Genes within Genes in Bacterial Genomes

SEZEN MEYDAN,1 NORA VÁZQUEZ-LASLOP,1 and ALEXANDER S. MANKIN1

1Center for Biomolecular Sciences, University of Illinois at Chicago, Chicago, IL 60607

ABSTRACT Genetic coding in bacteria largely operates via the "one gene-one protein" paradigm. However, the peculiarities of the mRNA structure, the versatility of the genetic code, and the dynamic nature of translation sometimes allow organisms to deviate from the standard rules of protein encoding. Bacteria can use several unorthodox modes of translation to express more than one protein from a single mRNA cistron. One such alternative path is the use of additional translation initiation sites within the gene. Proteins whose translation is initiated at different start sites within the same reading frame will differ in their N termini but will have identical C-terminal segments. On the other hand, alternative initiation of translation in a register different from the frame dictated by the primary start codon will yield a protein whose sequence is entirely different from the one encoded in the main frame. The use of internal mRNA codons as translation start sites is controlled by the nucleotide sequence and the mRNA folding. The proteins of the alternative proteome generated via the "genes-within-genes" strategy may carry important functions. In this review, we summarize the currently known examples of bacterial genes encoding more than one protein due to the utilization of additional translation start sites and discuss the known or proposed functions of the alternative polypeptides in relation to the main protein product of the gene. We also discuss recent proteome- and genome-wide approaches that will allow the discovery of novel translation initiation sites in a systematic fashion.

INTRODUCTION

The most common result of the translation of a gene is the production of a single protein product (Fig. 1a). However, the redundancy of the genetic code and the plasticity of the mRNA structure allow for expansions of the proteome by unorthodox interpreting of genetic information. A familiar strategy leading to unusual interpretation of genetic information is collectively known as recoding and has been discussed in several excellent reviews (1-4). The most conventional recoding involves programmed ribosomal framesshifting and can generate two gene products that are identical in their N-terminal segments but differ in the sequences of their C termini (Fig. 1b).

A much less studied and discussed type of unconventional translation strategy leads to the production of two (or in rare cases more than two) distinct proteins from one gene that differ in their N termini but have identical C-terminal structures. This scenario occurs when translation of a gene is initiated not only at the primary translation initiation site (pTIS) but also at an additional, in-frame, internal translation initiation site (iTIS) (Fig. 1c). Furthermore, if the iTIS directs the ribosome to begin protein synthesis at an out-of-frame (OOF) start codon, the amino acid sequence of the resulting protein would be principally different from that of the main-frame protein (Fig. 1c).
Internal initiation is controlled by the mRNA structure in the immediate neighborhood of the iTIS. Additionally, it may also be influenced by the ribosome traffic through the iTIS, which depends on the structure of a remote mRNA segment in the vicinity of the pTIS.

Internal initiation can lead to production of functional proteins that could play beneficial or even essential roles in cell physiology. Therefore, discovery and analysis of unconventional translation scenarios and identification of the functions of the alternative protein products are critical for understanding both the principles of translation regulation and the complexity of physiological networks. In the following pages we will discuss the known examples of bacterial genes that use internal initiation to encode more than one polypeptide. Most of these examples have been discovered serendipitously and have been analyzed with a varying degree of scrutiny. For consistency, we retained the designations of the protein isoforms used by the authors of the original papers. We will also discuss the emerging strategies that may pave the way for a more systematic identification of genes with more than one start codon. Although extensive literature exists about overlapping genes in bacteriophages (5–7), we are unaware of any published compilations of bacterial genes with alternative sites of translation initiation. We apologize to our colleagues whose findings relevant to the topic of this review have evaded our quest.

INITIATION OF TRANSLATION IN BACTERIA

Initiation of translation in bacteria is driven by the small ribosomal subunit (30S). The 30S subunit locates a ribosome binding site in mRNA and, with the help of translation initiation factors, binds the initiator fMet-tRNA and positions the start codon of the open reading frame...
(ORF) in the small subunit P-site (reviewed in reference 8). Upon the large subunit (50S) joining and dissociation of the initiation factors, an elongator tRNA binds in the ribosomal A-site, the first peptide bond is formed, and the ribosome transitions into the elongation phase of protein synthesis (8, 9).

Identifying the start codon of the protein-coding sequence in mRNA is one of the most critical tasks of the ribosome (8, 10). Not only does the start codon define the boundary of the mRNA segment to be translated, but it also sets the frame in which the ribosome reads the genetic message. AUG is the most commonly used start codon, but other triplets (GUG, UUG, CUG, AUU, AUC, and AUA) can also be employed for translation initiation (8, 11). Utilization of a codon as a translation initiation site (TIS) is often assisted by the Shine-Dalgarno (SD) sequence (with the consensus AGGAGG), which is complementary to the 3′ end of the 16S rRNA of the 30S subunit (12). The SD sequence is positioned 4 to 9 nucleotides (nt) upstream of the start codon and plays an important role in modulating the efficiency of initiation of translation. However, its presence is not sufficient or even required for defining the start codon of an ORF (8). In fact, some bacterial phyla make little use of the SD sequence (13). But even in bacteria that widely exploit the SD sequence, some SD-lacking mRNAs, as well as leaderless mRNAs devoid of the 5′ untranslated region (5′ UTR), can also be efficiently translated (14). The efficiency of translation initiation is also significantly affected by the accessibility of the TIS (2, 15). Sequestering the TIS in mRNA secondary or tertiary structure could dramatically reduce the initiation rate, whereas highly accessible TISs are more conducive to active translation. Binding of small, trans-acting RNAs or proteins to the TIS provides an additional layer of regulation of translation initiation (16).

The pTIS and the first in-frame stop codon delineate the boundaries of an ORF. The constellation of the structural elements favoring translation initiation, the known sequence of the encoded protein, the evolutionary conservation, and the codon-related nucleotide bias are usually employed to properly annotate the placement of the start codons of the bacterial genes. However, any of the in-frame or OOF internal codons that can be decoded by fMet-tRNAf Met may potentially define an additional site (iTIS) where translation could start. Utilization of unwanted iTISs can be precluded by occluding them in unfavorable higher-order structure of mRNA (17). Some iTISs, however, can be exploited for expanding the proteome by making two proteins from a single gene.

Throughout this review, we will refer to the proteins whose translation is initiated at the pTIS as the primary or full-size protein. We will call the polypeptide product translated from an iTIS the alternative protein, which in most cases means a second protein expressed in addition to, not instead of, the primary product (Fig. 1c). Some ORFs even carry multiple iTISs and therefore encode several alternative proteins.

Most of the known examples of internal initiation have been discovered inadvertently, when the production of two polypeptides from a single gene was noticed in vivo or in vitro. A special effort was required to demonstrate that the production of the smaller protein was not the result of proteolytic cleavage of the full-size polypeptide. In several cases the function of the internal initiation product has been established, but in many instances the purpose of production of the alternative protein remains obscure. We will present these examples according to the functionality of the alternative polypeptide in regard to that of the primary protein, with the full understanding that the assignment to a specific category is inevitably arbitrary.

