# **Comparative Genomics of Thiamin Biosynthesis in Procaryotes**

NEW GENES AND REGULATORY MECHANISMS\*

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

# Dmitry A. Rodionov<sup>‡</sup>§, Alexey G. Vitreschak<sup>¶</sup>, Andrey A. Mironov<sup>‡</sup>||, and Mikhail S. Gelfand<sup>‡</sup>||

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)<sup>1</sup> pathway of bacteria is outlined in Fig. 1. Thiamin monophosphate is formed by coupling of two inde-

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

pendently 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 adenvlation 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-

<sup>\*</sup> 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.

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

<sup>&</sup>lt;sup>1</sup> The abbreviations used are: TBS, thiamin biosynthesis; HMP, hydroxymethylpyrimidine; HET, hydroxyethylthiazole; TMS, transmembrane segment; SD, Shine-Dalgarno.

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<sup>TM</sup> (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 Gen-Bank<sup>TM</sup> 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  $\sim 150$  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.<sup>2</sup> 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

## Comparative Genomics of Thiamin Biosynthesis

The thiamin biosynthetic genes of E. coli (EC) and their counterparts in B. subtilis (BS)										
BS	Function	Similarity								
		%								
thiA	HMP biosynthesis	76								
yjbV	Phosphomethylpyrimidine kinase	43								
yjbT	Thiazole biosynthesis protein ThiG	52								
	Thiazole biosynthesis protein ThiH									
yjbR	Thiazole biosynthesis protein ThiO									
ytbJ	Thiazole biosynthesis protein ThiI	33								
yjbU	Adenylyltransferase	37								
yjbS	Sulfur carrier protein ThiS	31								
thiK	Hydroxyethylthiazole kinase	43								
thiC	Thiamin-phosphate synthase	39								
ydiA	Thiamin-monophosphate kinase	37								
	The thiamin biosynthetic BS thiA yjbV yjbT yjbR ytbJ yjbU yjbS thiK thiC ydiA	The thiamin biosynthetic genes of E. coli (EC) and their counterparts in B. subtilis (BS)BSFunctionthiAHMP biosynthesis $yjbV$ Phosphomethylpyrimidine kinase $yjbT$ Thiazole biosynthesis protein ThiG $yjbR$ Thiazole biosynthesis protein ThiH $yjbR$ Thiazole biosynthesis protein ThiO $ytbJ$ Thiazole biosynthesis protein ThiI $yjbV$ Adenylyltransferase $yjbS$ Sulfur carrier protein ThiS $thiK$ Hydroxyethylthiazole kinase $thiC$ Thiamin-phosphate synthase $ydiA$ Thiamin-monophosphate kinase								

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

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 biosynthesislinked 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 analvsis 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  $\alpha$ - and  $\beta$ -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  $\sim$ 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, Magneto-coccus 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

