BMC Microbiology

BioMed Central

Research article

Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria

Alexander V Seliverstov[1], Harald Putzer[2], Mikhail S Gelfand*[1] and Vassily A Lyubetsky[1]

Address: [1]Institute for Information Transmission Problems, RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia and [2]Institut de Biologie Physico-Chimique, CNRS UPR9073, 13, rue P. et M. Curie, 75005 Paris, France

Email: Alexander V Seliverstov - slvstv@iitp.ru; Harald Putzer - Harald.Putzer@ibpc.fr; Mikhail S Gelfand* - gelfand@iitp.ru; Vassily A Lyubetsky - lyubetsk@iitp.ru

* Corresponding author

Published: 03 October 2005

BMC Microbiology 2005, **5**:54 doi:10.1186/1471-2180-5-54

Received: 01 April 2005
Accepted: 03 October 2005

This article is available from: http://www.biomedcentral.com/1471-2180/5/54

Abstract

Background: Formation of alternative structures in mRNA in response to external stimuli, either direct or mediated by proteins or other RNAs, is a major mechanism of regulation of gene expression in bacteria. This mechanism has been studied in detail using experimental and computational approaches in proteobacteria and Firmicutes, but not in other groups of bacteria.

Results: Comparative analysis of amino acid biosynthesis operons in Actinobacteria resulted in identification of conserved regions upstream of several operons. Classical attenuators were predicted upstream of *trp* operons in *Corynebacterium* spp. and *Streptomyces* spp., and *trpS* and *leuS* genes in some *Streptomyces* spp. Candidate leader peptides with terminators were observed upstream of *ilvB* genes in *Corynebacterium* spp., *Mycobacterium* spp. and *Streptomyces* spp. Candidate leader peptides without obvious terminators were found upstream of *cys* operons in *Mycobacterium* spp. and several other species. A conserved pseudoknot (named LEU element) was identified upstream of *leuA* operons in most Actinobacteria. Finally, T-boxes likely involved in the regulation of translation initiation were observed upstream of *ileS* genes from several Actinobacteria.

Conclusion: The metabolism of tryptophan, cysteine and leucine in Actinobacteria seems to be regulated on the RNA level. In some cases the mechanism is classical attenuation, but in many cases some components of attenuators are missing. The most interesting case seems to be the *leuA* operon preceded by the LEU element that may fold into a conserved pseudoknot or an alternative structure. A LEU element has been observed in a transposase gene from *Bifidobacterium longum*, but it is not conserved in genes encoding closely related transposases despite a very high level of protein similarity. One possibility is that the regulatory region of the *leuA* has been co-opted from some element involved in transposition. Analysis of phylogenetic patterns allowed for identification of ML1624 of *M. leprae* and its orthologs as the candidate regulatory proteins that may bind to the LEU element. T-boxes upstream of the *ileS* genes are unusual, as their regulatory mechanism seems to be inhibition of translation initiation via a hairpin sequestering the Shine-Dalgarno box.

Background

Formation of alternative structures in 5'-leader regions of mRNAs is emerging as a major mechanism of gene regulation. There exist several possible variants of this mechanism whose common feature is the competition between two structures, one of which represses gene expression via premature termination of transcription or inhibition of translation initiation (reviewed in [1-6]). The energetically or kinetically more favourable structure forms by default, whereas the other one is stabilized by binding of a regulatory protein, tRNA, or a small cofactor, or is formed co-transcriptionally, as in classical attenuators.

RNA regulatory elements have been studied mainly in gamma-proteobacteria (*Escherichia coli*) and firmicutes (*Bacillus subtilis*). Computational analysis also has been mainly restricted to proteobacteria [7,8] and firmicutes [9-12]. Recently a new class of regulatory elements, riboswitches, has been described. These elements are highly conserved and were found in all major taxa of bacteria, as well as in some eukaryotes and archaea [13,14]. Comparative genomic analysis has played a major role in the discovery and analysis of T-boxes [9,15] and most riboswitches (reviewed in [4,5]). Several groups performed large-scale search for new RNA regulatory structures [16,17]. Analysis of RNA-based regulation often leads to non-trivial functional assignments for hypothetical genes and filling gaps in metabolic reconstruction (e.g. [11,14,18,19]).

Here we performed comparative analysis of candidate RNA regulatory elements in genomes of Actinobacteria. There are few known attenuators in these genomes. Those that have been experimentally studied are attenuators of the *trp* operons in *Corynebacterium glutamicum* [20] and *Streptomyces venezuelae* [21]. Studies of attenuator-like structures upstream of the *ilvB* and *leuA* genes of *Streptomyces coelicolor* produced somewhat ambivalent results. Indeed, although candidate leader peptides and alternative RNA structures were found upstream of the *ilvB* and *leuA* genes, reminiscent of the classical attenuators, the mutation analysis demonstrated that the regulatory mechanism is not attenuation in the strict sense: mutations in candidate regulatory codons in the leader peptide of the *ilvB* gene had no effect on regulation, and, although mutations in the leader peptide of *leuA* had some effect, it was not consistent with classical attenuation [22]. Computational analysis identified several types of riboswitches: THI-elements [14], RFN-elements [18], B12-elements [19], all of them regulating genes of cofactor metabolism by sequestering the Shine-Dalgarno box and start codon, and interfering with initiation of translation.

Results and discussion

Following an approach described previously [8], we systematically analysed the upstream regions of amino acid biosynthesis and aminoacyl-tRNA synthetase operons. Candidate regulatory structures were found upstream of genes involved in tryptophan, cysteine, and leucine metabolism. Candidate T-boxes were observed upstream of isoleucyl-tRNA synthetase genes. No conserved structures were observed upstream of genes from other amino acid biosynthesis pathways.

Tryptophan

The *trp* operons are preceded by classical candidate attenuators in all considered genomes of *Corynebacterium* spp. and *Streptomyces* spp. (Fig. 1). The leader peptides have double or triple repeats of regulatory UGG codons. All terminators are GC-rich and followed by poly-U-tracts. The antiterminator and terminator hairpins in all genomes contain complementary triples gGCC-rGCy-GGCC where absolutely conserved positions are set in capitals. This is analogous to the situation in proteobacteria, where the patterns involved in multiple interactions within attenuators are conserved at large evolutionary distances [8]. In *C. diphteriae*, candidate attenuators were found upstream of both biosynthetic operons $trpB_1EDGC$ and $trpB_2A$. A candidate attenuator was found upstream of the tryptophanyl-tRNA synthetase gene $trpS_2$ in *S. avermitilis*.

