ELSEVIER

FEMS Microbiology Letters 222 (2003) 211–220

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria

Ekaterina M. Panina [a], Alexey G. Vitreschak [b,c], Andrey A. Mironov [b,d], Mikhail S. Gelfand [b,d,*]

[a] *Graduate Program in Molecular, Cellular and Integrative Life Sciences, 172 Molecular Science Building, University of California at Los Angeles, Los Angeles, CA 90095-1570, USA*

[b] *Integrated Genomics, P.O. Box 348, Moscow 117333, Russia*

[c] *Institute for Problems of Information Transmission, Russian Academy of Science, Bolshoj Karetny per. 19, Moscow 101447, Russia*

[d] *State Scientific Centre GosNIIGenetica, 1-st Dorozhny pr. 1, Moscow 113545, Russia*

Received 26 February 2003; received in revised form 31 March 2003; accepted 31 March 2003

First published online 30 April 2003

Abstract

Computational comparative techniques were applied to analysis of the aromatic amino acid regulon in Gram-positive bacteria. A new candidate transcription regulation signal of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase and shikimate kinase genes was identified in *Streptococcus* and *Lactococcus* species. New T-boxes were found upstream of aromatic amino acid biosynthesis and transport genes in the *Bacillus*/*Clostridium* group. The substrate specificity of proteins from the PabA/TrpG family was assigned based on metabolic reconstruction and analysis of regulatory signals and phylogenetic patterns. New candidate tryptophan transporters were identified; their specificity was predicted by analysis of T-box regulatory sites. Comparison of all available genomes shows that regulation of genes of the aromatic amino acid biosynthesis pathway is quite labile and involves at least four regulatory systems, two at the DNA level and two more involving competition of alternative RNA secondary structures for transcription and/or translation regulation at the RNA level.

Keywords: Aromatic amino acid; Regulation; T-box; TRAP; ABC transporter

1. Introduction

Biosynthesis of aromatic amino acids is similar in Gram-positive and Gram-negative bacteria. It starts with the common pathway leading from phosphoenolpyruvate and erythrose 4-phosphate through 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) and shikimate to chorismic acid. Then the pathway divides into the terminal pathways, specific for each aromatic amino acid (Fig. 1, see [1] for a review). The only step that is performed by non-homologous enzymes in *Escherichia coli* and *Bacillus subtilis* is the first reaction of converting phosphoenolpyruvate and erythrose 4-phosphate to DAHP, which is catalyzed by three paralogous DAHP synthases in *E. coli* (AroF, AroG, AroH), and by a non-homologous DAHP synthase AroA in *B. subtilis*.

Although a significant number of aromatic amino acid transporters are known in Gram-negative bacteria, including AroP (general aromatic amino acid permease), Mtr (tryptophan transporter), TyrP (tyrosine transporter) and PheP (phenylalanine transporter) in *E. coli* [1], no transport proteins for aromatic amino acids were known in Gram-positive bacteria until recently a candidate tryptophan transporter YhaG was identified in *B. subtilis* [2]. The specificity of this transporter was established indirectly, by the discovery of *yhaG* regulation by TRAP, a Trp-dependent regulator of tryptophan biosynthesis in *B. subtilis* [2].

In Gram-negative bacteria, the biosynthesis of aromatic amino acids is regulated at both the DNA and RNA levels. TrpR and TyrR are two transcriptional repressors of this pathway known in *E. coli* [3,4]. At the RNA level, the *trpEDCBA* operon, encoding enzymes for the tryptophan terminal pathway, and the *pheA* gene, encoding choris-

* Corresponding author. Tel.: +7 (095) 3150156; Fax: +7 (095) 3150501.
E-mail address: gelfand@integratedgenomics.ru (M.S. Gelfand).

0378-1097/03/$22.00
doi:10.1016/S0378-1097(03)00303-3

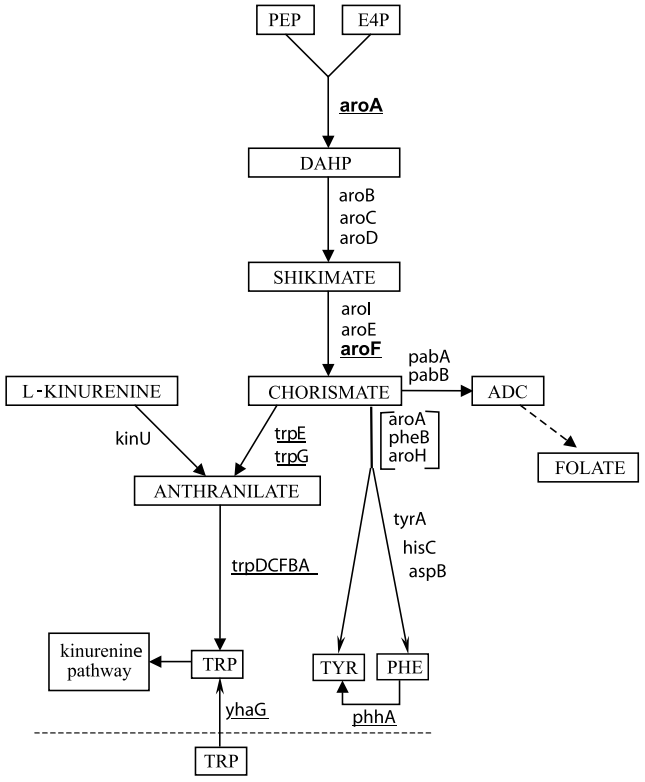

Fig. 1. Schematic representation of the aromatic amino acid biosynthesis and the folate biosynthesis pathways in the *Bacillus/Clostridium* group. The genes that encode enzymes for a single reaction are shown in square brackets. Known regulation in *B. subtilis* is shown: genes regulated at the DNA level are shown in bold; genes regulated at the RNA level are underlined. The dotted line represents the cell membrane. Abbreviations: PEP, phosphoenolpyruvate; DAHP, 3-deoxy-D-arabino-heptulosonate-7-phosphate; E4P, erythrose 4-phosphate; ADC, 4-amino-4-deoxy-chorismate.

