Genomic Analysis of the Genes Encoding Ribosomal Proteins in Eight Eubacterial Species and *Saccharomyces cerevisiae*

Katsutoshi Fujita[1] **Tomoya Baba**[2] **Katsumi Isono**[1]

fujita@biol.kobe-u.ac.jp baba@staff.or.jp isono@biol.kobe-u.ac.jp

[1] Graduate School of Science and Technology, Kobe University
Rokkodai, Kobe 657-8501, Japan
[2] STAFF, Kamiyokoba, Ippaizuka, Tsukuba 305-0854, Japan

Abstract

The complete genomic nucleotide sequence data of more than 10 unicellular organisms have become available. During the past years, we have been focusing our attention to the analysis of the structure and function of the ribosome and its protein components. By making use of the genomic sequence data, our work can now be extended to comparative analysis of the ribosomal components at the genomic level. Such analysis will contribute to our understanding of the structure-function relationship of the ribosome that is vital to the expression of genetic information. Bearing these in mind, the ribosomal protein genes of organisms whose genomic sequence data are available were analyzed, which included *Aquifex aeolicus; Archaeoglobus fulgidus; Borrelia burgdorferi; Bacillus subtilis; Escherichia coli; Haemophilus influenzae; Helicobacter pylori; Methanococcus jannaschii; Mycoplasma genitalium; Mycoplasma pneumoniae; Synechosystis* sp., and *Saccharomyces cerevisiae.* In addition, the amino acid sequence data of *Bacillus stearothermophilus* ribosomal proteins were used in the evolutionary evaluation. The results indicate that, in eubacteria including two species of *Mycoplasma*, the operon structure of ribosomal protein genes is well conserved, while their relative orientation and chromosomal location are diverged into several classes. The operon structure in *M. jannaschii* on the other hand is quite different from the eubacterial one and we noticed that its many genes show similarity to rat ribosomal protein genes. The degrees of sequence conservation differ from one ribosomal protein gene to another, but several genes encoding proteins that are considered to be of structural importance are conserved throughout the bacterial species including archaebacteria and further in *S. cerevisiae.*

1 Introduction

Recent remarkable progresses in the methodology and instrumentation for performing large scale nucleotide sequencing have resulted in the completion of genomic nucleotide sequencing of more than ten organisms and the data for additional ten or more organisms are expected to become available in the near future. Furthermore, genomic sequencing of organisms that are important in agriculture and medicine is also in progress. The resultant data will drastically affect the ways of performing research in biological sciences. With the vast amount of sequence data available, we can select genes of our interest and characterize them *in silico* before performing actual *in vitro* or *in vivo* experiments. By such cyber-biological analyses, we will be able to predict the features of the genes and their protein products and to determine the way of approaching to their experimental characterization.

Escherichia coli and *Haemophilus influenzae* are rather closely related, both belonging to the same branch within the gamma subdivision of purple bacteria. While *E. coli* K-12 is benign, *H. influenzae* as well as some wild *E. coli* strains are pathogenic, causing diseases of varying degrees of harmfulness. The differences must lie within their individual genomic contents: virulence related factors could be identified by subtracting from *H. influenzae* those genes that have a homolog(s) in *E. coli* [13]. Such analysis would be not only useful for therapeutic purposes but also for the identification of differences among organisms.

There are many structural entities that are involved in the translation of genetic messages within the cell. Of them, the ribosome is a pivotal apparatus which is composed of two subunits of unequal size and contains several RNA and more than 70 different protein molecules. Because of the high degree of functional importance, interaction between ribosomal subunits and their individual components must have been highly conserved during the course of evolution. A mutation in one of the components will affect its local conformation, changing its ability to be assembled with other components, and consequently the function of the whole ribosome may be impaired. Since all ribosomal proteins (r-proteins) of *E. coli* have been extensively characterized and their *in vitro* reconstitution into active ribosomes has been made possible [21], it would be interesting to analyze to what extent we might be able to correlate the evolutionary conservation and structural importance of individual r-proteins. Furthermore, comparative studies of r-protein genes at the genomic level would clarify whether any one or more r-proteins are, either partly or totally, dispensable or not, and if so, what would be a prerequisite for that to occur. As a first step to such an approach, we carried out extensive similarity search of the r-protein genes in the genomic sequence data of nine bacterial species and yeast as will be described below.

2 Materials and Methods

The genomic nucleotide sequence data were retrieved either directly from the DDBJ/EMBL/GenBank nucleic acid sequence databases or through the World Wide Web server at TIGR (http://www.tigr.org/). The amino acid sequence data were retrieved from the SWISSPROT and PIR databases. To compare individual ribosomal proteins, we used the FASTA program [24]. Since r-proteins have not been fully characterized in organisms other than *E. coli*, their genomic nucleotide sequences were translated into amino acid sequences in six reading frames and subjected to FASTA analysis using the r-protein amino acid sequence data of *E. coli* and *Bacillus stearothermophilus* [30]. The FASTA scores obtained were subsequently converted into 'degrees of conservation' by normalizing them with the corresponding 'self-examination' data of *E. coli* and *B. stearothermophilus* r-proteins.

