nature
biotechnology

Systematic discovery of analogous enzymes in thiamin biosynthesis

Enrique Morett[1,6], Jan O Korbel[2,3,6], Emmanuvel Rajan[1], Gloria Saab-Rincon[1], Leticia Olvera[1], Maricela Olvera[1], Steffen Schmidt[2–4], Berend Snel[2,5] & Peer Bork[2,3]

In all genome-sequencing projects completed to date, a considerable number of 'gaps' have been found in the biochemical pathways of the respective species. In many instances, missing enzymes are displaced by analogs, functionally equivalent proteins that have evolved independently and lack sequence and structural similarity. Here we fill such gaps by analyzing anticorrelating occurrences of genes across species. Our approach, applied to the thiamin biosynthesis pathway comprising approximately 15 catalytic steps, predicts seven instances in which known enzymes have been displaced by analogous proteins. So far we have verified four predictions by genetic complementation, including three proteins for which there was no previous experimental evidence of a role in the thiamin biosynthesis pathway. For one hypothetical protein, biochemical characterization confirmed the predicted thiamin phosphate synthase (ThiE) activity. The results demonstrate the ability of our computational approach to predict specific functions without taking into account sequence similarity.

Thiamin (vitamin B1) is a key nutrient for humans and other mammals, with a recommended daily dose of 1.4 mg and an annual industrial production of over 3,000 tons[1,2]. Its deficiency causes beriberi. Thiamin pyrophosphate (THI-PP) is the active form of this cofactor and has multiple functions, for example, in carbohydrate metabolism. The THI-PP biosynthesis pathway (**Fig. 1**; reviewed in refs. 3–7) has been studied in a variety of organisms including *Escherichia coli*, *Salmonella typhimurium*, *Bacillus subtilis* and *Saccharomyces cerevisiae*. Despite over 30 years of research on the pathway, not all of its catalytic steps are understood. Moreover, analysis of the genome sequences of microbial species that can grow on minimal media lacking thiamin did not reveal some of the characterized enzymes involved in even well-understood parts of the pathway (**Fig. 2a,b**).

Extensive sequence similarity searches are usually applied to fill these gaps, but such approaches are hampered by the displacement of proteins by analogous enzymes[8,9] (herein also referred to as gene displacement). Recently, several methods have been introduced that exploit the genomic context of a gene across various genomes to deduce functional information in the absence of sequence similarity. They analyze gene fusions[10,11], the co-occurrence of genes in putative operons[12,13] or the co-occurrence of genes across genomes[14] to predict 'functional associations' for the encoded protein—that is, which other protein it interacts with or which pathway it is involved in.

These methods exploit positive correlations such as the common occurrence of genes in a defined unit (such as an operon or genome). Here, we search for anticorrelations in the presence of genes across genomes. This approach does not predict functional associations but rather the displacement of functionally equivalent proteins, thereby deducing precise protein function. The underlying concept is that the mutually exclusive presence of genes (where if gene A occurs in a particular genome, then gene B is absent, and vice versa) indicates functional equivalence, as there is no need to encode the same function in a genome more than once. Lack of detectable homology within anticorrelating genes indicates that they might encode analogous enzymes.

The basic concept has been proposed previously[15,16]. Moreover, it was shown recently that a gene earlier reported to genetically complement the growth requirement of a thymidylate synthase (*thyA*)-deficient *Dictyostelium discoideum* strain[17], which also negatively correlates with *thyA* (ref. 15), indeed has thymidylate synthase activity[18]. However, anticorrelations are usually imperfect and thus cannot be retrieved 'by eye': pairs of analogous enzymes tend to be present or absent together in a fraction of genomes, and the anticorrelations may involve more than two genes. Moreover, the genes are typically absent in species lacking the corresponding pathway entirely. We have thus developed a computational approach to extract and evaluate the imperfect, or fuzzy, anticorrelations of gene occurrences across the various complete genomes available. A systematic examination of the THI-PP biosynthesis pathway demonstrates the predictive power of the method: so far, four of seven predicted displacements of functionally equivalent, but probably unrelated, enzymes have been tested and confirmed experimentally.

[1]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Av. Universidad 2001, Cuernavaca, Morelos, 62210, Mexico. [2]European Molecular Biology Laboratory, Meyerhofstraβe 1, 69117 Heidelberg, Germany. [3]Max-Delbrück-Center for Molecular Medicine, 13092 Berlin-Buch, Germany. [4]University of Heidelberg, Department of Parasitology, INF 324, 69120 Heidelberg, Germany. [5]Present address: Nijmegen Centre for Molecular Life Sciences, p/a CMBI, Toernooiveld 1, 6525 ED Nijmegen, Netherlands. [6]These authors contributed equally to this work. Correspondence should be addressed to E.M. (emorett@ibt.unam.mx) or P.B. (bork@embl-heidelberg.de).

