Computational Biology and Chemistry 49 (2014) 7–13

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism

Vassily A. Lyubetsky*, Semen A. Korolev, Alexandr V. Seliverstov, Oleg A. Zverkov, Lev I. Rubanov

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Russia

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form 6 January 2014
Accepted 9 January 2014

Keywords:
Mycobacteria
Sulfur metabolism
Sulfur-containing compounds transport
Cysteine
Rho-dependent transcription attenuation
Classic transcription attenuation

ABSTRACT

The paper studies proteins with domains PF00480 or PF14340, as well as some other poorly characterized proteins, encoded by genes associated with leader peptide genes containing a tract of cysteine codons. Such proteins are hypothetically regulated with cysteine-dependent transcription attenuation, namely the Rho-dependent or classic transcription attenuation. Cysteine is an important structural amino acid in various proteins and is required for synthesis of many sulfur-containing compounds, such as methionine, thiamine, glutathione, taurine and the lipoic acid. Earlier a few species of mycobacteria were predicted by the authors to have cysteine-dependent regulation of operons containing the *cysK* gene. In *Escherichia coli* this regulation is absent, and the same operon is regulated by the CysB transcription activator. The paper also studies Rho-dependent and classic transcription regulations in all annotated genes of mycobacteria available in GenBank and their orthologs in Actinomycetales. We predict regulations for many genes involved in sulfur metabolism and transport of sulfur-containing compounds; these regulations differ considerably among species. On the basis of predictions, we assign a putative role to proteins encoded by the regulated genes with unknown function, and also describe the structure of corresponding regulons, predict the lack of such regulations for many genes. Thus, all proteins with the uncharacterized Pfam domains PF14340 and PF00480, as well as some others, are predicted to be involved in sulfur metabolism. We also surmise the affinity of some transporters to sulfur-containing compounds. The obtained results considerably extend earlier large-scale studies of Rho-dependent and classic transcription attenuations.

1. Introduction

We study putative proteins encoded by genes associated with leader peptide genes containing a *tract of cysteine codons*, among them proteins with the PF00480 (ROK) and PF14340 (DUF4395) domains. We hypothesize that gene expression of such proteins is regulated with cysteine-dependent transcription attenuation, namely the Rho-dependent or classic transcription attenuation. Cysteine is an important structural amino acid in various proteins and is required for synthesis of many sulfur-containing compounds, such as methionine, thiamine, glutathione, taurine and the lipoic acid.

Studies of the Rho-dependent and classic attenuation regulations mediated by the concentration of tryptophan date back to the pioneer works by C. Yanofsky (Konan and Yanofsky, 2000; Yanofsky, 1981). These regulations are characterized by an open reading frame of the leader peptide gene (further referred to as the *leader gene*) that possesses a short tract of regulatory codons. The regulation mechanism is based on the transcription and translation coupling, a common phenomenon in prokaryotes.

In the case of classic attenuation the leader gene is separated from a downstream structural gene (or an operon) by a relatively short spacer region, which transcript sequence forms one or several terminator hairpins and often contains very short poly(U) (U-rich) regions positionally associated with the hairpins' 3′-shoulders. A complex of a terminator and its neighboring poly(U) is called the intrinsic *terminator*. The formation of the terminator hairpins is regulated by other (one or several) hairpins located upstream and called *antiterminators*, or by the ribosome attached to leader gene's codons.

The study (Yanofsky, 1981) established a mechanism of classic attenuation regulation based on the mutually exclusive formation of overlapping terminator and antiterminator on mRNA sequence (transcription attenuation sensu Yanofsky). We proposed alternatives of this mechanism that often involve the formation of pseudoknots, triplexes, and additional hairpins. For each alternative, large-scale predictions were obtained with biological data

* Corresponding author at: IITP RAS, 19 Bolshoy Karetny Lane, Moscow 127994, Russia. Tel.: +7 910 4646917.
E-mail address: lyubetsk@iitp.ru (V.A. Lyubetsky).

http://dx.doi.org/10.1016/j.compbiolchem.2014.01.001
1476-9271/

(Lopatovskaia et al., 2010). In particular, classic attenuation mediated by concentrations of tryptophan, phenylalanine, histidine, threonine or ramified amino acids is predicted in large-scale analyses using bioinformatic tools (Lopatovskaia et al., 2010; Seliverstov et al., 2005); further references are provided in the cited works.

In the case of Rho-dependent attenuation regulation, the stop codon of the leader gene is associated with a short {C,U}-rich site called the *Rho binding site*, which is specifically recognized by the Rho factor. A pyrimidine-rich region may be considered as a putative Rho binding site, which however does not suffice to draw further conclusion. For the Rho factor, the ATPase activity was measured for different nucleotide compositions of mRNA (Kim and Patel, 1999). It was shown to be 3–4-fold higher with poly(C) compared to poly(U), and tens of thousands times lower if no mRNA is available. This and some other indirect evidence did not so far bring more clarity to the definition of the Rho binding site.

Rho-dependent transcription attenuation was first described for expression regulation of the tryptophanase gene *tna* in *Escherichia coli* (Konan and Yanofsky, 2000). This case remains the best studied to date. Another example (Richardson, 2002), a Rho binding site located closely to the start codon in the coding region of the *lac* operon in *E. coli*. For a few species of mycobacteria we earlier predicted the cysteine-mediated Rho-dependent transcription attenuation of the operons containing the *cysK* gene (Seliverstov et al., 2005). Note that *E. coli* lacks this regulation, and the same operon is regulated with an alternative mechanism mediated by the CysB transcription activator. The Rho transcription factor is known to be present in all mycobacteria, with the exception of *Mycobacterium africanum* GM041182.

