Computational and Structural Biotechnology Journal 10 (2014) 41–50

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway

Valérie de Crécy-Lagard *

Department of Microbiology and Cell Science and Genetics Institute, University of Florida, Gainesville, FL, United States

ARTICLE INFO

Available online 11 June 2014

Keywords:
Non-orthologous displacements
Paralogs
Metabolic reconstruction

ABSTRACT

The availability of thousands of sequenced genomes has revealed the diversity of biochemical solutions to similar chemical problems. Even for molecules at the heart of metabolism, such as cofactors, the pathway enzymes first discovered in model organisms like *Escherichia coli* or *Saccharomyces cerevisiae* are often not universally conserved. Tetrahydrofolate (THF) (or its close relative tetrahydromethanopterin) is a universal and essential C_1-carrier that most microbes and plants synthesize *de novo*. The THF biosynthesis pathway and enzymes are, however, not universal and alternate solutions are found for most steps, making this pathway a challenge to annotate automatically in many genomes. Comparing THF pathway reconstructions and functional annotations of a chosen set of folate synthesis genes in specific prokaryotes revealed the strengths and weaknesses of different microbial annotation platforms. This analysis revealed that most current platforms fail in metabolic reconstruction of variant pathways. However, all the pieces are in place to quickly correct these deficiencies if the different databases were built on each other's strengths.

Contents

1. Introduction

In order to deal with the flood of data pouring from next-generation sequencing machines [1], robust and automated microbial genome annotation pipelines have become an acute necessity. The steps from gene calling to function prediction have been streamlined in annotation platforms, allowing laboratories with little bioinformatics capacity to annotate microbial genomes in a short amount of time [2–4]. Most of these pipelines base their function prediction calls on sequence similarity; however, this process is still far from perfect and high numbers of erroneous annotations remain [5–7]. Adding other types of information beyond sequence similarity, such as biological contexts by metabolic reconstruction, gene context by physical clustering, or phylogenetic conservation by co-distribution analyses, can greatly improve the quality of functional annotations [7–9]. These methods are slowly becoming part of the annotation pipelines [10,11], improving functional calls and also allowing the identification of gaps ('holes') also called "missing genes" in metabolic pathways [12,13]. Subsequent detailed comparative

* Department of Microbiology and Cell Science, University of Florida, P.O. Box 110700, Gainesville, FL 32611-0700, United States. Tel.: +1 352 392 9416; fax: +1 352 392 5922.
E-mail address: vcrecy@ufl.edu.

http://dx.doi.org/10.1016/j.csbj.2014.05.008
2001-0370/© 2014 de Crécy-Lagard.

genomics and experimental studies are then required to fill these pathway holes [14] because these are difficult to fill accurately using current automated gap-filling methods, even if a few success stories have been reported [15,16].

Tetrahydrofolate (THF) is a tripartite cofactor comprised of a pterin core attached to a *p*-aminobenzoate (*p*ABA) moiety and a glutamyl tail (Fig. 1). The THF synthesis pathway is complex and has been biochemically and genetically characterized extensively in *Escherichia coli* (black route in Fig. 1), with only one gene remaining to be identified (yellow highlight in Fig. 1). As the THF synthesis enzymes in yeast and *Arabidopsis thaliana* are very similar to the *E. coli* ones, this pathway was seen as an example of the uniformity of metabolism [17,18]. This view has been shattered with the advent of whole genome sequencing and the availability of thousands of genomes of diverse taxonomic origin has uncovered alternate solutions for nearly every step of the pathway. This diversity makes THF synthesis ideal to evaluate automated microbial functional annotation platforms. In this review, we provide a detailed description of all known pathway variations in THF synthesis in Bacteria and Archaea, then use these to check the annotations of the corresponding genes and the adequate calling of the THF pathway in the most common platforms used by experimentalists for gene functional annotation and pathway predictions (listed in Table 1).

1.1. Examples of non-orthologous displacements in the THF pathways

Several examples of non-orthologous enzymes catalyzing the same catalytic steps are found in the THF pathway. These can be analogous but non-homologous families, where totally different folds have been