### Internal Initiation from In-Frame Start Codons

#### The Function of the Alternative Polypeptide Is Closely Related to That of the Primary Protein

Alternative IF2 isoforms function as initiation factors

Production of more than one protein from the Escherichia coli infB gene, which encodes the translation initiation factor 2 (IF2), presents one of the best-characterized examples of exploiting internal initiation for generating functional isoforms of the main translation product. The existence of two IF2 isoforms, IF2-1 (formerly IF2α), which is the primary protein, and IF2-2 (IF2β), the alternative product, was noticed during the chromatographic analysis of the purified protein (18). Sequencing of the infB gene helped establish the GUC158 codon as the iTIS that directs translation of IF2-2 (19) (Table 1). The use of GUC158 as a true start codon was additionally verified when the N-terminal amino acid sequences of the purified IF2 isoforms were directly determined by Edman degradation (20). Subsequent experiments revealed the presence of the third, IF2-3 isoform (IF2γ), whose translation is initiated at AUG165, 7 codons downstream from the iTIS of IF2-2 (21, 22). The conservation of iTISs in infB and the expression of two or three IF2 isoforms have been demonstrated in a wide range of bacteria (23, 24) (Table 1).

---

**Table 1**

<table>
<thead>
<tr>
<th>Internal Initiation from In-Frame Start Codons</th>
<th>IF2 Isoforms</th>
<th>Function of Alternative Polypeptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closely Related to That of the Primary Protein</td>
<td>IF2-1 (IF2α)</td>
<td>IF2-2 (IF2β)</td>
</tr>
<tr>
<td>Function of the Alternative Polypeptide Is</td>
<td>IF2-1 (IF2α)</td>
<td>IF2-2 (IF2β)</td>
</tr>
</tbody>
</table>

**References**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Length of the primary protein (amino acids)</th>
<th>Length of the alternative protein (amino acids)</th>
<th>Primary SD sequence</th>
<th>Spacer length (nt)</th>
<th>Primary start codon</th>
<th>Internal SD sequence</th>
<th>Spacer length (nt)</th>
<th>Internal start codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>infB</td>
<td>E. coli</td>
<td>890</td>
<td>733</td>
<td>AAGGA</td>
<td>5</td>
<td>AUG</td>
<td>GGAAAAAG</td>
<td>5</td>
<td>GUG</td>
</tr>
<tr>
<td>infB</td>
<td>B. subtilis</td>
<td>716</td>
<td>623</td>
<td>GGGGUG</td>
<td>5</td>
<td>AUG</td>
<td>AAGAAAG</td>
<td>6</td>
<td>GUG</td>
</tr>
<tr>
<td>rpoS</td>
<td>E. coli K-12 ZK126</td>
<td>330</td>
<td>277</td>
<td>AGGAG</td>
<td>7</td>
<td>AUG</td>
<td>AGGAG</td>
<td>11</td>
<td>GUG</td>
</tr>
<tr>
<td>mrcB</td>
<td>E. coli</td>
<td>844</td>
<td>799</td>
<td>GGAG</td>
<td>8</td>
<td>AUG</td>
<td>GAAAAAA</td>
<td>15</td>
<td>AUG</td>
</tr>
<tr>
<td>mip</td>
<td>C. burnetii</td>
<td>230</td>
<td>140</td>
<td>AGGAG</td>
<td>8</td>
<td>AUG</td>
<td>AAAA</td>
<td>11</td>
<td>AUG</td>
</tr>
<tr>
<td>pikAV</td>
<td>S. venezuelae</td>
<td>1,346</td>
<td>1,146</td>
<td>GGAAG</td>
<td>11</td>
<td>AUG</td>
<td>GGAGGG</td>
<td>7</td>
<td>GUG</td>
</tr>
<tr>
<td>trfA</td>
<td>E. coli</td>
<td>382</td>
<td>285</td>
<td>AGGAGG</td>
<td>4</td>
<td>AUG</td>
<td>GAGG</td>
<td>4</td>
<td>AUG</td>
</tr>
<tr>
<td>clpB</td>
<td>E. coli</td>
<td>857</td>
<td>709</td>
<td>GGAGG</td>
<td>4</td>
<td>AUG</td>
<td>GGAGG</td>
<td>7</td>
<td>GUG</td>
</tr>
<tr>
<td>clpB</td>
<td>Synechococcus 7942</td>
<td>874</td>
<td>724</td>
<td>-</td>
<td>-</td>
<td>AUG</td>
<td>GAGG</td>
<td>10</td>
<td>GUG</td>
</tr>
<tr>
<td>clpA</td>
<td>E. coli</td>
<td>758</td>
<td>590</td>
<td>GGAGG</td>
<td>5</td>
<td>AUG</td>
<td>GGAGG</td>
<td>5</td>
<td>AUG</td>
</tr>
<tr>
<td>cvaA</td>
<td>Avian pathogenic E. coli</td>
<td>413</td>
<td>253</td>
<td>GGAGG</td>
<td>7</td>
<td>UUG</td>
<td>AGGAAG</td>
<td>11</td>
<td>AUG</td>
</tr>
<tr>
<td>lcnD</td>
<td>L. lactis</td>
<td>474</td>
<td>429</td>
<td>AGGAGG</td>
<td>6</td>
<td>AUG</td>
<td>AGGAG</td>
<td>0</td>
<td>UUG</td>
</tr>
<tr>
<td>mcrB</td>
<td>E. coli</td>
<td>459</td>
<td>298</td>
<td>AGGAAAG</td>
<td>9</td>
<td>AUG</td>
<td>AAGAG</td>
<td>13</td>
<td>AUG</td>
</tr>
<tr>
<td>ccmM</td>
<td>Synechococcus 7942</td>
<td>539</td>
<td>324</td>
<td>GGAGG</td>
<td>7</td>
<td>AUG</td>
<td>GAGG</td>
<td>8</td>
<td>GUG</td>
</tr>
<tr>
<td>ssaQ</td>
<td>S. Typhimurium</td>
<td>322</td>
<td>106</td>
<td>GGAGG</td>
<td>7</td>
<td>AUG</td>
<td>AGGA</td>
<td>8</td>
<td>AUG</td>
</tr>
<tr>
<td>cheA</td>
<td>E. coli</td>
<td>654</td>
<td>557</td>
<td>GAGG</td>
<td>7</td>
<td>GUG</td>
<td>AAGA</td>
<td>4</td>
<td>AUG</td>
</tr>
<tr>
<td>lysC</td>
<td>B. subtilis</td>
<td>408</td>
<td>163</td>
<td>AAAGG</td>
<td>10</td>
<td>AUG</td>
<td>AGGAAG</td>
<td>8</td>
<td>AUG</td>
</tr>
<tr>
<td>tilt/hprT</td>
<td>L. monocytogenes</td>
<td>648</td>
<td>179</td>
<td>AGGAAG</td>
<td>10</td>
<td>AUG</td>
<td>AGGAAG</td>
<td>10</td>
<td>AUG</td>
</tr>
<tr>
<td>pgaM</td>
<td>Streptomyces sp. PG64</td>
<td>784</td>
<td>253</td>
<td>GGAAAG</td>
<td>12</td>
<td>AUG</td>
<td>GAGA</td>
<td>7</td>
<td>AUG</td>
</tr>
<tr>
<td>incC</td>
<td>E. coli</td>
<td>364</td>
<td>259</td>
<td>GGAGG</td>
<td>6</td>
<td>AUG</td>
<td>AGGA</td>
<td>9</td>
<td>AUG</td>
</tr>
<tr>
<td>petH</td>
<td>Synechocystis 6803</td>
<td>413</td>
<td>301</td>
<td>GGAGG</td>
<td>10</td>
<td>AUG</td>
<td>-</td>
<td>-</td>
<td>AUG</td>
</tr>
<tr>
<td>fliO</td>
<td>S. Typhimurium</td>
<td>125</td>
<td>104</td>
<td>GAGG</td>
<td>6</td>
<td>AUG</td>
<td>GAGA</td>
<td>9</td>
<td>AUG</td>
</tr>
<tr>
<td>mwp</td>
<td>B. brevis 47</td>
<td>1,084</td>
<td>1,053</td>
<td>AGGAAG</td>
<td>10</td>
<td>UUG</td>
<td>AGGA</td>
<td>11</td>
<td>AUG</td>
</tr>
<tr>
<td>safA</td>
<td>B. subtilis</td>
<td>387</td>
<td>224</td>
<td>AGGAAG</td>
<td>8</td>
<td>UUG</td>
<td>AGGAAG</td>
<td>11</td>
<td>AUG</td>
</tr>
<tr>
<td>hlyB</td>
<td>E. coli</td>
<td>707</td>
<td>422</td>
<td>GGAGG</td>
<td>5</td>
<td>AUG</td>
<td>GCGG</td>
<td>1</td>
<td>AUG</td>
</tr>
<tr>
<td>plcR</td>
<td>P. aeruginosa</td>
<td>207</td>
<td>151</td>
<td>AGGAAG</td>
<td>8</td>
<td>AUG</td>
<td>CCGA</td>
<td>8</td>
<td>AUG</td>
</tr>
<tr>
<td>srfAB</td>
<td>B. subtilis</td>
<td>3,583</td>
<td>46</td>
<td>AGGAAG</td>
<td>6</td>
<td>AUG</td>
<td>AGGA</td>
<td>9</td>
<td>UUG</td>
</tr>
<tr>
<td>rnpA</td>
<td>T. thermophilus</td>
<td>163</td>
<td>77</td>
<td>GGAGG</td>
<td>3</td>
<td>AUG</td>
<td>GGAGG</td>
<td>7</td>
<td>AUG</td>
</tr>
<tr>
<td>gnd</td>
<td>E. coli</td>
<td>468</td>
<td>36–53</td>
<td>AGGAAG</td>
<td>6</td>
<td>AUG</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
</tbody>
</table>