mate mutase/prephenate dehydratase, are regulated by attenuation in *E. coli* [5–7]. In Gram-positive bacteria, no transcriptional regulation of aromatic amino acid biosynthesis has yet been experimentally discovered. However, phylogenetically conserved elements (PCEs) were recently identified upstream of *aroA* genes in *B. subtilis* (ACTTAAAAGCGTT) and *Bacillus halodurans* (ACTTAAAAGCGTc) and upstream of *aroF* genes in *B. subtilis* (ACTTAAAAGCGTT) and *Bacillus stearothermophilus* (ACTTAAAcGCGTT) [8]. These elements might play a role in the transcriptional regulation of aromatic amino acid biosynthesis in *Bacillus* species. The RNA regulation of this pathway in Gram-positive bacteria involves the *trp* RNA binding attenuation protein, TRAP [9], and the T-box antitermination mechanism [10]. In *B. subtilis*, TRAP plays the central role in controlling tryptophan metabolism [9]. It is responsible for binding to the mRNA leader or intercistronic regions in the presence of high levels of tryptophan. In the case of the *trp* operon, TRAP binding results in the termination conformation of the leader transcript, which leads to premature termination of transcription. In other cases TRAP is responsible for regulation of translation either by promoting formation of an RNA secondary structure that sequesters the Shine–Dalgarno (SD) sequence (*trpE*) or by binding to the RNA region overlapping the SD sequence (*trpG*, *ycbK*, *yhaG*). In *Lactococcus lactis*, the *trp* operon is regulated by the T-box antitermination mechanism, which is widely distributed in Gram-positive bacteria. T-box antitermination involves binding of uncharged tRNA to the RNA secondary structure (T-box) and promoting formation of antiterminator. The major role in regulation is played by the T-box 'specifier codon', which interacts with the anticodon of an uncharged tRNA. As the position of this regulatory codon in the T-box structure is fixed, one can predict the specificity of the regulatory signal [10].

Previously, we applied the comparative genomics approach to analysis of DNA and RNA level regulation of aromatic amino acid biosynthesis in γ-proteobacteria [11]. Here we apply the same approach to analysis of regulatory patterns involved in this pathway in Gram-positive bacteria including *Bacillus*, *Clostridium*, *Streptococcus*, *Enterococcus*, *Lactococcus*, *Staphylococcus*, *Listeria* and *Desulfitobacterium* spp. We describe the interchange of different types of DNA and RNA regulation of similar genes in various species. We suggest a new type of transcriptional regulation of DAHP synthase and shikimate kinase genes in *Streptococcus/Lactococcus* species and a new candidate tryptophan transporter in *Streptococcus*, *Lactococcus*, *Enterococcus* and *Desulfitobacterium* species.

2. Data and methods

2.1. Sequence data

Complete genome sequences of *B. subtilis*, *B. halodurans*, *Streptococcus pneumoniae*, *L. lactis*, *Streptococcus pyogenes*, *Clostridium acetobutylicum*, *Staphylococcus aureus*, and *Listeria monocytogenes* were downloaded from GenBank (http://www.ncbi.nlm.nih.gov). Partially sequenced genomes of *B. stearothermophilus*, *Streptococcus mutans*, *Enterococcus faecalis*, *Clostridium difficile* and *Desulfitobacterium hafniense* were extracted from the ERGO database (http://wit.mcs.anl.gov/WIT2/). The partially sequenced genome of *Enterococcus faecium* was obtained from the DOE Joint Genome Institute (http://www.jgi.doe.gov); the partially sequenced genome of *Bacillus anthracis* was obtained from The Institute for Genomic Research (http://www.tigr.org). The gene names in unfinished genomes were assigned based on the names of orthologous genes in related species.

FASTA sequences of all proteins with new or revised names are available from the authors.

2.2. Identification of regulatory signals

At the first step, groups of genes are formed that may contain a common regulatory element in their upstream

regions. Generally, either these genes belong to one genome and are expected to be co-regulated due to experimental data or their function in the common pathway, or they are orthologs in closely related species. Here we use a combination of these two approaches, compiling samples of functionally related genes represented by orthologs and paralogs in several genomes. The first sample was constructed based on the PCEs upstream of *aroA* genes in *B. subtilis* and *B. halodurans*, and upstream of *aroF* genes in *B. subtilis* and *B. stearothermophilus*, initially identified in [8]. For the second sample we extracted the upstream regions of two paralogous DAHP synthase genes from *L. lactis* (*aroF*, *aroH*) and the upstream regions of the operons encoding DAHP synthases from *S. pneumoniae* (*SP1701-SP1700*, referred here to as *aroG*1-*aroG*2) and *S. mutans* (*seqA-aroG*1-*aroG*2).

At the next step, a recognition rule was generated. If some regulatory sites had already been identified in experiment, a profile was constructed using the alignment of these known sites. If there were no known sites, an iterative procedure was performed in order to construct a profile. All L-mer words were selected in each upstream region. Each word was compared to all words in other regions, and one word, closest to the initial one, was selected in each region. These words were used to construct a profile. Thus we obtain as many profiles as there were words in the sample. Positional nucleotide weights in the profile were defined as:

$$W(b,k) = \log[N(b,k) + 0.5] - 0.25$$

$$\sum_{i=\mathrm{A,C,G,T}} \log[N(i,k) + 0.5]$$

where $N(b,k)$ is the count of nucleotide b at position k. The score of a candidate site was calculated as the sum of the respective positional nucleotide weights:

$$Z(b_1 \ldots b_L) = \sum_{k=1\ldots L} W(b_k, k)$$

The obtained profiles were used to scan the set of words again, and the procedure was iterated until convergence. Then the best profile was selected to be used as the recognition rule. The quality of a profile was defined as its information content [12]:

$$I = \sum_{k=1\ldots L} \sum_{i=\mathrm{A,C,G,T}} f(i,k) \log(f(i,k)/0.25)$$

where $f(i,k)$ is the frequency of nucleotide i at alignment position k.

Construction of a profile for an unknown signal requires specifying the site length L. In this study we considered $L = 10$, 11, 12, 13, 14, 15, 20, 21, 25, 30. The highest informational content per position was obtained for $L = 14$ (data for $L \neq 14$ are not shown).

Finally, the constructed profile was used to scan genomic sequences, which resulted in identification of candidate sites. A site upstream of a gene was accepted as a putative regulatory element if similar sites appeared upstream of orthologous genes in several analyzed genomes with significant Z scores.

The RNApattern program [13] was used to search for candidate T-boxes. The input pattern combined secondary structure and consensus sequence motifs. Each RNA secondary structure element was described by a set of the following parameters: the number of helices, the interval of accepted lengths for each helix, allowed loop lengths and description of the topology of helix pairs. The pattern for T-boxes was constructed using a training set of 10 known T-box structures from *B. subtilis*, *Lactococcus* spp. and *S. aureus*. [10,14,15]. The T-box pattern was used to scan each genome from the database. The pattern was absolutely specific: no candidate T-boxes were found upstream of genes not related to the amino acid functional systems and the T-boxes upstream of aromatic amino acid genes had appropriate specifier codons.

