From: *Bioinformatics and Genomes: Current Perspectives*
Edited by: Miguel A. Andrade

Chapter 4

High Rate of Gene Displacement in Vitamin Biosynthesis Pathways

Enrique Morett, Gloria Saab-Rincon, Enrique Merino, Peer Bork, Emmanuvel Rajan, Leticia Olvera, and Maricela Olvera

Abstract

One of the most challenging tasks in genomic science is the prediction of the function of genes for which there is no clear sequence similarity to annotated genes. However, it is even more challenging to assign the correct function to genes that display sequence similarity to genes of unrelated function: analogous enzymes perform the same biochemical reaction but they are not phylogenetically related, such that it is not possible to identify them by sequence similarity. Here we propose that the vitamin biosynthetic pathways have experienced multiple events of gene loss and recovery of function by unrelated genes, the so-called analogous gene displacement. We carried out an extensive search for the genes that participate in thiamin biosynthetic pathways in the completely sequenced genomes. We show that the great majority of these organisms lack from a few to many orthologs to the *Escherichia coli thi* genes. We searched for the analogous enzymes using gene neighbourhood, co-ocurrence in operons, identification of regulatory sequences, and anticorrelation strategies. Our strategy resulted in the identification of some possible analogous enzymes in this pathway.

Introduction

Generally, enzymes that catalyze the same biochemical reaction show a certain degree of sequence similarity andor are structurally closely related. This implies that, normally, they have a common phylogenetic origin and have evolved by divergence from an ancestral protein. Fitch (1970) coined the term orthologs to name such proteins when they are from different organisms, while paralog proteins denotes the same evolutionary origin but acquisition of new function by duplication and divergence. Paralog genes normally code for enzymes that have related activities and belong to the same structural family. However, since the early ages of enzymology, it has been documented that there are also apparently unrelated enzymes that present the same activity: the so-called analogous enzymes (Warburg and Christian, 1943). But, how common are analogous enzymes in nature?

In a recent analysis of all the enzyme sequences deposited in the public sequence data banks, Koonin and coworkers reported that analogous enzymes are more common that previously thought. They showed that more than 30 reactions are catalyzed by two or more proteins for which neither amino acid sequence nor structural similarity can be detected (Galperin *et al.*, 1998). Thus, it is inferred that these enzymes have then very likely evolved independently rather than from a common ancestral protein, and converged into the same catalytic function.

Is Orthologous Gene Displacement Frequent In Vitamin Biosynthesis Pathways?

The aim of this chapter is to propose that the presence of analogous enzymes is very common in pathways for the biosynthesis of compounds that are required in minute amounts in the cells, as vitamins. Our rationale is that since vitamins are required in so low amounts in the cell, but on the other hand they are absolutely required for cell growth, mutations that impair the activity of enzymes in their biosynthetic pathways could be suppressed by mutations in unrelated enzymes that alter their activity, such that they result in a weak catalytic activity towards the substrate of the affected enzyme. An extremely low catalytic activity of an enzyme that is expressed at high levels could be sufficient to provide the required limited amounts of the vitamin in question. Thus, very inefficient enzymes could provide enough activity and therefore be readily selected with the new activity. Later, via an evolutionary optimization process, the enzyme with the new activity would attain the required levels of activation and expression. It is likely that similar processes could also operate in other pathways, but the low amount of vitamins needed made them more plausible to evolve independently from unrelated enzymes. For example, pathways for secondary metabolism could also be targets of gene displacement. In such pathways the selective pressure to recover the activity of a mutant cell that is unable to synthesize a metabolite, not required for survival, would not be so intense as compared to vitamin biosynthesis.

As an example of the presence of multiple analog enzymes in vitamin biosynthetic pathways we present our analysis of the thiamin biosynthesis in the completely sequenced microbial genomes. It is shown that the great majority of organisms lack from one to several genes orthologs to the *thi* genes of *E. coli.* As fast divergence of enzymes in vitamin biosynthesis is rather unlikely, the absence of identifiable orthologs indicates that multiple events of gene displacement, that is the presence of analogous enzymes, have occurred in the biosynthetic pathway of thiamin.