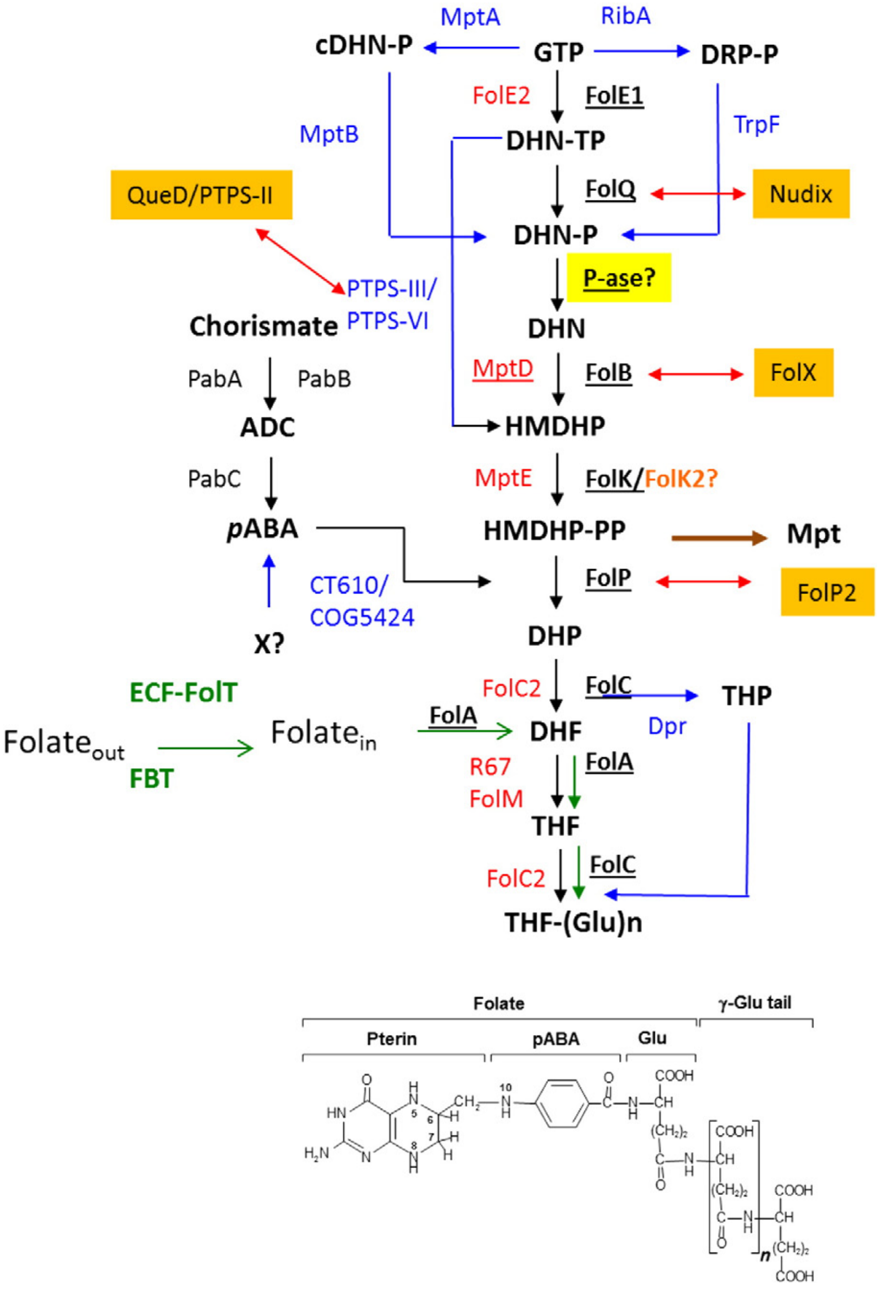


Fig. 1. Known variations and paralogs in the THF pathway. Code: underlined, canonical enzymes; red, non-orthologous displacements; blue, alternate pathways; green, salvage; yellow box, unknown gene; orange box, paralogs not in THF pathway; and orange, paralogs in folate pathway. Enzymes names are given in Table 2. Abbreviations: DHN-TP, dihydroneopterin triphosphate; DHN-MP, dihydroneopterin monophosphate; cDHNP, 7,8-dihydro-D-neopterin 2′,3′-cyclic phosphate; DRP-P, 2,5-diamino-6-ribosylamino-4(3*H*)-pyrimidinone 5′-phosphate; HMDHP, 6-hydroxymethyldihydropterin; HMDHP-PP 6-hydroxymethyldihydropterin diphosphate; *p*ABA, *p*-aminobenzoate; ADC, aminodeoxychorismate; DHP, dihydropteroate; THP, tetrahydropteroate; DHF, dihydrofolate; THF, tetrahydrofolate. THF-(Glu)$_n$, polyglutamylated THF; Mpt, methanopterin.

Table 1
Integrative microbial databases analyzed.

Database	Families[a]	Reactions[b]	Pathway reconstruction	Phenotype[c]	Location
Uniprot/Unipathway	Yes/HAMAP	Yes	Only a subset of genomes	No	http://www.uniprot.org/ http://www.unipathway.org
IMG	No	Yes	Yes	Yes	https://img.jgi.doe.gov/cgi-bin/w/main.cgi
PATRIC	Yes/FigFam	No	Yes	No	http://patricbrc.org/portal/portal/patric/Home
MicrobesOnline	No	No	Yes	No	http://www.microbesonline.org/
Microscope	Yes/Syntonome	Yes	Yes	No	http://www.genoscope.cns.fr/agc/microscope/home/index.php
BioCyc	No	Yes	Yes	No	http://biocyc.org/
KEGG	No	Yes	Yes	No	http://www.genome.jp/kegg/pathway.html
CMR	Yes/TIGRFAm	No	Yes	Yes	http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi

Abbreviations: High-quality Automated and Manual Annotation of Proteins (HAMAP); Integrated Microbial Genomes (IMG); PAThogen Resource Integration Center (PATRIC); Kyoto Encyclopedia of Genes and Genomes (KEGG); Comprehensive Microbial Resource (CMR).

[a] This is a family definition where annotations are going to be transferred to that family, not just a COG or Pfam membership, the name of these isofunctional families is listed.

[b] The chemical reaction is encoded in the database.

[c] Prediction of prototrophy or auxotrophy.

recruited to perform the same function, or very divergent members of the same superfamily.

The first step of the THF pathway is a complex reaction that transforms GTP into H_2-neopterin triphosphate (DHN-TP). The enzyme GTP cyclohydrolase I, encoded in *E. coli* by *folE*, catalyzes both the guanine ring cleavage and the subsequent Amadori rearrangement [19]. Around 20% of the bacteria that synthesize THF *de novo* lack a *folE* gene (Fig. 2A). In most of these organisms (11%, Fig. 2A), the same reaction is catalyzed by members of the COG1469 family, now called GTP cyclohydrolase IB or FolE2 [20]. Even though the FolE and FolE2 families are part of the Tunnel-Fold (or T-fold) superfamily, they have no detectable sequence similarities by BlastP [20].

Although some Archaea do use THF as C1-carrier, most use a very similar molecule, tetrahydromethanopterin (Mpt), that is synthesized through an analogous pathway, at least for the initial steps [21]. However, the reactions leading to the common 6-hydroxymethyl-7,8-dihydropterin diphosphate intermediate (HMDHP-PP) (Fig. 1) are catalyzed in Archaea by enzymes that are different from the bacterial ones. Only 28% of the THF/Mpt prototrophic Archaea use a FolE1 type GTP cyclohydrolase I while 72% use the FolE2 type (Fig. 2A). Similarly, the T-fold enzyme FolB has been replaced in 56% of these organisms by MptD, a member of the COG2098 family that bears no resemblance in sequence or structure to FolB (Fig. 2B) [22]. Finally, in most Archaea, the formation of HMDHP-PP is catalyzed by MptE from the TPK