2.3. Software

Signal identification and construction of recognition profiles was performed using the SIGNALX program [16]. T-box RNA structures were found by the RNApattern program [13]. Genomic analyses (protein similarity searches using the Smith–Waterman algorithm, analysis of orthology, identification of candidate sites in genomic sequences) were done using GenomeExplorer [16]. Multiple protein alignments were constructed using CLUSTAL [17]. Phylogenetic trees were constructed using the PHYLIP package programs [18]. Transmembrane segments in proteins were predicted by TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html).

3. Results

3.1. The pathway: genes and operons

The backbone of the aromatic amino acid biosynthesis pathway is conserved in most bacterial species, but some steps vary within the analyzed group. First, the complete genomes of *S. pyogenes* and *E. faecalis* lack the genes for the terminal tryptophan pathway. Second, in the *S. pyogenes* genome there are no homologs of the *B. subtilis* genes *pheA* and *tyrA* from the terminal phenylalanine and tyrosine pathways, respectively. Third, in the genomes of *S. pneumoniae*, *S. mutans* and *L. lactis* there are no homologs of the *B. subtilis* gene *aroA* whose product catalyzes the first step of the common pathway (DAHP synthesis). However, in each of these three genomes there are two genes homologous to DAHP synthases from Gram-negative bacteria (Table 1). Finally, in *B. anthracis* and *D. hafniense* there is a homolog of the *phhA* gene, whose product phenylalanine 4-hydroxylase catalyzes conversion of phenylalanine to tyrosine.

The predicted operon structure of aromatic amino acid

Table 1
Schematic representation of candidate operons of aromatic amino acid biosynthesis genes in the *Bacillus/Clostridium* group

Bacterium	Function	Candidate operons and regulation
Bacillus subtilis	B	(PCE) *aroFBH*–(TRAP) *trpECFBA-hisC-tyrA-aroE*; (PCE) *aroA*; (TRAP) *pabA**; *pheBA*; *aroI*; *aspB*; *aroC*; *aroD*
	T	(TRAP) *yhaG*; (T-Trp) yczA-(TRAP)ycbK
Bacillus halodurans	B	(**PCE**) *aroFBH*–(**TRAP**) *trpEDCFBA-hisC-tyrA-aroE* (PCE)*aroA*; (**TRAP**) *pabA*; *pheBA*; *aroI*; *aspB*; *aroD*
	T	–(?)
Bacillus stearothermophilus	B	(**PCE**) *aroFBH*–(**TRAP**) *trpE-trpD-trpC-trpF-trpB-trpA-hisC-tyrA-aroE*; *pheBA*; (**TRAP**) *pabA*; *aroI*; *aspB*; *aroC*; *aroD*; *aroA*
	T	*yhaG*; *trpXYZ*
Bacillus anthracis (*cereus*)	B	(**T-tyr**) *aroA*–(**T-tyr**) *aroF2-hisC2-tyrA-aroE*; *aroF1-aroB-hisC1*; (**T-trp**) (**T-trp**) *trpE-pabA2-trpD-trpC-trpF-trpB-trpA*; (**T-tyr**) (**T-tyr**) *phhA*; *pabA1*; *aspB*; *aroD*; *pheA*
	T	(**T-tyr**) *yheL(ZC)*; (**T-trp**) sdt1
Streptococcus pneumoniae	B	*ywbD-aroC-aroD-aroB-aroF-tyrA-yheA–aroE*-(**ARO**)-*aroI-pheA-psr*; (**ARO**) *aroG1*-aroG2**; (**T-trp**) *trpE-pabA-trpD-trpC-trpF-trpB-trpA*; *aspB*
	T	(**T-trp**) *trpXYZ*
Streptococcus mutans	B	*ywbD-aroC-aroD-aroB-aroF-tyrA-yheA–aroE*-(**ARO**)-*aroI-pheA-psr* (**ARO**)-*seqA-aroG1*-aroG2**; (**T**) *trpE-trpF-trpD-trpC-pabA-trpB-trpA*; *aspB*; *hisC*
	T	*trpXYZ*
Lactococcus lactis	B	*aroC*; *aroD-aroB*; *aroF*; *tyrA*; *aroE*-(**ARO**)-*aroI-pheA*; (**ARO**) *aroG**; (**ARO**) *aroG** (**T-box**) *trpE-pabA2-trpD-trpC-trpF-trpB-trpAaspB*; *hisC*; *aroF*; *pabA1*
	T	–(?)
Streptococcus pyogenes	B	*ywbD-aroC-aroF*; *aroE-aroI*; *aroD-ycgJ-*>< *-aroB-aroA*; *aspB*; *pabA*
	T	trpX-geneX-trpYZ
Enterococcus faecalis	B	*aroD-aroA-aroB-aroF-tyrA-aroE-aroI-pheA–psr*; *aspB*; *aroC*
	T	(**T-tyr**) *yheL*; *trpXYZ*
Clostridium acetobutylicum	B	*aroA1-tyrA-aroB-aroE-aroF-aroD-aroI-yqhS*; (**T-trp**) *trpE-pabA-trpD-trpC-trpF-trpB-trpA*; *aspB*; *hisC*; *pheA*; *aroA2*; *pheB*
	T	(**T-trp**) *yhaG*
Clostridium difficile	B	*aroA1–aroB-aroE-aroF-pheA-aroD1-aroI–tyrA aspB*; *pabA*; *aroC*; *hisC*; *aroA2*; *aroD2*
	T	(**T-trp**) *yhaG*
Staphylococcus aureus		*aroF-aroB-aroE-tyrA-aroB-aroE-aroF-aroD-aroI-yqhS*; (**T-trp**) *trpE-pabA-trpD-trpC-trpF-trpB-trpA-*>< *-tyrA hisC*; *pheA*; *pabA2*; *aroI*; *pheB*; *aroC*; *aroA*; *aroD*; *tyrA*
	T	–(?)
Listeria monocytogenes	B	*aroF-aroB-aroH-hisC-tyrA-aroE*; (**T-trp**)(**T-trp**) *trpE-pabA-trpD-trpC-trpF-trpB-trpA*; *aroD-aroC*; *pheA*; *aroI*; *aroA*
	T	–(?)
Desulfitobacterium hafniense	B	*aroA1-tyrA-aroE*; *aroL*-aroB*; (**T-trp**) *trpC-trpF-trpD-trpB-trpA-trpE-pabA*; *aroI*; *hisC*; *aroD*; (**T-phe**) *pheA*; *aroA2*; *aroG**; *phhA*
	T	(**T-trp**) trpX1[; (**T-trp**) trpX2Y2[