Thiamin, also known as vitamin B1, is a cofactor for several key enzymes involved in carbohydrate metabolism, therefore it is essential for growth (White and Spencer, 1996). Thiamin phosphate is made by the condensation of hydroxy methylpyrimidine pyrophosphate (HMP-PP) and methyl (-hydroxyethyl) thiazole phosphate (THZ-P) moieties, by the enzyme thiamin phosphate synthase, coded by the *thiE* gene (see Figure 1). Thiamin phosphate is converted to the pyrophosphate form, the active form of the vitamin, by thiamin kinase, the product of the *thiL* gene. In *E. coli* at least two gene products, *thiC* and *thiD* participate in the formation of HMP-PP from AIR, an intermediate in purine biosynthesis, while at least six gene products, *thiH, thil, this, thiG, thiF* and *thiM,* are needed for the synthesis of THZ-P (Begley *et al.,* 1998).

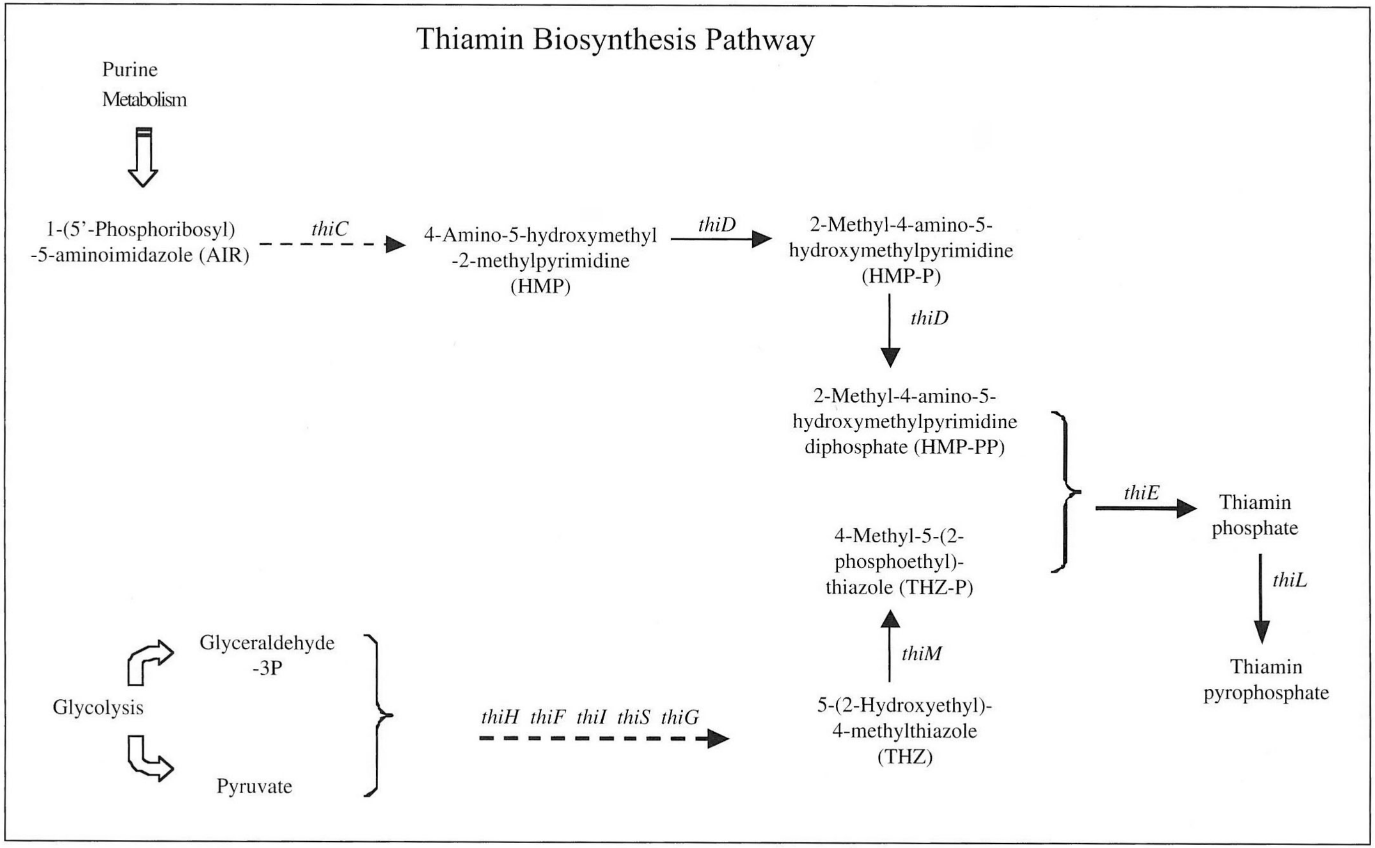
Thiamin Biosynthesis Pathway
Purine Metabolism
1-(5'-Phosphoribosyl) -5-aminoimidazole (AIR)
thiC
4-Amino-5-hydroxymethyl -2-methylpyrimidine (HMP)
thiD
2-Methyl-4-amino-5-hydroxymethylpyrimidine (HMP-P)
thiD
2-Methyl-4-amino-5-hydroxymethylpyrimidine diphosphate (HMP-PP)
thiE
Thiamin phosphate
thiL
Thiamin pyrophosphate
4-Methyl-5-(2-phosphoethyl)-thiazole (THZ-P)
thiM
5-(2-Hydroxyethyl)-4-methylthiazole (THZ)
Glycolysis
Glyceraldehyde -3P
Pyruvate
thiH thiF thiI thiS thiG