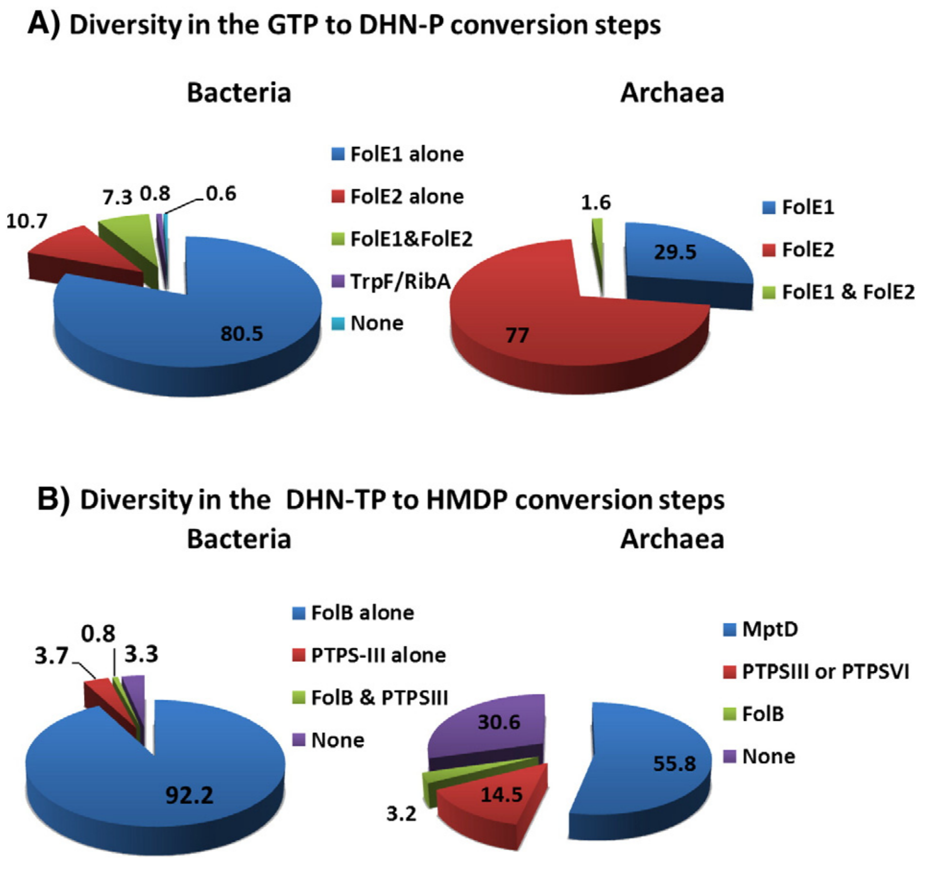


Fig. 2. Distribution of THF pathway variations. The percentage of each gene family was calculated based on a total number of 9327 bacteria that contained both a FolK and a FolB homolog and on 61 Archaea with an active THF or Mpt pathway. All the percentages in this figure, as well as those given throughout the text, were extracted from the "Folate biosynthesis" (http://pubseed.theseed.org/SubsysEditor.cgi?page=ShowSubsystem&subsystem=Folate_Biosynthesis.) and the "Early steps pterin biosynthesis Archaea" (http://pubseed.theseed.org/SubsysEditor.cgi?page=ShowSubsystem&subsystem=Early_Pterin_Biosynthesis_Steps_Archaea) SEED subsystems after downloading the corresponding excel files, eliminating duplicate genomes, and sorting using Excel tools. Abbreviations are defined in Table 2.

superfamily (COG1634) and not by FolK of the HPPK superfamily (COG0801) [22].

Another example of non-orthologous displacement is seen in 1% of the analyzed genomes (all Chlamydiae and a few Wolbachia species). In these bacteria, the enzyme that adds the glutamate moieties, a member of the FolC/COG0285 family in all other genomes analyzed, has been replaced by FolC2 [23], an enzyme homologous to the archaeal F420 glutamylation enzyme CofE [24] and part of the COG1478 family.

Finally four solutions are known to date for reducing the folate moiety from the dihydro to the tetrahydro form. Three are at the level of dihydrofolate (DHF): the canonical type I dihydrofolate reductase (DHFR) encoded in *E. coli* by *folA* [25], the type II trimethoprim R67 type DHFR (type II) found in both plasmids and integrons [26,27], and the short-chain dehydrogenase/reductase class DHFR encoded in *E. coli* by *folM* [28]. The fourth solution, consisting of an alternate route that reduces dihydropteroate (Fig. 1), is discussed below.

1.2. Variations in the THF pathway

Pathway variations, in contrast with non-orthologous displacement, use different chemical routes to get to the same end-points, and many variations are found in the THF/Mpt pathways. In approximately 20% of Archaea (mainly methanogens), the MptA subgroup of COG1469 performs a slightly different chemistry and produces 7,8-dihydroneopterin 2′,3′-cyclic phosphate [29] that needs that to be hydrolyzed by MptB [30] (Fig. 1). Another recently discovered variation in this first step of THF synthesis is the recruitment in Chlamydiae of enzymes of two other pathways to replace FolE and the DHN-TP pyrophosphatase (FolQ). The synthesis of DHN-P in these species relies on the first enzyme of riboflavin synthesis, GTP cyclohydrolase II (RibA) that produces 2,5-diamino-6-hydroxy-4-(5-phospho-D-ribosylamino)pyrimidine (DRP-P) that is then transformed into H_2-neoterin monophosphate (DHN-P) by the tryptophan synthesis enzyme phosphoribosylanthranilate isomerase (TrpF) [31] (Figs. 1 and 2A). Because the product of the RibA/TrpF pathway is DHN-P, and not DHN-TP [31], FolQ is not required in these bacteria (Fig. 1).

Another variation that eliminates FolQ is found at the next step of THF synthesis. The H_2-neopterin aldolase FolB (COG1539), found in *E. coli*, is absent in 7% of the bacterial genomes analyzed (Fig. 2). It has been replaced by the PTPS-III subgroups of the COG0720 family that directly converts the DHN-TP intermediate into dihydroneopterin (DHN), bypassing three steps of the standard pathway in 3.7% of the genomes analyzed [32,33] (Fig. 2B). This shunt is also found in Archaea, which use both PTPS-III and PTPS-VI variants, in 14.5% of the genomes analyzed (Fig. 2).

The three enzymes required to make *p*ABA from chorismate have been replaced by a single enzyme of the COG5424 family in nearly 2% of the genomes analyzed. This solution has been adopted also in chlamydial species [31,34], making them the most exotic bacteria in terms of folate synthesis, with three steps deviating from the canonical path [31].

