Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria

Yang Fu, Kaila Deiorio-Haggar, Jon Anthony and Michelle M. Meyer*

Department of Biology, Boston College, 140 Commonwealth Ave. Chestnut Hill, MA 02467, USA

Received November 2, 2012; Revised January 2, 2013; Accepted January 10, 2013

**ABSTRACT**

In *Escherichia coli*, 12 distinct RNA structures within the transcripts encoding ribosomal proteins interact with specific ribosomal proteins to allow autogenous regulation of expression from large multi-gene operons, thus coordinating ribosomal protein biosynthesis across multiple operons. However, these RNA structures are typically not represented in the RNA Families Database or annotated in genomic sequences databases, and their phylogenetic distribution is largely unknown. To investigate the extent to which these RNA structures are conserved across eubacterial phyla, we created multiple sequence alignments representing 10 of these messenger RNA (mRNA) structures in *E. coli*. We find that while three RNA structures are widely distributed across many phyla of bacteria, seven of the RNAs are narrowly distributed to a few orders of Gammaproteobacteria. To experimentally validate our computational predictions, we biochemically confirmed dual L1-binding sites identified in many Firmicute species. This work reveals that RNA-based regulation of ribosomal protein biosynthesis is used in nearly all eubacterial phyla, but the specific RNA structures that regulate ribosomal protein biosynthesis in *E. coli* are narrowly distributed. These results highlight the limits of our knowledge regarding ribosomal protein biosynthesis regulation outside of *E. coli*, and the potential for alternative RNA structures responsible for regulating ribosomal proteins in other eubacteria.

**INTRODUCTION**

In *Escherichia coli*, ribosomes and their associated cofactors comprise between 25 and 50% of cellular mass in actively growing cells (1), and the synthesis of ribosomal RNA and proteins is tightly coordinated to maintain stoichiometric levels of each component (2). Coordinating this synthesis is a complex task that is regulated at multiple levels (3). More than half of the ribosomal proteins (r-proteins) in *E. coli* are controlled by 12 distinct RNA autogenous regulatory elements (Figure 1) that occur within the mRNA transcripts, most frequently in the 5′-untranslated regions (UTRs) or leader sequences, of operons encoding r-proteins. Each regulatory element consists of a structured region of mRNA transcript that interacts with a specific r-protein to inhibit expression of an entire operon encoding multiple r-proteins (4). In many cases, the mRNA structure responsible for regulation is a mimic of the ribosomal RNA (rRNA)-binding site for the same protein (5). A variety of different mechanisms are used to inhibit gene expression including transcription termination (6), translation inhibition through ribosome-binding site occlusion (7) and ribosome entrapment (8,9). These regulatory RNAs provide direct feedback between the levels of ribosomal RNAs and the levels of r-proteins.

First identified in the late 1970s, r-protein autogenous regulators are some of the oldest known examples of effector-driven RNA-based gene regulation in bacteria, predating the discovery of riboswitches (10), T-boxes (11) and small RNAs (sRNAs) (12). Owing to the importance of ribosomal protein biosynthesis, the mechanism and structure for many of these regulators is well understood (4,13). However, most of the r-protein regulators described in *E. coli* are not annotated in standard genomic databases (14), and alignments for only 2 of the 12 RNAs are available in the RNA Families Database (Rfam) [Rfam: RF00140, RF00114] (15). While phylogenetic studies have been performed for several of these regulatory RNAs (16–20), such treatments were not systematic and the knowledge gained is not preserved in the databases used for genomic annotation. As a result, these important RNA molecules risk being overlooked as genome annotation moves further toward automation and online repositories. In addition, the extensive experimental data derived from *E. coli* sequences have not been...
put in context with homologous sequences identified through comparative genomic studies.

This work uses Infernal 1.0, an RNA-specific search tool (21), to identify homologues of the \textit{E. coli} r-protein autogenous regulatory RNAs. Multiple-sequence alignments representing each of the structured mRNAs responsible for gene regulation were either expanded from existing alignments, or created based on the \textit{E. coli} sequence and experimentally determined secondary structures in the literature. From these alignments, we integrated decades of experimental data with the RNA consensus secondary structures and characterized the phylogenetic distribution for each r-protein regulatory RNA. Additionally, for the L1-interacting RNA, we identified both a phyla-specific change in genomic locus and dual binding sites in a large proportion of sequenced Firmicute genomes. To verify the dual binding sites, we experimentally validated L1 interactions with the pair of RNA structures originating from \textit{Geobacillus kaustophilus}.

