Systematic discovery of analogous enzymes in thiamin biosynthesis

Enrique Morett1,6, Jan O Korbel2,3,6, Emmanuvel Rajan1, Gloria Saab-Rincon1, Leticia Olvera1, Maricela Olvera1, Steffen Schmidt2–4, Berend Snel2,5 & Peer Bork2,3

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 tons1,2. 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 enzymes8,9 (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 fusions10,11, the co-occurrence of genes in putative operons12,13 or the co-occurrence of genes across genomes14 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 previously15,16. Moreover, it was shown recently that a gene earlier reported to genetically complement the growth requirement of a thymidylate synthase (thyA)-deficient Dictyostelium discoideum strain17, which also negatively correlates with thyA (ref. 15), indeed has thymidylate synthase activity18. 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.

1Instituto de Biotecnología, Universidad Nacional Autónoma de México, Av. Universidad 2001, Cuernavaca, Morelos, 62210, Mexico. 2European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany. 3Max-Delbrück-Center for Molecular Medicine, 13092 Berlin-Buch, Germany. 4University of Heidelberg, Department of Parasitology, INF 324, 69120 Heidelberg, Germany. 5Present address: Nijmegen Centre for Molecular Life Sciences, p/a CMBI, Toernooiveld 1, 6525 ED Nijmegen, Netherlands. 6These authors contributed equally to this work. Correspondence should be addressed to E.M. (morett@ibt.unam.mx) or P.B. (bork@embl-heidelberg.de).
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 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)\(^1\) 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 thiC gene\(^2\) (found in S. typhimurium) encoding thiamin phosphate kinase can be displaced by THI80, encoding thiamin pyrophosphokinase\(^3\) (score = 0.86), which was isolated from S. cerevisiae; (iii) the E. coli thiG gene\(^1\) can be displaced by the THI4 gene (score = 0.82) found in S. cerevisiae\(^2\); (iv) the E. coli thiC gene\(^1\) can be displaced by tenA (score = 0.80), a B. subtilis gene reported to be involved (possibly indirectly) in transcriptional regulation\(^2\); (v–vi) THI4, the thiH gene originally described in E. coli\(^1\), and thiO, a gene from Rhizobium etli putatively involved in the pathway\(^2\), 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.\(^2\) 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 ThiiD (cloned and characterized\(^6\) in S. typhimurium; ThiiD catalyzes both phosphorylations from 2-methyl-4-amino-5-hydroxymethylpyrimidine (HMP to HMP-PP; see Fig. 1)) that is observed in some species. The ‘ThiiDN’ 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\(^7\).

Prediction (ii). ThiL and ThiO, 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 ThiO (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-hydroxethylthiazole phosphate (THZ-β-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 barrel and contains a putative FMN-binding motif, whereas Thi4, originally reported to be involved in DNA damage tolerance\(^28\), 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 \(\alpha\)-amino acid oxidase (the catalytic residues appear to be conserved\(^{26}\)). ThiO and ThiH 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 \(\alpha/\beta\)-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 ThiH in both a ‘one-to-one’ and a ‘ternary’ displacement prediction is not contradictory, as the enzyme could have more than one catalytic activity.

**Experimental verification of analogous proteins**

To study the reliability of the gene displacement predictions, we tested experimentally the functional equivalence of the putative analogous enzymes from different species as predicted between the ‘one-to-one relationship’ with the highest and lowest anti-correlation 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 \(\Delta\)thiE, E. coli \(\Delta\)thiC and E. coli \(\Delta\)thiH) showed thiamin auxotrophy that was specifically complemented by the respective gene (Fig. 3a 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 that was specifically complemented by the respective gene (Fig. 3a). The putative ThiD was then purified to homogeneity and detected in vitro thiamin phosphate synthase (TPS) activity (Fig. 3). 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 ThiH 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 \(\Delta\)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; COG1060\(^{*}\), COG0665\(^{*}\) and COG0715\(^{*}\) represent refined orthologous groups (see Methods)). 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).
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, ΔthiC and 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 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.

In contrast to previous gene context–based strategies10–14, 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 believed4,5.

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 reaction31. 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 operons32,33, whereas B. subtilis thiC mutants showed a thiamin requirement34. 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

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.

© 2003 Nature Publishing Group  http://www.nature.com/naturebiotechnology
of orthologous groups' (COG) database 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 (Fig. 2a,b). The putative thiO gene, previously suggested to replace the function of thiH in E. coli, 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. Using STRING, we detected gene fusions and conserved operon structures across genomes, both of which predict involvement of genes in a common pathway. More gene distributions across 44 species were extracted from the COG database. We assume that 35 species having at least two genes with a role unique to the THI-PP biosynthesis (as listed in EcoCyc) 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. 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 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 jbbt and tena, S. cerevisiae THI4 and PET18, E. coli thiO, 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 6 June 2003; doi:10.1038/nbt834