Finally, a bypass of DHFR has been described where a flavin-dependent dihydropteroate reductase (Dpr) that can be fused to FolP or FolC leads to THF through tetrahydropteroate instead of dihydrofolate [35,36] (Fig. 1).

1.3. Paralogs of THF genes that have other functions

Many mistakes in gene annotations stem from over-annotation of paralog families [5], and there are several such cases in the THF pathways. FolX is a paralog of FolB found in 17% of the bacterial genomes analyzed. It was shown that FolX, in combination with FolM, is involved in the synthesis of another cofactor, monapterin [37], and hence, neither is a THF enzyme *stricto senso*. There are cases where FolM has replaced FolA in some organisms, but not in others. In 2% of folate prototrophs, FolM is present when both FolA and FolX are absent. However, in other organisms FolM is not involved in THF, but monapterin synthesis. Only by including physical context and co-distribution analysis can the functional calls be made.

The issue of paralogs is a major problem for the PTPS-III/VI enzymes that belong to the COG0720 family. This family contains enzymes catalyzing slightly different reactions in the synthesis of the tRNA modification queuosine (PTPS-I/QueD family) or the other pterin cofactor biopterin (PTPS-II family), in addition to the members involved in folate synthesis [32]. It was only by combining physical clustering with motif analysis that the different members of the COG0720 family could be correctly annotated, and this analysis revealed that some members of the family are actually bifunctional PTPS-III/QueD enzymes involved in both the folate and the queuosine pathways [32].

FolQ has been identified and experimentally validated in *E. coli*, *Arabidopsis thaliana* and *Lactococcus lactis* [38,39], however, the corresponding gene remains unknown in many species. FolQ is a member of the Nudix superfamily, whose members are notoriously difficult to annotate [40]. It is, therefore, difficult to propagate the annotation beyond genomes closely related to the ones where the function was experimentally validated; hence, the *folQ* gene is still missing in the majority of genomes (73%).

We had previously noted that duplications of both *folK* and *folP* are found in specific bacteria [23]. This was confirmed in the current analysis, as 8.5% of the genomes analyzed contained at least two *folK* genes and 3.5% contained at least two *folP* genes. In *Mycobacterium tuberculosis*, only *folP1*, and not *folP2*, is active in THF synthesis [41], but the function of *folP2* is still to be uncovered. Also, one of the *folK* paralogs of *Acinetobacter baylyi* is essential (ACIAD3062) and the other is not (ACIAD2407) [42], so further characterization is required to decipher the respective roles of these two genes in THF synthesis.

1.4. Identifying a signature gene for the THF/Mpt pathways

An important step to help in automatic metabolic reconstructions is to identify signature genes for each pathway. Ideally, this gene has to be part of an isofunctional family (no paralogs), easily identifiable by blast scores alone (no motif analysis required), and has to be found in all organisms that synthesize the specific metabolite (no bypasses or orthologous displacements). For THF and Mpt synthesis, it seems that there is no signature gene that is valid in all kingdoms. Indeed, in previous analyses [23], we had considered *folP* and *folK* to be good candidates for signature genes. However, in Archaea, FolK is replaced by MptE, and in Bacteria, both *folK* and *folP* have paralogs of unknown function (Fig. 1), so both should be eliminated as signature genes. To circumvent this problem, kingdom-specific criteria can be used to predict an active THF or Mpt pathway. Bacteria that harbor both a *folK* and *folP* homolog should have a complete *de novo* THF pathway. In Archaea, *mptE* seems to be a good signature gene for THF/Mpt synthesis pathways, as it is missing in just one organism [22].

1.5. Predicting folate salvage

Around 10% of the Bacteria analyzed lacked a *folP* or a *folK* homolog, and thus, should lack a *de novo* THF pathway. However, most of these organisms need THF, as only a handful of bacteria, mainly intracellular pathogens, lack all T-dependent enzymes [23]. Most bacteria that do not synthesize this cofactor *de novo* salvage some form of folate. These organisms must have active folate transporters, as well as FolA and FolC enzymes to produce the final active form of the cofactor [23] (green arrows in Fig. 1). Different families of folate transporters have been identified in bacteria. One is part of the folate-biopterin transporter (FBT) family [43], and the other, FolT [44], is part of the ECF family of transporters [45]. There are clearly other transporters to be discovered, as only 4% of organisms that require THF but lack the *de novo* pathway harbor FolT or FBT homologs.

1.6. Remaining missing genes and open questions in THF synthesis

One THF synthesis step has yet to be linked to a gene in any organism. This globally missing gene is the DHN-P phosphatase (P-ase, Fig. 1). It has been postulated that this activity is carried out by a non-specific phosphatase in *E. coli* [46], but clear evidence is lacking and this does not rule out the existence of a specific phosphatase yet to be identified. There still might be a few locally missing genes in specific genomes for the other steps of the pathway beyond the cases discussed above, but at this stage it is difficult to separate them from problems in gene calling (an example is given in Table 3). These require case-by-case analysis that should be automated, as it is becoming difficult to perform manually even with the SEED subsystem annotation tools that were designed for this purpose [47]. For FolB however, there must be other enzymes to be discovered, as 3% of Bacteria and 30% of Archaea still lack a known path to produce DHN. Also, as we had already noted [23], many cases of locally missing DHFRs remain and these have yet to be solved. Finally, in Archaea, the *p*ABA synthesis pathway remains a mystery [48]. Of course, as genomes from bacterial taxa that have never been sequenced become available [49], novel variations in the THF pathway could emerge.

1.7. Comparing specific folate pathway reconstructions and gene annotations in different platforms

Many integrated databases reconstruct metabolic pathways based on the presence/absence of pathway genes and a few predict whether a given organism can make a specific compound (Table 1). The THF pathway predictions in the model bacteria *E. coli* K12 MG1655 and in a handful of organisms that use one (or several) of the variations described in Fig. 1 were compared using the major annotation databases used by experimentalist for gene annotations and pathway predictions: KEGG [50,51], IMG/JGI [3], PATRIC [52], Microscope [4], Microbesonline [53], CMR [54], and BioCyc [55] databases. Uniprot/Unipathway was also added, as Uniprot [56] is mainly a protein annotation database, but has links to Unipathway, a resource that performs gene/reaction/pathway mapping [57]. The SEED database [2] was not included in this comparison, because the analysis would be biased, since the "Folate biosynthesis" subsystem of the SEED database was curated by the author and used to perform all the analyses in this review. The functional roles for seven specific genes were also collected for comparison in the same set of databases. All the queries and comparisons were performed using the regular web-based tools specific to each platform. The results of the analysis, shown in Table 3 and discussed in more detail below, reveal the strengths and weaknesses of the different platforms and suggest possible strategies to improve annotations and pathway reconstructions.