**MATERIALS AND METHODS**

Alignment construction and homology searches

For the S4 and S15 binding RNAs, the seed multiple sequence alignments were downloaded from the Rfam database (Rfam families RF00140 and RF00114, respectively) (15). These alignments were manually examined to ensure that sequences were compatible with available experimental data and any alternative or pseudoknotted structures were represented by the structural annotation. The initial alignments for the L1-, L10(L12)-, and S2-binding RNAs derived from non-coding RNA discovery efforts (22–25). To generate multiple sequence alignments for the remaining RNAs, BLAST matches outside of the \textit{Escherichia} genus (exclude taxid: 0561) to the \textit{E. coli} sequence corresponding to a minimal binding element were collected and aligned. To this multiple sequence alignment, experimentally determined secondary structure information was manually added, and the alignment was adjusted as necessary.

Once initial alignments were obtained, Infernal 1.0 (cmsguild) was used to create co-variance models corresponding to the RNA. These models were calibrated (cmcalibrate) and additional homologues identified for each RNA (cmsearch) (21). Both global and local searches were performed against a custom sequence database that contains genomic regions proximal to ribosomal proteins from all complete bacterial genomes in refseq46 (14). This database is \(\sim 57\) MB, which is an \(\sim 100\)-fold reduction from the entire refseq46-microbial database. The use of this database significantly increased the speed and sensitivity of our search process. The complete database (refseq46-microbial) was used for a single search with the final alignments, and no additional homologues were identified. Prospective homologues were screened based on the appropriateness of the genomic context using a custom visualization tool, GenomeChart (26). Homologues were subsequently screened for their fit to the existing alignment and the consistency with experimental data. When necessary, alignments were manually...
adjusted as sequences were added, especially in cases with variable length helices. Searches that produced many potential sequences inconsistent with experimental data are discussed in the text (e.g. L4, S8 and S4 interacting autogenous regulatory RNAs). The search process was typically repeated two to three times to expand the sequence diversity present in the alignment. In cases with pseudoknotted or alternative structures, searches were conducted with each structure iteratively. The curated final alignments produced are available in Supplementary Datasets. Percentages of bacteria in each phyla containing each RNA were calculated based on the number of completed genomes within refseq46 and the alignments produced by this work. Consensus secondary structure diagrams were created from the alignments using GSC-weighting in R2R (27).

L1 RNA-protein binding assays

The gene encoding ribosomal protein L1 (rplA) was polymerase chain reaction (PCR) amplified from G. kaustophilus genomic DNA (28) using the following primers 5'-ggattccatatcttattgtgcaacagcaaccgtg-3' and 5'-cggagtccttcatgcaacatagcaaccgtg-3' and expressed in E. coli strain BL21 using T7 overexpression from pET-HT G. The PCR products were cloned using a TOPO-cloning kit (Invitrogen) and the inserts sequenced to identify the 5'-end of the transcripts.

for PCR with an oligonucleotide corresponding to the 5'-linker (5'-GACTGGAGACCGAGGACACTGA) and a second set of gene-specific primers (rplK, 5'-CCAATT GGTGGTGTGTTAGGTCG; rplA, 5'-CGTGCTACAA GCATTAGCAGTTC). PCR products were cloned using a TOPO-cloning kit (Invitrogen) and the inserts sequenced to identify the 5'-end of the transcripts.

RESULTS

Overview of distributions

We compiled alignments for 10 of the 12 RNA structures controlling ribosomal protein biosynthesis in E. coli. Of the regulatory RNAs examined, three are widely distributed over many subbacterial phyla, and seven are narrowly distributed to a few orders of Gammaproteobacteria (Figure 2). The three widely distributed RNAs interact with ribosomal proteins L1, L10(L12), and S2, and each was identified in species from many bacterial phyla. All bacterial phyla, with the exception of Acidobacteria, harbor at least one of the RNAs. This sets a precedent for RNA-based regulation of ribosomal proteins in a wide range of bacterial diversity. The scattered distribution of the RNAs, and the low frequency of identification of some regulatory RNAs may point toward horizontal transfer, multiple inventions or a lack of sensitivity in our homology search methodologies.