#### Comparative Genomics of Thiamin Biosynthesis

#### 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- $\alpha$ , - $\beta$ , - $\gamma$ , - $\varepsilon$ , Cyan, CFB, T/D, B/C, Actin, Ther, and A in the "Tax" column represent  $\alpha$ -,  $\beta$ -,  $\gamma$ ,  $\varepsilon$ -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 bicsynthetic genes	Other thiamine-related genes
P-	Mesorhizobium loti	MLO	&thiC-thiO-thiG-thiS-thiE-thiD	&thiB-thiP-thiQ / &ykoF-thiX-thiZ-thiY
	Agrobacterium turnefaciens	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 / hiD	
	Brucella melitensis	BME	&thID-thIO-thIS-thIG-thIE-thIY-tenA-thIX	&thiB-thiP-thiQ
	Rhodobacter sphaeroides #	RS	8thiM-thiE-thiD-thiY-thiZ-thiX	&thiB-thiP-thiQ
	Caulobacter crescentus	CO	&thiC / thiS-thiG / thiE / thiD	
P-	Bordetella pertussis	BP	8thiC / thiG-thiD / thiE / thiO / thiS	
	Burkholderia cepacia #	BU	SthiC / Sthio-this-thig-thiE / X-thip	
	Nitrosomonas europaea	NE	&thiC / thiD-thiE / thiS-thiG / thiO	
	Neisserie meningitidis	NM	SthiC / &cytX-thiO-thiE=thiS-thiG / thiD	thiB
	Methylobaclilus flagellatus #	MFL	EthIC / 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	8thiC-thiE-thiE-thiS-thiG-thiH / 8thiM-thiD	&thiB-thiP-thiQ / &tenA-thiZ-thiX-thiY
	Yersinia pestis	YP	&thiC-thiE-thiE-thiS-thiG-thiH / &thiD	&thiB-thiP-thiQ
	Haemophilus influenzae	н	&thiM-thiD-thiE-thiU	&thiB-thiP-thiQ / &thiZ-thiX-thiY-tenA
	Pasteurella multocida	VK	8th/Z-th/X-tenA-th/Y=th/M-th/D-th/E-th/U	&thiB-thiP-thiQ
	Mannheimia haemolytica #	PQ	8thiM-thiU=8(thiD-thiE)-cvtX	
	Actinobacillus actinomvostemcomitans #	AB		&thiB-thiP-thiQ
	Vibrio cholerae	VC	&thiC-thiE-thiE-thiS-thiG-thiH / &thiD	&thiB-thiP-thiQ
	Vibrio fischeri #	VEI	8thiC-thiE1-thiF-thiS-thiG-thiH / 8thiD-thiZ-thiX-thiY-tenA-thiM-thiE2	&thiB-thiP-thiQ
	Pseudomonas aeruginosa	PA	&thiC / thiD-thiE / thiS-thiG / X-thiO	
	Pseudomonas putida	PP	8thiC-cvtX / thiD-thiE / thiS-thiG / X-thiO	StenA2-tenA1
	Pseudomonas fluorescens#. P.svringiae #	PU.PY	&thiC-cvtX / thiD-thiE /thiS-thiG / X-thiO	and the second
	Shewanella putrefaciens	SH	&th/C-th/D-th/E-th/F-th/S-th/G-th/H	&omr2
	Xvlella fastidiosa	XFA	&thiC / thiD / thiS-thiG / thiE	-
	Legionelle pneumophila #	LP	nmt1-thiO-thiS-thiG-(thiD-thiE)-thiF	
P-	Helicobacter pylori	HP	X-thiM-thiD-thiE	tenA <> X-pnuT-tnr3
	Campylobacter lelun!	CJ	8th/C / X-thiD-thiE1 / X-th/S-th/F-th/G-th/H-th/E2	tenA
	Magnetococcus #	MCO	thiC / thiD / thiE / thiS-thiG / X-thiO-X	
Cvan	Anabaena sp.	AN	SthiC / thiE-thiS / (thiO-thiG) / thiD	tenA
	Prochlorococcus marinus	CK	SthiC / thiE-thiS / thiO / thiG / tanA-thiD	
	Synechocystissp., Synechococcussp.	CY. SN	8thiC / thiE-thiS / (thiO-thiG)	
<b>ČFB</b>	Porphyromonas gingivalis #	PG	&th/S-th/C-(th/E2-th/D-th/E1)-th/G-th/H	& 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< td=""><td></td></thie2<>	
T/D	Delnococcus radiodurans	DR	SthiC-thiE-thiS-thiG-thiD / X-thiO	SthiB / 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	ShiC / 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* elementregulated 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  $\alpha$ - and  $\gamma$ -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 thiBthiP-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  $\gamma$ -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	Bacilius subtilis	BS	&thIC / &tenA-thIE2-thIO-thIS-thIG-thIF-thID / ywb1-thIM-thIE1	&ykoF-ykoE-ykoD-ykoC / &yuaJ / &yimB			
	Bacillus cereus	ZC	&thiC / &tenA1-thiX1-thiY1-thiZ1-thiE2-thiO-thiS-thiG-thiF-thiD / &thiM-thiE1	&thiX2-thiY2-thiZ2 / &tenA2 / &yuaJ / ylmB			
	Bacilius halodurans	HD	&thiC / &thiE-thiS-thiG-thiO-thiD / thiM-X	&yimB-tenA-thiZ-thiX-thiY /&yuaJ			
	Bacillus steerothermophilus #	BE	&thiC / &lanA-ykoE-ykoD-ykoC-thiE2-thiO-thiS-thiG-thiF / thiM-thiD-thiE1				
	Staphylococcus aureus	SA	&ienA-thiD-thiM-thiE-orf11	&ykoE-ykoD-ykoC			