3 Results and Discussion

3.1 Ribosomal protein genes are conserved across kingdoms.

Putative r-proteins were identified in the amino acid sequence data translated from the genomic nucleotide sequences of individual organisms by comparing similarity to the r-protein sequences of *E. coli* and *B. stearothermophilus*. We calculated the degrees of conservation of individual r-proteins as described in Materials and Methods and used them for comparison. The genomic nucleotide sequences of eubacteria, *Bacillus subtilis* [19], *E. coli* [2], *H. influenzae* [5], *Helicobacter pylori* [28], *Mycoplasma genitalium* [6], *Mycoplasma pneumoniae* [10], *Synechosystis* sp. [15], and *Aquifex aeolicus* [4] were analyzed. Of these, *B. subtilis* is classified as Gram-positive, *E. coli*, *H. influenzae* and *H. pylori* as Gram-negative, *A. aeolicus* as an extreme thermophile, and *Synechosystis* as cyanobacteria. In addition to these eubacteria, the data of a spirochete *Borrelia burgdorferi* [7] as well as two archaebacteria, *Methanococcus jannaschii* [3] and *Archaeoglobus fulgidus* [16] and a unicellular fungus *Saccharomyces cerevisiae* [8] were included in the analysis. We treated ORFs whose amino acid translation data showed FASTA scores higher than 100 as putative r-protein genes. The results of analysis with *E. coli* r-proteins are summarized in Table 1. It is readily evident that r-proteins are widely conserved amongst eubacteria analyzed and to a much less extent in archaebacteria and yeast. Some proteins known to be important for ribosomal structure and function such as S4, S5, S8, S10, S12, S13, L1, L2, L5, and L11 are well conserved throughout the bacterial kingdom as expected. Moreover, some of the *E. coli* and/or *B. stearothermophilus* r-proteins show a considerable degree of similarity to the cytoplasmic r-proteins of *S. cerevisiae* as well.

Table 1: Comparison of r-proteins from various organisms [(a)]

Ec	Size	Hi	Bs	Hp	Ss	Bb	Aa	Mg	Mp	Af	Mj	Sc		
Prot	(a.a.)											mit	cyt-1	cyt-2
E-S1	557	+++	+	+	+	+	+	-	-	-	+	-	+	-
E-S2	241	+++	++	++	++	++	++	++	++	+	-	+	-	-
E-S3	233	+++	++	++	++	++	++	++	++	+	+	-	-	-
E-S4	206	+++	++	++	++	++	++	++	++	+	+	-	-	-
E-S5	166	+++	++	++	++	++	++	++	++	+	+	+	+	-
E-S6	135	+++	++	+	+	-	+	+	+	-	-	-	-	-
E-S7	178	+++	++	++	++	++	++	++	++	-	-	+	-	-
E-S8	130	+++	++	++	++	++	+	++	++	+	+	-	-	-
E-S9	128	+++	++	++	++	++	++	++	++	+	-	++	-	-
E-S10	103	+++	+++	+++	++	++	++	++	++	++	++	+	+	-
E-S11	129	+++	+++	++	++	++	++	++	++	++	+	-	+	+
E-S12	124	+++	++	+++	+++	+++	+++	++	++	+	+	++	-	-
E-S13	118	+++	+++	++	++	++	++	++	++	+	+	+	+	+
E-S14	98	+++	++	+	++	+	+	+	+	-	-	+	-	-
E-S15	89	+++	+++	++	++	++	++	++	++	-	-	++	-	-
E-S16	82	+++	++	++	++	++	-	++	-	-	-	++	-	-
E-S17	84	+++	++	++	++	++	++	++	-	+	++	-	++	++
E-S18	74	+++	++	++	++	++	++	++	++	-	-	-	-	-
E-S19	92	+++	+++	++	+++	++	++	++	++	+	+	++	+	-
E-S20	87	+++	++	+	+	+	++	+	-	-	-	-	-	-
E-S21	71	+++	++	++	++	++	++	-	-	-	-	-	-	-
E-S22	45	-	-	-	-	-	-	-	-	-	-	-	-	-
E-L1	234	+++	++	++	++	++	++	++	++	+	+	+	-	-
E-L2	273	+++	++	++	++	++	++	++	++	+	+	+	+	+
E-L3	209	+++	++	+	++	++	++	++	++	+	+	++	-	-
E-L4	201	+++	++	+	++	++	+	+	+	-	-	-	-	-
E-L5	178	+++	+++	++	+++	+++	+++	++	++	+	+	-	+	+
E-L6	176	+++	++	++	++	++	++	++	++	-	+	+	+	-
E-L9	148	+++	++	+	++	++	++	+	+	-	-	-	-	-
E-L10	165	+++	+	-	+	+	+	+	+	-	-	-	-	-
E-L11	142	+++	++	++	+++	++	++	++	++	++	++	++	-	-
E-L12	121	+++	+++	+++	++	++	+++	++	++	-	+	++	-	-
E-L13	142	+++	++	++	++	++	++	++	++	+	-	++	-	-
E-L14	120	+++	++	+++	++	++	++	++	++	+	+	-	+	+
E-L15	144	+++	++	++	++	++	++	++	++	-	+	++	-	-
E-L16	136	+++	++	++	+++	++	++	++	++	-	-	++	-	-
E-L17	127	+++	++	++	++	++	++	+	+	-	-	++	-	-
E-L18	117	+++	++	+	++	++	++	+	+	+	-	-	-	-
E-L19	115	+++	++	++	++	++	++	++	++	+	-	-	-	-
E-L20	118	+++	+++	++	++	++	++	++	++	-	-	-	-	-
E-L21	103	+++	++	++	++	++	++	+	+	-	-	-	-	-
E-L22	110	+++	++	++	++	++	++	++	++	-	+	-	-	-
E-L23	102	+++	+	+	++	++	++	-	-	+	+	-	+	-
E-L24	103	+++	++	+	-	++	++	+	-	+	+	-	+	+
E-L25	94	+++	-	-	-	+	-	-	-	-	-	-	-	-
E-L27	85	+++	++	++	++	++	++	++	++	-	-	++	-	-
E-L28	78	+++	+	-	++	-	-	-	-	-	-	-	-	-
E-L29	63	+++	++	++	-	-	++	++	++	+	++	-	-	-
E-L30	58	+++	++	-	-	++	-	-	-	-	-	+	-	-
E-L31	70	+++	++	++	++	+	++	++	++	-	-	-	-	-
E-L32	57	+++	-	-	-	+	-	++	++	-	-	-	-	-
E-L33	55	+++	++	++	++	++	++	++	++	-	-	++	+	-
E-L34	46	+++	+++	++	++	+++	++	+++	+++	-	-	++	-	-
E-L35	65	+++	++	++	+	++	++	++	++	-	-	+	-	-
E-L36	38	+++	+++	+++	+++	+++	++	+++	+++	-	-	++	-	-