Figure 1. *E. coli* thiamin biosynthesis pathway. Discontinuous arrows indicate unknown parts of the pathway. The functional implication of some genes to those unknown steps might be known (e.g. *thiC* is known to be somehow involved in the transformation of AIR into HMP, and it is not kwnown whether there are other proteins involved in the process). Thick arrows indicate connections with other important pathways. Note that the product of *thiD* catalyses two different consecutive steps.

Search For Orthologs Of The *E. coli thi* Genes In The Completely Sequenced Genomes By Sequence Similarity

We searched for the presence of ortholog genes to the *E. coli thi* genes in the complete microbial genomes, and analysed their genomic organization. Nine Archeal, 33 Eubacterial and the genomes of the yeast *Saccharomyces cerevisiae* and *Candida albicans* were analysed. It follows a list of the genes found in these completely sequenced genomes.

thiD is the gene that is present in more of the organisms analysed; it is present in all microbes but *Synechocystis* including those with small genomes, *Rickettsia*, *Chlamydia*, *Ureaplasma,* and *Mycoplasma,* which lack almost all the enzymes for *de novo* thiamin synthesis. The ThiD enzyme catalyzes the two consecutive phophorylation steps in the synthesis of HMP-PP. The products of *thiJ* andpdxP can also phosphorylate HMP to the mono phosphate form, but these proteins are involved mainly in scavenge of HMP from the media and we will not discuss them further.

thiE and *thiL* are the second most ubiquitous genes after *thiD*. *thiE* codes for the key enzyme in thiamin biosynthesis, thiamin phosphate synthase. This enzyme condensates the THZ-P and HMP-PP to produce thiamin monophosphate, which is converted to the pyrophosphate form by thiamin kinase, the product of *thiL*. *thiE* is present in yeast and in all the eubacterial species but *Thermotoga maritima*, and is missing in the archeobacteria *Halobacterium, Methanobacterium, Methanococcus, Pyrococcus* and *Aeropyrum.* Like *thiE, thiL* is missing in *Thermotoga maritima;* it is as well missing in *Deinococcus radiodurans, Lactococcus,* and *Streptococcus.* It is present in all the Archea but it is missing in yeast and *Candida.* Sequence similarity using PSI-BLAST searches (Altschul *et al.,* 1997) failed to detect orthologs of *thiL* in other Eukaryotes. It is very likely a gene only present in Archea and Eubacteria.

thiC is present in the three domains of life, being missing only in *Thermoplasma, Aeropyrum pernix, Lactococcus lactis, Streptococcus pyogenes, Haemophilus influenzae,* and *Pasteurella multocida.* The function of its product is not yet known. *thiM* is scattered in the three domains of life. It is present in the Archea *Archaeoglobus fulgidus* and *Pyrococcus,* and in yeast. In Eubacteria it is present in the firmicutes, *Enterobacteriae, Pasteurellacea, Helicobacter,* and *Campylobacter.* ThiM phosphorylates THZ to THZ-P.