The seven narrowly distributed r-protein regulatory RNAs (interacting with S4, S1, S7, S8, S15, L20 and L4 proteins) were identified only in Gammaproteobacterial species, and in many cases, only within a subset of the Gammaproteobacterial orders (Figure 2). The S1-interacting RNA shows the widest apparent distribution appearing in species from each order. However, the remaining regulatory RNAs are present only in the Enterobacteriales, Vibrionales, Pasteurellales, Aeromonadales and Alteromonadales orders of Gammaproteobacteria. This pattern of conservation is consistent with vertical inheritance within Gammaproteobacteria (33). In addition, the loss of several RNA elements, including L1 and S2, is apparent in many genera of obligate intracellular species within Enterobacteria, such as Wigglesworthia, Blochmania and Buchnera. While Buchnera species typically retain the S1- and L4-binding RNAs, most enterobacterial endosymbionts appear to have lost these control mechanisms during their genome reduction process, although they typically retain the ribosomal proteins implicated in regulation (34). While the general gene order for most ribosomal proteins is typically conserved (35), there are significant changes to the operon structures of the spe, alpha and S10 operons (regulated in E. coli by S8, S4 and L4 respectively, see Figure 1). In B. subtilis, the majority of genes within these operons are co-transcribed as a single transcript (36), thus possibly removing the need to regulate these operons individually.

L20-interacting and S20-interacting RNAs

There are two RNAs previously described in E. coli for which we did not create alignments: a second L20 interaction site, and the reported S20-interacting mRNA.
While the mRNA sequence preceding rpsT (encoding S20) has reported autogenous regulatory activity and apparent similarity to the ribosomal RNA, no secondary structure is reported for it in the literature (37, 38). Given its small size, without a secondary structure it is impossible to create a convincing starting alignment for RNA homology searches that rely heavily on RNA secondary structure. This region is highly conserved in most Enterobacteria, but BLAST searches detect no similarity outside of Enterobacteriales (NCBI TAXID 91347).

A second L20-binding RNA contains a long-range pseudoknot between bases within the infC and rpmF coding regions. This large insertion overlaps significantly with the coding region of infC, making convincing identification of the RNA difficult with the tools used here. A previous phylogenetic study identified this RNA in Enterobacteriales, Vibrionales, Pasteurellales, Alteromonadales, Legionales and Pseudomonadales (19).

### L1-interacting RNA

L1-interacting mRNAs have been reported in both Enterobacteria (E. coli) (39) and in several archaea including Methanococcus vannielii (40, 41). While the L1-binding sites in these organisms are similar, the genes regulated and the genomic loci of these sites are distinct in the two organisms. In E. coli, rplK and rplA are co-localized in the genome, and their translation is strongly coupled (Figure 1) (42). The L1-interacting structure precedes rplK, and L1 binding inhibits translation of both L1 and L11. In M. vannielii, rplK is not co-localized with rplA, instead rplA is co-transcribed with both rplJ and rplL, and the RNA structure directly precedes rplA. Consistent with the observed similarity between the binding sites, the E. coli L1 is able to regulate the M. vannielii production of L1 (40).

Our study identified L1-binding RNAs in the widest variety of different bacterial phyla. However, in many cases, only a few examples were identified in each phyla (Figure 2). Due to its small size (Figures 3 and 4) and minimal sequence conservation, the L1-binding site is somewhat difficult to track with computational tools, and it is likely that there are many individual occurrences that were not uncovered by our searches. The consensus figure produced from our alignment (Figure 4) for the

![Figure 2. Phylogenetic distribution of E. coli r-protein regulatory RNAs. (A) Distribution of ribosomal protein autogenous regulatory RNAs in bacterial phyla. (B) Distribution of regulatory RNAs in orders of Gammaproteobacteria.](https://academic.oup.com/nar/article-abstract/41/6/3491/2902762/3494)
L1-binding RNA is in good agreement with previous structural and biochemical studies. The consensus structure contains a 6–12-base-pair stem with an internal bulge that contains a potential non-canonical G–A pair followed by an unpaired adenosine. In structural studies, these nucleotides, and those in the base pairs immediately flanking the bulge, form a complex network of hydrogen bonds that is critical for protein binding (43).

Despite its widespread distribution, the genomic locus of the L1-interacting RNA is not consistent within eubacteria. In Cyanobacteria, Actinobacteria and Chloroflexi, the L1-interacting RNA appears directly preceding rplA (typically between rplA and rplK), and in Proteobacteria, Spirochaetes, Thermotoga and Tenericutes, it appears preceding rplK. Surprisingly, we identified two binding sites in >40% of completed Firmicutes genomes, one preceding rplK, and one preceding rplA.