[(a)] Symbols indicate: +, degree of conservation of less than 30%; ++, between 30 and 70%; +++, more than 70%. Abbreviations of organism names are: Aa, *Aquifex aeolicus*; Af, *Archaeoglobus fulgidus*; Bb, *Borrelia burgdorferi*; Bs, *Bacillus subtilis*; Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Mg, *Mycoplasma genitalium*; Mj, *Methnococcus jannaschii*; Mp, *Mycoplasma pneumoniae*; Sc, *Saccharomyces cerevisiae*; Ss, *Synechosystis* sp. Abbreviations mit and cyt, respectively, indicate mitochondrial and cytoplasmic ribosomes.

Protein S22 was identified as an r-protein in *E. coli* [29], but its counterpart could not be identified in other organisms including the most closely related bacterium, *H. influenzae*. S22 may be an *E. coli*-specific r-protein. *H. influenzae* harbors a highly conserved L25 homologue. However, this was not the case with other bacteria except for *B. burgdorferi* which showed the presence of a protein of weak similarity to L25. In *B. subtilis* both S14 and L31 are encoded by two genes which are not identical. Similarly, the gene for S15 is duplicated in *H. influenzae*. More surprisingly, the extreme thermophile, *A. aeolicus*, contains two sets of S7 and S12 genes and a gene which shows a weaker homology to S1 in addition to the one listed in Table 1. Although *A. aeolicus* has been placed closest to the branch point of eubacteria and archaebacteria [4], it appears to a typical eubacteria without any doubt.

Of the highly conserved r-proteins, *E. coli* S3 protein has been shown to interact with 16 S rRNA and was found to be located in the head of the small subunit near the site at which the polypeptide release factor RF-2 binds. Together with S4 and S5, protein S12 participates in a region designated as the recognition complexes [22, 23] (see below). S4 in turn is one of the primary rRNA-binding proteins that initiates assembly of the 30 S subunit and is known to protect several 16 S rRNA regions from chemical modification during assembly. Proteins S8 and S11 interact with a highly conserved region in the central domain of the 16 S rRNA termed the platform ring [23]. S19 interacts with proteins S7, S9 and S14. S10 and S13 are later-assembling proteins and has been shown to rely relatively less on protein-RNA interactions and more on protein-protein interaction.