this codes for a very short protein of 66 amino acids which has some sequence similarity with MoaD. Both proteins have two glycines at the COOH terminus that are involved in sulfur transfer via the formation of a thiocarboxylate at the COOH end of the protein. This donates the sulfur for the formation of thiazole in a reaction catalyzed by both ThiI and IscS proteins. *this* is present in the Archea *Archaeoglobus* and *Methanobacterium.* It is widely present in Eubacteria although it is missing in *Thermotoga, Aquifex,* and *Haemophilus.*

thiF is highly similar to *moeB,* a gene involved in the biosynthesis of the molibdopterin cofactor. The high level of similarity has lead to several miss-annotations in many organisms. We identified unambiguously *thiF* genes not only by their higher sequence similarity to *E. coli thiF* than to *E. coli moeB,* but also by being located in operons with other *thi* genes in the following organisms: *Campylobacter, Bacillus, Escherichia coli, Aquifex, Neisseria, Vibrio, Xylella, Helicobacter, Archaeoglobus, Aeropyrum,* and *Methanobacterium.* ThiF is involved also in the sulfur transfer process to form THZ. It has recently been found that a covalently linked protein-protein conjugate is formed between ThiF and This similar to the ubiquitin-El conjugate (Xi *et al.,* 2001).

thil has a dualrole in *Escherichia coli.* Its product is not only involved in thiamin biosynthesis, transferring sulfur to This, but also participates in the formation of thiouridine in response to UV light, a protective mechanism that allows the cells to survive to the DNA damage. *thil* is found in many organisms including the small genomes but *Chlamydia.* However, we did not find *thil* homologues in *Campylobacter, Aquifex, Mycobacterium, Neisseria, Xylella, Helicobacter,* and *Haemophilus,* although almost all these organisms have *this,* as shown above. Then, there must be a different protein involved in the sufphur transfer to This in these bacteria.

thiG is only present in Eubacteria but no orthologs were found in *Thermotoga, Lactococcus,* and *Haemophilus.* The exact role of ThiG is still not clear. *thiH* is the least represented gene. However, Eubacteria as diverse as *Thermotoga, Escherichia, Vibrio,* and *Helicobacter* have this gene. *thiH* has sequence similarity to several genes involved in the biosynthesis of sulfur containing cofactors such as biotin *(bioB),*lipoic acid *(lipA),*PQQ *(pqqE),* and molibdopterin *(moaA).*The products of all these genes share a highly conserved cysteine-rich motif. These observations suggest that a similar sulfur incorporation takes place in the synthesis of all these vitamins.

Search For *thi* Genes Not Present In *E. coli*

As shown above, many organisms prototrophs for thiamin lack from one to several of the known *thi* genes. Actually, only the organisms closely related to E. *coli* show the complete set. This could imply that they utilize alternative pathways for thiamin synthesis or that many events of non-orthologous gene displacements have occurred and the *thi* genes have been reinvented several times in nature.

How the nonortholog genes could be identified? Methods of finding these missing genes include the search for physical linkage of open reading frames (Lathe *et al., 2000)* and patterns of ortholog co-ocurrence in different species (Pellegrini *et al.,* 1999). Also, the presence of highly conserved regulatory sequences, the *thi* box (Miranda-Rios *et al.,* 1997; *2001),* could be used to detect specifically *thi* genes. Using these tools, we have identified several ORFs that might be involved in thiamin biosynthesis.

Search For *thi* Genes In Complete Genomes Using Gene Neighbourhood And Anticorrelation Of Occurrence

In the search for analog genes of *thiE* we found a gene (that we denote *thiE*)* present in the Archeal as well as *Thermotoga,* organisms in which we could not find a *thiE* homologue gene. We did not find any homolog of *thiE** in organisms that have bona fide *thiE,* except *Pyrococcus.* This gene is actually fused to *thiD* forming a predicted polypeptide of about *400* amino acids. The extended region of this unusual *thiD* gene did not show sequence similarity to any other gene in the public data banks searched using PSI-BLAST. This anticorrelation and the fact that in these organisms, *thiE* is the only gene missing in common in the biosynthetic pathway of thiamin, makes very tempting to speculate that the *thiE** gene codes for an analog thiamin synthase enzyme. It will be very interesting to determine whether this gene can complement a *thiE* mutant strain of *E. coli.* We are currently constructing such a strain and cloning the extended *thiD* gene of *T. maritima* to check this hypothesis.