Genes forming one candidate operon (with spacer less than 100 bp) are separated by dashes '-'. Larger spacers between genes are marked by '–'. Operons from different loci, if shown in one column, are separated by a semicolon ';'. Known and predicted regulatory elements (PCE, TRAP, T-trp, T-tyr and T-phe) are shown in light and bold fonts, respectively. The latter three regulatory elements stand for tryptophan-, tyrosine- and phenylalanine-specific T-boxes, respectively. ARO stand for a new DNA regulatory signal (Table 2). Genes with closest homologs in γ-proteobacteria rather than in the *Bacillus/Clostridium* group are marked by asterisks (*). Paralogs are numbered (e.g. *aroA*1, *aroA*2) by decreasing similarity to the respective *B. subtilis* genes. Gene names in all genomes are given based on orthologous genes in *B. subtilis*, with the exception of genes marked by the asterisks, which are named after respective *E. coli* genes. Designations in the 'Function' column are as follows: 'B' – biosynthetic genes, 'T' – transporter genes. Transporter genes are underlined.

biosynthesis genes varies significantly within the studied group of genomes (Table 1). The only conserved feature is the *trp* operon encoding enzymes for the tryptophan terminal pathway, which is absent or present as a whole in each genome. In *B. subtilis*, *B. halodurans* and *B. stearothermophilus*, this candidate operon is a part of a larger locus containing other *aro* genes: *aroF-aroB-aroH-trpEDCFBA-tyrA-hisC-aroE*. The operon structure in *B. anthracis* differs from that in other bacilli: the *trp* operon lies separately, and the remaining genes form two more loci: *aroF*1-*aroB-hisC*1 and *aroA-aroF*2-*hisC*2-*tyrA-aroE*, where *aroF*1,2 and *hisC*1,2 denote pairs of paralogs (*aroF*2 and *hisC*2 display less identity to *aroF* and *hisC* from *B. subtilis* than *aroF*1 and *hisC*1 do, respectively). In *S. pneumoniae*, *S. mutans* and *L. lactis*, the *trp* operon is also isolated. In *S. pneumoniae* and *S. mutans*, there is one more large locus with *aro* genes: *aroC-aroD-aroB-aroF-tyrA-yheA-aroE-aroI-pheA* (the gene names are as in *B. subtilis*). In *L. lactis*, several outsider genes are inserted into this gene cluster. Pairs of DAHP synthase genes in *S. pneumoniae*, *S. mutans* and *L. lactis*, homologous to DAHP synthases from Gram-negative rather than Gram-positive bacteria, form candidate operons in *S. pneumoniae* and *S. mutans*, but are located separately in the genome of *L. lactis*. Large *aro* gene loci are also present in the genomes of *E. faecalis* (*aroD-aroA-aroB-aroF-tyrA-aroE-*

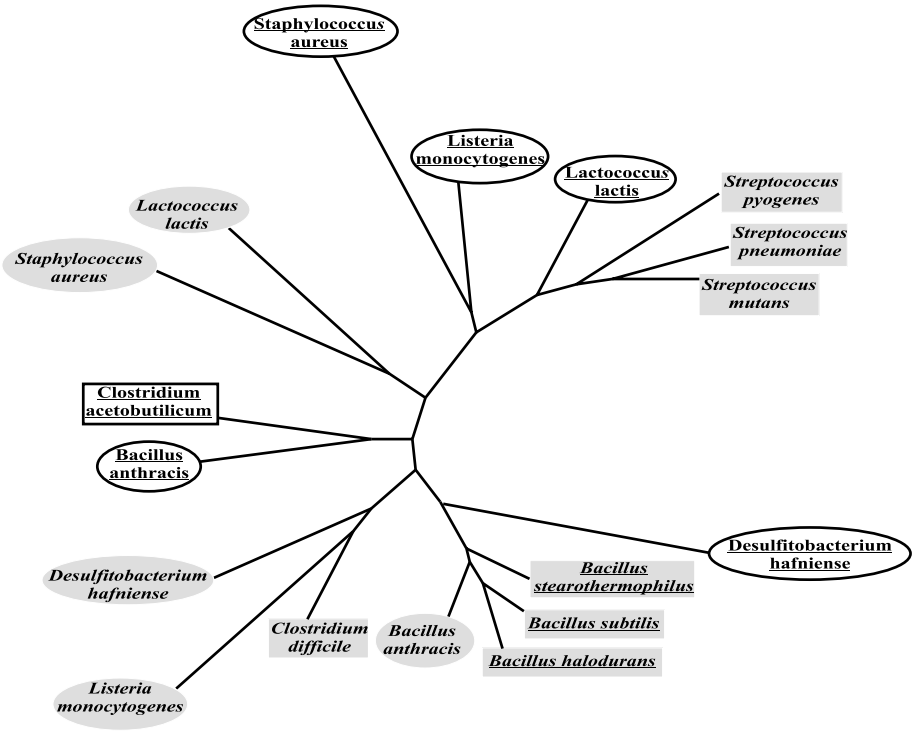

Fig. 2. Phylogenetic tree of the PabA/TrpG family proteins. Squares: one paralog per genome; ovals: two paralogs per genome. Filled frames: genes located within the *pab* operons; empty frames: genes located within the *trp* operons. Italics: folate-specific enzymes; underlined: tryptophan-specific enzymes (the specificity is identified in this study); italics and underlined: bi-functional enzymes.

aroI-pheA), *C. acetobutylicum* (*aroA*1-*tyrA-aroB-aroE-aroF-aroD-aroI*), *C. difficile* (*aroA*1-*aroB-aroE-aroF-pheA-aroD*1-*aroI-tyrA*), and *L. monocytogenes* (*aroF-aroB-aroH-hisC-tyrA-aroE*).

Two genes were found in aromatic amino acid operons in *Streptococcus* spp.: *ywbD* (*SP1378*) and *psr* (*SP1368*) (Table 1). As they co-localize with aromatic amino acid genes in several genomes, their function may be somehow linked to the aromatic amino acid metabolism.

3.2. TrpG/PabA: assignment of specificity

Anthranylate synthase component II (TrpG), which catalyzes conversion of chorismate into anthranylate (the tryptophan terminal pathway), and *p*-aminobenzoate synthase component I (PabA), which catalyzes conversion of chorismate into 4-amino-4-deoxy-chorismate (the folate biosynthesis pathway) [19], are encoded either by two paralogous genes, as in *E. coli*, or by one bi-functional gene, as in *B. subtilis*. The orthology relationships between members of the TrpG/PabA family can hardly be resolved by protein similarity analysis alone. The scheme of the folate biosynthesis and the step catalyzed by PabA are shown in Fig. 1.

