THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 2002 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 277, No. 50, Issue of December 13, pp. 48949–48959, 2002
Printed in U.S.A.

Comparative Genomics of Thiamin Biosynthesis in Procaryotes

NEW GENES AND REGULATORY MECHANISMS*[S]

Received for publication, September 3, 2002
Published, JBC Papers in Press, October 9, 2002, DOI 10.1074/jbc.M208965200

Dmitry A. Rodionov‡§, Alexey G. Vitreschak¶, Andrey A. Mironov‡‖, and Mikhail S. Gelfand‡‖

From the ‡State Scientific Center GosNIIGenetika, Moscow 113545, Russia, ¶Institute for Problems of Information Transmission, Moscow 101447, Russia, and ‖Integrated Genomics, Moscow, P.O. Pox 348, Moscow 117333, Russia

Vitamin B_1 in its active form thiamin pyrophosphate is an essential coenzyme that is synthesized by coupling of pyrimidine (hydroxymethylpyrimidine; HMP) and thiazole (hydroxyethylthiazole) moieties in bacteria. Using comparative analysis of genes, operons, and regulatory elements, we describe the thiamin biosynthetic pathway in available bacterial genomes. The previously detected thiamin-regulatory element, *thi* box (Miranda-Rios, J., Navarro, M., and Soberon, M. (2001) *Proc. Natl. Acad. Sci. U. S. A.* 98, 9736–9741), was extended, resulting in a new, highly conserved RNA secondary structure, the *THI* element, which is widely distributed in eubacteria and also occurs in some archaea. Search for *THI* elements and analysis of operon structures identified a large number of new candidate thiamin-regulated genes, mostly transporters, in various prokaryotic organisms. In particular, we assign the thiamin transporter function to *yuaJ* in the *Bacillus/Clostridium* group and the HMP transporter function to an ABC transporter *thiXYZ* in some proteobacteria and firmicutes. By analogy to the model of regulation of the riboflavin biosynthesis, we suggest thiamin-mediated regulation based on formation of alternative RNA structures involving the *THI* element. Either transcriptional or translational attenuation mechanism may operate in different taxonomic groups, dependent on the existence of putative hairpins that either act as transcriptional terminators or sequester translation initiation sites. Based on analysis of co-occurrence of the thiamin biosynthetic genes in complete genomes, we predict that eubacteria, archaea, and eukaryota have different pathways for the HMP and hydroxyethylthiazole biosynthesis.

Thiamin pyrophosphate (vitamin B_1) is an essential cofactor for several important enzymes of the carbohydrate metabolism (1). Many microorganisms, as well as plants and fungi, synthesize thiamin, but it is not produced by vertebrates. The thiamin biosynthetic (TBS)[1] pathway of bacteria is outlined in Fig. 1. Thiamin monophosphate is formed by coupling of two independently synthesized moieties, HMP-PP and HET-P. In *Escherichia coli* and *Salmonella typhimurium*, this enzymatic step is mediated by the ThiE protein. At the next step, thiamin monophosphate is phosphorylated by ThiL to form thiamin pyrophosphate. The pyrimidine moiety of thiamin, HMP-PP, is synthesized from aminoimidazole ribotide, an intermediate of the purine biosynthesis pathway. ThiC produces HMP-P, which is then phosphorylated by the bifunctional HMP kinase/HMP-P kinase ThiD. The thiazole moiety of thiamin in *E. coli* is derived from tyrosine, cysteine, and 1-deoxy-D-xylulose phosphate in an unresolved chain of reactions involving the *thiF*, *thiS*, *thiG*, *thiH*, and *thiI* gene products. 1-Deoxy-D-xylulose phosphate, whose production is catalyzed by the *dxs* gene product, the latter utilizing thiamin pyrophosphate as a co-factor, is also used in the nonmevalonate pathway and the pyridoxal biosynthesis (2). ThiF catalyzes adenylation of the sulfur carrier protein ThiS by ATP. In addition, ThiI and IscS, enzymes shared by the thiamin and 4-thiouridine biosynthetic pathways, may play a role in the sulfur transfer chemistry. Three distinct kinases, ThiM, ThiD, and ThiK, are involved in the salvage of HET, HMP, and thiamin, respectively, from the culture medium. Thiamin, thiamin phosphate, and thiamin pyrophosphate are actively transported in enteric bacteria using the ABC transport system ThiBPQ (3). No other thiamin transporters, neither HET nor HMP transport systems, have been identified in bacteria. A gene for the thiamin kinase ThiK has not yet been identified in the complete genome of *E. coli*, although the genes for other mentioned proteins are known.

A similar TBS pathway exists in *Bacillus subtilis*, but instead of *thiH* it involves another probable thiazole biosynthesis gene, *yjbR*, which is most similar to the *thiO* gene from *Rhizobium etli* (4). It has been proposed that ThiO may have the amino acid oxidase activity in the thiazole biosynthesis (5). The traditional gene names are different in *E. coli* and *B. subtilis* (Table I). HMP biosynthesis protein ThiC, thiamin-phosphate pyrophosphorylase ThiE, and hydroxyethylthiazole kinase ThiM from *E. coli* have their counterparts in *B. subtilis* named ThiA, ThiC, and ThiK, respectively. Moreover, the bifunctional gene *thiD* from *E. coli* has two orthologs in *B. subtilis*, *yjbV* and *ywdB*, which separately could encode the biosynthetic and salvage HMP kinases (4). For consistency, unless specified otherwise, we use the *E. coli* gene names throughout.

No thiamin-regulatory genes have been identified in bacteria, but it has been shown that thiamin pyrophosphate is an effector molecule involved in the regulation of TBS genes. In *S. typhimurium*, the TBS operons *thiCEFSGH* and *thiMD* and the thiamin transport operon *thiBPQ* are transcriptionally regulated by thiamin pyrophosphate, whereas the *thiI* and *thiL* genes are not (3, 6–9). *B. subtilis* has a thiamin-regulated gene, *thiA*, and the *ywbI-thiKC* operon whose transcription is partially repressed by thiazole but not by thiamin (10, 11). Re-

* This study was partially supported by INTAS Grant 99-1476 and Howard Hughes Medical Institute Grant 55000309. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains one figure and one table.

§ To whom correspondence should be addressed: State Scientific Center GosNIIGenetika, Moscow 113545, Russia. Fax: 7-095-315-05-01; E-mail: rodionov@genetika.ru.

[1] The abbreviations used are: TBS, thiamin biosynthesis; HMP, hydroxymethylpyrimidine; HET, hydroxyethylthiazole; TMS, transmembrane segment; SD, Shine-Dalgarno.

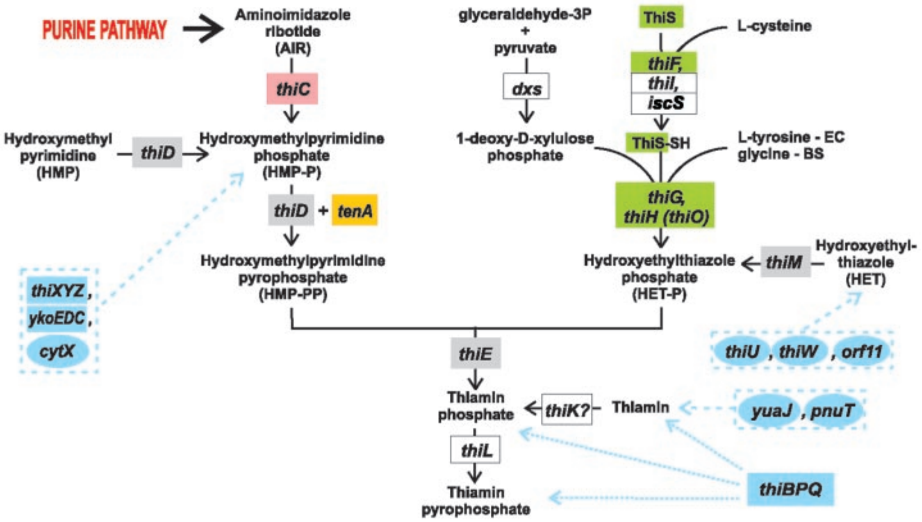

PURINE PATHWAY
Aminoimidazole ribotide (AIR)
thiC
Hydroxymethyl pyrimidine (HMP)
thiD
Hydroxymethylpyrimidine phosphate (HMP-P)
thiD + tenA
Hydroxymethylpyrimidine pyrophosphate (HMP-PP)
thiXYZ, ykoEDC, cytX
glyceraldehyde-3P + pyruvate
dxs
1-deoxy-D-xylulose phosphate
ThiS
L-cysteine
thiF, thiI, iscS
ThiS-SH
L-tyrosine - EC glycine - BS
thiG, thiH (thiO)
Hydroxyethylthiazole phosphate (HET-P)
thiM
Hydroxyethyl-thiazole (HET)
thiU, thiW, orf11
thiE
Thiamin phosphate
thiK?
Thiamin
yuaJ, pnuT
thiL
Thiamin pyrophosphate
thiBPQ

FIG. 1. **The thiamin biosynthesis pathway in bacteria.** The standard *E. coli* gene names are used throughout except *tenA-tenI*, *ykoFEDC*, *yuaJ*, and *ywdB* of *B. subtilis* and *thiO* of *R. etli* (see the Introduction for explanation and Table I for the *B. subtilis* equivalents). HET-P is synthesized using either L-tyrosine and ThiH (as in *E. coli*) or glycine and ThiO (as in *B. subtilis*). Proposed and known transport routes are shown by *dashed* and *arrowed lines*, respectively. Primary and ATP-dependent transporters are in *circles* and *rectangles*, respectively.

cently, a new thiamin-regulated operon, *tenA-tenI-yjbR-thiSGF-yjbV*, was detected in *B. subtilis* by the expression microarray analysis (12). Sequence analysis revealed the existence of putative Rho-independent transcriptional terminator sites in the upstream regions of the *B. subtilis* thiamin-regulated operons (4). Deletion of one such site located upstream of the *tenA-tenI-yjbR-thiSGF-yjbV* operon increased the expression level of *tenA* (13).

The 5′-untranslated region of the *R. etli thiCOGE* operon contains a 39-bp sequence, *thi* box, that is highly conserved in the upstream regions of the TBS genes from several bacterial genomes, and an additional stem-loop structure that would mask the ribosome binding site of *thiC* (5). Involvement of these two RNA structural elements in the thiamin-mediated translational regulation of the *R. etli* TBS operon has been demonstrated using deletion analysis (14). The exact mechanism by which thiamin inhibits translation initiation of the *thiC* gene remains to be determined. RNA elements similar to the *thi* box of *R. etli* have been observed upstream of the *thiC* genes from *E. coli*, *S. typhimurium*, *B. subtilis*, *Mycobacterium tuberculosis*, *Synechocystis* sp., and the *thiMD* operon from *S. typhimurium* (5).

