Dynamic evolution of translation initiation mechanisms in prokaryotes

So Nakagawa[a], Yoshihito Niimura[b], Kin-ichiro Miura[c,1], and Takashi Gojobori[a,d,2]

[a]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima 411-8540, Japan; [b]Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo 113-8510, Japan; [c]Department of Medical Genome Science, Graduate School of Frontier Science, University of Tokyo, Kashiwa 277-8562, Japan; and [d]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

Communicated by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, February 23, 2010 (received for review December 7, 2009)

It is generally believed that prokaryotic translation is initiated by the interaction between the Shine-Dalgarno (SD) sequence in the 5′ UTR of an mRNA and the anti-SD sequence in the 3′ end of a 16S ribosomal RNA. However, there are two exceptional mechanisms, which do not require the SD sequence for translation initiation: one is mediated by a ribosomal protein S1 (RPS1) and the other used leaderless mRNA that lacks its 5′ UTR. To understand the evolutionary changes of the mechanisms of translation initiation, we examined how universal the SD sequence is as an effective initiator for translation among prokaryotes. We identified the SD sequence from 277 species (249 eubacteria and 28 archaebacteria). We also devised an SD index that is a proportion of SD-containing genes in which the differences of GC contents are taken into account. We found that the SD indices varied among prokaryotic species, but were similar within each phylum. Although the anti-SD sequence is conserved among species, loss of the SD sequence seems to have occurred multiple times, independently, in different phyla. For those phyla, RPS1-mediated or leaderless mRNA-used mechanisms of translation initiation are considered to be working to a greater extent. Moreover, we also found that some species, such as Cyanobacteria, may acquire new mechanisms of translation initiation. Our findings indicate that, although translation initiation is indispensable for all protein-coding genes in the genome of every species, its mechanisms have dynamically changed during evolution.

dynamic evolution | Shine-Dalgarno sequence | ribosomal protein S1

Translation initiation is fundamentally important for all protein-coding genes in the genome of every organism. Initiation, rather than elongation, is usually the rate-limiting step in translation, and proceeds at very different efficiencies depending on the sequences in the 5′ UTRs of mRNAs (1). In prokaryotes (for both eubacteria and archaebacteria), the Shine-Dalgarno (SD) sequence in an mRNA is well known as the initiator element of translation (2, 3). The SD sequence, typically GGAGG, is located approximately 10 nucleotides upstream of the initiator codon. The SD sequence pairs with a complementary sequence (CCUCC) in the 3′ end of a 16S rRNA. In the 16S rRNA, the sequence is called the anti-SD sequence in the 3′ tail of which region is single-stranded. The interaction between the SD and the anti-SD sequences (called the SD interaction) augments initiation by anchoring the small (30S) ribosomal subunit around the initiation codon to form a preinitiation complex (4). The importance of the SD interaction for efficient initiation of translation has been experimentally verified for both eubacteria and archaebacteria. Alterations of the SD sequence or the anti-SD sequence strongly inhibit protein synthesis, both in eubacteria including *Escherichia coli* (5) and *Bacillus subtilis* (6, 7) and in archaebacteria such as *Methanocaldococcus jannaschii* (8). For this reason, the SD interaction is thought to be the universal mechanism of translation initiation in prokaryotes (9, 10).

Although translation initiation is essential for all protein-coding genes in the genome of every species, its mechanisms are quite different between prokaryotes and eukaryotes. In eukaryotes, translation is generally initiated by a scanning mechanism. The small (40S) ribosomal subunit, with several initiation factors, binds the 7-methyl guanosine cap (11) at the 5′ end of an mRNA. It moves along the mRNA until it encounters an AUG codon that is surrounded by a particular sequence such as the Kozak sequence (12). Hernández (13) hypothesized that the emergence of a nucleus led to the disappearance of the SD interaction and establishment of other mechanisms of translation initiation in eukaryotes. Thus, it appears that the mechanism of translation initiation in eubacteria and archaebacteria have not changed during evolution as a result of the absence of a nucleus.

However, two exceptional mechanisms of translation initiation have been identified in prokaryotes. One is translation initiation mediated by a ribosomal protein S1 (RPS1), which is a component of the 30S ribosomal subunit. In *Escherichia coli*, RPS1 interacts with a 5′ UTR of an mRNA, initiating translation efficiently, regardless of the presence of the SD sequence (14, 15). RPS1 of *E. coli* contains six S1 domains that are essential for RNA binding, although the number of domains is different among species (16). Recently, Salah and colleagues (17) analyzed the molecular diversity of RPS1s, and classified them into four types depending on their functional reliability of translation initiation, suggesting that the function of RPS1 in translation initiation is different among prokaryotes.

The other mechanism of translation initiation is for leaderless mRNAs that lack their 5′ UTR. A leaderless mRNA directly binds a 70S ribosome including an N-formyl-methionyl-transfer RNA, where translation is initiated (18–20). Leaderless mRNAs have been found in various species of prokaryotes, particularly in archaebacteria (21–23). For example, in *Halobacterium salinarum*, which belongs to the Euryarchaeota, leaderless mRNAs show a 15-fold higher activity in translation than mRNAs with the SD sequence (24). This suggests that the SD interaction might not necessarily be effective for translation initiation in some species of prokaryotes. Rather, the presence of these two mechanisms implies the possibility that the mechanisms of translation initiation have diversified among prokaryotes (25). However, the evolutionary changes of these mechanisms of translation initiation among prokaryotes are still unclear. In particular, it is interesting to know how universal the SD sequence is as the effective initiator for translation among prokaryotes.