Under the excess of a regulating amino acid the operon is active, because the Rho binding site is shielded by a ribosome. When the amino acid is in deficiency (e.g., during starvation), ribosomes do not occupy the Rho binding site on mRNA, and it becomes available to the Rho factor that terminates transcription even when the 5′-end of the structural operon has already been transcribed. Functioning of this mechanism can cause the opposite (decreasing instead of increasing) correlation depending on the distance between the regulatory codons and the Rho binding site. The Rho factor can also interact with riboswitches (Hollands et al., 2012). Elsewhere (Konan and Yanofsky, 2000; Yanofsky, 1981; Heery and Dunican, 1993; Lin et al., 1998), mentioned types of the regulations are corroborated experimentally.

In the case of classic transcription attenuation (Lopatovskaia et al., 2010; Seliverstov et al., 2005), the increase of amynoacyl-tRNA concentration is associated with the decrease of transcription continuation frequency caused by the RNA polymerase passing the putative transcription termination site of the structural gene (usually a U-rich tract within the regulatory region). This frequency is called the *transcription level* and is measured in unit fractions (or in percent) as $1-p(c)$, where c is the amynoacyl-tRNA concentration, and $p(c)$ – the frequency of transcription termination (DNA·RNA-duplex dissociation) usually occurring on a U-rich tract. In regulations of catabolism, transport, etc., the transcription level increases with the increase of amynoacyl-tRNA concentration. The dependency $p(c)$ is estimated according to a model of classic attenuation (Lyubetsky et al., 2006, 2007; Rubanov and Lyubetsky, 2007). This model is systematically applied herein to predict regulation on the basis of a noticeable change of the structural gene transcription level $1-p(c)$.

Proteobacteria lack cysteine-dependent transcription attenuation (Lopatovskaia et al., 2010) and possess other regulation types (Kredich, 1992; Lynch et al., 1994). The regulation of sulfur metabolism substantially differs across proteobacterial lineages. In *E. coli*, certain pathways, including that of cysteine synthesis, are regulated by the CysB protein from the LysR family, which binds DNA close to the promoter. In proteobacteria *Salmonella typhimurium*, *E. coli* and *Klebsiella aerogenes* the transcription of the operons *cysPTWAM*, *cysK*, *cysJIH*, *cysDNC*, *sbp*, and the L-cysteine transport system is also activated by the CysB protein (Kredich, 1992; Lynch et al., 1994). In these proteobacteria, self-repression of transcription is described for the gene *cysB*. The CysB·DNA binding is sensitive to concentration of the *O*-acetylserine cysteine precursor but is not sensitive to that of sulfuric compounds. In *E. coli* and other γ-proteobacteria the methionine synthesis is known to be regulated by the MetJ factor, which suppresses transcription upon binding to S-adenosylmethionine (Saint-Girons et al., 1984; Augustus and Spicer, 2011). In some Gram-positive bacteria, including *Bacillus subtilis*, *Clostridium acetobutylicum*, and *Staphylococcus aureus*, the regulation of sulfur metabolism is mediated by the S-box riboswitch (Grundy and Henkin, 1998).

The Rho protein is a homohexamer. Each of its subunits contains two domains, one binding mRNA, and the other being an ATPase. Among the six identical subunits of the hexamer only two possess the ATPase activity due to the molecule's asymmetry and assembly-dependent conformational properties of adjacent subunits. In in vitro experiments the Rho protein binds a 78(C)-long region.

Remember that at the initial stage of cysteine synthesis the serine hydroxyl is acetylated by the *cysE*-encoded serine acetyltransferase. At the next stage, *O*-acetylserine reacts with hydrogen sulfide to form the cysteine. This reaction is catalyzed by the *cysK*-encoded cysteine synthase.

Cysteine is a donor of sulfur in taurine synthesis. In mycobacteria, genes *tauC* and *tauA* involved in the taurine transport system likely belong to the same operon, as their separating spacer is only about 10 nt long (with the exception of *Mycobacterium massiliense* str. GO 06).

Transcriptional repression of plastid genes *cysT* and *cysA* involved in sulfate transport was proposed in the Viridiplantae (Lyubetsky et al., 2013). In this case, a single-box repressor binding conserved motif with the consensus TAAWATGATT was found close to the promoters in many species of algae.

Both *cysK* and *cysE* genes usually belong to the same operon, which in *Mycobacterium avium* and *Mycobacterium leprae* also contains proteins with *unknown function*.

The ROK (Repressor, ORF, Kinase) domain PF00480 belongs to a diverse family of proteins that unites transcription factors (repressors *xylR* in *B. subtilis*, *Lactobacillus pentosus*, *Staphylococcus xylosus*, and *nagC* in *E. coli* that possess a helix-turn-helix DNA-binding motif absent from other members of the family), sugar kinases and uncharacterized proteins (Titgemeyer et al., 1994).

The Pfam domain PF14340 (DUF4395) of unknown function is frequently found across bacteria and eukaryotes. It possesses two conserved cysteic residues likely to be functionally important. The role of PF00480 and PF14340 proteins remained unclear.