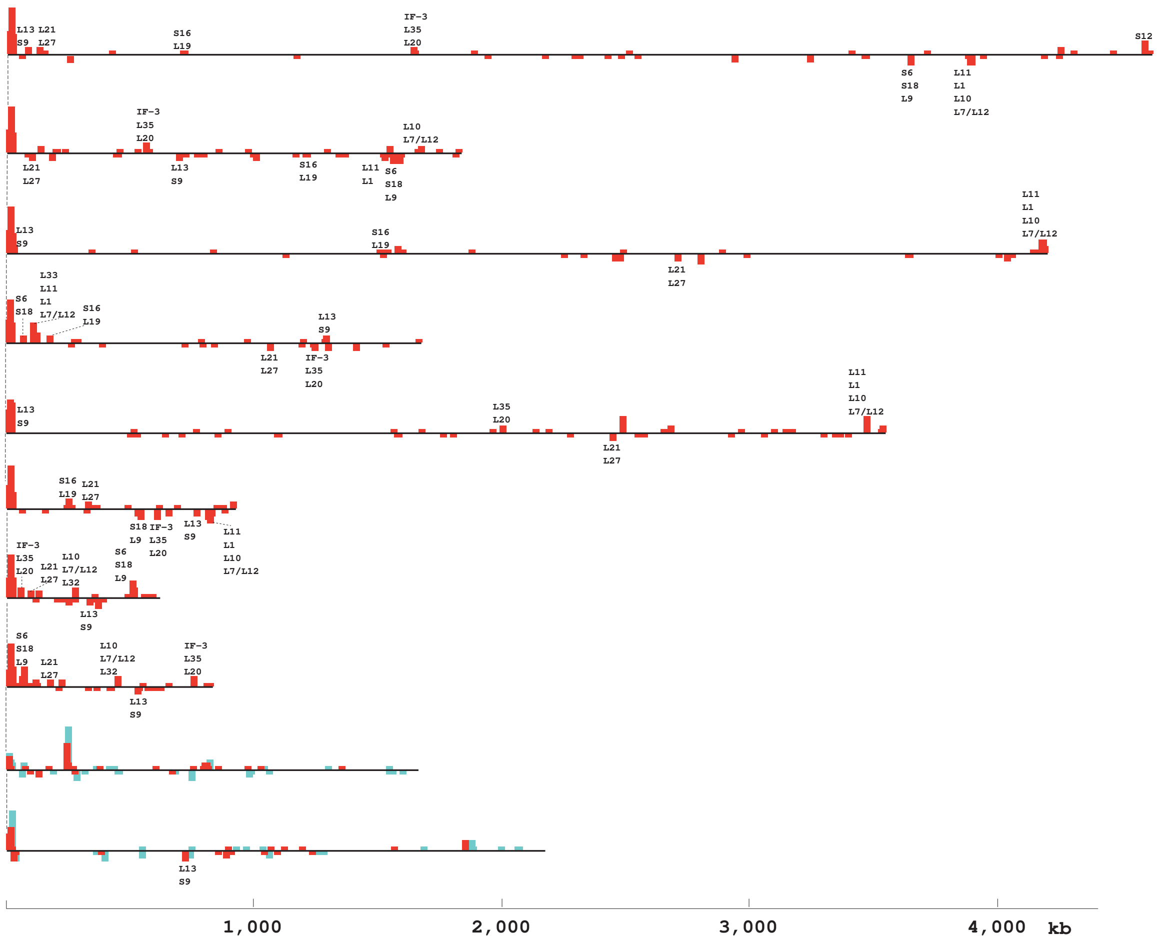

Figure 1: The genomic distribution of transcription/translation-related protein genes in individual organisms. Boxes representing the genes/ORFs homologous to those of *E. coli* are placed one after another in the vertical direction on the **w** (above) and **c** strand (below), respectively, of the horizontal line indicating the chromosome of each organism drawn to scale. In addition, those of *M. jannaschii* and *A. fulgidus* showing similarity to rat r-proteins are similarly

placed in light grey. The left end of each chromosome is arbitrarily set at the position of the L3 gene of the S10 operon. Some of the genes are marked with the name of their protein products for easier identification.

In *E. coli*, the three highly conserved proteins L1, L2 and L11 belong to the early-assembly proteins of the large subunit [21]. L1 has been shown to take part in the formation of the ridge on the large subunit [23]. Proteins L2 and L3 form part of the peptidyltransferase center and L5 binds to 5 S rRNA. The highly conserved L14 proteins is one of the late-assembly proteins and does not bind directly to 23 S r RNA [21].

S. cerevisiae has two sets of ribosomes, functioning in the cytoplasm and mitochondria, respectively. Considering the putative prokaryotic origin of mitochondria during evolution, it is expected that mitochondrial ribosomes contain bacterial r-protein homologues that are structurally well conserved. The results obtained were largely as expected. However, some bacterial r-proteins such as S11 and S17 showed no discernible similarity to any of the mitochondrial r-proteins, but rather to cytoplasmic ones.

3.2 Genomic distribution of the genes for r-proteins and related protein factors.

Since in *E. coli* the genes for r-proteins are often clustered with those for transcription and translation-related proteins such as RNA polymerase subunits, initiation factors, elongation factors, and so on, we surveyed not only r-protein genes but also the genes for other transcription/translation-related proteins. In addition, we analyzed the genes encoding r-protein modifying enzymes. The distribution of these genes was compared in the genome of each organism. The relative positions and orientation of r-protein genes were aligned with each other by placing the gene for L3 protein of the S10 operon at the left end as shown in Fig. 1.

The L13 and S9 genes are located near the S10 operon in *E. coli*, *B. subtilis* and *Synechocystis*, but they are not in the corresponding region in other organisms and its relative orientation are also different. The genes encoding L21 and L27, IF3 and L35, as well as L20 are similarly located and clustered in the eubacterial species analyzed. From these data it can be generalized that, while individual genes for r-proteins and related protein factors have been evolved by forming separate gene clusters which are well conserved amongst eubacteria, the locations and relative orientations of these gene clusters are quite diverged in other bacteria. There are exceptions: the genes for L11, L1, L10 and L7/L12 proteins, for example, compose a single operon in most eubacteria, but in *H. influenzae* they are split into two operons and their relative orientation is also different. Although *E. coli* and *H. influenzae* are most closely related amongst the eubacterial species analyzed, there are several differences in the organization of genes involved in transcription and translation in the two species. Unlike other bacteria *H. influenzae* has an ORF (see bellow) inserted between the S17 and L14 genes of the S10 operon and the S12, S7, EF-G and EF-Tu gene cluster is located on the opposite strand with respect to the S10 operon. *H. influenzae* has already been compared at the genomic scale with *E. coli* and *M. genitalium*. A comparison of the positions of the orthologous genes in the *E. coli* and *H. influenzae* chromosomes clearly showed lack of a long-range colinearity as reported by Tatusov *et al.* [27]. In comparison, lack of conservation of the gene localization in the genome of *M. genitalium* is striking [17].