To ensure that the dual binding sites were not false-positives resulting from the small size of the L1-binding site and the sensitivity of our searches, we experimentally validated the interaction between ribosomal protein L1 and the two RNA structures originating from *G. kaustophilus* using filter-binding assays (Figure 3A and B). These studies show that each of the RNA structures has nM affinity for the protein (Kd ~25–50 nM) that is abolished by point mutations to the consensus-binding site (Figure 3C) (31). To verify that *rplK* and *rplA* are co-transcribed in Firmicutes (as they are in *E. coli*) we performed 5′-RACE on total RNA extracted from *B. subtilis*. We found that all the products isolated from gene-specific primer within *rplA* (encoding L1) contain the entirety of *rplK*, indicating that the two genes are co-transcribed (Supplementary Figure S1). In addition, ~40% of the sequences isolated from gene specific primers within *rplK* extend to include the second RNA element that precedes *rplK*. The combination of these findings suggests that both sites are biologically relevant and play a role in regulating these genes in Firmicutes.

### L10(L12)_e-interacting RNA

The L10(L12)_e-interacting autogenous regulatory RNA is also widely distributed, and the RNA has been described in both Proteobacteria and Firmicutes (20). Our study expands the phylogenetic spread of this RNA considerably, identifying it in more than half of sequenced Fusobacteria, Actinobacteria, Cyanobacteria and Chloroflexi. In addition, examples were identified in Tenericutes, Thermotogae, Aquificae and Deinococcus-Thermus. The RNA is responsible for regulating *rplL* and *rplL* in response to the L10(L12)_e complex (44,45) (Figure 1). The minimal binding site identified using phylogenetic analyses and *in vitro* assays is characterized by a kink-turn motif followed by a bulged cytosine in the non-canonical stem (20). This motif is a direct mimic of the rRNA binding site for L10 (20). In the alignment produced by this study, the kink-turn is always present, and the bulged cytosine is present in >90% of examples (Figure 4). However, examples from Epsilonproteobacteria, Aquificae and Cyanobacteria are often missing this bulged cytosine. The stem is topped with a hairpin that contains an internal loop with two unpaired adenosines. These adenosines are present in all sequences examined, even in those where the bulge is extended into a second hairpin of varying length (Figure 4). Previous experimental work has shown that the two bulged adenosines and the kink-turn are both critical for L10(L12)_e binding (20).

The mechanism that the L10(L12)_e-binding RNA leader from *E. coli* uses to control gene expression is dependent on complex folding patterns with other portions of the mRNA leader (46–49) and appears to act post-transcriptionally, but before translation (50). However, based on the current Rfam alignment (which includes a terminator), this RNA is followed by a rho-independent intrinsic terminator in Firmicutes, Cyanobacteria and Fusobacteria (25,51), suggesting that there may be different regulatory mechanisms in different bacterial species, similarly to many riboswitch aptamers, which may have distinct expression platforms depending on the species of bacteria (52).

### S2-interacting RNA

Despite its wide distribution (Figure 2), the S2-interacting regulatory RNA in *E. coli* was discovered relatively recently (53), and no mechanistic studies have been...
Figure 4. Consensus sequence and secondary structures for E. coli r-protein regulatory RNAs. The alignments used to produce these figures are available as a Supplementary Dataset (Alignments.zip). Diagrams indicate the conserved secondary structure and sequence of the RNA structures. Base pairs supported by co-variation or compatible mutations are indicated by green or blue shading only if Watson–Crick pairing occurs in >95% of the sequences. Start codons (AUG or GUG) that occur within an RNA structure are boxed with solid lines, putative protein binding sites are indicated by dashed lines and helix numbering schemes are consistent with pre-existing literature for each RNA.
performed to date. The RNA appears to control the synthesis of both ribosomal protein S2 and EF-Ts (Figure 1), and the RNA was independently discovered through two additional avenues. It was first described in a study of the E. coli transcriptome as sRNA t44 [Rfam: RF00127] (54), and subsequently was identified in a de novo search for non-coding RNAs in marine Alphaproteobacteria using comparative genomics (22). A good alignment of the S2-binding RNA is already available (22); this work re-evaluated the existing alignment and identified the RNA in a wider range of organisms by incorporating data from more recent sequencing projects.

Although the native transcript in E. coli is significantly longer than the conserved S2-binding site, deletion experiments have shown that only the conserved portion of the RNA represented by the alignment is necessary for regulation (53). In addition, deletion of the GGU internal loop (predicted to form a pseudoknot by this alignment) is sufficient to abolish activity (Figure 4). Subsequent experiments have also shown that instances of this RNA from other Gammaproteobacteria, including Pseudomonas, function as regulatory elements in response to the E. coli S2 protein (17). The S2 interacting regulatory RNA also stands in contrast to the other widely distributed RNA autogenous regulatory elements in that it is not an obvious mimic of the ribosomal RNA, and its binding partner is not a primary RNA binding protein (55).