The main purpose of this article is to answer this question to understand the evolutionary processes of translation initiation. We examined the genomes of 277 prokaryotes belonging to 14 phyla of eubacteria and three phyla of archaebacteria (Dataset S1).

Author contributions: S.N., Y.N., K.-i.M., and T.G. designed research; S.N. performed research; S.N. contributed new reagents/analytic tools; S.N. analyzed data; and S.N., Y.N., and T.G. wrote the paper.

The authors declare no conflict of interest.

[1]Deceased September 21, 2009.

[2]To whom correspondence should be addressed. E-mail: tgojobor@genes.nig.ac.jp.

This article contains supporting information online at www.pnas.org/cgi/content/full/1002036107/DCSupplemental.

Our comparative analysis of a wide variety of genomes provides a comprehensive picture of the evolution of the mechanisms of translation initiation.

Results

Conservation of the Anti-SD Sequence in 16S rRNA. To identify the SD sequence in the genomes of 277 prokaryote species, we first determined the 3′ terminal sequence of a 16S rRNA, which includes the anti-SD sequence, in each species. We defined the last 13 bases of a 16S rRNA as the 3′ tail sequence, which is the same as that of *E. coli* (GAUCACCUCCUUA). We found that the annotated sequences of 16S rRNAs in 98 of the 277 species do not contain the nucleotide sequence CCUCC, the anti-SD sequence of *E. coli*, in their 3′ tails. It is known that the annotation of 16S rRNAs is often dubious (26). Therefore, to determine whether the absence of the anti-SD sequence is an annotation error, we extended the terminal sequences of each of the 16S rRNAs using the genome sequences of the 277 prokaryotes. We then constructed a multiple alignment from these sequences. As a result, we identified a highly conserved motif in the extended 3′ end of 16S rRNAs of each species corresponding to the sequence of 3′ tail of a 16S rRNA of *E. coli* (Fig. 1). In particular, the anti-SD sequence was completely conserved in all species examined, except for three archaebacteria in which the cytosine at the last position is substituted by adenine. We also found several other highly conserved motifs in the upstream region of the 3′ tail (Fig. S1 shows alignment of 16S rRNAs). We therefore used the obtained 3′ tails of 16S rRNAs of each species for identification of the SD sequence in this study. The presence of this highly conserved anti-SD sequence indicates that the SD interaction functions as the initiator of translation in various prokaryotes. Dataset S1 includes the 3′ tails of 16S rRNA sequences for all species examined, which were annotated and obtained in this study.

Interspecific Variation in the Proportion of SD-Containing Genes. We examined the presence or absence of the SD sequence in each mRNA of the 277 species by calculating the interaction energy between the 3′ tail of a 16S rRNA and the SD region of an mRNA sequence, from –20 (i.e., 20 bases before the initiation codon) to –5. The SD sequence (such as GGAGG) is GC-rich, and therefore a genome with a higher GC content tends to artificially show a higher proportion of SD-containing genes (we named the proportion of SD-containing genes R_{SD}). Therefore, we calculated the R_{SD} value of the random sequences with a given GC content in a species (rR_{SD}, 0.24 ± 0.11). We then defined an SD index $dR_{SD} = R_{SD} - rR_{SD}$ (see *Materials and Methods* for details). We found that the dR_{SD} values vary greatly among species, ranging from 0.836 to –0.229, suggesting that the usage of the SD sequence is highly diversified among prokaryotes. We summarized both values of dR_{SD} and R_{SD} for each phylum in Table 1 (data set for all species examined).

Fig. 1. Highly conserved sequence in the 3′ end of 16S rRNA The sequence logo was obtained from the multiple alignment of 16S rRNAs of 277 species. Positions with information content contain a stack of nucleotide characters (A, U, G, and C). The overall height of the stack indicates the sequence conservation at that position, whereas the height of symbols within the stack indicates the relative frequency of each nucleotide at that position (*Materials and Methods*). An asterisk indicates the position corresponding to the 3′ end of the 16S rRNA of *E. coli*. See Fig. S1 for the whole alignment.

Table 1. Number of species and mean dR_{SD} and R_{SD} in the 17 phyla used in this study

Phylum	No. of species	dR_{SD}	R_{SD}
Prokaryotes	277	0.390	0.627
Eubacteria	249	0.404	0.635
Actinobacteria	22	0.383	0.673
Aquificae	1	0.358	0.515
Bacteroidetes	4	–0.149	0.143
Chlamydiae	7	0.361	0.541
Chlorobi	3	0.095	0.338
Chloroflexi	2	0.560	0.752
Cyanobacteria	9	0.012	0.386
Deinococcus-Thermus	3	0.258	0.588
Firmicutes	51	0.636	0.770
Mollicutes	14	0.414	0.513
non-Mollicutes	37	0.720	0.868
Fusobacteria	1	0.607	0.697
Planctomycetes	1	0.086	0.335
Proteobacteria	138	0.368	0.620
α	43	0.335	0.578
β	25	0.260	0.591
δ	11	0.450	0.710
ε	6	0.587	0.741
γ	53	0.405	0.636
Spirochaetes	6	0.498	0.626
Thermotogae	1	0.543	0.871
Archaea	28	0.259	0.553
Crenarchaeota	5	0.186	0.463
Euryarchaeota	22	0.288	0.592
Nanoarchaeota	1	–0.024	0.153