*Indicates the OOF internal start sites.
In vitro assays suggested that not only IF2-1 but also the IF2-2/3 variants are true initiation factors: they all associate with the 30S subunit and stimulate binding of fMet-tRNA (25). However, the activities of IF2-1 and IF2-2/3 are likely not redundant. Both major isoforms (IF2-1 and IF2-2) are necessary for optimal cell fitness, since deletion of infB is fully complemented only when both IF2 isoforms are expressed. Consistently, synthesis of either only IF2-1 or IF2-2 is insufficient to support maximal cell growth and results in cold sensitivity (22).

The infB gene belongs to the cold stress regulon, and its expression is additionally stimulated when cells are exposed to low temperatures. While IF2-1 and IF2-2 proteins are produced in approximately equimolar amounts in E. coli grown at 37°C, shifting temperature to 20°C results in higher expression of IF2-2 compared to IF2-1 (26). This suggests that usage of the infB TISs could be regulated in order to adjust the relative abundance of the IF2 isoforms for specific needs of the cell.

It is not known, however, whether there are differences in the functional properties of the IF2 variants. Conceivably, the IF2 isoforms could play particular roles in expressing mRNAs with different translation initiation signals, e.g., SD-containing cistrons versus leaderless ones or those exploiting certain start codons. It is also possible that the reasons for the existence of IF2 variants are not even related to the function of the factor in translation: IF2 could be also involved in facilitating the restart of DNA replication after exposing cells to DNA-damaging agents (27, 28). The exact role of IF2-1 in this process is unclear, but it is remarkable that only IF2-1, but not IF2-2/3, is capable of promoting replication restart (29).

Several attempts have been made to understand the regulation of internal initiation in infB. It has been proposed that the presence of rare codons upstream of the iTIS could result in reduced ribosome traffic through the internal start site, thereby increasing the probability of translation initiation at the iTIS (19). However, ribosome profiling (Ribo-seq) experiments yielded no evidence of such an effect (30). The propensity of mRNA to adopt an unfolded state (23) or its possible folding into a pseudoknot (31) have also been considered as factors facilitating initiation at the infB iTISs, but no experimental evidence supporting these proposals has been reported. Thus, at the moment, the principles of regulation of the internal initiation event in infB, in spite of it being one of the best-studied examples, remain unclear.

A truncated RNA polymerase σ\(^5\) factor allows survival of mutant cells

The rpoS gene codes for the RNA polymerase σ\(^5\) factor, a key transcription master regulator of the genes whose products are needed in the stationary phase and under stress. Puzzlingly, the rpoS gene of several E. coli K-12 strains carries a nonsense mutation replacing the CAG\(_{33}\) (or, in some strains, GAG) codon with the UAG stop codon (32, 33). Nevertheless, growth of the mutant cells was unaffected despite the presence of the premature stop codon in rpoS. This is possible because the ribosomes can initiate translation from an iTIS located in rpoS a short distance downstream from the premature stop codon (at GUG\(_{54}\)). Internal initiation leads to production of an N-terminally truncated, but nevertheless functional, variant of the σ\(^5\) factor (34) (Table 1). No expression of truncated σ\(^5\) was observed in the strains expressing rpoS lacking the premature stop codon (35). This may indicate that the iTIS becomes available for initiation only in the absence of translating ribosome traffic. However, no systematic analysis of potential co-expression of two rpoS isoforms under different physiological conditions has been undertaken, and it remains possible that the presence of an iTIS even in the wild-type rpoS gene is physiologically meaningful.

The redundant PBP-1b isoforms

The E. coli mrcB gene encodes one of the major penicillin binding proteins, known as PBP-1b or MrcB, a peptidoglycan synthase with transglycosylation and transpeptidation activities. Analysis of the purified MrcB protein by SDS-PAGE revealed three distinct bands designated as MrcB-α, -β, and -γ (36, 37). While the presence of the β variant was likely an artifact of isolation resulting from protease cleavage of MrcB-α (38), MrcB-γ was shown to be the product of internal initiation at the mrcB AUG\(_{46}\) codon (39) (Table 1). The α and γ isoforms assemble into α\(_2\) or γ\(_2\) homodimers (but not an αγ heterodimer) that are found in the membrane fraction (37, 40). They also show similar enzymatic activities (37). However, the biological reasons for coproduction of the full-length MrcB-α and its truncated MrcB-γ variant remain enigmatic, moreover that expression of the γ variant is sufficient to fully compensate for the mrcB deletion (36).

Virulence factor Mip variants with similar activities but distinct cellular localization

Macrophage infectivity potentiator (Mip) proteins are virulence factors of many pathogenic bacteria. These proteins possess peptidyl-prolyl cis/trans isomerase ac-
tivity and are involved in protein folding, maturation, and targeting. Bacteria often carry several different mip genes, some of which encode proteins that are retained in the cytoplasm whereas the products of others are targeted to the outer membrane or the periplasm (41). In contrast, the compact genome of the obligate intracellular parasite Coxiella burnetii contains only one mip gene (cbmip), encoding a Mip-like protein, CbMip (42). Expression of cbmip in E. coli led to the production of three polypeptides (42). Sequencing of the N termini of these products and mutational analysis of the gene demonstrated that the two shorter forms of CbMip were generated due to initiation of translation at two iTISs present within cbmip (42) (Table 1). Similar CbMip isoforms were also detected in the native host. Even though all three CbMip variants have comparable enzymatic activities, the full-size CbMip carries a signal sequence whereas the products of internal initiation lack it. Therefore, the primary CbMip is exported from the cell, but the shorter isoforms are retained in the cytoplasm (42) (Fig. 2). Being an obligate intracellular parasite, C. burnetii has a rather streamlined genome, encoding merely 2,000 genes (43). Therefore, it could be that C. burnetii employs internal initiation to diversify the repertoire of the Mip proteins without expanding the size of its genome. It is yet to be investigated whether the relative expression of the intracellular and secreted variants of CbMip is regulated.