Comparative analysis of many bacterial genomes is a powerful approach to reconstruction of metabolic pathways and their DNA or RNA regulation (for a review, see Ref. 15). In particular, analysis of the regulation of the riboflavin and biotin biosynthesis has shown that these vitamin regulons are highly conserved among unrelated bacteria (16, 17). In the former study, a model for the riboflavin-mediated regulation based on formation of alternative RNA structures involving the *RFN* elements has been suggested. To construct a single conserved structure of an RNA regulatory element, analysis of complementary substitutions in aligned sequences is used (18). In addition, analysis of positional clustering of genes on the chromosome helps in detection of functionally coupled genes (19). Simultaneous analysis of probable operon structures and regulatory elements is the most effective theoretical method of functional annotation when the standard homology-based methods are insufficient.

In this study, we analyzed the TBS pathway and the thiamin regulon in all available bacterial genomes by the comparative genomics approach. After extension of the *thi* box, we found a new RNA structure, the *THI* element, which is highly conserved on the sequence and structural levels. A possible mechanism of the *THI*-element-mediated regulation involving either transcriptional or translational attenuation was proposed for different groups of bacteria. Analysis of the candidate *THI* elements and positional clustering of the TBS genes resulted in identification of new thiamin-related genes, most of which are hypothetical transport systems. Finally, using metabolic reconstruction of the TBS pathway, we described some radical differences of the HET and HMP biosynthetic pathways in eubacteria, archaea, and eukaryota.

EXPERIMENTAL PROCEDURES

Complete and partial sequences of bacterial genomes were downloaded from GenBank™ (20). Preliminary sequence data were also obtained from the World Wide Web sites of the Institute for Genomic Research (www.tigr.org) the University of Oklahoma's Advanced Center for Genome Technology (www.genome.ou.edu/), the Wellcome Trust Sanger Institute (www.sanger.ac.uk/), the DOE Joint Genome Institute (jgi.doe.gov), and the ERGO data base (ergo.integratedgenomics.com/ERGO/) (21). Gene identifiers from the ERGO data base and GenBank™ are used throughout.

The RNA-PATTERN program (22) was used to search for conserved RNA regulatory elements. The input RNA pattern included both the RNA secondary structure and the sequence consensus motifs. The RNA secondary structure was described as a set of the following parameters: the number of helices, the length of each helix, the loop lengths, and the description of the topology of helix pairs. The initial RNA pattern of the *thi* box was constructed using the training set of eight *thi* boxes (5). Each genome was scanned with the *thi*-box pattern, resulting in detection of ~150 new *thi* boxes. Using multiple alignment of these *thi* boxes with flanking regions, additional conserved helices and sequence motifs were revealed, resulting in an extended RNA secondary structure, named the *THI* element. The RNA secondary structures of the *THI* elements, antiterminators, and antisequestors were predicted using Zuker's algorithm of free energy minimization (23) implemented in the Mfold program (available on the World Wide Web at bioinfo.math.rpi.edu/~mfold/rna).

Protein similarity search was done using the Smith-Waterman algorithm implemented in the *GenomeExplorer* program (24). Orthologous proteins were initially defined by the best bidirectional hits criterion (25) and, if necessary, confirmed by construction of phylogenetic trees. The phylogenetic trees were created by the maximum likelihood method implemented in PHYLIP (26) and drawn using the GeneMaster program.[2] Distant homologs were identified using PSI-BLAST (27). Transmembrane segments (TMSs) were predicted using the TMpred program (www.ch.embnet.org/software/TMPRED_form.html). Multiple sequence alignments were constructed using ClustalX (28).

RESULTS

THI Elements and Genes of Thiamin Biosynthesis and Transport

Orthologs of the thiamin biosynthesis and transport genes from *E. coli* and *B. subtilis* have been identified in all available

[2] A. Mironov, unpublished results.

TABLE I
The thiamin biosynthetic genes of E. coli (EC) and their counterparts in B. subtilis (BS)

EC	*BS*	Function	Similarity
			%
thiC	*thiA*	HMP biosynthesis	76
thiD	*yjbV*	Phosphomethylpyrimidine kinase	43
thiG	*yjbT*	Thiazole biosynthesis protein ThiG	52
thiH		Thiazole biosynthesis protein ThiH	
	yjbR	Thiazole biosynthesis protein ThiO	
thiI	*ytbJ*	Thiazole biosynthesis protein ThiI	33
thiF	*yjbU*	Adenylyltransferase	37
thiS	*yjbS*	Sulfur carrier protein ThiS	31
thiM	*thiK*	Hydroxyethylthiazole kinase	43
thiE	*thiC*	Thiamin-phosphate synthase	39
thiL	*ydiA*	Thiamin-monophosphate kinase	37

bacterial genomes by similarity search (Table I). We have not considered the *dxs*, *iscS*, and *thiI* genes because they are shared between the TBS and other pathways. Then we scanned 103 genomic sequences by the RNA-PATTERN program and found 170 *THI* elements in 78 genomes. It has been demonstrated that the thiamin biosynthesis is a widely distributed metabolic pathway in bacteria and it is usually regulated by the *THI* element. Note that the fact of gene absence is reliable only for complete genomes. Among all complete genomes, only spirochetes, mycoplasmas, chlamydiae, and rickettsiae have neither TBS genes nor *THI* elements. Two streptococci lack the TBS genes but have *THI* elements. In contrast, *Aquifex aeolicus*, *Helicobacter pylori*, *Lactococcus lactis*, *Legionella pneumophila*, *Magnetococcus* sp., and almost all archaeal genomes lack *THI* elements but have the TBS genes. The detailed phylogenetic and positional analysis of the TBS genes and the *THI* elements is given below.

At the first step, we have considered genomes that have no genes for the initial steps of the TBS pathway, namely *thiC* for the HMP biosynthesis and *thiS-thiG-thiH* (or *thiS-thiG-thiO*) for the HET biosynthesis. Most Gram-positive pathogens from the *Bacillus*/*Clostridium* group, all Pasteurellaeceae, and *H. pylori* lack both HMP and HET biosynthetic genes but have the *thiM*, *thiD*, and *thiE* genes. Thus, the TBS pathway in these organisms is incomplete and possibly uses exogenously supplied HET and HMP. Using analysis of the *THI* elements, we tried to identify candidate genes for the HMP and HET transport (see below). Homologs of genes for the HET biosynthesis are absent in all archaeal genomes as well as in the genome of *Thermotoga maritima*. However, all of these microorganisms except for *Aeropyrum pernix* and *Thermoplasma* sp. have the HMP biosynthetic gene *thiC*. In this work, we predict that archaea and *T. maritima*, like eukaryota, have a different pathway of HET biosynthesis (see below).

The similarities between the ThiF/ThiS and MoeB/MoaD proteins involved in the initial steps of TBS and molybdopterin biosynthesis, respectively, have already been described (29). In bacterial genomes containing the HET biosynthetic genes, we have identified either one or two ThiF/MoeB homologs per genome. Interestingly, all of these genes have been found in the loci containing either TBS or molybdopterin biosynthesis genes. However, the phylogenetic tree of the ThiF/MoeB family (data not shown) has several branches represented by both TBS-linked proteins (ThiF) and molybdopterin biosynthesis-linked proteins (MoeB). Thus, it is likely that the sulfur transfer chemistry of these two biosynthetic pathways can be shared in bacteria with only one ThiF/MoeB homolog. Alternatively, these organisms could have an unidentified ThiS-activating enzyme. Because of that, we do not consider *thiF* during analysis of the HET biosynthetic genes.

Two distinct enzymes, ThiH and ThiO, are involved in the HET biosynthesis in *E. coli* and *B. subtilis*, respectively. Similarity search in bacterial genomes has showed that aerobic microorganisms including α- and β-proteobacteria, pseudomonads, bacilli, actinomycetes, members of the *Thermus*/*Deinococcus* group, *A. aeolicus*, *L. pneumophila*, and *Magnetococcus* sp. have ThiO, whereas enterobacteria, clostridia, bacteria of the CFB group, *Shewanella putrefaciens*, *Campylobacter jejuni*, *Chlorobium tepidum*, and *Fusobacterium nucleatum*, which are mostly anaerobic microorganisms, have ThiH. This diversity in one enzyme of the HET biosynthesis can be explained by the use of different substrates for the synthesis of the thiazole moiety of thiamin by aerobes and anaerobes. Indeed, in two experimentally studied cases, an aerobe *B. subtilis* and a facultative anaerobe *E. coli* require glycine and tyrosine, respectively (30).

The *thiE* gene, which is required for coupling of the HET and HMP moieties of thiamin, has been identified in almost all organisms containing the TBS pathway except *T. maritima* and seven archaebacteria. The *thiD* gene encoding HMP kinase is the most widely distributed TBS gene, which is absent only in *Synechocystis* sp. Interestingly, the ThiD proteins from *T. maritima* and most archaea have an additional C-terminal domain of ~130 amino acids, whereas this domain is encoded by a separate gene in *Methanobacterium thermoautotrophicum*. The additional ThiD domain, named here ThiN, is not similar to any known protein and contains no conserved motifs. In all cases when ThiE is absent and ThiD is present, there is the ThiN domain, although in many cases ThiN and ThiE co-exist. We suggest that this conserved domain is somehow involved in the TBS, possibly replacing the ThiE function in the genomes of some archaea and *T. maritima*. The least common gene of the TBS pathway is the *thiM* gene encoding HET kinase from the thiazole salvage pathway. *thiM* was found only in the *Bacillus*/*Clostridium* group, enterobacteria, Pasteurellaecae, *Vibrio fischeri*, *H. pylori*, *Agrobacterium tumefaciens*, *Rhodobacter sphaeroides*, *Corynebacterium glutamicum*, and some archaea.

The operon structures of the TBS genes are quite diverse (Table II). Some genomes (*e.g. Corynebacterium diphtheriae*) have all TBS genes clustered in one putative operon, whereas the genomes of *A. aeolicus*, *Caulobacter crescentus*, *Magnetococcus* sp., and *Xylella fastidiosa* contain single TBS genes.