In *Rhizobium etli* there is *a* gene, *thi0,* for which there is no ortholog in *E. coli,* that has been proposed to participate in thiamin biosynthesis (Miranda-Rios *et al.,* 1997). This gene forms part of a *thiCOSGE* operon. We searched for ortholog genes to *thiO* in the completely sequenced genomes. We identified likely ortholog genes to *thiO* in *Aquifex, Mycobacterium, Bacillus,* and *Neisseria.* None of these species have *thiH.* Interestingly, *Bacillus subtilis* has almost the complete set of *thi* genes identified in *E. coli,* with the exception of *thiH.* Thus, we found that *thiO* and *thiH* are mutually exclusive; in other words, their presence anticorrelates in the completely sequenced microorganisms. Thus, we speculate that *thiO* and *thiH* have equivalent functions being a case of non-ortologous gene displacement. Neither the function of the *thiO* product nor the function of *thiH* gene product is known. Both *thiO* and *thiH* are distributed among different Eubacterial taxa. Then it is difficult to predict which gene is the ancestral *thi* gene. Interestingly neither Archea nor the microbial Eukaryotes have any of these genes, implying that there must be a third analog gene coding for the same function in these organisms. We are constructing an E. *coli* strain devoid of *thiH* to try to complement its function with *thiO* from R. *etli*.

Search For The Regulatory *thi* Box

The *thi* box is a conserved regulatory sequence of about 38 nucleotides present in front of several *thi* genes in both Eubacteria and Archea organisms. About one half of the positions of the *thi* box are almost invariant, but all can fold in a similar RNA structure due to correlated substitutions that maintain base-pair complementarity (Miranda-Rios *et al.,* 2001). It has been proposed that this sequence, once transcribed to mRNA, can directly bind thiamin and down-regulate the expression of *thi* genes when the vitamin is present. No other regulatory sequence is conserved in such diverse group of bacteria.

We searched for possible *thi* boxes in the leader region of all the genes of the completely sequenced genomes in an attempt to identify new putative analog *thi* genes. We developed a program that scans the 500 nt sequence upstream of every gene, according to the GenBank database annotations. A 30 nt sliding window is used to search for the acctga conserved sequence in the 5' side of the *thi* box element. If a window has a perfect match or only one mismatch, then the program searches for the cgnngg sequence at the **3'** end of the *thi* box element. Since the analysis of *thi* box sequences present in different genomes revealed that the 3' end sequence may vary importantly, our program allows the following variants: gnngg, gnnngg, cnnngg, cnnnngg, cgnnng, cgnnnng, cgnng, cgnnng. If both 3' and 5' conserved strings are found, then the program predicts the secondary structure of the 30 nt window using the FoldRNA program of the GCG package. In case that the predicted mRNA fold corresponds to a stem-and-loop structure the program verifies that its free energy is smaller than zero. If the structure meets this requirement, the program verifies that all of the bases found in the 5' (acctga) and 3' (cgnngg) searches, are part of a stem or a loop, accordingly with its position within the consensus *thi* box mRNA fold. If a window meets all of these considerations, its sequence is therefore considered as a putative *thi* box element.

Our search strategy resulted in the identification of several of the *thi* boxes used to construct the searching matrix and of *thi* boxes corresponding to already identified *thi* genes for which no *thi* box had previously been described. Additionally, several other genes having *thi* boxes were found.

Several seem to be false positives, since they are involved in non vitamin-related functions. In general, these *thi* boxes showed a low score. However, many other quite interesting hits were obtained which support a common role of thiamin in the regulation of at least some genes involved in the molibdopterin cofactor and biotin biosynthesis. Then, we speculate that thiamin can, in some organisms regulate the expression not only of the genes involved in its synthesis but also of genes involved in other sulfur-containing vitamins. Following we describe some of the most relevant examples.