**Polyketide synthase isoforms produce two different antibiotics**

*Streptomyces venezuelae* produces two major macrolide antibiotics, the 14-atom macrolactone-ring pikromycin and a smaller, 12-atom macrolactone, methymycin (Fig. 3) (44). Both antibiotics inhibit protein synthesis by targeting the bacterial ribosome (45). The pikAIV to pikBIV genes of the biosynthetic operon encode modular polyketide synthases that are required for generating the macrolactone core. The product of the pikAIV gene is responsible for the last condensation step during biosynthesis of the 14-atom macrolactone of pikromycin. How the smaller methymycin core is generated remains an enigma for many years. The puzzle was solved when genetic and biochemical experiments revealed that initiation of translation from an iTIS within the pikAIV gene generates an N-terminally truncated version of the PikAIV protein (46) (Table 1). The truncated PikAIV contains intact acyl carrier protein, acyltransferase, and thioesterase domains of the full-size PikAIV, but only a portion of the first, ketosynthase domain. As a result, the last condensation step of the macrolactone synthesis does not take place and, instead, the 12-atom macrolactone of methymycin is produced (46) (Fig. 3). Remarkably, the composition of the growth media defines which of the two antibiotics is preferentially synthesized in *S. venezuelae* cells (47, 48), suggesting that the use of the pTIS or the iTIS of pikAIV is regulated in response to environmental cues. However, the mechanism controlling such regulation remains unknown.

**RK2 plasmid copy number in different hosts is controlled by TrfA isoforms**

Replication of the broad-range Gram-negative bacterial plasmid RK2 relies on the coordinated activity of plasmid-encoded replication initiation protein TrfA and host proteins (reviewed in reference 49). TrfA is expressed in two versions, the full-size TrfA-44 and its N-terminally truncated TrfA-33 isoform, whose translation is initiated at the internal trfA codon AUG₉₈ (50, 51) (Table 1). Each of the individual isoforms is sufficient to support plasmid replication in several bacterial species (52). Nevertheless, TrfA-44 and TrfA-33 probably play distinct roles in controlling plasmid copy number in specific bacterial hosts. For example, the lack of TrfA-44 negatively affects RK2 replication in *Pseudomonas aeruginosa*, likely because the DnaB protein of this host that...
is required for RK2 replication is unable to properly interact with TrfA-33 (53–55). Conversely, the lack of TrfA-33 reduces the RK2 copy number in E. coli and Azotobacter vinelandii (56), suggesting that in these bacteria the internal initiation product fulfills an important supportive function. It is unknown if the relative efficiency of utilization of the pTIS and iTIS is regulated and whether it is host specific.

Alternative Protein Facilitates the Function of the Primary Protein

Isoforms of ClpB and ClpA aid in handling the misfolded and aggregated proteins

Chaperones and proteases help cells deal with misfolded and aggregated proteins that accumulate under stress conditions (57). One such protein is ClpB, a chaperone that utilizes the energy of ATP hydrolysis to disaggregate polypeptides. Gel-electrophoretic analysis of the E. coli proteins induced during heat shock identified a shorter isoform of ClpB, ClpB79, in addition to the full-length ClpB93. Mutagenesis studies demonstrated that an iTIS at codon GUG<sup>149</sup> of clpB was required for production of ClpB79 (58, 59) (Table 1). While both isoforms retain the two ATP-binding sites, ClpB79 lacks the N-terminal domain of its full-size counterpart, which is involved in substrate recognition (58). Nevertheless, the ATPase, chaperone, and oligomerization activities of ClpB79 seem to be largely unaffected (60). ClpB functions as a homotetrameric complex in which both isoforms could be present simultaneously (58). Both ClpB versions are also produced in several other bacteria (60, 61) (Table 1). The mechanistic advantage of their coproduction is unclear, but it has been shown that E. coli or Synechococcus sp. handles thermal stress more efficiently when both variants are coordinately expressed (62–64), while in other species, the internal initiation product alone can afford the same level of thermoprotection as the simultaneous expression of both ClpBs (63, 65). Interestingly, the ratio of the ClpB isoforms varies depending on the severity of the heat shock, suggesting that the relative activity of the pTIS and iTIS could be regulated (62).

A protein-dependent ATPase, ClpA, shows significant sequence similarity to ClpB. ClpA associates with the proteolytic subunit ClpP to form the Clp protease. Similar to the existence of the ClpB isoforms, two ClpA variants are expressed from the clpA gene, the main protein ClpA84 and the alternative polypeptide ClpA65, whose translation is initiated at the AUG<sub>169</sub> codon of clpA (66) (Table 1). In the absence of substrates, the Clp complex is prone to autodegradation because idling ClpP can cleave the ClpA84 subunit. However, ClpA65 represses Clp self-proteolysis by acting as a decoy.
suggested that the product of internal initiation plays a regulatory role in maintaining protein homeostasis (67).

While the precise physiological roles of the internal initiation products of clpA and clpB genes are yet to be fully elucidated, the conservation of the iTIS in these genes in several bacterial species argues that the ability to produce N-terminally truncated versions of these heat shock proteins is important for coping with thermal shock. An attractive hypothesis is that temperature-dependent alterations in the mRNA structure could affect the relative efficiency of translation initiation at the primary and internal start codons (68). However, this possibility has not been explored.

Alternative transport proteins facilitate secretion of bacteriocins

Several E. coli strains secrete the antibacterial toxin colicin V (ColV). The secretion of ColV requires the activity of the membrane-associated transport proteins CvaA and CvaB (69, 70). Besides full-size CvaA, the shorter CvaA* variant is translated from the same cvaA gene, due to translation initiation at the AUG161 codon (69, 70) (Table 1). Coexpression of CvaA and CvaA* is required for optimal secretion of ColV. Similar to the differential targeting of the Mip isoforms discussed above (Fig. 2), the primary protein CvaA is bound to the membrane, but CvaA*, which lacks the hydrophobic N-terminal segment of CvaA, remains in the cytoplasm (70). How the CvaA variants cooperate in secretion of the ColV toxin is unknown, but reminiscent of the stabilization of ClpA by its shorter isoform, CvaA* protects the full-length CvaA from degradation (70). The regulation of relative expression of CvaA and CvaA* has not been yet investigated.

A comparable mechanism is likely involved in production of another bacteriocin, lactococcin A (LcnA), secreted by strains of Lactococcus lactis. Export of LcnA relies on the activity of two membrane proteins, LcnC and LcnD. As for CvaA*, the existence of an iTIS in lenD leads to expression of a full-size membrane-associated LcnD and a shorter version, LcnD*, which is retained in the cytoplasm (71) (Table 1). Functions of LcnD variants in LcnA secretion have not been characterized.