Cysteine

The upstream regions of the *cys* operon in *Mycobacterium* spp. and *Propionibacterium acnes* and the *cbs* gene of *Bifidobacterium longum* contain short open reading frames encoding candidate leader peptides with runs of cysteine codons near the stop codon (Fig. 2a). The upstream regions of *Mycobacterium* spp. are very similar and can be aligned (Fig. 2b). However, they do not contain any conserved hairpins that could serve as terminators of transcription. One possibility is that this region contains rho-dependent terminators similar to the situation in the tryptophanase operon *tna* of *E. coli* [23]. Indeed, *Mycobacteium* spp. have few rho-independent terminators [24,25]. On the other hand, all *Mycobacterium* genomes contain the components of the rho-dependent termination mechanism, *rho*, *nusG*, *nusA*, *nusB*. The region between the candidate leader peptide ORFs and the first genes in the *cys* operons contain polyY motifs that could serve as Rho-binding sites [26-28]. However, these motifs are not conserved, and thus this prediction is rather weak.

The cysteine operons in *M. avium* and *M. leprae* contain additional hypothetical genes, *MAP2122* and *ML0840* respectively, that are 62% identitical but have no other reliable homologs.

a)

Bacterium	Locus	Gene	Gene coordinates	Protein
C. diphtheriae	NC_002935	*trpB1*	2456701..2458032	NP_940652
		trpB2	2465139..2466365	NP_940660
C. efficiens	NC_004369	*trpE*	3052837..3054504	NP_739478
C. glutamicum	NC_003450	*trpE*	3233404..3234960	NP_602223
S. avermitilis	NC_003155	*trpS2*	complement(5757496..5758491)	NP_825902
		trpE1	complement(7320283..7322268)	NP_827260
S. coelicolor	NC_003888	*trpE*	2276703..2278607	NP_626374

b)

Bacterium	Operon	Leader peptide	
C. diphtheriae	*trpB1EGDC1*	2456514 ------MNAHN**WWW**RA-----	2456543
C. diphtheriae	*trpB2A*	2464983 -----MNAAFKF**WW**RA-----	2465015
C. efficiens	*trpEGDCBA*	3052621 VNNFCQSQGTQ**WWW**RAR----	3052671
C. glutamicum	*trpEGDCBA*	3233152 VNNSCLSQSTQ**WWW**RAN----	3233199
S. avermitilis	*trpS2*	5758647 ---MTTRTCTQQ**WW**AA-----	5758609
S. avermitilis	*trpE1*	7322414 ---MFAHSIQN**WWW**TAHPAAH	7322361
S. coelicolor	*trpE*	2276540 ---MFAHSTRN**WWW**TAHPAAH	2276593

c)

Bacterium	Operon	Attenuator
C. diphtheriae	*trpB1EGDC1*	**ugguggugg**cgcgcu**uaa**cc.gcgggcc.guuuu...cacgcauucauuuc.
C. diphtheriae	*trpB2A*	uuc**ugguggg**cgcgcc**uag**caggcgggcccc**uuuu**gugugagcauucaccaca
C. efficiens	*trpEGDCBA*	**ugguggugg**cgcgcuaga**uaa**gcgggcccacggaucaccaaguuguuuucac
C. glutamicum	*trpEGDCBA*	**ugguggugg**cgcgcuaac**uaa**gcgagccugacaccucaaguuguuuucacuu
S. avermitilis	*trpS2*	cag**ugguggg**gccgcc**uga**.cggcg.gccguacacacguauguacuc.....
S. avermitilis	*trpE1*	**ugguggugg**accgcucauccggcg.gcccac**uga**cugcgcgu.....
S. coelicolor	*trpE*	**ugguggugg**accgcucacccggcg.gcccac**uga**cugcgcgcg.....
S. venezuelae	*trpE*	**ugguggugg**accgcucacccggcg.gcccac**uga**ucgcgcgu.....

C. diphtheriae .....aac..aggcucgccuugucca....ac.aagcagcgggccuuuuuguuagc
C. diphtheriae .caacuuuuggaaacacaagcccgcguau.....c.gcgggcuuuuucguauau
C. efficiens .acugaagauuu...caaggcucguguacuucguucgacgaagcagcgggccuuuu.gugguuca
C. glutamicum ...ugaugaauuuuuuaaggcucgu..acuucguucgacgaagaagcgggccuuuu.gugguuuuu
S. avermitilis .....aacggccgccgccu.....cggcggccguucucguuucuc
S. avermitilis .acgcaagacuucgcgaaggccgccc.....gaggggcggccuuucgugguuuccg
S. coelicolor ..acucaagacucgcgaaggccgccc.....gaggggcggccuucgguguuuucg
S. venezuelae acacggaucacacgcacaggccgccc.....gaggggcggccuuucucg

Figure 1
Leader peptides and candidate attenuators upstream the *trp* operons in *Corynebacterium* and *Streptomyces* spp. a) Coordinates and protein identifiers of the first genes in the operons. b) Alignment of the leader peptides. The numbers denote genome positions of the aligned fragments. c) Alignment of the attenuators. Tryptophan and stop codons are shown in bold. The terminator hairpins are highlighted in grey, the antiterminator hairpins are underlined. The alignment contains fragments between the tryptohan codons and the terminator hairpin followed by poly-U-tracts. The numbers denote genome positions of the aligned fragments.

a)

Bacterium	Locus	Gene	Gene coordinates	Protein
M. avium	NC_002944	*MAP2122*	2351330..2352622	NP_961056
M. bovis	NC_002945	*cysK1*	2586392..2587324	NP_856011
M. tub CDC1551	NC_002755	*cysK*	2604640..2605572	NP_336875
M. tub H37Rv	NC_000962	*cysK*	2608794..2609726	NP_216850
M. leprae	NC_002677	*ML0840*	complement(997285..998589)	NP_301634
M. marinum	gnl|Sanger_216594|mar22d05.p1c	*cysK*	complement(136548..137477)	(unfinished)
P. acnes	NC_006085	*cysK*	1047389..1048324	YP_055674
B. longum	NC_004307	*cbs*	1006495..1007721	NP_696325

b)

Bacterium	Operon		Leader peptide	
M. avium	*XcysKE*	2351124	MQHRLQPRFAPSRCLVVACCCCCCR	2351177
M. bovis	*cysK1E*	2586122	MQQAIQLRFILPRRLAVGCCCC---	2586187
M. tub CDC1551	*cysKE*	2604371	MQQAIQLRFILPRRLAVGCCCC---	2604436
M. tub H37Rv	*cysKE*	2608526	MQQAIQLRFILPRRLAVGCCCC---	2608591
M. leprae	*XcysKE*	0998791	MHQSTQPRFVFTRRFTVDCYCRCC-	0998742
M. marinum	*cysKE*	0138059	MQQAAQLSFVLTRCPAVDCCCC---	0137994
P. acnes	*cysK*	1047061	MTSAMMVCICRCCC-	1047102
B. longum	*cbs*	1007876	MQIISCCCR-	1007850

