A Search for Genes Encoding Histidine-Containing Leader Peptides in Actinobacteria

Semen A. Korolev, Sergei A. Lyzhin, Oleg A. Zverkov, Alexandr V. Seliverstov, and Vassily A. Lyubetsky

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
{korolev,zverkov,slvstv,lyubetsk}@iitp.ru, serodger@yandex.ru

Abstract. A large-scale search for leader peptides was conducted in Actinobacteria made it possible to predict a mechanism of regulation of translation initiation. This mechanism relies on the interaction between the ribosome translating the leader peptide and the RNA helix potentially overlapping the ribosome-binding site.

Keywords: Actinobacteria · gene expression regulation · translation · leader peptide.

1 Introduction

The phylum Actinobacteria includes agents of socially important diseases (tuberculosis, paratuberculosis, leprosy, diphtheria, etc.), plant pathogens, producers of antibiotics, components of the normal human intestinal microflora, and free-living species suitable for sewage treatment including radiation-resistant ones.

Polycistronic mRNAs encoding several proteins in a row are typical for bacteria. Hereafter, we refer to the *first* protein in a row. Normally, the ribosome binds the Shine-Dalgarno sequence in the 5'-untranslated region of RNA not far from the start codon of the gene to initiate translation. This polypurine sequence with the consensus GGAGGA is complementary to a sequence at the 3' end of 16S rRNA. There are well-documented examples of regulation of translation initiation based on masking the Shine-Dalgarno sequence by the RNA secondary structure. These include riboswitches, whose structure depend on ligand binding to the RNA [1,2], T-box regulation of the *ileS* gene encoding isoleucyl-tRNA synthetase, and leucine regulation of the *leuA* gene encoding α-isopropylmalate synthase in many Actinobacteria [3,4]. Interestingly, leucine regulation involves a leader peptide whose rate of translation depends on leucine concentration. This rate is the key factor in the classical attenuation regulation of gene expression commonly dependent on the concentration of tryptophan or histidine. Similar regulatory mechanisms relying on the concentrations of phenylalanine, branched-chain amino acids, and threonine are also known [4]. The latter publication also proposed a number of radically new variants of this regulation. We have studied a

similar regulation relying on the concentration of cysteine in Actinobacteria [5]. The ribosome masks 30-40 nucleotides of mRNA, 10-12 of which are downstream of the current codon. Here, special attention was given to proteins that can unwind RNA and contain with domains in the Pfam database that belong to the DEAD (PF00270) and Helicase_C (PF00271) families [6,7,8]. They are involved in RNA metabolism including its transcription, translation, and degradation. In particular, proteins with poorly explored domains such as Pfam-B_340 were considered in detail. Special focus was placed on the agents of tuberculosis [9] and diphtheria [10].

2 Materials and Methods

Bacterial genomes were retrieved from GenBank. Leader peptides were identified using the original algorithm described elsewhere [5]. It selects open reading frames (presumably encoding leader peptides) in the 5'-leader regions of structural genes with the local density of histidine codons in these open reading frames exceeding a threshold (set equal to 1.4 in this work). Gene annotations were verified using the Pfam database [11]. The results were visualized using WebLogo [12]. RNA secondary structures were predicted using RNAstructure [13].

3 Results

3.1 Pattern of Histidine-Containing Leader Peptides

Our program for the search of leader peptides produced the following results for Actinobacteria and histidine as the regulatory amino acid. The distance between the start codon of the first structural gene and the stop codon of the presumable leader peptide in the range from 1 to 50 nucleotides has a very pronounced peak for the distance of 10-11-12 nucleotides; the less pronounced peak is observed at 5 (Fig. 1). No pronounced peaks have been found within this range in other groups of bacteria including Cyanobacteria and Proteobacteria.

3.2 Analysis of Domain Structure and Nucleotide Composition of Typical Leader Peptides

The domain structure was analyzed in actinobacterial proteins encoded by structural genes with upstream histidine-rich leader genes at a distance of 6-18 nucleotides. Such structural genes often code for transcription factors of the LysR (PF00126) and TetR (PF13972) families, cytochrome P450 (PF00067), subunits of ABC transporters (PF00005), proteins with Helicase_C (PF00271), DEAD (PF00270), and Phage_integrase (PF00589) domains or with conserved domains of unknown function: UPF0182, Pfam-B_340, Pfam-B_671, and Pfam-B_11008.