The alternative McrB protein tunes the activity of a restriction enzyme

The E. coli McrBC restriction enzyme cleaves foreign DNA containing 5-methylcytosine (72). The active enzyme is formed by association of the nucleolytic McrC polypeptide with seven copies of the GTPase subunit McrB (73) (Fig. 4).

The mcrB gene directs production of two polypeptides, the full-size protein McrB1 and its N-terminally truncated variant McrB5, whose translation is initiated at the AUG162 codon (74) (Table 1). Because McrB5 lacks the DNA-binding domain of McrB1, its complex with McrC is enzymatically inactive, which raised the question of the biological role of McrB5 (73). The possible solution to the puzzle came from the observation that an excess of McrC leads to reduced activity of the enzyme, whereas the simultaneous production of McrB5 remedies the problem (75, 76). Overproduction of McrC distorts the ratio of the subunits in the McrB complex, thus decreasing its overall enzymatic activity. McrB5 may trap the excessive amounts of McrC, thereby restoring the optimal stoichiometry of the functional McrBC enzyme (Fig. 4).

A delicate balance of the amounts of McrB1, McrB5, and McrC is likely required for optimal activity of the restriction nuclease in the cell (75), but whether such balance is achieved by regulating the relative expression of the McrB isoforms is unknown.

The internal initiation product of CcmM is required for organizing the structure of β-carboxysomes

RubisCO plays the central role in CO2 fixation during photosynthesis. To stimulate this reaction, several cyanobacterial strains form specialized microcompartments called β-carboxysomes that help increase the local concentration of CO2 (77). Immunoblot analysis of the β-carboxysome polypeptides in Synechococcus sp. strain PCC 7942 revealed the presence of two isoforms of the protein CcmM, full-size CcmM-58 and the N-terminally truncated CcmM-35. They are translated from the same gene due to the presence of an iTIS at the ccmM GUG216 codon (78, 79) (Table 1). The CcmM isoforms carry out specific and distinct structural functions (78). The N-terminal domain of CcmM-58 facilitates the formation of protein trimers and localizes them to the inner shell of the carboxysome; the C-terminal segments of CcmM-58 interlink RubisCO holoenzymes (Fig. 5). The shorter CcmM-35, which lacks the N-terminal domain of CcmM-58, is located in the lumen of the β-carboxysome, where it organizes RubisCO in a paracrystalline array (77, 78, 80) (Fig. 5). The ccmM genes of several other cyanobacterial species also contain a similarly positioned iTIS (72), suggesting that the use of internal initiation is a successful strategy for expressing polypeptides with related functions but unique patterns of interprotein interactions.
Assembly of type III secretion systems relies upon expression of the C-ring protein variant

Some pathogenic bacteria, including *Salmonella enterica* serovar Typhimurium, use a type III secretion system (T3SS), also known as an injectisome, to export effector proteins into the invaded cells. The core of the T3SS is a highly complex, multiprotein, needle-like structure with rings of specific proteins in the inner and outer membranes (reviewed in reference 81). The protein product of the *ssaQ* gene oligomerizes to form the cytoplasmic ring (C-ring) of the injectisome. However, two polypeptides are expressed from *ssaQ*, the full-size SsaQL and the shorter SsaQS, the latter due to internal initiation at the AUG217 codon (82) (Table 1). *Salmonella* strains unable to express SsaQS are attenuated in their virulence (82). A similar organization of the genes analogous to *ssaQ* of *S. Typhimurium* has been found in *Yersinia*, *Shigella*, and *Thermotoga* strains (83, 84) (Table 1). In all these bacteria, expression of the N-terminally truncated protein is necessary for the formation of a functional T3SS. Although the exact function of SsaQS remains obscure, it has been shown that the internal
initiation product stabilizes SsaQ₁ and augments the activity of the T3SS (82). An unexpected validation of the importance of coexpression of the two isoforms of the C-ring protein came from the attempts of synthetic biologists to assemble a functional injectisome using a minimal set of synthetic genes resembling the T3SS genes of the Salmonella pathogenicity island 1 (85). While the initial gene assembly was inactive, subsequent debugging pinned down the problem to the inadvertent elimination of an iTIS in the spaO gene (the functional homolog of ssaQ). Subsequent restoration of the iTIS allowed the expression of the functional T3SS (85).

The T3SS injectisome is evolutionarily and structurally related to the flagellar machinery of Gram-negative bacteria (81). Interestingly, while SsaQ₁ and SsaQ₅ (and the equivalent injectisome components in other bacteria) are expressed from a single gene due to the presence of iTISs, their structural equivalents FliM/FliN of the flagellar C-ring are expressed from two individual genes. It remains to be investigated why coexpression of the C-ring components from a single gene is beneficial for the formation and function of the injectisome but not the flagella.

Involvement of CheA₁ and CheA₅ in chemotaxis

Many Gram-negative bacteria sense and respond to gradients of chemical attractants or repellents by functionally coupling membrane-bound receptors with the rotation of their flagella. A central regulatory hub of this mechanism is the signal transduction system, which operates through phosphorylation of the response regulator CheY. CheA is the sensory histidine kinase responsible for CheY phosphorylation, and CheZ is a phosphatase that dephosphorylates CheY (reviewed in reference 86). The CheY-regulated complex biochemical cascade eventually helps bacteria swim toward the attractant.

The cheA gene encodes two polypeptides, the full-size CheA₁ and the smaller CheA₅, whose translation is driven by the internal AUG₉₈ codon (87, 88) (Table 1). Although CheA₅ lacks His₄₈, the site of CheA₁ auto-phosphorylation, various functions have been attributed to this CheA isoform: from facilitating phosphorylation of CheA₁ (89) to mediating protein-protein interactions important for chemotaxis (90). Nevertheless, despite decades of research, the precise role of CheA₅ in chemotaxis remains obscure. Interestingly, the internal start codon of E. coli cheA is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87).

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.

The alternative product of the lysC gene expands the amino acid-mediated regulation of aspartate kinases

Aspartate kinases (AKs) catalyze the first step of the pathways leading to the synthesis of several essential amino acids, including lysine and threonine. The catalytic domain of AKs is linked to the regulatory domain responsible for the negative feedback regulation by specific amino acids, the end product(s) of the pathway. In many organisms, AKs assemble into multimers, where the binding site(s) for the regulatory amino acid(s) is located within the loop of a hairpin formed by the mRNA segments flanking the iTIS, suggesting that the access to it and the iTIS usage could be modulated by mRNA structure (87). To the best of our knowledge, no evidence for such regulation has been reported so far.
The function of the TilS-HprT protein complex in Listeria relies on the iTIS of the fused genes. The genes of the tRNA\textsuperscript{Ile} lysidine synthetase, TilS, involved in biogenesis of tRNA\textsuperscript{Ile}, and of hypoxanthine-guanine phosphoribosyltransferase HprT, which participates in the purine salvage pathway, are often organized into an operon (Fig. 7). In B. subtilis, these two proteins form a complex, which binds to the promoter and regulates the expression of the \textit{ftsH} gene, whose product is involved in membrane quality control (97). In many bacterial genomes, the \textit{hprT} start codon overlaps with the stop codon of \textit{tilS}. However, in Listeria \textit{monocytogenes}, the stop codon of \textit{tilS} is missing and translation of the in-frame \textit{tilS-hprT} ORFs results in the expression of the single hybrid protein TilS-HprT (Fig. 7). While the hybrid protein can perform the functions of both individual enzymes (97), proteomics analysis showed that HprT is additionally expressed from the internal AUG start codon of the fused \textit{tilS-hprT} gene (61). Interestingly, it was impossible to isolate mutants with inactivated iTIS (61), suggesting that the presence of individual (not TilS-fused) HprT may be required for stabilizing the TilS-HprT hybrid in \textit{L. monocytogenes} (97). Alternatively, translation from the iTIS may increase the stability of the hybrid mRNA.