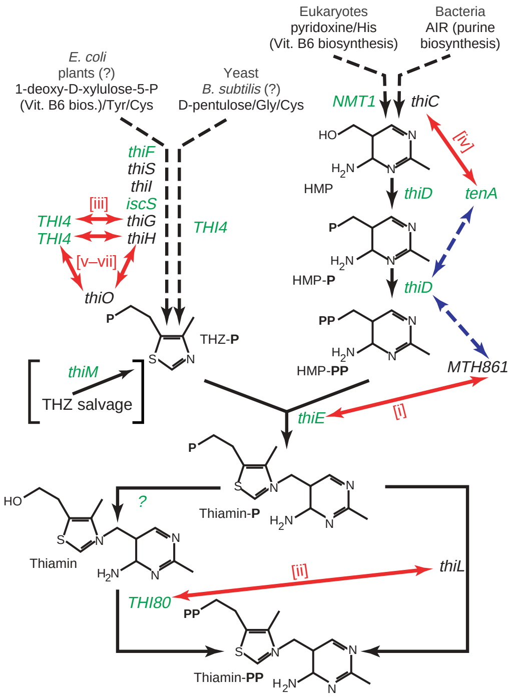

Figure 1 The THI-PP biosynthesis pathway[3–5,36]. Gene names are applied according to the first gene described from a group of orthologs (a table with synonyms can be found as **Supplementary Table 1** online). Solid black arrows represent known or proposed reaction steps and dashed black arrows indicate unknown reactions. In addition, significant anticorrelations in the occurrence of genes across species (red arrows), and relevant *in silico* predicted protein-protein interactions (blue dashed arrows) are illustrated. Distinct precursors have been proposed for different species[3–5] (indicated in gray). Genes with orthologous sequences[35] in eukaryotes and prokaryotes are in green; genes assumed to be prokaryote-specific are black. Interestingly, significant 'one-to-one' anticorrelations usually involve a prokaryote-specific and a 'ubiquitous' gene. Abbreviations: AIR, 5-aminoimidazole ribonucleotide; Cys, cysteine; Gly, glycine; His, histidine; HMP, 2-methyl-4-amino-5-hydroxymethylpyrimidine; THZ, 4-methyl-5-β-hydroxyethylthiazole; Tyr, tyrosine; Vit. B6, Vitamin B6.

RESULTS

Computational prediction of analogous enzymes

The prediction of analogous enzymes in the THI-PP biosynthesis pathway first required the definition of proteins associated with that biochemical process. There are 14 genes for which an essential role in the biosynthesis has been reported (**Fig. 2a,b**), and another 20 genes are predicted to be associated with the pathway by established gene context prediction methods (**Fig. 2c**; we use gene names according to the first described gene from a group of orthologs; capital letters denote genes originally isolated in eukaryotes). We applied a computational method to identify and score fuzzy, anticorrelating phylogenetic distributions within the 34 THI-PP–associated genes. Seven gene displacement predictions were found to be relevant (**Figs. 1** and **2**): four 'one-to-one relationships' (where one gene is directly displaced by another one) and three instances within a 'ternary relationship' (where each of three distinct genes may displace any of the others). These are: (i) the *thiE* gene encoding thiamin phosphate synthase (a gene first described in *E. coli*[19]) can be displaced by genes orthologous to the hypothetical *Methanobacterium thermoautotrophicum* ORF *MTH861* ('anticorrelation score' = 0.89, in a range from –1 to +1 with scores above 0.75 being significant (see Methods)); (ii) the *thiL* gene[20] (found in *S. typhimurium*) encoding thiamin phosphate kinase can be displaced by *THI80*, encoding thiamin pyrophosphokinase[21](score = 0.86), which was isolated from *S. cerevisiae*; (iii) the *E. coli thiG* gene[19] can be displaced by the *THI4* gene (score = 0.82) found in *S. cerevisiae*[22]; (iv) the *E. coli thiC* gene[19] can be displaced by *tenA* (score = 0.80), a *B. subtilis* gene reported to be involved (possibly indirectly) in transcriptional regulation[23]; (v–vii) *THI4, the thiH* gene originally described in *E. coli*[19]*,* and *thiO,* a gene from *Rhizobium etli* putatively involved in the pathway[24], can all displace each other ('ternary displacement'; score = 0.80). Although, for predictions (iii) to (vii), the precise enzymatic activities of the corresponding gene products remain to be characterized, we propose that they are functionally equivalent to the genes they displace.

Sequence and structural analysis of the predicted analogs

For further characterization of the candidates, we undertook sequence and structural analyses (see **Supplementary Methods** online), with the following results.

Prediction (i). We have used the designation *thiN* for the observed anticorrelation of *thiE* and *MTH861*. This name was also suggested by Rodionov *et al.*[25] while this manuscript was in preparation; the authors predicted displacement of *thiE* and *MTH861* following manual anticorrelation analysis. The anticorrelation is supported by the fusion of orthologs of MTH861 to the bifunctional enzyme ThiD (cloned and characterized[26] in *S. typhimurium*; ThiD catalyzes both phosphorylations from 2-methyl-4-amino-5-hydroxymethylpyrimidine (HMP to HMP-PP; see **Fig. 1**)) that is observed in some species. The 'ThiDN' fusion protein might thus catalyze three subsequent steps of the pathway. Although no significant sequence similarity between MTH861 and any known proteins was detected, secondary structure predictions do not rule out a similarity with the tertiary structure of ThiE[27].