Protein functions can be predicted on the basis of their regulation mechanisms. With this notion, we discuss the putative involvement of proteins in sulfur metabolism in Actinomycetales. To detect regulations, we performed a large-scale search for putative leader genes containing poly(Cys) regions in all currently available genomes of the Mycobacteriaceae (species of *Mycobacterium* and *Amycolicicoccus subflavus*) and Actinomycetales. Among many leader genes detected, the majority is associated with structural genes involved in sulfur metabolism, and some – with genes encoding less characterized proteins, including proteins with Pfam domains PF14340 and PF00480. The length of poly(Cys) tracts in leader genes reaches 8 (e.g., upstream of the protein YP_006570365.1[1] gene in *Mycobacterium smegmatis* str. MC2 155), and 4 upstream of the PF00480-domain protein genes.

[1] Hereafter, sequence/protein accession numbers refer to GenBank.

2. Methods

The search for leader genes was conducted with an original computer program, which conceptual description is provided below. The search was constrained to 5′-leader regions of genes annotated in GenBank (as per 2013) to detect open reading frames (ORFs) that usually do not overlap with the regulated gene (although they do overlap in some cases, Seliverstov et al., 2005, and so the overlap is allowed). If a leader gene was found in several mycobacterial species, orthologs of other structural genes in the Actinomycetales were included in the analysis.

The PF00480 and PF14340 domains were identified according to the Pfam database (Finn et al., 2010). Annotations were verified with PROSITE (Sigrist et al., 2010). Protein alignment was performed with Clustal 2.0.3 (Thompson et al., 1997). Phylogenetic trees were constructed with the MEGA5 software using the neighbor-joining algorithm (Tamura et al., 2011). Additional materials to this paper are available and referred to as Appendices hereafter.

Whenever a leader gene is found (ref. to Appendix 1), modeling of attenuation regulation was done to detect classic attenuation according to the earlier described method (Lyubetsky et al., 2007). The method description, the corresponding mathematical model and the computer program mentioned in the Background are freely available on the Web site (Model of RNA, 2014). All positive model predictions are given in Table 1 and further detailed in Section 3; negative predictions are exemplified in selected cases only.

In modeling, we used the average elongation rates of 45 nt/s for the ribosome, and 40 nt/s for the RNA polymerase (Lyubetsky et al., 2006, 2007; Rubanov and Lyubetsky, 2007). All U-rich regions were considered within the range of 100 nt from the leader gene stop codon. The maximum allowed length of the helix loop was 90 nt; the U-rich region parameters were as follows: minimum number of U entries – 3 nt, maximum distance between adjacent poly(U)s – 2 nt. For statistical confidence, the results were averaged over 1000 modeling trajectories.

2.1. The leader gene search algorithm

We proposed an algorithm to detect leader genes upstream structural genes. The algorithm recognizes a leader gene as an ORF with a predefined amino acid composition; this ORF must have a high local density of regulatory codons. The codon density is defined as follows. All codons in the current ORF are visited and change the counter value initially set to zero. If a codon encodes a regulatory amino acid, the counter increases by 1.0, and otherwise reduces by 0.5. The density of the entire ORF is its maximum counter value. The algorithm detects short ORFs (potential leader genes) with a high density of pre-defined regulatory codons upstream of annotated genes.

The algorithm parameters are listed below (used values shown in parentheses):

(1) minimum leader region length (100 nt), shorter intergenic regions are ignored;
(2) maximum leader region length (1400 nt), for longer intergenic regions only the 1400 nt range upstream from the gene start is considered;
(3) minimum structural gene length (200 nt), shorter genes are ignored;
(4) maximum structural gene length (10,000 nt), longer genes are ignored;
(5) minimum ORF length (five codons), shorter ORFs are ignored;
(6) minimum allowed density of regulatory codons for the entire ORF (3.0);
(7) the sets of regulatory amino acids, start and stop codons of the leader gene.

The algorithm starts with extracting non-coding leader regions on both DNA strands that meet the conditions (1)–(4). Each extracted leader region is searched for ORFs, and the density of specified regulatory amino acids is computed for such ORFs. The algorithm outputs ORFs (of candidate leader genes) that satisfy the specified parameters. If several leader genes with the same density are found for a structural gene, the shortest one is selected.

3. Results

The search for leader genes and Rho-dependent or classic attenuations was conducted for all annotated genes of mycobacteria (the genus Mycobacteriaceae) and their orthologs in other actinomycetes (the order Actinomycetales). The found regulations are grouped according to the gene products: cysteine synthases, thiosulfate sulfurtransferases, PF00480(ROK)-domain proteins, PF14340(DUF4395)-domain proteins, taurine transport, other genes. Negative predictions are generally not shown. The "other proteins" group contains miscellaneous homologous proteins, with the exception of the last paragraph. Phylogenetic analyses of the Rho protein and its binding sites are provided thereafter.

The discrimination between Rho-dependent and classic attenuations in our modeling is not straightforward. When one of the alternative regulations was to be selected, the well substantiated model of classic attenuation was applied. Classic attenuation was assumed if so predicted and if no suitable pyrimidine-rich region was found. Rho-dependent regulation was assumed otherwise.

3.1. Cysteine synthases

Many cysteine synthase associated leader genes are found in the actinomycetes listed in Appendix 1. The stop codon of the leader gene adjoins or overlaps with the putative Rho factor binding site, a degenerate repeat of a pyrimidine-rich region. In *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *M. africanum*, and *Mycobacterium canettii* the leader genes are separated from the cysteine synthase gene *cysK* by another gene of the same operon, which contributes to the earlier observation (Seliverstov et al., 2005). The *cysK1* gene in *M. tuberculosis* KZN 605 contains a frame shift, while the leader gene is preserved.