c)

```
                      RBS   Start
M. avium          uauaguggugac**aug**caacaccgccuacagccgcgcuuu
M. bovis, tub     uauaguggggccc**aug**caacaggccauacagcugcgcuuu
M. leprae         uauaguggaccu**aug**caucaguccacacagccacgcuuu
M. marinum        uauaguagagcc**aug**caacaggccgcacagcugagcuuu

                                                Cys tract
M. avium          gccccgucgcgcugccuugucguggcc**uguuguugcuguuguugu**cgu
M. bovis, tub     auccucccgcgccgccucgccgugggc**uguuguuguugu**.........
M. leprae         gucuuuacgcgccgcuuuaccguggac**ugu**uau**ugu**cgc**uguugc**...
M. marinum        guccucacgcgcugccccgccguggac**uguuguuguugcugu**......

                  Stop and putative Rho binding site
M. avium          **ugA**UUUCCgcaaGCCCUCugacgcuguagaaAUCCCCgcgcucGCCCCUgcccg
M. bovis, tub     **ugA**UUCCUg.gcguccacagcaAUUCCUcgcGCUCUUgcccg
M. leprae         **ugA**UUCCUgac.ACCUUUuaacGCUCUCagcaaaucauucacGUUCUCgccua
M. marinum        **ugA**UUCCUgac.gcguucugaccguccaguaaucgucGCCUCUgucgccucaugg
```

Figure 2

Leader peptides upstream the *cys* operons in *Mycobacterium* spp. and *P. acnes* and *cbs* operon in *B. longum*. a) Coordinates and protein identifiers of the first genes in the operons. b) Alignment of the leader peptides. The numbers denote genome positions of the aligned fragments. c) DNA alignment of the leader peptide genes. Start, cysteine and stop codons are shown in bold; candidate Rho-binding sites are shown in capitals.

Leucine

The upstream regions of the *ilvB* genes (operons *ilvBNC, ilvBHC, ilvBserA*$_1$) in *Corynebactecterium, Mycobacterium, Streptomyces* species contain short ORFs with runs of isoleucine, valine and leucine codons overlapping the candidate terminator hairpins followed by polyU-runs (Fig. 3). However, the exact mode of regulation is not clear, as experimental substitution of possible regulatory codons upstream of the *ilvBNC* operon in *S. coelicolor* had no effect on regulation or expression of *ilvB* [23].

a)

Bacterium	Locus	Gene	Gene coordinate	Protein
C. diphtheriae	NC_002935	*ilvB*	1082013..1083971	NP_939459
C. efficiens	NC_004369	*ilvB*	1432330..1434327	NP_737975
C. glutamicum	NC_003450	*ilvB*	1338131..1340011	NP_600493
M. tuberculosis H37Rv	NC_000962	*ilvB*	complement(3361127..3362983)	NP_217519
M. tuberculosis CDC1551	NC_002755	*ilvB*	complement(3355506..3357362)	NP_337598
M. bovis	NC_002945	*ilvB1*	complement(3317745..3319601)	NP_856673
M. leprae	NC_002677	*ilvB*	complement(2044335..2046212)	NP_302166
M. avium	NC_002944	*ilvB1*	complement(3379032..3380900)	NP_961972
M. marinum gnl|Sanger_216594|mar755h11.p2k1114			complement(164709..166565)	(unfinished)
S. avermitilis	NC_003155	*ilvB*	complement(3354433..3356283)	NP_823909
S. coelicolor	NC_003888	*ilvB*	6003117..6004958	NP_629647

b)

Bacterium	Operon	Leader Peptide		
C. diphtheriae	*ilvBHC*	1081747	MNIIRLVVITTRRLP	1081791
C. efficiens	*ilvBHC*	1432212	MTSIRPVVIVAARRLP-	1432259
C. glutamicum	*ilvBHC*	1337840	MTIIRLVVVTARRLP	1337884
M. tuberculosis H37Rv	*ilvBNC*	3363152	MDKAGKPGMLVVIGRRVGA	3363096
M. tuberculosis CDC1551	*ilvBNC*	3357528	MDKAGKPGMLVVIGRRVGA	3357472
M. bovis	*ilvB1NC*	3319767	MDKAGKPGMLVVIGRRVGA	3319711
M. leprae	*ilvBNC*	2046378	MLVVICQRVGG	2046346
M. avium	*ilvB1N*	3381051	MLVVI-RRVGA	3381022
M. marinum	*ilvB*	166742	MDTAGTPGKLVVLGRRVVA	166686
S. avermitilis	*ilvBNC*	3356481	MRTRILVLGKRVG	3356443
S. coelicolor	*ilvBNC*	6002909	MRTRILVLGKRVG	6002947

c)

Bacterium	Terminator
C. diphtheriae	aaaagcg...cccucgacag....caccacacaugcugagcgggggcuuuccuuau
C. efficiens	caa.gcg...cccucgacaguaccaccacagugcuguuucgagggcuuuguugu.
C. glutamicum	caa.gcg...cccucgacaacacucaccacaguguuggaacgagggcuuucuuguu
M. tuberculosis	caacgcg..acccucgugcagcagc.......ugagcuggcga.ggguuuuuucuu
M. bovis	caacgcg..acccucgugcagcagc.......ugagcuggcga.ggguuuuuucuu
M. leprae	caacgcgcaacccucgugcagcuag.......ucagcugucga.ggguuuuuuguu
M. avium	caacgcgcaacccucgugcagcaca.........agcugucg.ggggguuuuuguu
M. marinum	caacgcgcaacccucgugcagcag.......cugagcugacg.ggggguuuuuguu
S. avermitilis	cggcgcgcuccccucgcuugcc..........ucacggcacgaggggguuuuuguu
S. coelicolor	cgacgcgcuccccucgcuugcc..........uuacggcacgaggggguuuuuguu

Figure 3

Candidate leader peptides and terminators upstream the *ilv* opreron in Actinobacteria. a) Coordinates and protein identifiers of the first genes in the operons. b) Alignment of the leader peptides. The numbers denote genome positions of the aligned fragments. c) Alignment of the terminators. The terminator hairpins are highlighted in grey.

Classical candidate attenuators were found upstream of *leuS* (leucyl-tRNA-synthetase) in *S. avermitilis* and *S. coelicolor*. Each of them contains an ORFs encoding the leader peptide, as well as the antiterminator and terminator hairpins (Fig. 4).

