Published online 23 December 2009 *Nucleic Acids Research, 2010, Vol. 38, No. 6* ***2089–2098***
doi:10.1093/nar/gkp1170

Structural motifs of the bacterial ribosomal proteins S20, S18 and S16 that contact rRNA present in the eukaryotic ribosomal proteins S25, S26 and S27A, respectively

Alexey A. Malygin and Galina G. Karpova*

Institute for Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090, Russia

Received June 25, 2009; Revised November 27, 2009; Accepted November 29, 2009

ABSTRACT

The majority of constitutive proteins in the bacterial 30S ribosomal subunit have orthologues in Eukarya and Archaea. The eukaryotic counterparts for the remainder (S6, S16, S18 and S20) have not been identified. We assumed that amino acid residues in the ribosomal proteins that contact rRNA are to be constrained in evolution and that the most highly conserved of them are those residues that are involved in forming the secondary protein structure. We aligned the sequences of the bacterial ribosomal proteins from the S20p, S18p and S16p families, which make multiple contacts with rRNA in the *Thermus thermophilus* 30S ribosomal subunit (in contrast to the S6p family), with the sequences of the unassigned eukaryotic small ribosomal subunit protein families. This made it possible to reveal that the conserved structural motifs of S20p, S18p and S16p that contact rRNA in the bacterial ribosome are present in the ribosomal proteins S25e, S26e and S27Ae, respectively. We suggest that ribosomal protein families S20p, S18p and S16p are homologous to the families S25e, S26e and S27Ae, respectively.

INTRODUCTION

Ribosomal proteins (r-proteins) are constitutive components of the ribosome—a complicated ribonucleoprotein, which is a central participant of the translation machinery (1). Each of several dozen r-proteins binds to a definite site in the ribosomal subunit, small or large and participates in the formation of its unique spatial structure (2). Naturally related r-proteins have similar structures, bind to similar sites on the ribosome, and compose one r-protein family. The exact number of the r-protein families varies among the three phylogenetic domains of life and even within domains (3). Bacteria are characterized by a smaller set of r-protein families than Archaea and Eukarya, while differences in the sets of the r-protein families between archaeal and eukaryotic ribosomes are comparatively small (3). Remarkably, several r-protein genes in Archaea are absent in some taxons, providing evidence of reductive evolution of r-proteins within this domain (3).

The r-proteins of families presented in all three phylogenetic domains are in general rather conserved, indicating their indispensable role. The majority of the bacterial r-proteins have orthologues in archaea and eukaryotes, but some, especially in the large subunit, appear to be specific for bacteria (3). Thus, of 19 small subunit r-protein families present in all bacterial species with sequenced genomes, 15 have orthologues among 32 human small subunit r-proteins and only 4 (namely S6p, S16p, S18p and S20p) have not (3). These four families might be truly restricted to bacteria (in other words, they were either gained by bacteria or lost by eukaryotes during the evolution) or they might have diverged so much that their orthologues cannot be identified by normal alignments.

Based on the natural properties of the ribosomal proteins, we devised several criteria for the alignment of r-proteins and made an attempt to find orthologues by analysing the conservation of r-protein–RNA contacts based on the high resolution structure of the thermophilic bacteria *Thermus thermophilus* 30S ribosomal subunit. Our investigation revealed that r-proteins S16p, S18p and S20p are homologous to S27Ae, S26e and S25e, respectively.

MATERIALS AND METHODS

Datasets

The amino acid sequences of r-proteins S16p, S18p and S20p for *Anabaena* sp., *Aquifex aeolicus, Bacillus*

*To whom correspondence should be addressed. Tel: +7 383 363 5140; Fax: +7 383 363 5153; Email: karpova@niboch.nsc.ru *halodurans, Bacillus subtilis, Bifidobacterium longum, Bordetella parapertussis, Chlamydia muridarum, Desulfovibrio vulgaris, Escherichia coli, Helicobacter pylori, Listeria innocua, Mycobacterium leprae, Mycoplasma genitalium, Porphyromonas gingivalis, Rickettsia conorii, Rhodopirellula baltica, Streptococcus pneumoniae, T. thermophilus, Thermotoga maritime, Treponema pallidum, Vibrio cholerae, Wolbachia* sp.; r-proreins S25e, S26e and S27Ae for *Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Dictyostelium discoideum, Drosophila melanogaster, Homo sapiens, Mus musculus, Neurospora crassa, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe*; r-proteins S25e and S27Ae for *Ictalurus punctatus, Solanum lycopersicum, Spodoptera frugiperda*; r-proteins S26e and S27Ae for *Oryza sativa*, r-proteins S25e and S26e for *Ovis aries*; r-protein S26e for *Anopheles gambiae, Brugia pahangi, Cricetus cricetus, Macaca fascicularis, Mustela vison, Octopus vulgaris, Oxytricha nova, Schizophyllum commune, Sus scrofa*; r-protein S25 for *Amaranthus cruentus, Ashbya gossypii, Branchiostoma belcheri, Candida glabrata, Danio rerio, Encephalitozoon cuniculi, Leishmania infantum*; r-protein S27A for *Cavia porcellus, Gallus gallus, Hordeum vulgare, Kluyveromyces lactis, Lupinus albus, Plutella xylostella, and Zea mays* were taken from the UniProtKB database (http://www.uniprot.org).

Protein sequence alignment

The MUSCLE program (4) was used to align sequences of homologous proteins. Routine fitting was performed after the alignment using criteria described in the Results section. Espript 2.2 (5) was used for presentation of the alignments. WebLogo (6) was used for creating a graph of aligned amino acid sequences of r-proteins for 22 bacterial and 22 eukaryotic species. Analysis of the contacts between side chains of r-protein amino acids and rRNA was performed on the *T. thermophilus* 30S ribosomal subunit (7) (PDB number 2J02) using program PyMol (DeLano) (8). Nucleotide conservation in the secondary structure of the small subunit rRNAs for three phylogenetic domains was taken from the Comparative RNA Web site (http://www.rna.icmb.utexas.edu) (9).

RESULTS

Criteria for r-protein orthologues and protein multiple alignment

In elaborating criteria for the search for eukaryotic orthologues in bacterial r-proteins, we started from the following assumptions. First, it is known that r-proteins are located mostly on the periphery of the ribosomal subunit and bound preferably to rRNA at that long extensions of many r-proteins penetrate deeply into the subunit stabilizing its core (2). We assumed that r-protein regions forming contacts with rRNA should be more conserved than regions exposed to solvent. This follows from the supposition that any amino acid mutation resulting in the loss of an rRNA–protein contact (polar, hydrophobic, etc.) weakens the total strength of the protein binding, and accumulation of this kind mutations may cause the loss of protein ability to bind to rRNA resulting in assembly of defective ribosomal subunits. Mutations of the solvent exposed amino acids should not have such a drastic effect, though in this case disorders in binding of ligands by ribosome are possible. Next, any bulky insertion in the r-protein site contacting rRNA disturbs the overall fit of the protein RNA-binding surface to the rRNA and reduces its affinity to the rRNA. Hence, the overall homology of r-proteins has to be determined by the structural homology of their regions facing the rRNA and interacting with it. Finally, since protein regions involved in secondary structure (α-helices and β-strands) formation are usually less prone to mutations than the unstructured parts (10), we infer that the most highly conserved amino acid residues of those contacting the rRNA would be the residues located in regions of secondary structure.