We anticipated that the distribution pattern of these protein genes would be conserved in closely related species such as *M. pneumoniae* and *M. genitalium*, both belonging to *Mycoplasmataceae*. However, their distribution was found to be rather different from each other: while the genes for IF-3, L35 and L20 are located between the S10 operon and the cluster of L21 and L27 genes in *M. genitalium*, another cluster containing the S6, S18 and L9 genes is located in place of the IF-3-L35-L20 cluster in *M. pneumoniae*. Thus, the positions of the IF-3-L35-L20 and S6-S18-L9 clusters are exchanged between the two *Mycoplasmataceae* bacteria. Recently, comparative analysis of the two bacteria was performed by Himmelreich *et al.* [11], suggesting that the two genomes could be subdivided into six segments. The order of orthologous genes was well conserved within individual segments but the order of these segments was different. They explained that the different organization of the segments has been brought in by translocation via homologous recombination. A similar feature has been reported by Kunisawa [18] concerning *E. coli* and *B. subtilis*. There are many instances in which a contiguous r-protein gene segment in one genome is split into two or more segments that are located at different positions in the other. Unlike the two mycoplasma species mentioned above, *E. coli* and *B. subtilis* are rather distantly related: a phylogenetic study of their 5 S rRNA sequences has given an estimate of their divergence time of ca. 1.2 billion years [12].

Despite various differences mentioned above, there is one feature that is common to the eubacterial genomes we have analyzed; namely, the genes for many r-proteins are clustered in the region near the S10 operon. Since the expression of all ribosomal components must be regulated coordinately and the amount of each gene product be equal to each other, it would be beneficial that many r-protein genes are clustered with each other so that the regulation of their transcription would be rather easily facilitated.

Since we were interested in characterizing the r-protein genes from an evolutionary point of view, we then performed a similar analysis with the *M. jannaschii* and *A. fulgidus* r-protein genes. It was evident that the degree of similarity of their r-proteins to those of eubacteria was not very high. Therefore, the data for translation-related proteins of rat were also included in subsequent FASTA analysis. The genes in these archaebacteria have been diverged in clusters as in the case of those of eubacteria. The clusters are not identical, but some of the r-protein genes contained are similar to

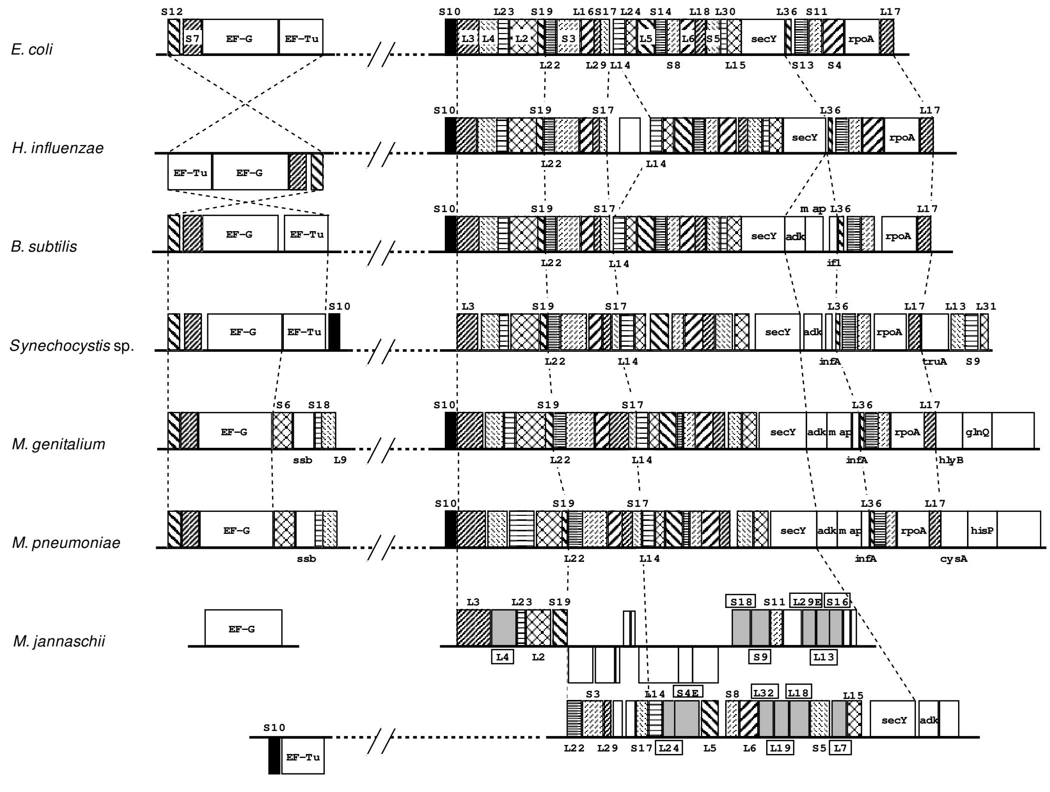

Figure 2: Alignment of the S12-S10-*rpoA*-operons. The r-protein genes (marked by *E. coli* protein names) are indicated by boxes of various patterns for easier identification and other protein genes are indicated by white boxes. The r-protein names that are boxed indicate those similar to rat r-proteins.

those in eubacteria, while many others are similar to those of rat which are often inserted between eubacteria-type r-protein genes. Fig. 2 shows the results obtained with the r-protein genes of *M. jannaschii*. It was also noticed that many of the eubacteria-type r-protein genes are either smaller or larger in size than their eubacterial counterparts.