	Staphylococcus epidermidis #	ZY	&tenA-thID-thIM-thIE-orf11	&ykoE-ykoD-ykoC / &th/Y / &oarX			
	Listeria monocytogenes	LM	&tenA-thiM-thiD-thiE	&yuaJ			
	Clostridium acetobutylicum	CA	&thiC / thiE1 / &thiM-thiD / &thiS-thiF-thiG-thiH-thiE2	&thiX-thiY-thiZ / &vuaJ			
	Clostridium perfringes	CI	&thiC / &thiD-thiM-thiE1 / &thiS-thiF-thiG-thiH-thiE2	StenA-vua.11 / vue.12			
	Clostridium botuilnum #	CB	thiC / &thiD-thiM1-thiE / &thiW-thiM2	vkoE-vkoD-vkoC / EvuaJ			
	Clostridium difficile #	DF	&thiC-thiS-thiF-thiG-thH-thiE2 / &thiD-thiM-thiE1	&thiX-thiY-thiZ			
	Thermoanserobacter tenacongensis	TTE	&thiC / &(thiD1-thiE1)-thiW-cvtX-thiM1 / &thiD2-oerX-thiE2-thiM2	&vuaJ			
	Enterococcus faecalis	EF	&thiW-thiM-thiE-thiD	&vkoE-vkoD-vkoC-tenA / vuaJ			
	Enterococcus faecium #	ZZ		&thiX-thiY-thiZ / Roug.			
	Lactococcus lactis	LLX	thiM-thiD-thiE	SvuaJ / tenA			
	Streptococcus pneumoniae	PN	&tenA1=thiM1-thiE1-&vkoE-vkoC-tenA2-thiW-thiM2-thiE2 >< thiD	RthiX-thiX-thiZ			
	Strentococcus nyogenes, S mutans	ST MN	10	Rough			
	Desulfitobacterium baliniense #	DHA	&thiC-thiM-thiE-X-thiD	Tth/Y-th/Z			
Actin	Corvnebacterium alutamicum	CGL	Rihic / RihiE-IniO-IniS-IniG-IniF / RihiM-(IniE-IniD-IenA)	AvkoE-vkoD-vkoC / BoarX			
- Notali	Corvnebacterium diphtheriae	DI	&thiC-thiE-thiO-thiS-thiG-thiF-thiD	&vkoE / &vkoD-vkoC			
	Mycobacterium spp.	MT. MB. ML	SthiC-thiD /thiE / SthiO-thiS-thiG				
	Rhodococcus str. #	RK	SthiC-thiD /thiE<>SthiO-thiS-thiG / B-thiC	BhiX / &tenA			
	Streptomyces coelicolor	SX	Rthic / hiF-X <rthio-this-thio-x td="" thil-thip<=""><td colspan="3">&amp;thiB-thiP-thiQ</td></rthio-this-thio-x>	&thiB-thiP-thiQ			
	Thermomorporp fusce #	TEU	thiE-X <> 8thiO-this-thic-X / 18thiD	SykaE-ykaD-ykaC / tanA			
	Atoonhim minutem #	AMI		RihiX-thiZ-thi7			
Ther	Thermologe medime	TM	Ribid_th(C_Y_(thiD_thR))	thiy_thiy_thiz			
	Petrotore miotherne #	PMI		2thip.thip			
_	Chlomflevus aurantiacus #	CALL	7/60	RihiX-thiZ / RoytX 1 / Riand			
_	Europectertum puckeatum	EN	RthD th/51.th/C.th/S.th/C.th/A.th/F2	Shly hly hly			
Arch	Thermoniasma son	TAC TVEAC	X-thiE / thin	Schirt / Smirz / tone			
AUGIN	Methanosarcine bakeri M mazei	MRA MMZ	thic1/thic2/thiAthiE / thiD_thiAD / thid	autil / autil / tene			
	Holobactarian en			HIP HIP HIP			
	Helefamy unleanli #	NO		ton A 1 ton A 2 thill			
	Archaeoglobus fulgidus	AG		toria regnazare			
	Accession and a second s	AD		fond			
	Aeropyrum pernix	AP NIL NIL		torus.			
	Methanococcus jannaschii, M. Kanolen						
	Memanooacterium	DK	(1)(1)/(1)(2-X / X-(1)(1)/(1)(1)/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1)(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X-(1))/(X	thip thips thip / tond Y			
	Pyrobaculum aerophilum	PK					
	Pyrococcus spp.	PF, PH, PO	thiC / tenA1-tenA2-cytX-thiM-thiE-(thiD-thiN) / thi4 /	X-INIP-INIQ> <inib< td=""></inib<>			
	Suitolobus soffataricus	510	(niG1-X /(niG2-X/X-(miD-thiN)1-X/X-(miD-thiN)2-X/hiN1/thiN2 / thi4	IONAT-X /X-IONAZ-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  $\alpha$ - and  $\gamma$ -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 substratebinding 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, cytXeither 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. in-fluenzae* 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 thiaminrelated 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 thiDand 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-ykoDykoC), 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 omr1pnuT-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  $\Delta G$  smaller than  $\Delta G$  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 Gramnegative 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 Gramnegative 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 SDbox 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.