L4-interacting RNA

The S10 operon in E. coli consists of 11 genes (rpsJ, rplC, rplD, rplW, rplB, rpsS, rpsV, rpsC, rplP, rpmC, rpsQ) and is autogenously controlled by L4, encoded by rplD (Figure 1). A structured portion of mRNA preceding all of the coding genes allows both transcription termination (56) and translation inhibition (57) in response to L4. The L4-interacting RNA is perhaps the most comprehensively studied of the r-protein regulatory RNAs. Genetic and in vitro assays have shown that translation inhibition and transcription termination are conferred by distinct, although slightly overlapping, portions of the mRNA and together can account for a nearly 30-fold decrease in gene expression in the presence of excess L4 (58).

The L4-interacting RNA from E. coli consists of five hairpins, one of which overlaps the start codon of S10 (59) (Figure 4). The overall structure is largely conserved, and homologous sequences from Vibrio cholerae and Haemophilus influenzae have been experimentally demonstrated to regulate gene expression in response to E. coli L4 (16). The alignment of L4-interacting RNAs in this work indicates that helices HA-HC are highly variable in both sequence and in length. Without the annotation of transcription start sites, in many organisms, it is challenging to identify all of the hairpins based solely on genomic sequences. This is consistent with experimental results indicated that these helices are not necessary for regulation (60).

In contrast, helices HD-HG, which are required for transcription termination (61), show a much greater degree of sequence conservation allowing greater confidence in their alignment. In HD, the most conserved area encompasses the loop and the bases proximal to the loop, and this region has been implicated in L4 binding (62). While HE is present in every sequence in the alignment, there is a significant amount of sequence variation and many non-conserved single-nucleotide bulges. Portions of HE are required for transcription termination, and have been implicated in the NusA-dependent RNA polymerase pausing that precedes termination (61,63). While the string of uridines necessary for rho-independent transcription termination is not rigorously conserved in position, there are strings of uridines present in the lower part of HE or the linker between predicted HE and HG in all the sequences in the alignment. Helix HG is the most highly conserved region of the RNA, likely due in part to overlap with the S10 coding sequence (Figure 4). While we identified examples of the full-length RNA in the Pseudomonas family Moraxellaceae, we also identified sequences with similarity to HG alone in several species of Xanthomonas and the Pseudomonas family Pseudomonadaceae (16). As these regions have not been demonstrated to control gene expression in the absence of the upstream regions of the RNA, sequences from these species were omitted from the alignment.

The L4-interacting RNA translation inhibition mechanism has received less attention than its transcription termination mechanism. Deletion of HA-HD or HE is known to have little effect on translation regulation, while point mutations in HE have significant effects on translational efficiency (58). The proposed mechanism involves formation of an alternative stem that incorporates bases within both HG and HE (59). However, this stem is not exceedingly stable in the E. coli sequence and there is no evidence from the alignment for a conserved stem in this area.

S8-interacting RNA

The spc operon consists of twelve genes (rplN, rplX, rplE, rpsN, rpsH, rplF, rplR, rpsE, rpmD, rpoO, secY, rpmJ) and in E. coli is controlled by the binding of ribosomal protein S8 to an mRNA structure that overlaps the start of rplE (64,65) (Figure 1). The binding of S8 to this region inhibits translation of L5 and the proteins encoded by the subsequent genes (rpsN-rplO) that are translationally coupled to rplE (66,67). Regulation of rplN and rplX, which occur 5’ of the S8 binding site, has also been reported (68). However, the mechanism of this retroregulation is not known.

Like several r-protein autogenous regulatory RNAs, the S8 binding site bears a strong resemblance to the S8 binding site on the 16S rRNA (64). Consistent with previous mutagenesis results (64,69) and three-dimensional structure information (70), the alignment indicates that the distal portion of the stem can be variable despite the large overlap between the RNA structure and the L5 coding sequence. In contrast to variability of the distal stem, the core S8 binding site within the internal bulge is highly conserved with few or no mutations (Figure 4). Bases that comprise the core S8-binding site and directly contact S8 in a crystal structure of the mRNA–protein complex (70) are well-conserved in our alignment.
The binding site is directly adjacent to the conserved pair of bulged adenosines (A8 and A9 in our diagram where numbering starts at the AUG) and extends through the conserved G12-C79 base pair. The bulged C5 and U2 have been implicated in reducing the S8 affinity for the mRNA over the rRNA (69). These bases did not contact the S8 protein in a crystal structure of the mRNA–S8 complex, and displayed high beta factors suggesting local flexibility (70). Our alignment shows that these bases are well conserved. However, it is not clear whether this conservation is due to functional constraints on the RNA or overlap with the L5-coding region.