Sequences upstream of the isopropylmalate synthase genes *leuA* contain a number of candidate regulatory sequences, together named the LEU element (Fig. 5, 6). Firstly, there is an upstream ORF encoding a candidate leader peptide with a run of leucine codons (Fig. 7).

a)

Bacterium	Locus	Gene	Gene coordinates	Protein
S. avermitilis	NC_003155	*leuS*	6661895..6664783	NP_826665
S. coelicolor	NC_003888	*leuS*	complement(2775536..2778436)	NP_626809

b)

```
                          M  R  A  V  R  L  L  L  S  E  P  R
S. avermitilis 6661741 augcgugccguacgccuucugcuuagcgagccgcgcugaucagcccagaccacugacga
S. coelicolor  2778624 augcgugccguacgccuucugcuuagcgagccgcgcugaucagucccgaccccggucgu

S. avermitilis ..uuc.guggucggaaucggcgcggcgucccccuccugugcgaggggguuuuuucauu 6661852
S. coelicolor  aguccgguggccggaaucggcgcggcgucccccuccugugcgaggggauuuuucauu 2778510
```

Figure 4

Candidate attenuators upstream the *leuS* opreron in *Streptomyces* spp. a) Coordinates and protein identifiers of the *leuS* genes. b) Alignment of the attenuators. Start, leucine and stop codons are shown in bold. The terminator hairpins are highlighted in grey, the antiterminator hairpins are underlined. The alignment contains fragments between the leader peptide ORFs and the terminator hairpin followed by poly-U-tracts.

Secondly, this region may fold into a pseudoknot with an additional stem at its base formed by pairing of the leucine codon run with the Shine-Dalgarno box of the *leuA* gene (Fig. 5, 8). Finally, the same region may form an alternative hairpin with the same base stem (Fig. 6).

A similar pseudoknot was found in *B. longum* within a gene encoding a transposase. The latter is homologous to the IS1554 transposase of *M. tuberculosis* and *M. bovis* (66% identity), a putative transposase in *C. efficiens* (40% identity), putative IS256 family transposases of *S. avermitilis* (31% identity), hypothetical protein MAP2274 of *M. avium* (29% identity), and some other putative transposases from *B. longum*, *C. efficiens*, *M. tuberculosis*, *M. bovis*, *R. xylanophilus*, *S. avermitilis*, *S. coelicolor* (Fig. 9a). However, only the *B. longum* transposase contains a fragment that may fold into the pseudoknot (Fig. 9b), whereas other transposases, although highly similar on the protein level in the corresponding region, contain a number of non-complementary mismatches in synonymous codon positions and thus have lost the pseudoknot folding potential.

T-boxes

Candidate T-box structures were found upstream of the *ileS* genes from several Actinobacteria. They are unusual, as instead of terminators, they contain hairpins sequestering the Shine-Dalgarno boxes of the *ileS* genes (Fig. 10). Thus it is likely that the regulatory mechanism involves inhibition of translation initiation. To our knowledge, this is the first example of a T-box acting on the level of translation.

Conclusion

Candidate regulatory elements were found upstream of genes involved in the tryptophan, cysteine and branched chain amino acids metabolism. No conserved RNA regulatory structures were observed upstream of histidine, threonine, phenylalanine, tyrosine, arginine, lysine, methionine operons, although orthologous genes involved in the latter pathways are regulated on the RNA level in other species: methionine and lysine by the S-box and L-box riboswitches respectively [3-5], histidine, threonine and phenylalanine by attenuators [7,8], tyrosine and arginine by T-boxes [12].

Attenuators of the classical type were observed upstream of the aminoacyl-tRNA-synthetase genes *trpS* and *leuS* in some *Streptomyces* genomes, similar to those observed in gamma-proteobacteria, (e.g. the *pheST* operon) [7]. In contrast, in Firmicutes, most aminoacyl-tRNA-synthetase genes are regulated by tRNA-dependent antitermination (T-boxes) and none by classical attenuation [2,9,15]. No classical T-boxes were found in Actinobacteria, but unusual T-boxes, possibly regulating initiation of translation,

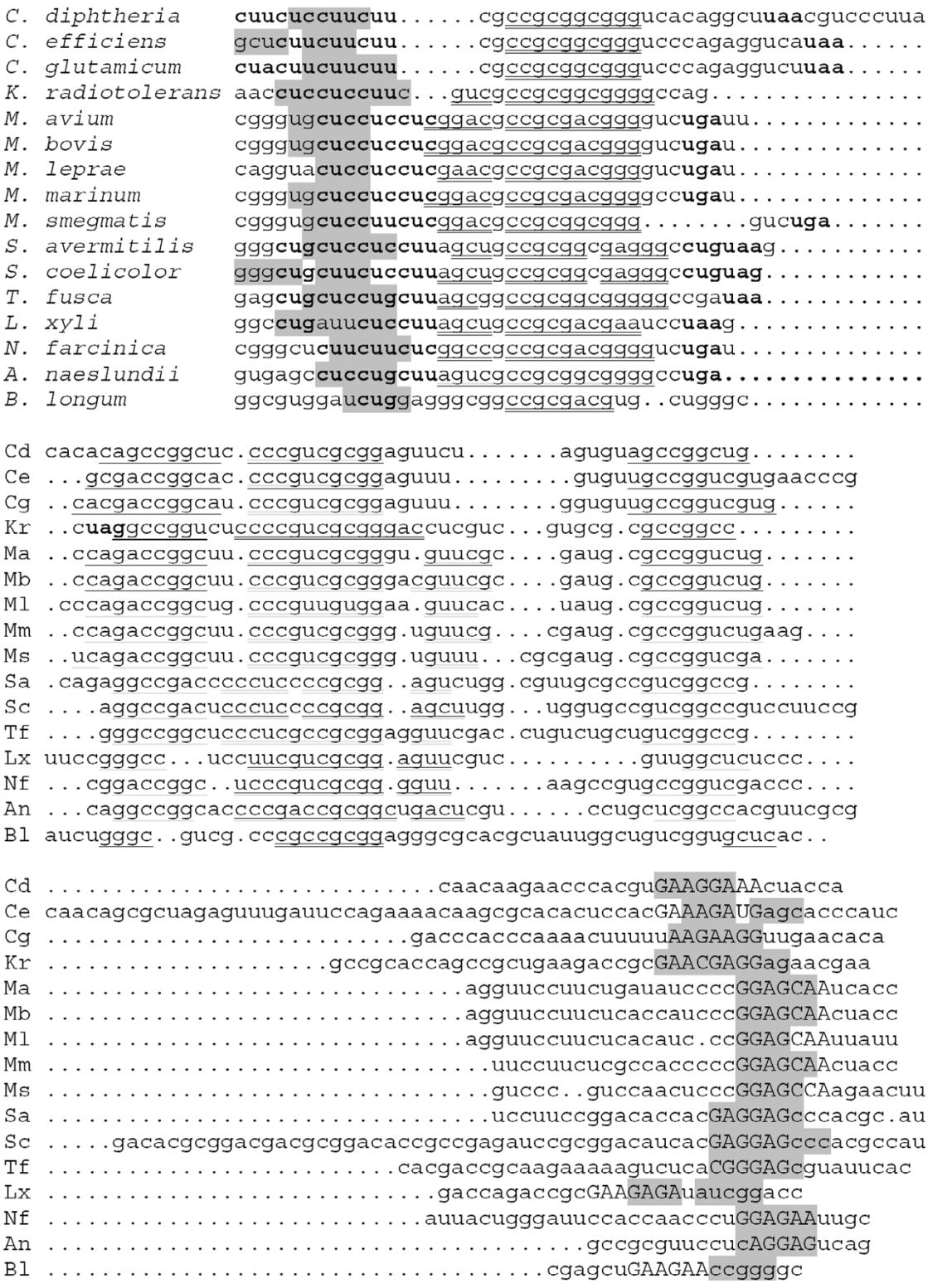

Figure 5

Alignment and RNA secondary structures of the *leuA* upstream regions (LEU elements). The stem at the base is highlighted in grey, helices forming the pseudoknot are underlined and double underlined, leucine and stop codons are set in bold, the candidate Shine-Dalgarno boxes of the *leuA* are set in capitals. The last sequence is that of the transposase from *B. longum* (see the text). Sequences for *M. bovis* (Mb) and *M. tuberculosis* spp. (Mt and Rv) coincide.