First, it is important to stress that retrieving this information required several hours, instead of the several minutes expected. Not all databases harbor the same sets of genomes, so close relatives had to be used. For example, some databases contain the *Halobacterium salinarum* R1 genome and not the *Halobacterium* sp. NRC1 genome, or *vice versa*. Not all databases consistently use locus tags as identifiers; some use a version with underscore, whereas others use a version without underscore (VNG1901c or VNG_1901c), vitiating the use of locus tags as universal identifiers. Some genes were not called because of an annotation mistake (*e.g.*, the MptD homolog in *H. salinarum* R1 can be found by tblastn but the gene is not called in this genome). The only way to quickly find target genes in KEGG was through the blast search entry point. Second, it is clear that, with the exception of Uniprot, the databases are not capturing new knowledge in a reasonable time frame. The only functional annotations that are correct in nearly all databases are the annotations of the FolE2/MptA homologs for which the original experimental work was published in 2006. The others, with original publication dates ranging from 2007 to 2013, are very poorly annotated.

Table 2
Enzymes of the THF pathways.

Abbreviation	Enzyme name	COG number
FolE	GTP cyclohydrolase I (EC 3.5.4.16) type 1	COG0302
FolE2 or MptA	GTP cyclohydrolase I (EC 3.5.4.16) type 2	COG1469
RibA	GTP cyclohydrolase II (EC 3.5.4.25)	COG0807
TrpF	Phosphoribosylanthranilate isomerase (EC 5.3.1.24)	COG0135
FolQ	Dihydroneopterin triphosphate pyrophosphatase	COG1051
Nudix	Nudix hydrolase superfamily	COG1051
P-ase	Dihydroneopterin monophosphate phosphatase	?
FolB	Dihydroneopterin aldolase (EC 4.1.2.25)	COG1539
PTPS-III	6-Hydroxymethyldihydropterin synthase, PTPS-III type	COG0720
PTPS-IV	6-Hydroxymethyldihydropterin synthase, PTPS-VI type	COG0720
QueD	6-Carboxytetrahydropterin synthase (EC 4.1.2.50)	COG0720
PTPS-II	6-Pyruvoyl tetrahydrobiopterin synthase (EC 4.2.3.12)	
MptB	7,8-Dihydro-D-neopterin 2′,3′-cyclic phosphate phosphodiesterase	COG3481
MptD	MptD, dihydroneopterin aldolase archaeal type	COG2098
FolK	FolK, hydroxymethyldihydropterin pyrophosphokinase (EC 2.7.6.3)	COG0801
MptE	Hydroxymethyldihydropterin pyrophosphokinase archaeal type	COG1634
FolP	Dihydropteroate synthase (EC 2.5.1.15)	COG0294
FolC	Bifunctional dihydrofolate synthase (EC 6.3.2.12) folylpolyglutamyl synthase (EC 6.3.2.17)	COG0285
FolC2	Bifunctional dihydrofolate synthase (EC 6.3.2.12) folylpolyglutamyl synthase (EC 6.3.2.17) type 2	COG1478
FolX	Dihydroneopterin triphosphate epimerase	COG1539
FolA	Dihydrofolate reductase type I	COG0262
FolM	Dihydropterin reductase	COG1028
R67	Dihydrofolate reductase type II	pfam06442
Dpr	Flavin-dependent dihydropteroate reductase	No named domain
PabA	Para-aminobenzoate synthase, aminase component (EC 2.6.1.85)	COG0147
PabB	Para-aminobenzoate synthase, amidotransferase component (EC 2.6.1.85)	COG0512
PabC	Aminodeoxychorismate lyase (EC 4.1.3.38)	COG0115
CT610	Alternate p*ABA* synthase	COG5424
FBT	Folate-biopterin transporter	COG2111
ECF-FolT	Substrate-specific component FolT of folate Energy coupling factor (ECF) transporter	pfam12822

When no "Cluster of Orthologous Group" (COG) number exists, a pfam number is given when it is available.

Table 3
Comparison THF pathway reconstruction and THF gene annotations in different annotations databases.

Correct Prediction	Year[a]	Uniprot/Unipathway	IMG	Patric	MicrobesOnline	Microscope	BioCyc	KEGG	CMR
THF pathway correctly predicted		**EcSaHs**		**EcSa**	**Ec**	**EcSa**	**EcSa**	**EcSa**	**Ec**
THF pathway incorrectly		Ct[b]	EcSaCbCtHs	CbCt[c]	SaCbCtHs[d]	CbCtHs[d]	BtCtHs	CbCtHs	SaCbCtHs
FolE2 (SACOL0613/Q5HIA9[e]) in Sa	2006	Yes	Yes	Yes	No	Yes	Yes	Yes	No
FolE2 (OE3673F/VNG1901c[f]/B0R6L9) in Hs	2006	Yes	Yes	NA[c]	Yes	Yes	Yes	Yes	No
FolC2 (CT611/CTA0664/Q3KL84) in Ct	2007	No	No[g]	Yes	No	No	No	No	No
QueD/PTPS-III (CBO0827/CLC_0882/A5I019) in Cb	2009	Only QueD	No[g]	Yes	No	No	No	No/PTPS-II	No/PTPS-II
MptD (VNG0127C/Q9HSQ4) in Hs	2012	Yes	No	NA[c]	No	No	[h]	No	No
MptE (OE2919R/VNG1343C/B0R5E3) in Hs	2012	Yes	No[i]	NA[c]	No	No	No	No	No
CT610 (CTA0663/Q3KL85) in Ct	2013	No[k]	No	No[j]	No	No	No	No	No

Abbreviations
Ct, *Chlamydia trachomatis* D/UW-3/CX (sv D) or *C. trachomatis* A/HAR-13 in BioCyc.
Cb, *Clostridium botulinum* A str. ATCC 3502 or Hall.
Ec, *Escherichia coli* K-12 MG1655.
Hs, *Halobacterium salinarum* R1 DSM 671 or *Halobacterium* species NRC-1.
Sa, *Staphylococcus aureus* subsp. aureus COL or *Staphylococcus aureus* subsp. aureus NCTC 8325.