The modeling predicts possible classic attenuation for the cysteine synthase genes only in *Jonesia denitrificans* DSM 20603 and *Propionibacterium freudenreichii* subsp. *shermanii* CIRM-BIA1; in these cases only a slight increase of the transcription level is predicted for higher concentrations of cysteine (ref. to Table 1).

3.2. Thiosulfate sulfurtransferases

In *M. tuberculosis*, *M. bovis*, *M. africanum*, and *M. canettii* the leader genes with Cys codon tracts are located upstream the *sse* (*cysA2*) genes that encode the thiosulfate sulfurtransferase. A leader gene containing six consecutive Cys codons is also found upstream the thiosulfate sulfurtransferase gene in *M. massiliense* str. GO 06 (YP_006520168.1). The regulation mechanism is unclear: the Rho binding site is not detected, nor is the terminator hairpin with a neighboring poly(Cys) tract. Modeling also does not predict classic attenuation (ref. to Table 1).

Among actinomycetes, the leader genes upstream *cysA2* are predicted in the species listed in Appendix 1.

In most cases, modeling shows no classic attenuation, with the exception of *Gordonia polyisoprenivorans* VH2, which was predicted

Table 1
Modeling results for classic attenuation regulation. Data shown are percentage of transcription level against relative cysteine concentration. Rows with predicted regulations are indicated by up and down arrows for positive and negative dependencies, respectively. Abbreviation *P.* means *Propionibacterium*, *C.* – *Corynebacterium*, *M.* – *Mycobacterium*.

		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CysK												
Jonesia denitrificans DSM 20603	↗	48	49	51	52	57	59	59	61	63	64	63
P. freudenreichii ssp. shermanii CIRM-BIA1	↗	47	47	49	51	54	55	56	59	63	62	62
CysA2												
M. tuberculosis, *bovis*, *africanum*, *canettii*		89	82	85	85	86	85	81	82	78	78	77
Mycobacterium massiliense str. GO 06		34	35	32	32	33	33	30	28	27	26	27
Gordonia polyisoprenivorans VH2	↘	76	65	63	58	55	53	56	53	56	53	53
PF00480 (ROK)												
Mycobacterium sp. JLS, KMS, MCS	↗	55	51	51	52	58	58	62	65	65	68	73
Mycobacterium sp. MOTT36Y	↗	72	73	73	71	73	74	76	76	78	80	83
Mycobacterium chubuense	↗	41	41	45	47	50	50	53	55	54	52	56
Mycobacterium gilvum	↗	19	18	27	28	37	40	42	48	52	50	53
Mycobacterium massiliense	↗	74	76	77	81	80	84	83	84	87	87	87
Mycobacterium smegmatis	↗	61	75	80	79	82	86	88	89	90	91	92
Mycobacterium avium	↘	58	50	42	39	36	35	32	33	36	35	34
Mycobacterium vanbaalenii	↘	76	74	72	65	60	56	58	55	54	56	54
Mycobacterium abscessus		61	72	70	70	69	71	73	70	72	72	72
Gordonia bronchialis DSM 43247	↘	43	29	23	21	19	17	16	16	17	16	17
Gordonia polyisoprenivorans VH2	↘	67	56	51	45	43	43	42	42	44	45	43
Nocardia brasiliensis ATCC 700358	↗	51	51	54	58	61	62	63	64	65	66	69
PF14340 (DUF4395)												
Mycobacterium abscessus ATCC 19977		59	57	59	60	58	60	56	54	52	54	50
Mycobacterium avium		32	34	36	41	43	43	45	44	46	46	44
Mycobacterium intracellulare		35	37	39	44	44	48	49	50	49	52	52
Mycobacterium leprae		68	61	60	65	66	68	69	72	73	74	75
Mycobacterium rhodesiae NBB3		74	63	64	62	62	64	69	70	70	70	73
Mycobacterium sp. JDM601		46	42	35	33	31	29	29	28	25	26	27
Mycobacterium sp. MOTT36Y		32	34	36	41	43	43	45	44	46	46	44
Mycobacterium ulcerans Agy99		50	43	37	39	44	48	50	53	56	57	59
Mycobacterium vanbaalenii PYR-1	↘	83	75	72	68	72	71	74	75	74	72	71
Mycobacterium smegmatis	↘	83	73	66	65	64	67	65	67	68	69	68
Nocardia brasiliensis ATCC 700358	↗	35	31	40	41	46	47	48	49	52	56	57
Nocardia cyriacigeorgica GUH-2	↗	43	54	57	55	58	58	61	61	62	63	63
Nocardia farcinica IFM 10152	↗	60	71	73	77	76	82	82	86	84	86	89
P. freudenreichii ssp. shermanii CIRM-BIA1	↗	46	64	70	77	79	79	83	85	84	86	87
Rhodococcus equi 103S	↗	63	68	73	75	76	78	79	81	81	82	84
Rhodococcus erythropolis PR4	↗	36	37	40	48	56	57	61	64	67	68	72
Rhodococcus opacus B4	↗	46	49	54	53	59	62	62	67	67	70	70
Rhodococcus jostii RHA1		33	36	34	31	32	31	31	34	31	33	33
TauC												
C. pseudotuberculosis 1/06-A	↘	79	57	52	46	47	48	46	49	51	51	53
tetR family												
Gordonia bronchialis DSM 43247	↗	38	47	54	58	63	64	64	67	67	68	67
Arthrobacter phenanthrenivorans Sphe3	↗	18	28	30	35	40	44	49	53	55	57	57
Nocardia brasiliensis ATCC 700358	↗	36	51	55	60	60	60	62	63	63	64	63
C. variabile DSM 44702	↗	34	38	40	44	45	52	52	53	56	58	56
C. diphtheriae PW8, 31A, BH8, HC02		30	35	40	40	44	42	41	41	42	41	39
Rhodococcus equi 103S	↘	69	68	62	57	52	51	51	50	49	50	50
P. propionicum F0230a	↘	78	60	50	49	52	51	52	53	56	57	57