Figure 6

Alternative RNA secondary structure in LEU elements. The stem at the base is highlighted in grey, two internal helices are underlined and double underlined, other notation as in Fig 5.

a)

Bacterium	Locus	Gene	Gene coordinates	Protein
C. diphtheria	NC_002935	*leuA*	complement(228555..230372)	NP_938656
C. efficiens	NC_004369	*leuA*	complement(233589..235439) (adding 105 nucleotides)	NP_736826
C. glutamicum	NC_003450	*leuA*	complement(266151..268001)	NP_599502
K. radiotolerans	AAEF02000060	*leuA*	complement(3238..4965)	EAM73829
M. avium	NC_002944	*leuA*	333789..335633	NP_959246
M. bovis	NC_002945	*leuA*	4091088..4093193	NP_857375
M. tub CDC1551	NC_002755	*leuA*	4145949..4147928	NP_338367
M. tub H37Rv	NC_000962	*leuA*	4153737..4155671	NP_218227
M. leprae	NC_002677	*leuA*	2754640..2756463	NP_302512
M. marinum	gnl|Sanger_216594|mar428a07.p1k		192528..194345	(unfinished)
M. smegmatis	gnl|TIGR_246196|contig:3563:m_smegmatis		6334690..6336495	(unfinished)
S. avermitilis	NC_003155	*leuA2*	6774328..6776049	NP_826778
S. coelicolor	NC_003888	*leuA*	complement(2725480..2727201)	NP_733575
T. fusca	NZ_AAAQ02000002	*leuA*	349237..350943 (adding 27 nucleotides)	ZP_00293601
L. xyli	NC_006087	*leuA*	complement(1501628..1503400)	YP_062368
N. farcinica	NC_006361	*leuA*	complement(322994..324787)	YP_116514
A. naeslundii	gnl|TIGR_240017|contig:1063:a_naeslundii		594374..596211	(unfinished)

b)

Bacterium		Leader peptide	
C. diphtheria	230506	MNRANLLLLRRGGSQA-	230459
C. efficiens	235612	MFSSHERSALLLRRGGSQRS	235553
C. glutamicum	268124	MTSRANLLLLRRGGSQRS	268095
K. radiotolerans	5097	VARLENLLLRRRGGAS-	5050
M. avium	333705	VADVQRVLLLGRRDGV--	333752
M. bovis	4090959	VLHVQRVLLLGRRDGV--4091006	
M. tub CDC1551	4145866	VLHVQRVLLLGRRDGV--4145913	
M. tub H37Rv	4153611	VLHVQRVLLLGRRDGV--4153658	
M. leprae	2754521	VQQVLLLERRDGV--2754559	
M. marinum	192399	VLCVQRVLLLGRRDG---	192443
M. smegmatis	6334564	VLGVQRVLLLGRRGGV--6334611	
S. avermitilis	6774199	MRFGLLLLSCRGEGL-6774243	
S. coelicolor	2727361	MRFGLLLLSCRGEGL-2727317	
T. fusca	349104	MLRELLLLSGRGGGR-	349148
L. xyli	1503533	MRVTLGLVYGLILLSCRDES--1503474	
N. farcinica	324906	MQRALLLGRRDGV--	324868
A. naeslundii	594266	VSLLLSRRGGA--	594298

Figure 7
Candidate leader peptides in the LEU elements.

were observed upstream of the *ileS* genes in several genomes.

Despite the presense of conserved leader peptides upstream of some cysteine and leucine operons, the mode of regulation is unknown, as other attenuator elements are missing. One possible explanation is that attenuation of the *cys* operons in *Mycobacterium* spp. and *P. acnes* and the *cbs* operon in *B. longum* involves Rho-dependent termination, similar to the *tna* operon of *E. coli* [23,29].