[a] The year when protein/gene characterization was first published.
[b] Cb genome not in Unimap.
[c] No Archaea in PATRIC.
[d] The FolE2/MptA was correctly linked to the pathway in Hs.
[e] Uniprot numbers were used because locus tags were erratic for BioCyc.
[f] Some databases require the underscore version of the locus tag such as VNG_1901c.
[g] But correct SEED annotation on gene page.
[h] Cannot be evaluated as gene not called in *H. salinarum* R1 and *Halobacterium* NRC-1 not in BioCyc.
[i] But correct GO term on gene page.
[j] But correct reference on gene page.
[k] No *C. botulinum* in Unipathway.

Uniprot is the most accurate database with 4/7 correct annotations. One incorrect annotation is for CT610, for which a function was published only in August of 2013 [34]. The other for CT611/FolC2 was published in 2007, but it is difficult to capture, as the annotation was buried in the text of the paper and not mentioned in the abstract or keywords [23]. The third incorrect annotation is actually a miscalling of a dual function protein; the bifunctional QueD/PTPS-III protein of *Clostridium botulinum* was annotated only as QueD. Of note, multiple orthologs of the two genes MptD and MptE, which were characterized only in 2012, were annotated correctly in Uniprot. This shows that: 1) new annotations are captured within a period of six months to a year; 2) Uniprot's protein family annotation tool HAMAP (High-quality Automated and Manual Annotation of Proteins) [58] is very efficient and accurate in annotation propagation.

PATRIC was the second best in this analysis with 3/7 correct annotations, mostly because the annotation source is the SEED database that, for this pathway at least, has been continuously updated as the papers were published. Even though the roles were incorrectly predicted for all genes but the FolE2/MptA pair in all the other databases, correct SEED or GO annotations were visible on three of the IMG gene pages, and the correct 2013 reference was captured for CT610 in PATRIC. BioCyc was also able to capture the annotation on the MptD and MptE genes, but only in *Methanocaldococcus jannaschii*, the organism for which there was experimental evidence. This annotation was not transferred to its orthologs (Fig. 3B). Finally, the problem of paralogs was revealed in the erroneous annotation in KEGG and CMR of CBO0827 as a PTPS-II instead of a QueD/PTPS-III.

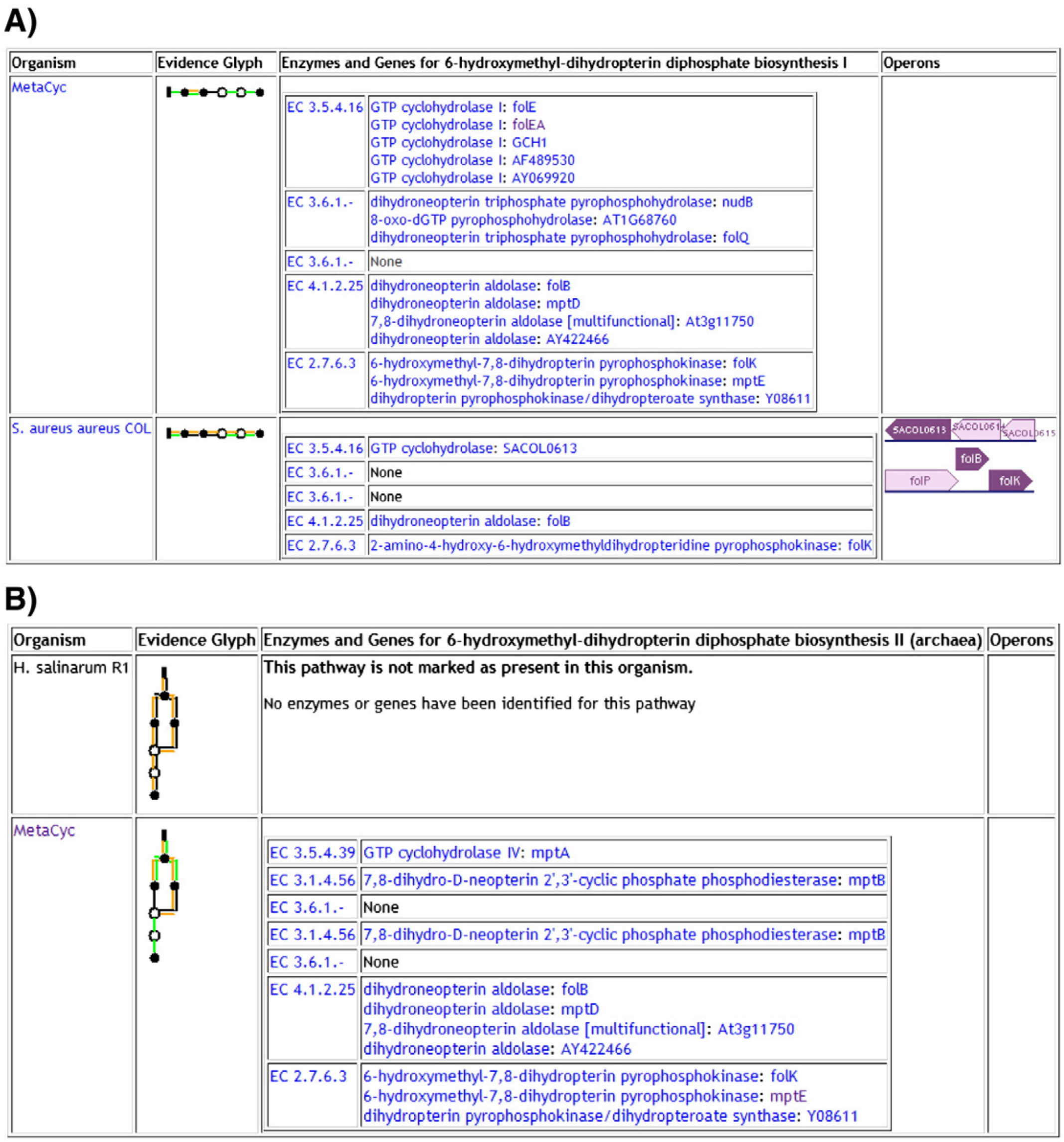
For pathway prediction, the results were quite disappointing for all databases except UniPathway. UniPathway correctly linked the gene to the pathways for all the genes it had correctly functionally called. However, its use as a pathway reconstruction database is limited by the small number of genomes covered (only 18 Archaea, 201 Bacteria for the folate pathway, with no *C. botulinum* genome). Also the way folate synthesis is split, with the first step in one sub-pathway (Unipath id: ULS00410) and all the other steps in another pathway (Unipath id: UPA00077), is confusing and does not allow the quick prediction of an active THF pathway in a given organism. UniPathway has been designed more as a resource to map genes to reactions to pathways in a structured way rather than to be used as a metabolic reconstruction tool.

