Comparative Genomics of Thiamin Biosynthesis in Prokaryotes

NEW GENES AND REGULATORY MECHANISMS*§

Received for publication, September 3, 2002
Received, 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. Box 348, Moscow 117333, Russia

Vitamin B1, 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, 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 yuad 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 hydroxymethylthiazole biosynthesis.

Thiamin pyrophosphate (vitamin B1) 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) 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 aminomimidazole 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-β-xylulose phosphate in an unresolved chain of reactions involving the thiF, thiS, thiO, thiH, and thiI gene products. 1-Deoxy-β-xylulose phosphate, whose production is catalyzed by the dxa 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, Thi 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 thiBFQ 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 ywbd-thiKC operon whose transcription is partially repressed by thiazole but not by thiamin (10, 11). Re-
Fig. 1. The thiamin biosynthesis pathway in bacteria. The standard E. coli gene names are used throughout except tenA-tenI, ykoFEDC, ynaO, 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 ThiI (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.

Recently, 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 euabacteria, archaea, and eukaryota.

EXPERIMENTAL PROCEDURES

Complete and partial sequences of bacterial genomes were downloaded from GenBankTM (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 database (ergo.integratedgenomics.com/ERGO/). (21). Gene identifiers from the ERGO database and GenBankTM 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 antisense genes 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 PHYLP (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.
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*, 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-thiI* (or *thiS-thiG-thiO*) for the HET biosynthesis. Most Gram-positive pathogens from the *Bacillus/Clostridium* group, all Pasteurellaceae, 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/moeD* 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 shown 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 teditum*, 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 aerobic *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 archaeabacteria. 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 thermoadaptomorophum*. 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 genome 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, Pasteurellaceae, *Vibrio fisheri*, *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 diptheriae*) 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>
<thead>
<tr>
<th>EC</th>
<th>BS</th>
<th>Function</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>thiC</td>
<td>thiA</td>
<td>HMP biosynthesis</td>
<td>76</td>
</tr>
<tr>
<td>thiD</td>
<td>ybhV</td>
<td>Phosphomethylpyrimidine kinase</td>
<td>43</td>
</tr>
<tr>
<td>thiG</td>
<td>ybhT</td>
<td>Thiazole biosynthesis protein ThiG</td>
<td>52</td>
</tr>
<tr>
<td>thiH</td>
<td>ybhR</td>
<td>Thiazole biosynthesis protein ThiO</td>
<td>-</td>
</tr>
<tr>
<td>thiI</td>
<td>ybhJ</td>
<td>Thiazole biosynthesis protein ThiI</td>
<td>33</td>
</tr>
<tr>
<td>thiF</td>
<td>ybhU</td>
<td>Adenylyltransferase</td>
<td>37</td>
</tr>
<tr>
<td>thiS</td>
<td>ybhS</td>
<td>Sulfur carrier protein ThiS</td>
<td>31</td>
</tr>
<tr>
<td>thiM</td>
<td>thiK</td>
<td>Hydroxyethylthiazole kinase</td>
<td>43</td>
</tr>
<tr>
<td>thiE</td>
<td>thiC</td>
<td>Thiamin-phosphate synthase</td>
<td>39</td>
</tr>
<tr>
<td>thiL</td>
<td>ydiA</td>
<td>Thiamin-monophosphate kinase</td>
<td>37</td>
</tr>
</tbody>
</table>
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 elements; one-fourth of the genomes contain only one THI-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 Pasteurellaceae. For instance, the ThiD proteins from Hemophilus influenzae, Pasteurella multocida, and Mannheimia hemolytica are close to ThiD from the Pyrobaculum species. The ThiD proteins from the Pasteurellaceae never cluster with TBS genes.
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 perfringens 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 thiX-thiY-thiZ, was identified in bacteria from various taxonomic divisions, such as γ- 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, M. genitalium, 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, named ykoE-ykoD-ykoC, which 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. hemofulica, Chloroflexus aurantiacus, Thermoaerobacter tengcongensis, pseudomonomads, 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 thiMD-EHI0418 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 Pasteurellaeae, P. multocida and M. hemofulica. Since HI0418 and all of its orthologs are clustered with the thiMD 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 Methylobacterium 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-ribosylcinotaminde 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 periplasms, 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, clostridia, 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 (yclmB in Bacillus sp.) belongs to the ArgE/dapE/ACY1/CPG2/yscF family of metallopectidases. 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-inr3. 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 B12 (34). In another bacteria from the CFB group, Polaribacter filamentosus, omr1 is a single gene that is preceded by a THI element. The pnuT gene
encodes a transporter that is similar to PnuC, N-ribosylcyanamide transporters from enterobacteria (35). The PnuT protein forms 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-trn3 operon is weakly similar to the C-terminal part of the thiamin pyrophosphokinase TNR3 from yeast. Based on these data, we propose that the hypothetical THI-regulated omr1-pnuT-trn3 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-trn3 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-ribosylcyanamide but not nicotinamide mononucleotide), which are then phosphorylated by specific kinases (TNR3, RibF, and NadR for thiamin, riboflavin, and N-ribosylcyanamide, respectively) to make transport.