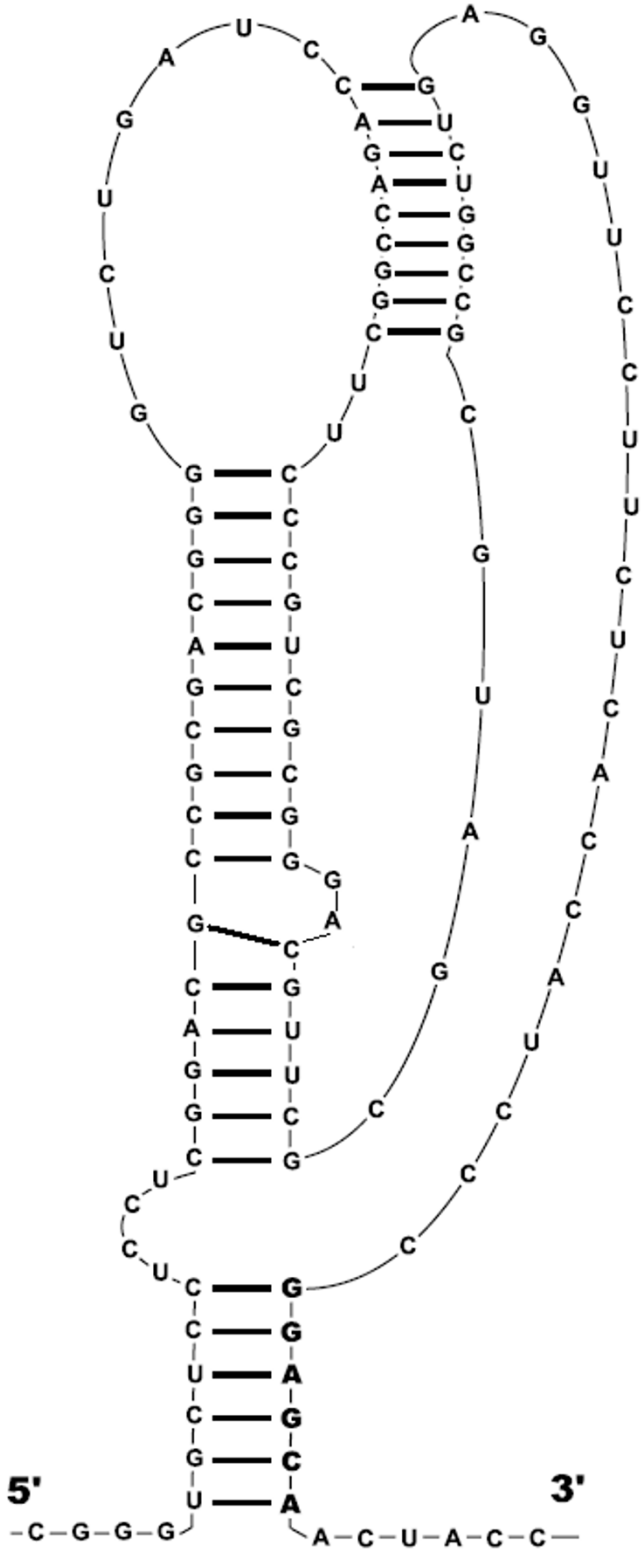

Figure 8
Candidate RNA pseudoknot upstream of the *leuA* operon in *M. bovis*. The corresponding alignment is given Fig. 5. Boldface: the candidate Shine-Dalgarno box.

The most interesting case seems to be that of the *leuA* genes. The upstream regions of these genes contain several conserved elements (referred to as the LEU element) that can be interpreted in different ways. There are some architectural similarities with riboswitches, in particular, a compact structure with a stem at the base [5,30,31]. The latter is formed by interaction of a run of leucine codons and the Shine-Dalgarno box. Indeed, Actinobacteria seem to be the only taxonomic group where the base stems of riboswitches directly overlap the translation initiation site, without additional regulatory hairpins [5]. However, the LEU element differs from all known riboswitches, as the alignment of LEU elements does not contain conserved unpaired nucleotides that would be involved in tertiary interactions and form the ligand-binding pocket, as in the purine riboswitches whose spatial structure has been resolved [30,31] and in other riboswitches [5]. Thus direct binding of a small molecule to LEU elements seems unlikely. On the other hand, there is experimental evidence that mutations in the leucine codons do not influence the regulation [22] and thus classical attenuation involving translation of a leader peptide also is an unlikely mechanism of regulation.

The above considerations make it likely that the LEU element is a binding site of some regulatory protein. To test for this possibility, we compared the pattern of phylogenetic distribution of LEU elements to phylogenetic distributions of all actinobacterial genes. The closest phylogenetic pattern was observed for orthologs of ML1624 from *M. leprae*: homologs of this protein with E-values $<10^{-170}$ were found in all genomes containing LEU elements, but not outside Actinobacteria. The only unexplained fact is the presence of a homolog with the E-value $\sim 10^{-108}$ in *P. acnes*, which does not have a LEU element. The structure of the ML1642 protein is consistent with an RNA-binding regulatory role, as the protein contains an N-terminal DEAD-box helicase domain (ProFam family PF00270, E-value $3.6 \cdot 10^{-6}$) that may be involved in unwinding of nucleic acids.

An additional enigma is the presence of a LEU element-like sequence within a transposase gene. On the other hand, it may be a clue to the origin of LEU elements. One possibility is that the *B. longum* transposase represents an ancestral state where the LEU element was involved in maintenance or regulation of transposition. Situations when a regulatory site occurs within a regulatory and/or regulated gene are not very common, but they happen in mobile elements [32]. Other transposase genes may have lost the ability to form this structure due to mutations; notably, the protein sequence has not changed much (Fig. 9), as most mutations occurred in synonymous codon positions. A plausible scenario is that the transposase gene was inserted upstream of the *leuA* gene in the ancestral

a)

Bacterium	Locus	Coordinates	Protein
B. longum	NC_004307	2124903.. 2126108	NC_004307
M. bovis	NC_002945	complement(1025963..1027282)	NP_854601
M. tuberculosis CDC1551	NC_002755	complement(1025510..1026829)	NP_335380
C. efficiens	NC_004369	complement(1561522..1562694)	NP_738106

b)

```
B. longum
M. bovis                                                                MDAAQVIEPAHAGQDVDEAAVAA
M. tuberculosis CDC1551                                                 MDAAQVIEPAHAGQDVDEAAVAA
C. efficiens

Bl                    MAKEKGLDLTGPDGLLKQFTKSVLETALDEEMTEHLGR**AKHKKSKDGRAANTRNGTTAKTVVTDSVGPVGIEVPRDRDGS
Mb     RELSGAERALVGDLVRQARAEGVALTGPDGLLKALTKTVLEAALQEEMTEHLGY***DRHAAAGRGSGNSRNGSRNKKVITDACGQVEIAVPRDRNGT
Mt     RELSGAERALVGDLVRQARAEGVALTGPDGLLKALTKTVLEAALQEEMTEHLGY***DRHAAAGRGSGNSRNGSRNKKVITDACGQVEIAVPRDRNGT
Ce                                                MNAEMDAHLGYGHSDRDGKTAAGQGNHRNGYYPK*RVDSNYGPIDVAVPRDRNGS
                      ******************************EM**HLG**************N*RNG***K*******G*****VPRDR*GS

Bl    FEPVVVRKRQRRLPGVDEVVLSLYARGLTTGEISAHFQEIYGADVSRETVSRITERVVAEKDEWCSRPLDRVYAAVFIDATVVKVRDG*QVANRAFYVAV
Mb    FEPVIVGKRKRRVTDVDRVVLSLYAKGLTTGEIAAHFADVYGVSVSKDTISRITDRVIEEMQAWWSRPLEKVYAAVFIDAIMVKIRDG*QVRNRPVYAAI
Mt    FEPVIVGKRKRRVTDVDRVVLSLYAKGLTTGEIAAHFADVYGVSVSKDTISRITDRVIEEMQAWWSRPLEKVYAAVFIDAIMVKIRDG*QVRNRPVYAAI
Ce    FLPTMVPKGSRRLTDVDDMII*LYAGGMTVRDIQHHMITSMGVDISHETISAITDAVLDEVMIWQNRQLDDFYPVIFLDALRIKVRDGGRVVNKSVYLAI
      F*P**V*K**RR***VD*****LYA*G*T***I**H*****G***S**TIS*IT*AV**E***W**R*L***Y***F*DA***K*RDG**V*N***Y*A*

      ================== ==================
Bl    GVDLEGGRDVLGIWASPA*AEGARYWLSVLTELKNRGVDDVFFLICDGLKGLPDAVGAVWPLAIVQTCVVHLLRNTFRYASKKDWDAIKRDVKPIYTAPS
Mb    GVDLDGHKDILGMWAGEGDGESAKFWLAVLTELRNRGVKDIFFLVCDGLKGLPDSVSAAFPLATVQTCIIHLIRNTFRYASRKYWDKISVDLKPIYTAAS
Mt    GVDLDGHKDILGMWAGEGDGESAKFWLAVLTDLRNRGVKDIFFLVCDGLKGLPDSVSAAFPLATVQTCIIHLIRNTFRYASRKYWDKISVDLKPIYTAAS
Ce    GVDIDGIKHILGIWLAKE**EGASFWANVCANLATRGVQDVFIVCCDGLKGLPQAVEATWPDSMVQTCVVHLIRAANRWVAYGDRKAVSAQLRKIYTAPT
      GVD**G****LG*W******E*A*FW**V***L**RGV*D*F***CDGLKGLP**V*A**P***VQTC**HLIR***R***************IYTA**

Bl    **AAAAAAARDAMLDKWEARYPAIRRLWMDAWERFIPFLDYDVEIRRVICTTNAIESLNARFKRSIRARGHFPDEQAALKCMYLTVRSLDPTGKGRIRWS
Mb    **AAEARLRYEEFAEKWGKPYPAITRLWDSAWEEFIPFLDYDVEIRRVPCSTNAIESLNARYRRAVRARGHFPNEQSALKTLYLVTRSLDPKGTGQTKWA
Mt    **AAEARLRYEEFAEKWGKPYPAITRLWDSAWEEFIPFLDYDVEIRRVPCSTNAIESLNARYRRAVRARGHFPNEQSALKTLYLVTRSLDPKGTGQTKWA
Ce    EDTAIAALEEFEASELGVK*YPQSAKVWRDAWDRFIPFLQFPPMARKVIYTTNSIESMNNELRKATRNRVQFTNDESAIKTLWLMICNIEDKRAAKRAKQ
      ***A*A**************YP*****W**AW**FIPFL******R*V***TN*IES*N*******R****F*****AIK***L***************

Bl     ARWKPALNAFAITFADRWPSEGTQQ
Mb     VRWKPALNALAITFADRMPAAEER
Mt     VRWKPALNALAITFADRMPAAEER
Ce     GKRVAASSGRLIEGRKVANWKQAINQMAVAFPDRFEAYL
       **********I*************
```