Cooperation of the alternative products of the \textit{pgaM} gene is required for optimal synthesis of the antibiotic gaudimycin C

The biosynthetic gene cluster \textit{pgaEFLM} of \textit{Streptomyces} sp. PGA64 is responsible for production of the angucycline polyketide antibiotic gaudimycin C (98). Similar to the \textit{tilS-hprT} fusion (Fig. 7), the gene \textit{pgaM} of the cluster is the result of the fusion of two independent genes, where the upstream ORF lost its stop codon and

---

**FIGURE 6** The alternative LysC isoform expands the amino acid-mediated regulation of the enzyme. The full-size (α) and N-terminally truncated (β) isoforms of LysC associate into a functional α\textsubscript{2}β\textsubscript{2} tetramer, in which the binding sites for the regulatory amino acids lysine (orange) and threonine (red) are formed at the interface of the α-β subunits. Shown is the structure of \textit{Corynebacterium glutamicum} LysC (PDB ID: 3AAW) (95).

**FIGURE 7** Elimination of the stop codon creates a fused gene with two translation starts. Two genes, \textit{tilS} and \textit{hprT}, which encode functionally distinct proteins and associate to form a functional complex, are organized in an operon in \textit{B. subtilis}. Due to the elimination of the \textit{tilS} stop codon in \textit{L. monocytogenes}, the two ORFs are fused into a single gene and can be expressed from the primary and the internal TIS.
the downstream ORF retained its own TIS. The full-size protein, PgaM, possesses both oxygenase (in its N-terminal portion) and reductase (in its C-terminal segment) activities. In contrast, the shorter variant, PgaMred, whose translation starts at the iTIS, contains only the reductase portion (99) (Table 1). Expressed together, PgaM and PgaMred form a stable heterotetramer, conducive to the production of gaudimycin C (99). In contrast, if iTIS is mutated and expression of PgaMred is abolished, PgaM undergoes self-oligomerization, forming a less active complex. Unexpectedly, the iTIS mutation decreased the yield of PgaM by 90%. Conversely, inactivation of the pgaM pTIS abolished production not only of the full-size protein but also of PgaMred (99). This is a clear demonstration of interdependent expression of nested genes that could be potentially regulated either at the level of translation or due to altered mRNA stability.

Putative iTISs were also identified between the domains of fused angucycline oxygenase-reductase genes in other Streptomyces species (99). This observation, together with the example of the pikAIV gene in S. venezuelae discussed earlier, suggests that optimization of secondary metabolite biosynthesis by producing two proteins from one gene could be a widespread strategy.

The alternative IncC variant facilitates plasmid partitioning

We discussed earlier the isoforms of the TrfA protein encoded in the RK2 plasmid that help control plasmid copy number in different hosts. Another RK2 gene, incC (also known as parA), also directs synthesis of two polypeptides involved in plasmid partitioning, IncC1 and IncC2, via the pTIS and an iTIS, respectively (100) (Table 1). Both proteins are detected in bacterial cells carrying the plasmid, but their relative production is host specific (101). The N-terminal domain of IncC1, absent in IncC2, plays a role in the ADP-stimulated dimerization of the protein required for its activity. Interestingly, however, mixing IncC1 and IncC2 leads to protein polymerization as well as improved DNA binding. Thus, cooperation of both variants may be required for optimal plasmid partitioning (101).

The Function of the Alternative Protein Differs Significantly from That of the Primary Protein

FNR enzymes with opposite functions in cyanobacteria

Ferredoxin:NADP oxidoreductases (FNRs) facilitate CO₂ fixation during photosynthesis by reducing NADP⁺ to NADPH. However, nonphotosynthetic plastids contain a distinct version of this protein that is able to perform the opposite reaction and oxidize NADPH to provide electrons for nitrogen assimilation (102, 103). During nitrogen starvation, the cyanobacterium Synechocystis sp. strain PCC 6803 can switch from phototrophic to chemoheterotrophic metabolism. Remarkably, instead of carrying two distinct genes encoding the photosynthetic and the heterotrophic versions of FNR, Synechocystis sp. has a single gene, petH, which codes for two FNR variants capable of carrying the opposite reactions (Fig. 8). During photosynthetic growth, the major product of the petH gene is the full-length FNR₄, while under conditions of heterotrophic metabolism, the alternative AUG₁₁₃ codon is used to produce the shorter variant FNR₅ (104, 105) (Table 1). The relative expression of the FNR isoforms is controlled using alternative promoters. A transcript initiated at the upstream promoter has a long 5′ UTR, which folds into a stable structure that hinders access to the pTIS, downregulating synthesis of FNR₄ and favoring production of FNR₅ (Fig. 8). In the transcript with a shorter 5′ UTR, initiated at the promoter more proximal to the primary start codon of the gene, the pTIS becomes accessible and FNR₄ is preferentially synthesized. Similar mechanisms are used for controlled utilization of alternative initiation sites in the petH genes in a range of cyanobacteria (106). This elegant regulatory scheme illustrates how mRNA conformations enable ribosomes to choose between different TISs.

Alternative Translation Products with an Unknown Raison d’Être

Either of the two FliO variants can help assemble functional flagella

The transmembrane protein FliO is involved in assembly of the flagella of Gram-negative bacteria. Two FliO isoforms are produced from the alternative TISs proposed to operate in the fliO gene: the primary one that overlaps with the stop codon of the upstream fliN gene or an iTIS located 20 (in E. coli) or 21 (in S. Typhimurium) codons downstream from the pTIS (107, 108) (Table 1). The fliO null mutants could be complemented by either version of FliO, although the shorter version was less efficient (108). While the physiological role of the short FliO is unclear, it is interesting that the iTIS is located within the loop of an mRNA hairpin, whose presence, but not the sequence, is conserved between the fliO genes of E. coli and S. Typhimurium (108).
Two translation start sites could be operating in the middle wall protein gene of *Bacillus brevis*

The cell wall of some *B. brevis* strains is composed of three layers: outer wall protein, middle wall protein (MWP), and peptidoglycan (109). Two TISs seem to operate within the *mwp* gene of the *B. brevis* 47 strain, leading to production of the putative longer and shorter variants of the MWP polypeptide (109) (Table 1). A similar arrangement of two TISs is found in the cell wall protein genes of several other *B. brevis* strains (110). The N terminus of the short MWP contains a conventional signal sequence, while the N-terminal extension present in the longer MWP variant is rich in charged residues and does not resemble any signal sequence. Deletion of the pTIS does not preclude expression of a functional MWP (109), arguing that the shorter MWP variant produced from iTIS is actively used in the cell. Mutating the iTIS from AUG to AUC reduced, but did not preclude, the expression of the protein (111). Although this result was interpreted as an indication that the functional protein could be also translated from the upstream pTIS, this conclusion must be taken with caution because AUC can function as initiator codon in bacteria (8). To the best of our knowledge, the expression of two MWP variants has not been directly demonstrated.

