Sequence and evolutionary analysis of bacterial ribosomal S1 proteins

Evgenia Deryusheva[1], Andrey Machulin[2], Maxim Matyunin[3], and Oxana Galzitskaya[4]

[1]Institute for Biological Instrumentation, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences"
[2]Institute for Biological Instrumentation, Russian Academy of Sciences
[3]3 Institute of Protein Research, Russian Academy of Sciences
[4]Institute of Protein Research

November 19, 2020

Abstract

The multi-domain bacterial S1 protein is the largest and most functionally important ribosomal protein of the 30S subunit, which interacts with both mRNA and proteins. The family of ribosomal S1 proteins differs in the classical sense from a protein with tandem repeats and has a "bead-on-string" organization, where each repeat is folded into a globular domain. Based on our recent data, the study of evolutionary relationships for the bacterial phyla will provide evidence for one of the proposed theories of the evolutionary development of proteins with structural repeats: from multiple repeats of assembles to single repeats, or vice versa. In this comparative analysis of 1333 S1 sequences that were identified in 24 different phyla; we demonstrate how such phyla can independently/dependently form during evolution. To our knowledge, this work is the first study of the evolutionary history of bacterial ribosomal S1 proteins. The collected and structured data can be useful to computer biologists as a resource for determining percent identity, amino acid composition and logo motifs, as well as dN/dS ratio in bacterial S1 protein. The obtained research data suggested that the evolutionary development of bacterial ribosomal proteins S1 evolved from multiple assemblies to single repeat. The presented data are integrated into the server, which can be accessed at http://oka.protres.ru:4200.

Sequence and evolutionary analysis of bacterial ribosomal S1 proteins

Running title: Bacterial ribosomal S1 proteins evolution

Evgeniya I. Deryusheva[1], Andrey V. Machulin[2], Maxim A. Matyunin[3] and Oxana V. Galzitskaya[3,4*]

[1] Institute for Biological Instrumentation, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", 142290, Pushchino, Moscow Region, Russia

Skryabin Institute of Biochemistry and Physiology of Microorganisms, Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", 142290, Pushchino, Moscow Region, Russia

Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Moscow Region, Russia

[4]Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142290, Pushchino, Moscow Region, Russia.

*Correspondence to: Oxana Galzitskaya, Laboratory of Bioinformatics and Proteomics, Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russia. E-mail: ogalzit@vega.protres.ru

The multi-domain bacterial S1 protein is the largest and most functionally important ribosomal protein of the 30S subunit, which interacts with both mRNA and proteins. The family of ribosomal S1 proteins differs in the classical sense from a protein with tandem repeats and has a "bead-on-string" organization, where each repeat is folded into a globular domain. Based on our recent data, the study of evolutionary relationships for the bacterial phyla will provide evidence for one of the proposed theories of the evolutionary development of proteins with structural repeats: from multiple repeats of assembles to single repeats, or vice versa. In this comparative analysis of 1333 S1 sequences that were identified in 24 different phyla; we demonstrate how such phyla can independently/dependently form during evolution. To our knowledge, this work is the first study of the evolutionary history of bacterial ribosomal S1 proteins. The collected and structured data can be useful to computer biologists as a resource for determining percent identity, amino acid composition and logo motifs, as well as dN/dS ratio in bacterial S1 protein. The obtained research data suggested that the evolutionary development of bacterial ribosomal proteins S1 evolved from multiple assemblies to single repeat. The presented data are integrated into the server, which can be accessed at*http://oka.protres.ru:4200*.

Keywords: ribosomal S1 proteins, evolutionary analysis, residue conservation, S1 server

Introduction

The discovery of specific signatures of the evolution of bacterial ribosomal proteins is an actual task, which allows a new insight at the emergence and evolution of not only the protein component of ribosomes, but also bacterial evolution [1–5]. The family of ribosomal S1 proteins is a unique family of proteins characterized by a different number of S1 structural domains, each of which has a number of specific characteristics [6,7]. The family of ribosomal proteins S1 makes up about 20% of all bacterial proteins containing the S1 domain [8]. The number of structural S1 domains in bacteria varies within a strictly limited range from one to six[9]. Proteins of this family interact with mRNAs, participate in the initiation and translation of mRNAs *in vivo*and interact with the mRNA-like part of the tmRNA molecule[10,11]. Like some other ribosomal proteins, ribosomal S1 protein is an autogenic repressor of its own synthesis[12]. In addition, S1 can function outside the ribosome.

The role of the separate S1 domains is also being actively studied. Thus, a partial functional specialization of the S1 protein domains of*Escherichia coli* has been identified by many studies. It is known that the first two domains are responsible for binding to ribosomes[13], while the next four are involved in interactions with mRNA. The sixth domain has been shown to be optional for translation initiation [14]. In addition, the first two domains are responsible for the binding of S1 to RNA replicase of the Qb phage, while the sixth domain is not required for its activity in phage replication [15]. The S1 fragment, formed by the third, fourth and fifth domains, increases the activity of ribonuclease RegB of the T4 phage as efficiently as the whole protein [16]. Accordingly, S1 appears to consist of three main regions: the N-terminal region formed from the first and second domains and involved in interaction with other S1 partners in the cell (ribosome, Qb replicase), the intermediate region formed by the third, fourth and fifth domains and involved in the interactions with RNAs (translation or replication initiation region, RegB substrates) and, finally, the sixth domain, the role of which remains to be elucidated [9,17].

Several attempts have been made to classify ribosomal S1 proteins taking into account different numbers of sequences. 13 bacterial phyla were studied by Salah et al. [17]. This work was carried out on 26 bacterial sequences. The authors used the number and pairwise alignment of S1 domains in the family of ribosomal S1 proteins to investigate the relationship between Gram-positive and Gram-negative bacteria. Of the 273 S1 sequences, 12 phyla were identified[18]. The authors of another work [19]used the rpsA gene encoding the ribosomal protein S1 as a biomarker for the main 8 types of mycobacteria, the differences between which were not revealed in the analysis of 16S rDNA.

We have recently shown that the number of domains in S1 is a distinctive characteristic of the phylogenetic grouping of bacteria in the main phyla. The studied data, containing 1453 S1 sequences made it possible to identify bacterial ribosomal S1 proteins in 25 different phyla according to the List of Prokaryotic Names with Standing in Nomenclature. In addition, we searched for a conserved domain in the family of 30S ribosomal S1 protein and hypothesized a possible evolutionary development of the family of 30S ribosomal S1 proteins. The obtained data made it possible to group some bacterial phyla into superphyla according to the number of S1 domains [9].

Here we collect and structured data about features of the family of ribosomal S1 protein and expand and analyze them with data on the percentage identity, amino acid composition and logo motifs, as well as dN/dS ratios. The presented data are integrated in the server, which can be accessed at*http://oka.protres.ru:4200*.

Materials and methods

Construction of ribosomal S1 proteins dataset

A representative dataset of records was selected as described in[9]. The analyzed dataset consists of 1333 records (File S1).

Realization

Algorithms for searching, collecting, presenting and analyzing data were implemented using the freely available programming language Python 3 (https://www.python.org/), implemented in PyCharm v.2018. Various modules from Biopython (a set of biological computing tools written in Python) [20] were used for bioinformatics analysis of bacterial S1 protein sequences.

Taxonomic diversity of bacteria

The bacteria were classified according to the main taxonomic categories (phylum, class, family, genus, type) according to the NCBI Taxonomic database (*http://www.ncbi.nlm.nih.gov/taxonomy*). Gram stain information for studied bacteria was taken from the GOLD: Genomes Online Database (*https://gold.jgi.doe.gov/*)[21].

Number and identification of structural domains in protein sequences

For each analyzed record, the values of the number of S1 domains corresponding to the SMART database (about 1200 domains) were selected[22]. If there was no data on the number of domains in one of the analyzed databases (None), this number was taken equal to zero (these records were deleted from the analyzed dataset). The exact boundaries for each S1 domain for each record were taken from the UniProt database (position, domain, and field of repeats)[23].