Starting from these assumptions, we have analysed motifs in the structures of bacterial r-proteins S16p, S18p and S20p, which make multiple contacts with rRNA in the 30S ribosomal subunit [protein S6p was not analysed since it has only few contacts with RNA (2)], and sought similar motifs in those eukaryotic r-proteins that lacked orthologues in bacteria. Initially, with these three proteins we compared the sets of amino acid residues contacting rRNA in the 30S ribosomal subunit structures deposited in the RCSB Protein Data Bank by independent researcher groups [PDB numbers 2J02 (7), 2HGP (11), 2ZM6 (unpublished data) and 1VS5 (12)] and ensured that these sets mostly coincided. Therefore, the following analysis and alignment were based on the spatial structure of the *T. thermophilus* 30S ribosomal subunit which has the highest (2.8 Å) resolution (PDB number 2J02) (7) of all ribosome structures in the RCSB Protein Data Bank. We used the following criteria for the potential orthologues: first, the amino acids contacting rRNA have to be conserved among members of the same protein family; second, those amino acids located in α-helices and β-strands that contact the rRNA have to be more conserved than those that are not involved in secondary structures; third, the lengths of the secondary structure elements have to be retained for the members of the same protein family, i.e. these elements should not contain internal insertions or deletions; fourth, conservation of the amino acids that contact rRNA, accompanied by the conservation of the respective rRNA nucleotides.

The search was performed as follows. Initially, a multiple sequence alignment using the algorithm MUSCLE (4) was performed with 10 members of each eukaryotic small subunit r-protein family with unknown homology in bacteria. While each family contains a few dozens of members with known sequences, only those from the most evolutionary distant organisms were chosen for the alignment. The results of the alignment were compared with the results of multiple sequence alignment that was performed by the same manner for 10 bacterial representatives of r-protein families S16p, S18p and S20p, to find similar fragments in the protein sequences. In this analysis, representatives from species with a remote relationship were used, where possible. In the event of matching pairs, five bacterial and five eukaryotic proteins were taken from each r-protein family set and a joint alignment was performed for them. Finally, manual fitting was carried out with the use of the criteria mentioned above, ensuring that the alignment within each phylogenetic domain remained immutable if possible. The final fitting was checked for the preservation of the RNA–protein contacts.

Ribosomal proteins S18p and S26e

R-protein S18p is located close to the platform in the *T. thermophilus* 30S ribosomal subunit (Figure 1A). It has only one extensive secondary structure element representing an α-helix of five turns in the C-terminal part of the molecule (helix α3). Two other α-helices (α1 and α2), in the central part of the protein, are rather short. Together, these helices form a globule that contacts rRNA alongside helix α3. The position of the N-terminal part of the protein was not determined by X-ray analysis. Apparently, this part of the protein lacks secondary structure and is rather flexible. Contacts of S18p with 16S rRNA helices 22 and 23a are formed by the conserved positively charged arginines and lysines located on the one side of helix α3 (Figure 1A); the residue R64 in this helix contacts also helix 26 in 16S rRNA. Remarkably, all these arginines and lysines are situated in helix α3 at a distance of about one turn of the helix (K-x2-R-x3-K-x2-KR-x-R). A search in eukaryotic small subunit r-proteins revealed a similar motif in the C-terminal region of S26e. Alignment of the bacterial S18p and eukaryotic S26e family sequences exhibited a high level of conservation of the positively charged amino acids (Figure 1B). Thus, arginines in the positions 64 and 74 (hereafter the numbering of amino acids corresponds to *T. thermophilus* proteins) are practically invariant, whereas the substitutions of R by K and vice-versa may have happened in the other positions. Moreover, the multiple alignments showed conserved elements in other positions of the protein. Among them are a positively charged tripeptide in the N-terminal part of the protein (the region not resolved by X-ray crystallography) and residues D33 and R54 (substituted by lysine in some proteins). The side group of D33 contacts the peptide group at K35 and this bond seems to be important for the maintenance of the spatial structure in this protein region. The structural role of R54 is not obvious yet. The Logo graph of the multiple alignment of the helix α3 region for S26e from 22 eukaryotes and S18p from the same number of the evolutionary distant bacterial species (Figure 1C) illustrates the overall conservation of this region and the near invariance of R64, R74, K/R71 and R/K72.