to slightly decrease the transcription level under elevated cysteine concentrations (Table 1). This decrease is well explained by the role of thiosulfate sulfurtransferase in cysteine synthesis. However, classic attenuation is not found in closely related *G. bronchialis* DSM 43247 and *G.* sp. KTR9.

3.3. PF00480(ROK)-domain protein

PF00480(ROK)-domain protein genes preceded by leader genes were found in the actinomycetes listed in Appendix 1. In the protein YP_006565684.1 from *M. smegmatis* str. MC2_155 the PF00480 domain is detected between residues 128–280 with the expect value of 8.3×10^{-10}.

The joint alignment of the 5′-leader region and partial PF00480-domain encoding sequences from mycobacteria *Mycobacterium gilvum*, *Mycobacterium rhodesiae*, *M. smegmatis*, *Mycobacterium vanbaalenii* (Fig. 1) demonstrates conservativity of the leader gene between its Cys and stop codons, a short region after the stop codon, and the region after the leader gene start codon that encodes a long helix in the transcript mRNA. Conversely, the sequence upstream of the Cys codons is little conserved. This observation well agrees with functional importance of the Cys codons and helix structures encoded downstream. The corresponding regions in *M.* sp. MOTT36Y and *M. avium* align well but possess a lower similarity with other mycobacteria.

The modeling detects classic attenuation in many species (ref. to Table 1). The transcription level increases with cysteine concentration in *M.* sp. JLS, KMS, MCS, MOTT36Y, *Mycobacterium chubuense*, *M. gilvum*, *M. massiliense*, *M. smegmatis*. The exceptions are *M. avium*, *M. vanbaalenii* that lack the Rho binding site and exhibit a monotonously decreasing dependency between the transcription level and concentrations of cysteinyl-tRNA; and *Mycobacterium abscessus*, *M. rhodesiae* that show no evident correlation. In the latter cases, the C-rich regions flanking the leader gene stop codon

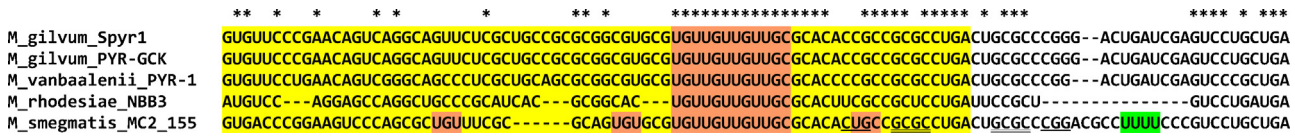

Fig. 1. Alignment of the leader gene and the putative termination region in *M. smegmatis* and close mycobacteria (1, *M. gilvum* Spyr1; 2, *M. gilvum* PYR-GCK; 3, *M. vanbaalenii* PYR-1; 4, *M. rhodesiae* NBB3; 5, *M. smegmatis* MC2_155). The start and stop codons are shown in lower case, regulatory (Cys) codons – in light gray, poly(U) – in dark gray, the termination region is located at the 5′-end of the poly(U). Classic attenuation is predicted in the lowest line, the terminator shoulders underlined. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

represent putative Rho binding sites, which suggests the presence of Rho-dependent attenuation.

In *M. smegmatis* and *M. massiliense* the intrinsic terminators are located in low conserved regions surrounded by higher conserved sequences. In these cases, a classic attenuation without antiterminator is expected. Indeed, specific modeling strongly suggests classic regulation in *M. smegmatis* (Fig. 2a), where the termination is almost not observed under high cysteine concentrations because the terminator hairpin is disintegrated by the ribosome. Under low cysteine contents the terminator almost always assembles, although the hairpin bulges reduce the termination frequency in the model. The transcription level against cysteine concentration is monotonously decreasing (Fig. 2a). Modeling suggests that termination can be triggered by three different hairpins assembling upstream the poly(U) region. The mRNA secondary structures formed between the ribosome and polymerase under zero cysteinyl-tRNA concentration trigger either termination (T) or continuation of transcription (A), as shown in Fig. 3 (Appendix 2). One of the terminator hairpins closest to the poly(U) is depicted in Fig. 4 (Appendix 2).

The same modeling was performed for *M. massiliense*. Mutual arrangement of the leader gene and the U-rich terminator region is shown in Fig. 5 (Appendix 2). In this case, the curve of the transcription level against cysteine concentration is also approximately monotonous, Fig. 2b. One of the putative terminator hairpins closest to the U-rich region is depicted in Fig. 6 (Appendix 2).

Among other actinomycetes, the modeled gene transcription level decreases with the increase of cysteine concentration in *G. bronchialis* DSM 43247 and *G. polyisoprenivorans* VH2 (similarly to mycobacteria), does not change in *Gordonia* sp. KTR9, *Nocardia cyriacigeorgica* GUH-2, *S. avermitilis* MA-4680, *T. paurometabola* DSM 20162, and slightly increases in *Nocardia brasiliensis* ATCC 700358 (Table 1). Classic attenuation regulation is predicted in these cases.