Alignment and sequence analysis

Multiple Sequence Alignment was implemented by the MEGA service (*https://www.megasoftware.net*) (ClustalW). In our work, we used the standard parameters of this program. The sequence logos were created by the WebLogo 3 server (*http://weblogo.threeplusone.com*)[24]. To calculate the PID (percent sequence identity), each pair of amino acid sequences was aligned using the Bio.pairwise2.align.globalds function with the Bio.SubsMat.MatrixInfo.blosum62 matrix (Biopython).The PID was calculated as described in [9]. The Bio.codonalign.build and Bio.codonalign.codonseq.cal_dn_ds modules (using the Goldman and Yang model [25]) were used to calculate dN/dS (ratio of non-synonymous to synonymous substitutions). Bio.SeqUtils.ProtParam.ProteinAnalysis was used to count the number of amino acids in a protein sequence.

Web development

We used the Flask web framework (*https://palletsprojects.com/p/flask/*) to develop the websites. MongoDB was used as a document-oriented cross-platform database program (*https://www.mongodb.com*). The Elm language (*https://elm-lang.org*) was used for declarative creation of graphical user interfaces based on a web browser. The S1 server can be accessed at*http://oka.protres.ru:4200*.

Results and discussion

Features of phylogenetic distribution of ribosomal S1 proteins

As mentioned above, automated extended exhaustive analysis of 1453 S1 sequences allowed us to demonstrate that the number of structural domains in S1 is a hallmark for the phylogenetic grouping of bacteria in main phyla [9]. Considering 1453 S1 sequences, we obtained that about 62% of all records were identified as six-domain S1 proteins, which belong to the Proteobacteria phylum. Records with four S1 domains were found in 33% of cases. Almost all analyzed bacteria in this group belong to the Actinobacteria phylum (50% of all four-domain proteins S1) and Firmicutes (47% of all four-domain S1 proteins). Records belonging to these phyla make up 33% of all records. The least represented two-domain proteins S1 make up about 0.6% of all records. S1 proteins, containing one domain, account for only 0.8% of all studied ribosomal S1 proteins. The most represented in this group is the Tenericutes phylum. Cyanobacteria have three S1 domains[9].

For this study, we used a dataset containing 1333 records identified in 24 different bacterial phyla. Phylum is the highest-level group in bacterial domains [26] and is therefore a useful rank for reviewing prokaryotic diversity. Features of the phylogenetic distribution of ribosomal S1 proteins (the number of structural domains, phylum and superphyla) in the studied dataset are shown in Table 1.

The distribution of the number of S1 structural domains in ribosomal S1 proteins and the percentage of representation of various phyla differ little from previous results [9]. 55% of all studied sequences belong to Proteobacteria, 16% and 17% of sequences belong to Firmicutes and Actinobacteria, respectively, and 6% to Bacteroidetes.

Here, for further analysis, we have added the existing in the literature classifications of the studied phyla into supergroups and their division by the Gram staining method (Table 1, File S2). For some phyla, taxonomic classes have been added to Table 1. So, the Bacteroidetes phylum is combined with the Chlorobi phyla, and Fibrobacteres into the FCB group[27]. Our data (Table 1) demonstrate that the ribosomal S1 protein of this group almost always contains six S1 domains (Chlorobi, Fibrobacteres, Gemmatimonadates and Ignavibacteriae). An exception is the Bacteroidetes phylum, which has one, four, or six structural S1domains. However, proteins containing six domains cover 98% of the sequences belonging to this phylum.

Analysis of 16S rRNA and characteristic conserved indels in some proteins is used to group the phyla Planctomycetes, Verrucomicrobia, Chlamydiae into the PVC clan [28]. As our data show (Table 1), bacteria of the phyla Chlamydiae and Verrucomicrobia basically contain six S1 domains, while Planctomycetes can have four, five, and six S1 domains. The proposed superphylum Terrabacteria includes Actinobacteria, Cyanobacteria, Deinococcus-Thermus, and Firmicutes [29,30]. This supergroup unites the two most representative phyla Actinobacteria and Firmicutes, which can contain a different number of structural S1 domains ranging from one to four. In addition, these two phyla are Gram-positive bacteria (G+). Cyanobacteria always have three S1 domains, the Chloroflexi phylum has four S1 domains, and the Deinococcus-Thermus phylum has five S1 domains. It has been suggested that some classes of the phylum Proteobacteria may be a phylum in themselves, which would make Proteobacteria a superphylum [31]. For example, the Deltaproteobacteria group does not always form a monophyletic lineage with other Proteobacteria classes [32]. According to our data, bacterial ribosomal proteins S1 of this phylum can contain a different number of structural S1 domains (from one to six). However, the predominant number of sequences in this group contains six S1 domains (98%).

Bacterial ribosomal S1 proteins biodiversity

The obtained data make it possible to estimate the prevalence of groups containing different numbers of structural S1 domains in the family of the bacterial S1 proteins. Thus, one-, two-, three-, and five-domain S1 proteins account for 1%, 0.8%, 2% and 1.2% of all studied sequences, respectively. Four- and six domain proteins are most represented: 33% and 62%, respectively (Fig. 1a.). At the same time, as we showed above, 55% of all studied bacterial S1 sequences belong to the Proteobacteria, 16% and 17% belong to Firmicutes and Actinobacteria, respectively, and 6% to Bacteroidetes (Fig. 1b.).

Numerous studies showed that >88% of all bacterial isolates belong to four phyla of bacteria (Big Four):

Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes [33,34](Fig. 1c.). In fact, obtaining isolates that do not belong to the Big Four is challenging, and therefore these four phyla dominate our current understanding of microbiology [34]. At the same time, the number of microorganisms belonging to the Proteobacteria phyla was highlighted in most studies determining the diversity of microorganisms, with a range from 40 to 90%, either for isolation analysis or for analysis of microbiomes [35–37]. In general, the dataset we study reflects the percentage of major bacterial phyla and can be considered representative. The most representative groups are the six-domain containing proteins S1 from Proteobacteria and Bacteroidetes and the four-domain containing S1 proteins from Actinobacteria and Firmicutes.

Percent Sequence Identity of bacterial ribosomal S1 proteins

For all available bacterial ribosomal S1 sequences, we collected and calculated the percent identity for each phyla and group (according to the number of structural domains) and the results of pairwise alignment within the bacterial phyla. The presented data are integrated into the server, which can be accessed at *http://oka.protres.ru:4200*(Analysis PID). An example of a PID analysis of four-domain containing S1 proteins is given in Figure 2a. An example of a PID analysis for five-domain containing S1 proteins within the Deinococcus-Thermus phylum is shown in Figure 2b. The S1 server also allows the user to obtain amino acid sequence information for any records of the dataset by clicking on the appropriate UniProt code (Figure 1c).

As mentioned in [9], a relatively low percent of sequence identity both within individual phyla and between them was revealed by aligning the sequences of bacterial S1 proteins.

For S1 proteins containing one-domain, the highest percent of sequence identity within individual phyla belongs to the Actinobacteria phylum (58%), the smallest, to the Tenericutes phylum (25%). Other phyla in this group (S1 proteins containing one-domain) are mono-representatives. Between taxonomic phyla in this group, the percent of the sequence identity ranges from 10% (for example, Actinobacteria and Bacteroidetes; Tenericutes and Actinobacteria, etc.) to 18% (Bacteroidetes and Firmicutes) (http://oka.protres.ru:4200/protein/5eb71e488886fe5b65803db9/pid). For this group, Actinobacteria *Amycolatopsis vancoresmycina* and*Actinoplanes friuliensis* (24%) have the highest percent of sequence identity. Thus, based on our data, it seems possible to classify the one-domain containing S1 proteins as a unique group of S1 proteins. The uniqueness of such proteins is also mentioned[17].

In all studied phyla, only a few bacteria (0.8% of all sequences, Fig.1) were found containing two S1 domains (some bacteria from the Actinobacteria, Firmicutes, and Proteobacteria phyla). For S1 proteins containing two domains, the highest percent of sequence identity within individual phyla belongs to the Actinobacteria phylum (24 %), the smallest, to the Proteobacteria phylum (15%). Between taxonomic phyla in this group, the percent of sequence identity is 13% (Actinobacteria and Firmicutes; Proteobacteria and Firmicutes) and 15% (Actinobacteria and Proteobacteria) (http://oka.protres.ru:4200/protein/5eb71e4b07439c8b4d90c98a/pid). For this group, Actinobacteria *Streptomyces rimosus* and*Amycolatopsis mediterranei* have the highest percent of sequence identity (85%).