	0	1	2	202	2'	3		31	1'		4	5	5'	4'	0'
		>==>		·×·		>		<	- <			>	<	<b>~</b>	<
		Magag	WOYY		RECTORGA	13	10	1 1 2	NACCORE		ACCTGA	toygentaa	TRCorgCo	NAGOGA	
MLO THIC	CATTON	CLOGOG	ACTC		GOTGAGAT	6	7	7	GACCOGT	2	AACCTOR	TCCAGTTCA	TACTOGOG	TACCOL	COGTOCAAG
MLO THIB	OCCTCTA	ACCECC	TCCC	20	CCTCACA	3	6	3	AACCCCC	2	AACCTGA	TOCOCTTE	TACCOGOG	CACCCA-	TTAGACGT
MLO YROF	GCCCATCCA	C10000	IGCT	12	GCTGAGA	7	5	8	ACCCTA	1	ANGCIGA	TCTOGOTAA-	TACCAGO	GAGCGA-	-agcaagca
AU_THIC	CATTCAC	CIGOGG	IGTC	9	GOTGAGA	6	6	6	TGACCCOT	2	AACCTGA	TCCAGTTCA-	TACTOOCO-	TAGEGA	COGTOCAAL
AU_TEIB	TCATCTA	ACCOCC	ICIC	20	CCTCACAGC	CI		TC	CCAACCCCA	2	AACCTGA	TCEOGTTCA-	TACCEGOG-	GAGGGA-	TTACACGC
AU_THIX	CTCTTCC	616000	AGCA	8	TGCTGAGA	4	5	4	GACCCTT	1	AACCTGA	TCCGGGTCA	TOCCODCO	TAGGAL	-CGGRACTG
SM THIC	CATTCAL	CAGGGG	GGTC	20	GCTGAGAT	5		5	GACCCOT	2	AACCTGA	TCCAGTTCA	-CACTEGOG-	TAGEGAC	GGTGCAGA
SM THID	GCATTO	GIGGGG	AGCA	8	TGCTGAGA	2	5	4	GACCOLT	1	AACCTGA	TCCGGGTCA	TECCEGOG	TAGGAL	-CGGAAAGA
RPA THIO	CCGTTCC	GAGGGG	GGCT	9	AGCTGAGA	22	3	22	ACCCTT	2	AACCTGA	TCCGGGTCA	TECCEGOG	ALGEGAC	AGGGATGCA
BME THIB	OGCTCTA	ACGOGG	IGCC	16	GCTGAGA	29	5	38	AACCOSC	2	AACCTGA	TCCGGTTTG-	-CACCEGOG-	GAGGGA-	TTAGACGC
BME THID	ALATTOS	TIGGGG	ICCC	8	GCTGAGA	4	8	4	ANCCOLT	2	AACCTGA	TCCGGGTAA	TACCIGOG	TAGGGA-	-ACGGAGTT
RS_THIM	GGAGTGA	CIGGGG	OGTC	13	CGCTGAGAA		-		GCACCCTT	2	AACCTGA	ACCAGGTCA-	TECTEGOG	GAGGA1-	-GTOGOGAT
AS THIS	TECEACE	TCGGGG	TGCC		GCTGAGA		2	3	AACCCGT	2 2	AACCIGA	CCCGGTTAG	GACCGGGG	GAGEGA	AGGTGCAT
BP THIC	GATCACZ	C10000	TACT	28	TUTTUAGAGAGA		•	-	GTCCCTT	2	TACCAG-	TACGGATAA	TECCORTE	CTORGA	-CCTATTC
BU THIC	TOCTARC	GCGGGG	GTCC	34	GOTGAGAA				ATACCCTT	2	AACCTGA	TCTOGATAA-	TOCCAGO	CAGGGAZ	-GCGTACGG
BU THIO	GACGRAR	CIGOGG	TOCT	39	GOCTGRGAG				AGACCOTT	2	CACCOGA	TCCOGGTAA	TACCOGCG	CGGGBB-	-GTTTCCGG
THIC	CTCTCCC	TACCCC	ICCG	23	COOTGAGAG				AGTCCCTT	2	TACCTGA	TCCCCATRA-	-CCCCCCCC-	GACCCA-	-AGCCAGGA
MM_THIC	ATTGAAA	C10000	IGCT	15	GCTGAGAA				ATACCCTT	2	CACCCGA	TOGGATAA	TACCIGCO	TOOGGL	-GTTTTCAC
M CYTX	CTCCTTC	TCGEAG	TCCC	10	GCTGAGAT	з	•	з	CAATCOUT	2	AACCTGT		TECCTOCE-	TAGEAL	-ACAAACCG
MFL CARY	CTTOGOC	ACGOGG	TGCC	26	GETGAGA				CACCOUT	1	TACCTGA	TCCBGATCA	TACTOGOG	GAGGA A-	TGOCTETC
WEL THIY	CTANCE	TAGGGG	AGCC	6	GCTGAGAA	5	4	5	GACCOTT	2	TACTTGA	TCCAGACCA	TECTOGOG	ALLGGA-	AGCAAACA
MEL OMR3	GCCACGC	TAGOGG	TCCT	15	TGOTGACA		-		ATACCCT T	1	AACCTGA	ACTOGATAA	TECCAGOG	TAGELL-	AGOGCAAN
RAL_THIC	TEGATGAAA	CIGGGG	ICCC	16	GGCTGAGAG				AGTCCCT T	2	CACCOGA-	TCCGGTTCG-	TACCEGOG	TGGGAL-	-GTITCITC
EC_THIC	TTTCTTG	TCGGAG	ICCC	6	GCTGAGA	3	6	з	GATCOCC	2	AACCTGA	TCAGGCTAA-	TACCIGOG	ALGGGA	-ACAAGAGT
EC_THIB	GTTCTCA	ACGGGG	IGCC	13	CGCTGAGAAA	•			ATACCCCT	2	AACCTGA	TCCGGATAA	CCCCCCCCCC-	ALGEGAT	-TTGAGGCT
EC_THIN	AAACGAC	TCGGGG	IGCC	12	GCTGAGAAA				TACCOT	2	CACCIGA	TCTGGATAA-	TECCAGOG	TAGGGA	-AGTCACGG
TY TEIN	TGACGAC	TCGGGG	TGCC	12	GCTGAGA	3	.0	3	ATACCCOT	2	CACCIGA	TCTGGATAA	TECCIGOG	TAGOGA	-AGTCTGAC
TY TELB	GAACTCA	ACGGGG	TGCC	13	COCTGAGAAA				ATACCCOT	2	AACCTGA	TOCOGATAA	TECCOGOG	AAGGGAT	TTGAGGCT
RP_THIC	CATCTTC	TCGGAG	ICCC	30	CCTCACA.	з	6	3	GATCCCC	2	AACCTGA	TCACCTRA-	TACCIGOG-	AAGCCA-	ACAAGAGT
KP_THIM	TTTTGAC	TCGGGG	IOCC	13	GCTGAGAA				TACCCOT	2	CACCTGA	TCTOGATAA-	TOCCAGO	TAGGGA	-AGTCAGAG
KP_THIB	CTTCTCA	ACGCCCC	ICCT	133	GCTGAGAAA			-	ATACCCCT	2	AACCTGA	TCCGGATRA	-CCCCCCCCC-	AAGEGAT	-TTCACCCT
VD TENA	OCTAC	TCGG G	TACC	31	CTGAGA	3		3	GAANCCOLT	2	BACCING.	TCTGGGTTA	TACCORCE	ANOGGA	-CGGGTCGG
YP TEID	TTCTGAC	TCGGGG	TCCC	19	GCTGAGAGA		•	-	TEACCOLT	2	TACCTGA	TCTOGATTA	TECCAGOG	TAGGA	-AGTCTCGG
YP TEIB	COTCTCA	ACGOGG	TGCA		COCTONOAGT				ARACCOST	2	AACCTGA	TCCOCTTAN	-ccccescs-	GACCONT	TTCACAAT