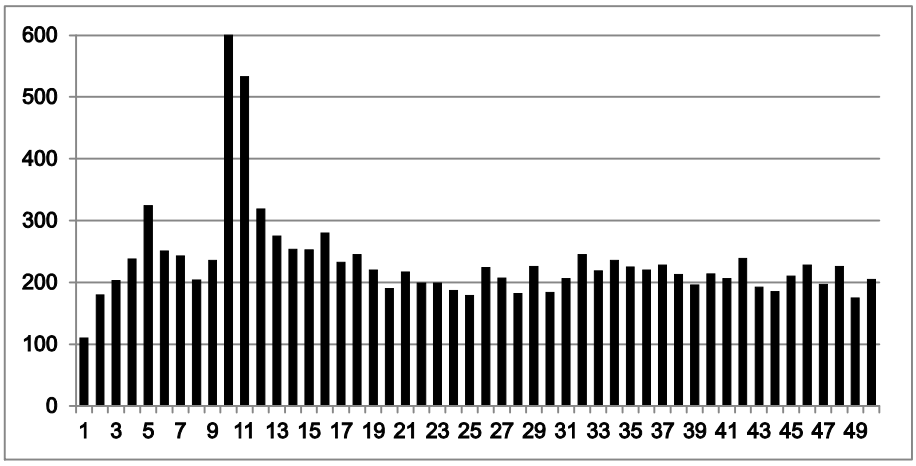

Fig. 1. Number of actinobacterial genes as a function of the distance between the stop codon of a putative leader peptide with regulatory histidine codons and the start codon of a structural gene

Nucleotide sequence analysis of the 5'-untranslated region and the adjacent coding regions demonstrated that the region usually occupied by the Shine-Dalgarno sequence often contains pyrimidines (U or C) instead of the typical purines (A or G). Thus, direct translation initiation is impossible and ribosome reinitiation is required after the translation of the leader peptide. Long degenerate palindromic repeats were found near the stop codon of the leader gene upstream of the structural genes encoding proteins with Helicase_C and DEAD domains, which can give rise to a RNA helix. It was found in distant actinobacterial species *Bifidobacterium animalis*, *Corynebacterium diphtheriae*, *Corynebacterium glutamicum*, and *Streptomyces griseus*.

For example, mRNAs encoding proteins with Helicase_C and DEAD domains *Corynebacterium diphtheriae* contain the palindromic sequence GCCUUgAAGGC overlapping the stop codon of the leader gene by one nucleotide and the whole region upstream of the start codon of the structural gene (positions -11 to -1 relative to the start codon). This RNA region forms a single duplex with the free energy of -4.6 kcal/mol shown in Fig. 2. The corresponding proteins are listed in Table 1.

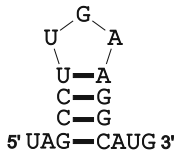

Fig. 2. RNA duplex in the region from the stop codon of the leader gene to the start codon of the structural gene encoding helicase in *Corynebacterium diphtheriae*

Table 1. Helicases with Helicase_C and DEAD domains in *Corynebacterium diphtheria* whose genes are preceded by the palindromic sequence GCCUUgAAGGC and the leader gene

Species and strain	Locus	Protein
Corynebacterium diphtheriae 241	NC_016782	YP_005125436.1
Corynebacterium diphtheriae 31A	NC_016799	YP_005157952.1
Corynebacterium diphtheriae BH8	NC_016800	YP_005160307.1
Corynebacterium diphtheriae CDCE 8392	NC_016785	YP_005133681.1
Corynebacterium diphtheriae HC01	NC_016786	YP_005135967.1
Corynebacterium diphtheriae HC02	NC_016802	YP_005164930.1
Corynebacterium diphtheriae HC03	NC_016787	YP_005138191.1
Corynebacterium diphtheriae HC04	NC_016788	YP_005140463.1
Corynebacterium diphtheriae INCA 402	NC_016783	YP_005127653.1
Corynebacterium diphtheriae NCTC 13129	NC_002935	NP_939594.1
Corynebacterium diphtheriae PW8	NC_016789	YP_005142778.1
Corynebacterium diphtheriae VA01	NC_016790	YP_005145018.1

The RNAs encoding proteins with the same domains in *Corynebacterium glutamicum* have the Shine-Dalgarno sequence overlapping the hairpin CCgACUAgaguUAGUgGG with the free energy of -6.3 kcal/mol. The duplex is shown in Fig. 3 and the list of corresponding proteins is shown in Table 2. In this case, the distance between the stop codon of the leader gene and the start codon of the helicase gene is 12 nt.

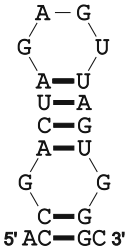

Fig. 3. RNA hairpin overlapping the Shine-Dalgarno sequence in the helicase in *Corynebacterium glutamicum*

Table 2. Helicases with Helicase_C and DEAD domains in *Corynebacterium glutamicum*. The Shine-Dalgarno sequence overlaps the RNA hairpin with the free energy of -6.3 kcal/mol.