In all cases, the phylum Cyanobacteria (Terrabacteria superphylum) has three S1 domains; also some representatives of the Actinobacteria phyla (G+ Terrabacteria) and Proteobacteria (mono-representative) have three-domain S1 proteins. As a rule, three-domain S1 proteins are identified in 2% of cases (Figure 1). Within the Cyanobacteria phylum the percent of sequence identity is 38%. Between taxonomic phyla in this group, the percent of the sequence identity is 14% for Actinobacteria and Cyanobacteria and 15% for the Actinobacteria and Proteobacteria phyla and for the Proteobacteria and Cyanobacteria phyla. For this group, Cyanobacteria strains *Microcystis aeruginosaTAIHU98* and *Microcystis aeruginosa DIANCHI905* have the highest percent of sequence identity (99%).

Records with four S1 domains were identified in 33% cases of all ribosomal S1 proteins studied. Almost all analyzed bacteria in this group belong to the phyla Actinobacteria (52% of all four-domain proteins S1) and Firmicutes (45% of all four-domain proteins S1) (Figure 2). Phyla Bacteroidetes and Caldiserica are mono-representatives in this group. For S1 proteins containing four domains, the highest percent of sequence identity within individual phyla belongs to the Actinobacteria phylum (64%), the smallest to the Chloroflexi phylum (27%). Between taxonomic phyla in this group, the percent of sequence identity varies from 14% (for example, Proteobacteria and Planctomycetes) to 23% (Actinobacteria and Firmicutes). For this group, Actinobacteria strains *Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)* and *Mycobacterium tuberculosis (strain CDC 1551 / Oshkosh)* have the highest percent of sequence identity (100%).

Bacteria of the monotypic (consisting of one Deinococci class) phylum Deinococcus-Thermus, have always five S1 domains. Five S1 domains are also found in bacteria of the Synergistetes, Haloplasmatales, Verrucomicrobia, Proteobacteria, Planctomycetes and Chlamydiae phyla. As a rule, five-domain S1 proteins make up 1.2% of all studied ribosomal S1 proteins (Fig.1). For S1 proteins containing five domains, the highest percent of sequence identity within individual phyla belongs to the Deinococcus-Thermus phylum (53%) and the smallest to the Proteobacteria phylum (26%). Between taxonomic phyla in this group, the percent of sequence identity ranges from 14% (for example, Planctomycetes and Haloplasmatales) to 31% (Verrucomicrobia and Chlamydiae) (http://oka.protres.ru:4200/protein/5eb71e57aab9de6c16cb9d0d/pid). For this group, the phylum Deinococcus-Thermus *Thermus parvatiensis*and *Thermus thermophiles* have the highest percent of sequence identity (97%).

About 62% of the records were identified as proteins containing six S1 domains (Figure 1). As a rule, these proteins belong to the Proteobacteria phylum (86% of all six-domain S1 proteins). The distribution of taxonomic classes within the phylum Proteobacteria is 23%, 15%, 55%, 3%, and 4% for the alpha, beta, gamma, delta, and epsilon classes, respectively. Also ribosomal proteins S1 from bacteria of the phylum Chlorobi (green sulfur bacteria), Acidobacteria, Aquificae, Deferribacteres, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Ignavibacteriae, Nitrospirae, Oligoflexia, Planctomycetes and Verrucomicrobia have six S1 domains. Gram-negative bacteria containing six S1 domains include the Spirochaetes, Bacteroidetes, Chlamidia phyla. For S1 proteins containing six domains, the highest percent of sequence identity within individual phyla belongs to the Chlamidia phylum (69%), the smallest to the Spirochaetes phylum (37%). Between taxonomic phyla in this group, the percent of the sequence identity ranges from 15% (Fusobacteria and Acidobacteria, Fusobacteria and Nitrospirae) to 45% (beta and gamma Proteobacteria). For S1 proteins containing six domains, bacteria from the phylum Chlamydiae *Chlamydia trachomatis* and *Chlamydia muridarum*have the highest percent of sequence identity (95%).

Logo motifs

The logo motifs in each study group (by the number of structural domains) were analyzed with alignment of the S1 protein sequences. An example of Logo motif profiles for S1 proteins containing two domains is shown in Figure 3 (Logo analysis motif, http://oka.protres.ru:4200). Moreover, the user can create logo profiles for full-size sequences using the S1 server.

Some specific patterns were identified when considering the analysis of the S1 protein profiles. Thus, the most conserved are the sequence regions corresponding to β-strands, which correlate with our earlier data on the high conservatism of the secondary structure in such proteins [38]. In addition, an increase in the number of structural S1 domains correlates with an increase in conservatism within each individual domain. Proteins containing five domains are an exception, possibly due to the small sequence representation.

As shown in [9], single domain S1 proteins have a not very high percentage of identity with each other (27%). The strict presence of conserved residues F19, F22, H34, N64, and R68[39], which form RNA binding site in other bacterial, archeal, and eukaryotic protein containing the S1 domain[7] was not revealed taking into account analysis of the logo motif of this group (*http://oka.protres.ru:4200/protein/5eb71e488886fe5b65803db9/logo*). For this group, residues F19, F22, and R68 are conserved only for some bacteria. At the same time, as is known, single-domain S1 proteins of parasitic bacteria of the Mollicutes class (the Tenericutes phylum) effectively perform the main RNA-binding function [40]. It is possible that for these bacteria the RNA binding site is formed by specific amino acid residues or the RNA binding mechanism differs from other proteins containing the S1 domain.

The first and second domains in S1 proteins, containing two structural domains, also have a low percentage of identity within domains: 27% and 30%, respectively. The first and the second domains from S1 proteins containing two structural domains have 38% identity, while pairs with the maximum and minimum values of identity have been identified for the remaining domains [9]. For the first domain in this group, F19, F22 and R68 residues of the RNA binding site are conserved. F19 and H34 are conserved residues for the second domain in this group (Figure 3a).

For S1 proteins containing three structural domains, the maximal value of identity was found between the first and third domains (53%) and the minimum value between the first and the second domains (42%). Moreover, the third domain has the maximum percentage of identity (57%) among other domains for this group [9]. For the first domain in this group of bacteria, N64 residue of the RNA-binding site is conserved. N64, R68, and R34 (at the position of the conserved residue H34) seem to form the RNA-binding site of the second domain in the three-domain containing bacterial S1 proteins. F19, H34, and R68 residues are conserved for the third domain. It can be assumed that for this group, the first domain is characterized by a lower degree of RNA binding efficiency.

For S1 proteins containing four structural domains, the maximum identity value was found between the third and fourth domains (78%) and the minimum identity value between the second and third domains. The third domain also has also the maximum percentage of homology (66%) among other domains in this group. F19, F22, H34 and R68 residues are highly conserved for this domain. These residues are also conserved for the fourth S1 domain in this group. For the second domain F22, N64, and R68 residues formed an RNA binding site. For the first domain, only R34 residue (at the position of the conserved H34 residue) is retained.

The third and fourth domains in the group of S1 proteins containing five structural domains have the maximum percentage of identity (66%), while the second and fifth domains have the lowest percentage of identity (43%). In this group, the fourth domain has the maximum percentage of identity among other domains (49%) [9]. The first domain has no specific conserved motif residues; for the second domain, only R68 residue from the RNA-binding site is retained. Despite the small representativeness of the sequence of bacteria of this group, for the remaining three domains F19, F22, H34 and R68 residues apparently form an RNA binding site.

For the most abundant S1 proteins containing six structural domains, as well as, for S1 proteins with four and five domains, the maximum values of identity are determined between the third and fourth domains (71%) and the minimum values are between the first and the second (39%). The third domain has the highest percentage of identity among other domains in this group (68%) [9]. For this domain, the RNA binding site is formed by five residues: F19, L22 (conserved for F22), H34, N64, and R68 (Figure 3b). The first and sixth domains have no specific conserved residues that can form a RNA binding site. For the second domain, F22, N64, and R68 are retained. Four residues, F19, L22 (in the position of the conserved F22 residue), H34 and R68 are specific for the fourth domain in this group (Fig 3b). the obtained data are in a good agreement with the experimental data confirming that cutting off one S1 domain from the C-terminus or two S1 domains from the N-terminus of the protein reduces only the efficiency of the protein functions, but not its functional capabilities [14,41].