We analyzed the functional specificity of the TrpG/PabA family members in the *Bacillus/Clostridium* group using positional analysis. We found that *E. faecalis* lacks members of this family; the *Bacillus*, *Streptococcus*, and *Clostridium* genomes, excluding only *B. anthracis*, each contain one copy of the *trpG/pabA* genes, whereas *L. lactis*, *L. monocytogenes*, *S. aureus*, *B. anthracis*, and *D. hafniense* each have two paralogous genes. Moreover, out of two paralogs, one always lies within the *trp* operon, while the other co-localizes with the *pab* operon (Fig. 2). This suggests that the first class of paralogous genes (*trpG* (LL), *trpG* (LM), *trpG* (SA), *trpG* (BQ, DH), see appendix) is specific for tryptophan biosynthesis, whereas the second class of paralogs (*pabA* (LL), *SAV0700*, *lmo2749*, *pabA* (DH)) is specific for folate biosynthesis.

A phylogenetic tree of all members of the TrpG/PabA family from the analyzed genomes was constructed (Fig. 2). We found that the single member of this family from *C. acetobutylicum* is localized in the *trp* operon in the genome and clustered with the tryptophan-specific paralog from *B. anthracis* in the tree. Besides, *C. acetobutylicum* lacks other *pab* genes. Thus, we propose tryptophan-related specificity rather than bi-functionality for this single protein. A similar situation was observed in *C. difficile*: it has a single gene, positioned in the *pab* operon, and clustered with folate-specific paralogs in the tree. Thus, we propose folate specificity for the protein from *C. difficile*. The unfinished genome of *C. difficile* lacks the *trp* operon. We expect that if this operon exists in the unsequenced portion of the genome, there should be one more, tryptophan-specific member of the TrpG/PabA family.

The complete genome of *S. pyogenes* lacks the tryptophan terminal pathway, and it has only one representative of the TrpG/PabA family, which is positioned within the

pab operon. Thus we suggest that it is folate-specific. In contrast, the complete genome of *S. pneumoniae* and the partial genome of *S. mutans* lack *pab* operons, and they have one representative of the TrpG/PabA family each, both positioned in the *trp* operons. Thus we suggest that they are tryptophan-specific.

3.3. DNA level regulation: new candidate signals

Pairs of DAHP synthase genes of *S. pneumoniae*, *S. mutans* and *L. lactis*, encoding proteins homologous to DAHP synthases (AroA) from Gram-negative bacteria, form gene clusters in *S. pneumoniae* and *S. mutans*, but are located separately in *L. lactis*. We found a conserved 14-bp sequence ATGGAGGCANATAA upstream of the DAHP synthase operons in *S. pneumoniae* and *S. mutans*, and upstream of both DAHP synthase genes in *L. lactis*. Moreover, a similar sequence was found in the upstream regions of the shikimate kinase genes (*aroI*) in all three species (Table 2). Notably, the reactions catalyzed by shikimate kinase and DAHP synthase are the only two irreversible steps within the common pathway of the biosynthesis of aromatic amino acids, and only these genes of the common pathway are regulated at the transcriptional level in γ-proteobacteria. Thus, we propose that the new conserved signal sequence plays a role in transcriptional regulation of the DAHP synthase and shikimate kinase genes in the genomes of *Streptococcus* and *L. lactis*.

We also constructed a profile based on the PCEs described in [8]. Using this profile we found a new candidate site ACTTAAccaCGTT upstream of the *aroF* gene in *B. halodurans*.

3.4. RNA level regulation

A number of new candidate T-boxes were found upstream of genes involved in aromatic amino acid biosynthesis (Fig. 3). Expression of the *aroF* and *aroA* genes is predicted to be regulated at the DNA level in *B. subtilis*, *B. halodurans*, and *B. stearothermophilus* (see above). In contrast, tyrosine-specific T-boxes were found upstream of these genes in *B. anthracis* (Table 1). The *aroA-aroF-hisC-tyrA-aroE* locus in *B. anthracis* appears to be strictly regulated by the T-box antitermination mechanism, as two possible tyrosine-specific T-boxes are located upstream of the *aroA* gene and one more tyrosine-specific T-box is located upstream of the *aroF* gene. The *aroF* gene in *E. coli* is known to be regulated by the tyrosine-specific repressor TyrR [4]. Finally, a tyrosine-specific T-box was observed in *B. anthracis* upstream of the *phhA* gene. PhhA catalyzes conversion of phenylalanine to tyrosine (Fig. 1) and the *phhA* gene is possibly regulated by tyrosine via the T-box antitermination mechanism in this bacterium.

The *trp* operons are known to be regulated at the RNA level by two different mechanisms, TRAP-mediated repression in *B. subtilis* [9] and T-box antitermination in *L. lactis* [14]. Additional candidate TRAP binding sites were found upstream of the *trp* operons and the *trpG* genes in *B. halodurans* and *B. stearothermophilus* (Fig. 4). Tryptophan-specific T-boxes were found upstream of the *trp* operons in *B. anthracis*, *S. pneumoniae*, *S. mutans*, *L. lactis*, *C. acetobutylicum*, *S. aureus*, and *L. monocytogenes*. Thus, TRAP-mediated regulation was observed only in three *Bacillus* species, *B. subtilis* [9], *B. halodurans*, and *B. stearothermophilus* (this study), whereas in other Gram-positive bacteria, including *B. anthracis*, only the T-box antitermination mechanism was detected. Moreover, the *mtrB* gene, which encodes subunits of TRAP, is present only in these three *Bacillus* genomes. Interestingly, *B. anthracis* has at least twice as many T-boxes as other Gram-positive bacteria (A. Vitreschak, unpublished). We also observed a phenylalanine-specific T-box site upstream of the *pheA* gene in *D. hafniense*.

3.5. New candidate transporters of aromatic amino acids

The only known tryptophan transporter in the *Bacillus/Clostridium* group is YhaG of *B. subtilis*, whose translation is regulated by the TRAP protein. We found orthologs of the *yhaG* gene in *B. stearothermophilus*, *C. acetobutylicum* and *C. difficile*. No homologs of *yhaG* were observed in the genomes of *E. faecalis* and *S. pyogenes* that lack the tryptophan biosynthesis pathway, and thus should transport tryptophan from the environment. We identified tryptophan-specific T-boxes upstream of the *yhaG* orthologs in both *Clostridium* species; in *B. stearothermophilus* the upstream region of this gene is unavailable. Thus, *yhaG* is regulated in *B. subtilis* and *Clostridium*

Table 2
A new DNA signal regulating DAHP synthase and shikimate kinase genes of several species identified in this study

Genome	Gene/operon	Candidate site	Position
L. lactis	*aroG1*	AaGGAGGCAcATAA	−93
L. lactis	*aroG2*	ATGGAcGCAaATAA	−80
L. lactis	*aroI*	ATGGGGGCcaAAAT	−256
S. mutans	*secA-aroG-aroG*	ATGGGGGCAGAAAA	−113
S. mutans	*aroI*	ATGGGGGCtaAgAT	26
S. pneumoniae	*aroG-aroG*	tTaGAGGCgGATAT	−67
S. pneumoniae	*aroI*	ATGGGaGCAGATAT	−216

Position is given relative to the translation start site. Paralogous *aroG* genes are numbered for convenience.