The nearly 100% overlap of the S8 regulatory RNA with the L5 coding sequence introduces challenges for the computational homology searches. Purely sequence-based searches (e.g. BLAST) return spurious hits that lack a recognizable hairpin structure. The examples included in our alignment were stringently selected to preserve not only a highly conserved binding site, but also extended secondary structure in the variable region. In addition, there are several sequences originating from various species of Blochmania and Baumania that display a weak hairpin structure, but also contain mutations in the conserved binding site that are not likely to be compatible with binding. These sequences were also excluded from the alignment.

S7-interacting RNA

The str operon in E. coli consists of four genes, rpsL, rpsG, fusA and tuflA, that encode for ribosomal proteins S12, S7 and elongation factors EF-G and EF-tu (Figure 1). The translation of S12 and S7 is coupled, with independent S7 translation accounting for only 10–20% of protein production (71). S7 binds to the mRNA region between rpsL and rpsG and inhibits only the coupled translation of S12 and S7 (72). Retroregulation of S12 has been observed (73). The exact mechanism of action is unknown, but the mRNA structure formed in the absence of S7 binding is proposed to facilitate the coupled translation of S12 and S7 (74). On S7 binding, the mRNA structure and the translational coupling are disrupted.

Based on in vitro nuclease protection assays, there are two proposed binding sites for S7. Both of these sites resemble S7-binding sites on the rRNA (73,75). The first is composed of the 4–8 nucleotide bulge in H-III (Figure 4), and bases within the lower portion of H-III (5'-UGUAA and 5'-UGAAU in E. coli, 5'-AGUAA and 5'-UGAAN in our consensus diagram). However, the bulge is poorly conserved in the alignment compiled here. The second proposed binding site consists of the 5'-CCA and 5'-UUGGA sequences at the three-stem junction in the E. coli S7-binding mRNA (5'-CCA and 5'-UUGGR in our consensus diagram) (Figure 4). Bases implicated in S7-binding by UV-crosslinking (the UU from E. coli 5'-UUGGA) (74) are well-conserved. However, two of the three helices that form the three-stem junction (H-IV and H-V) are absent from the RNA in many Shewanella species, and there is a great deal of variability in both helices outside the regions proposed for S7-binding. In particular, H-IV is often not thermodynamically stable.

It has been reported that nearly one-third of all bacterial genomes contain an extended distance between rpsL and rpsG, and a sequence with the same genomic location and some similarity to the 16S ribosomal RNA has been described in Cyanobacteria (76). Our investigations of other bacterial phyla reveal that while extended distances between S12 and S7 exist in many species, the RNA structure observed in E. coli is not obviously conserved in these species.

S1-interacting RNA

The S1-binding RNA is somewhat unique because ribosomal protein S1 interactions with the ribosome are still a subject of debate (77). S1 is known to interact nonspecifically with many RNAs and is required for translation of many transcripts (78), including its own (18). The mRNA regulatory structure is found in the 5'-UTR of rpsA (Figure 1) and regulates S1 synthesis at the translational level (79). Ternary complexes composed of the RNA, S1 protein and 30S particles have been reconstituted in vitro (18).

Our alignment is consistent with previous phylogenetic analysis (18), indicating that the S1-interacting regulatory RNA is narrowly distributed (Figure 2). The conserved GG sequences in loops L1 and L2 are important for regulation (Figure 4) (18). The sequences of H2 and H3 are somewhat conserved, but the linker between them (single-stranded region 2, ss-2) is variable in sequence and in length, but it is nearly always AU rich. These findings are consistent with mutagenesis studies showing that the lower portion of helix H2 is important for regulation, but that the sequence and length of the second single-stranded region (SS-2) is not important for regulation (18). Toe-printing assays show that the S1 protein largely interacts with the AU-rich single stranded regions that occur between the helices and in the loop region of helix H3.

There are no additional experimentally verified examples of S1-binding autogenous regulatory motifs in other organisms. However, a putative RNA structure identified through comparative genomics precedes rpsA in many species of Cyanobacteria (23). This RNA bears no resemblance to the RNA in E. coli, and these RNAs are not included in our alignment.

S4-interacting RNA

The alpha operon in E. coli consists of five genes: rpsM, rpsK, rpsD, rpoA and rplQ, encoding ribosomal proteins S13, S11, S4, the α component of RNA polymerase and ribosomal protein L17, respectively (Figure 1). Ribosomal protein S4 regulates the synthesis of rpsM, rpsK, rpsD and rplQ as a translational repressor, but does not affect expression of the intervening rpoA (80,81). The RNA structure responsible for this regulation occurs in the 5'-UTR, partially overlaps rpsM and is a complex double pseudoknot (Figure 4). The start codon and a significant stretch of the rpsM protein coding sequence are contained within the structure. The mRNA and rRNA structures

...
bound by *E. coli* are different from one another (82), but the same domain of S4 protein appears to be responsible for binding both (83). The mechanism of action for this RNA involves the 30S particle and the RNA inhibits translation through ribosomal entrapment (9).