Species and strain	Locus	Protein
Corynebacterium glutamicum ATCC 13032	NC_003450	NP_600667.1
Corynebacterium glutamicum ATCC 13032	NC_006958	YP_225735.1
Corynebacterium glutamicum R	NC_009342	YP_001138404.1

The RNAs encoding proteins with Helicase_C and DEAD domains in *Bifidobacterium animalis* and *Streptomyces griseus* have the start codon for the helicase overlapping the RNA helix containing a G-U pair with the free energy of -1.4 kcal/mol. The hairpin is shown in Fig. 4 and the list of corresponding proteins is shown in Table 3.

The hairpin has a long loop as well as a long distance of 17 nt between the stop codon of the leader gene and the start codon of the helicase.

The palindromic sequence GUAGaCUGC with a U-G pair was found in mycobacteria related to *Mycobacterium tuberculosis* at positions from -10 to -1 relative to the start codon of the structural genes encoding proteins with the Pfam-B_340 domain. It is possible that this RNA region forms a stable helix within a more complex secondary structure. The list of proteins containing the Pfam-B_340 domain and associated with leader peptides containing histidines is shown in Table 4.

4 Discussion

We have predicted new cases of regulation of translation initiation in Actinobacteria for genes with an upstream gene encoding leader peptide with histidines. In some cases, a helix is formed in the RNA near the stop codon of the leader peptide to prevent initiation of structural protein translation. The formation of a more complex RNA structure overlapping the ribosome-binding site can be anticipated in other cases.

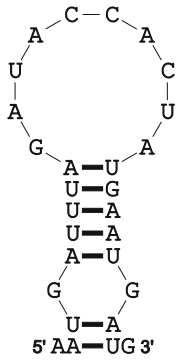

Fig. 4. RNA hairpin overlapping two nucleotides of the helicase start codon in *Bifidobacterium animalis* and *Streptomyces griseus*

Table 3. Helicases with Helicase_C and DEAD domains in *Bifidobacterium animalis* and *Streptomyces griseus*. The helicase start codons overlap the RNA hairpin AUgAUUUAgauaccacuaUGAAUgAU with the free energy of -1.4 kcal/mol.

Species and strain	**Locus**	**Protein**
Bifidobacterium animalis subsp. lactis B420	NC_017866	YP_006300983.1
Bifidobacterium animalis subsp. lactis Bi-07	NC_017867	YP_006302567.1
Bifidobacterium animalis subsp. lactis Bl-04	NC_012814	YP_002968467.1
Bifidobacterium animalis subsp. lactis BLC1	NC_017216	YP_005579811.1
Bifidobacterium animalis subsp. lactis DSM 10140	NC_012815	YP_002970034.1
Bifidobacterium animalis subsp. lactis V9	NC_017217	YP_005581376.1
Streptomyces griseus subsp. griseus NBRC 13350	NC_010572	YP_001825934.1

Table 4. Proteins containing Pfam-B_340 and associated with the leader peptides

Species and strain	Locus	Protein
Mycobacterium tuberculosis H37Rv	NC_000962	NP_218168.1
Mycobacterium tuberculosis CDC1551	NC_002755	NP_338300.1
Mycobacterium bovis AF2122/97	NC_002945	NP_857314.1
Mycobacterium bovis BCG str. Pasteur 1173P2	NC_008769	YP_979788.1
Mycobacterium tuberculosis H37Ra	NC_009525	YP_001285037.1
Mycobacterium tuberculosis F11	NC_009565	YP_001289607.1
Mycobacterium marinum M	NC_010612	YP_001853401.1
Mycobacterium bovis BCG str. Tokyo 172	NC_012207	YP_002646750.1
Mycobacterium tuberculosis KZN 1435	NC_012943	YP_003033692.1
Mycobacterium africanum GM041182	NC_015758	YP_004725285.1
Mycobacterium canettii CIPT 140010059	NC_015848	YP_004747076.1
Mycobacterium tuberculosis KZN 4207	NC_016768	YP_005102183.1
Mycobacterium bovis BCG str. Mexico	NC_016804	YP_005173158.1
Mycobacterium tuberculosis UT205	NC_016934	YP_005309871.1
Mycobacterium tuberculosis RGTB327	NC_017026	YP_005362185.1
Mycobacterium tuberculosis CCDC5180	NC_017522	YP_005911166.1
Mycobacterium tuberculosis CCDC5079	NC_017523	YP_005914806.1
Mycobacterium tuberculosis CTRI-2	NC_017524	YP_005918738.1
Mycobacterium tuberculosis KZN 605	NC_018078	YP_006475172.1
Mycobacterium tuberculosis H37Rv	NC_018143	YP_006517138.1

We propose the following regulation mechanism. For brevity, we assume that the rate of leader peptide translation depends on histidine concentration. If it is deficient, the ribosome translating the leader peptide does not reach the stop codon and an RNA hairpin is formed to prevent initiation of structural gene translation. If histidine is excessive, the ribosome rapidly translates the leader peptide and unwinds or prevents formation the RNA helix. After reaching the stop codon of the leader peptide, the ribosome overlaps the start codon of the structural gene, which likely favors the reinitiation.