Ratio of non-synonymous to synonymous substitutions

For all available sequences of bacterial ribosomal S1 proteins, we calculated the ratio of non-synonymous to synonymous substitutions (using the Goldman and Yang model) for each group (in accordance with the number of structural domains) (dN/dS analysis, http://oka.protres.ru:4200). As is known, the dN/dS ratio is used to assess the balance between neutral mutations, purifying selection and beneficial mutations acting on a set of homologous genes encoding a protein [42]. This ratio measures the strength and mode of natural selection acting on protein genes, with $dN/dS > 1$ indicating positive (adaptive or diversifying) selection, $dN/dS = 1$ indicating neutral evolution, and $dN/dS < 1$ indicating negative (purifying or cleaning) selection. The dN/dS ratio summarizes the evolutionary rates of genes and can be an informative feature, since it can determine which genes are the most (or least) conserved, as well as identify genes that may have gone through periods of adaptive evolution [43]. However, in real data, positive selection does not occur, because such selection is usually observed only in a certain region of the protein (for example: a specific domain) and/or within one branch of phylogeny (some, but not all species)[44–46]. For S1 single domain proteins, negative selection ($dN/dS < 1$) was most often observed. However, for some representatives of the phylum Actinobacteria, positive selection was revealed relative to the phylum Proteobacteria ($dN/dS = 1.46$) and Tenericutes ($dN/dS = 1.26$). Also, a relatively high dN/dS ratio (0.75) was found for the phylum Bacteroidetes relative to the phylum Tenericutes. For two-domain S1 proteins, negative selection ($dN/dS < 1$) also predominates. Relatively high dN/dS ratios were found for *Eubacterium hallii* (Actinobacteria) and*Actinoplanes friuliensis* (Firmicutes) – 0.83 and for the pair*Beggiatoa sp.* (Firmicutes) and *Tyzzerella nexilis*(Proteobacteria). For three-domain and five-domain S1 proteins, negative selection ($dN/dS < 1$) is characteristic of all pairs of the nucleic acid sequence in this group. For the four-domain S1 proteins, for some representatives of the phylum Actinobacteria, a relatively positive selection was revealed (for example, $dN/dS = 1.05$:*Rhodococcus wratislaviensis* and *Parascardovia denticolens*). For other representatives of the four-domain S1 proteins (within and between phyla) negative selection ($dN/dS < 1$) is observed. For six-domain S1 proteins (within and between phyla) only negative selection ($dN/dS < 1$) is observed. For individual S1 domains (within and between bacteria phyla), analysis of the dN/dS ratio, unfortunately, gives ambiguous results that are difficult to interpret.

Amino acid composition

For all available sequences of bacterial ribosomal S1 proteins, we collected and calculated the percentage of amino acid residues for each group (in accordance with the number of structural domains) (Analysis Amino acid composition,*http://oka.protres.ru:4200*). The data are presented in Table 2.

As mentioned above, the basic structural unit of ribosomal S1 proteins is the S1 domain, which is represented by a β-barrel with an additional α-helix between the third and fourth β-strands [47]. Detailed analysis of the position-specific tendencies of amino acids in β-strands [48,49] revealed a predominance of large aromatic residues (Y, F, W) and β-branched amino acids (T, V, I). As can be seen from Table 2, V, I, F and T are specific for structural β-strands in ribosomal proteins S1. A and L are characteristics of an additional α-helix in the structure of the S1 domain. The predominance of 'disorder-promoting' residues E and K is explained by flexible linkers and terminus, as well as a flexible region in the S1 structural domain [7]. As a rule, for the most representative phyla of S1 proteins containing four and six domains, the percentage of different amino acid residues is almost the same, which is associated with the conservatism of the S1 domain. Note that the alignment of the sequences between the individual domains in each group reveals a rather low percentage of identity, indicating that the structure scaffold (S1 domain) is obviously more important for the overall functioning of these proteins [9].

Discussion

Possible evolutionary development of the family of bacterial 30S ribosomal S1 proteins

The problem of understanding the nature of protein repeats, the corresponding functions for each repeat, and their evolution is still unclear. These repeats evolved from a common ancestor, which necessarily contained a single repeat [50]. Some authors suggested that the common ancestor of the family was indeed a single repeat that formed homo-oligomers for effective functional activity[51]. The homo-oligomeric structure of an ancestor may reflect the intrachain repeating structure of its modern homologue, with the exception of its multi-chain character. However, there are examples of homologous multiple repeats, which are formed both from oligomers with single repeats and from one chain of several repeats (Andrade et al., 2001).

For the investigated bacterial proteins, the maximum number of repeats of the S1 domain (six) is sufficient to perform all the necessary functions. The third domain in this group has the highest identity (68%) among other domains. In addition, this domain has the highest identity with the S1 domain from PNPase (*E. coli*) and the S1 domains from S1 single domain proteins (Tenericutes, Mollicutes)[9], and the RNA binding site is formed by five residues: F19, L22, H34, N64, and R68, which once again confirms the uniqueness of this repeat and allows us to consider it as the strongest RNA binding site. Thus, the central part of proteins (third and fourth domains) appears to be vital for the activity and functionality of these proteins. This suggestion is consistent with experimental data. One of the well-studied proteins with six repeats of the S1 domain is the bacterial 30S ribosomal protein S1 from *E. coli* . It was shown that cutting off one S1 domain from the C-terminus or two S1 domains from the N-terminus of the protein reduces only the efficiency of the protein functions, but not its functionality [14,41].

As mentioned above, the Proteobacteria consists of 55% of all proteins S1 (Figure 1b). Within this, phylogenetic classes are represented by a different number of sequences and structural S1 domains (Figure 4). Thus, Acidithiobacillia and Epsilonproteobacteria have six S1 domains, Alpha − and Deltaproteobacteria consist of five or six S1 domains. Note that Epsilonproteobacteria is considered to be the oldest class in this phylum [29,52]. The Oligoflexia class is characterized by the presence of four or six S1 domains; for Beta and Gamma proteobacteria, the number of S1 domains ranges from one to six. Betaproteobacteria are evolutionary most closely related to Gamma-proteobacteria and Acidithiobacillia, and together they make up a taxon called Chromatibacteria [53]. However, the Acidithiobacillales class was previously classified as part of the Gamma-proteobacteria [54]. Our data also confirm the separation of this class into a separate one for a constant number of structural S1 domains (Figure 4). Phylogenetic analyses of various proteins suggest that that Beta-proteobacteria and Gamma-proteobacteria branched out later than most other phyla of Bacteria along with Proteobacteria [55,56].

Alphaproteobacteria branched out at the same time as Deltaproteobacteria[55,56]. Note that these classes have five and six domains, with Beta-proteobacteria and Gamma-proteobacteria having different numbers of S1 domains. According to our data, these classes within Proteobacteria (in addition to the Actinobacteria, Bacteriodites and Firmicutes phyla) have the greatest diversity in the number of S1 domains in comparison with other phyla, where this number constantly or insignificantly changes. The specific relationship of the phylum Aquificae to the Epsilonproteobacteria is supported by the conserved indel signature in inorganic pyrophosphatase, which is uniquely found in the species of the two phyla [57]. In[58], the authors also suggested that Aquificae are closely related to Proteobacteria. This closeness is due to frequent horizontal gene transfer due to common ecological niches. According to our data, bacteria from the phylum Aquificae and class Epsilonproteobacteria have strictly six S1 domains. The evolutionary development of representatives of the Acidobacteria phylum is often considered to be associated with Alphaproteobacteria[59,60] due to the fact that both bacteria belonging to these phyla were associated with a copiotrophic lifestyles[61]. According to our data, the phyla Acidobacteria and the class Alphaproteobacteria have six S1 domains. The evolutionary independent development of such phyla as Caldiserica, Deferribacteres, Fusobacteria, Spirochaetes, Nitrospirae, Nitrospinae/Tectomicrobia is apparently reflected in the constant number of structural S1 domains in these bacteria. Moreover, the phylum Spirochaetes in the literature is considered a phylogenetically ancient and distinct group of microorganisms [62]. This phylum contains six S1 domains (Figure 4).

As mentioned above, the analysis of 16S rRNA and characteristic conserved indels in some proteins is used to group the phyla Planctomycetes, Verrucomicrobia, Chlamydiae in the PVC clan[28]. Bacteria of the Chlamydiae and Verrucomicrobia phyla generally contain six S1 domains, while Planctomycetes can have four, five, and six S1 domains (Figure 4). According to some published data, the genome of organisms of the phylum Planctomycetes, in comparison with other phyla of superphylum PVC, is the largest and most susceptible to evolutionary changes [63]. Phyla Clamydiae and Verrucomicrobia are considered evolutionarily closer to each other [64].