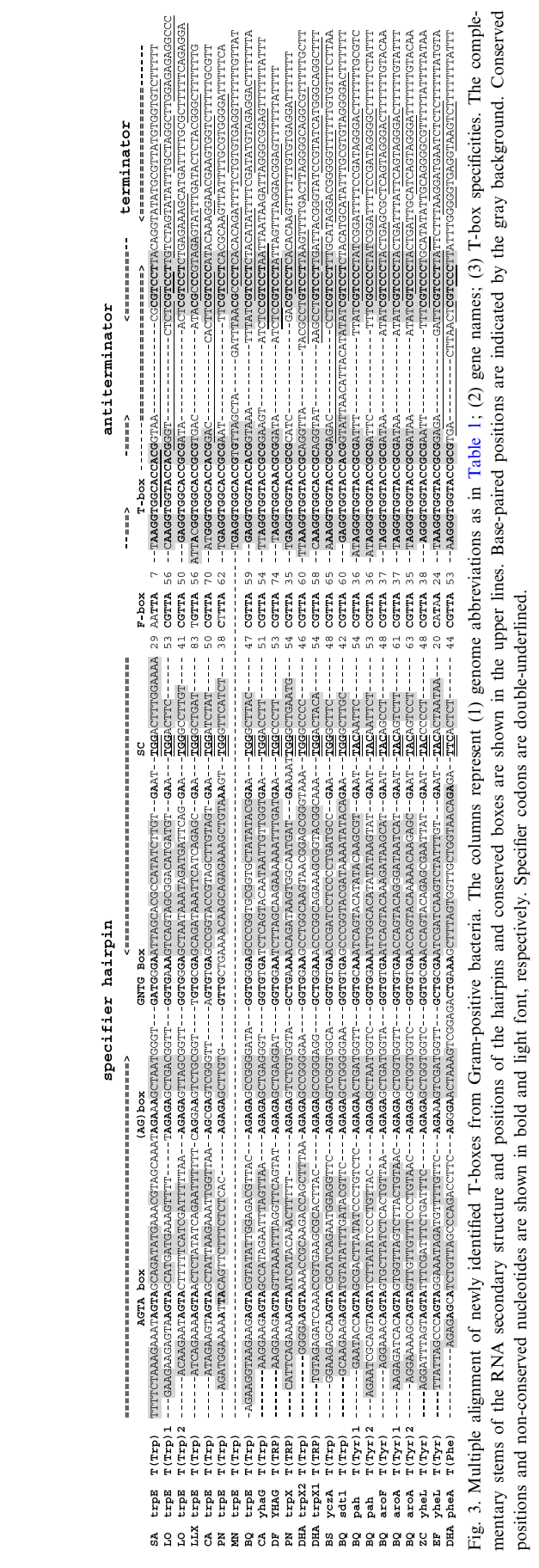

Fig. 3. Multiple alignment of newly identified T-boxes from Gram-positive bacteria. The columns represent (1) genome abbreviations as in Table 1; (2) gene names; (3) T-box specificities. The complementary stems of the RNA secondary structure and positions of the hairpins and conserved boxes are shown in the upper lines. Base-paired positions are indicated by the gray background. Conserved positions and non-conserved nucleotides are shown in bold and light font, respectively. Specifier codons are double-underlined.

at the RNA level by two different mechanisms, tryptophan-mediated TRAP repression and tryptophan-specific T-box antitermination, respectively.

Analyzing the predicted T-box regulatory sites and positional gene clustering we identified a new candidate tryptophan ABC transporter, named *trpXYZ*, in the genomes of *S. pneumoniae*, *S. mutans*, *S. pyogenes*, *S. equi*, *E. faecalis*, *E. faecium*, *B. stearothermophilus*, *D. hafniense*, *B. cepacia*, and *M. loti* (the latter two are α-proteobacteria). The genes in the *S. pneumoniae* genome are *SP1069*, *SP1070*, *SP1071*. Fig. 5 shows the phylogenetic tree of the substrate binding components of this transporter from all available genomes. *D. hafniense* has three *trpXYZ* paralogs, and two of them have tryptophan-specific T-boxes in the upstream regions. Additionally, *trpXYZ* is preceded by a tryptophan-specific T-box in *S. pneumoniae*. Moreover, *trpXYZ* is located in one candidate operon with the ortholog of the *kynU* gene in *M. loti*. *kynU* encodes L-kynurenine hydrolase, which catalyzes conversion of L-kynurenine into anthranylate (Fig. 1). Thus co-induction of the *trpXYZ-kynU* operon in tryptophan-depleted conditions leads to the transport of tryptophan from the medium and the concurrent accumulation of anthranylate, a tryptophan biosynthetic precursor. Additionally, *trpXYZ* is co-localized with the *aroD* gene in *E. faecium*. These pieces of evidence allow us to ascribe tryptophan specificity to all but one major clades of the *trpXYZ* family members on the phylogenetic tree (Fig. 5). Note that this assignment fills the above-mentioned gap in the *E. faecalis* and *S. pyogenes* metabolic maps.

We identified another candidate tryptophan transporter in *B. anthracis*. Four members of the sodium transporter family (homologous to *yocR* and *yhdH* in *B. subtilis*) are present in this bacterium and two of them are regulated by the T-box antitermination mechanism. We assigned specificities based on the T-box regulatory elements. We predict that one of these genes, named here *sdt1*, encodes a tryptophan-specific transporter, and the other gene, *sdt2*, is serine-specific. Homologous genes of this transporter family were identified in the genomes of *B. subtilis*, *B. halodurans*, *B. anthracis*, *S. aureus*, *L. monocytogenes* and *S. pneumoniae*. Additionally, a homologous transporter in *Haemophilus ducrei* forms an operon with genes of the tryptophan biosynthesis.