Two SafA protein isoforms in the *B. subtilis* spore coat

Upon nutrition stress, *B. subtilis* differentiates into two cell types. One of these, the endospore, matures within the mother cell to sequester and protect the genome until the environmental conditions improve. SafA protein is incorporated in the endospore coat at the early stages of development. However, a shorter SafA-C30 isoform, translated from the AUG164 codon of the *safA* gene, can additionally be found in the coat of the mature spore (112) (Table 1). While mutating the iTIS does not affect spore coat formation, overexpression of SafA-C30 blocks sporulation (112), hinting that the internal initiation product may play a yet to be determined regulatory role in spore formation.

Protein isoforms might be involved in the export of hemolytic enzymes

Several pathogenic Gram-negative bacteria secrete hemolytic enzymes (113). For example, hemolysin A is...
produced by uropathogenic *E. coli* strains. The export machinery of hemolysin A is composed of several proteins, including the ABC-type transporter HlyB. However, the *hlyB* gene from the uropathogenic *E. coli* plasmid pGLO70 could direct expression of two HlyB variants, with the synthesis of the shorter polypeptide initiating at the internal AUG196 codon (114, 115) (Table 1). The function of the shorter HlyB isoform or its importance for secretion of hemolysin A is unknown. It is noteworthy that the mRNA secondary structure in the vicinity of the pTIS was proposed to attenuate translation of the full-length HlyB (116) and thus could be a part of the mechanism controlling the relative expression of the HlyB isoforms.

Another hemolytic enzyme, phospholipase C (PlcH), is secreted by *P. aeruginosa* and is important for the virulence of this organism (reviewed in reference 117). The operation of the PlcH secretion apparatus is not fully understood, but the product of the *plcH* gene is known to be important for this process because its deletion interferes with PlcH secretion (118). Two polypeptides are translated from the *plcH* gene, the full-length PlcR1 and the product of internal initiation at the AUG57 codon, PlcR2 (119) (Table 1). Like the differential localization of the CbMip (Fig. 2) and CvaA isoforms mentioned earlier, PlcR1 and PlcR2, when expressed in *E. coli*, are localized in the periplasm and the cytoplasm, respectively, because the signal sequence present in PlcR1 is lacking in PlcR2. On the basis of differential compartmentalization, it was proposed that PlcR2 might act as a PlcH chaperone, helping its folding or translocation through the inner membrane, while the proline-rich segment modulates the expression of PlcR2 by slowing down ribosomes immediately upstream of the iTIS.

### INTERNAL INITIATION OF TRANSLATION THAT DIRECTS SYNTHESIS OF A PROTEIN FROM AN ALTERNATIVE READING FRAME

When an iTIS directs translation from an OOF start codon, the sequence of the translated protein will be entirely different from the sequence of the polypeptide encoded in the main frame (Fig. 1). The stop codon of the frame corresponding to the OOF iTIS is often located at a relatively short distance from it, and as a result, the product of internal initiation would be a fairly short protein. Because identification and characterization of small proteins is experimentally difficult and bioinformatically challenging, to the best of our knowledge, only three examples of this type of genetic coding in bacteria are known.

#### The Competence Protein ComS Is Encoded within the srfA Gene of *B. subtilis*

The ability to take up exogenous DNA allows *B. subtilis* to survive unfavorable conditions. The timing of establishing the competent state is controlled by quorum sensing and is regulated by several genes (reviewed in reference 121). Genetic mapping suggested that one of the determinants of competence, *comS*, is located within the *srfA* operon, which contains genes responsible for production of the antibiotic surfactin (122). Although secretion of surfactin is coordinated with the establishment of competence, none of the activities associated with surfactin production are important for competence per se. Instead, an OOF ORF encoding the 46-amino-acid-long ComS protein was identified within the second gene of the *srfA* operon (122, 123). Translation of the nested *comS* gene is initiated at an OOF UUG codon preceded by a strong SD sequence (Table 1) (note that UUG is a common start codon in *B. subtilis* genes [11]). Embedding the *comS* gene within one of the surfactin operon genes is an elegant strategy to coregulate two distinct but related activities contributing to DNA uptake: starvation induces expression and secretion of surfactin, which causes lysis of the cells of other competing bacterial species, whereas ComS facilitates uptake and utilization of the released genetic material.

#### The Gene of Ribosomal Protein L34 Is Nested within the RNase P Protein Gene of *Thermus thermophilus*

RNase P directs maturation of the tRNA 5′ end. In bacteria, this enzyme is composed of a catalytic RNA and a small accessory protein encoded by the *rpmA* gene (124). In many bacterial genomes, the *rpmA* gene is positioned immediately downstream of the *rpmH* gene, encoding ribosomal protein L34, and both genes are usually cotranscribed within one operon (125). In *T. thermophilus*, however, no *rpmH* ORF was found upstream of the unusually long *rpmA* gene. Instead, the entire 49-codon-long *rpmH* is completely embedded within the *rpmA* ORF, but in an alternative reading frame (126) (Fig. 9). The start codons of *rpmH* and *rpmA* are separated by only one nucleotide, and expression of both
proteins is likely supported by the same SD sequence (126) (Table 1).

Because bacterial cells contain a large (~100-fold) excess of ribosomes over RNase P, rpmH is likely translated much more efficiently than rnpA. One of the factors contributing to the difference in expression of the two overlapping ORFs is the distance of their respective start codons from SD: the AUG start codon of rpmH is located at the optimal distance of 7 nt from the GGAGG SD sequence, whereas the rnpA AUG start codon is only 3 nt away from it (126) (Fig. 9). Short spacing between the SD sequence and the start codon likely decreases the efficiency of initiation of rnpA translation (127). Furthermore, the highly efficient initiation of rpmH translation might compete with initiation at the rnpA start codon. Similar to other ribosomal protein genes, the codon usage in rpmH has been evolutionarily optimized for highly active translation. In contrast, the codon usage of rnpA is less favorable. The presence of rare codons in the vicinity of the rnpA TIS could lead to abortive translation, which would additionally decrease the expression level of RNase P protein relative to that of L34.

Occlusion of rpmH within the extended version of the rnpA gene likely occurred due to a mutation leading to appearance of a start codon upstream of rpmH (in frame with the downstream rnpA, but OOF relative to rpmH) and subsequent elimination of the rnpA authentic start. A similar overlapping arrangement of rnpA and rpmH genes has been found in other members of the genus Thermus (126). It is worth noting that occlusion of an OOF gene within another gene should be more frequent in the high-GC-content organisms, including the Thermus species, due to reduced frequency of occurrence of OOF stop codons.

A Heat Shock Protein Is Encoded in an Alternative Reading Frame within the gnd Gene of E. coli

To survive abrupt increase in temperature, bacteria express heat shock proteins. However, not all heat shock proteins have been fully characterized, especially low-abundance, small, membrane-associated polypeptides. A recently developed approach for quantitative membrane proteomics identified a tryptic peptide apparently belonging to a previously unknown small heat shock protein named GndA (128). The identified tryptic peptide mapped to a short ORF located entirely within the gnd gene encoding 6-phosphogluconate dehydrogenase, but in the -1 frame (128). At what codon the translation of the small gndA ORF is initiated is unknown because two in-frame ATG codons are found upstream of the sequence encoding the mapped peptide (Table 1). It was previously suggested that sequestration of the pTIS of the main gnd ORF in the mRNA secondary structure could modulate the expression of 6-phosphogluconate dehydrogenase in response to changes in the cell growth rate (129). The proposed mRNA folding may also involve the gndA ORF start site, and it is possible that regulation of two genes is interconnected.