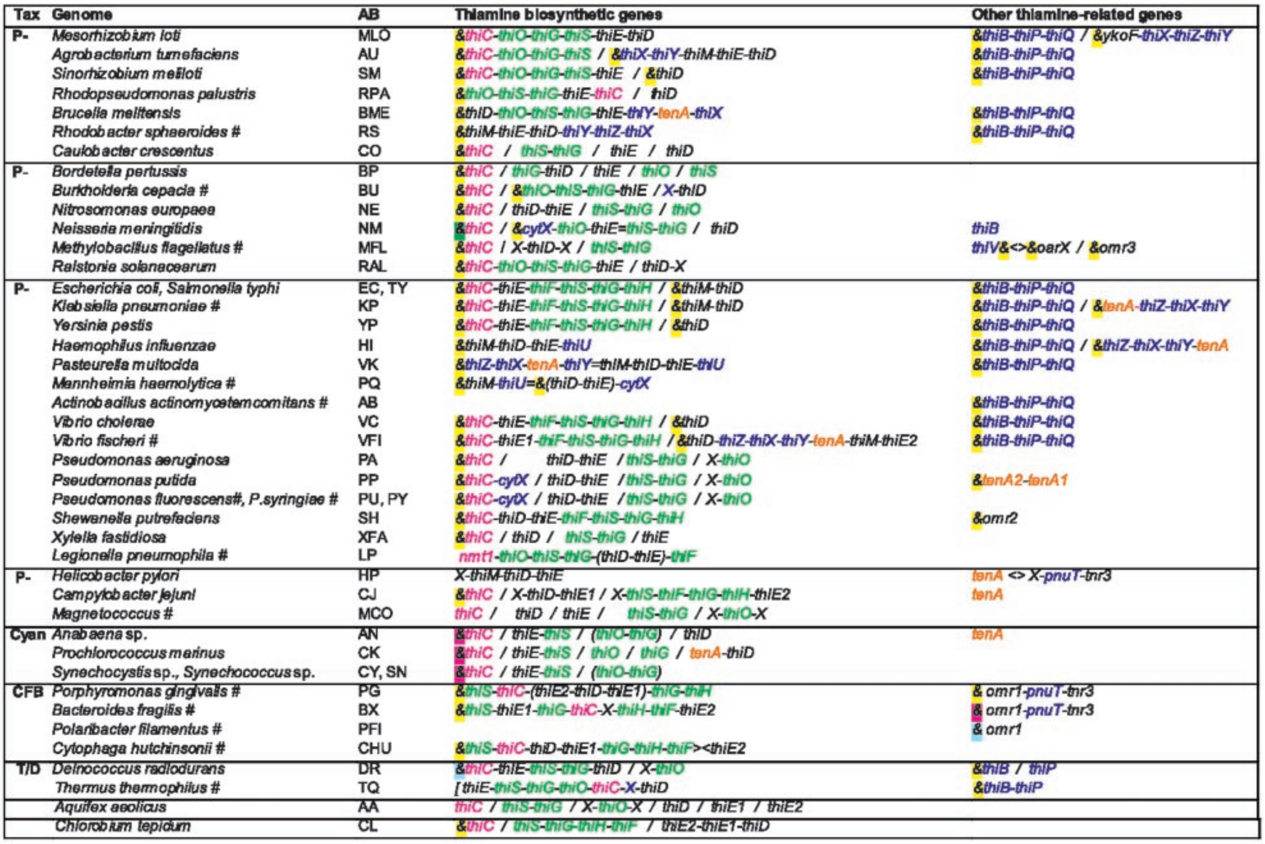
The *thiM*, *thiD*, and *thiE* genes, encoding adjacent enzymatic steps of the TBS pathway, often form clusters (probably operons) in a bacterial chromosome or can even be fused. The fused *thiE-thiD* genes were found in three bacteria, *C. glutamicum*, *L. pneumophila*, and *Porphyromonas gingivalis*, and one eukaryote, plant *Brassica napus*. In addition, several yeast genomes contain a single gene encoding the fused protein ThiE-ThiM.

Another frequently occurring gene cluster includes genes of

TABLE II

Thiamin biosynthesis and transport genes and THI elements in bacteria

The standard *E. coli* names of the TBS genes are used throughout (see the Introduction for the explanation and Table I for the *B. subtilis* equivalents). Genes of the HMP and thiazole biosynthesis are shown in magenta and green, respectively. Genes encoding transport proteins and the hypothetical TenA protein are shown in blue and orange, respectively. Parentheses denote gene fusions. Genes forming one candidate operon (with spacer less than 100 bp) are separated by a hyphen. Larger spacers between genes are marked by an equals sign. Operons from different loci, if shown in one column, are separated by slashes. Non-TBS genes are shown as *X*. Ampersands denote *THI* elements, and the background color indicates the proposed regulatory mechanism: yellow, sequestor; blue, terminator; green, dual terminator/sequestor; magenta, *THI* element is able to directly sequester the SD sequence. The contig ends are marked by square brackets. P-α, -β, -γ, -ε, Cyan, CFB, T/D, B/C, Actin, Ther, and A in the "Tax" column represent α-, β-, γ-, ε-proteobacteria, cyanobacteria, the CFB group, the *Thermus/Deinococcus* group, the *Bacillus/Clostridium* group, actinomycetes, thermotogales, and archaea, respectively. The genome abbreviations are given in column "AB" with unfinished genomes marked by #. Additional genome abbreviations as follows: MT, *Mycobacterium tuberculosis*; MB, *Mycobacterium bovis*; ML, *Mycobacterium leprae*; TAC, *Thermoplasma acidophilum*; FAC, *Ferroplasma acidarmanus*; TVO, *Thermoplasma volcanium*; MK, *Methanopyrus kandleri.*



Tax	Genome	AB	Thiamine biosynthetic genes	Other thiamine-related genes
P-	*Mesorhizobium loti*	MLO	&*thiC-thiO-thiG-thiS-thiE-thiD*	&*thiB-thiP-thiQ* / &*ykoF-thiX-thiZ-thiY*
	Agrobacterium tumefaciens	AU	&*thiC-thiO-thiG-thiS* / &*thiX-thiY-thiM-thiE-thiD*	&*thiB-thiP-thiQ*
	Sinorhizobium meliloti	SM	&*thiC-thiO-thiG-thiS-thiE* / &*thiD*	&*thiB-thiP-thiQ*
	Rhodopseudomonas palustris	RPA	&*thiO-thiS-thiG-thiE-thiC* / *thiD*	
	Brucella melitensis	BME	&*thiD-thiO-thiS-thiG-thiE-thiY-tenA-thiX*	&*thiB-thiP-thiQ*
	Rhodobacter sphaeroides #	RS	&*thiM-thiE-thiD-thiY-thiZ-thiX*	&*thiB-thiP-thiQ*
	Caulobacter crescentus	CO	&*thiC* / *thiS-thiG* / *thiE* / *thiD*	
P-	*Bordetella pertussis*	BP	&*thiC* / *thiG-thiD* / *thiE* / *thiO* / *thiS*	
	Burkholderia cepacia #	BU	&*thiC* / &*thiO-thiS-thiG-thiE* / *X-thiD*	
	Nitrosomonas europaea	NE	&*thiC* / *thiD-thiE* / *thiS-thiG* / *thiO*	
	Neisseria meningitidis	NM	&*thiC* / &*cytX-thiO-thiE=thiS-thiG* / *thiD*	*thiB*
	Methylobacillus flagellatus #	MFL	&*thiC* / *X-thiD-X* / *thiS-thiG*	*thiV*&<>&*oarX* / &*omr3*
	Ralstonia solanacearum	RAL	&*thiC-thiO-thiS-thiG-thiE* / *thiD-X*	
P-	*Escherichia coli, Salmonella typhi*	EC, TY	&*thiC-thiE-thiF-thiS-thiG-thiH* / &*thiM-thiD*	&*thiB-thiP-thiQ*
	Klebsiella pneumoniae #	KP	&*thiC-thiE-thiF-thiS-thiG-thiH* / &*thiM-thiD*	&*thiB-thiP-thiQ* / &*tenA-thiZ-thiX-thiY*
	Yersinia pestis	YP	&*thiC-thiE-thiF-thiS-thiG-thiH* / &*thiD*	&*thiB-thiP-thiQ*
	Haemophilus influenzae	HI	&*thiM-thiD-thiE-thiU*	&*thiB-thiP-thiQ* / &*thiZ-thiX-thiY-tenA*
	Pasteurella multocida	VK	&*thiZ-thiX-tenA-thiY=thiM-thiD-thiE-thiU*	&*thiB-thiP-thiQ*
	Mannheimia haemolytica #	PQ	&*thiM-thiU*=&*(thiD-thiE)-cytX*	
	Actinobacillus actinomycetemcomitans #	AB		&*thiB-thiP-thiQ*
	Vibrio cholerae	VC	&*thiC-thiE-thiF-thiS-thiG-thiH* / &*thiD*	&*thiB-thiP-thiQ*
	Vibrio fischeri #	VFI	&*thiC-thiE1-thiF-thiS-thiG-thiH* / &*thiD-thiZ-thiX-thiY-tenA-thiM-thiE2*	&*thiB-thiP-thiQ*
	Pseudomonas aeruginosa	PA	&*thiC* / *thiD-thiE* / *thiS-thiG* / *X-thiO*	
	Pseudomonas putida	PP	&*thiC-cytX* / *thiD-thiE* / *thiS-thiG* / *X-thiO*	&*tenA2-tenA1*
	Pseudomonas fluorescens#, *P.syringiae* #	PU, PY	&*thiC-cytX* / *thiD-thiE* / *thiS-thiG* / *X-thiO*	
	Shewanella putrefaciens	SH	&*thiC-thiD-thiE-thiF-thiS-thiG-thiH*	&*omr2*
	Xylella fastidiosa	XFA	&*thiC* / *thiD* / *thiS-thiG* / *thiE*	
	Legionella pneumophila #	LP	*nmt1-thiO-thiS-thiG-(thiD-thiE)-thiF*	
P-	*Helicobacter pylori*	HP	*X-thiM-thiD-thiE*	*tenA* <> *X-pnuT-tnr3*
	Campylobacter jejuni	CJ	&*thiC* / *X-thiD-thiE1* / *X-thiS-thiF-thiG-thiH-thiE2*	*tenA*
	Magnetococcus #	MCO	*thiC* / *thiD* / *thiE* / *thiS-thiG* / *X-thiO-X*	
Cyan	*Anabaena* sp.	AN	&*thiC* / *thiE-thiS* / *(thiO-thiG)* / *thiD*	*tenA*
	Prochlorococcus marinus	CK	&*thiC* / *thiE-thiS* / *thiO* / *thiG* / *tenA-thiD*	
	Synechocystis sp., *Synechococcus* sp.	CY, SN	&*thiC* / *thiE-thiS* / *(thiO-thiG)*	
CFB	*Porphyromonas gingivalis* #	PG	&*thiS-thiC-(thiE2-thiD-thiE1)-thiG-thiH*	& *omr1-pnuT-tnr3*
	Bacteroides fragilis #	BX	&*thiS-thiE1-thiG-thiC-X-thiH-thiF-thiE2*	& *omr1-pnuT-tnr3*
	Polaribacter filamentus #	PFI		& *omr1*
	Cytophaga hutchinsonii #	CHU	&*thiS-thiC-thiD-thiE1-thiG-thiH-thiF*><*thiE2*	
T/D	*Deinococcus radiodurans*	DR	&*thiC-thiE-thiS-thiG-thiD* / *X-thiO*	&*thiB* / *thiP*
	Thermus thermophilus #	TQ	[*thiE-thiS-thiG-thiO-thiC-X-thiD*	&*thiB-thiP*
	Aquifex aeolicus	AA	*thiC* / *thiS-thiG* / *X-thiO-X* / *thiD* / *thiE1* / *thiE2*	
	Chlorobium tepidum	CL	&*thiC* / *thiS-thiG-thiH-thiF* / *thiE2-thiE1-thiD*	