c)

```
B. longum       G  V  D  L  E  G  G  R  D  V  L  G  I  W  A  S  P  A  *  A  E
                ggcgtggatctggagggcggccgcgacgtgctgggcatctgggcgtcgcccgcc***gcggag
M. bovis        G  V  D  L  D  G  H  K  D  I  L  G  M  W  A  G  E  G  D  G  E
                ggcgtcgacctcgacggccacaaggacatcctggggatgtgggccggcgaaggcgacggtgag
M. tuberculosis G  V  D  L  D  G  H  K  D  I  L  G  M  W  A  G  E  G  D  G  E
                ggcgtcgacctcgacggccacaaggacatcctggggatgtgggccggcgaaggcgacggtgag

B. longum       G  A  R  Y  W  L  S  V  L  T  E  L  K  N  R  G
                ggcgcacgctattggctgtcggtgctcaccgagctgaagaaccggggc
M. bovis        S  A  K  F  W  L  A  V  L  T  E  L  R  N  R  G
                tcagccaaattttggctggcagtgctcaccgaactgcgcaatcgtggg
M. tuberculosis S  A  K  F  W  L  A  V  L  T  D  L  R  N  R  G
                tcagccaaattttggctggcagtgctcaccgacctgcgcaatcgtggg
```

Figure 9

Multiple alignments of transposases. a) Coordinates and protein identifiers of putative transposases. b) Protein alignment. The fragment marked by the double line above corresponds to the *B. longum* fragment homologous to candidate pseudoknot and shown in the last line of Fig. 5. c) Nucleotide alignment of the region shown by the double line in (b).

a)

Bacterium	Locus	The *ileS* gene coordinates	Protein
A. naeslundii	gnl|TIGR_240017|contig:1063:a_naeslundii	complement(1311947..1315252)	unfinished
C. diphtheriae	NC_002935	complement(1617227..1620385)	NP_939931
C. efficiens	NC_004369	complement(2160737..2164195) 49 codons removed	NP_738653
C. glutamicum	NC_003450	complement(2270986..2274150)	NP_601350
M. avium	NC_002944	1324371..1327532	NP_960180
M. bovis	NC_002945	1720532..1723657	NP_855215
M. tub H37Rv	NC_000962	1736519..1739644	NP_216052
M. tub CDC1551	NC_002755	1736672..1739797	NP_336040
M. marinum	gnl|Sanger_216594|mar288e12.s1k	complement(184205..187372)	unfinished
M. leprae	NC_002677	1410785..1413964	NP_301871
N. farcinica	NC_006361	1932119..1935247	YP_117986
P. acnes	NC_006085	268050..271394	YP_054935
R. xylanophilus	NZ_AAEB01000029	complement(26358..29492)	ZP_00187197
S. avermitilis	NC_003155	complement(7371348..7374491)	NP_827306
S. coelicolor	NC_003888	2227237..2230380	NP_626335
T. fusca	NZ_AAAQ02000011	complement(75752..78934)	ZP_00291779

b)

Figure 10

Multiple alignment of T-box structures upstream of the *ileS* genes. a) Coordinates and protein identifiers of the *ileS* genes. b) Nucleotide alignment of the 5' untranslated regions. T-box hairpins are underlined and T-box sequences are set in capitals. The sequestor hairpin is shaded in grey. Candidate Shine-Dalgarno boxes are set on capitals. Anti-sequestor hairpins are set in bold.

actinobacterial genome. The main fraction of the coding sequence was subsequently deleted, whereas the structural element was co-opted for regulation of the downstream *leuA* gene.

Methods

Genomes of Actinobacteria *Actinomyces naeslundii* (An), *Bifidobacterium longum* (Bl), *Corynebacterium diphtheriae* (Cd), *Corynebacterium efficiens* (Ce), *Corynebacterium glutamicum* (Cg), *Kineococcus radiotolerans* (Kr), *Leifsonia xyli* (Lx), *Mycobacterium avium* (Ma), *Mycobacterium bovis* (Mb), *Mycobacterium leprae* (Ml), *Mycobacterium marinum* (Mm), *Mycobacterium smegmatis* (Ms), *Mycobacterium tuberculosis* (Rv and Mt), *Nocardia farcinica* (Nf), *Propionibacterium acnes* (Pa), *Rubrobacter xylanophilus* (Rx), *Streptomyces avermitilis* (Sa), *Streptomyces coelicolor* (Sc), *Thermobifida fusca* (Tf), *Tropheryma whipplei* (Tw) were downloaded from the NCBI web site. We also used sequences of *Streptomyces venezuelae* (Sv) from [21].