The *ycz-ycbK* operon of *B. subtilis* is known to be regulated by TRAP-mediated repression and tryptophan-specific T-box antitermination [20]. A TRAP site and a tryptophan-specific T-box are located in the intergenic region of the *yczA-ycbK* operon and in the leader region of the *yczA* gene respectively. The *yczA* gene is known to encode the anti-TRAP protein, an inhibitor of TRAP activity [20], but the function of YcbK is unknown. As this protein has a predicted 10 transmembrane segments and the gene *ycbK* is likely regulated by tryptophan, YcbK can be involved in transport of tryptophan or tryptophan-related compounds.

```
BS  pabA GAGCATTAGAGCTGAGCG-AAGAAGAGACAAAAATTAG-ATGAGGTGAGCG-GAGAAATGATT
BE* pabA AAGCGAAAGAGCTGAGCG-AAGCAGAGGCATTATTTCCGAGCATGAGGTGAGAATGATGATC
HD* pabA GAGTATAGACGAGCAAAAGCAAAGAATAGA-AAAGTAGAGC-TGAGGAGGAATCACGATG

BS  trpE AAGCAATTAGAATGAGTTGAGTTAGAGAATAGGGTAGCAGAGAATGAGTTTAGTTGAGCTGAG
BE* trpE AAGTGGAGCGAGAGTGGAGAGCGAGCGTAGGGTAGATGAGAATGAGC-GAGTTTAGCTGAGGTTGAG
HD* trpE TAGTAAAGCTTAGTTTACCAGTTTAGTTGAGATGAGAATGAAGAGTTGAGGAGAG
```

Fig. 4. TRAP binding sites in the leader regions of the *trp* and *pabA* genes in the *Bacillus* spp. The (G/T/A)AG repeats, which are recognized by the TRAP repressor, are highlighted by the gray background. Start codons of the *pabA* genes are underlined. Newly identified TRAP sites are indicated by asterisks.

4. Discussion

4.1. Evolution of the PabA/TrpG protein family

Enzymes of the PabA/TrpG protein family are present in one or two copies per genome. The single-copy proteins in the *Bacillus* spp. are bi-functional. One could suggest that originally there existed a bi-functional protein, which had been independently duplicated in *B. anthracis*, *L. lactis*, *L. monocytogenes*, *S. aureus*, and *D. hafniense*. However, our analysis demonstrates that the only copy of the enzyme in the genomes of *C. acetobutylicum*, *S. pyogenes*, and *S. pneumoniae* is likely mono-functional (either folate- or tryptophan-specific). Thus, the assumption of universally distributed bi-functionality would require too many independent duplication and loss-of-function events. The most parsimonious scenario of the evolution of the PabA/TrpG family seems to be as follows. Initially there were two enzymes of the TrpG/PabA family, one belonging to the tryptophan biosynthesis pathway, and the other belonging to the folate pathway. The corresponding genes were co-localized with other genes of the respective pathways. This situation is still conserved in the genomes of *B. anthracis*, *L. lactis*, *L. monocytogenes*, *S. aureus*, and *D. hafniense*. However, some species have eliminated one of the pathways (e.g. *C. acetobutylicum*, *S. pyogenes*, *S. pneumoniae*) as well as the corresponding TrpG/PabA protein. *E. faecalis* has eliminated both pathways, and accordingly, it has no representatives of this family. At the time of the *B. subtilis-B. halodurans* branch formation, the *trpG* gene was lost from the *trp* operon, and the remaining PabA protein encoded in the *pab* operon acquired additional function in tryptophan biosynthesis, thus becoming bi-functional. Simultaneously, the *pabA* gene acquired the TRAP-dependent regulation of translation. In these genomes *pabA* is the second gene in the *pab* operon, thus the tryptophan-dependent regulation is 'wedged' inside the folate biosynthesis operon. We propose that TRAP bound to the upstream region of *pabA* blocks de novo initiation of *pabA* translation, whereas it does not prevent the ribosomal re-initiation at the *pabA* start codon, as the ribosome coming from the upstream *pabB*

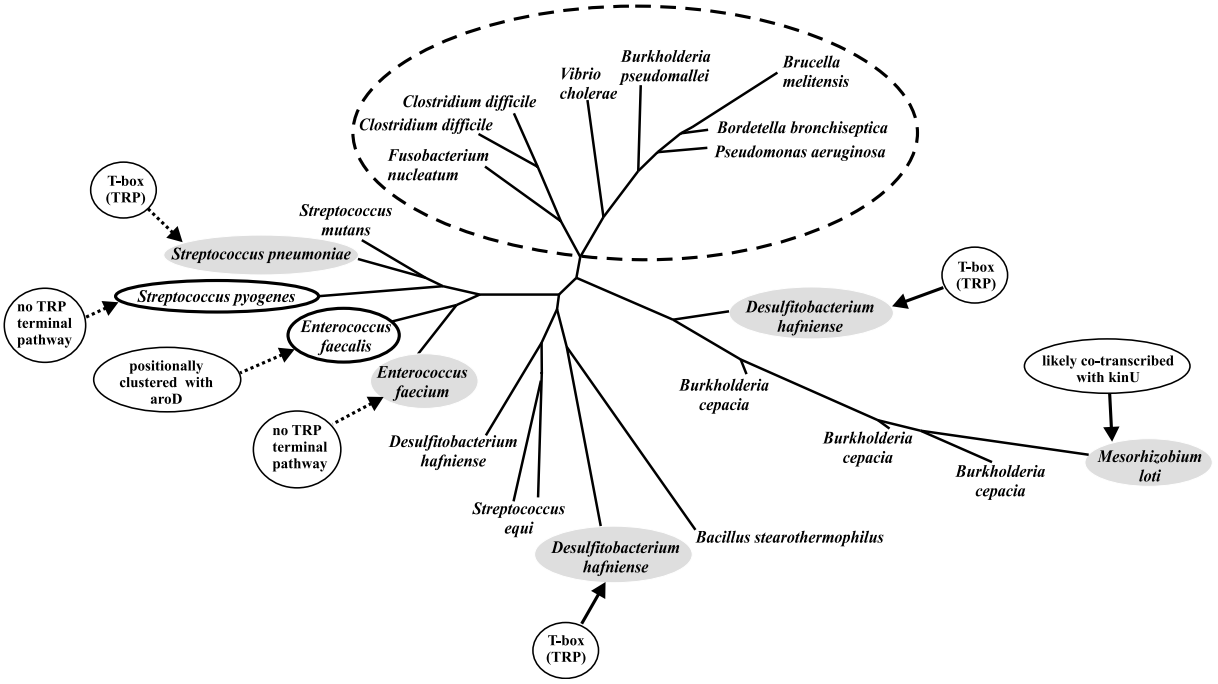

Fig. 5. Phylogenetic tree of the substrate binding component TrpX of the TrpXYZ transporter in the *Bacillus*/*Clostridium* group. Filled ovals: genes that are either regulated by tryptophan-specific T-boxes or positionally clustered with genes involved in tryptophan metabolism. Empty ovals: genes from the genomes that lack the tryptophan terminal pathway and thus should transport tryptophan from the environment. We predict tryptophan specificity for all but one major clades (the latter is circled by a dotted line) that contain genes with evidence for tryptophan specificity.

gene removes TRAP from the RNA. In this way, the tryptophan-dependent regulation of the bi-functional *pabA* gene does not interfere with folate synthesis.