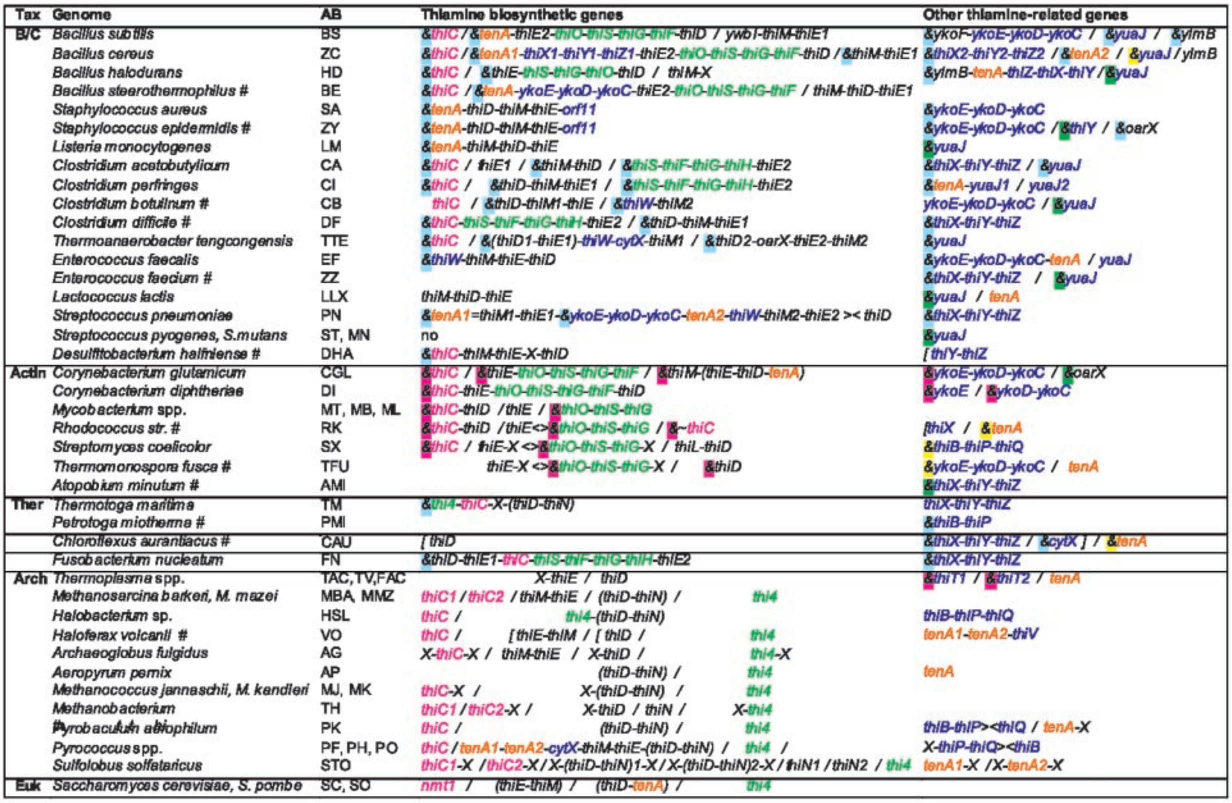
the HET biosynthesis: *thiF*, *thiS*, *thiG*, and *thiH* (or *thiO*). Again, we have observed a single gene encoding the fused protein ThiO-ThiG in two cyanobacteria.

A search for *THI* elements upstream of TBS genes showed that the TBS pathways of all eubacteria, except *A. aeolicus*, *H. pylori*, *L. pneumophila*, *L. lactis*, and *Magnetococcus* sp., are regulated by *THI* elements (Table II). Moreover, the TBS pathways in about half of these bacteria seem to be completely regulated, since all TBS operons have upstream *THI* elements; about one-fourth of the genomes contain only one *THI* element-regulated gene *thiC*, and the remaining bacteria apparently have partially regulated TBS pathways. The *thiC* gene is the most tightly *THI*-regulated gene of the TBS pathway, since only *Clostridium botulinum* has *THI* regulation, but not of *thiC*. Finally, the archaeal TBS operons apparently are not regulated by *THI* elements.

The *thiB-thiP-thiQ* operon encoding an ATP-dependent transport system for thiamin has been identified in most α- and γ-proteobacteria and *Streptomyces coelicolor*, and in all of these cases it is preceded by *THI* elements. In addition, bacteria from the *Thermus/Deinococcus* group and *Petrotoga miotherma* have incomplete *thiB-thiP* loci, which are also *THI*-regulated (*cf*. discussion of the ThiX-ThiY-ThiZ system below). The *thiB-thiP-thiQ* loci without *THI* elements were detected in several archaea, namely *Halobacterium* sp., *Pyrobaculum aerophilum*, and *Pyrococcus* species. The *thiB-thiP-thiQ* genes never cluster with TBS genes.

Comparison of TBS protein phylogenetic trees with the standard trees for ribosomal proteins reveals some unusual branches. The most interesting observation is a likely horizontal transfer of the *thiM-thiD-thiE* genes from *Listeria* species to three Pasteurellaeceae. For instance, the ThiD proteins from *Hemophilus influenzae*, *Pasteurella multocida*, and *Mannheimia hemolytica* are close to ThiD from the *Bacillus/Clostridium* group, showing the highest similarity to *Listeria* species, and the same holds for other phylogenetic trees (data not shown). Among γ-proteobacteria, only Pasteurellaeceae have an incomplete TBS pathway (*i.e.* ThiM-ThiD-ThiE), which is widely distributed in Gram-positive pathogens from the *Bacillus/Clostridium* group. Another example of possible horizontal transfer is the *thiM-thiD-thiE* operon of *H. pylori.* Again, the TBS proteins of this bacterium are similar to the proteins from the *Bacillus/Clostridium* group.

TABLE II—*continued*



Tax	Genome	AB	Thiamine biosynthetic genes	Other thiamine-related genes
B/C	*Bacillus subtilis*	BS	*&thiC / &tenA-thiE2-thiO-thiS-thiG-thiF-thiD / ywbI-thiM-thiE1*	*&ykoF-ykoE-ykoD-ykoC / &yuaJ / &ylmB*
	Bacillus cereus	ZC	*&thiC / &tenA1-thiX1-thiY1-thiZ1-thiE2-thiO-thiS-thiG-thiF-thiD / &thiM-thiE1*	*&thiX2-thiY2-thiZ2 / &tenA2 / &yuaJ / ylmB*
	Bacillus halodurans	HD	*&thiC / &thiE-thiS-thiG-thiO-thiD / thiM-X*	*&ylmB-tenA-thiZ-thiX-thiY / &yuaJ*
	Bacillus stearothermophilus #	BE	*&thiC / &tenA-ykoE-ykoD-ykoC-thiE2-thiO-thiS-thiG-thiF / thiM-thiD-thiE1*	
	Staphylococcus aureus	SA	*&tenA-thiD-thiM-thiE-orf11*	*&ykoE-ykoD-ykoC*
	Staphylococcus epidermidis #	ZY	*&tenA-thiD-thiM-thiE-orf11*	*&ykoE-ykoD-ykoC / &thiY / &oarX*
	Listeria monocytogenes	LM	*&tenA-thiM-thiD-thiE*	*&yuaJ*
	Clostridium acetobutylicum	CA	*&thiC / thiE1 / &thiM-thiD / &thiS-thiF-thiG-thiH-thiE2*	*&thiX-thiY-thiZ / &yuaJ*
	Clostridium perfringes	CI	*&thiC / &thiD-thiM-thiE1 / &thiS-thiF-thiG-thiH-thiE2*	*&tenA-yuaJ1 / yuaJ2*
	Clostridium botulinum #	CB	*thiC / &thiD-thiM1-thiE / &thiW-thiM2*	*ykoE-ykoD-ykoC / &yuaJ*
	Clostridium difficile #	DF	*&thiC-thiS-thiF-thiG-thiH-thiE2 / &thiD-thiM-thiE1*	*&thiX-thiY-thiZ*
	Thermoanaerobacter tengcongensis	TTE	*&thiC / &(thiD1-thiE1)-thiW-cytX-thiM1 / &thiD2-oarX-thiE2-thiM2*	*&yuaJ*
	Enterococcus faecalis	EF	*&thiW-thiM-thiE-thiD*	*&ykoE-ykoD-ykoC-tenA / yuaJ*
	Enterococcus faecium #	ZZ		*&thiX-thiY-thiZ / &yuaJ*
	Lactococcus lactis	LLX	*thiM-thiD-thiE*	*&yuaJ / tenA*
	Streptococcus pneumoniae	PN	*&tenA1=thiM1-thiE1-&ykoE-ykoD-ykoC-tenA2-thiW-thiM2-thiE2 >< thiD*	*&thiX-thiY-thiZ*
	Streptococcus pyogenes, S.mutans	ST, MN	no	*&yuaJ*
	Desulfitobacterium halfniense #	DHA	*&thiC-thiM-thiE-X-thiD*	*[thiY-thiZ*
Actin	*Corynebacterium glutamicum*	CGL	*&thiC / &thiE-thiO-thiS-thiG-thiF / &thiM-(thiE-thiD-tenA)*	*&ykoE-ykoD-ykoC / &oarX*
	Corynebacterium diphtheriae	DI	*&thiC-thiE-thiO-thiS-thiG-thiF-thiD*	*&ykoE / &ykoD-ykoC*
	Mycobacterium spp.	MT, MB, ML	*&thiC-thiD /thiE / &thiO-thiS-thiG*	
	Rhodococcus str. #	RK	*&thiC-thiD /thiE<>&thiO-thiS-thiG / &~thiC*	*[thiX / &tenA*
	Streptomyces coelicolor	SX	*&thiC / thiE-X <>&thiO-thiS-thiG-X / thiL-thiD*	*&thiB-thiP-thiQ*
	Thermomonospora fusca #	TFU	*thiE-X <>&thiO-thiS-thiG-X / &thiD*	*&ykoE-ykoD-ykoC / tenA*
	Atopobium minutum #	AMI		*&thiX-thiY-thiZ*
Ther	*Thermotoga maritima*	TM	*&thi4-thiC-X-(thiD-thiN)*	*thiX-thiY-thiZ*
	Petrotoga miotherma #	PMI		*&thiB-thiP*
	Chloroflexus aurantiacus #	CAU	*[thiD*	*&thiX-thiY-thiZ / &cytX] / &tenA*
	Fusobacterium nucleatum	FN	*&thiD-thiE1-thiC-thiS-thiF-thiG-thiH-thiE2*	*&thiX-thiY-thiZ*
Arch	*Thermoplasma* spp.	TAC,TV,FAC	*X-thiE / thiD*	*&thiT1 / &thiT2 / tenA*
	Methanosarcina barkeri, M. mazei	MBA, MMZ	*thiC1 / thiC2 / thiM-thiE / (thiD-thiN) / thi4*	
	Halobacterium sp.	HSL	*thiC / thi4-(thiD-thiN)*	*thiB-thiP-thiQ*
	Haloferax volcanii #	VO	*thiC / [thiE-thiM / [thiD / thi4*	*tenA1-tenA2-thiV*
	Archaeoglobus fulgidus	AG	*X-thiC-X / thiM-thiE / X-thiD / thi4-X*	
	Aeropyrum pernix	AP	*(thiD-thiN) / thi4*	*tenA*
	Methanococcus jannaschii, M. kandleri	MJ, MK	*thiC-X / X-(thiD-thiN) / thi4*	
	Methanobacterium	TH	*thiC1 / thiC2-X / X-thiD / thiN / X-thi4*	
	Pyrobaculum aerophilum	PK	*thiC / (thiD-thiN) / thi4*	*thiB-thiP><thiQ / tenA-X*
	Pyrococcus spp.	PF, PH, PO	*thiC / tenA1-tenA2-cytX-thiM-thiE-(thiD-thiN) / thi4 /*	*X-thiP-thiQ><thiB*
	Sulfolobus solfataricus	STO	*thiC1-X / thiC2-X / X-(thiD-thiN)1-X / X-(thiD-thiN)2-X / thiN1 / thiN2 / thi4*	*tenA1-X /X-tenA2-X*
Euk	*Saccharomyces cerevisiae, S. pombe*	SC, SO	*nmt1 / (thiE-thiM) / (thiD-tenA) / thi4*	