The FCB group is a superphylum of bacteria named after the main member phyla Fibrobacteres, Chlorobi, and Bacteroidetes. Some authors also include the phyla Gemmatimonadates and Ignavibacteriae in this group[27]. It should be noted, that these phyla on phylogenetic trees are often at the same level, while the phylum Fibrobacteres is considered a phylogenetically more ancient group. Our data show that the ribosomal S1 protein in this group almost always contains six S1 domains (constant number for the Gemmatimonadates, Ignavibacteriae, Fibrobacteres, Chlorobi and class Bacteroidia phyla). The class Cytophagia has one, four, and six domains within the phylum Bacteroidetes (Figure 4).

Phylum Bacteroidetes, along with Proteobacteria, Firmicutes, and Actinobacteria, are also among the most common bacterial groups in the rhizosphere [65]. They have been found in soil samples from various locations, including cultivated fields, greenhouse soils, and unexploited areas [66]. Note that for these phyla, the number of structural S1 domains can vary from one to six (Figure 4).

Terrabacteria are a supergroup containing the Actinobacteria, Tenerecutes, and Firmicutes phyla, as well as the Cyanobacteria, Chloroflexi, and Deinococcus-Thermus phyla [29,52]. It is widely accepted that oxygenic photosynthesis devoloped in ancient lineages of Cyanobacterial [67], but very little is known about the nature and evolutionary history of anoxygenic phototrophy, and much of the understanding is based on assumptions and hypotheses based on few existing bacterial taxa, in which this metabolism occurs. However, a number of studies have argued that one of the earliest forms of anoxygenic photosynthesis arose in the Chloroflexi phylum before the invention of oxygenic photosynthesis during the Archean Eon [68,69]. Our data revealed three S1 domains in the phylum Cyanobacteria and four S1 domains in the phylum Chloroflexi. According to another version, the phyla Actinobacteria and Chloroflexi are more evolutionarily close [32]. Note, that Actinobacteria predominantly have four S1 domains. Evolutionary close to the phyla Actinobacteria, Cyanobacteria, Chloroflexi, and Deinococcus-Thermus, and the phylum Firmicutes according to our data, it also predominantly has four S1 domains [70,71]. meanwhile, according to [32,70] the phylum Deinococcus-Thermus (five S1 domains) is more ancient than other phyla in the supergroup Terrabacteria.

Note that the bacterial 30S ribosomal S1 protein from the parasitic bacteria Mollicutes (phylum Tenerecutes) effectively performs the basic functions of RNA binding [40]. There is an assumption in the literature that mycoplasmas (Mollicutes) are a regressive branch of the evolution of some Gram-positive bacteria or Firmicutes[72]. This hypothesis was confirmed experimentally and is considered in two possible variants: all mycoplasmas originate either from a common ancestor with Gram-positive bacteria, or from different bacteria [72]. Based on a comparison of the 16S rRNA oligonucleotide sequences of several species of mycoplasmas and Gram-positive bacteria from the genera Clostridium, Bacillus, Lactobacillus, and Streptococcus, a reasonable assumption was made about their evolutionary relationship with the phylum Firmicutes[73,74]. A more detailed analysis of 16S RNA sequences showed that mycoplasmas are phylogenetically closest to clostridia[75]. In turn, the most likely ancestors of clostridia (Firmicutes) are Gram-positive bacteria with a low G+C content in their DNA. According to our data, the 30S ribosomal S1 protein from the phylum Tenerecutes has one S1 domain.

Summarizing all the above, it can be argued that, firstly, the number of structural S1 domains in bacteria of different phyla may coincide during symbiotic life and secondly, more phylogenetic ancient divisions have a greater number of structural domains (basically six). Moreover, the earlier in the phylogenetic respect the microorganism, the greater the likelihood of decreasing and ranking the number of structural S1 domains in it.

Conclusions

Here we have collected and analyzed the percentage identity, amino acid composition and logo motifs, and the dN/dS ratio in the largest and functionally important ribosome multidomain bacterial S1 protein. The collected and structured data were integrated into the server, which can be accessed at*http://oka.protres.ru:4200*. In this comparative analysis of 1333 S1 protein sequences that have been identified in 24 different phyla, we demonstrate how such phyla can be independently/dependently formed during evolution in accordance with the change in the number of structural RNA-binding S1 domains, the presence of a conserved RNA binding site in each individual domain, as well as the specificity of the amino acid composition and the ratio of non-synonymous and synonymous substitutions. Based on the obtained data, the study of the evolutionary relationships for the bacterial phyla will suggest that for the bacterial ribosomal S1 proteins, the evolutionary development of proteins evolved from multirepeat assemblies to single repeat. At the same time, it is more likely that the terminal S1 domains are “cut off”, while the more conservative central ones are preserved.

Acknowledgements

This research was funded by the Russian Science Foundation, grant number 18-14-00321.

References

1. Yutin N, Puigbò P, Koonin E V, Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. Lespinet O, ed. *PLoS One* . 2012;7(5):e36972. doi:10.1371/journal.pone.0036972

2. Pilla SP, Bahadur RP. Residue conservation elucidates the evolution of r-proteins in ribosomal assembly and function. *Int J Biol Macromol* . 2019;140:323-329. doi:10.1016/j.ijbiomac.2019.08.127

3. Roberts E, Sethi A, Montoya J, Woese CR, Luthey-Schulten Z. Molecular signatures of ribosomal evolution. *Proc Natl Acad Sci U S A* . 2008;105(37):13953-13958. doi:10.1073/pnas.0804861105

4. Lecompte O, Ripp R, Thierry J-C, Moras D, Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res* . 2002;30(24):5382-5390. doi:10.1093/nar/gkf693

5. Ramulu HG, Groussin M, Talla E, Planel R, Daubin V, Brochier-Armanet C. Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol Phylogenet Evol* . 2014;75(1):103-117. doi:10.1016/j.ympev.2014.02.013

6. Machulin A, Deryusheva E, Lobanov M, Galzitskaya O. Repeats in S1 proteins: flexibility and tendency for intrinsic disorder. *Int J Mol Sci* . 2019;20(10):2377. doi:10.3390/ijms20102377

7. Deryusheva EI, Machulin A V., Matyunin MA, Galzitskaya O V. Investigation of the relationship between the S1 domain and its molecular functions derived from studies of the tertiary structure.*Molecules* . 2019;24(20):3681. doi:10.3390/molecules24203681

8. Deryusheva EI, Machulin A V., Selivanova OM, Galzitskaya O V. Taxonomic distribution, repeats, and functions of the S1 domain-containing proteins as members of the OB-fold family.*Proteins Struct Funct Bioinforma* . 2017;85(4):602-613. doi:10.1002/prot.25237

9. Machulin A V, Deryusheva EI, Selivanova OM, Galzitskaya O V. The number of domains in the ribosomal protein S1 as a hallmark of the phylogenetic grouping of bacteria. *PLoS One* . 2019;14(8):e0221370. doi:10.1371/journal.pone.0221370

10. Sørensen MA, Fricke J, Pedersen S. Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. *J Mol Biol* . 1998;280(4):561-569. doi:10.1006/jmbi.1998.1909

11. Okada T, Wower IK, Wower J, Zwieb CW, Kimura M. Contribution of the second OB fold of ribosomal protein S1 from Escherichia coli to the recognition of tmRNA. *Biosci Biotechnol Biochem* . 2004;68(11):2319-2325. doi:10.1271/bbb.68.2319

12. Skouv J, Schnier J, Rasmussen MD, Subramanian AR, Pedersen S. Ribosomal protein S1 of Escherichia coli is the effector for the regulation of its own synthesis. *J Biol Chem* . 1990;265(28):17044-17049. Accessed January 19, 2017. http://www.ncbi.nlm.nih.gov/pubmed/2120211

13. Subramanian AR. Structure and functions of ribosomal protein S1.*Prog Nucleic Acid Res Mol Biol* . 1983;28:101-142. Accessed November 2, 2012. http://www.ncbi.nlm.nih.gov/pubmed/6348874

14. Boni I V, Artamonova VS, Dreyfus M. The last RNA-binding repeat of the Escherichia coli ribosomal protein S1 is specifically involved in autogenous control. *J Bacteriol* . 2000;182(20):5872-5879. doi:10.1128/JB.182.20.5872-5879.2000