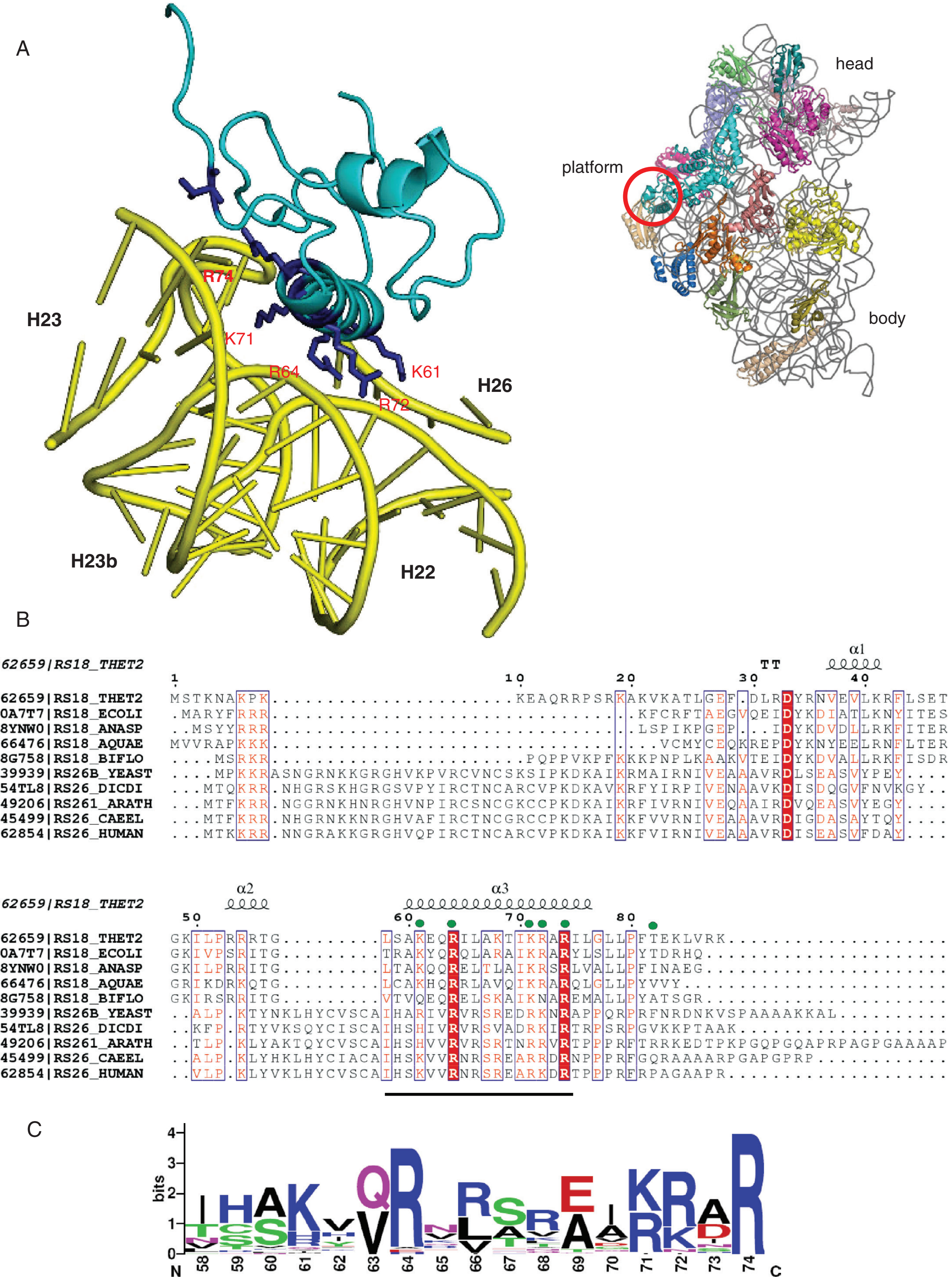
The secondary structure of the small subunit rRNA region formed by helices 22 and 23a is similar in Bacteria and Eukarya and lacks long insertions or deletions (9). Hence, one may suppose that the tertiary structures of this rRNA region in these two phylogenetic domains are also similar. If so, the organization of the RNA–protein contacts in the ribosome region formed by these helices has to be conserved too. Exploration of some of these contacts revealed that nucleotide C720 (the numbering of nucleotides corresponds to *E. coli* 16S rRNA) contacting the side chain of K71 by atom O2 is conserved in 90–98% of Bacteria and Archaea, and in 98–100% of Eukarya [taken from the Comparative RNA website (9)]. Nucleotide C719 is bound to R72 by two hydrogen bonds at the positions O2 and N3, and it is conserved in 90–98% of Bacteria and Archaea and replaced by adenine in 98–100% of Eukarya. Obviously, in the last case, the contact formed by atom N3 in C719 may be replaced by the contact of atom N1 in adenine. Finally, R64 is necessary for clamping helices 22 and 26, because it contacts both backbones. All of these may be evidence that S26e is an orthologue for S18p.

The N-terminal region of S26e has no homologue in S18p but it is nonetheless conserved. The presence of lysines and arginines repeated with an interval of 2–4 amino acids makes this region similar to helix α3 and suggests that this region forms the α-helix that contacts 18S rRNA too.

Ribosomal proteins S16p and S27Ae

R-protein S16p of *T. thermophilus* is located in the lower part of the ‘body’ of the 30S subunit and makes numerous contacts with a number of helices in the 5′-terminal domain of 16S rRNA and helix H21 in the central domain that together form a ‘pocket’ for this protein (Figure 2A). The structure of the protein comprises four anti-parallel β-strands and two adjacent α-helices. Contacts of the protein with the rRNA are made by polar amino acid residues lying basically between the β-strands and in helix α2 (Figure 2B). A search among eukaryotic small subunit r-proteins revealed that r-proteins of family S27Ae are most similar to the S16p family (Figure 2B). Alignment of these two protein families showed that at least 9 of 21 amino acids contacting rRNA in S16p of *T. thermophilus* are conserved. Among these amino acids, one may highlight the following: a cluster of positively charged amino acids positioned in the loop between strands β1 and β2, contacting helices 15 and 21 and clamping them in the ribosome; a tyrosine residue contacting helix 15; arginines 25 and 28 in the loop between strands β2 and β3 contacting helices 7 and 15, respectively; and residues R42 and R47 contacting helix 17. The Logo graph for families S16p and S27Ae for strand β1 and the flanking loop regions shows (Figure 2C) highly conserved amino acids in the positions 12, 13, 28 and invariant Y17.

Of the four helices in the *T. thermophilus* small subunit rRNA that S16p binds substantially (H7, H15, H17 and H21), only helix H15 is conserved in all three phylogenetic domains and, therefore, we analysed the contacts of S16p only to this helix. Atom O2 of nucleotide U375, conserved in 98–100% among all three domains, forms a hydrogen bond with the side chain of R28. The ribose moieties in the same nucleotide and in the adjacent nucleotide A374, which is conserved in 98–100% of Bacteria and 90–98% of Eukarya, but not in Archaea, are bound to Y17. The sugar-phosphate backbone of G391–A393 contacts the side chains of amino acids 8, 12, 13 and 28.



```
P62659|RS18_THET2                                                 TT   α1
                   1        10        20        30        40
P62659|RS18_THET2  MSTKNAKPK......................KEAQRRPSRKAKVKATLGEF.DLRDYRNVEVLKRFLSET
P0A7T7|RS18_ECOLI  .MARYFRRR.................................KFCRFTAEGVQEIDYKDIATLKNYITES
Q8YNW0|RS18_ANASP  ..MSYYRRR.................................LSPIKPGEP...IDYKDVDLLRKFITER
O66476|RS18_AQUAE  MVVRAPKKK...................................VCMYCEQKREPDYKNYEELRNFLTER
Q8G758|RS18_BIFLO  ....MSRKR.......................PQPPVKPFKKKPNPLKAAKVTEIDYKDVALLRKFISDR
P39939|RS26B_YEAST ....MPKKRASNGRNKKGRGHVKPVRCVNCSKSIPKDKAIKRMAIRNIVEAAAVRDLSEASVYPEY....
Q54TL8|RS26_DICDI  ...MTQKRR.NHGRSKHGRGSVPYIRCTNCARCVPKDKAVKRFYIRPIVENAAVKDISDQGVFNVKGY..
P49206|RS261_ARATH ...MTFKRR.NGGRNKHNRGHVNPIRCSNCGKCCPKDKAIKRFIVRNIVEQAAIRDVQEASVYEGY....
O45499|RS26_CAEEL  ...MTFKRR.NHGRNKKNRGHVAFIRCTNCGRCCPKDKAIKKFVVRNIVEAAAVRDIGDASAYTQY....
P62854|RS26_HUMAN  ...MTKKRR.NNGRAKKGRGHVQPIRCTNCARCVPKDKAIKKFVIRNIVEAAAVRDISEASVFDAY....

                      α2                α3
P62659|RS18_THET2
                   50        60        70        80
P62659|RS18_THET2  GKILPRRRTG.........LSAKEQRILAKTIKRARILGLLPFTEKLVRK....................
P0A7T7|RS18_ECOLI  GKIVPSRITG.........TRAKYQRQLARAIKRARYLSLLPYTDRHQ......................
Q8YNW0|RS18_ANASP  GKILPRRITG.........LTAKQQRELTLAIKRSRLVALLPFINAEG......................
O66476|RS18_AQUAE  GRIKDRKQTG.........LCAKHQRRLAVQIKRARQLGLLPYVVVY.......................
Q8G758|RS18_BIFLO  GKIRSRRITG.........VTVQEQRELSKAIKNAREMALLPYATSGR......................
P39939|RS26B_YEAST ..ALP.KTYNKLHYCVSCAIHARIVRVRSREDRKNRAPPQRPRFNRDNKVSPAAAAKKAL..........
Q54TL8|RS26_DICDI  ..KFP.RTYVKSQYCISCAIHSHIVRVRSVADRKIRTRPSRPGVKKPTAAK...................
P49206|RS261_ARATH ..TLP.KLYAKTQYCVSCAIHSHVVRVRSRTNRRVRTPPPRFTRRKEDTPKPGQPGQAPRPAGPGAAAAP
O45499|RS26_CAEEL  ..ALP.KLYHKLHYCIACAIHSKVVRNRSREARRDRNPPPRFGQRAAAARPGAPGPRP............
P62854|RS26_HUMAN  ..VLP.KLYVKLHYCVSCAIHSKVVRNRSREARKDRTPPPRFRPAGAAPR....................
```