New Thiamin-regulated Genes

Transporters—A search for *THI* elements in bacterial genomes complemented by analysis of the putative operon structure of the TBS genes has allowed us to detect a number of new thiamin-related genes. Most of these genes encode new transport systems.

The single *THI*-regulated gene *yuaJ* (the *B. subtilis* name) was found in all complete genomes of the *Bacillus*/*Clostridium* group except *Staphylococcus aureus* and *Streptococcus pneumoniae* (Table II). It is always preceded by a *THI* element with only one exception in *Enterococcus faecalis* and is never clustered with TBS genes. *Clostridium perfringes* has two *yuaJ* paralogs, with and without an upstream *THI* element. YuaJ has six predicted transmembrane segments (TMSs) and is not similar to any known protein. *yuaJ* is the only thiamin-regulated gene in the complete genomes of *Streptococcus mutans* and *Streptococcus pyogenes*, which have no genes for the TBS pathway. These observations strongly suggest that YuaJ is a thiamin transporter, which, in contrast to ThiB-ThiP-ThiQ, is obviously not ATP-dependent. In support of this prediction, the thiamin uptake in *Bacillus cereus*, which has *yuaJ*, is coupled to the proton movement (31).

A hypothetical thiamin-related ABC transporter, named here *thiX-thiY-thiZ*, was identified in bacteria from various taxonomic divisions, such as α- and γ-proteobacteria, the *Bacillus/Clostridium* group, and Thermotogales. The first gene, *thiX*, encodes the transmembrane component of the ABC transport system, whereas the second (*thiY*) and the third (*thiZ*) genes encode the substrate- and ATP-binding components, respectively. These genes have upstream *THI* elements in all cases with only one exception in *T. maritima*. In *A. tumefaciens*, *R. sphaeroides*, *B. melitensis*, *Pasteurella multocida*, *V. fischeri*, and *B. cereus*, the *thiX-thiY-thiZ* genes are clustered with the *thiD* gene that encodes HMP kinase. In contrast to *yuaJ*, the *thiX-thiY-thiZ* operon is not found in the genomes without TBS genes, but sometimes it occurs in genomes with the incomplete TBS pathway. The need of HMP and HET moiety for the thiamin biosynthesis is obvious. However, pathways other than TBS that could supply these compounds are not known. The putative substrate-binding protein ThiY is similar to enzymes for the HMP biosynthesis from yeasts, namely Thi3 of *Schizosaccharomyces pombe* and Thi5 of *Saccharomyces cerevisiae*. All found ThiY orthologs are predicted to have an N-terminal transmembrane segment, which is common for substrate-binding components of ABC transporters. Thus, we predict that ThiX-ThiY-ThiZ is a HMP transport system that substitutes for missing HMP biosynthesis in some bacteria. Unusually, *Brucella melitensis* and *A. tumefaciens* have ThiX-ThiY but miss the ATPase component ThiZ. A similar situation with incomplete ThiB-ThiP systems in some bacteria has been described above. Based on the experimental fact that ATPases of different ABC transport systems can be functionally exchangeable (32), we suggest that the incomplete ThiXY and ThiBP systems could use another ATPase component. The HMP specificity of the ThiXYZ system is further supported by the observation that *B. melitensis* lacks the HMP pathway but not the HET pathway.

In some Gram-positive bacteria, we have found another thiamin-related ABC transporter, YkoE-YkoD-YkoC. It consists of two transmembrane components (YkoE and YkoC) and an ATPase component (YkoD). We could not identify a substrate-binding component for this system. Similarly to *thiX-thiY-thiZ*, the *ykoE-ykoD-ykoC* genes always co-occur with the TBS genes and are preceded by a *THI* element. They have also been found in genomes with the incomplete TBS pathway. In *B. subtilis*, the first gene of the *THI*-regulated *ykoF-ykoE-ykoD-ykoC* operon is not similar to any known protein and has only one ortholog in *Mesorhizobium loti*, where it clusters with the above described candidate HMP transporter, forming a *THI*-regulated cluster, *ykoF-thiX-thiY-thiZ*. Thus, the new ABC transport system YkoE-YkoD-YkoC is obviously thiamin-related and most likely is involved in the HMP transport for TBS. This prediction is based on positional clustering and on the following fact: when YkoEDC occurs in genomes lacking both HMP and HET pathways, there always is a candidate HET transporter (see below) but not other HMP transporters.

The first gene of the TBS operon in *Neisseria meningitidis*, *NMB2067*, encodes a hypothetical transporter with 12 predicted TMSs, which is similar to the cytosine permease CodB from *E. coli*. Orthologs of this gene, named *cytX*, exist in *M. hemolytica*, *Chloroflexus aurantiacus*, *Thermoanaerobacter tengcongensis*, pseudomonads, and pyrococci. In all cases, *cytX* either clusters with the TBS genes or has upstream *THI* elements or both. Based on positional analysis and similarity to the pyrimidine transporter, the new thiamin-related transporter CytX is most likely involved in the HMP transport.

The last gene of the TBS operon *thiMDE-HI0418* in *H. influenzae* encodes a hypothetical transmembrane protein with 12 predicted TMSs. This gene is similar to transporters from the MFS family and has orthologs in two other Pasteurellaecae, *P. multocida* and *M. hemolytica*. Since *HI0418* and all of its orthologs are clustered with the *thiMDE* genes and *THI*-regulated, we named this new thiamin-related transporter *thiU*. *H. influenzae* and *P. multocida* lack both HMP and HET pathways. The former is accounted for by the ThiXYZ system (see above). This, together with positional analysis, suggests that ThiU is a HET transporter.

The TBS operons of *S. pneumoniae*, *C. botulinum*, *T. tengcongensis*, and *E. faecalis* contain a new gene (*SP0723* in *S. pneumoniae*) encoding yet another thiamin-related transporter with five predicted TMSs. This gene, named *thiW*, is not similar to any known protein and has no homologs in other genomes. It is always *THI*-regulated and located immediately upstream of the *thiM* gene in all cases. Similarly, the last gene of the TBS operon in three *Staphylococcus* species (*orf11* in *S. carnosus*) encodes a hypothetical transporter with five TMSs. Since ThiW seems to complement the absence of the HET pathway in *T. tengcongensis* and *S. pneumoniae* and since Orf11 does the same in *S. aureus*, we tentatively predict that these proteins are involved in transport of the thiazole moiety of thiamin in the above bacteria.

In *Methylobacillus flagellatus*, one of the detected *THI* elements precedes a new thiamin-related gene, named *thiV*, which encodes a hypothetical transmembrane protein with 13 predicted TMSs. ThiV is similar to the pantothenate symporter PanF from *E. coli* and has only one ortholog in archaeon *Haloferax volcanii*, where it is clustered with the thiamin-related gene *tenA* (see below).

Bacteria from the CFB group, *Bacteroides fragilis* and *P. gingivalis*, contain candidate transporter *pnuT*. It encodes a protein with six predicted TMSs and is homologous to *pnuC* of enterobacteria, encoding the *N*-ribosylnicotinamide transporter (for other details, see "Enzymes").

Finally, we have observed the first example of *THI* element regulation in archaea. The archaeal *THI* elements were found upstream of two paralogous genes, named *thiT1* and *thiT2*, in each of the three *Thermoplasma* genomes. These genes encode hypothetical transmembrane proteins with nine predicted TMSs similar to transporters of the MFS family. The specificity of these transporters is not clear because of incompleteness of the TBS pathways in thermoplasmas, which have only *thiD* and *thiE* genes. However, based on the assumption that these transporters are the only thiamin-regulated genes in thermoplasmas, we propose their possible involvement in the thiamin transport.

Enzymes—*tenA* and *tenI*, two genes of unknown function located in the TBS operon of *B. subtilis*, were previously described as hypothetical regulators of extracellular enzyme production. However, they were not essential for the cell growth and the production of extracellular enzymes (13).

A similarity search demonstrated that *tenA* is a widely distributed *THI*-regulated gene in eubacteria and archaea, which usually is positionally linked to thiamin-related genes. It is never observed in genomes without the thiamin biosynthetic pathway (Table II). Analysis of the operon structure shows that *tenA* often forms one putative operon with either TBS genes or thiamin-related transporters (*thiX-thiY-thiZ* or *ykoE-ykoD-ykoC*), which are always *THI*-regulated. A single *tenA* gene can be *THI*-regulated (*B. cereus*, *Rhodococcus*, *C. aurantiacus*, *P. putida*) or not (*L. lactis*, *Nostoc* sp., *Thermomonospora fusca*, *H. pylori*, *C. jejuni*, and some archaea). In bacterial genomes, *tenA* always co-occurs with *thiD*, whereas in available genomes of thiamin-producing eukaryotes there are single genes encoding a fused protein ThiD-TenA. However, the C-terminal part of the ThiD-TenA protein is not required for the HMP-P kinase activity (33). In addition, three genes, *thiE*, *thiD*, and *tenA*, are fused in bacterium *C. glutamicum*. A BLAST search did not reveal similarity of TenA to any known protein except eukaryotic ThiD-TenA fusions. Thus, TenA is somehow associated with ThiD and may play an auxiliary role in the thiamin metabolism.