ET_TE DA	CTTTTAG	TCGGGG	ICCC	4	GCTGAGAT				GATACCCCT	1	AACCTGA	-AACIGTTAG	-CACTGACG-	TAGGAL-	-ACTANTAT
EI_TEIB	AGCCTAG	TCGGGG	IGCA	3	CGCTGAGAT				CATACCCET	1	AACCTGA	-AACAGTTAA-	TACTCACG	TAGGAL-	-ACTAGGAA
W THIS	CTOTTAG	TCGGGG	TOCT	8	GCTGAGATG				ATACCOUT	1	AACCINCA	TOCIOTTAL	TACTOROS	TAGGALA	-ACTACTAT
VK THIZ	ACTITAG	TCGGGG	TGCT	19	ACTGAGATC				ATACCOT	1	AACCTGA	TACAGCTAA	CICTGICG	TAGGAL	ACTANTTA
PO TEIM	CATTING	TCGGGG	TOCT	12	AGCTGAGAAA				TACCCOT	2	AACCTGA	TOCAGTTAG	CACTGACG	TAGGAL-	-ACTAAAGG
PQ_CYTX	LCGCTTG	TCGGAG	IGCC	4	GCTGAGAT	3	4	3	AGATCOGT	2	AACCTGT	AGGTTAG-	TACCIGOG	TAGGGA	ACARGETT
AB_THIB	TTATTAG	TCGGGG	IGCT	3	TOCTORGAL				CATACCCC T	1	AACCTGA	TACAGTTAA-	TACTOACG-	TAGGAA-	-ACTRACAG
VC THIC	OCACTTG	TCGGAG	ICCC	4	GCTGAGA	3	6	3	GATCCOT	2	AACCTGA	TCAGGTTAA	TACCTOCO	AAGGGA	-ACARGAGA
VC TETA	OCCACIO	ACCCCC	ance	18	CECTGACA	-	3	-	GACCOGC	-2	1-ACCTCA	ACCAGATAA	TECTOGOG	TACCAN	TCACTAC
VEI THIC	AAACTTG	TCCCAC	TCCT	19	ACCTGAGA	4	3	4	GATCOLT	2	AACCTGA	TCACCTAA	TACCTOC	AACCCA	ACAAGAGA
VFI THIB	CTTCTCA	TCGGGG	AGCT	2	GCTGAGA	8	12	9	ACCOUT	2	CACCTGA-	ACCAGATAA-	TECTEGOG	TAGGAAT	TGAGATGA
VFI_THID	CATTAG	TCGCGG	ACCC	4	CCTCACA	4	4	4	ACCOUT	2	AACCTGA	TTCAGTTAG	TACTOROG	TACCCA	-ACTAATGG
PA_THIC	GTTCTTG	TCGGGG	ICCC	10	GCTGAGAT	4	4	4	GATCCCCT	2	AACCTGA	TCGGGCTAG	-32CCCGCG-	TACCEA	-ACAAGATG
PP_THIC	GTTCTTG	TCGGGG	IGCC	10	GCTGAGA	5	13	5	ATCCCCT	2	AACCTGA	TCAGGTTAG	TOCCOOC	TAGGGA	-ACAACATT
PU THIC	GTTCTTG	TCGGGG	TGCC	9	GETGAGA	5	4	5	ATCCCCT	2	AACCTGA	TCAGETTAG	CECCTECE	TAGGGA	ACAAGATT
PY THIC	GTTCTTG	TCGGGG	TGCC	9	GCTGAGA	5	3	5	ATCCCCT	2	AACCTGA	TCAGGTTAG	CECCIGOG-	TAGGGA	-ACAAGATT
SE_THIC	CATCHIC	TCGGAG	TOCC	4	GCTGAGA.	з	6	3	GATCCOT	2	AACCTGA	TCAGGTTAA-	ALCCTOCO-	AAGGAL-	-ACAAQCAT
SH_OMR2	OCCOCCC	ALGEGE	IGTT	15	CACTGAGA	4	4	4	AACCCTT	2	AACCTGA	TCC GGCTAA	TACCEGOG	TAGGAL-	-TGGGCCAA
XFA_THIC	TTTGAAG	C00000	TACC	13	GTTGAGAC				CACCCTT	2	AACCTGA	TCCOGTTIA	-CACCGGCG-	TAGGAN	-OCTTCOTO
CY THIC	CATARAT	TACCCC	TCTC	10	CCTCACA	¢	3	9	AAACCCTT	2	AACCUGA	-TCTGGATAA-	TACCAGOG	CACCCA	-ACCTCACCA
AN THIC	TCCATCC	TAGGGG	TOCT	10	GCTGAGATT				ACACCCTT	2	CACCTGA	ACT OG GTAA-	TACCAGOG	AAGEGA	AGCTOTTTA
CK THIC	ATATAAC	TAGOGG	TOCT	13	AGCTGAGATC				ATACCCTT	2	AACCTGA	ANCAGTTAA-	AACTGACG	CAGGAA	AGTTTCAAT
SM_THIC	ACACCAC	TAGGGG	ICCC	23	GCTGAGATC				ACACCCTC	2	AACCTGA	-CCCCCCCTCA-	TECCERCE	AAGGGA-	-AGTGACCAG
PG_OMR1	TTGGGAG	ARCCCC	ICCT	26	CCTCACAAC				AAACCCTC	2	CACCTGA	ACCOGATAA-	TACCOCC	TACCAN	-CTCTCCGTC
PG_THIS	ACCECTA	CCCCCCC	TGCT	25	GCTGAGAT				ANTACCCAT	1	GACCTGA	TCCGGATAA	TACCEGCE	CACCENT	-GTAGAATCG
BA ORRI	CACTTAG	T GGGGG	TGCC	7	GCTGAGAAC				ATACCOTT	1 2	ANCONCE	TOCOCTAL	TACTOCOC	AGOOR	TTOTOTATT
PFI ONRI	TCTCATA	ALGOGG	TGCT	78	GCTGAGATO	2			TACCOLT	2	AACCTAC	ACAGE-TAA-	TCCTCTT	TAGENA	TTTATTGCA
CEU THIS	CACOUT	TAGGGG	TGTC	22	CTGAGATC				ATACCCTT	2	AACCTGA	TCAGGCTGA	TACCTGCG	AAGELA	-BANGTACAG
DR TEIB	GCGTCAC	COGOGO	TOCC	13	GOTGAGAAC				ACACCCCA	2	AACCTGA	ACCOGGTCA-	TTCCGGCG-	GAGGGAG	TGTGATGCTC
DR_THIC	TCGTCAA	CLEGEG	ICCC	17	GETGAGAG	3	7	3	TAACCCTA	2	AACCTGA	ACIGGTIAG	CACCAGOG	GAGGGA-	-GIGIGACGG
TO_THIS	GOOCGTCAC	C00000	TOCC	7	GCTGAGAGC	-			ATACCCTT	N	AACCTOR	TOCOLOTCA	TOCCOOCO	TAGOGA	-GGTGLACGGCC
TAC THIT?	CTAGGO	AGGGGG	AGCT	9	GCTGAGAG	3	2	5	GACCCCT	2	AACCTUA	TCCQQQTAA	TACCORCE	CAGGGA	TCGTA TGA TG
FAC THITI	AGTTATA	CCGOGG	AGCT	4	TOCTGAGAG	3	4	3	GACCCOT	2	AACCTGA	TCC GGACAA-	TACCORCO	GAOGGAC	ATGGATACGA
FAC THIT2	ANTATOC	AGGGGG	AGCT	5	AGCTGAGAG	3	4	3	GACCCCG	2	AACCTGA	TCCGGCCAG	TACCORCO	GAOGGAT	TOGATGAANS
TV_TEIT1	GGACGAT	ARCCCC	AGCT	8	CCTCACAC	5	5	5	GACCCTT	2	AACCTGA	TCC OGGEAA-	TECCEGOG	GACCCAR	CHARGINGT
IV_THIT2	GTTCATA	ANCOCC	ACCT	8	CCTCACAC	5	5	5	GACCCT T	rC	-AACCTGA	TCCOCCTAA-	TECCEGOG	GACCCAN	ATTATCTCCC