3.3 Comparison of the S10 and S12 operon structures.

The fundamental structure of the S10-*spc*-*rpoA* operons is well conserved amongst the eubacterial species analyzed along with the relative gene order within the operons (Fig. 2). However, there are cases of insertions of other genes into the r-protein operons as well as translocations of some of the genes within the operon. In *H. influenzae*, for example, an ORF of hitherto unknown function has been inserted between the S17 and L14 genes. In *B. subtilis*, three protein genes exist between *secY* and the L36 gene, and the S4 gene is translocated between the S11 gene and *rpoA* encoding the α-subunit of RNA polymerase. The size of each gene is nearly identical from one species to another in most cases. A notable exception is the case of the L23 gene of *M. pneumoniae* which contains an extra stretch at the 3'-terminus which shows no similarity to any other known bacterial proteins. It is likely that the extra stretch of L23 has been acquired after the diversion of *M. pneumoniae*, although the precise branching point is not known. In *Synechocystis*, the genes for L13, S9 and L31 have been translocated at the end of the S10-*spc*-*rpoA* operons. That the L13 and S9 genes are clustered can be observed in other eubacteria. The S10 gene of *Synechocystis* is not in the S10-*spc*-*rpoA* operons, but it is located in the S12 operon. Since the S10 gene leads the S10-*spc*-*rpoA* operons in *E. coli* and *B. subtilis* and is just downstream of the S12 operon, during the course of genomic evolution, the S10 gene of *Synechocystis* have apparently moved to the 3'-end of the S12 operon. The order of the genes for S12, S7 and EF-G is also conserved, but the genes following the operon differs from one species to another: in *E. coli*, *H. influenzae*, *Synechocystis*, and *B. subtilis* the gene encoding EF-Tu is located immediately downstream of the EF-G gene, while it is the S6-*ssb*-S18-L9 cluster that follows the S12 operon (Fig. 2). The S6-*ssb*-S18-L9 cluster is also conserved as such, though at a different locus, except for *Synechocystis* sp. in which these protein genes are dispersed at various loci. Furthermore, in Gram-negative bacteria including *Synechocystis*, the *adk* gene encoding adenylate kinase exists at the end of the S10 operon, but another copy exists at a different locus in *Synechocystis*. Therefore, the genomic organization of the genes for ribosomal and related proteins in *Synechocystis* is more different than expected from their nucleotide sequence divergence. Recently, the structure of the S10-*spc*-*rpoA* operons in *B. subtilis* have been analyzed and the promoters identified [20]. While the three operons in *E. coli* are transcribed from individual promoters, the *B. subtilis* counterparts are transcribed from one and the same promoter. Thus they form a single operon. We have compared the sequence between the S7 and EF-G genes and found that they are quite diverse (data not shown). Similarly, the size of the inter-cistronic regions vary a great deal from one organism to another.

In *M. jannaschii*, the genes in the S10 operon are not conserved, although the order of identifiable r-protein genes are relatively well conserved as shown in Fig. 2. There are many genes the protein products of which are different from eubacterial r-proteins: they are rather similar to eukaryotic proteins. Thus, in *M. jannaschii* the r-protein gene clusters are composed of genes some of which are similar to eubacterial, while some others to eukaryotic r-protein genes. The corresponding region in another archaebacterium, *A. fulgidus*, resembles that of *M. jannaschii*. However, the S10 homolog itself is not included in the region in both archaebacteria as in the case of *Synechocystis* sp. The genes encoding S12, S7, S10 proteins and the elongation factor 1a of a third archaebacterium, *Sulfolobus solfataricus*, are located in regions comparable to those of the corresponding genes in *E. coli* [14]. Interestingly, the genes for L11, L1, L10 and L12 proteins in *S. solfatarcius*, *S. acidocaldarius* and *Halobacterium cutirubrum* are located in the same order as in the case of their *E. coli* counterparts [25].