Prediction (ii). ThiL and Thi80, which are predicted to be functionally equivalent, both catalyze the conversion to THI-PP, albeit in alternative reaction steps (**Fig. 1**). An ancient gene displacement seems plausible, assuming broader substrate specificity for at least one of the enzymes. Comparison of the known structure of Thi80 (ref. 28) with the predicted structure of ThiL (an enzyme homologous to aminoimidazole ribonucleotide synthetase) reveals different folds. Thus, two distinct folds might catalyze the same reaction.

Prediction (iii). The predicted gene displacement involving ThiG and Thi4 is supported by the role of both enzymes in the synthesis of 4-methyl-5-β-hydroxyethylthiazole phosphate (THZ-P[19,22]), involving a two-electron oxidation step. Notably, the proteins are similar to oxidases of different fold types: ThiG shows sequence similarity to TIM barrels and contains a putative FMN-binding motif, whereas Thi4, originally reported to be involved in DNA damage tolerance[29], has two predicted FAD/NAD(P)-binding domains (Rossmann fold).

Prediction (iv). We also predict functional equivalence between ThiC and TenA (**Fig. 2**). ThiC was implicated in HMP synthesis[19]. TenA was previously reported to be involved (possibly indirectly) in the transcriptional regulation of extracellular degradative enzymes[23]. As far as we know, no experimental evidence indicating a role for TenA in the THI-PP biosynthesis pathway has been reported. It is unlikely that these proteins are homologous, as a purely α-helical structure is predicted for TenA, whereas ThiC seems to be an α/β-protein.

Predictions (v–vii). The ternary anticorrelation observed for ThiH, ThiO and Thi4 suggests that each of the three proteins can displace any of the others. Like Thi4, the two other proteins are putative oxidases. Like Thi4, ThiH was shown to be involved in THI-PP biosynthesis (that is, in the synthesis of THZ-P)[19], but ThiO was only suggested to be similarly involved (the protein was predicted to be involved in the pathway based on its occurrence in a THI-PP biosynthesis operon[24]). ThiO is highly similar to D-amino acid oxidase (the catalytic residues appear to be conserved[24]). ThiO and Thi4 share putative Rossmann folds and limited sequence similarity, within a 30-amino-acid region matching a predicted NAD binding pattern. ThiH has an Fe-S cluster motif and is related to Fe-S oxidoreductases and synthases involved in other cofactor biosyntheses. Secondary structure predictions reveal a putative α/β-fold for ThiH. Although they may be homologous, all three putative enzymes have clearly evolved independently from the different families with which they share recognizable sequence similarity. The involvement of Thi4 in both a 'one-to-one' and a 'ternary' displacement prediction is not contradictory, as the enzyme could have more than one catalytic activity.

a

		Organisms putatively having a functional THI-PP biosynthesis pathway																																			THI-PP pathway unlikely										
Gene name	**COG id**	Cda	Sce	Afu	Ape	Hbs	Mja	Mth	Pab	Pho	Tac	Tvo	Aae	Bha	Bsu	Ccr	Cje	Dra	Eco	EcZ	Hin	Hpy	jHp	Lla	Mle	Mlo	Mtu	NmA	Nme	Pae	Pmu	Spy	Syn	Tma	Vch	Xfa	Bbu	Buc	Cpn	Ctr	Mge	Mpn	Rpr	Uur	Tpa	**Interact. with**	**STRING score**
thiD	COG0351	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	+	*thiE*	7
thiF	COG0476	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
thiS	COG2104	-	-	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	*thiG*	8
thiC	COG0422	-	-	+	-	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	*thiD,E*	3
thiE	COG0352	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	-	-	-	-	-	-	-	-	-	*thiD*	7
thiG	COG2022	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	*thiS*	8
thiH	COG1060*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-
thiL	COG0611	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	+	-	+	+	+	+	+	-	+	-	+	+	-	+	-	-	-	-	-	-	-	-	-

b

		Organisms putatively having a functional THI-PP biosynthesis pathway																																			THI-PP pathway unlikely										
Gene name	**COG id**	Cda	Sce	Afu	Ape	Hbs	Mja	Mth	Pab	Pho	Tac	Tvo	Aae	Bha	Bsu	Ccr	Cje	Dra	Eco	EcZ	Hin	Hpy	jHp	Lla	Mle	Mlo	Mtu	NmA	Nme	Pae	Pmu	Spy	Syn	Tma	Vch	Xfa	Bbu	Buc	Cpn	Ctr	Mge	Mpn	Rpr	Uur	Tpa	**Interact. with**	**STRING score**
THI80	COG1564	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	+	+	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-
thiO	COG0665*	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	*thiG*	4
THI4	COG1635	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
thiI	COG0301	-	-	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	+	+	+	-	-	+	-	-	-	-	-	+	+	+	-	+	+	-	-	-	-	-	+	+	-	+	+	-	-
iscS	COG1104	+	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	-	+	-	+	*thiI*	3
NMT1	COG0715*	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	+	+	+	+	-	-	-	-	+	-	-	-	+	+	-	+	+	+	-	-	-	+	+	-	-	-	-	-	-	-