The consensus structure of the S4-interacting regulatory RNA produced by our alignment (Figure 4) is in good agreement with previous experimental studies. Helix H1 shows almost no sequence conservation, but does have extensive co-variation indicating the structure rather than the sequence is important (82). There is no co-variation apparent in the pseudoknotted structure, likely due to the overlap of these regions with the coding sequence. Mutagenesis studies have shown that compensatory mutations in H2 that maintain the stem do not restore S4-binding (82).

The existing Rfam seed alignment for this RNA [Rfam: RF00140] includes several examples that are not consistent with the available experimental evidence. These include examples from *Bordetella pertussis*, *Bordetella bronchiseptica*, *Dechloromonas aromatica* and *Pseudomonas syringae*. All of these sequences contain base changes inconsistent with one or more of the pseudoknots, and are missing two or more bases of the H2 stem, which is highly conserved and critical for S4 binding (82). After careful consideration, these sequences were removed and not included in the alignment. As a result of the large overlap between the RNA structure and the coding region of S4, purely sequenced-based search tools often return such spurious examples.

In many eubacteria, *rpsD* occurs in isolation in the genome, rather than as a portion of the alpha operon. As a result, studies to examine *rpsD* regulation in *B. subtilis* lead to the discovery of an mRNA structure that interacts with ribosomal protein S4 in this organism (84). However, the mRNA structure present in *B. subtilis* bears no obvious resemblance to the mRNA structure in *E. coli*.

**S15-interacting RNA**

The mRNA structure interacting with S15 only regulates *rpsO*, encoding S15 (Figure 1). The RNA structure partially overlaps the coding sequence for *rpsO* (85) and like the S4-interacting regulatory RNA, the S15-interacting RNA is pseudoknotted (86) and appears to act through a ribosomal entrapment mechanism (87). There are two possible structures (88), one of which is stabilized by the binding of S15 (Figure 4). The current Rfam alignment for this RNA [Rfam: RF00114] only reflects the alternative unbound structure for this RNA.

A great deal of mutagenesis as well as structural probing has been performed to elucidate the secondary and tertiary structure for the S15-interacting regulatory RNA (89–91), and our alignment is consistent with this data. Conserved structures corresponding to both possible conformations of the RNA are observed, although the pseudoknotted structure stabilized by S15 binding is better supported by the alignment with more co-varying positions and fewer non-conical pairs and insertions (Figure 4). The nucleotides in the lower portion of the H1 helix are variable, but base-pairing is largely maintained. Consistent with previous mutagenesis studies, the C-G/U=G base pairs in the upper portion of H1 and the purine between H1 and H2 are rigorously conserved (89) and likely play an important role in recognition. Deletion studies have shown that the L2 loop may be reduced to 7 nucleotides (including the AUG) (91), and our alignment shows that L2 is variable in length.

Two additional mRNA structures that interact with S15 to regulate *rpsO* in other organisms are known, one in *Geobacillus stearothermophilus* (92) and the other *Thermus thermophilus* (93). Neither of these RNAs bears any resemblance to the RNA structure in *E. coli*, and cross-species RNA–protein binding assays indicate that the S15 proteins in *E. coli* and *G. stearothermophilus* may have distinct determinants for binding their regulator RNAs (94).

**L20-interacting RNA**

There are two L20 binding sites in *E. coli*, and an alignment was generated for only one of these sites for reasons discussed above. L20 regulates the translation of itself and L35, and both the L20 regulatory sites are at the junction of *infC* and *rmpi*, encoding initiation factor IF3 and L35, respectively (95,96) (Figure 1). The two L20 binding sites are independent (7) and have similar in vitro affinity for L20 (19). We compiled an alignment for the L20 site encompassing the *infC/rmpi* intergenic region and some flanking sequence that overlaps the coding regions. While it is known that regulation occurs at the translational level, the exact mechanism of action is not currently known (7).

Like the binding sites for L1 and S8, the L20 binding site is proposed to act as a mimic of the ribosomal RNA (19). The bulge containing a rigorously conserved pair of adenosines (Figure 4) is similar to the L20 binding site on the 23S RNA (19). The pairing elements show little base-conservation, and the length of the helix is variable. All of these aspects are consistent with previous mutagenesis and truncation experiments (7,19).