Figure 1. Homology between r-protein families S18p and S26e. (**A**, left) RNA binding site of r-protein S18 in the *T. thermophilus* 30S ribosomal subunit. Amino acid side chains contacting the rRNA are shown by sticks and colored more intensively; 16S rRNA helices are marked; conserved amino acids contacting RNA are marked in red, invariant ones in bold (A, right) Structure of the *T. thermophilus* 30S ribosomal subunit with encircled position of protein S18. (**B**) Alignment of five bacterial proteins from the S18p family and five eukaryotic proteins from the S26e family. Positions of elements of the secondary structure of S18 from *T. thermophilus* are shown over its sequence. The numbering corresponds to S18 from the *T. thermophilus* sequence. The S18 residues contacting rRNA in the *T. thermophilus* 30S ribosomal subunit by their side chains are marked by green points. (**C**) WebLogo graph for the underlined portion in (B) of the aligned amino acid sequences of r-proteins for 22 evolutionary distant bacterial and 22 eukaryotic species.

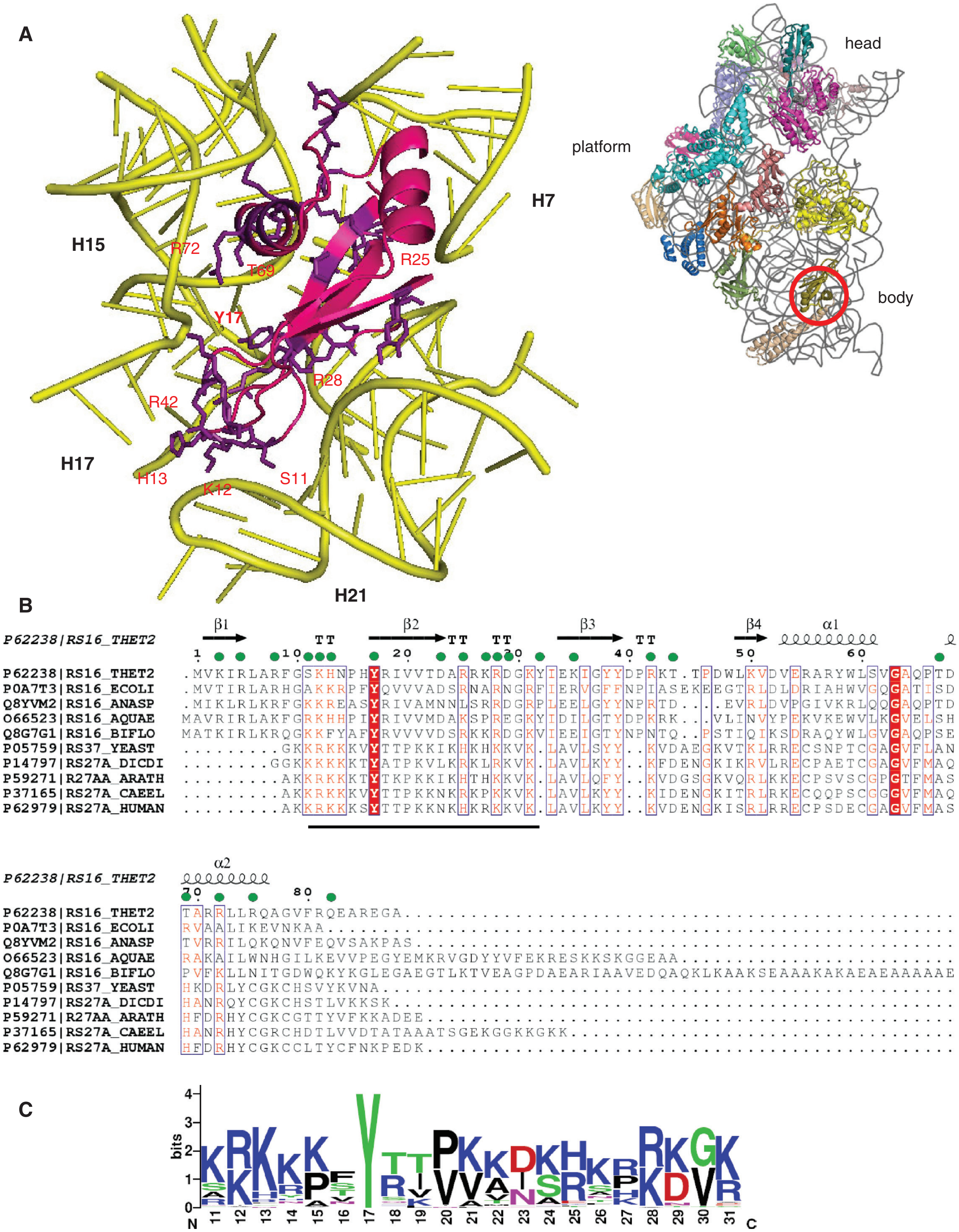

A
H15
H7
H17
H21
R72
T69
R25
Y17
R28
R42
H13
K12
S11
head
platform
body
B
P62238|RS16_THET2
β1
β2
β3
β4
α1
α2
TT
1
10
20
30
40
50
60
70
80
P62238|RS16_THET2 .MVKIRLARFGSKHNPHYRIVVTDARRKRDGKYIEKIGYYDPRKT.TPDWLKVDVERARYWLSVGAQPTD TARRLLRQAGVFRQEAREGA
P0A7T3|RS16_ECOLI .MVTIRLARHGAKKRPFYQVVVADSRNARNGRFIERVGFFNPIASEKEEGTRLDLDRIAHWVGQGATISD RVAALIKEVNKAA
Q8YVM2|RS16_ANASP .MIKLRLKRFGKKREASYRIVAMNNLSRRDGRPLEELGYYNPRTD....EVRLDVPGIVKRLQQGAQPTD TVRRILQKQNVFEQVSAKPAS
O66523|RS16_AQUAE MAVRIRLAKFGRKHHPIYRIVVMDAKSPREGKYIDILGTYDPKRK...VLINVYPEKVKEWVLKGVELSH RAKAILWNHGILKEVVPEGYEMKRVGDYYVFEKRESKKSKGGEAA
Q8G7G1|RS16_BIFLO MATKIRLKRQGKKFYAFYRVVVVDSRKKRDGKVIEEIGTYNPNTQ..PSTIQIKSDRAQYWLGVGAQPSE PVFKLLNITGDWQKYKGLEGAEGTLKTVEAGPDAEARIAAVEDQAQKLKAAKSEAAAKAKAEAEAAAAAE
P05759|RS37_YEAST .........GKKRKKKVYTTPKKIKHKHKKVK.LAVLSYY..KVDAEGKVTKLRRECSNPTCGAGVFLAN HKDRLYCGKCHSVYKVNA
P14797|RS27A_DICDI ........GGKKKKKKTYATPKVLKRKLRKVK.LAVLKYY..KFDENGKIKRVLRECPAETCGAGVFMAQ HANRQYCGKCHSTLVKKSK
P59271|R27AA_ARATH .........AKKRKKKTYTKPKKIKHTHKKVK.LAVLQFY..KVDGSGKVQRLKKECPSVSCGPGTFMAS HFDRHYCGKCGTTYVFKKADEE
P37165|RS27A_CAEEL .........AKKRKKKVYTTPKKNKRKPKKVK.LAVLKYY..KIDENGKITRLRKECQQPSCGGGVFMAQ HANRHYCGRCHDTLVVDTATAAATSGEKGGKKGKK
P62979|RS27A_HUMAN .........AKKRKKKSYTTPKKNKHKRKKVK.LAVLKYY..KVDENGKISRLRRECPSDECGAGVFMAS HFDRHYCGKCCLTYCFNKPEDK
C
bits
4
3
2
1
0
N
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
C

Figure 2. Homology between r-protein families S16p and S27Ae. Notations are as for Figure 1.