3.4 A high degree of conservation of the S12 gene.

S12 protein was found to be one of the most highly conserved r-proteins amongst the organisms we have analyzed including *S. cerevisiae*. A yeast ORF termed YNR036c is likely to encode its mitochondrial homolog which showed a very high degree of similarity to *E. coli* S12 (57.4 % identity in FASTA). In *E. coli*, S12 is essential for translational accuracy and is known to be a streptomycin-targeted protein [1]. For these reasons, we compared S12 genes and their deduced protein products from individual organisms in detail (Fig. 3). The amino acid residues that are known to be altered by streptomycin resistance mutations in *E. coli* were found to be highly conserved even in *S. cerevisiae* and archaebacteria. Therefore, we thought that these residues would also be important for the ribosomal function in other organisms. In fact, in *S. cerevisiae*, the amino acid residues of the cytoplasmic S12 homolog, RPS28 (indicated as Sc-c), were reported to be related to paromomycin resistance and translational accuracy [1]. However, there are differences between Gram-positive and Gram-negative bacteria. The S12 proteins

```
                                         *************
Ec      1 MATVNQLVR-KPRARKVAKSNVPALE-------------ACPQKRGVCTRVYTTTPKKPNSALRKVCRVRL-TNGFEVTSYIGGE
Hi      1 ...I.....-...VK..V........-------------.......................I..-............
Ss      1 .P.IQ..I.-SE.SKVQK.TKS....-------------Q...R....................A....-.S.....A..P.I
Bb      1 .P.I...I.-...KSQTEKTAS...Q-------------N...R..IC...M.V..........A....-S......A..P.I
Hp      1 .P.I...I.-.E.KKV.K.TKS...V-------------E...R....................AK...-.SK...I..IP..
Bt      1 .P.I.....-.G.EK..F..KS...NKGYNSFKKEQTNVAS.........G.M...........YA....-...I...A..P.I
Bs      1 .P.I...IR-.G.VS..EN.KS...NKGYNSFKKEHTNVSS.........G.M...........YA....-...I...A..P.I
Mp      1 ...IA..I.-...KK.KV..KS...HYNLNLLNKKVTNVYS.L.......G.M...........YAK...-......LT..P..
Mg      1 ...IA..I.-...QK.KV..KS...HYNLNLLNKKTTNVYS.L.......G.M..R........YAK...-......LA..P..
Sc-m   29  ..L..IK.GSGPP.RKKI.TAPQLD-------------Q...RK..VL..MVLK.......Q..A.....-...NV.SA..P..
Af     36                   AKADP..-------------GA.MA..IVLEKIGIEARQ....I..AV..Q.IK..RQI.AFCP.D
Mj     38                  .EKYDP..-------------GA.MA..IVIEKVGLEA.Q....I..CV..Q.IK..RV..AFCP.N
Sc-c   41                      SPFG-------------GSSHAK.IVLEKLGIES.Q....I..CV..Q.IK..KK..AFVPND

Ec     71 G--HNLQEHSVILIR--GGR-----VKDLPGVRYHTVRGALDC--SGVKDRKQARSKYGVKRPKA
Hi     71 .--........V...--...-----.................--A.......G..........
Ss     71 .--........V...--...-----.............I...T..A--T.......G.....TREKAKK
Bb     71 .--........V...--...-----.............I....K.T--L..NN..KG.....T.K...
Hp     71 .--.......IV.V.--...-----..........K..I......T--A..NK.TVS.....T.KA..TDKKATDNKKK
Bt     84 .--........V...--...-----....R.....II..G..T--A..AN.M.G.....A.K...AKK
Bs     84 .--........V...--...-----..N..R....I......T--A..EN.A.G.P...T.K...K
Mp     84 .--........T.L.--...-----..............I...T..T--V..EK.R.Q..A..A.KP..KS
Mg     84 .--........T.L.--...-----..............I...T..T--V..DK.R.Q..A..A.K..PKS
Sc-m   99 .--.DA....IVYVR--...-----CQ........K..VI...G.L--...VN.ISS.....A.K.SKS
Af     89 .AINFID..DEVIVEKI...MGRS-MG.I......KV.KVNNTSLREL.RG..EK.LR
Mj     95 HAINFID..DEVI.EGI..PKGPRAKG.I.....K.KVIMVGRNSLREL.RG.QEKIKR
Sc-c   91 .CLNFVD.NDEV.LA--.FGRKGKAKG.I.....FKV.KVSGVSLLALWKEK.EKPRS
```

Figure 3: Alignment of S12 proteins. The amino acid residues altered by streptomycin resistance mutations in *E. coli* are boxed. An extra stretch found in Gram-positive bacteria is indicated by asterisks. In *S. cerevisiae*, mitochondrial and cytoplasmic homologues are indicated as Sc-m and Sc-c, respectively. Dots represent residues identical to those of *E. coli* and hyphens indicate residues inserted for maximal alignment.

of the latter contain additional 13 amino acid residues in the N-terminal region which are well conserved in the four species belonging to *Bacillaceae* and *Mycoplasmataceae*. This feature is apparently characteristic to Gram-positive bacteria. In spite of this rather long insertion, the *E. coli* S12 can be completely substituted with the *B. stearothermophilus* S12 in poly(U)-dependent poly-Phe synthesis *in vitro* [9]. This might suggest that the extra residues in this protein are not essential for the ribosomal function. Interestingly, the stretch does not exist in archaebacteria and in either of the two homologues in *S. cerevisiae*. It might indicate that the stretch emerged after the diversion of Gram-positive and Gram-negative bacteria. Similarly, in several other highly conserved r-proteins, functionally important amino acid residues were found to be conserved across kingdom (data not shown).

Concluding Remarks

In this study, we analyzed the genes encoding r-proteins as well as other transcription/translation-related proteins of unicellular organisms whose genomic sequence has been completed. Many of the genes encoding these proteins are clustered (most likely forming operons) and their relative orientation is generally rather well conserved even in archaebacteria, although the genomic positions of r-protein gene clusters in the two mycoplasmas analyzed, for example, were found not to be identical. Furthermore, some of the r-protein genes are missing, while some others are duplicated, in some bacterial species. It remains to be investigated further to what extent the regulation of the expression of these r-protein gene clusters is conserved. Ribosomal proteins, especially those playing pivotal roles in translation such as S12, are structurally highly conserved in all organisms we have analyzed in this work. Since the genetical and biochemical aspects of r-proteins and their genes have been extensively characterized in *E. coli*, but not so much in other organisms, we hope that the analysis of r-proteins and their genes reported here will be useful to obtain important clues as to how the genomic architecture of essential genes such as those encoding r-proteins was established and has been evolved since then and what will be their functional relevance.