Candidate operons were defined as chains of genes transcribed in the same direction with intergenic regions not exceeding 150 nucleotides. Multiple alignments of genes were used to verify and, if necessary, revise annotated gene starts [33]. The revisions included adding 105 nucleotides (35 codons) to the *leuA* gene from *C. efficiens*, adding 27 nucleotides (9 codons) of the *leuA* gene from *T. fusca*, and removing 147 nucleotides (49 codons) of the *ileS* gene from *C. efficiens*.

RNA sequence and structure alignments were constructed using MultAlign (A.A. Mironov, personal communication) and the program GL [34]. Search for RNA structural patterns was performed using the PAT program (A.V.Seliverstov, unpublished). Search for conserved sequence fragments was done using the CLIQUE program [35]. Multiple protein sequence alignments were constructed using MultAlign.

Authors' contributions

AVS and VAL developed algorithms. AVS wrote the programs and performed sequence analysis. HP and AVS identified translational T-boxes. AVS, VAL, and MSG analyzed LEU elements. AVS and MSG performed functional annotation and wrote the paper. VAL and MSG conceived and supervised the project.

Acknowledgements

We are grateful to Andrei Mironov for the MultAlig program and to K.Yu. Gorbunov for useful discussion. This work was partially supported by grants from ISTC (2766), Howard Hughes Medical Institute (55000309), Russian Academy of Sciences (programs "Molecular and Cellular Biology" and "Origin and Evolution of the Biosphere"), Fund for Support of Russian Science.

References

1. Henkin TM, Yanofsky C: **Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions.** *Bioessays* 2002, **24:**700-707.
2. Grundy FJ, Henkin TM: **The T box and S box transcription termination control systems.** *Front Biosci* 2003, **8:**d20-31.
3. Grundy FJ, Henkin TM: **Regulation of gene expression by effectors that bind to RNA.** *Curr Opin Microbiol* 2004, **7:**126-131.
4. Mandal M, Breaker RR: **Gene regulation by riboswitches.** *Nat Rev Mol Cell Biol* 2004, **5:**451-463.
5. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Riboswitches: the oldest mechanism for the regulation of gene expression?** *Trends in Genetics* 2004, **20(1):**44-50.
6. Yanofsky C: **The different roles of tryptophan transfer RNA in regulating trp operon expression in *E. coli* versus *B. subtilis*.** *Trends in Genetics* 2004, **20:**367-74.
7. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria.** *J Mol Microbiol Biotechnol* 2001, **3:**529-543.
8. Vitreschak AG, Lyubetskaya EV, Shirshin MA, Gelfand MS, Lyubetsky VA: **Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis.** *FEMS Microbiology Letters* 2004, **234:**357-370.
9. Grundy FJ, Henkin TM: **Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria.** *J Mol Biol* 1994, **235:**798-804.
10. Grundy FJ, Henkin TM: **The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria.** *Mol Microbiol* 1998, **30:**737-749.
11. Murphy BA, Grundy FJ, Henkin TM: **Prediction of gene function in methylthioadenosine recycling from regulatory signals.** *J Bacteriol* 2002, **184:**2314-2318.
12. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria.** *FEMS Microbiol Lett* 2003, **222:**211-220.
13. Sudarsan N, Barrick JE, Breaker RR: **Metabolite-binding RNA domains are present in the genes of eukaryotes.** *RNA* 2003, **9:**644-7.
14. Rodionov DA, Vitreschak AA, Mironov AA, Gelfand MS: **Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes.** *J Biol Chem* 2003, **277:**48949-48959.
15. Henkin TM, Glass BL, Grundy FJ: **Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes.** *J Bacteriol* 1992, **174:**1299-1306.
16. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR: **New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control.** *Proc Natl Acad Sci USA* 2004, **101:**6421-6426.
17. Abreu-Goodger C, Ontiveros-Palacios N, Ciria R, Merino E: **Conserved regulatory motifs in bacteria: riboswitches and beyond.** *Trends Genet* 2004, **20:**475-9.
18. Vitreschak AA, Rodionov DA, Mironov AA, Gelfand MS: **Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation.** *Nucleic Acids Research* 2002, **30:**3141-3151.
19. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: **Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element.** *RNA* 2003, **9:**1084-1097.
20. Heery DM, Dunican LK: **Cloning of the *trp* gene cluster from a tryptophan-hyperproducing strain of *Corynebacterium glutamicum*: Identification of a mutation in the *trp* leader sequence.** *Applied and Environmental Microbiology* 1993, **59:**791-799.
21. Lin C, Pradkar AS, Vining LC: **Regulation of an antranilate synthase gene in *Stryptomyces venezuelae* by *trp* attenuator.** *Microbiology* 1998, **144:**1971-1980.
22. Craster HL, Potter CA, Baumberg S: **End-product control of branched-chain amino acid biosynthesis genes in *Streptomyces coelicolor* A3(2): paradoxical relationships between DNA sequence and regulatory phenotype.** *Microbiology* 1999, **145:**2375-2384.

23. Konan KV, Yanofsky C: **Rho-dependent transcription termination in the *tna* operon of *Escherichia coli*: Roles of the boxA sequence and the *rut* site.** *Journal of Bacteriology* 2000, **182:**3981-3988.
24. Washio T, Sasayama J, Tomita M: **Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination.** *Nucleic Acids Research* 1998, **26:**5456-5463.
25. Unniraman S, Prakash R, Nagaraja V: **Conserved economics of transcription termination in eubacteria.** *Nucleic Acids Research* 2002, **30:**675-684.
26. Richardson JP: **Rho-dependent termination and ATPases in transcript termination.** *Biochimica et Biophysica Acta* 2002, **1577:**251-260.
27. Richardson JP: **Structural organization of transcription termination factor Rho.** *The Journal of Biological Chemistry* 1996, **271:**1251-1254.
28. Kaplan DL, O'Donne M: **Rho factor: transcription termination in four steps dispatch.** *Current Biology* 2003, **13:**R714-R716.
29. Gong F, Yanofsky C: **Rho's role in transcription attenuation in the *tna* operon of *E. coli*.** *Methods Enzymol* 2003, **371:**383-391.
30. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ: **Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs.** *Chem Biol* 2004, **11:**1729-1741.
31. Batey RT, Gilbert SD, Montange RK: **Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine.** *Nature* 2004, **432:**411-415.
32. Koonin EV, Ilyina TV: **Computer-assisted dissection of rolling circle DNA replication.** *BioSystems* 1993, **30:**241-268.
33. Baytalulk MV, Gelfand MS, Mironov AA: **Exact mapping of prokaryotic gene starts.** *Briefings in Bioinformatics* 2002, **3:**181-194.
34. Gorbunov KYu, Mironov AA, Lyubetsky VA: **Search for conserved secondary structures of RNA.** *Mol Biol* 2003, **37:**850-860.
35. Lyubetsky VA, Seliverstov AV: **Selected algorithms related to finite groups.** *Information Processes* 2003, **3:**39-46. (in Russian)