It should be noted that the overall length of the r-proteins from S27Ae family is similar to the length of the proteins from the family S16p (with except for S16p of *B. longum*), i.e. the representatives of S27Ae family lack insertions (expansion segments), which are usually characteristic for the eukaryotic orthologues of bacterial r-proteins. At the same time, proteins from S27Ae family lack the region corresponding to the N-terminal fragment in S16p, which forms strand β1 and contacts rRNA helix 15. Divergence between S16p and S27Ae in their central and the C-terminal regions, which in the case of S16p from *T. thermophilus* contact mainly the 16S rRNA helices 7, 17 and 21, correlates with low conservation of these helices. Obviously, the change of rRNA structure in these areas has to lead to the modification of protein structure in the adjacent areas and to the substitution of some RNA–protein contacts by others.

Ribosomal proteins S20p and S25e

The structure of r-protein S20p of *T. thermophilus* comprises three long parallel α-helices, which form a rather elongated globule buried deeply into the 'body' of 30S subunit. The protein makes numerous contacts with helices and loops in the 5′-terminal domain of 16S rRNA and a tip of helix 44 (Figure 3A). Most of the contacts are made by helix α1, whereas helices α3 and especially α2 make poor contacts with the rRNA (Figure 3B). Among the eukaryotic small subunit r-protein families, the S25e family was the best match for the S20p family proteins (Figure 3B). Both of these protein families have several conserved regions. In S20p, there are a cluster of positively charged amino acids in the N-terminus of helix α1 interacting with the base of helix 6 of the 16S rRNA, residue K38 interacting with helix 44 and a cluster of amino acids between helices α2 and α3 located in a very cramped rRNA environment.

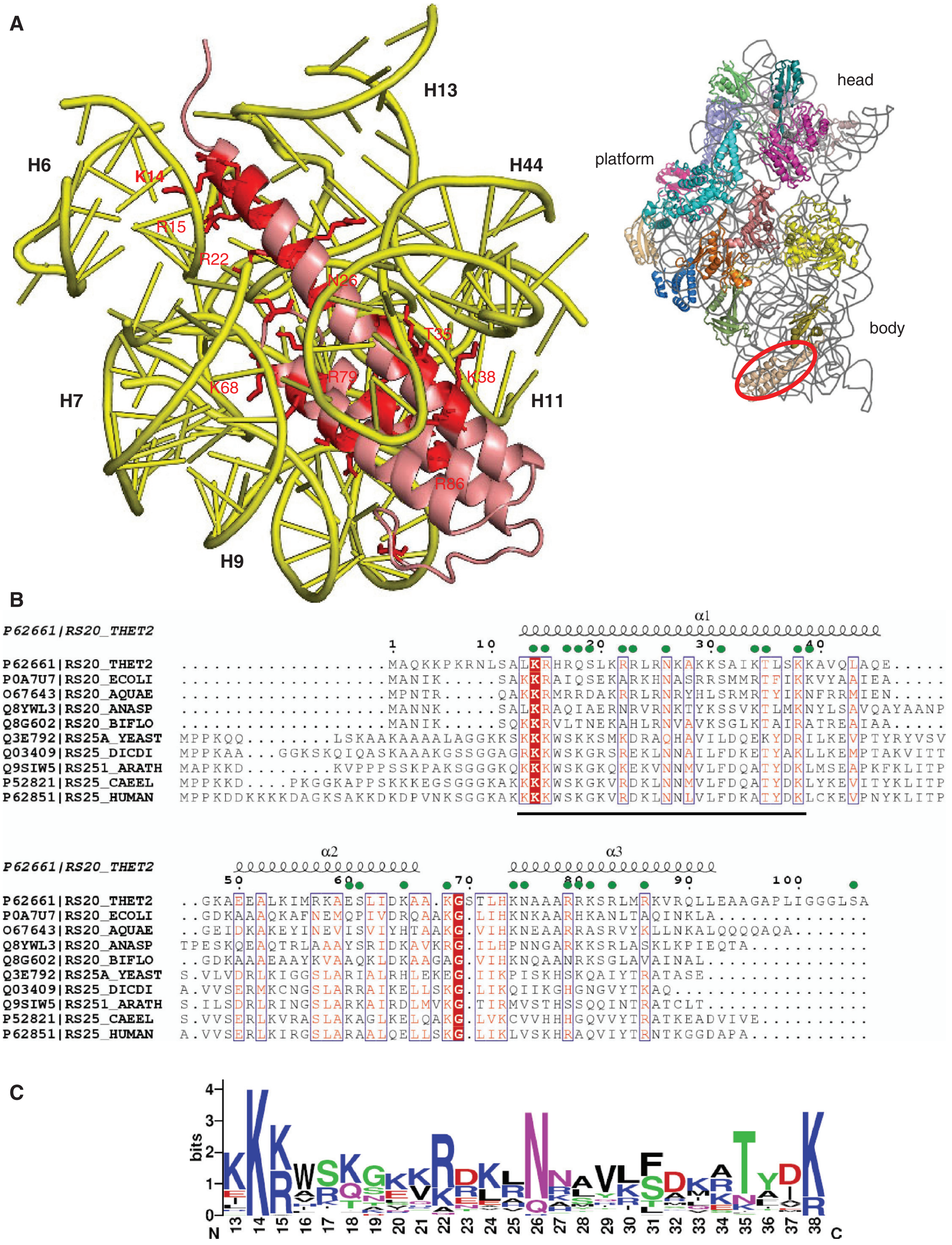
The Logo graph for the alignment of the S20p and S25e families in the region of helix α1 shows (Figure 3C) highly conserved amino acids in positions 15, 22, 26, 35, 38 and an invariant residue, K14. From them, in the 30S ribosomal subunit of *T. thermophilus* residues T35 and K38 contact helix 44, helping to attach it to the 5′-terminal domain of 16S rRNA; R22 and N26 contact the sugar-phosphate backbone at U323 and G324 in the highly conserved loop of helix 13; residue R15 binds the base of G107, which is conserved in 90–98% of Bacteria and Eukarya and in 80–90% of Archaea. Nucleotide G104 (helix 6) binds to K14 via hydrogen bonds at positions O6 and N7, but it is not conserved in Eukarya and Archaea. In other X-ray structure of the *T. thermophilus* 30S ribosomal subunit (11) (PDB number 2HGP), the position of the side chain nitrogen in K14 is closer to position O6 in the nucleotide G105 (3.7 Å) than to position O6 in the nucleotide G104 (4.2 Å). In contrast to G104, nucleotide G105 is conserved in 98–100% of Bacteria and Archaea, and in 90–98% of Eukarya. Therefore, K14 in the 30S ribosomal subunit seems to contact the conserved G105 rather than the non-conserved G104.