The local density of histidine is higher than that of tryptophan in leader peptides involved in the classical attenuator and similar regulations [4]. Sometimes, the regulation relies on the concentration of several rather than one amino acids or aminoacyl-tRNAs. The regulatory mechanism proposed here is simple compared to riboswitches or leucine regulation since no complex RNA structures are involved. On the other hand, such regulation is hard to reveal from an individual sequence; it requires statistical analysis of the 5'-leader regions of orthologous genes in many species, which was hardly possible until the recent expansion of GenBank.

The regulation of expression of genes encoding helicases can trigger a complex cascade of regulatory events associated with specific RNA degradation in conditions of excessive amino acids. On the other hand, the proposed regulation of expression of helicase genes involves RNA secondary structures, which suggests a negative feedback effect of the concentration of cytoplasmic helicases on their expression. In this

case, the leader peptide can include any amino acids with relatively low normal concentration.

Since the genes potentially regulated by this mechanism include transcription factors of the LysR [14] and TetR [15] families, a regulatory cascade repressing transcription of certain genes in response to excessive amino acids with relatively low normal concentration can be proposed in Actinobacteria.

The identification of the putative regulation only in Actinobacteria suggests that it emerged after the separation of this bacteria, the regulation of translation initiation applies to many actinobacterial genes [3].

Acknowledgements. This work was supported by the Russian Scientific Fund (project no. 14–50–00150).

References

1. Mandal, M., Breaker, R.R.: Gene regulation by riboswitches. Nat. Rev. Mol. Cell. Biol. 5, 451–463 (2004)
2. Suna, E.I., Rodionov, D.A.: Computational analysis of riboswitch-based regulation. Biochimica et Biophysica Acta. 1839(10), 900–907 (2014)
3. Seliverstov, A.V., Putzer, H., Gelfand, M.S., Lyubetsky, V.A.: Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. BMC Microbiology. 5(54), 14 p. (2005)
4. Lopatovskaya, K.V., Seliverstov, A.V., Lyubetsky, V.A.: Attenuation Regulation of the Amino Acid and Aminoacyl-tRNA Biosynthesis Operons in Bacteria: A Comparative Genomic Analysis. Molecular Biology. 44(1), 128–139 (2010)
5. Lyubetsky, V.A., Korolev, S.A., Seliverstov, A.V., Zverkov, O.A., Rubanov, L.I.: Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism. Computational Biology and Chemistry. 49, 7–13 (2014)
6. Johnson, E.R., McKay, D.B.: Crystallographic structure of the amino terminal domain of yeast initiation factor 4A, a representative DEAD-box RNA helicase. RNA. 5(12), 1526–1534 (1999)
7. de la Cruz, J., Kressler, D., Linder, P.: Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. Trends Biochem. Sci. 24(5), 192–198 (1999)
8. Aubourg, S., Kreis, M., Lecharny, A.: The DEAD box RNA helicase family in Arabidopsis thaliana. Nucleic Acids Res. 27(2), 628–636 (1999)
9. Camus, J.C., Pryor, M.J., Medigue, C., Cole, S.T.: Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology (Reading, Engl.). 148(10), 2967–2973 (2002)
10. Cerdeno-Tarraga, A.M., Efstratiou, A., Dover, L.G., Holden, M.T., Pallen, M., Bentley, S.D., Besra, G.S., Churcher, C., James, K.D., De Zoysa, *et al.*: The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res. 31(22), 6516–6523 (2003)
11. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.*: Pfam: The protein families database. Nucleic Acids Res. 42, D222–D230 (2014)
12. Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E.: WebLogo: A sequence logo generator. Genome Research. 14, 1188–1190 (2004)

13. Reuter, J.S., Mathews, D.H.: RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics, 11, 129 (2010)
14. Maddocks, S.E., Oyston, P.C.: Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins, Microbiology, 154(12), 3609–3623 (2008)
15. Ramos, J.L., Martinez-Bueno, M., Molina-Henares, A.J., Teran, W., Watanabe, K., Zhang, X., Gallegos, M.T., Brennan, R., Tobes, R.: The TetR family of transcriptional repressors. Microbiol Mol Biol Rev. 69, 326–356 (2005)