Fig. 2 shows the phylogenetic trees of eubacteria (249 species) and archaebacteria (28 species) based on their 16S rRNA sequences. In this study, we classified the species depending on its phylum provided by the Gene Trek in Prokaryote Space 2006 (GTPS2006) database in DDBJ (27). The species in Proteobacteria are subdivided into five classes as exceptions because of the large number of species (138 species). Firmicutes are also subdivided into two classes; (*i*) Mollicutes, including Mycoplasmas, and (*ii*) the other Firmicutes, because of distinct biological features of Mollicutes (e.g., ref. 28). Phylogenetic analyses suggested that, regardless of the variation in dR_{SD} values among species, the dR_{SD} values are relatively constant within each phylum. However, the dR_{SD} values of Euryarchaeota or Mollicutes varied within phylum or class, respectively (*Discussion*). The box plots of dR_{SD} classified by phylogenetic relationships (29) showed that the phyla with low dR_{SD} values (such as Bacteroidetes, Nanoarchaeota, and Cyanobacteria) have no close relationships to each other (Fig. S2*A*).

RPS1: An Alternative Mechanism of Translation Initiation. As noted in here earlier, RPS1 can initiate translation without the presence of an SD sequence in an mRNA, and the molecular structure of RPS1 varies among phyla. We therefore hypothesized that a variation of the SD sequence is related to molecular diversity of the RPS1s among prokaryotes. To verify this hypothesis, we classified all species examined into five types depending on the reliability of the RPS1 function for translation initiation, according to Salah et al. (17) as follows: type I is Aquificae, Bacteroidetes, Chlamydiae, Chlorobi, Deinococcus-Thermus, Planctomycetes, Proteobacteria, Spirochetes, and Thermotogae; type II is Actinobacteria; type III is Chloroflexi and Clostridia in Firmicutes; type IV is Cyanobacteria, Fusobacteria, and two

EVOLUTION

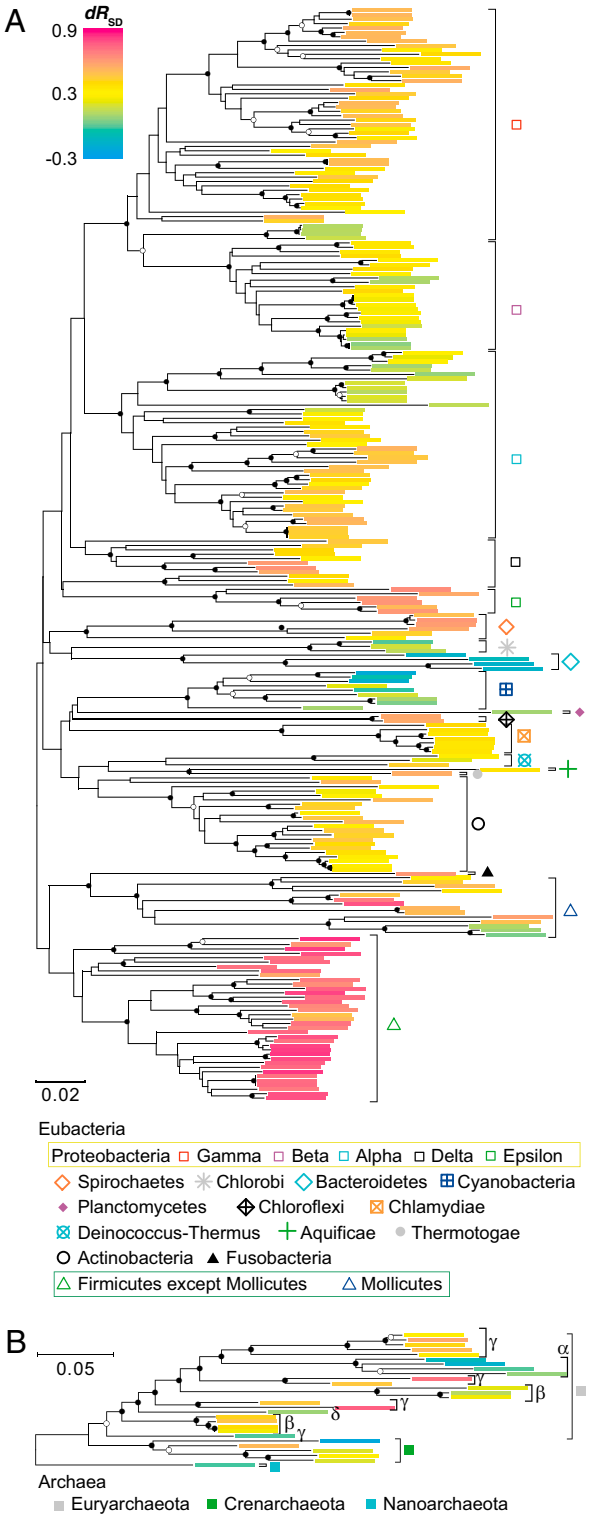

Fig. 2. Phylogenetic trees showing dR_{SD}. Neighbor-joining phylogenetic trees were constructed based on the 16S ribosomal RNA sequences from eubacteria (*A*) and archaebacteria (*B*). A colored bar at the branch shows the dR_{SD} value for each species. The diagram at the upper left indicates the color scheme for dR_{SD} values. Nodes supported with high bootstrap values, which were obtained from 1,000 resamplings, are shown by a black circle (≥90%) and an open circle (≥80%). Symbols represent each taxonomic group of species.