Orthologs of the *tenI* gene have been detected in bacilli, clostridiae, members of the CFB group, *C. jejuni*, *F. nucleatum*, *C. tepidum*, and *A. aeolicus*. They are mostly clustered with the TBS genes and are *THI*-regulated. Furthermore, a single gene encoding fused protein TenI-ThiD-ThiE was observed in the TBS operon of *P. gingivalis*. *tenI* shows significant similarity to *thiE*, and the *tenI* genes do not form a separate branch on the phylogenetic tree of the ThiE-TenI protein family (data not shown). Thus, we suggest that *tenI* genes are recent paralogs of *thiE*, and we use the notation *thiE1* and *thiE2* (Table II).

In *M. flagellatus*, *C. glutamicum*, and *Staphylococcus epidermidis*, the *THI* elements were found upstream of a single gene encoding a protein from the short-chain dehydrogenase/reductase superfamily. We named this gene *oarX* because of its high similarity to 3-oxoacyl-(acyl-carrier protein) reductase *fabG* from *B. subtilis*. Importantly, one more ortholog of *oarX*, *fabG5*, belongs to the *THI*-regulated *thiD-oarX-thiE-thiM* operon of *T. tengcongensis*, a recently sequenced bacterium from the *Bacillus*/*Clostridium* group. The obtained data seem to be sufficiently strong to warrant experimental analysis of the functional role of the *oarX* gene product in the thiamin metabolism.

Another new gene of the thiamin regulon (*ylmB* in *Bacillus* sp.) belongs to the ArgE/dapE/ACY1/CPG2/yscS family of metallopeptidases. The single *ylmB* gene in *B. subtilis* and the putative *ylmB-tenA-thiX-thiY-thiZ* operon in *B. halodurans* are preceded by *THI* elements, but no regulatory element was found upstream of the single *ylmB* gene in *B. cereus*.

In two bacteria from the CFB group, *B. fragilis* and *P. gingivalis*, *THI* elements precede a hypothetical operon, named *omr1-pnuT-tnr3*. The *omr1* gene, named by abbreviation of "outer membrane receptor," is similar to TonB-dependent outer membrane receptors of Gram-negative bacteria that perform high affinity binding and energy-dependent uptake of specific substrates into the periplasmic space. Among known substrates of the TonB-dependent receptors are various siderophores, bacteriocins and vitamin B_{12} (34). In another bacterium from the CFB group, *Polaribacter filamentus*, *omr1* is a single gene that is preceded by a *THI* element. The *pnuT* gene encodes a transporter that is similar to PnuC, *N*-ribosylnicotinamide transporters from enterobacteria (35). The PnuT proteins form a single branch on the phylogenetic tree of the PnuC family of transporters (data not shown). Another separate branch on this tree includes the recently identified riboflavin transporters PnuX (16). The last gene of the *omr1-pnuT-tnr3* operon is weakly similar to the C-terminal part of the thiamin pyrophosphokinase TNR3 from yeast *S. pombe*. Based on these data, we propose that the hypothetical *THI*-regulated *omr1-pnuT-tnr3* operon could be involved in the thiamin transport and its subsequent phosphorylation up to thiamin pyrophosphate. In confirmation, the hypothetical *HP1290-HP1291* operon of *H. pylori*, which is similar to *pnuT-tnr3* from the CFB bacterial group, forms a divergon with the thiamin-related gene *tenA*. In general, the PnuC family of transporters seems to transfer nonphosphorylated precursors (such as thiamin but not thiamin phosphate, riboflavin but not FMN, *N*-ribosylnicotinamide but not nicotinamide mononucleotide), which are then phosphorylated by specific kinases (TNR3, RibF, and NadR for thiamin, riboflavin, and *N*-ribosylnicotinamide, respectively) to make transport "vectorial." More thiamin-related TonB-dependent receptors have been found in *S. putrefaciens* and *M. flagellatus.* Each of these bacteria has a hypothetical *THI*-regulated TonB-dependent receptor, named *omr2* in *S. putrefaciens* and *omr3* in *M. flagellatus*. The amino acid identity between the Omr1, Omr2, and Omr3 proteins is about 20%.

Possible Attenuation Mechanism for the THI-mediated Regulation

Using the alignment of 170 *thi* boxes with flanking regions, additional conserved helices and sequence motifs were revealed. It resulted in an extended RNA secondary structure, named the *THI* element (Fig. 2). The conserved secondary structure has five helices and a single base stem of at least three base pairs (Fig. 3). Among them, only the first, fourth, and fifth helices are conserved on the sequence level. Of 14 conserved base pairs of the above helices, nine are invariant on the sequence level, whereas the remaining positions are confirmed by compensatory substitutions. In addition, 23 nonpaired positions are strongly conserved in the *THI* element. The conserved region including the fourth and fifth helices comprises the previously defined *thi* box (5). The internal loop between stem-loops 2 and 3 contains an absolutely invariant segment, UGAGA.

The base stem of the *THI* element can form in most bacteria, with several exceptions in thermoplasmas and actinomycetes (see below), but it is not conserved on the sequence level. The *THI* element contains two other nonconserved structure elements, an additional stem-loop extending stem-loop 2, and facultative stem-loop 3. The maximum observed length of the additional and facultative stem-loops are 180 and 102 nucleotides, respectively. The presence of these two stem-loops and their lengths do not seem to be correlated with function or phylogeny of *THI* elements, genes, or genomes. The loop between the first and fourth helices is also variable and its length usually equals one or two nucleotides. The maximum length of this loop, 22 nt, was observed in the *THI* element upstream of the *Vibrio cholerae* operon *thiBPQ.* Other internal and hairpin loops of the *THI* element are highly conserved. The only exception is the 36-bp hairpin loop in stem-loop 5 of the *THI* element upstream of the *Pseudomonas aeruginosa* gene *thiC*. Notably, all unusually long additional loops of the *THI* element can form a stable secondary structure.

Recent experiments (14) demonstrated that thiamin-mediated regulation of the TBS operon of *R. etli* involves the *thi* box and the region immediately upstream of the first gene in the operon, *thiC*. This region can form a hairpin that would sequester the ribosome-binding site. Here we use the comparative analysis of nucleotide sequences downstream of 170 *THI* elements and an analogy with the previously proposed regulatory mechanism for the riboflavin regulon (16), and we propose a possible mechanism for *THI*-mediated regulation of genes of the thiamin biosynthesis and transport (Fig. 4).

Downstream of all *THI* elements, except those of actinomycetes, cyanobacteria, and thermoplasmas, there are potential hairpins that are either followed by runs of thymines (and thus are candidate Rho-independent terminators of transcription) or overlap the SD box of the first gene in the regulated operon (and thus are candidate sequestors, which prevent the ribosome binding to the SD sequence). In addition, we have found complementary fragments of RNA sequences that partially overlap both the proposed regulatory hairpin (terminator or sequestor) and one of conserved helices in the *THI* element (see Supplemental Fig. 5). Furthermore, these complementary fragments always form the base stem of a new, more stable alternative secondary structure with ΔG smaller than ΔG of the *THI* element. We predict that this structure functions as an antiterminator/antisequestor, alternative to both the *THI* element and the terminator/sequestor hairpin. Thus, two different types of regulation by competing of alternative RNA structures are suggested, attenuation of transcription via an antitermination mechanism and attenuation of translation by sequestering of the Shine-Dalgarno box.

Most Gram-positive bacteria from the *Bacillus/Clostridium* group, Thermotogales, *F. nucleatum*, and *C. aurantiacus* are predicted to have a terminator hairpin, whereas most Gram-negative bacteria (proteobacteria and the CFB group), bacteria from the *Thermus/Deinococcus* group, and *Chlorobium tepidum* have SD-sequestering hairpins downstream of the *THI* elements (Table II). The phylogenetic distribution of the proposed terminators and sequestors in Gram-positive and Gram-negative bacteria, respectively, is similar to the previously observed distribution of the regulatory hairpins for the riboflavin regulon (16). Analysis of the 5′-noncoding RNA regions of the *yuaJ* genes reveals two possibilities for the regulation. In most cases, the predicted terminator hairpin overlaps the SD-box of the *yuaJ* gene. Therefore, this hairpin can function both as a terminator and a sequestor. Again, this is reminiscent of the dual action candidate regulatory hairpins upstream of riboflavin transporter genes *ypaA* (16).

Most *THI* elements in actinomycetes, cyanobacteria, and thermoplasmas overlap the SD-boxes directly (see Supplemental Fig. 5). We predict that in these bacteria, the *THI* element regulates translation without additional RNA elements. In the presence of thiamin pyrophosphate, the stabilized *THI* element represses initiation of translation. Conversely, *THI* element is not stable in the absence of thiamin pyrophosphate that results in opening of the SD-box and release from repression. It is interesting that in most cases of the direct SD sequestering, the base stem of the *THI* element is absent.

The left part of the base stem of the predicted antiterminator (or antisequestor) overlaps either the base stem or a part of the conserved *thi* box sequence of the *THI* element. At that, the antiterminator structure can either partially or completely include the *THI* element helices. When the spacer between *THI* element and terminator (or sequestor) is long enough, it can potentially fold into an additional secondary structure. In particular, in the case of overlap with the left part of the base *THI*-element stem, the antiterminator is formed by this spacer and intact *THI* helices.