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 conserved on the sequence level. In addition, 23 nonpaired positions of the above helices 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 is 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 a 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. aurantius 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.
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
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 systems. 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 pathways 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.

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 replacements 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 thiL, are never regulated by the THI
element and are never positionally clustered with other thia-
mamin biosynthetic genes. More surprisingly, thiamin-monophos-
phate kinase thiL is not clustered with other thi genes and is
not regulated by the THI element.

The proposed model for the thiham 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 forma-
tion of the terminator hairpin and premature termina-
tion of transcription. Without thiamin pyrophosphate, unstable
THI element is replaced by the more energy-favorable antitermina-
tion conformation that allows for the transcription read-
through. In Gram-negative bacteria, stabilization of the THI
element leads to formation of the SD sequencer hairpin, which
represses initiation of translation. In the derepressing con-
ditions, THI element is replaced by the antisecuquencer that
relaxes the SD-box and allows for initiation of translation.

It is known that various nucleotides, such as flavin mononu-
cleotide or nicotinamide mononucleotide, can specifically bind
to RNA aptamers (36). Thiamin pyrophosphate, containing py-
rimidine and thiazole moieties, is a regulatory molecule for the
regulation of expression of the thiamin biosynthetic and trans-
gene (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.

Comparative Genomics of Thiamin Biosynthesis

element due to possible

REFERENCES


Taylor, S., Campobasso, N., Chiu, H. J., Kuslinad, C., Reddick, J. J., and Xi,


279–293, American Society for Microbiology, Washington, D. C.

6. Miranda-Rios, J., Morera, C., Toda, H., Davalos, A., Encarnacion, S., Mora,


179, 3030–3035


Sci. U. S. A. 98, 9736–9741

Brief. Bioinform. 1, 357–371

Nucleic Acids Res. 30, 3141–3151


22. Overbeck, B., Larsen, N., Pusch, G. D., D’Souza, M., Selkov, E. Jr., Kyrpides,
28, 123–125

Problems,” Samara, Russia, September 4–9, 2001, pp. 623–625, The Insti-
tute of Control of Complex Systems, Samara, Russia

440–445

222–231


28. Alsulaiman, S., Madsen, T., Schaffer, A., Zhang, Y., Zhang, Z., Miller, W.,

29. Thompson, J. D., Gibson, T. J., Prolliak, J., Jeannoufing, F., and Higgins,

30. Taylor, S. V., Kelleher, N. L., Kuslinad, C., and Costello, C. T., Backstrom, A. D.,

Biochem. Int. 14, 153–160


36. Kurnasov, O., Polanuyer, B., Ananta, S., Sloutsky, R., Tam, A., Gerdes, S., and


Acknowledgments—We are grateful to Andrew Osterman for attention,
advice, and encouragement and to Tadzh Begley, Pietro Durrestein, Iain
Anders, and Olga Vasseva 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. ther-
mophilus genome sequence for this analysis.

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 reglons (16,
17).