15. Guerrier-Takada C, Subramanian AR, Cole PE. The activity of discrete fragments of ribosomal protein S1 in Q beta replicase function. *J Biol Chem* . 1983;258(22):13649-13652. http://www.ncbi.nlm.nih.gov/pubmed/6358207

16. Bisaglia M, Laalami S, Uzan M, Bontems F. Activation of the RegB endoribonuclease by the S1 ribosomal protein is due to cooperation between the S1 four C-terminal modules in a substrate-dependant manner.*J Biol Chem* . 2003;278(17):15261-15271. doi:10.1074/jbc.M212731200

17. Salah P, Bisaglia M, Aliprandi P, Uzan M, Sizun C, Bontems F. Probing the relationship between gram-negative and gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res* . 2009;37(16):5578-5588. doi:10.1093/nar/gkp547

18. Deryusheva EI, Selivanova OM, Serdyuk IN. Loops and repeats in proteins as footprints of molecular evolution. *Biochemistry (Mosc)* . 2012;77(13):1487-1499. doi:10.1134/S000629791213007X

19. Duan H, Liu G, Wang X, et al. Evaluation of the ribosomal protein S1 gene (rpsA) as a novel biomarker for Mycobacterium species identification. *Biomed Res Int* . 2015;2015:271728. doi:10.1155/2015/271728

20. Cock PJA, Antao T, Chang JT, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics.*Bioinformatics* . 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163

21. Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Res* . 2019;47(D1):D649-D659. doi:10.1093/nar/gky977

22. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* . 2018;46(D1):D493-D496. doi:10.1093/nar/gkx922

23. Bateman A, Martin MJ, O'Donovan C, et al. UniProt: A hub for protein information. *Nucleic Acids Res* . 2015;43(D1):D204-D212. doi:10.1093/nar/gku989

24. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* . 2004;14(6):1188-1190. doi:10.1101/gr.849004

25. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* . 1994;11(5):725-736. doi:10.1093/oxfordjournals.molbev.a040153

26. Ludwig W, Klenk H-P. Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. In: *Bergey's Manual® of Systematic Bacteriology* . Springer New York; 2001:49-65. doi:10.1007/978-0-387-21609-6_8

27. Gupta RS. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. *Crit Rev Microbiol* . 2004;30(2):123-143. doi:10.1080/10408410490435133

28. Gupta. Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. *Front Microbiol* . 2012;3:327. doi:10.3389/fmicb.2012.00327

29. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* . 2004;4:44. doi:10.1186/1471-2148-4-44

30. Sekiguchi Y, Ohashi A, Parks DH, Yamauchi T, Tyson GW, Hugenholtz P. First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. *PeerJ* . 2015;3(1):e740. doi:10.7717/peerj.740

31. Yarza P, Yilmaz P, Pruesse E, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* . 2014;12(9):635-645. doi:10.1038/nrmicro3330

32. Hug LA, Baker BJ, Anantharaman K, et al. A new view of the tree of life. *Nat Microbiol* . 2016;1(5):16048. doi:10.1038/nmicrobiol.2016.48

33. Nikolaki S, Tsiamis G. Microbial diversity in the era of omic technologies. *Biomed Res Int* . 2013;2013:958719. doi:10.1155/2013/958719

34. Hugenholtz P. Exploring prokaryotic diversity in the genomic era.*Genome Biol* . 2002;3(2):REVIEWS0003. doi:10.1186/gb-2002-3-2-reviews0003

35. Bertani I, Abbruscato P, Piffanelli P, Subramoni S, Venturi V. Rice bacterial endophytes: Isolation of a collection, identification of beneficial strains and microbiome analysis. *Environ Microbiol Rep* . 2016;8(3):388-398. doi:10.1111/1758-2229.12403

36. Kolton M, Sela N, Elad Y, Cytryn E. Comparative genomic analysis indicates that niche adaptation of terrestrial Flavobacteria is strongly linked to plant glycan metabolism. Robinson DA, ed. *PLoS One* . 2013;8(9):e76704. doi:10.1371/journal.pone.0076704

37. Hartman K, van der Heijden MGA, Roussely-Provent V, Walser J-C, Schlaeppi K. Deciphering composition and function of the root microbiome of a legume plant. *Microbiome* . 2017;5(1):2. doi:10.1186/s40168-016-0220-z

38. Grishin SY, Deryusheva EI, Machulin A V., et al. Amyloidogenic Propensities of Ribosomal S1 Proteins: Bioinformatics Screening and Experimental Checking. *Int J Mol Sci* . 2020;21(15):5199. doi:10.3390/ijms21155199

39. Agrawal V, Kishan KVR. OB-fold: growing bigger with functional consistency. *Curr Protein Pept Sci* . 2003;4(3):195-206. doi:10.2174/1389203033487207

40. Sirand-Pugnet P, Lartigue C, Marenda M, et al. Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome.*PLoS Genet* . 2007;3(5):744-758. doi:10.1371/journal.pgen.0030075

41. Amblar M, Barbas A, Gomez-Puertas P, Arraiano CM. The role of the S1 domain in exoribonucleolytic activity: substrate specificity and multimerization. *RNA* . 2007;13(3):317-327. doi:10.1261/rna.220407

42. Jeffares DC, Tomiczek B, Sojo V, dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol* . 2015;1201:65-90. doi:10.1007/978-1-4939-1438-8_4

43. Kosiol C, Vinar T, da Fonseca RR, et al. Patterns of positive selection in six Mammalian genomes. Schierup MH, ed. *PLoS Genet* . 2008;4(8):e1000144. doi:10.1371/journal.pgen.1000144

44. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* . 2000;15(12):496-503. doi:10.1016/S0169-5347(00)01994-7

45. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A* . 2001;98(5):2509-2514. doi:10.1073/pnas.051605998

46. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution.*Mol Biol Evol* . 2001;18(8):1585-1592. doi:10.1093/oxfordjournals.molbev.a003945

47. Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell* . 1997;88(2):235-242. doi:10.1016/S0092-8674(00)81844-9

48. Bhattacharjee N, Biswas P. Position-specific propensities of amino acids in the β-strand. *BMC Struct Biol* . 2010;10(1):29. doi:10.1186/1472-6807-10-29

49. Richardson JS, Richardson DC. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A* . 2002;99(5):2754-2759. doi:10.1073/pnas.052706099

50. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. *J Struct Biol* . 2001;134(2-3):117-131. doi:10.1006/jsbi.2001.4392

51. Ponting CP, Russell RB. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J Mol Biol* . 2000;302(5):1041-1047. doi:10.1006/jmbi.2000.4087

52. Battistuzzi FU, Hedges SB. A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol* . 2009;26(2):335-343. doi:10.1093/molbev/msn247

53. Robertson L a., Kuenen JG. *The Prokaryotes* . Vol Volume 5. (Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, eds.). Springer New York; 2006. doi:10.1007/0-387-30745-1

54. Williams KP, Kelly DP. Proposal for a new class within the phylum Proteobacteria, Acidithiobacillia classis nov., with the type order Acidithiobacillales, and emended description of the class Gammaproteobacteria. *Int J Syst Evol Microbiol* . 2013;63(PART8):2901-2906. doi:10.1099/ijs.0.049270-0

55. Gupta RS. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev* . 2000;24(4):367-402. doi:10.1111/j.1574-6976.2000.tb00547.x

56. Gupta RS, Sneath PHA. Application of the character compatibility approach to generalized molecular sequence data: Branching order of the proteobacterial subdivisions. *J Mol Evol* . 2007;64(1):90-100. doi:10.1007/s00239-006-0082-2

57. Griffiths E, Gupta RS. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales.*Int Microbiol* . 2004;7(1):41-52. doi:10.2436/im.v7i1.9443

58. Boussau B, Guéguen L, Gouy M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. *BMC Evol Biol* . 2008;8(1):272. doi:10.1186/1471-2148-8-272

59. Kielak AM, Barreto CC, Kowalchuk GA, van Veen JA, Kuramae EE. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front Microbiol* . 2016;7(MAY):744. doi:10.3389/fmicb.2016.00744

60. Spring S, Schulze R, Overmann J, Schleifer K-H. Identification and characterization of ecologically significant prokaryotes in the sediment of freshwater lakes: molecular and cultivation studies. *FEMS Microbiol Rev* . 2000;24(5):573-590. doi:10.1111/j.1574-6976.2000.tb00559.x