The *ydbH* gene of *Lactococcus* presents at its untranslated region a highly probable *thi* box. This gene codes for a protein of unknown function but it is located in an operon with *bioY* which codes for biotin synthase. The *Thermoplasma acidophilum* gene *Ta0442* also has a putative *thi* box. This gene is in an operon with the molibdopetrin synthesis gene *moaA*.

Another gene for which we identifyed a *thi* box is the *ylmB* gene of *B. subtilis*. This gene is paralog to *argE,* and in *B. halodurans* the ortholog of this gene is in an operon with *thi4,* a gene first identified in *Neurospora crassa* as involved in thiamin biosynthesis. It is very intriguing that *B. subtilis ylmB* has a *thi* box and that its ortholog in the closely related *B. halodurans* species is in an operon with a thiamin related gene, although we did not detected a *thi* box in the latter gene.

Interestingly, we found *thi* boxes in the intergenic region of mRNAs. For example there is a *thi* box exactly before the start point of the *B. subtilis thiO* gene. This gene is the third gene in an operon with other *thi* genes. In *Synechocystis* the *brkB* gene is the second gene in an operon with *thiC* and there is a thi box in front of it. The *Thermoplasma* species have a gene with sequence similarity to membrane transporters, TVN1053, which is the second gene in an operon with *moaA*. The fact that we found *thi* boxes in these intergenic regions strongly supports the hypothesis that these regulatory sequences exert their role at the level of translation (Miranda-Rios *et al.*, 2001). The lack of terminator sequences and the low stability of the *thi* boxes makes them unlikely to play a role in transcriptional termination.

CONCLUSIONS

In this chapter we analysed the thiamin biosynthesis pathways in the completely sequenced microorganisms. Almost all of them lack at least one of the *thi* genes reported for E. *coli*. The fact that the majority of these organisms do not require thiamin for growth indicates that there are multiple analogous genes for the biosynthesis of this vitamin. Some genes likely to participate in this process were identified by gene neighbourhood, co-occurrence, and for the presence of *thi* boxes. These strategies can certainly be used not only for the identification of genes that participate in vitamin biosynthesis but also for the identification of many other cellular functions in the completely sequenced organisms.

Acknowledgments

This work has been supported in part by grants from CONACyT and DGAPA, UNAM. EM was an Alexander von Humboldt Fellow.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 2289-3402.

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. Thiamin biosynthesis in prokaryotes. Arch. Microbiol. 171: 293-300.

Galperin, M.Y., Walker, D.R., and Koonin, E.V. 1998. Analogous enzymes: independent inventions in enzyme evolution. Genome Res. 8: 779-790.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99-113.

Lathe, W.C. 3rd, Snel, B., and Bork, P. 2000. Gene context conservation of a higher order than operons. Trends Biochem. Sci. 25: 474-479.

Miranda-Rios, J., Morera, C., Taboada, H., Davalos, A., Encarnacion, S., Mora, J., and Soberon, M. 1997. Expression of thiamin biosynthetic genes (thiCOGE) and production of symbiotic terminal oxidase cbb(3), in *Rhizobium etli*. J. Bacteriol. 179: 6887-6893.

Miranda-Rios, J., Navarro, M., and Soberon M. 2001. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. Proc. Natl. Acad. Sci. USA. 98: 9736-9741.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.0. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA. 96: 4285-4288.

Warburg, O., and Christian, W. 1943. Isolierung und kristallization des garungsferments zymohexase. *Biochem. Z* 314: 149-176.

White, R.L., and Spencer, I.D. 1996. Biosynthesis of thiamin. In: *Escherichia coli* and *Salmonella.* Cellular and Molecular Biology. F.C. Neidhard *et al.,* eds. ASM Press, Washington. p. 680-686.

Xi, J., Ge, Y., Kinsland, C., McLafferty, F.W., and Begley, T.P. 2001. Biosynthesis of the thiazole moiety of thiamin in *Escherichia coli:* identification of an acyldisulfide-linked protein-protein conjugate that is functionally analogous to the ubiquitin/E1 complex. Proc. Natl. Acad. Sci. USA. 98: 8513-8518.