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

	0	1 2		2'	3		3'	1'		4	5	5'	4'	0'
	===;	MEGGENGY	Y Y	ROCTGAGA	>	•	<	NACCCkt		AACCTGA	tcygONTaa	TRCergCG	NAGOgA	ç
BS_TRIC	TAGTTACT	00000100	C 9	GOCTGAGA	6	5	6	AACCCTT	2	GACCTOA-	TCTOGTTCG	TACCAGO	TOOOGA-	-AGTAGAGG
BS_TINA	TRACCACI		59	GCTGAGA	7	3	7 5	GACCOLC	2 0	AACTTGA-	ACAGGITCA	-GACCIGCG-	TAGEGA-	-AGTGGAGC
BS TROP	ARAQCACT		5	GCTGAGA	9	7	9	ATCCCTT	2	GACCCGA-	TCTOGATAA	TACCAGOG	TOOGA-	-AGTOCAGG
BS_YUAJ	TGACCACA	AGGGGAGC	A 4	AGCTGAGAGT	5	4	5	TGACCCTT	3	AACCTGT-	TAGTTAA	-COCTOOCO-	TAGOGA-	-TGTGGCAR
BE TENA	CATCCACI		C 8	GCTGAGA	6	5	6	GACCOLT	2	CACCTGA-	TCTGGGTAA	TCCCAGOG	TAGGGA	-AGIGGAGG
ED_TUAJ	TAACCACT	AGGGGTGT	C 4	CACTGAGA	4	5	4	AACTCTT	2	AACCTGA-	TCTAGTTCA	TACTAGOG	ALGGGA-	-AGTGGCGC
HD_THIE	AGGAGACT	AGGGGTGT	C 10	CACTGAGA	6	6	6	-ACCCTC	1	TACCIGA-	TCIGGATCA	TECCAGOG	TACCGA	-AGTEGACT
ED_YLMB	AATCCACT		A 3	COCTGAGA	8	4	8	AACCOLC	3	CACCIGA-	TCLAGGIA	TACTAGOG	AAQQGA-	-AGTOGCCA
SC_THIC	GATAAACI		C 13	GCEGAGA	6	8	6	AACCCTT	2	GACCTGA-	TCTGGCTCG	TACCAGO	TAGGGA-	-AGTTAACG
SC_TENA	CATCCACT		C 7	CGCTGAGA	4	6	4	TTGATCCTT	25	TACCIGA-	TCIGGTEAA	TOCTOOG	TAGGGA	-ACTOCCAA
ZC_THIX	TTANGGAO	CGGGGAG	C 4	GCTGAGAG	5	6	5	GACCOTC	0	AACCTGA	TCTOGATAA	TOCCAGOG	TAGGGA-	-GTTACTTAL
SC_TRIM	AGAAACAC	ATOGGAGT	1 10	AACTGAGAGT	5	9	5	TGACCATT	2	AACCTGT-	TOGATAA	TOCCAGO	TAGGGAG	AGTGTAAA
SA TENA	GAACTACT	AGOGGAG	6 9	GCTGAGA	4	3	4	GACCOTT	2	GACCTGA-	TTTGGTTAG	TACCARCO	TAGGAA	-AGTAGTTA
SA YKOE	CACACACI	AGGGGTGT	т 3	TACTGAGA	5	4	5	ABCCCTT	2	AACCTGA-	TCLAGCITG	-AACTAGOG-	TAGGAA-	-AGIGITAC
TENA	TIGCTACT		4	CGCTGAGA	5	2	5	GACCOTT	2 2	AACCTGA	TTIGGTTAA	TACCAROS	TAGGAA-	-AGTAGTTA
EY THIY	GTATCACT	AGGGGTGC		TECTGAGA	6	2	6	AACCCTT	1	AACCTGT	TGGTTAG	-CACCEGOG-	TAGGAA-	-AGTGAGCA
ET_IKOS.	AACOCACI	10000101	A 5	TACTGAGA	5	4	5	AACCCTT	2	AACCTGA-	TCTAGCTAG	TTACTAGO	TAGGAA-	-AGTGTTGT
IM YUAJ	TTACCAC	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	6	CCEGAGAT	7	6	5	TGACCOLT	2	AACCTGT-	TT-GTTAG	-TCCARGOG-	TAGGGA-	-ACIGAATC
CA THIS	ATAGTTA	CGGGGAG	C 8	GCTGAGAG	6	3	6	GACCOTC	2	AACCTGA	TTEGATAA	TOCCALCG	TAGGGA-	-GTTAATGC
CA YUAJ	TATOTOCI	AGOGGIGC	C 4	GCTGAGA	8	4	?	AACCOLT	0	AACCTGA-	TCTOGATAA	TACCAGOG	TAGOGA-	-AGCAGTTT
CA THIC	TTTTAGC		56	GCTGAGA	3	-	3	AACCOLT	2	AACTTGA-	TOTAGTTAA	TACTACO	TAGGGA-	-ACCHITIT
CA_THIN	TATTTTC	AGOGGAG	1 8	GCTGAGAG	5	6	5	GATCCAT	1	TACCIGA-	TTTOGATAA	TOCCARCO	TAGGRAA	TACOTTACT
DF_THID	AAATAQCI	16066 AG	C 3	AGCEGAGAGO	TA		TA	CCGACCCTT	1	CACCTGA-	TCTOGATAA	TACCAGO	TAGGAA-	AGCITAGTA
DF THIX	ATATATT	CGGGGGGGG	T 4	TGCTGAGAG	5	3	5	GACCOTT	2	AACCTGT	TGGATAA	-TGCCAGTG-	TAGGGAG	ACCAGTTAA
CI_TENA	TTATAGCT	AGGGGTGC	C 3	GCTGAGA	6	5	6	ARCCOTT	2	AACCTGA-	TCTGGATAA	TACCAGOG	TAGGAA-	-AGOCTGTA
CI_THIC	TAACTCCT		8 4	TECTGAGA	4	3	1	AACCOTT	2	AACCTGA-	TCIACTIAC	TACTACOG	TACCCA-	-ACCANAAG
CI THIS	ATATTCCI	AGGGGGGG	1 6	GCTGAGA	5	5	5	AACTCTA	2	AACCTGA	TTIGGTTAA	TACCAGOG	TAGGGA	-AGTATATT
CB THIN TAL	TTTTCAAP	AGOGGAG	T 9	GCEGAGAG	5	4	5	GACCCAT	0	AACTTGA-	TTCOGATAA	TOCCOLCO	TANGGAN	ATATTATT
CB_THID	TATGTTGT		6	GCTGAGA	4	6	4	AACTCIA	2	AACCTGA-	TCCAGATAA	TACTOGOG	TAGGGAA	GCAGCATAT
TTE THIC	CAAGTOCI	GGGGAG	c 6	GCTGAGAG	4	5	4	GACCCTT	2	AACCTGA-	CCAGGGTAA	TCCCTGCG	ANGGGA-	-AGCACGTT
TTE THID	AAAGTGCI	AGGGAAGC	C 6	GCTGAGAG	4	5	1	GACCCTT	2	AACCTGA-	CCAGGETAA	TECCTECE	AAGGGA-	-AGCACGTT