Like S15, L20 interacts with a different mRNA structure in *B. subtilis* (97). In *B. subtilis*, the RNA structure occurs upstream of *infC*, and regulation occurs at the transcriptional level rather than the translational level. The RNA itself bears only a faint resemblance to the one described in *E. coli*, and thus is not included in our alignment. However, both RNAs are considered mimics of the ribosomal RNA and the C-terminal half of the *E. coli* L20 protein has the similar effects to L20 from *B. subtilis* during in vitro single-round transcription attenuation assays (97). These results indicate that the necessary binding determinates are replicated in each of the RNA structures despite their dissimilar structures.

**DISCUSSION**

Structured mRNA elements autogenously controlling ribosomal protein biosynthesis are common in *E. coli*, and their discovery and characterization represents some of the initial work performed on RNA-based gene regulation. Despite their importance for an essential process and
the extensive mechanistic studies performed, the regulatory RNAs responsible for ribosomal protein regulation are not well annotated in genomic databases. Owing to the technological limitations associated with RNA annotation and homology searches (98), and the lack of a complete archive of known RNAs, it is often difficult to determine whether a putative RNA structure represents a new finding or is a homologue of one that has been extensively studied in the past.

This study examines important RNAs in the light of new genomic data, places the consensus sequences and structures in the context of experimental data collected in the past and documents them for modern databases. In addition to creating alignments for many autogenous regulatory RNAs controlling r-protein synthesis, we have made several corrections to existing alignments to ensure that they are consistent with experimental data. We also experimentally validated some of our most surprising findings by verifying the transcription of dual L1-binding sites in *G. kaustophilus* and the interaction of both RNAs with L1 protein.

The exceedingly narrow distribution for most of the RNAs is striking. The presence of individual RNA structures within a few orders of Gammaproteobacteria has been remarked on in the past (16,17). However, there is no clear explanation for the explosion of regulatory mechanisms in the more recently evolved branches of Gammaproteobacteria (33). We speculate that most other bacteria do not lack such mechanisms, but rather a large number of distinct RNA structures are yet to be identified. The presence of RNA-based regulation of ribosomal protein biosynthesis in the vast majority of bacterial phyla and the sporadic identification of distinct RNA control structures in other bacterial phyla (84,91,97) support this hypothesis. In addition, as new sequence data continue to fuel comparative genomic-based non-coding RNA discovery efforts, RNA structures associated with ribosomal proteins are frequently encountered. For example, structured RNAs have been reported preceding the coding region of S1 and between the coding regions of S7 and S12 in Cyanobacteria (23,76), and several structured RNAs associated with ribosomal protein genes have been reported in *B. subtilis* and other Firmicutes (24,25).

The growing diversity of distinct RNA regulators using various mechanisms to regulate ribosomal protein genes across different bacterial phyla when operon structure is largely preserved suggests that there is significant pressure to tightly regulate ribosomal protein synthesis. In addition, most r-proteins are RNA-binding proteins, potentially increasing the probability of developing RNA-based autogenous regulatory mechanisms. However, it also seems that many distinct RNAs may solve the same biological problem. Indeed, diversity rather than conservation is becoming the dominant theme in RNA-based regulation in bacteria. RNA-based regulation of processes such as methionine (99) and sugar metabolism (100) is extraordinarily diverse, and trans-acting RNAs such as sRNAs evolve rapidly (101). This study illustrates the limited extent of our knowledge regarding ribosomal protein regulation in bacteria and sets the stage to allow the discovery and characterization of diverse RNA structures and mechanisms of regulation for ribosomal proteins in other microorganisms.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online: Supplementary Figure 1 and Supplementary Datasets.

**ACKNOWLEDGEMENTS**

We thank Dr. Zasha Weinberg and Prof. Ming Hammond for the discussions that lead to this work and Prof. Ronald Breaker for the freedom to work on it. We also thank the Yale University Life Sciences High Performance Computing Center and the High Performance Computing Center at Boston College. Genomic DNA for *G. kaustophilus* was a kind gift from Hideo Takami at the Japan Agency of Marine-Earth Science and Technology, Microbial Genome Research Group.

**FUNDING**

Boston College; the Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation and the Alfred P. Sloan Foundation. Funding for open access charge: Alfred P. Sloan Foundation.

**Conflict of interest statement.** None declared.

**REFERENCES**


Zengel, J.M. and Lindahl, L. (1992) Ribosomal protein L4 and


Downloaded from https://academic.oup.com/nar/article-abstract/41/6/3491/2902762 by guest on 14 March 2019