As a whole, the r-proteins of S25e family are somewhat longer than the proteins of S20p family due to the N-terminal extension. The total positive charge of this extension might enable it to interact with 18S rRNA in the eukaryotic ribosome. Helix 9 is not conserved, so amino acids in the protein helices α3 and α2 contacting H9 are not conservative too. As in the case of S16p, the divergence of helix 9 obviously led to changes in S20p sequence and structure so that old RNA–protein contacts were lost and new ones arose.

Verification of homologous proteins search approach and validation of protein similarity

To examine our approach for search of homologues among r-proteins, we applied it to the r-proteins from large ribosomal subunit, bacterial L16p and archaeal L10e, as example. Initially, these proteins were not considered as relatives, however, determination of structures of archaeal (13) and bacterial (14,15) large ribosomal subunits revealed that these proteins have a high structural similarity and are located in similar positions of 50S subunits (16) (Supplementary Figure S1A). Currently, both proteins are classified in one family (see the Pfam database, http://pfam.sanger.ac.uk/family?acc = PF00252). Indeed, multiple sequence alignment of L16p and L10e families reveals groups of conserved amino acid residues located predominantly in helices α1 (numbering of structural elements and amino acids is given for L10e from *Haloarcula marismortui*) and α2 and strands β6 and β7 (Supplementary Figure S1B). Among nine almost invariant residues, there are only three (E66, R69 and K159) with polar side chains. Remarkably, K159 in helix α2 (Supplementary Figure S1C) is practically a single residue contacting a nucleotide base but not RNA backbone. This base, C2483 (numbering is given for 23S rRNA from *H. marismortui*), is universal (>98%) in all three domains of life that indicates great importance of this contact for the protein binding. Helix α2 also contains conserved positively charged amino acids important for the protein binding to the 23S rRNA and strongly conserved hydrophobic residues concentrated on the helix side, which forms a hydrophobic contact with helix α1. The remaining amino acid residues contacting backbone of 23S rRNA are arranged in the related regions of the proteins L16p and L10e. Many of these contacts are conserved, some of others are formed by neighbour residues suggesting substitution of one important contact by another. As for invariant residues E66 and R69 in the helix α1, they are exposed outside of the subunit that indicates their possible functional role. This examination clearly indicates that contacting rRNA amino acids residues of homologues r-proteins have a tendency to conservation.

To validate the protein similarity, we determined the amino acid sequence signature specific for each pair of the protein families and scanned the Prosite database (http://expasy.org/prosite) to find how many proteins have these sequences signatures. Sequence signatures were determined from the sequences, which have maximal homology (Figures 1C, 2C and 3C).



```
P62661|RS20_THET2

P62661|RS20_THET2    ...................MAQKKPKRNLSALKRHRQSLKRRLRNKAKKSAIKTLSKKAVQLAQE.....
P0A7U7|RS20_ECOLI    ...................MANIK.....SAKKRAIQSEKARKHNASRRSMMRTFIKKVYAAIEA.....
O67643|RS20_AQUAE    ...................MPNTR.....QAKKRMRRDAKRRLRNRYHLSRMRTYIKNFRRMIEN.....
Q8YWL3|RS20_ANASP    ...................MANNK.....SALKRAQIAERNRVRNKTYKSSVKTLMKNYLSAVQAYAANP
Q8G602|RS20_BIFLO    ...................MANIK.....SQKKRVLTNEKAHLRNVAVKSGLKTAIRATREAIAA.....
Q3E792|RS25A_YEAST   MPPKQQ........LSKAAKAAAALAGGKKSKKKWSKKSMKDRAQHAVILDQEKYDRILKEVPTYRYVSV
Q03409|RS25_DICDI    MPPKAA...GGKSKQIQASKAAAKGSSGGAGRKKWSKGRSREKLNNAILFDKETYAKLLKEMPTAKVITT
Q9SIW5|RS251_ARATH   MAPKKD........KVPPPSSKPAKSGGGKQKKKWSKGKQKEKVNNMVLFDQATYDKLMSEAPKFKLITP
P52821|RS25_CAEEL    MPPKKD....PKGGKAPPSKKKEGSGGGKAKKKKWSKGKVRDKLNNMVLFDQATYDKLYKEVITYKLITP
P62851|RS25_HUMAN    MPPKDDKKKKDAGKSAKKDKDPVNKSGGKAKKKKWSKGKVRDKLNNLVLFDKATYDKLCKEVPNYKLITP

P62661|RS20_THET2

P62661|RS20_THET2    ..GKAEEALKIMRKAESLIDKAA.KGSTLHKNAAARRKSRLMRKVRQLLEAAGAPLIGGGLSA
P0A7U7|RS20_ECOLI    ..GDKAAAQKAFNEMQPIVDRQAAKG.LIHKNKAARHKANLTAQINKLA..............
O67643|RS20_AQUAE    ..GEIDKAKEYINEVISVIYHTAAKG.VIHKNEAARRASRVYKLLNKALQQQQAQA.......
Q8YWL3|RS20_ANASP    TPESKQEAQTRLAAAYSRIDKAVKRG.ILHPNNGARKKSRLASKLKPIEQTA...........
Q8G602|RS20_BIFLO    ..GDKAAAEAAYKVAAQKLDKAAGAG.VIHKNQAANRKSGLAVAINAL...............
Q3E792|RS25A_YEAST   S.VLVDRLKIGGSLARIALRHLEKEG.IIKPISKHSKQAIYTRATASE...............
Q03409|RS25_DICDI    A.VVSERMKCNGSLARRAIKELLSKG.LIKQIIKGHGNGVYTKAQ..................
Q9SIW5|RS251_ARATH   S.ILSDRLRINGSLARKAIRDLMVKG.TIRMVSTHSSQQINTRATCLT...............
P52821|RS25_CAEEL    S.VVSERLKVRASLAKAGLKELQAKG.LVKCVVHHHGQVVYTRATKEADVIVE..........
P62851|RS25_HUMAN    A.VVSERLKIRGSLARAALQELLSKG.LIKLVSKHRAQVIYTRNTKGGDAPA...........
```