References

[1] Anthony, R. A., and Liebman, S. W., Alterations in ribosomal protein RPS28 can diversely affect translational accuracy in *Saccharomyces cerevisiae. Genetics*, 140:1247-1258, 1995.

[2] Blattner, F. R., Plunkett, G. III, Bloch, C. A., *et al.*, The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453-1474, 1997.

[3] Bult, C. J., White, O., Olsen, G. J., *et al.*, Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science*, 273:1058-1073, 1996.

[4] Deckert, G., Warren, P. V., Gaasterland, T., *et al.*, The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature*, 392:353-358, 1998.

[5] Fleischmann, R. D., Adams, M. D., White, O., *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269:496-512, 1995.

[6] Fraser, C. M., Gocayne, J. D., White, O., *et al.*, The minimal gene complement of *Mycoplasma genitalium. Science*, 270:397-403, 1995.

[7] Fraser, C. M., Casjens, S., Huang, W. M., *et al.*, Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature*, 390:580-586, 1997.

[8] Goffeau, A., Aert, R., Agostini-Carbone, M.L., *et al.*, The yeast genome directory. *Supplement to Nature*, 387:sup1-sup105, 1997.

[9] Higo, K., Held, W., Kahan, L., and Nomura, M., Functional correspondence between 30S ribosomal proteins of *Escherichia coli* and *Bacillus stearothermophilus. Proc. Nat. Acad. Sci. USA*, 70:944-948, 1973.

[10] Himmelreich, R., Hilbert, H., Plagens, H., *et al.*, Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.*, 24:4420-4449, 1996.

[11] Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R., Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium. Nucleic Acids Res.*, 25:701-712, 1997.

[12] Hori, H., and Osawa, S., Origin and evolution of organisms as deduced from 5 S rRNA sequences. *Mol. Biol. Evol.*, 4:445-472, 1987.

[13] Huynen, M. A., Differential genome display. *Trends Genet.*, 13:389-490, 1997

[14] Ianniciello G., Gallo, M., Arcari, P., and Bocchini, V., Organization of a *Sulfolobus solfataricus* gene cluster homologous to the *Escherichia coli str* operon. *Biochem. Mol. Biol. Internat.*, 33:927-937, 1994.

[15] Kaneko, T., Sato, S., Kotani, H., *et al.*, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, 3:109-136, 1996.

[16] Klenk, H.-P., Clayton, R. A., Tomb, J.-F., *et al.*, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature*, 390:364-370, 1997.

[17] Kolstö, A-B, Dynamic bacterial genome organization. *Mol. Micro.*, 24:241-248, 1997.

[18] Kunisawa, T, Identification and chromosomal distribution of DNA sequence segments conserved since divergence of *Escherichia coli* and *Bacillus subtilis. J. Mol. Evol.*, 40:585-593, 1995.

[19] Kunst, F., Ogasawara, N., Moszer, I., *et al.*, The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis. Nature*, 390:249-256, 1997.

[20] Li, X., L. Lindahl, L., Y. Sha, Y., and J. M. Zengel, J.M., Analysis of the *Bacillus subtilis* S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase S10-*spc*-α cluster. *J. Bacteriol.*, 179:7046-7054, 1997.

[21] Nierhaus K. H., The assembly of prokaryotic ribosomes. *Biochimie*, 73:739-755, 1991.

[22] Noller, H. F., Moazed, D., Stern, S., em et al., Structure of rRNA and its function interactions in translation. In: em The Ribosome: Structure, Function and Evolution pp. 73-92, Hill, W. E. (ed.), Amer. Soc. Microbiol., Washington, DC., 1990.

[23] Oakes, M. I., Scheinman, A., Atha, T., Shankweiler, G., and Lake, J.A., Ribosome structure: three-dimensional locations of rRNA and proteins. In: *The Ribosome: Structure, Function and Evolution* pp. 180-193, Hill, W. E. (ed.), Amer. Soc. Microbiol., Washington, DC., 1990.

[24] Pearson, W. R., and Lipman, D.J., Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444-2448, 1988.

[25] Ramirez, C., Shimmin, L.C., Leggatt, P., and Matheson, A.T., Structure and transcription of the L11-L1-L10-L12 ribosomal protein gene operon from the extreme thermophilic archaeon *Sulfolobus acidocaldarius. J. Mol. Biol.*, 244:242-249, 1994.

[26] Sanchez, C., Blanco, G., Mendez, C., and Salas, J.A., Cloning, sequencing and transcriptional analysis of a *Streptomyces coelicolor* operon containing the *rplM* and *rpsl* genes encoding ribosomal proteins ScoL13 and ScoS9. *Mol. Gen. Genet.*, 257:91-96, 1997.

[27] Tatusov, R. L., Mushegian, A.R., Bork, P., *et al.*, Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Curr. Biol.*, 6:279-291, 1996.

[28] Tomb, J-F., White, O., Kerlavage, A.R., *et al.*, The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature*, 388:539-547, 1997.

[29] Wada, A., Analysis of *Escherichia coli* ribosomal proteins by an improved two dimensional gel electrophoresis. I. Detection of four new proteins. *J. Biochem.* (Tokyo), 100:1583-1594, 1986.

[30] Wittmann-Liebold, B., Ribosomal proteins: their structure and evolution. In: em Structure, Function and Genetics of Ribosomes, pp. 326-361, Hardesty, B., and Kramer, G. (eds.), Springer-Verlag, New York, 1986.