NEW TOOLS FOR IDENTIFYING ALTERNATIVE TRANSLATION START SITES

Most of the examples of internal initiation discussed in the previous sections have been discovered serendipitously, mainly while analyzing the purified protein or characterizing its functions. Although many bacterial genes could potentially contain functional iTISs, their experimental identification is a challenging problem.

Powerful proteomics techniques potentially could be employed for detecting the N termini of the proteins and thus mapping iTISs (130). Unfortunately, neither conventional top-down proteomics nor the traditional bottom-up techniques can reliably distinguish the products of internal initiation from the protein fragments generated by proteolysis. Newly emerging proteomics methodologies are better suited for revealing iTISs.
The most promising of them relies on the use of peptide deformylase inhibitors, such as actinomycin. Formylmethionine at the N termini of peptides helps to distinguish translation initiation products from those of proteolytic degradation and to determine the true pTISs or iTISs (131). Nineteen iTISs have been recently mapped in the genome of L. monocytogenes using this technique (61). This approach, however, has its intrinsic limitations because neither very short nor very long peptides can be easily detected by mass spectrometry and, in addition, low-abundance proteins (a common scenario for many of the internally initiated polypeptides) evade detection. A recently developed methodology combines bottom-up mass spectrometry with analysis of hypothetical polypeptides encoded in all six frames in the genome of interest (128). Applying this approach to membrane proteins helped to identify the novel OOF internal ORF gndA within the gnd gene of E. coli (128) that we described earlier. The downside of the approach is that it does not immediately determine the codon used for initiation of translation.

A more systematic and sensitive strategy is to use the gene-oriented ribosome-profiling approach (30, 132). This technique involves isolation and deep sequencing of ribosome-protected mRNA fragments (“ribosome footprints”). Mapped to the genome, ribosome footprints reflect the distribution of ribosomes along the translated mRNAs (Fig. 10a). If the iTIS is sufficiently strong, it should increase the ribosome traffic through the downstream portion of the ORF and lead to higher occupancy of the distal mRNA segment (133). The changes in the ribosome density could be, however, too subtle when the iTIS promotes translation with moderate or, moreover, low efficiencies. For example, ribosome profiling does not immediately reveal internal initiation within the E. coli genes infB or mrcB that we discussed earlier (30) (Fig. 10b).

Fortunately, using ribosome-targeting antibiotics that specifically inhibit translation initiation can increase the sensitivity of ribosome profiling for detecting TISs. Such an approach has been successfully used for mapping TISs in eukaryotic genomes and identified thousands of previously unknown in-frame and OOF iTISs as well as upstream TISs that would result in N-terminal extensions of known ORFs (134–136). The paucity of antibiotics specifically targeting the initiation step of bacterial translation delayed the application of this approach for bacteria. Nevertheless, it has been shown that tetracycline could be potentially suitable for mapping TISs in bacterial genes (137). This finding was unexpected because tetracycline has been known as an inhibitor of translation elongation and yet, in profiling experiments, the majority of the ribosome density peaks in the tetracycline-treated cells were observed at the start codons of the genes (137). Tetracycline-assisted ribosome profiling helped correct the assignment of some of the previously misannotated TISs of E. coli genes. Furthermore, several of the observed peaks of ribosome density also suggested that some E. coli genes have more than one start codon, with the additional sites located commonly in the vicinity of the pTIS (137). Nevertheless, the uncertainty of why tetracycline preferentially inhibits translation initiation leaves open the possibility that some of the peaks of ribosome density observed at the internal codons of the gene could result from inhibition of elongation rather than initiation of translation.

A more promising approach relies on the use of an authentic translation initiation inhibitor, retapamulin, in ribosome-profiling experiments. Retapamulin belongs to the pleuromutilin class of protein synthesis inhibitors. It targets the peptidyltransferase center, and its binding is incompatible with the presence of a nascent peptide within the ribosome (138). Therefore, retapamulin is capable of acting exclusively during translation initiation and efficiently arrests the ribosome at the start codons by blocking formation of the first peptide bond, probably by altering the placement of the CCA end of initiator tRNA bound at the ribosomal P-site (139). Indeed, we have found that retapamulin can be used as an excellent tool for revealing both pTISs and iTISs in bacteria (unpublished data). In retapamulin-treated cells, well-defined ribosome density peaks could be readily seen at previously known iTISs in the E. coli genome (Fig. 10). In addition, a number of other iTIS candidates have been identified and are being currently analyzed. Additional tools, for example, the use of specific inhibitors of translation termination (140) in conjunction with initiation inhibitors, might facilitate the detection of OOF internal ORFs. In addition, advanced RNA-sequencing methods, such as differential RNA sequencing (141), can be used to identify alternative transcript isoforms that can also result in expression of internal ORFs. In the future, combining antibiotic-enhanced ribosome profiling with transcriptome sequencing and actionin-in-basedproteomics could form an even more robust framework for revealing the alternative proteome originated from unusual translation events.

**CONCLUDING REMARKS**

In our review, we attempted to present a snapshot of the currently known examples of internal translation...
FIGURE 10  Retapamulin-assisted ribosome profiling illuminates sites of internal translation initiation. (a) (Left) Ribosome profiling in untreated cells shows the distribution of translating ribosomes along the mRNAs [30]. (Right) Brief pretreatment of cells with the translation initiation inhibitor retapamulin arrests the ribosomes at the translation start site of the ORFs. (b) Examples of primary and internal TISs revealed by retapamulin-assisted ribosome profiling in E. coli genes known to contain functional internal start codons. Note that the presence of the iTISs is difficult to detect when ribosome profiling is performed in untreated cells.
initiation in bacteria. We have intentionally left out bacteriophages, which present many more examples of genes-within-genes encoding as well as unconfirmed suggestions that internal initiation occurs within some archaeal genes, and instead focused our attention primarily on intrinsic bacterial genes.

In this review, we decided to keep the names given to the internal initiation products by the authors who originally discovered them. For some of the in-frame cases, the molecular weights (in kDa) of the proteins were indicated next to their names (for example, ClpB93 and ClpB79). We find this to be the most straightforward designation for identification of the different protein isoforms and suggest using this unifying nomenclature to annotate the products of in-frame alternative initiation.

Most of the few currently known examples of translation initiation at internal start codons of genes have been identified serendipitously. Many questions about the pervasiveness of this mode of gene expression, functions of the alternative protein products, or the mechanisms of regulation remain unanswered. We do not know how the ribosome traffic influences the utilization of the iTIS or what the influence is of the local and global mRNA structure on the relative use of the primary and secondary start codons. We also poorly understand the advantage of expressing a second protein from an internal codon of a gene instead of maintaining its independent gene in a bacterial genome.

Emerging tools will likely rapidly expand our knowledge about the scenarios of unconventional protein coding in a more systematic and time- and cost-effective way. Studies of the alternative proteome cryptically encoded in the genome will most likely illuminate new layers of gene regulation and inform us of the possible functions of previously unknown proteins. This new knowledge may contribute to developing new medicines that could specifically target the alternative proteome or the mechanisms of unconventional decoding of genetic information.

ACKNOWLEDGMENTS

We thank Dorota Klepacki and James Marks for help in performing and interpreting the ribosome-profiling experiments mentioned in this review and Elizabeth Woods for proofreading the manuscript. The work on translation regulation in the Mankin/Vázquez-Laslop laboratory is supported by grant MCB 161583 from the National Science Foundation.

REFERENCES