Figure 3. Homology between r-protein families S20p and S25e. Notations are as for Figure 1.

The scanning revealed that specific for S18p/S26e pair signature [KRHQ]-x-[QV]-[RK]-x-[RLV]-x(2)-[EA]-x-[RK]-[RK]-x-R presents in 566 proteins, 537 of them belong to S18p and S26e families. The signature [KRSA]-[RK]-[KRH]-x(3)-Y-x(2)-[PV]-x(4)-[HR]-x(2)-[RK]-x(2)-[RK] was found in 381 proteins, and only one from them did not belong to S16p or S27Ae families. The signature K-[KR]-x(6)-[RKAN]-x(3)-[NQ]-x(8)-[TNK]-x(2)-[KR]-x(10,40)-[KRAE]-G-x(0,1)-[LIVT]-[LI]-[KRH] was specific for 354 proteins, and only four of them were not S20p or S25e. This suggests that protein similarities found by us are not stochastic.

DISCUSSION

In contrast to bacterial ribosomal subunits, high resolution crystallographic data for eukaryotic ribosomal subunits and the small subunit of an archaeal ribosome are unavailable, so the exact coordinates of the r-proteins S25e, S26e and S27Ae in the ribosome structure are unknown. Cryo-electron microscopy (cryo-EM) provides reliable information about the location in the 40S ribosomal subunit only for those r-proteins whose homology with bacterial r-proteins has been established. Nevertheless, we have tried to verify our findings about the homology of bacterial and eukaryotic r-proteins with the recent cryo-EM map of the mammalian 40S ribosomal subunit at 8.7 Å resolution (17).

Chandramouli *et al.* (17) observed protein electron density sites in the canine 40S ribosomal subunit that did not correlate with any known r-protein. Among them, the protein density sites denoted by the authors as S-X, S-IV and S-VII correspond approximately to the sites of the r-proteins S16p, S18p and S20p, respectively, in the 30S subunit. Thus, protein density near positions of helices 17 and 21, site S-X, may correspond to S27Ae, since the same helices of 16S rRNA neighbour S16p in the 30S subunit; site S-IV is close to helices 23 and 23a (like S18p, which contacts the same 16S rRNA helices in the 30S subunit) and therefore may correspond to S26e; finally, site S-VII is close to helices 9, 11 and 44 (like S20p in the 30S subunit) and may correspond to S25e. Assuming that site S-IV corresponds to S26e, we can speculate that the conserved part of this protein occupies nearly the same position in the 40S subunit as S18p in the 30S subunit, whereas two stretched out the subunit platform α-helices visible on the cryo-EM map may correspond to the long N-terminal extension of S26e.

Several of our observations are consistent also with biochemical data. In the human ribosome, r-protein S26e was found to be the major target for cross-linking to short mRNA analogs bearing a cross-linking group at specific location and phased in the mRNA-binding centre by tRNA cognate to the triplet directed to the P-site (18,19). In these mRNA analogues, the length of the spacer that linked a modifying group and mRNA moiety was 11 Å. Cross-linking to both S26e and 18S rRNA was observed when the derivatized mRNA nucleotide was located in positions +1 to −3 (position +1 corresponds to the first nucleotide in the P-site bound codon), but S26e was the only target for cross-linking if this nucleotide was in positions –4 to –9. The same protein was also cross-linked to mRNAs containing thiouridines in positions –7 to –10 in the 48S/80S initiation complexes (20). It should be noted also that in early studies on bacterial ribosomes with the use of mRNA analog containing thiouridines in its 5′-half bound to the subunit without phasing, protein S18p was found among cross-linked proteins (21). All these suggest that S26e, like S18p, is located close to or directly on the platform of the small ribosomal subunit and participates in the formation of the site of binding of mRNA upstream of the E-site codon. According to X-ray data, r-protein S18p in the 30S subunit interacts with an mRNA position −15 located upstream of the Shine-Dalgarno sequence (22), this is more distant from the E-site codon than mRNA nucleotides neighboring S26e in the 40S subunit. Therefore, we speculate that the eukaryote specific N-terminal part of S26e, rich in arginines and lysines, may protrude from its globular part (homologous to S18p) to the region of the 40S subunit neighbouring the area corresponding to the Shine–Dalgarno interaction area in the 30S subunit. This position of the N-terminal part of S26e points to its possible interaction with 5′UTR of mRNAs and thereby to participation of S26e in the translation regulation pathway of specific mRNAs similarly to S18p, which is involved in the interaction with a regulatory element located at the 5′ UTR of *rpsO* mRNA in *E. coli* (23).

Suggested close location of S26e to the platform of the small ribosomal subunit may also imply participation of this protein in translation initiation through its interaction with eukaryotic translation initiation factors bound to the subunit like S18p participates in binding of the C-terminal domain of IF3 (24). Indeed, S26e was found among proteins capable of cross-linking to eIF3 under 2-iminothiolan treatment of the complex of 40S subunit with this factor (25) that possibly suggests neighbourhood (or even contact) of the protein with eIF3. This conclusion is also supported by data indicated that the same nucleotide positions (–7 and –10) of mRNA bound to 40S subunit in 48S initiation complex contacted both S26e and eIF3d (20).

Archaeal r-proteins as a rule are more similar to their eukaryotic counterparts than bacterial ones. However, S25e and S26e are absent in the majority of known archaeal taxons as a result of reductive evolution (3). The length of archaeal r-proteins from S27Ae family is only about half that of the eukaryotic homologues. One may suggest that proteins in these families in Archaea have felt high mutagenic pressure that caused their high divergence (or even missing). This r-protein divergence seems to be accompanied by changes in the structure of the protein binding sites on rRNA.