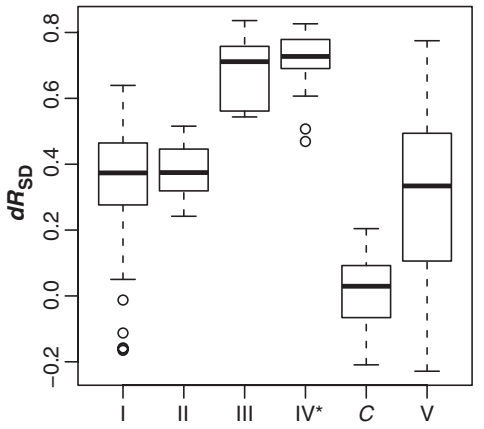

Fig. 3. Box plot of dR_{SD} depending on RPS1 function for translation initiation The box plot represents dR_{SD} values of each RPS1 type (I ~ V), as described in *Results*. Type IV (indicated by an asterisk) does not contain Cyanobacteria species. The dR_{SD} values of Cyanobacteria are shown in the column indicated by the italic letter *C*. The middle line indicates the median and the upper and lower edges of the boxes represent the first and third quartiles, respectively. The ends of the vertical lines indicate the minimum and maximum data values, unless outliers are present, in which the lines extend to a maximum of 1.5 times the interquartile range.

classes of Firmicutes (Bacillales and Lactobacillales). Among the first four types, type I is the most reliable and type IV is the least. As for the taxonomic groups that do not have an RPS1, we unified them into one type, type V, which represents all archaebacteria and Mollicutes in Firmicutes. We then examined the dR_{SD} values in each type, and found that the species in type I or II tend to show low values of dR_{SD}, whereas those in type III or IV, except Cyanobacteria, tend to show high values (Fig. 3; $P < 10^{-10}$, Wilcoxon rank-sum test of types I/II vs. types III/IV except Cyanobacteria). This result shows that species with a functional RPS1 for translation initiation tend to show low values of dR_{SD}. However, we also found that the dR_{SD} values of type V varied (Fig. 3). The exception of Cyanobacteria in type IV and the species in type V may represent the possibility of other mechanisms of translation initiation, including leaderless mRNAs (*Discussion*).

Evaluation of SD Interaction for Efficient Initiation of Translation. The results of our analysis revealed that the universality of the SD interaction as the effective initiator for translation is debatable, and one might wonder whether the SD interaction is really functional for an effective initiator of translation, particularly in those species with low dR_{SD} values. To answer this question, we categorized all species examined into three groups depending on their dR_{SD} values: high SD (dR_{SD} >0.5; 78 species), middle SD (0.5 ≥ dR_{SD} > 0.1; 170 species), and low SD (dR_{SD} ≤0.1; 29 species; Fig. S3). For each group, the efficiencies of translation initiation were compared between SD-containing genes and non–SD-containing genes. It has recently been shown that mRNA folding around the initiation codon is associated with the efficiency of translation initiation and plays a predominant role in determining the amount of protein produced (30, 31). The rate of translation initiation is thought to be high for an mRNA whose secondary structure around the initiation codon is unfolded. In addition, codon biases in a coding region are correlated with protein production (32, 33). For this reason, we evaluated the efficiency of translation initiation of each gene by the energy of an mRNA folding around the initiation codon and the index of codon usage bias.

As a result, in the high SD group, SD-containing genes showed significantly lower folding energies or stronger codon usage biases than the non–SD-containing genes (Fig. 4; $P < 10^{-5}$ in both cases, Wilcoxon signed-rank test with Bonferroni correction), indicating

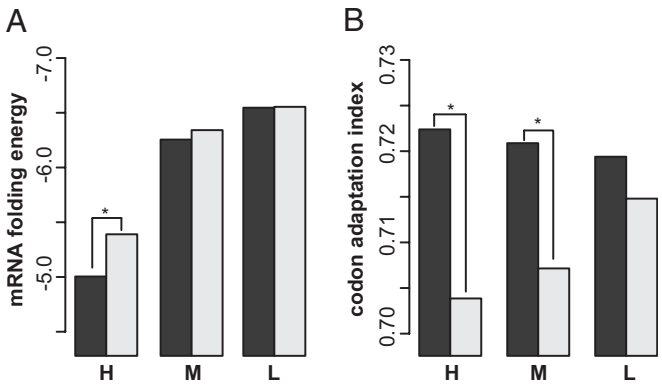

Fig. 4. Differences in the efficiencies of translation initiation between SD-containing genes and non–SD-containing genes. The mRNA folding energy (*A*) or codon adaptation index (*B*) of SD-containing genes and non–SD-containing genes are shown as black and gray bars, respectively. H, M, and L represent high SD, middle SD, and low SD groups, respectively (*Results*). An asterisk indicates a P value $<10^{-5}$ (Wilcoxon signed-rank test with Bonferroni correction), and no mark indicates no significant differences (i.e., $P > 0.05$).

efficient initiation of translation in SD-containing genes. In the middle SD group, codon usage biases in the SD-containing genes were significantly larger than those in the non–SD-containing genes ($P < 10^{-5}$), whereas folding energies between SD-containing and non–SD-containing genes were not statistically different ($P > 0.05$). In the low SD group, there were no significant differences between SD-containing genes and non–SD-containing genes in the folding energy or the codon usage bias ($P > 0.05$ in both cases). These results indicate that the SD interaction is not an efficient mechanism of translation initiation in species with a small proportion of SD-containing genes. These results further suggest that the loss of the SD sequence might be a result of the loss of function of enhancing translation initiation.