c

		Organisms putatively having a functional THI-PP biosynthesis pathway																																			THI-PP pathway unlikely										
Gene name	**COG id**	Cda	Sce	Afu	Ape	Hbs	Mja	Mth	Pab	Pho	Tac	Tvo	Aae	Bha	Bsu	Ccr	Cje	Dra	Eco	EcZ	Hin	Hpy	jHp	Lla	Mle	Mlo	Mtu	NmA	Nme	Pae	Pmu	Spy	Syn	Tma	Vch	Xfa	Bbu	Buc	Cpn	Ctr	Mge	Mpn	Rpr	Uur	Tpa	**Interact. with**	**STRING score**
MTH861	COG1992	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	*thiD*	5
tenA	COG0819	+	+	-	+	-	-	-	+	+	+	+	-	+	+	-	+	-	-	-	+	+	+	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*thiD*	3
pspE	COG0607	+	+	+	-	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	+	-	-	*thiI*	4
iscU	COG0822	+	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	+	+	+	+	+	+	+	+	*iscS*	8
iscA	COG0316	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	-	+	+	+	+	-	-	-	+	+	+	+	+	+	+	-	+	-	+	+	-	+	-	-	-	-	+	-	-	*iscS*	7
yfhP	COG1959	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	-	-	+	-	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-	-	+	-	-	*iscS*	7
hscB	COG1076	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	+	-	-	-	-	+	-	+	+	+	+	-	+	-	+	-	-	+	-	-	-	-	+	-	-	*iscS*	4
lasT	COG0565	-	-	+	+	+	+	+	-	-	+	+	-	-	-	+	-	+	+	+	+	-	-	-	-	+	-	+	+	+	+	-	+	-	+	+	-	-	-	-	-	-	-	-	-	*iscS*	3
fdx	COG0633	+	+	+	-	+	-	-	-	-	-	-	+	-	-	+	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	-	+	-	+	-	-	+	+	+	-	-	+	-	-	*iscS*	3
yfhJ	COG2975	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	+	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	*iscS*	3
tauC	COG0600	-	-	+	-	-	+	-	-	-	+	+	-	+	+	+	-	+	+	+	+	-	-	-	-	+	-	-	-	+	+	-	+	+	-	+	-	-	+	+	-	-	-	-	-	*NMT1*	10
tauB	COG1116	-	-	+	-	-	+	-	-	-	+	+	-	+	+	+	-	+	+	+	+	-	-	-	-	+	-	-	-	+	+	-	+	+	-	+	-	-	-	-	-	-	-	-	-	*NMT1*	9
ribH	COG0054	-	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	-	-	-	-	-	*thiL*	5
ribC	COG0307	+	+	-	-	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	-	-	-	-	-	*thiL*	3
ribA	COG0807	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	-	-	-	-	-	*thiL*	3
ribB	COG0108	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	-	-	-	-	-	*thiL*	3
pgpA	COG1267	-	-	-	+	-	+	+	+	+	-	-	+	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	+	+	+	+	-	-	+	-	-	-	-	-	-	-	+	-	-	*thiL*	4
nusB	COG0781	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	*thiL*	5
moeA	COG0303	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	*thiF*	3
moaD	COG1977	-	-	+	-	+	-	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	-	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	*thiF*	3

Figure 2 Distribution across species of genes associated with THI-PP biosynthesis. (For species abbreviations, see **Supplementary Methods** online.) Negatively correlating gene occurrences are highlighted using the same colors. Species having at least two genes with a role unique to THI-PP biosynthesis[38] are predicted to possess the functional pathway. The column 'STRING score' shows the most significant interaction for each gene, predicted using the STRING server[32,33]. Predicted interaction partners are listed in the column 'Interact. with'. COG1060*, COG0665* and COG0715* represent refined orthologous groups (see Methods). (**a**) Essential THI-PP biosynthesis enzymes, which are unique to the pathway[38]. (**b**) Essential THI-PP biosynthesis enzymes, which have been implicated in more than one biological process[28,29,35,36,38]. The *thiO* gene, suggested to play a role in the pathway[24], was also added to that list. (**c**) Proteins predicted *in silico* to be involved in the pathway (see Methods).

Experimental verification of analogous proteins

To study the reliability of the gene displacement predictions, we tested experimentally the functional equivalence of the putative analogous genes from different species as predicted both from the 'one-to-one relationships' with the highest and lowest anticorrelation scores and from the more complex 'ternary displacement' prediction. For this purpose, we constructed three MC1061 *E. coli* strains by precise deletion of the *thiE*, *thiC* or *thiH* genes. The resulting strains (*E. coli ΔthiE*, *E. coli ΔthiC* and *E. coli ΔthiH*) showed thiamin auxotrophy that was specifically complemented by the respective gene in *trans*. In strains where the *thi* genes were completely removed, we never observed reversion (**Fig. 3** and data not shown).