All the other databases, with the exception of IMG (discussed below), correctly predicted the known enzymes of the *E. coli* THF pathway. Most databases, except IMG and Microbesonline, correctly predicted the known enzymes of the *S. aureus* pathway that uses the alternate *folE2* gene. The reasons for failure were as follows: either (i) a linkage between the pathway reconstruction and the gene annotation was not established (IMG) or (ii) the correct annotation had not been captured (Microbesonline). Also, Microbesonline and Microscope correctly included the MptA step in the *H. salinarum* R1 pathway.

The PATRIC database failed to correctly predict the *C. botulinum* PTPS-III or the *C. trachomatis* FolC2 in the reconstructed pathways, mainly because the functional roles in the annotations were not mapped to the reactions in the pathway, and this was also the major problem for all IMG reconstructions. Indeed, IMG has recently developed a series of tools ("Phenotypes") to address the very issue of capturing pathway variations in metabolic reconstruction and predict whether a pathway is active or not in a specific organism [59]. For example, one can retrieve the organisms that are predicted prototrophic or auxotrophic for specific amino acids. Unfortunately, the only vitamins for which phenotype predictions are available to date are biotin and coenzyme A, so this tool could not be evaluated using the THF pathway as a benchmark. However, several pathways that capture some of the variations in THF metabolism described here are encoded in IMG (Table 3 and Fig. 4) and have the potential to produce high quality annotations. Unfortunately, the failure to correctly link the genes with the reactions in these pathways is hampering this process, as even the *E. coli* THF pathway could not be correctly predicted (Fig. 4 and Table 4).

2. Discussion

In order to generate high quality metabolic reconstructions in microbial databases, several key features have to be implemented and correctly integrated. The first step is to capture functional annotations from the literature in a timely fashion (less than six months or a year). The most successful methods use professional curators (such as in Uniprot or BioCyc). SEED with its expert-based Subsystem annotation system seems also quite efficient. The second step is to accurately transfer the annotation from the experimentally validated gene/protein to its functional orthologs and this is efficiently done through well-curated



A)

Organism	Evidence Glyph	Enzymes and Genes for 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I		Operons
MetaCyc		EC 3.5.4.16	GTP cyclohydrolase I: folE GTP cyclohydrolase I: folEA GTP cyclohydrolase I: GCH1 GTP cyclohydrolase I: AF489530 GTP cyclohydrolase I: AY069920	
		EC 3.6.1.-	dihydroneopterin triphosphate pyrophosphohydrolase: nudB 8-oxo-dGTP pyrophosphohydrolase: AT1G68760 dihydroneopterin triphosphate pyrophosphohydrolase: folQ	
		EC 3.6.1.-	None	
		EC 4.1.2.25	dihydroneopterin aldolase: folB dihydroneopterin aldolase: mptD 7,8-dihydroneopterin aldolase [multifunctional]: At3g11750 dihydroneopterin aldolase: AY422466	
		EC 2.7.6.3	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase: folK 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase: mptE dihydropterin pyrophosphokinase/dihydropteroate synthase: Y08611	
S. aureus aureus COL		EC 3.5.4.16	GTP cyclohydrolase: SACOL0613	
		EC 3.6.1.-	None	
		EC 3.6.1.-	None	
		EC 4.1.2.25	dihydroneopterin aldolase: folB	
		EC 2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase: folK	

B)

Organism	Evidence Glyph	Enzymes and Genes for 6-hydroxymethyl-dihydropterin diphosphate biosynthesis II (archaea)		Operons
H. salinarum R1		This pathway is not marked as present in this organism. No enzymes or genes have been identified for this pathway		
MetaCyc		EC 3.5.4.39	GTP cyclohydrolase IV: mptA	
		EC 3.1.4.56	7,8-dihydro-D-neopterin 2′,3′-cyclic phosphate phosphodiesterase: mptB	
		EC 3.6.1.-	None	
		EC 3.1.4.56	7,8-dihydro-D-neopterin 2′,3′-cyclic phosphate phosphodiesterase: mptB	
		EC 3.6.1.-	None	
		EC 4.1.2.25	dihydroneopterin aldolase: folB dihydroneopterin aldolase: mptD 7,8-dihydroneopterin aldolase [multifunctional]: At3g11750 dihydroneopterin aldolase: AY422466	
		EC 2.7.6.3	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase: folK 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase: mptE dihydropterin pyrophosphokinase/dihydropteroate synthase: Y08611	

Fig. 3. BioCyc pathway comparisons. (A) The "enzymes and genes for 6-hydromethyl dihydropterin diphosphate biosynthesis II (archaea)" pathway was opened in BioCyc (BioCyc: http://biocyc.org/), and the pathway comparison tool was used choosing *S. aureus* COL in the genome list (B) The "Enzymes and genes for 6-hydromethyl dihydropterin diphosphate biosynthesis II (Archaea)" pathway was opened in BioCyc and the pathway comparison tool was used choosing *H. salinarum R1* in the genome list.