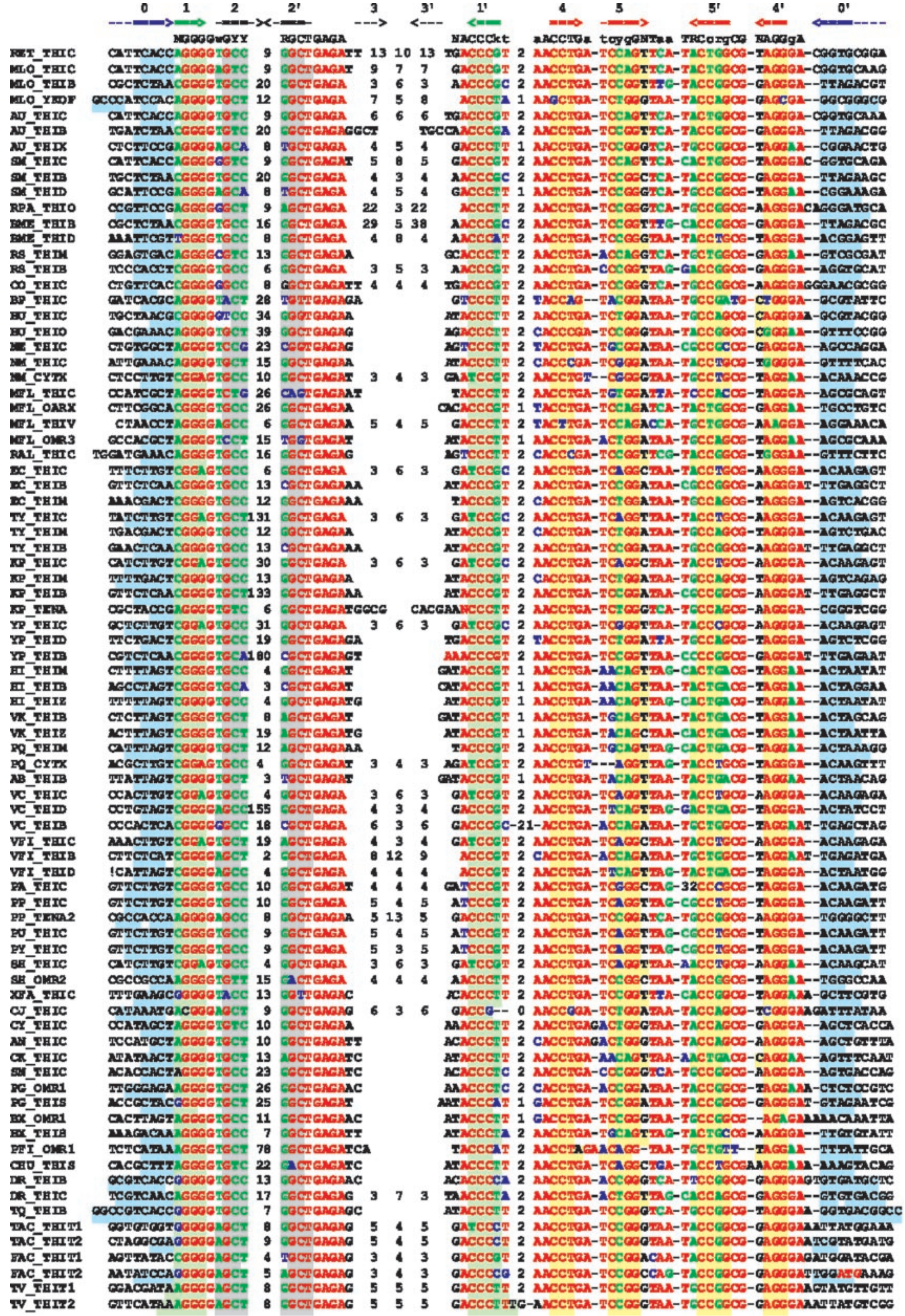


FIG. 2. **Alignment of the *THI* element sequences.** The *first column* contains the genome abbreviations and the names of the proximal downstream genes (see Table II). The complementary stems of the RNA secondary structure are shown by *arrows* in the *upper line*. Base-paired positions are highlighted. Conserved positions are set in *red*, degenerate conserved positions in *green*, nonconserved positions in *black*, and nonconsensus nucleotides in conserved positions in *blue*. The lengths of additional and facultative stem-loops are given.

DISCUSSION

Identification of the TBS genes and new *THI*-regulated genes allows us to reconstruct and compare the TBS pathway in various organisms (see Supplemental Table III). The most conserved part of the TBS pathway in bacteria, archaea, and eukaryota is the synthesis of thiamin monophosphate from HET and HMP moieties involving the ThiD, ThiE, and ThiM proteins. In contrast, the HET and HMP biosynthesis are the

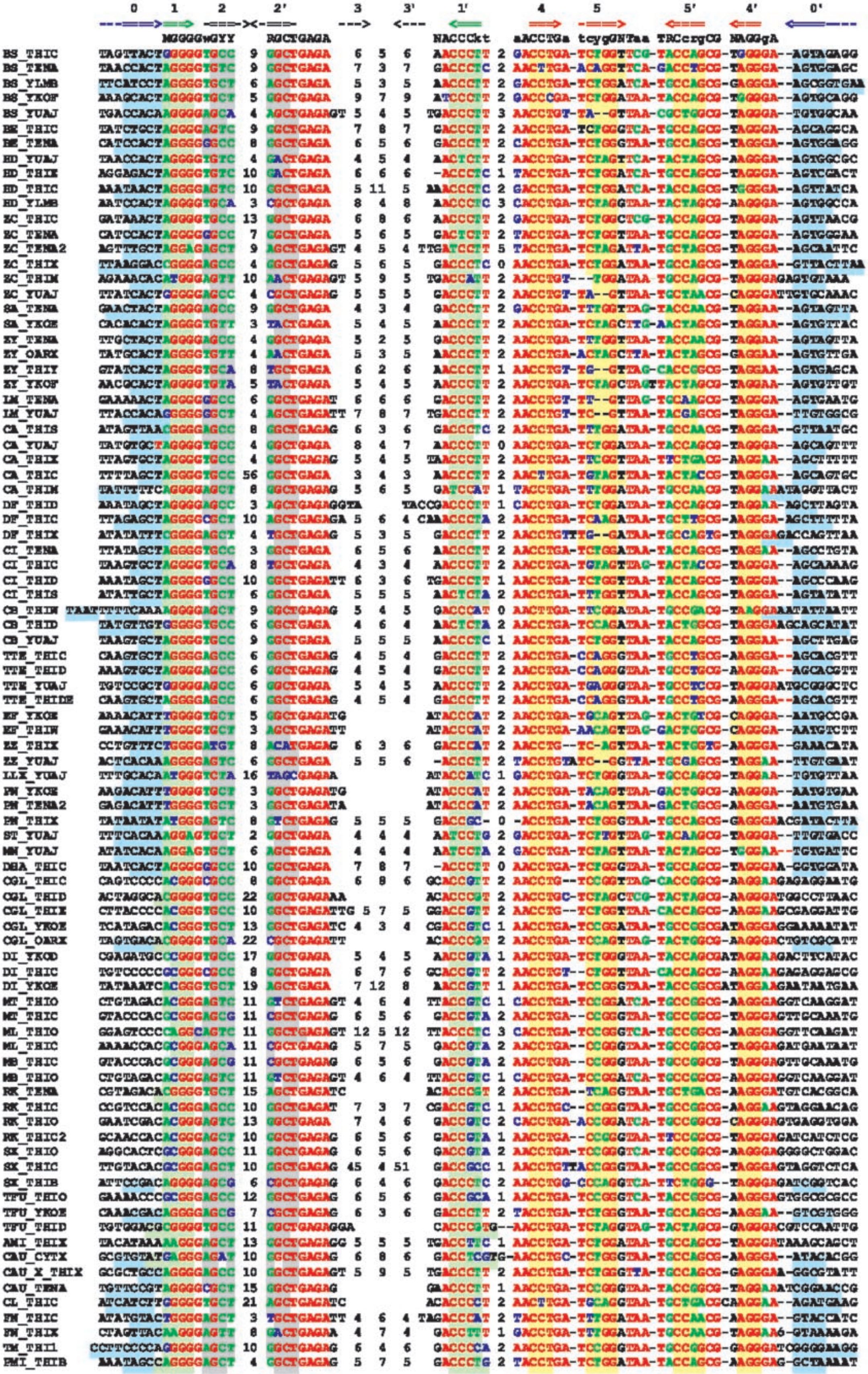


FIG. 2—*continued*

most variable part of the TBS pathway in prokaryotes.

The incomplete TBS pathway, which includes the ThiD, ThiE, and ThiM proteins only, was found in Gram-positive pathogens from the *Bacillus*/*Clostridium* group and in some Gram-negative pathogens, namely Pasteurellaceae and *H. pylori*. These eubacteria are unable to synthesize HET and HMP and are forced to uptake these thiamin precursors via specific transport system. Using analysis of operon structures and *THI*-mediated regulation, we have identified several candidate thiamin-related transporters. We predict that two groups of probable transporters, namely ThiX-ThiY-ThiZ/YkoE-YkoD-YkoC/CytX and ThiU/ThiW/Orf11, may be involved in the HMP and HET uptake, respectively, substituting the missing biosynthetic pathways in bacteria. The remaining puzzle is the presence of the *thiM-thiD-thiE* genes in the complete genomes of *L. lactis*, *Listeria monocytogenes*, and *H. pylori*, that have no candidate HMP or HET transporters. The thiamin biosynthetic genes seem to be not essential in these genomes, since they contain candidate thiamin transporters, *yuaJ* and *pnuT*. Thus, the possible explanations are as follows: existence of unidentified HMP and HET transporters; erroneous assignment of thiamin, but not HMP and/or HET specificity, to YuaJ and PnuT; and cryptic (nonfunctional) state of *thiM-thiD-thiE* in these organisms.

The HET biosynthesis in eubacteria uses the ThiF, ThiS, ThiG, and ThiH (ThiO) proteins. One interesting exception is *T. maritima*, whose complete genome revealed genes of HMP, but not HET, biosynthesis. The first gene of the *THI*-regulated TBS operon in *T. maritima* encodes a protein from the Thi4 family of eukaryotic enzymes involved in the thiazole biosynthesis. This family includes Thi4 from *S. cerevisiae*, Thi2 from *S. pombe*, Thi1 from *Zea mays*, and STI35 from *Fusarium* sp. A similarity search shows that the *thi4* gene is present in all available archaeal genomes (Table II). However, among eubacteria, only *T. maritima* has the *thi4* gene. Therefore, we concluded that the HET biosynthesis of archaea, eukaryota, and *T. maritima* differs from that in most eubacteria and uses the *thi4* gene product.

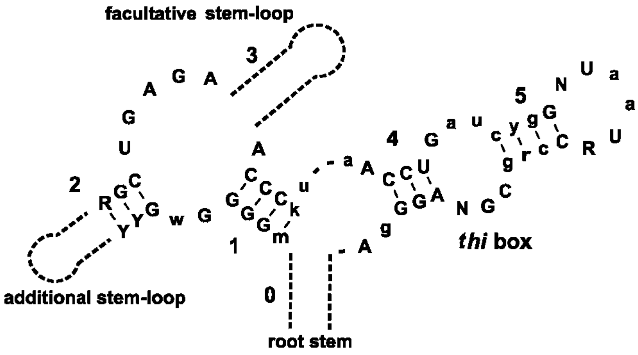


FIG. 3. **The conserved structure of the *THI* element.** *Capital letters* indicate invariant positions. *Lowercase letters* indicate strongly conserved positions. Degenerate positions are as follows: *R*, A or G; *Y*, C or U; *K*, G or U; *M*, A or C; *N*, any nucleotide.

The bacterial TBS pathway uses the ThiC protein for the HMP biosynthesis. The ThiC orthologs were identified in all archaeal genomes, except *A. pernix* and *Thermoplasma* species (Table II). The phylogenetic distribution of ThiC is restricted to bacteria and archaea. The HMP biosynthesis in eukaryota uses other proteins that are not similar to ThiC and belong to the NMT1 family. This family includes Thi5 from *S. cerevisiae*, Thi3 from *S. pombe*, and NMT1 from *Aspergillus parasiticus.* As mentioned above, the substrate-binding component of the predicted HMP transport system ThiY from various bacteria is highly similar to the proteins from the NMT1 family. The main difference between the ThiY and NMT1 proteins is the absence of the N-terminal transmembrane segment in the latter. Strikingly, the first gene of the TBS operon in *L. pneumophila*, a pathogenic γ-proteobacterium, is not *thiC*, as in most γ-proteobacteria, but a gene encoding an NMT1 family protein. This protein has no predicted transmembrane segments and is strongly linked to the eukaryotic NMT1 proteins in the phylogenetic tree of the NMT1/ThiY proteins. Thus, in contrast to other bacteria, the HMP biosynthesis in *L. pneumophila* is similar to the eukaryotic pathway.