4.2. Shuffling of regulatory systems

Table 1 summarizes the known and predicted regulatory elements of aromatic amino acid biosynthesis from the *Bacillus/Clostridium* group. The DNA-dependent regulation prevails in γ-proteobacteria [11] whereas in the *Bacillus/Clostridium* group the major type of regulation is RNA-dependent. However, the cores of regulons in Gram-negative and Gram-positive species coincide. These include the *trp* operon, DAHP synthase and shikimate kinase genes, and the *phhA* gene. Interestingly, even the type of regulation of these genes is almost conserved: in both groups the *trp* operon is regulated at the RNA level (although it is additionally regulated by a DNA binding repressor in γ-proteobacteria), whereas the DAHP synthase and shikimate kinase genes are regulated at the DNA level. In contrast, group-specific members of regulons, e.g. transporters *yhaG*, *trpXYZ*, *mtr*, *tyrP*, *aroP*, are regulated by variable mechanisms: by DNA-dependent regulation in Gram-negative genomes, and by RNA-dependent regulation in the Gram-positive group. The same pattern of regulation was observed for the *phhA* gene.

One notable exception to this rule is provided by *B. anthracis*. In contrast to other bacilli that display DNA level regulation of the *aroA* and *aroF* genes, *B. anthracis* has acquired T-boxes upstream of both genes, and thus shifted to RNA level regulation.

So far there seem to be four types of regulation of aromatic amino acid biosynthesis in the *Bacillus/Clostridium* group. The most general one is the T-box-dependent transcriptional regulation, which is present in all studied species. Another type of RNA-dependent transcriptional regulation, TRAP-mediated regulation, is unique to the *Bacillus* group except for *B. anthracis*, which lacks the TRAP protein. In *B. subtilis*, *B. halodurans*, and *B. stearothermophilus*, TRAP regulates transcription of the *trp* operon, which is regulated by tryptophan-specific T-boxes in all other species. The third type of regulation, PCE with consensus ACTTAAAAGCGTT, is also specific to *B. subtilis*, *B. halodurans*, and *B. stearothermophilus*, where it appears to regulate transcription of DAHP synthase and chorismate synthase genes. In *B. anthracis*, the same genes are regulated by tyrosine-specific T-boxes. Finally, in *S. pneumoniae*, *S. mutans*, and *L. lactis*, DAHP synthase and shikimate kinase genes seem to be under transcriptional regulation by ARO boxes identified in this study. The consensus of the ARO boxes is ATGGAGGCANATAA.

However, we could not identify the transcription factor responsible for this regulation, as no candidate sites or RNA elements were observed upstream of genes encoding proteins with potential DNA binding domains. This means that, unlike TrpR and TyrR of γ-proteobacteria, these hypothetical factors are not subject to auto-regulation.

Acknowledgements

We are grateful to Dmitry Rodionov and Andrey Osterman for useful discussion. This study was partially supported by grants from the Howard Hughes Medical Institute (55000309) and the Ludwig Institute for Cancer Research (CRDF RBO-1268).

References

[1] Pittard, A.J. (1996) Biosynthesis of aromatic amino acids. In: *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology (Neidhardt, F.C. et al., Eds.), pp. 458–484. ASM Press, Washington, DC.
[2] Sarsero, J.P., Merino, E. and Yanofsky, C. (2000) A *Bacillus subtilis* gene of previously unknown function, *yhaG*, is translationally regulated by tryptophan-activated TRAP and appears to be involved in tryptophan transport. J. Bacteriol. 182, 2329–2331.
[3] Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A. and Marmorstein, R.Q. et al. (1988) Crystal structure of the *trp* repressor/operator complex at atomic resolution. Nature 335, 321–329.
[4] Pittard, A.J. and Davidson, B.E. (1991) TyrR protein of *Escherichia coli* and its role as repressor and activator. Mol. Microbiol. 5, 1585–1592.
[5] Jackson, E.N. and Yanofsky, C. (1973) The region between the operator and first structural gene of the tryptophan operon of *Escherichia coli* may have a regulatory function. J. Mol. Biol. 76, 89–101.
[6] Yanofsky, C. (1981) Attenuation in the control of expression of bacterial operons. Nature 289, 751–758.
[7] Keller, E.B. and Calvo, J. (1979) Alternative secondary structures of leader RNAs and the regulation of the *trp*, *phe*, *his*, *thr*, and *leu* operons. Proc. Natl. Acad. Sci. USA 76, 6186–6190.
[8] Terai, G., Takagi, T. and Nakai, K. (2001) Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. Genome Biol. 11, research/0048.1.
[9] Babitzke, P. and Gollnick, P. (2001) Posttranscriptional initiation control of tryptophan metabolism in *Bacillus subtilis* by the trp RNA-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. J. Bacteriol. 183, 5795–5802.
[10] Henkin, T.M. (1994) tRNA-directed transcription antitermination. Mol. Microbiol. 13, 381–387.
[11] Panina, E.M., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2001) Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. J. Mol. Microbiol. Biotechnol. 3, 529–543.
[12] Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415–431.
[13] Vitreshchak, A., Mironov, A.A. and Gelfand, M.S. (2001) Computer prediction of RNA secondary structure. The RNApattern program: searching for RNA secondary structure by the pattern rule. Proceedings 3rd Int. Conf. 'Complex Systems: Control and modeling problems', Samara, pp. 623–625.
[14] Delorme, C., Ehrlich, S.D. and Renault, P. (1999) Regulation of expression of the *Lactococcus lactis* histidine operon. J. Bacteriol. 181, 2026–2037.
[15] Grundy, F.J., Haldeman, M.T., Hornblow, G.M., Ward, J.M.,

Chalker, A.F. and Henkin, T.M. (1997) The *Staphylococcus aureus ileS* gene, encoding isoleucyl-tRNA synthetase, is a member of the T-box family. J. Bacteriol. 179, 3767–3772.

[16] Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) Software for analysis of bacterial genomes. Mol. Biol. (Moscow) 34, 222–231.

[17] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgons, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

[18] Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266, 418–427.

[19] Green, J.M., Nichols, B.P. and Matthews, R.G. (1996) Folate biosynthesis, reduction, and polyglutamylation. In: *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology (Neidhardt, F.C. et al., Eds.), pp. 665–673. ASM Press, Washington, DC.

[20] Valbuzzi, A. and Yanofsky, C. (2001) Inhibition of the *B. subtilis* regulatory protein TRAP by the TRAP-inhibitory protein, AT. Science 293, 2057–2059.