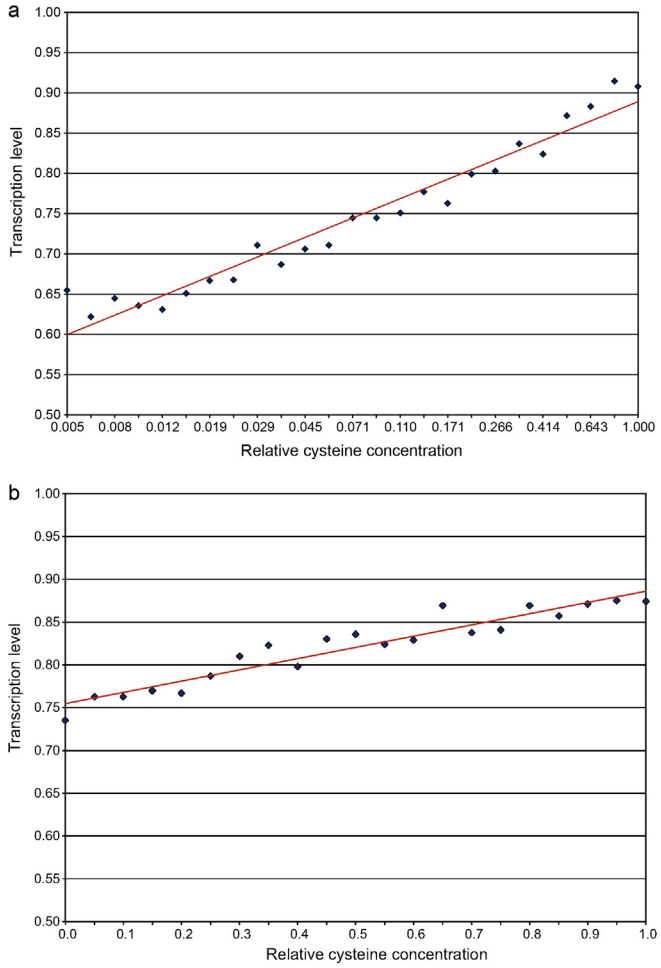

Fig. 2. Transcription level against cysteine concentration for *M. smegmatis* (a) and *M. massiliense* (b). Transcription level estimated over 1000 trajectories and shown as decimal fraction, not percentage. Data on the *X* axis in (a) is a geometric sequence with denominator 1.247.

3.4. PF14340(DUF4395)-domain protein

In mycobacteria, the PF14340(DUF4395)-domain protein genes are always associated with leader genes. Such associations are found in the actinomycetes listed in Appendix 1. In these cases, the leader gene stop codon is always preceded by a C-rich region, the putative Rho factor binding site on the transcript mRNA. In modeling classic attenuation, the transcription level is in most cases almost independent of cysteine concentration. This departure of mycobacteria from other actinomycetes may be explained by the existence of Rho-dependent regulation of these genes in mycobacteria.

The exceptions are *M. vanbaalenii* and *M. smegmatis*, where a decrease is observed. In *M. smegmatis* the predicted transcription level is 83% under zero cysteine content and drops to 64% under its higher concentrations (Table 1).

The transcription level is predicted to increase with cysteine concentration in *N. brasiliensis* ATCC 700358 (the increase in transcription level of the structural gene is 26%), *N. cyriacigeorgica* GUH-2 (20%), *Nocardia farcinica* IFM 10152 (29%), *P. freudenreichii* subsp. shermanii CIRM-BIA1 (41%), *Rhodococcus equi* 103S (21%), *Rhodococcus erythropolis* PR4 (36%), and *Rhodococcus opacus* B4 (24%). *Rhodococcus jostii* RHA1 showed no correlation (Table 1). Therefore, the classic attenuation is assumed.

3.5. Taurine transport

Proteins found to be associated in mycobacterial genomes (listed in Appendix 1) with leader genes possessing four consecutive Cys codons belong to the two TauC transport systems, the transport of taurine (TauC, TauA) and related compounds (taurocholate, sulfonates, sulfate esters).

In *M. abscessus* the Cys codons are distanced from the leader gene stop codon. In other species, on the contrary, the leader genes possess Cys codons in the 3′-terminus. In *M. rhodesiae* the region downstream the stop codon is enriched with uracil, in *M. smegmatis* – with cytosine. Other species do not exhibit similar enrichment patterns.

Leader genes and 5′-leader regions of mRNA in mycobacteria are depicted in Fig. 7 (Appendix 2). Proteins homologous to TauC are found in *M. gilvum* PYR-GCK (YP_001131997.1) and *M. vanbaalenii* PYR-1 (YP_950978.1), although these are not associated with leader genes containing poly(Cys) tracts.

In *M. massiliense* a leader gene with four consecutive Cys codons is associated with the TauA protein (YP_006520902.1), a putative taurine-binding periplasmic factor not homologous to TauC. Another 4-(Cys) containing leader gene is found upstream the transmembrane subunit of the ABC transporter (YP_004491913.1) in *A. subflavus* DQS3-9A1, which affinity to taurine is however unknown. Classic attenuation is eliminated due to the lack of intrinsic terminators. Rho-dependent attenuation may be hypothesized in these cases.