Analysis of phylogenetic patterns results in both strong and weak functional predictions for the TBS genes. The former involves nonorthologous displacements within the HET and HMP biosynthetic pathways. In contrast, preliminary prediction of the possible nonorthologous replacement of ThiE to ThiN in archaea is tentative.

Thus, based on comparative and phylogenetic analyses, we have shown key differences in the initial steps of the TBS pathway in eubacteria, archaea, and eukaryota. Moreover, the predicted HMP and HET transporters complement for the absence of corresponding biosynthetic pathways of the TBS in bacteria.

Using the global analysis of the *THI* elements in available bacterial genomes, we have found that this conserved RNA regulatory element is widely distributed in eubacteria and regulates most TBS genes. In contrast, among 17 available archaeal genomes, *THI* elements could be observed only upstream of newly identified thiamin-related transporters in *Thermoplasma* species. Among all bacterial TBS genes, only the *thiC* gene is always *THI*-regulated. The only two exceptions are two bacteria, *Magnetococcus* and *A. aeolicus*, that have no *THI* elements at all. Most thiamin-related transport systems, both known and predicted, are also regulated by *THI* elements. Interestingly, some genes required for the HET biosynthesis, namely *iscS*, *dxs*, and *thiI*, are never regulated by the *THI*

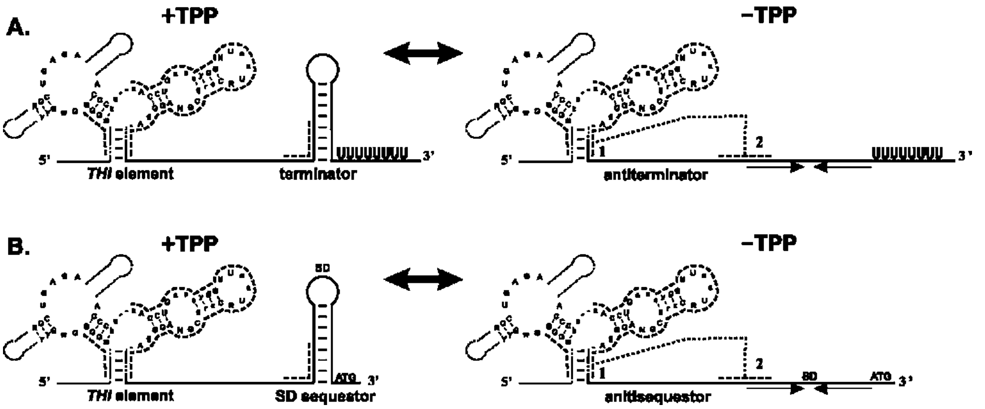


FIG. 4. **The predicted mechanism of the *THI*-mediated regulation of thiamin genes.** *A*, transcription attenuation; *B*, translation attenuation. *Dashed lines* show the location of complementary regions. *Point lines* show interactions in derepressed conditions. *SD*, the Shine-Dalgarno box; *ATG*, start codon; *UUUU*, poly(U) tract in the terminator.

element and are never positionally clustered with other thiamin biosynthetic genes. More surprisingly, thiamin-monophosphate kinase *thiL* is not clustered with other *thi* genes and is not regulated by the *THI* element.

The proposed model for the thiamin regulation is based on competition between alternative RNA secondary structures. In the repressing conditions, thiamin pyrophosphate stabilizes the *THI* element. In Gram-positive bacteria, it leads to formation of the terminator hairpin and premature termination of transcription. Without thiamin pyrophosphate, unstable *THI* element is replaced by the more energy-favorable antitermination conformation that allows for the transcription read-through. In Gram-negative bacteria, stabilization of the *THI* element leads to formation of the SD sequestor hairpin, which represses initiation of translation. In the derepressing conditions, *THI* element is replaced by the antisequestor that releases the SD-box and allows for initiation of translation.

It is known that various nucleotides, such as flavin mononucleotide or nicotinamide mononucleotide, can specifically bind to RNA aptamers (36). Thiamin pyrophosphate, containing pyrimidine and thiazole moieties, is a regulatory molecule for the regulation of expression of the thiamin biosynthetic and transport genes (3, 8, 9). The *THI* element was previously shown to be absolutely necessary for the high level expression of the TBS operon in *R. etli*, but the exact mechanism of the thiamin-mediated regulation was not clear (14). Our analysis shows that this regulation apparently requires high conservation of the sequence and structure of the *THI* element due to possible direct binding of thiamin pyrophosphate to this site.

The proposed mechanisms of regulation for the thiamin and riboflavin regulons, which are mediated by conserved RNA structural elements (the *THI* element and the *RFN* element, respectively), show striking similarities to each other. First of all, the secondary structures of these elements are strictly conserved, often contain complementary substitutions, and have a similar tree-like topology with one central base stem. Second, the nucleotide sequences of the *THI* and *RFN* elements contain a large number of invariant positions that can be involved in the binding of effectors, thiamin pyrophosphate and FMN, respectively. Third, the regulation probably involves the same mechanisms of either transcriptional or translational attenuation, which are based on the competition of alternative RNA secondary structures, terminator/antiterminator or sequestor/antisequestor, respectively. Finally, the phylogenetic distribution of regulatory hairpins is the same for both regulons, with terminators and sequestors occurring in Gram-positive and Gram-negative bacteria, respectively.

From the practical point, this study once again demonstrates the power of comparative genomics for functional annotation of genomes, especially when experimental data are limited. In particular, analysis of regulatory elements is a powerful tool for prediction of missing transport genes, as demonstrated here and in our analyses of the riboflavin and biotin regulons (16, 17).

Acknowledgments—We are grateful to Andrei Osterman for attention, advice, and encouragement and to Tadhg Begley, Pieter Dorrestein, Iain Anders, and Olga Vassieva for critical reading of the manuscript and useful discussions. We thank the Institute of Microbiology and Genetics, Georg-August-University (Göttingen, Germany) for use of the *T. thermophilus* genomic sequence for this analysis.

REFERENCES

1. Schowen, R. L. (1998) in *Comprehensive Biological Catalysis*, Vol. 2 (Sinnott, M. L., ed) pp. 217–266, Academic Press, London
2. Begley, T. P., Downs, D. M., Ealick, S. E., McLafferty, F. W., Van Loon, A. P., Taylor, S., Campobasso, N., Chiu, H. J., Kinsland, C., Reddick, J. J., and Xi, J. (1999) *Arch. Microbiol.* **171,** 293–300
3. Webb, E., Claas, K., and Downs, D. M. (1997) *J. Bacteriol.* **179,** 4399–4402
4. Perkins, J. B., and Pero, J. G. (2001) *Bacillus subtilis and Its Relatives: From Genes to Cells* (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 279–293, American Society for Microbiology, Washington, D. C.
5. Miranda-Rios, J., Morera, C., Taboada, H., Davalos, A., Encarnacion, S., Mora, J., and Soberon, M. (1997) *J. Bacteriol.* **179,** 6887–6893
6. Petersen, L. A., and Downs, D. M. (1997) *J. Bacteriol.* **179,** 4894–4900
7. Webb, E., and Downs, D. (1997) *J. Biol. Chem.* **272,** 15702–15707
8. Webb, E., Claas, K., and Downs, D. (1998) *J. Biol. Chem.* **273,** 8946–8950
9. Webb, E., Febres, F., and Downs, D. M. (1996) *J. Bacteriol.* **178,** 2533–2538
10. Zhang, Y., and Begley, T. P. (1997) *Gene* (*Amst.*) **198,** 73–82
11. Zhang, Y., Taylor, S. V., Chiu, H. J., and Begley, T. P. (1997) *J. Bacteriol.* **179,** 3030–3035
12. Lee, J. M., Zhang, S., Saha, S., Santa Anna, S., Jiang, C., and Perkins, J. (2001) *J. Bacteriol.* **183,** 7371–7380
13. Pang, A. S., Nathoo, S., and Wong, S. L. (1991) *J. Bacteriol.* **173,** 46–54
14. Miranda-Rios, J., Navarro, M., and Soberon, M. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 9736–9741
15. Gelfand, M. S., Novichkov, P. S., Novichkova, E. S., and Mironov, A. A. (2000) *Brief. Bioinform.* **1,** 357–371
16. Vitreshchak, A., Rodionov, D., Mironov, A. A., and Gelfand, M. S. (2002) *Nucleic Acids Res.* **30,** 3141–3151
17. Rodionov, D., Mironov, A. A., and Gelfand, M. S. (2002) *Genome Res.* **12,** 1507
18. Eddy, S. R., and Durbin, R. (1994) *Nucleic Acids Res.* **22,** 2079–2088
19. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 2896–2901
20. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000) *Nucleic Acids Res.* **28,** 15–18
21. Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E. Jr., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. (2000) *Nucleic Acids Res.* **28,** 123–125
22. Vitreshchak, A., Mironov, A. A., and Gelfand, M. S. (2001) *Proceedings of the 3rd International Conference on "Complex Systems: Control and Modeling Problems," Samara, Russia, September 4–9, 2001*, pp. 623–625, The Institute of Control of Complex Systems, Samara, Russia
23. Lyngso, R. B., Zuker, M., and Pedersen, C. N. (1999) *Bioinformatics* **15,** 440–445
24. Mironov, A. A., Vinokurova, N. P., and Gelfand, M. S. (2000) *Mol. Biol.* **34,** 222–231
25. Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000) *Nucleic Acids Res.* **28,** 33–36
26. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376
27. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) *Nucleic Acids Res.* **25,** 3389–3402
28. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882
29. Taylor, S. V., Kelleher, N. L., Kinsland, C., Chiu, H. J., Costello, C. A., Backstrom, A. D., McLafferty, F. W., and Begley, T. P. (1998) *J. Biol. Chem.* **273,** 16555–16560
30. Tazuya, K., Morisaki, M., Yamada, K., Kumaoka, H., and Saiki, K. (1987) *Biochem. Int.* **14,** 153–160
31. Toburen-Bots, I., and Hagedorn, H. (1977) *Arch. Microbiol.* **113,** 23–31
32. Hekstra, D., and Tommassen, J. (1993) *J. Bacteriol.* **175,** 6546–6552
33. Llorente, B., Fairhead, C., and Dujon, B. (1999) *Mol. Microbiol.* **32,** 1140–1152
34. Moeck, G. S., and Coulton, J. W. (1998) *Mol. Microbiol.* **28,** 675–681
35. Kurnasov, O., Polanuyer, B., Ananta, S., Sloutsky, R., Tam, A., Gerdes, S., and Osterman, A. (2002), in press
36. Hermann, T., and Patel, D. J. (2000) *Science* **287,** 820–825