TTE THIDE	CAAGTGCT	AGGGGAG	6	CGCTGAGAG	4	5	4	GACCOTT	2	AACCTGA-	CCAGEGTAA	-TCCCTCCG	AAGGGA-	-AGCACGTT
EF_YKCE	ARAACATI	TGOGGTGC	1 5	GCTGAGATG				ATACCCAT	2	AACCTGA-	TOCAGTIAG	TACTOTOG	CAOGGA-	-AATOCCGA
EF_THIN	CARACATI	TEGECIC	3	AGCTGAGATT			6	GACCOAT	2	AACCTGA-	TC-ACTTA A	-GACTOGOG-	-CAGGGA-	-AATGTCTT
ZE YUAJ	ACTCACA	GGGGAG	c 6	GCTGAGA	5	5	6	ACCCTT	2	TACCTOT	TCGGTTA	TECGAGOG	TAGGAA-	-TTGTGLAT
ILX YUAJ	TTTGCACZ	ATOGGICI	A 16	TROCGAGAA				ATACCATC	1	GACCTGA-	TCTOGGTAA	TOCCAGOO	TAGGAA-	-TGTGTTAR
PH TENA2	GAGACAT	TGGGGTG	3	GCTGAGATA				ATACCCAT	2	AACCTGA-	TACAGTTAA	-GACTGGCG- GACTGGCG-	AAGOGA-	-AATGTGAA
PN_THIX	TATATAT	ATGGGAGT	C 8	GTCTGAGAG	5	5	5	GACCGC-	0	-ACCTGA-	TCTGGGTAA	TECCAGOG-	-GAGGGAA	CGATACTTA
ST_TUAJ	TTTCACAP		2	GCTGAGA	4	4	1	ANTCOIG	2	GACCTGA-	TCITCITAC	TACANGCO	TAGGGA	-TTGTGACC
DEA THIC	TAATCACT	ACCCCCC	c 10	CECTGAGA	7		;	-ACCCTT	ō	AACCTGA-	TCEGGCEAA	TCCCAGOG	TAGGGAN	-CGTCCATA
COL_THIC	CAGTCCCC		C 8	COCT GAGA	6	8	6	OCACCOTT	2	ANCCTG-	TCCOGTTAG	-CACCOGCG-	ANGGRAG	AGAGGAATG
CGL THID	CTTACCCC		C 22	GCTGAGAAA	G	5 7	5	ACACCOT CCACCOT	2	AACCTGC-	TCTAGCTCG	-TACTAGOG-	ANGGAT	CACENTIC
CGL YROE	TCATAGAC	CGGGTGC	1 13	CECTGAGATC	4	3	4	CGACCOTC	1	AACCTGA-	TCCGGATAA	TECCEGOG	TAGGAG	CRAMATAT
COL OARE	TAGTGACI	CGGGGTGC	A 22	CGCTGAGATT				ACACCOST	2	AACCTGA-	TCCAGTTAG	TACTOGCO	ANOGGAC	TOTOCATT
DI TRIC	TGTCCCCC		C 8	CCTGAGA	6	7	5	GCACCOTT	2	AACCTGT-	CIGCITAA	-CACCAGOG-	ALGERAG	ACAGGACCC
DI_YKCE	TATAAATC	CGGGTGC	1 19	AGCTGAGA	7	12	8	AACCGTT	1	AACCTGA	TCCGGGTAA	TACCEGOG	TAGGAAG	ALTRA TGAA
MT_THIO	CTGTAGAC		C 11	CECTGAGAGT	4	6	4	CACCOLO	1 2	CACCTGA-	TCCGGATCA	TCCCCCCCC	- AAGGGAG	GTCALGGAT
ML_THIO	GGAGTCCC	CAGECAGT	c 11	CECTEMENET	12	5	12	TTACCCTC	3	CACCTGA-	TOCOGGTCA	TOCCOGOG	ALOGGAG	GTTCALGAT
ML_THIC	ARABCCAC	CCCCCCA	A 11	CGCTGAGAG	5	7	5	GACCOTA	2	ANCCIGA	-CCGGGTAA	-Teccesce-	TAGGGAG	ATGALTAAT
MB_THIC	CTGTACAC			COCTGAGAG	6	5	6	GACCOTA	2	CACCTGA-	TCCCCATCA	-TOCCOOCO-	- TAGGGAG	TICAATG
RE TENA	CGTAGACI	CGGGGTGC	1 15	GCTGAGATC				ACACCOST	2	AACCTGA	-TCAGGTAA	TOCTOLOG	AAGGGAT	GTCACGGCA
RK THIC	CCGTCCAC		C 10	CECTENENT	7	3	7	CGACCOTC	1	AACCTGC-	-CCGGGTAA	TECCEGOG	ALGERAG	TAGGALCAG
RK THIC2	CCAACCAC	ACGEGAGO	1 10	CECTGAGAC	6	5	6	GACCOLA	1	AACCTGA	-CCCCCCTAA	TTCCCCCCC	TAGEGAG	ATCATCTCC
SX_THIO	AGGCACTO	COCOGARGO	C 11	GOCTGAGAG	6	5	6	GACCOTA	2	AACCTGA-	TCCGGGTCA	TOCCOGCO	ANOGGAG	GGGCTGGAC
SX THIC	TTGTACAC		1 10	COCTGACAG	45	4	51	GACCOCC	1	AACCTGT	CORRECTAN	TICCCCCCCCC-	TAGGGAG	TAGGTCTCA
TTT THIO	GAAAACCO	CCCCCC	c 12	CECTGAGAG	6	5	6	GACCOCA	1	AACCTGA	TCCGGGTCA	TACCOGCG	ANGGGAG	TGGOGCGCC
TPU YROE	CAAACGAC	AGGGGAG	G 7	CECTGAGAG	6	3	6	GACCOTT	2	TACCTGA-	TCTGEGTAA	TECCIECE	AAGGAA-	-GTOGTGGG
ANI THIN	TACATAA	ALCOUGIC	11	GCTGAGAGG	5	5	5	TGACCTUC	1	AACCTGA	TCIGCATA	TCCCAGCO	TAGEGAT	AAAGCAGCT
CAU_CYTX	GCGTGTAT	GAGEGAGA	10	CGCTGAGAG	6	8	6	GACCICC	rG	-AACCTGC-	TCIGGGTAA	TECCIGOG	AAGGGA-	-ATACACGG
CAU X THIX	GCGCTGCC	AGGGGAGC	C 10	CGCTGAGAGT	5	9	5	TGACCCTT	2	AACCTGA-	TCTGGGTTA	TECCAGOG	GAGGGAN	-GGOGTATT
CL THIC	ATCATCT	CCCCCCCCCC	21	AGCTGAGATC				ACACCOCT	2	AACTTGA	TOCAGCIAA	TCCTCACG	AAGGAA-	-AGATCAAG
FN_THIC	ATATGTAC	TGGGGAG	7 3	TGCTGACATT	4	6	4	TAGACCCAT	2	TACCTGA-	TTTGGATAA	TOCCARCO	AAGEGA-	-GTACCATC
TH THIN	CTACTTAC	GGGGGAG	10	GGCEGAGA	4	7	4	GACCOTT	1	BACCTGA-	TOCCOTA	TOCCAACG	GAGGGAT	COGGODAGC
PMI_TEIB	AATAGCO	AGGGG.GC	7 4	CECTGACAG	5	7	5	GACCCTG	2	TACCTGA-	TCIGGATAA	TACCAGOG	ANGEGAG	GCTAALAT