In actinomycetes, leader genes upstream *tauC* are predicted in the species listed in Appendix 1. Classic attenuation was predicted with confidence only for *Corynebacterium pseudotuberculosis* 1/06-A. The transcription level decreases with the increase of cysteine concentration (ref. to Table 1), which agrees well with a reduced role of transport when taurine is produced in cytoplasm under the excess of cysteine.

3.6. Other proteins

Leader genes with 5-(Cys) tracts are associated with the two proteins that are 138 residues long and differ by a single serine-to-asparagine substitution at position 86 in two mycobacterial species: *Mycobacterium marinum* M (YP_001851915.1), *Mycobacterium ulcerans* Agy99 (YP_905462.1). Hereafter, the notation contains the species and the strain, the protein accession number, and the maximum size of the poly(Cys) tract.

Leader genes associated with transposases possess 5 Cys codons in *M. ulcerans* Agy99 (YP_906498.1), 4 in *M. gilvum* Spyr1 (YP_004077019.1), 4 in *M. gilvum* Spyr1 (YP_004078569.1). In *M. ulcerans* this protein is a transposase required for the insertion of the mobile element IS2404.

Leader genes with several Cys codons are found upstream from genes of the tetR transcription factor family in the actinomycetes listed in Appendix 1. Modeling shows (Table 1) a pronounced increase of the transcription level with cysteine concentration in *G. bronchialis* DSM 43247, *Arthrobacter phenanthrenivorans* Sphe3, *N. brasiliensis* ATCC 700358, and *Callithamnion variabile* DSM 44702. In *R. equi* 103S and *Probionibacterium propionicum* F0230a a slight decrease is observed for the protein YP_006510480.1 gene. Classic attenuation is predicted in all species, except for *Corynebacterium diphtheriae* PW8, 31A, BH8, HC02.

In *M. massiliense* str. GO 06, the *ssb* gene encoding the protein YP_006522383.1 with a ss-DNA binding domain was found to be associated with the leader gene containing a 5-(Cys) tract. Rarely the leader genes are associated with *lipL* (esterase), *pstP* (serine/threonine phosphatase), *proA* (gamma-glutamyl phosphate reductase involved in proline synthesis), and some other less characterized genes. Their regulation mechanism remains unclear. The lack of the terminator hairpin excludes classic attenuation, which presence is also not shown in modeling.

3.7. Phylogeny of the Rho proteins and Rho binding sites on mRNA

The phylogenetic relationship of the Rho proteins in actinomycetes is shown in Appendix 3.

This transcription factor was known to be omnipresent in mycobacteria, with the exception of *M. africanum* GM041182 (NC_015758.1). Protein alignment of the Rho sequence from *M. bovis* and its predicted homolog from *M. africanum* shows highly conserved regions at the C-terminus (residues 217–602) and N-terminus (a shorter region between residues 1–114), ref. to Fig. 9 in Appendix 2. A 100 aa-long internal region of the *M. bovis* protein is lost in *M. africanum*. This observation suggests that the Rho protein of *M. africanum* modified closer to the N-terminus may still retain function, while its gene is considered a pseudogene in the current annotation. Alternatively, the Rho protein of *M. africanum* may consist of two different subunits translated independently in two different frames.

In *M. tuberculosis*, *M. bovis*, and *M. canettii* the Rho proteins are very similar, albeit being diverged from those in other mycobacteria, which is also evident from corresponding phylogenies. *M. leprae*, *M. ulcerans*, and *M. marinum* as well express diverged patterns.

The alignment and LOGO profiles of the 5′-untranslated mRNA regions close to the leader gene stop codon demonstrate their high conservativity, especially for *cysA2*, *cysK* and PF00480 (ROK) (Fig. 9 in Appendix 2). The sites depicted in Fig. 9 possess pyrimidine-rich regions, the putative Rho binding sites. This observation conforms well with our presumption that these sites play an important role in regulation, although their high divergence does not allow to precisely delimit the Rho binding sites upstream the corresponding genes.

4. Conclusion

Our analysis allowed to predict the involvement of proteins with the poorly characterized PF00480 (ROK) and PF14340 (DUF4395) domains, as well as other proteins listed above, in sulfur metabolism. The cysteine-dependent regulation was predicted for many genes. These results are obtained for all annotated genes of mycobacteria available in GenBank and their orthologs from other Actinomycetales.

The regulon contains subunits of the transporters of taurine and other sulfur-containing compounds, like taurocholate, sulfonates, sulfate esters. The presence of transposases in the regulons of *M. gilvum* and *M. ulcerans* may suggest a later acquisition of the leader genes and associated regulations in the result of horizontal gene transfers between mycobacteria; in *M. marinum* and *M. ulcerans* the regulated genes might have also been horizontally transferred.

This study also predicts regulation patterns for many genes. The obtained results considerably extend earlier large-scale studies (Lopatovskaia et al., 2010; Seliverstov et al., 2005) of classic and Rho-dependent attenuations mediated by amino acids other than cysteine.