Gene Function Related to SD Sequence. Gene function might be related to the diversity in the SD indices among prokaryotes. Fig. S4 indicates the relative fraction of the SD-containing genes in each functional category. We found that metabolic-related genes, especially for energy production and conversion, tend to show a higher proportion of SD-containing genes than the other functional categories ($P < 0.01$, Wilcoxon rank-sum test). This result indicates that the presence or absence of an SD sequence in a gene may depend on the gene function. We also examined the correlation of dR_{SD} values with genomic or environmental features of species, including the genome size, the number of genes, gene densities, and living temperatures. However, we did not detect any significant correlations (Fig. S5). Therefore, gene function rather than genomic or environmental features may be partially responsible for the diversification of the mechanisms of translation initiation during evolution.

Discussion

Our analysis clearly shows that the SD index in a species is highly dependent on its phylum. However, Euryarchaeota and Mollicutes are exceptions to this result. The phylogenetic trees in Fig. 2 and the box plot of dR_{SD} in Fig. S2*A* show the diversification of dR_{SD} in these two groups. Interestingly, those species do not have any RPS1. A possible explanation of the variability of dR_{SD} is related to large proportions of leaderless mRNAs in these groups. Although it is difficult to distinguish between a leader mRNA (i.e., an mRNA with a 5′ UTR) and a leaderless mRNA from genomic sequences, it has been reported that leaderless transcripts are often found in archaebacteria, but rarely seen in eubacterial species except Mollicutes (21–23, 34). As for Euryarchaeota, the diversity of the SD indices can be also related with the high diversity within a phylum (Fig. 2*B*). The Euryarchaeota consists of eight heterogeneous classes ("*eury-*" means "broad") such as extreme halophilic species including Halobacteria (indicated by "α" in Fig. 2*B*), extreme thermophilic species including Methanopyri, Thermococci, and Thermoplasma (β), methanogenic species including Methanobacteria, Methanococci, Methanomicrobia, and Methanopyri (γ), and sulfate reducers including Archaeoglobi (δ) (35). Indeed, the dR_{SD} values in each class are relatively constant, and species living in similar environments tend to show similar dR_{SD} values (Fig. 2*B* and Fig. S2*B*).

Our results also revealed that the diversity of RPS1 function for translation initiation correlates with the proportion of SD-containing genes of each phylum. This might be related with a gain of function in RPSs of the species whose RPS1s are not used for translation initiation. The RPS1 of *Fusobacterium nucleatum* (Fusobacteria) was reported to be a fusion between protein LytB (residues 1–286) and four S1 domains (residues 450–800) (17). Moreover, the RPS1 of Cyanobacteria was reported to be nonfunctional, because its S1 motifs seem to be unable to bind the 30S ribosome (17). However, we found two copies of RPS1 with four S1 domains in seven of nine species of Cyanobacteria genomes (Dataset S1). These results might indicate that several RPS1 proteins have gained a new function other than translation initiation.

Moreover, as shown in Fig. 3, Cyanobacteria show low values of dR_{SD} (mean, 0.012), which are totally different from the dR_{SD} of the other species whose RPS1 is also not used for translation initiation (the mean value of type IV, except Cyanobacteria, is 0.716). This observation can be explained by assuming that another mechanism of translation initiation is also used in Cyanobacteria. In fact, we found a strong cytosine bias immediately before the initiation codon (CCaug, with "aug" representing the initiation codon) in Cyanobacteria species, especially those belonging to the Chroococcales class, using the *G*-test (Fig. S6; see *SI Materials and Methods* for details). It might be reasonable to assume that this bias is related to translation initiation, considering the position of the bias. Interestingly, the Kozak sequence [GCC(A/G)CCaugG, A/G represents A or G] observed around the initiation codon in eukaryotes is also characterized by a CC dinucleotide immediately before the initiation codon. Although the same nucleotide bias was detected in other species, such as *E. coli* belonging to Proteobacteria (Fig. S7), the tendency is not as strong as in Cyanobacteria. Conversely, nucleotide biases of G and A at the SD region were weak but significantly observed in Cyanobacteria (Fig. S6), indicating that the SD sequence may be used for translation initiation in some genes of those species. Experimental verification of these mechanisms, however, are required.