We first proved the functional equivalence of the putative *thiE* analog. The hypothetical TM0790 was isolated by PCR from a total DNA preparation of *Thermotoga maritima*. The complete protein (a fusion protein comprising ThiD and a C-terminal sequence orthologous to MTH861, which we have designated ThiN) and its C-terminal domain (starting at position 196) were cloned into a multicopy plasmid. Both constructions efficiently complemented the thiamin auxotrophy of the *E. coli ΔthiE* strain. Interestingly, the complete protein promoted a slightly faster growth (**Fig. 3a**), perhaps because of its higher stability (data not shown). We then purified TM0790 to homogeneity and detected *in vitro* thiamin phosphate synthase (TPS) activity (**Fig. 4**). Unlike the *E. coli* TPS, the *T. maritima* protein was active at 72 °C (data not shown). The activity was enhanced by the addition of cell extract (prepared from *E. coli ΔthiE* to avoid contamination), suggesting that TM0790 has some extra substrate or cofactor requirements, although the addition of the TPS substrates HMP-PP and THZ-P increased the rate of THI-P formation (**Fig. 4**). Our results clearly show TPS activity for the novel ThiN protein, though we cannot exclude broader substrate specificity.

To test the 'one-to-one' gene displacement prediction with the lowest score, we isolated and cloned *B. subtilis tenA*, the gene that negatively correlates with *thiC*. The resulting clone complemented the growth of the *E. coli* Δ*thiC* strain on minimal medium without thiamin (**Fig. 3b**), such that isolated colonies were detected after 72 h of incubation. After genetic complementation using genes from distinct species, slow growth rates of the complemented strains are not unusual[30]. To confirm that the genetic complementation was due to functional equivalence of TenA and ThiC, we carried out two different experiments. In the first experiment, we isolated and cloned a second *tenA* ortholog, *PET18*, from *S. cerevisiae* (yeast has four genes with high sequence similarity to *tenA*: *PET18*, *THI20*, *THI21* and *THI22*; the latter three copies are fused to *thiD*). *PET18* complemented the thiamin requirement of the *E. coli ΔthiC* strain at about the same rate as *B. subtilis tenA* (**Fig. 3b**). In the second experiment, we assayed the specificity of the functional complementation. For this purpose the *E. coli ΔthiH* (**Fig. 3b**) and *ΔthiE* (data not shown) strains were transformed with both *B. subtilis tenA* and yeast *PET18* clones. Neither gene complemented the growth of these strains (**Fig. 3b,c**) even when the plates were incubated for more than 10 d. Although the catalytic activities of both TenA and ThiC have not been elucidated, the fact that *tenA* from both *B. subtilis* and yeast specifically complement *E. coli ΔthiC* strongly suggests functional equivalence of the proteins.

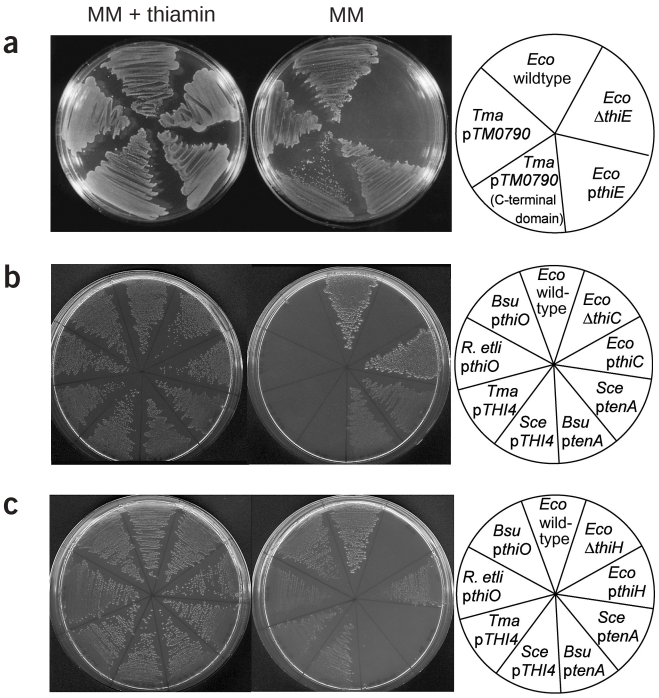

Figure 3 Genetic complementation of putative analogous enzymes. Strains (for abbreviations, see **Supplementary Methods** online) were grown on minimal medium (MM) or minimal medium supplemented with 10 µM thiamin (MM+thiamin). Colonies were visible after 1 d, except for the strains carrying the *tenA, THI4* or *thiO* genes, which were visible after 2–4 d. The untransformed *E. coli ΔthiE, E. coli ΔthiC,* and *E. coli ΔthiH* and the same strains transformed with pUC18 carrying the *E. coli thiE, thiC* or *thiH* genes were used as negative and positive controls, respectively. (**a**) Phenotypic complementation of the *E. coli ΔthiE* strain with *T. maritima TM0790* (which we have designated *thiDN*) and with its C-terminal domain (the sequence orthologous to *MTH861*). (**b**) Phenotypic complementation of *E. coli ΔthiC* with *tenA*, but not with *THI4* or *thiO*. (**c**) Phenotypic complementation of *E. coli ΔthiH* with *THI4* or *thiO*, but not with *tenA*.