We propose a roadmap to study structural genes (operons) preceded by leader peptide-coding genes ("leader genes"). In short, the leader gene is defined as an ORF with codons encoding certain pre-defined amino acids, which play a regulatory role with respect to this operon ("regulatory codons") and somehow relate to the operon's functionality. The connection of the leader gene with the operon is verified against certain requirements specified in Sections 2 and 3. In many cases (also in this work) the operon's function is not known a priori, and multiple amino acids are tried in turn for possible regulatory role. If the leader gene is present, the operon is regulated by a ribosome delaying at the regulatory codons during translation. This delay provides a mechanism to regulate transcription or/and translation. The regulation is mediated by secondary structures forming in the region between the leader gene and the operon, or protein factors usually binding mRNA near the first start codon of the operon. Such are cases of classic (Rho-independent) and Rho-dependent attenuation. Presence of the leader gene not only suggests the regulation mechanism but also the function of the operon. For example, the involvement of PF00480 or PF14340-domain proteins in sulfur metabolism was not clear before. Although only cysteine codons were considered in this work as regulatory, the method is equally applicable to other amino acids as well as other species. As an example of the latter,

the result of searching for leader genes with poly(Cys) tracts across all sequenced actinomycetes is provided in Appendix 4.

Acknowledgments

We thank Leonid Yu. Rusin for substantial contribution to the discussion and the manuscript preparation. We gratefully thank the anonymous reviewer for the profound and very useful review. This research is partly funded by the Ministry for Education and Science of Russia (grants 8481 and 14.740.11.1053), and the Russian Foundation for Basic Research (grant 13-04-40196-H).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.01.001.

References

Augustus, A.M., Spicer, L.D., 2011. The MetJ regulon in gammaproteobacteria determined by comparative genomics methods. BMC Genomics 12, 558.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., 2010. The Pfam protein families database. Nucl. Acids Res. 38, 211–222.

Grundy, F.J., Henkin, T.M., 1998. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. Mol. Microbiol. 30 (4), 737–749.

Heery, D.M., Dunican, L.K., 1993. Cloning of the trp gene cluster from a tryptophan-hyperproducing strain of *Corynebacterium glutamicum*: identification of a mutation in the trp leader sequence. Appl. Environ. Microbiol. 59, 791–799.

Hollands, K., Proshkin, S., Sklyarova, S., Epshtein, V., Mironov, A., Nudler, E., Groisman, E.A., 2012. Riboswitch control of Rho-dependent transcription termination. Proc. Natl. Acad. Sci. U. S. A. 109 (14), 5376–5381.

Kim, D.-E., Patel, S.S., 1999. The mechanism of ATP hydrolysis at the noncatalytic sites of the transcription termination factor Rho. J. Biol. Chem. 274 (46), 32667–32671.

Konan, K.V., Yanofsky, C., 2000. Rho-dependent transcription termination in the *tna* operon of *Escherichia coli*: roles of the boxA sequence and the rut site. J. Bacteriol. 182 (14), 3981–3988.

Kredich, N.M., 1992. The molecular basis for positive regulation of cys promoters in *Salmonella typhimurium* and *Escherichia coli*. Mol. Microbiol. 6 (19), 2747–2753.

Lin, C., Pradkar, A.S., Vining, L.C., 1998. Regulation of an antranilate synthase gene in *Stryptomyces venezuelae* by *trp* attenuator. Microbiology 144, 1971–1980.

Lopatovskaia, K.V., Seliverstov, A.V., Lyubetsky, V.A., 2010. Attenuation regulation of amino acid and amino acyl-tRNA biosynthetic operons in bacteria: comparative genomics analysis. Mol. Biol. (Mosk.) 44 (1), 140–151.

Lynch, A.S., Tyrrell, R., Smerdon, S.J., Briggs, G.S., Wilkinson, A.J., 1994. Characterization of the CysB protein of *Klebsiella aerogenes*: direct evidence that N-acetylserine rather than O-acetylserine serves as the inducer of the cysteine regulon. Biochem. J. 299 (1), 129–136.

Lyubetsky, V.A., Rubanov, L.I., Seliverstov, A.V., Pirogov, S.A., 2006. Model of genes expression regulation in bacteria by means of formation of secondary RNA structures. Mol. Biol. (Mosk.) 40 (3), 497–511.

Lyubetsky, V.A., Pirogov, S.A., Rubanov, L.I., Seliverstov, A.V., 2007. Modeling classic attenuation regulation of gene expression in bacteria. J. Bioinform. Comput. Biol. 5 (1), 155–180.

Lyubetsky, V., Seliverstov, A., Zverkov, O., 2013. Transcription regulation of plastid genes involved in sulfate transport in Viridiplantae. BioMed Res. Int. (Curr. Adv. Molec. Phylogenet.) 2013, 413450.

Model of RNA-related regulation in bacteria. http://lab6.iitp.ru/en/rnamodel/

Richardson, J.P., 2002. Rho-dependent termination and ATPases in transcript termination. Biochim. Biophys. Acta 1577, 251–260.

Rubanov, L., Lyubetsky, V., 2007. RNAmodel web server: modeling classic attenuation in bacteria. In Silico Biol. 7 (3), 285–308.

Saint-Girons, I., Duchange, N., Cohen, G.N., Zakin, M.M., 1984. Structure and autoregulation of the metJ regulatory gene in *Escherichia coli*. J. Biol. Chem. 259 (22), 14282–14285.

Seliverstov, A.V., Putzer, H., Gelfand, M.S., Lyubetsky, V.A., 2005. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. BMC Microbiol. 5 (54), 1–14.

Sigrist, C.J.A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N., 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucl. Acids Res. 38, 161–166.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molec. Biol. Evol. 28, 2731–2739.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids Res. 25, 4876–4882.

Titgemeyer, F., Reizer, J., Reizer, A., Saier Jr., M.H., 1994. Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. Microbiology 140 (9), 2349–2354.

Yanofsky, C., 1981. Attenuation in the control of expression of bacterial operons. Nature 289 (5800), 751–758.