The terminal sequence of 16S rRNAs is conserved among prokaryotes. Therefore, the SD interaction is thought to play an important role in translation initiation in essentially all prokaryote species that are descended from the last universal common ancestor. However, our results clearly show a diversity of mechanisms of translation initiation in prokaryotes during evolution. We also reported diversity in translation initiation mechanism in eukaryotes (36). One might then wonder why the SD sequence is considered to be the universal mechanism for translation initiation in prokaryotes. One possible reason is the large proportion of genes having the SD sequence in the well studied species (5–8). Those species in which the functionality of the SD sequence was confirmed by experimental evidences, such as *E. coli*, *B. subtilis*, or *M. jannaschii*, tend to show large positive values of dR_{SD} (53.7, 78.0, and 44.9, respectively). Although the genome of *H. salinarum*, in which leaderless mRNAs initiated translation efficiently (24), is not available, the dR_{SD} value of the other *Halobacterium* (*Halobacterium* sp. NRC-1) is negative (−3.0), indicating a lack of functionality of the SD sequence for translation initiation. Sakai and colleagues (37) analyzed a correlation between the codon usage bias and the Gibbs energy of the interaction between an upstream sequence of an mRNA and the 3′ end of the 16S rRNA in the species. They reported that a correlation was observed

for the following species: *E. coli*, *B. subtilis*, *M. jannaschii, Methanobacterium thermoautotrophicum* (dR_{SD} of 45.3), *Haemophilus influenzae* (41.5), and *Archaeoglobus fulgidus* (7.7). Meanwhile, no correlation was found in the following species: *Synechocystis* sp. (−3.9), *Mycoplasma genitalium* (2.7), and *Mycoplasma pneumoniae* (8.0). This result supposes that the SD sequence of the species with large dR_{SD} values tend to be effective for translation initiation. Therefore, we believe that the SD interaction has been considered the universal mechanism for effective initiation of translation in prokaryotes because the organisms used for most of the experiments have high SD presences.

According to our results and those of preceding studies, the origin of the mechanisms seems to use both the SD sequence and a leaderless mRNA, considering the evolutionary conservation of the anti-SD sequence and broad usage of leaderless mRNAs including all three domains, respectively (25, 38). It is known that the origin of eukaryotes is a hybrid of bacteria and archaebacteria, and translation-related proteins are shared by eukaryotes and archaebacteria (39). Indeed, three eukaryotic translation initiation factors are found in archaebacteria, not in eubacteria (40, 41). However, the details on molecular function of these homologous proteins in archaebacteria for translation initiation remain unclear (41). Additional experimental and comparative genomic studies are required to investigate the relationship of the mechanisms of translation initiation among three domains. In eubacteria, an RPS1 gene appeared in its root because Aquificae, which is reported to be the closest to the root of eubacteria, has an RPS1 gene, whereas neither archaebacteria nor eukaryotes have an RPS1 gene. The variation of RPS1 function might be related to the diversification of the proportion of SD-containing genes in a species depending on its phylum. The loss of the SD sequence might be accelerated when it is not essential for efficient translation initiation. Moreover, some species, such as Cyanobacteria, might acquire new mechanisms of translation initiation. All these results show that the mechanisms of translation initiation dynamically changed during evolution.

Materials and Methods

Genomic Data. All genome sequences and annotations were downloaded from the GTPS2006 database (27) (http://gtps.ddbj.nig.ac.jp/). For those species annotated with more than one strain, such as *E. coli* str. K12 substr. W3110 and *E. coli* str. K12 substr. MG1655, the strain having the largest number of genes was chosen as the representative one. We examined 277 species in this study (Dataset S1). For each species, we obtained protein-coding genes and 16S rRNA on the basis of the annotation. The protein-coding genes, which start from an AUG, GUG, UUG, AUA, AUU, or AUC codon and end with a stop codon, were used in this study. Information on gene functions and living temperatures of all organisms was obtained from the Clusters of Orthologous Groups (COG) Database of the National Center for Biotechnology Information (42) (http://www.ncbi.nlm.nih.gov/COG/) and the German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de/), respectively. The obtained living temperatures are described in Dataset S1.

Determination of 3′ End of 16S rRNAs and the SD Sequence. To detect the SD sequence in the 5′ UTR of mRNAs, we calculated the free energy for base pairing between the upstream sequence of an mRNA and the complementary sequence at the 3′ end of a 16S rRNA. As noted (*Results*), we searched the conserved elements of all species examined corresponding to the 3′ tail of 16S rRNA in *E. coli*. We constructed a multiple alignment of the sequences using the alignment program Q-INS-i in MAFFT (43). The Shannon entropy at position i, $I = 2 - (-\sum_n o_n^{(i)} \log_2 o_n^{(i)})$, where $o_n^{(i)}$ is the fraction of the observed number of nucleotide n (A, U, G, and C) at position i, was calculated by using WebLogo (44).

The change in the Gibbs free energy, ΔG, which is required to connect the two strands of nucleotides, the 3′ tail of a 16S rRNA and the SD region (position from −20 to −5) of an mRNA, was calculated using free_scan (45). Free_scan is based on individual nearest-neighbor hydrogen bonding methods (46). It is difficult to determine the terminal sequence of 16S rRNA. Therefore, the sequence corresponding to the 3′ tail of the *E. coli* 16S rRNA was used for each species in this study (Dataset S1). If the ΔG between the 3′ tail of a 16S rRNA and the SD region of an mRNA was smaller than −3.4535, the gene was assumed to have an SD sequence (45). The threshold for the identification of the SD sequence in an mRNA was the mean energy value of the four-base interactions between the SD and the anti-SD sequences (45). The proportion of the SD-containing genes in a species was calculated as R_{SD} as the number of the SD-containing genes divided by the number of all genes.