61. Smit E, Leeflang P, Gommans S, van den Broek J, van Mil S, Wernars K. Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. *Appl Environ Microbiol* . 2001;67(5):2284-2291. doi:10.1128/AEM.67.5.2284-2291.2001

62. Olson CM, Fikrig E, Anguita J. Host defenses to spirochetes. In:*Clinical Immunology: Principles and Practice: Fourth Edition* . Elsevier Inc.; 2013:338-345. doi:10.1016/B978-0-7234-3691-1.00016-7

63. Kamneva OK, Knight SJ, Liberles DA, Ward NL. Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol Evol* . 2012;4(12):1375-1390. doi:10.1093/gbe/evs113

64. Cho JC, Vergin KL, Morris RM, Giovannoni SJ. Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae.*Environ Microbiol* . 2004;6(6):611-621. doi:10.1111/j.1462-2920.2004.00614.x

65. Mendes R, Garbeva P, Raaijmakers JM. The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol Rev* . 2013;37(5):634-663. doi:10.1111/1574-6976.12028

66. Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G. Environmental and gut bacteroidetes: the food connection. *Front Microbiol* . 2011;2(MAY):93. doi:10.3389/fmicb.2011.00093

67. Shih PM, Hemp J, Ward LM, Matzke NJ, Fischer WW. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* . 2017;15(1):19-29. doi:10.1111/gbi.12200

68. Xiong J. Molecular Evidence for the Early Evolution of Photosynthesis. *Science (80-)* . 2000;289(5485):1724-1730. doi:10.1126/science.289.5485.1724

69. Blankenship RE. Origin and early evolution of photosynthesis.*Photosynth Res* . 1992;33(2):91-111. doi:10.1007/BF00039173

70. Lang JM, Darling AE, Eisen JA. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. Planet PJ, ed. *PLoS One* . 2013;8(4):e62510. doi:10.1371/journal.pone.0062510

71. Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* . 2013;499(7459):431-437. doi:10.1038/nature12352

72. Neimark H. Phylogenetic relationships between mycoplasmas and other prokaryotes. *The mycoplasmas* . 1979;1:43-61. doi:10.1099/00207713-42-2-226

73. Bhugra B, Dybvig K. High-frequency rearrangements in the chromosome of Mycoplasma pulmonis correlate with phenotypic switching. *Mol Microbiol* . 1992;6(9):1149-1154. doi:10.1111/j.1365-2958.1992.tb01553.x

74. Borkhsenius SN, Chernova OA, Chernov VM, Vonsky MS.*Mycoplasmas [in Russian]* . Nauka; 2002.

75. Blanchard A. Ureaplasma urealyticum urease genes; use of a UGA tryptophan codon. *Mol Microbiol* . 1990;4(4):669-676. doi:10.1111/j.1365-2958.1990.tb00636.x

Figure legends

Figure 1. Pie charts showing the distribution of bacterial S1 proteins, **a** in the number of structural S1 domains, **b**S1 protein phylum-level sequences distribution, **c** phylogenetic distribution of bacterial genome projects in progress of 2020 (*https://gold.jgi.doe.gov/statistics*)

Figure 2. (**a**) PID analysis of four-domain containing S1 proteins for bacterial phyla. (**b**) PID analysis for five-domain containing S1 proteins within phylum Deinococcus-Thermus. (**c**) UniProt ID and amino acid sequence for the five-domain containing S1 protein from *Thermus thermophilus* .

Figure 3 . **(a, b)** Logo motif representation for S1 proteins, containing two domains. The conserved residues F14, F17, R66 in the first domain and F17 and H29 in the second domain correspond to the RNA binding site. **(c,d)** Images of the Logo motif for the third and fourth domains of S1 proteins containing six domains. The conserved residues F17, L20, H34, N68 and R70 correspond to the RNA binding site. Logo profiles are generated by the WebLogo 3 server. The positions are given taking into account the shift due to the termini of the structural domains.

Figure 4. Schematic representative of the bacterial phyla, containing different number of structural S1 domains.

Table 1. Features of phylogenetic distribution of ribosomal S1 proteins

	Gram staining method	Supergroups	Phylum	Class	Number of S1 domains	Ratio of sequence number from dataset, %
1	G-		Acidobacteria	Acidobacteriia	6	<1
2	G+	Terrabacteria	Actinobacteria	Actinobacteria	1, 2, 3, 4	18
				Coriobacteriia	4	<1
3	G-		Cyanobacteria	Gloeobacteria	3	<1
				-	3	<1
4	G-		Deinococcus-Thermus	Deinococci	5	<1
5	G-		Chloroflexi	Caldilineae	4	<1
6			Tenericutes	Mollicutes	1	<1
7	G+		Firmicutes	Bacilli	1, 2, 4	15
				Clostridia	2, 4	<1
8			Aquificae	Aquificae	6	<1
9			Caldiserica	Caldisericia	4	<1

	Gram staining method	Supergroups	Phylum	Class	Number of S1 domains	Ratio of sequence number from dataset, %
10		FCB	Chlorobi	Chlorobia	6	<1
11			Fibrobacteres	Fibrobacteria	6	<1
12			Gemmatimonadetes	Gemmatimonadetes	6	<1
13			Ignavibacteriae	Ignavibacteria	6	<1
14	G-		Bacteroidetes	Bacteroidia	6	3
				Cytophagia	1, 4, 6	<1
				Flavobacteriia	6	2
				Sphingobacteriia	6	<1
				-	6	<1
15			Deferribacteres	Deferribacteres	6	<1
16			Fusobacteria	Fusobacteriia	6	<1
17			Nitrospinae/Tectomicrobia	Nitrospinia	6	<1
18			Nitrospirae	Nitrospira	6	<1
		PVC	Planctomycetes	Planctomycetia	4, 6	<1
19				Phycisphaerae	5, 6	<1
20			Verrucomicrobia	Methylacidiphilae	5, 6	<1
21	G-		Chlamydiae	Chlamydiia	5, 6	<1
22	G-	Proteobacteria	Proteobacteria	Alpha proteobacteria	5, 6	12
				Betaproteobacteria	2, 3, 4, 6	8
				Gammaproteobacteria	1, 2, 5, 6	30
				Delta proteobacteria	5, 6	2
				Epsilonproteobacteria	6	2
				Acidithiobacillia	6	<1
				Oligoflexia	4, 6	<1
23	G-		Spirochaetes	Spirochaetia	6	<1
24			Synergistetes	Synergistia	5	<1
Total amount of records: 1333	**Total amount of records:** 1333	**Total amount of records:** 1333	**Total amount of records:** 1333	**Total amount of records:** 1333	**Total amount of records:** 1333	**Total amount of records:** 1333

Table 2. Content of amino acid residues (%) in sequences of S1 ribosomal proteins by the number of structural domains.

Amino acid letter code	Number of structural S1domains	Number of structural S1domains	Number of structural S1dom
	one	two	three
E	9.6	10.17	10.66
K	9.4	7.517	4.799
V	9.1	11.82	8.232
L	8.7	9.618	9.427
I	7.3	6.135	6.005
N	6.6	3.758	3.306
A	6.3	6.467	7.865
G	6.0	7.407	7.681

Amino acid letter code	Number of structural S1domains	Number of structural S1domains	Number of structural S1dor
F	5.9	3.095	3.766
D	5.1	6.135	6.211
S	4.5	6.412	6.188
T	3.9	4.864	4.788
R	3.6	4.367	6.590
P	3.5	3.316	4.282
Q	3.0	2.653	3.995
M	2.3	1.824	2.353
Y	2.0	1.990	1.492
H	1.9	1.271	1.756
W	0.7	0.773	0.367
C	0.5	0.386	0.218

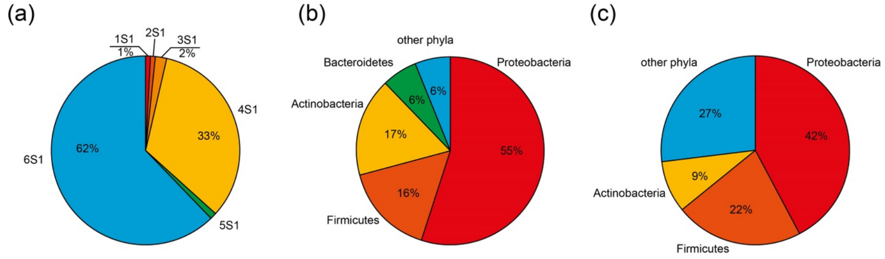
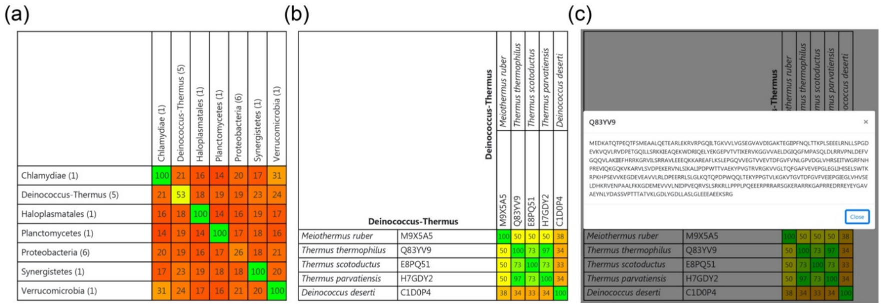


Figure 1.



