin. J Cell Sci 93:227–232

Kagan R, Clarke S (1994) Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes. Arch Biochem Biophys 310:417–427

Kasperaitis MA, Voorma HO, Thomas AA (1995) The amino acid sequence of eukaryotic translation initiation factor 1 and its similarity to yeast initiation factor SUI1. FEBS Lett 365:47–50

Koonin EV (1994) Prediction of an rRNA methyltransferase domain in human tumor-specific nucleolar protein P120. Nucleic Acids Res 22:2476–2478

Koonin EV (1996) Pseudouridine synthases: Four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res 24:2411–2415

Koonin EV, Aravind L (1998) Re-evaluation of translation machinery evolution. Curr Biol 8:R266–R269

Koonin EV, Tatusov RL, Rudd KE (1995) Sequence similarity analysis of *Escherichia coli* proteins: Functional and evolutionary implications. Proc Natl Acad Sci USA 92:11921–11925

Kyrpides N, Woese C (1998) Universally conserved translation initiation factors. Proc Natl Acad Sci USA 95:224–228

Kyrpides NC, Woese CR, Ouzounis CA (1996) KOW: A novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem Sci 21:425–426

Lane BG, Ofengand J, Gray MW (1995) Pseudouridine and O2′-methylated nucleosides. Significance of their selective occurrence in rRNA domains that function in ribosome-catalyzed synthesis of the peptide bonds in proteins. Biochimie 77:7–15

Limbach PA, Crain PF, McCloskey JA (1994) Summary: The modified nucleosides of RNA. Nucleic Acids Res 22:2183–2196

Murzin A (1993) OB(oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. EMBO J 12:861–867

Naranda T, MacMillan E, Donahue T, Hershey J (1996) SUI1/p16 is required for the activity of eukaryotic translation initiation factor 3 in *Saccharomyces cerevisiae.* Mol Cell Biol 16:2307–2313

Navaratnam N, Bhattacharya S, Fujino T, Patel D, Jarmuz AL, Scott J (1995) Evolutionary origins of apoB mRNA editing: Catalysis by a cytidine deaminase that has acquired a novel RNA-binding motif at its active site. Cell 81:187–195

Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. Protein Sci 4:1618–1632

Ofengand J, Bakin A, Wrzesinski J, Nurse K, Lane BG (1995) The pseudouridine residues of ribosomal RNA. Biochem Cell Biol 73: 915–924

Ogura M, Tanaka T (1996) Transcription of *Bacillus subtilis* degR is sigma D dependent and suppressed by multicopy proB through sigma D. J Bacteriol 178:216–222

Romier C, Reuter K, Suck D, Ficner R (1996) Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange. EMBO J 15:2850–2857

Romier C, Meyer JE, Suck D (1997) Slight sequence variations of a common fold explain the substrate specificities of tRNA-guanine transglycosylases from the three kingdoms. FEBS Lett 416:93–98

Rost B, Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72

Rost B, Schneider R, Sander C (1997) Protein fold recognition by prediction-based threading. J Mol Biol 270:471–480

Rottman FM, Bokar JA, Narayan P, Shambaugh ME, Ludwiczak R (1994) N6-Adenosine methylation in mRNA: substrate specificity and enzyme complexity. Biochimie 76:1109–1114

Schnuchel A, Wiltscheck R, Czisch M, Herrler M, Willimsky G, Graumann P, Marahiel MA, Holak TA (1993) Structure in solution of the major cold-shock protein from *Bacillus subtilis.* Nature 364: 169–171

Schuler GD, Altschul SF, Lipman DJ (1991) A workbench for multiple alignment construction and analysis. Proteins 9:180–190

Schwedock JS, Liu C, Leyh TS, Long SR (1994) *Rhizobium meliloti* NodP and NodQ form a multifunctional sulfate-activating complex requiring GTP for activity. J Bacteriol 176:7055–7064

Shamoo Y, Abdul-Manan N, Williams (1995) Multiple RNA binding domains (RBDs) just don't add up. Nucleic Acids Res 23:725–728

Smith HC, Sowden MP (1996) Base-modification mRNA editing through deamination—The good, the bad and the unregulated. Trends Genet 12:418–424

St Johnston D, Brown NH, Gall JG, Jantsch M (1992) A conserved double-stranded RNA-binding domain. Proc Natl Acad Sci USA 89:10979–10983

Tang C, Draper D. (1989) Unusual mRNA pseudoknot structure is recognized by a protein translational repressor. Cell 57:531–536

Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved

segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci USA 91:12091–12095

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

ter Huurne AA, Muir S, van Houten M, van der Zeijst BA, Gaastra W, Kusters JG (1994) Characterization of three putative *Serpulina hyodysenteriae* hemolysins. Microb Pathog 16:269–282

Thompson JD, Higgins DG, Gibson TJ (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

Tollervey D, Kiss T (1997) Function and synthesis of small nucleolar RNAs. Curr Opin Cell Biol 9:337–342

Tzermia M, Katsoulou C, Alexandraki D (1997) Sequence analysis of a 33.2 kb segment from the left arm of yeast chromosome XV reveals eight known genes and ten new open reading frames including homologues of ABC transporters, inositol phosphatases and human expressed sequence tags. Yeast 13:583–589

Watanabe M, Matsuo M, Tanaka S, Akimoto H, Asahi S, Nishimura S, Katze JR, Hashizume T, Crain PF, McCloskey JA, Okada N (1997) Biosynthesis of archaeosine, a novel derivative of 7-deazaguanosine specific to archaeal tRNA, proceeds via a pathway involving base replacement on the tRNA polynucleotide chain. J Biol Chem 272:20146–20151

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–573

Wrzesinski J, Bakin A, Nurse K, Lane BG, Ofengand J (1995a) Purification, cloning, and properties of the 16S RNA pseudouridine 516 synthase from *Escherichia coli.* Biochemistry 34:8904–8913

Wrzesinski J, Nurse K, Bakin A, Lane BG, Ofengand J (1995b) A dual-specificity pseudouridine synthase: An *Escherichia coli* synthase purified and cloned on the basis of its specificity for psi 746 in 23S RNA is also specific for psi32 in tRNA(phe). RNA 1:437–448

Yoon H, Donahue T (1992) The suil suppressor locus in Saccharomyces cerevisiae encodes a translation factor that functions during tRNA(iMet) recognition of the start codon. Mol Cell Biol 12:248–260