The determination of the SD sequence by calculating the interaction energy is, however, affected by the GC content in a species (from 74.9% in *Anaeromyxobacter dehalogenans* to 22.5% in *Wigglesworthia glossinidia*), because the SD sequence (GGAGG) is GC-rich. Therefore, to estimate the proportion of false-positive SD-containing genes resulting from GC content in a species, we generated 20,000 randomized sequences with the GC content calculated from the 5′ UTR (position from −100 to −1), excluding the SD region, of a given species. We then found that the proportion of the sequences recognized as SD sequences (named rR_{SD}) was strongly correlated with genomic GC content (Fig. S8*A*; Pearson correlation coefficient $r = 0.75$; $P < 10^{-10}$). To compare the proportion of SD-containing genes unaffected from the GC content in a species, we defined the SD index dR_{SD} to be R_{SD} minus rR_{SD}.

To validate dR_{SD} as the index for the proportion of the SD-containing genes in a species, we applied the *G*-test method, which can evaluate position-dependent nucleotide biases without the effect of variation in GC content (*SI Materials and Methods*) (36, 47, 48). Application of this method to the genomic data of *E. coli*, for example, led to successful identification of the nucleotide biases (G and A) in the SD region (Fig. S7). It seems reasonable to suppose that the strongest nucleotide bias in the SD region (g^{max}; *SI Materials and Methods*) is correlated with the proportion of SD-containing genes. The strong correlation between dR_{SD} and g^{max} (Fig. S8*B*; $r = 0.80$; $P < 10^{-10}$) indicates that dR_{SD} is applicable to the evaluation of a proportion of the SD-containing genes in a species, and that the use of the conserved terminal sequences of 16S rRNA is also suitable for this purpose. In addition, this result also suggests that the biases detected in the SD region are mainly caused by the SD sequence (i.e., the sequence corresponding to the 3′ tail of 16S rRNA). The slightly improved correlation (between g^{max} and R_{SD}, $r = 0.78$; $P < 10^{-10}$; Fig. S8*C*) might also support the use of dR_{SD}. We therefore used dR_{SD} as the index for the proportion of the SD-containing genes in a species. The values (g^{max}, dR_{SD}, R_{SD}, and rR_{SD}) for each species are summarized in Dataset S1.

Phylogenetic Analysis. Phylogenetic trees were constructed by first generating multiple alignments of 16S rRNAs of eubacteria and Archaebacteria using MAFFT (Q-INS-i) (43). The evolutionary distances were then computed using the Maximum Composite Likelihood Method (49). The phylogenetic trees were constructed from these distances by the neighbor-joining method as implemented in the program MEGA4 (50).

Calculation of Secondary Structure. Following the approach by Kudla et al. (29), we calculated the minimum Gibbs energy of a secondary structure from –4 to +37 in an mRNA of each species using the hybrid-ss-min program (version 3.5; NA = RNA, t = 37, $[Na^+] = 1$, $[Mg^{2+}] = 0$, maxloop = 30, prefilter = 22) (51).

Calculation of Codon Use. The codon bias in a gene was calculated as the geometric mean of the relative synonymous codon usage values corresponding to each of the codons used in that gene, divided by the maximum possible codon bias for a gene of the same amino acid composition (52). The codon usage was based on all of the protein-coding genes in the genome of each species.

ACKNOWLEDGMENTS. This article is respectfully dedicated to the late Kinichiro Miura. We thank José C. Clemente, Tomoya Baba, Todd S. Gorman, Mitch D. Day, Sonoko Kinjo, and Yoshiyuki Suzuki for their comments on this work. This research was financially supported by Grant 20770192 from the Ministry of Education, Culture, Sports, Science and Technology, Japan (to Y.N.), and Grant 2009-A80 from the National Institute of Genetics Cooperative Research Program.

1. Jacques N, Dreyfus M (1990) Translation initiation in *Escherichia coli*: old and new questions. *Mol Microbiol* 4:1063–1067.
2. Shine J, Dalgarno L (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* 71:1342–1346.
3. Shine J, Dalgarno L (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature* 254:34–38.
4. Dontsova O, Kopylov A, Brimacombe R (1991) The location of mRNA in the ribosomal 30S initiation complex; site-directed cross-linking of mRNA analogues carrying several photoreactive labels simultaneously on either side of the AUG start codon. *EMBO J* 10:2613–2620.