To verify the predicted 'ternary displacement,' we followed a procedure similar to the one described for *thiC*, cloning the putative analogous genes from two phylogenetically distant species and demonstrating the specificity of the complementation. We cloned *thiO* from *R. etli* and its ortholog from *B. subtilis* (*yjbR*), as well as *THI4* from *S. cerevisiae* and its ortholog from *T. maritima* (*TM0787*). All four genes complemented the thiamin auxotrophy of the *E. coli ΔthiH* strain (**Fig. 3c**). The complementation was specific for *E. coli ΔthiH*, as neither *E. coli ΔthiC* (**Fig. 3b**) nor *E. coli ΔthiE* (data not shown) were complemented by any of these genes. The *THI4* genes complemented the growth of the *E. coli ΔthiH* faster than the *thiO* genes, developing colonies at about 36 h versus 48 h. ThiH, Thi4 and ThiO thus seem to be functional analogs (probably oxidases, given the results of the sequence analyses). It is remarkable to note that analogous genes from such diverse species as yeast, *T. maritima* and *B. subtilis* can efficiently complement the thiamin requirement of the different *E. coli* strains tested.

DISCUSSION

Taken together, the four experimental confirmations strongly suggest that fuzzy anticorrelations in the occurrence of genes across species are a good indicator for the displacement of missing enzymes by analogs. For the novel ThiN protein, we have demonstrated functional equivalence to ThiE biochemically, whereas for both ThiC and ThiH, two proteins with uncharacterized biochemical activities, we have shown genetic complementation with predicted analogs from two widely different species, respectively. A fifth prediction, functional equivalence of Thi4 and ThiO, also seems likely to be borne out because both proteins can genetically complement the lack of ThiH.

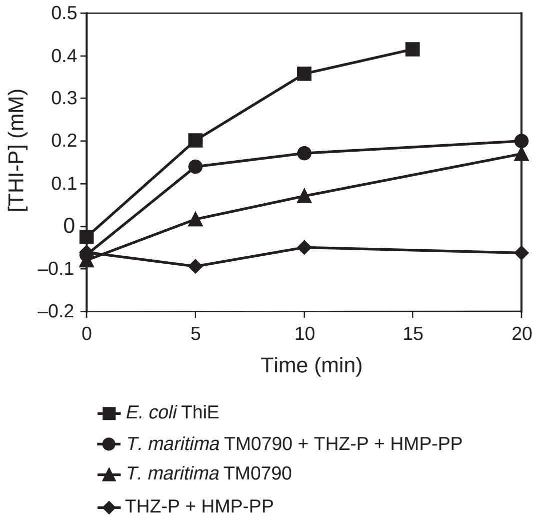

Figure 4 Thiamin phosphate synthase (TPS) activity of the thermophilic *T. maritima TM0790* gene product. 150 µl of freshly prepared cell extract of *E. coli ΔthiE* were added to a 500 µl mixture reaction and incubated at 55 °C.

In contrast to previous gene context–based strategies[10–14], the approach described here predicts exact enzymatic functions (that is, enzymes having identical Enzyme Commission numbers)—or at least functionally equivalent enzymatic reactions leading to the same compound (**Fig. 1**). Thus, the three genes *tenA*, *thiN* and *thiO* are not only generally assigned to the THI-PP biosynthesis pathway, but are placed into particular steps on the basis of their functional equivalence to known enzymes. As a result of our findings, in at least three prokaryotic species all missing steps of the pathway can now be filled in. Moreover, the fact that the yeast genes *THI4* and *PET18* complement auxotrophy caused by the deletion of prokaryotic genes suggests that early steps of thiamin biosynthesis might be more similar in eukaryotes and prokaryotes than currently believed[4,5].

The fuzziness of the bioinformatics method also reveals organisms that contain both anticorrelating genes. Such potential genetic redundancy has been described before: for instance, the *B. subtilis* genome encodes two analogous 3-dehydroquinases believed to catalyze the same reaction[31]. In regard to THI-PP biosynthesis, an interesting scenario is the presence of both *tenA* and *thiC* in *Pyrococcus abyssi*, *Campylobacter jejuni* and the genus *Bacillus.* Surprisingly, in these genomes, *tenA* but not *thiC* occurs with other THI-PP biosynthesis genes in putative operons[32,33], whereas *B. subtilis thiC* mutants showed a thiamin requirement[34]. We believe that this reflects a differential regulation of *thiC* and *tenA* as well as a low expression level of *tenA* in this species, and we predict that increased *tenA* expression can complement the growth defect.

Because there are probably about 15 enzymatic steps in the pathway (**Fig. 1**), and we predict that seven pairs of analogous enzymes are responsible for five independent activities, the rate of gene displacement must be considered substantial (on the order of 30%, at least in THI-PP biosynthesis). We expect that as the number of genomes sequenced increases (the 44 species included in the 'clusters of orthologous groups' (COG) database[35] were used here), more gene displacements will be detectable as additional anticorrelations become significant. It should then become possible to identify many more analogous proteins in a genome-wide analysis. Such analogs not only may represent promising targets for drug design (if one of the variants mainly occurs in pathogenic species), but also may open new avenues for finding catalysts for biotechnological processes.