	Chlamydiae (1)	Deinococcus-Thermus (5)	Haloplasmatales (1)	Planctomycetes (1)	Proteobacteria (6)	Synergistetes (1)	Verrucomicrobia (1)
Chlamydiae (1)	100	21	16	14	20	17	31
Deinococcus-Thermus (5)	21	53	18	19	19	23	24
Haloplasmatales (1)	16	18	100	14	16	19	17
Planctomycetes (1)	14	19	14	100	17	18	16
Proteobacteria (6)	20	19	16	17	26	18	21
Synergistetes (1)	17	23	19	18	18	100	20
Verrucomicrobia (1)	31	24	17	16	21	20	100

Deinococcus-Thermus		Deinococcus-Thermus *Meiothermus ruber* M9X5A5	*Thermus thermophilus* Q83YV9	*Thermus scotoductus* E8PQ51	*Thermus parvatiensis* H7GDY2	*Deinococcus deserti* C1D0P4
Meiothermus ruber	M9X5A5	100	50	50	50	38
Thermus thermophilus	Q83YV9	50	100	73	97	34
Thermus scotoductus	E8PQ51	50	73	100	73	33
Thermus parvatiensis	H7GDY2	50	97	73	100	34
Deinococcus deserti	C1D0P4	38	34	33	34	100

Q83YV9 ×

MEDKATQTPEQTFSMEAALQETEARLEKRVRPGQILTGKVVLVGSEGVAVDIGAKTEGIIPFNQLTTKPLSEEELRNLLSPGD
EVKVQVLRVDPETGQILLSRKKIEAQEKWDRIQELYEKGEPVTVTIKERVKGGVVAELDGIQGFMPASQLDLRRVPNLDEFV
GQQVLAKIIEFHRRKGRVILSRRAVLEEEQKKAREAFLKSLEPGQVVEGTVVEVTDFGVFVNLGPVDGLVHRSEITWGRFNH
PREVIQKGQKVKARVLSVDPEKERVNLSIKALIPDPWTTVAEKYPVGTRVRGKVVGLTQFGAFVEVEPGLEGLIHISELSWTK
RPKHPSEVVKEGDEVEAVVLRLDPEERRLSLGLKQTQPDPWQQLTEKYPPGTVLKGKVTGVTDFGVFVEIEPGIEGLVHVSE
LDHKRVENPAALFKKGDEMEVVVLNIDPVEQRVSLSRKRLLPPPLPQEEERPRRARSGKERARRKGAPRREDRREYEYGAV
AEYNLYDASSVPTTTATVKLGDLYGDLLASLGLEEEAEEKSRG

Close

Meiothermus ruber	M9X5A5	100	50	50	50	38
Thermus thermophilus	Q83YV9	50	100	73	97	34
Thermus scotoductus	E8PQ51	50	73	100	73	33
Thermus parvatiensis	H7GDY2	50	97	73	100	34
Deinococcus deserti	C1D0P4	38	34	33	34	100

Figure 2.

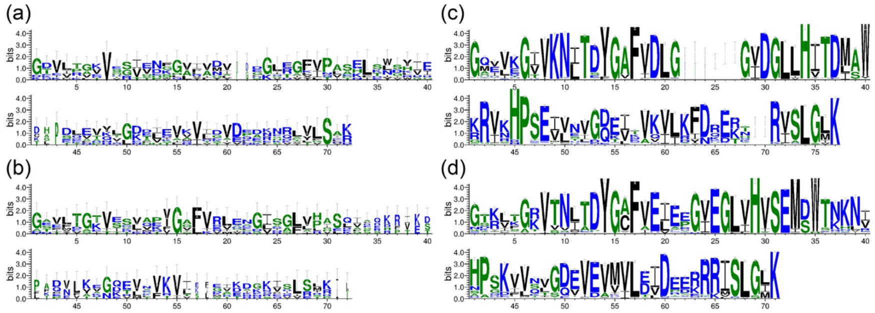


Figure 3.

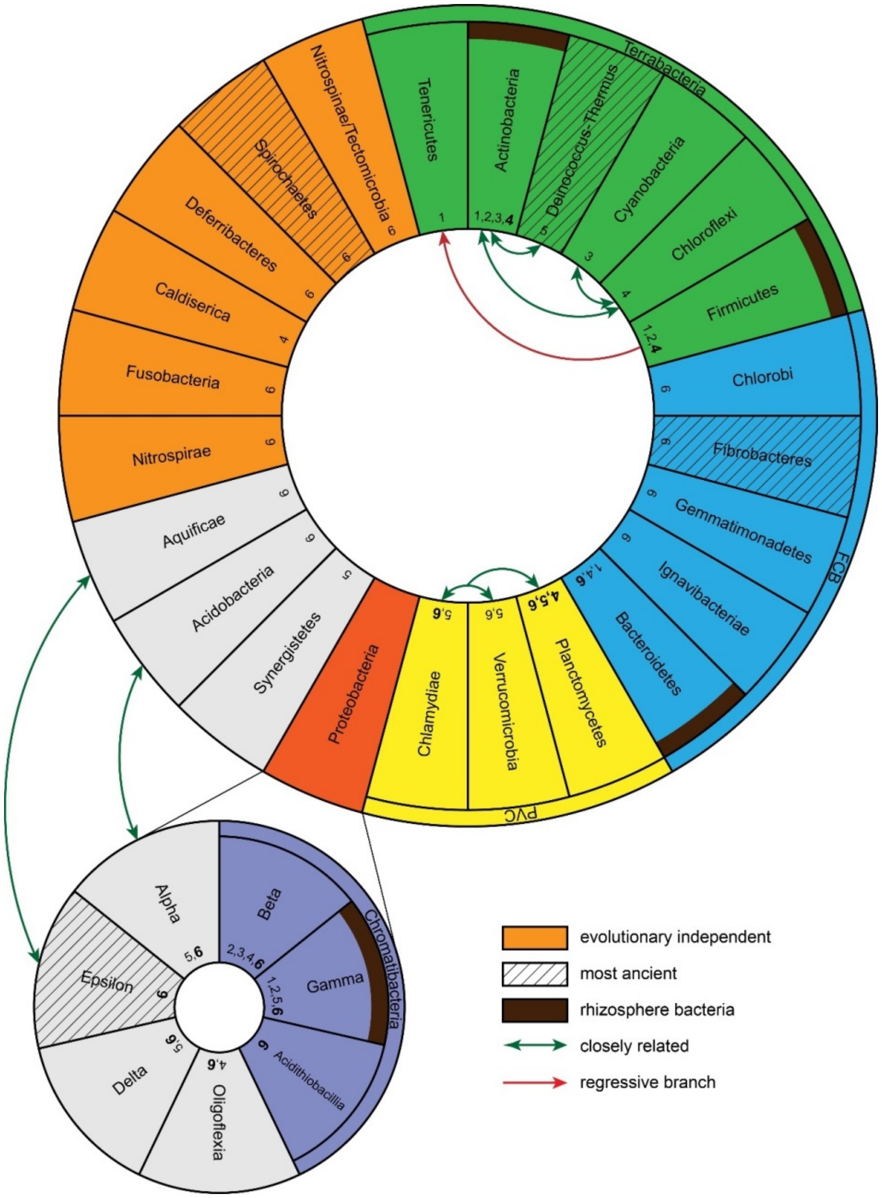


Figure 4.