5. Jacob WF, Santer M, Dahlberg AE (1987) A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. *Proc Natl Acad Sci USA* 84:4757–4761.
6. Band L, Henner DJ (1984) *Bacillus subtilis* requires a "stringent" Shine-Dalgarno region for gene expression. *DNA* 3:17–21.
7. Zhou J, Petracca R (2000) Influence of single base change in Shine-Dalgarno sequence on the stability of *B. subtilis* plasmid PSM604. *J Tongji Med Univ* 20:183–185.
8. Dennis PP (1997) Ancient ciphers: translation in Archaea. *Cell* 89:1007–1010.
9. Myasnikov AG, Simonetti A, Marzi S, Klaholz BP (2009) Structure-function insights into prokaryotic and eukaryotic translation initiation. *Curr Opin Struct Biol* 19: 300–309.
10. Schmeing TM, Ramakrishnan V (2009) What recent ribosome structures have revealed about the mechanism of translation. *Nature* 461:1234–1242.
11. Furuichi Y, Miura K (1975) A blocked structure at the 5′ terminus of mRNA from cytoplasmic polyhedrosis virus. *Nature* 253:374–375.
12. Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* 234: 187–208.
13. Hernández G (2009) On the origin of the cap-dependent initiation of translation in eukaryotes. *Trends Biochem Sci* 34:166–175.
14. Boni IV, Isaeva DM, Musychenko ML, Tzareva NV (1991) Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res* 19:155–162.
15. Komarova AV, Tchufistova LS, Dreyfus M, Boni IV (2005) AU-rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J Bacteriol* 187:1344–1349.
16. Subramanian AR (1983) Structure and functions of ribosomal protein S1. *Prog Nucleic Acid Res Mol Biol* 28:101–142.
17. Salah P, et al. (2009) Probing the relationship between Gram-negative and Gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res* 37:5578–5588.
18. Moll I, et al. (2001) Evidence against an Interaction between the mRNA downstream box and 16S rRNA in translation initiation. *J Bacteriol* 183:3499–3505.
19. Moll I, Grill S, Gualerzi CO, Bläsi U (2002) Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol Microbiol* 43:239–246.
20. Udagawa T, Shimizu Y, Ueda T (2004) Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria. *J Biol Chem* 279:8539–8546.
21. Balakin AG, Skripkin EA, Shatsky IN, Bogdanov AA (1992) Unusual ribosome binding properties of mRNA encoding bacteriophage lambda repressor. *Nucleic Acids Res* 20: 563–571.
22. O'Connor M, Asai T, Squires CL, Dahlberg AE (1999) Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA. *Proc Natl Acad Sci USA* 96:8973–8978.
23. La Teana A, Brandi A, O'Connor M, Freddi S, Pon CL (2000) Translation during cold adaptation does not involve mRNA-rRNA base pairing through the downstream box. *RNA* 6:1393–1402.
24. Sartorius-Neef S, Pfeifer F (2004) *In vivo* studies on putative Shine-Dalgarno sequences of the halophilic archaeon *Halobacterium salinarum*. *Mol Microbiol* 51:579–588.
25. Boni IV (2006) Diverse molecular mechanisms for translation initiation in prokaryotes. *Mol Biol (Mosk)* 40:658–668.
26. Lin YH, Chang BC, Chiang PW, Tang SL (2008) Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes. *Gene* 416:44–47.
27. Kosuge T, et al. (2006) Exploration and grading of possible genes from 183 bacterial strains by a common protocol to identification of new genes: Gene Trek in Prokaryote Space (GTPS). *DNA Res* 13:245–254.
28. Balish MF, Krause DC (2006) Mycoplasmas: a distinct cytoskeleton for wall-less bacteria. *J Mol Microbiol Biotechnol* 11:244–255.
29. Olsen GJ, Matsuda H, Hagstrom R, Overbeek R (1994) fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput Appl Biosci* 10:41–48.
30. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258.
31. Seo SW, Yang J, Jung GY (2009) Quantitative correlation between mRNA secondary structure around the region downstream of the initiation codon and translational efficiency in *Escherichia coli*. *Biotechnol Bioeng* 104:611–616.
32. Looman AC, et al. (1987) Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *Escherichia coli*. *EMBO J* 6:2489–2492.
33. Faxén M, Plumbridge J, Isaksson LA (1991) Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471-1480 in 16S ribosomal RNA affects expression of glnS. *Nucleic Acids Res* 19:5247–5251.
34. Madeira HMF, Gabriel JE (2007) Regulation of gene expression in Mycoplasmas: contribution from *Mycoplasma hyopneumoniae* and *Mycoplasma synoviae* genome sequences. *Genet Mol Biol (Brazil)* 30:277–282.
35. Cavalier-Smith T (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52: 7–76.
36. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* 36:861–871.
37. Sakai H, et al. (2001) Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. *J Mol Evol* 52:164–170.
38. Andreev DE, Terenin IM, Dunaevsky YE, Dmitriev SE, Shatsky IN (2006) A leaderless mRNA can bind to mammalian 80S ribosomes and direct polypeptide synthesis in the absence of translation initiation factors. *Mol Cell Biol* 26:3164–3169.
39. Ribeiro S, Golding GB (1998) The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* 15:779–788.
40. Bell SD, Jackson SP (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol* 6:222–228.
41. Londei P (2005) Evolution of translational initiation: new insights from the archaea. *FEMS Microbiol Rev* 29:185–200.
42. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33–36.
43. Katoh K, Toh H (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 9: 212.
44. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190.
45. Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLOS Comput Biol* 2:e57.
46. Xia T, et al. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37: 14719–14735.
47. Watanabe H, Gojobori T, Miura K (1997) Bacterial features in the genome of *Methanococcus jannaschii* in terms of gene composition and biased base composition in ORFs and their surrounding regions. *Gene* 205:7–18.
48. Niimura Y, Terabe M, Gojobori T, Miura K (2003) Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res* 31:5195–5201.
49. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* 101:11030–11035.
50. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599.
51. Markham NR, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* 33 (Web Server issue):W577–W581.
52. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15: 1281–1295.