METHODS

Genes and species associated with the pathway. We compiled a list of essential THI-PP biosynthesis genes (*i.e.*, those whose deletion causes auxotrophy) from the literature[3–5,36] (**Fig. 2a**,**b**). The putative *thiO* gene, previously suggested[3,24] to replace the function of *thiH* in *R. etli*, was also included. Additional genes likely to be involved in THI-PP biosynthesis (**Fig. 2c**) were collected using the initial collection as query in the STRING server[32,33]. Using STRING, we detected gene fusions and conserved operon structures across genomes, both of which predict involvement of genes in a common pathway[10–13,32,37] (see **Supplementary Methods** and **Supplementary Table 3** online).

The distributions of genes across 44 species were extracted from the COG database[35]. We assume that 35 species having at least two genes with a role unique to the THI-PP biosynthesis (as listed in EcoCyc[38]) are likely to possess the functional pathway (**Fig. 2a**; see **Supplementary Methods** online for more details).

Computational prediction of analogous enzymes. To predict instances in which known proteins have been displaced by analogs, we search for anticorrelating occurrences among the genes associated with THI-PP biosynthesis (**Fig. 2a–c**). An additional criterion was the absence of candidate genes in species lacking the pathway. 'Anticorrelation scores' were determined as follows. (i) In species having the pathway, a score of 1 is given when only one of the genes compared is present. (ii) If more than one of the genes is present, a penalty of –1 is assigned. (iii) If none of the genes is present in a species having the pathway, no penalty is given (to account for a true ortholog that escapes detection owing to sequence divergence, or for the presence of another nonorthologous gene carrying out the same function). (iv) In species that lack the pathway, a score of 1 is assigned, if none of the genes is present. (v) If one or more genes are present in a species lacking the pathway, a penalty of –1 is assigned. The anticorrelation score is defined as the sum of the scores divided by the number of species and thus ranges between –1 and 1. As a final filter, we require predictions to be each other's 'bidirectional best hit' (that is, if using A as a query gene results in the prediction of the top-scoring candidate B, the prediction is considered relevant only if using B as a query in turn reveals gene A).

To estimate the significance of a score, we calculated scores for all 5,010,195 possible pairs of orthologous groups in the COG database[35]. A score of 0.75 or higher was found in fewer than 0.5% of the comparisons and hence was considered as relevant. When applied to the genes associated with THI-PP biosynthesis, seven significant predictions have scores between 0.8 and 0.89, whereas the next best score is 0.59.

Splitting of inclusive COGs. Orthologous relationships are difficult to resolve if enzymes catalyzing distinct reactions share substantial sequence similarity. Accordingly, the COG database[35] lists some inclusive 'groups of orthologs' that contain largely paralogous genes. As suggested by the authors of COG, we revised orthologous relationships for genes with an assumed role in THI-PP biosynthesis (affected are the inclusive COG1060, COG0665 and COG0715; see **Supplementary Table 2** online).

Strain selection for genetic complementation. For gene replacement experiments, the wild-type MC1061 *E. coli* strain was used. All strains were grown in LB medium with the appropriate concentration of antibiotics and thiamin. All plasmid constructions were electroporated and propagated in the JM109 *E. coli* strain (see **Supplementary Methods** online for details of ligation, electroporation and PCR).

Construction of the *E. coli Δthi* strains. To obtain reliable *E. coli* mutants for genetic complementation, the *thiE*, *thiC* and *thiH* genes were precisely removed from the chromosome as described in the **Supplementary Methods**, **Supplementary Table 5** and **Supplementary Figures 1–6** online.

Cloning of predicted analogs. *T. maritima THI4* and *TM0790* (the complete gene and its *thiN* domain), *B. subtilis yjbR* and *tenA*, *S. cerevisiae THI4* and *PET18*, *R. etli thiO*, and *E. coli thiE*, *thiC* and *thiH* were isolated by PCR and cloned into the pUC18 vector (see **Supplementary Methods** and **Supplementary Table 4** online). The complete nucleotide sequence of each cloned gene was determined to confirm that no mutations had been introduced by the PCR procedure.

Determination of TPS activity of TM0790. A His-tag derivative of TM0790 was highly purified by affinity chromatography, and its TPS activity determined (for details on expression, purification and the TPS activity assay see **Supplementary Methods** online).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
This work was supported in part by grants from the Consejo Nacional de Ciencia y Tecnología (Mexico), Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México (Mexico), Netherlands Organization for Scientific Research, the Deutsche Forschungsgemeinschaft (Germany) and Bundesministerium für Forschung und Bildung (Germany). E.M. thanks the Alexander von Humboldt Stiftung. We thank R. Hernandez and P. Gaytan for technical assistance. We are indebted to T.P. Begley for providing HMP and a plasmid with the *thiD* gene from *E. coli* and to D.M. Downs, M. Soberón, J. Miranda and R. Russell for critical reading of the manuscript.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 21 February; accepted 2 April 2003
Published online 8 June 2003; doi:10.1038/nbt834