The experimental validations of our findings are out of this study. One of possible approaches to such validation could be based on assembly of hybrid 30S ribosomal subunits containing eukaryotic r-proteins replacing their bacterial counterparts, with the following analysis of protein topography and functional activity of the subunits. Nevertheless, similar assembly of hybrid 40S ribosomal subunits is hardly feasible because methods of *in vitro* assembly of 40S subunits are yet unknown and salt-wash protein depletion of the subunits does not result in step-wise dissociation of r-proteins (26).

Based on the data obtained, we may conclude that all constitutive (present in all taxons) r-proteins present in bacterial 30S ribosomal subunit are likely to have orthologues in eukaryotes (while as for S6p, it remains unclear whether it has orthologues in eukaryotes because the protein was not analysed in this work due to its few contacts with RNA). This suggests that all these r-proteins have a substantial significance for the small ribosomal subunit structure and are essential for the ribosome functionality. The comparisons of S25e/S20p, S26e/S18p and S27Ae/S16p suggest that the structural function of r-proteins is to clamp helices of rRNA together. In the absence of r-proteins these helices would be pushed apart by electrostatic repulsion, making the ribosomal structure fluffy, unstable and as a result non-functional. Since the functionally important amino acids in any protein are usually conserved, the r-protein amino acids responsible for contacts with rRNA and the maintenance of ribosome structure are conserved too. Changes in the r-protein sites contacting rRNA have to be associated with changes in the protein binding sites on rRNA. In this respect the presence of additional helices (expansion segments) in eukaryotic rRNA requires the presence of additional r-proteins or expanded regions within the existing r-proteins that will fix these helices in the ribosome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are deeply indebt to I.C. Eperon and D.M. Graifer for critical reading of the article.

FUNDING

Russian Foundation for Basic Research (No 08-04-00593-a to A.M.); program of Presidium of the Russian Academy of Sciences ‘Molecular and cell biology’ (to G.K.). Funding for open access charge: Russian Foundation for Basic Research (No 08-04-00593-a).

Conflict of interest statement. None declared.

REFERENCES

1. Ramakrishnan,V. (2002) Ribosome structure and the mechanism of translation. *Cell*, **108**, 557–572.
2. Brodersen,D.E., Clemons,W.M. Jr, Carter,A.P., Wimberly,B.T. and Ramakrishnan,V. (2002) Crystal structure of the 30 S ribosomal subunit from *Thermus thermophilus*: structure of the proteins and their interactions with 16 S RNA. *J. Mol. Biol.*, **316**, 725–768.
3. Lecompte,O., Ripp,R., Thierry,J.C., Moras,D. and Poch,O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.*, **30**, 5382–5390.
4. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
5. Gouet,P., Courcelle,E., Stuart,D.I. and Metoz,F. (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.
6. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
7. Selmer,M., Dunham,C.M., Murphy,F.V. 4th, Weixlbaumer,A., Petry,S., Kelley,A.C., Weir,J.R. and Ramakrishnan,V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935–1942.
8. DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.
9. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D’Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
10. Lesk,A.M. (2004) *Introduction to Protein Science: Architecture, Function, and Genomics*. Oxford University Press, Oxford.
11. Yusupova,G., Jenner,L., Rees,B., Moras,D. and Yusupov,M. (2006) Structural basis for messenger RNA movement on the ribosome. *Nature*, **444**, 391–394.
12. Schuwirth,B.S., Day,J.M., Hau,C.W., Janssen,G.R., Dahlberg,A.E., Cate,J.H. and Vila-Sanjurjo,A. (2006) Structural analysis of kasugamycin inhibition of translation. *Nat. Struct. Mol. Biol.*, **13**, 879–886.
13. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
14. Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F. and Yonath,A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679–688.
15. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.
16. Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Bartels,H., Agmon,I. and Yonath,A. (2002) Protein structure: experimental and theoretical aspects. *FEBS Lett.*, **525**, 76–78.
17. Chandramouli,P., Topf,M., Menetret,J.F., Eswar,N., Cannone,J.J., Gutell,R.R., Sali,A. and Akey,C.W. (2008) Structure of the mammalian 80S ribosome at 8.7 A resolution. *Structure*, **16**, 535–548.
18. Graifer,D., Molotkov,M., Styazhkina,V., Demeshkina,N., Bulygin,K., Eremina,A., Ivanov,A., Laletina,E., Ven’yaminova,A. and Karpova,G. (2004) Variable and conserved elements of human ribosomes surrounding the mRNA at the decoding and upstream sites. *Nucleic Acids Res.*, **32**, 3282–3293.
19. Malygin,A.A., Graifer,D.M., Bulygin,K.N., Zenkova,M.A., Yamkovoy,V.I., Stahl,J. and Karpova,G.G. (1994) Arrangement of mRNA at the decoding site of human ribosomes. 18S rRNA nucleotides and ribosomal proteins cross-linked to oligouridylate derivatives with alkylating groups at either the 3′ or the 5′ termini. *Eur. J. Biochem.*, **226**, 715–723.
20. Pisarev,A.V., Kolupaeva,V.G., Yusupov,M.M., Hellen,C.U. and Pestova,T.V. (2008) Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. *EMBO J.*, **27**, 1609–1621.
21. Dontsova,O., Kopylov,A. and Brimacombe,R. (1991) The location of mRNA in the ribosomal 30S initiation complex; site-directed cross-linking of mRNA analogues carrying several photo-reactive labels simultaneously on either side of the AUG start codon. *EMBO J.*, **10**, 2613–2620.
22. Yusupova,G.Z., Yusupov,M.M., Cate,J.H. and Noller,H.F. (2001) The path of messenger RNA through the ribosome. *Cell*, **106**, 233–241.

23. Marzi,S., Myasnikov,A.G., Serganov,A., Ehresmann,Ch., Romby,P., Yusupov,M. and Klaholz,B.P. (2007) Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. *Cell*, **130**, 1019–1031.
24. Pioletti,M., Schlünzen,F., Harms,J., Zarivach,R., Glühmann,M., Avila,H., Bashan,A., Bartels,H., Auerbach,T., Jacobi,C. *et al.* (2001) Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. *EMBO J.*, **20**, 1829–1839.
25. Tolan,D.R., Hershey,J.W. and Traut,R.T. (1983) Crosslinking of eukaryotic initiation factor eIF3 to the 40S ribosomal subunit from rabbit reticulocytes. *Biochimie*, **65**, 427–436.
26. Malygin,A.A., Shaulo,D.D. and Karpova,G.G. (2000) Proteins S7, S10, S16 and S19 of the human 40S ribosomal subunit are most resistant to dissociation by salt. *Biochim. Biophys. Acta*, **1494**, 213–216.