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 Pasteurellaeceae and *H. py-lori*. 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 THImediated 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*.



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.

*maritima* differs from that in most eubacteria and uses the *thi4* gene product.

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  $\gamma$ -proteobacterium, is not *thiC*, as in most  $\gamma$ -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 readthrough. 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 thiaminmediated 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

- Schowen, R. L. (1998) in *Comprehensive Biological Catalysis*, Vol. 2 (Sinnott, M. L., ed) pp. 217–266, Academic Press, London
- 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
- 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.
- 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
- 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
- Miranda-Rios, J., Navarro, M., and Soberon, M. (2001) Proc. Natl. Acad. Sci. U. S. A. 98, 9736–9741
- Gelfand, M. S., Novichkov, P. S., Novichkova, E. S., and Mironov, A. A. (2000) Brief. Bioinform. 1, 357–371
- 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
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999) Proc. Natl. Acad. Sci. U. S. A. 96, 2896–2901
- 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
- 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
- Lyngso, R. B., Zuker, M., and Pedersen, C. N. (1999) *Bioinformatics* 15, 440-445
- Mironov, A. A., Vinokurova, N. P., and Gelfand, M. S. (2000) Mol. Biol. 34, 222–231
- 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
- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Nucleic Acids Res. 25, 3389–3402
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) Nucleic Acids Res. 25, 4876–4882
- 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
- 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 Duion, B. (1999) Mol. Microbiol. 32, 1140–1152
- 34. Moeck, G. S., and Coulton, J. W. (1998) Mol. Microbiol. 28, 675-681
- 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