1. Burdick, D. in *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 25 (ed. Howe-Grant, M.) 152–171 (Wiley, New York, 1998).
2. Fenster, R. *Feed Additives: A Global Market Study* (PJB Publications, Richmond, Surrey, UK, 2001).
3. Begley, T.P. *et al.* Thiamin biosynthesis in prokaryotes. *Arch. Microbiol.* **171**, 293–300 (1999).
4. Hohmann, S. & Meacock, P.A. Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast *Saccharomyces cerevisiae*: genetic regulation. *Biochim. Biophys. Acta* **1385**, 201–219 (1998).
5. White, R.L. & Spenser, I.D. in *Escherichia coli* and *Salmonella. Cellular and Molecular Biology*, edn. 2, vol. 1 (ed. Neidhardt, F.C.) 680–686 (ASM Press, Washington, DC, 1996).
6. Xi, J., Ge, Y., Kinsland, C., McLafferty, F.W. & Begley, T.P. 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 (2001).
7. Allen, S., Zilles, J.L. & Downs, D.M. Metabolic flux in both the purine mononucleotide and histidine biosynthetic pathways can influence synthesis of the hydroxymethyl pyrimidine moiety of thiamine in *Salmonella enterica. J. Bacteriol.* **184**, 6130–6137 (2002).
8. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).
9. Koonin, E.V., Mushegian, A.R. & Bork, P. Non-orthologous gene displacement. *Trends Genet.* **12**, 334–336 (1996).
10. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
11. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
12. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
13. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).
14. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
15. Galperin, M.Y. & Koonin, E.V. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18**, 609–613 (2000).
16. Daugherty, M., Vonstein, V., Overbeek, R. & Osterman, A. Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J. Bacteriol.* **183**, 292–300 (2001).
17. Dynes, J.L. & Firtel, R.A. Molecular complementation of a genetic marker in *Dictyostelium* using a genomic DNA library. *Proc. Natl. Acad. Sci. USA* **86**, 7966–7970 (1989).

18. Myllykallio, H. *et al.* An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* **297**, 105–107 (2002).
19. Vander Horn, P.B., Backstrom, A.D., Stewart, V. & Begley, T.P. Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in *Escherichia coli* K-12. *J. Bacteriol.* **175**, 982–992 (1993).
20. Webb, E. & Downs, D. Characterization of thiL, encoding thiamin-monophosphate kinase, in *Salmonella typhimurium. J. Biol. Chem.* **272**, 15702–15707 (1997).
21. Nosaka, K., Kaneko, Y., Nishimura, H. & Iwashima, A. Isolation and characterization of a thiamin pyrophosphokinase gene, THI80, from *Saccharomyces cerevisiae. J. Biol. Chem.* **268**, 17440–17447 (1993).
22. Praekelt, U.M., Byrne, K.L. & Meacock, P.A. Regulation of THI4 (MOL1), a thiamine-biosynthetic gene of *Saccharomyces cerevisiae. Yeast* **10**, 481–490 (1994).
23. Pang, A.S., Nathoo, S. & Wong, S.L. Cloning and characterization of a pair of novel genes that regulate production of extracellular enzymes in *Bacillus subtilis. J. Bacteriol.* **173**, 46–54 (1991).
24. Miranda-Rios, J. *et al.* Expression of thiamin biosynthetic genes (thiCOGE) and production of symbiotic terminal oxidase cbb3 in *Rhizobium etli. J. Bacteriol.* **179**, 6887–6893 (1997).
25. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.* **277**, 48949–48959 (2002).
26. Petersen, L.A. & Downs, D.M. Identification and characterization of an operon in *Salmonella typhimurium* involved in thiamine biosynthesis. *J. Bacteriol.* **179**, 4894–4900 (1997).
27. Chiu, H.J., Reddick, J.J., Begley, T.P. & Ealick, S.E. Crystal structure of thiamin phosphate synthase from *Bacillus subtilis* at 1.25 A resolution. *Biochemistry* **38**, 6460–6470 (1999).
28. Baker, L.J., Dorocke, J.A., Harris, R.A. & Timm, D.E. The crystal structure of yeast thiamin pyrophosphokinase. *Structure (Camb)* **9**, 539–546 (2001).
29. Machado, C.R. *et al.* Dual role for the yeast THI4 gene in thiamine biosynthesis and DNA damage tolerance. *J. Mol. Biol.* **273**, 114–121 (1997).
30. Kim, Y.S. *et al.* A Brassica cDNA clone encoding a bifunctional hydroxymethylpyrimidine kinase/thiamin-phosphate pyrophosphorylase involved in thiamin biosynthesis. *Plant Mol. Biol.* **37**, 955–966 (1998).
31. Gourley, D.G. *et al.* The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nat. Struct. Biol.* **6**, 521–525 (1999).
32. Snel, B., Lehmann, G., Bork, P. & Huynen, M.A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).
33. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
34. Vandeyar, M.A. & Zahler, S.A. Chromosomal insertions of Tn917 in *Bacillus subtilis. J. Bacteriol.* **167**, 530–534 (1986).
35. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
36. Schwartz, C.J., Djaman, O., Imlay, J.A. & Kiley, P.J. The cysteine desulfurase, IscS, has a major role in *in vivo* Fe-S cluster formation in *Escherichia coli. Proc. Natl. Acad. Sci. USA* **97**, 9009–9014 (2000).
37. Huynen, M.A. & Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA* **95**, 5849–5856 (1998).
38. Karp, P.D. Pathway databases: a case study in computational symbolic theories. *Science* **293**, 2040–2044 (2001).