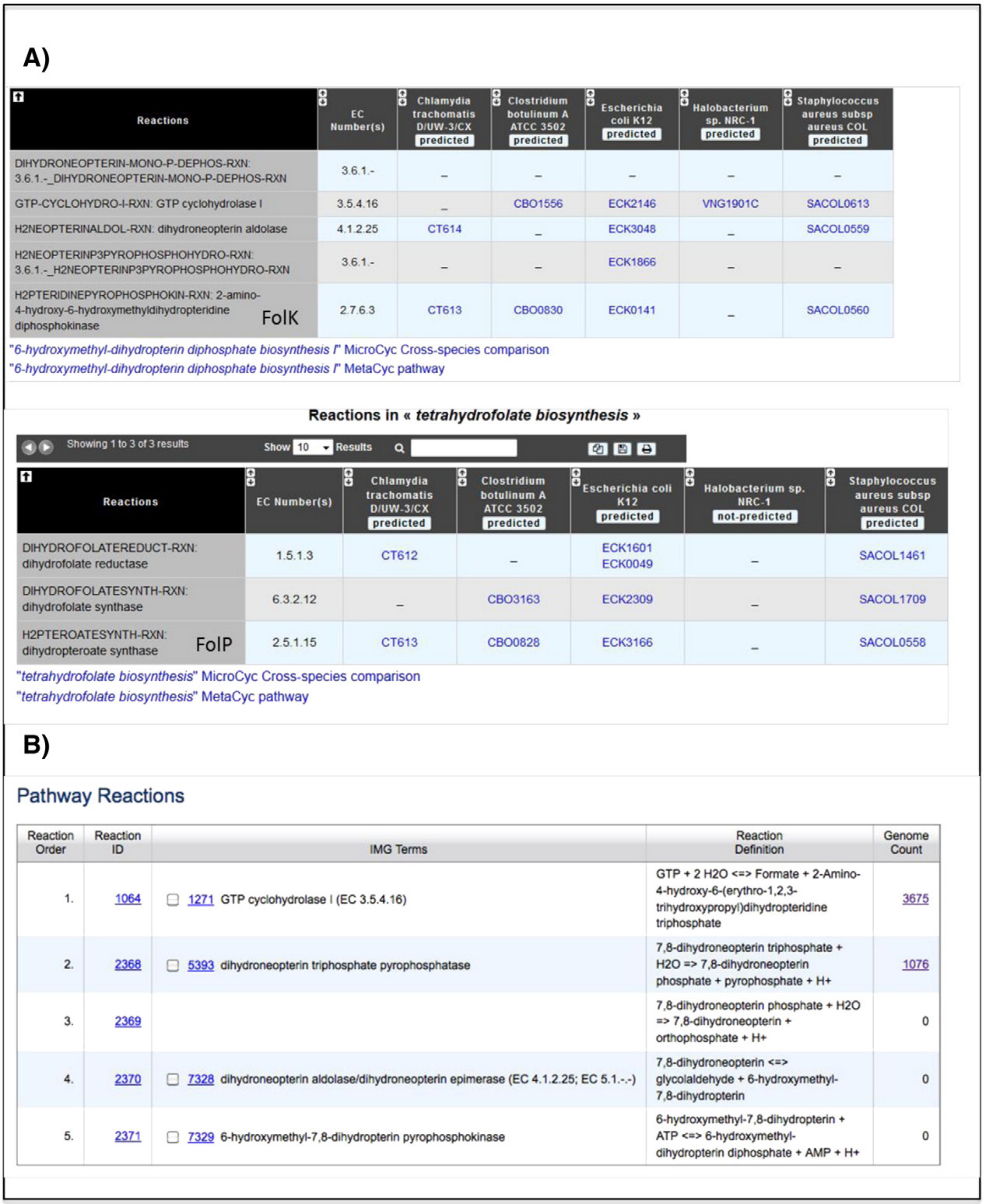
families such as FigFam, HAMAP or Syntonome (Table 1). In parallel, pathway variations need to be encoded. First, formalized as alternate enzymes and variant codes in the initial SEED manifesto (http://www.theseed.org/wiki/Annotating_1000_genomes) and paper [47], these are now being efficiently captured in a formal way as specific variant pathways in BioCyc or IMG (Table 4, Figs. 3 and 4C). The final step is to accurately link gene annotations to a pathway reaction, and it seems that all databases, except PATRIC, are set up to do this if the gene is correctly annotated.

Uniprot seems to be the only database analyzed that combines an efficient literature capture capacity with an accurate propagation of the annotation in a format that is recognized by the Unipathway tool. BioCyc seems to both capture the literature efficiently and create up-to-date pathway variants, but clearly lacks tools to propagate annotations among orthologs (Fig. 3). The other databases, such as Microscope and IMG, that have powerful metabolic reconstruction platforms and pathway curation could capitalize on the Uniprot HAMAP rules and curation power to quickly improve the quality of their predicted pathways.

Another key feature, required to accurately predict whether a pathway is active or not in a given genome, is the use of signature genes. As discussed above, many folate genes are globally or locally missing. A call can still be made on the capacity of an organism to synthesize this cofactor or not. Based on the signature suggested above for bacteria (the presence of FolK and FolP), one can predict that all four bacterial genomes analyzed here are prototrophic (see Fig. 4A), even if not all the genes have been called and these genes could be flagged automatically as locally missing. CMR is the only database that is close to using signature genes, correctly calling the pathways as active even when not all the genes have been identified (Table 5). Missing genes are identified by differential coloring on KEGG pathways in most databases, but only Microscope specifically generates the list of missing genes for a given genome (although this tool only works erratically).

3. Conclusion

The diversity of metabolic solutions observed in the THF pathway is not unique. It is more the rule than an exception, and this had been



Reactions	EC Number(s)	Chlamydia trachomatis D/UW-3/CX predicted	Clostridium botulinum A ATCC 3502 predicted	Escherichia coli K12 predicted	Halobacterium sp. NRC-1 predicted	Staphylococcus aureus subsp aureus COL predicted
DIHYDRONEOPTERIN-MONO-P-DEPHOS-RXN: 3.6.1.-_DIHYDRONEOPTERIN-MONO-P-DEPHOS-RXN	3.6.1.-	–	–	–	–	–
GTP-CYCLOHYDRO-I-RXN: GTP cyclohydrolase I	3.5.4.16	–	CBO1556	ECK2146	VNG1901C	SACOL0613
H2NEOPTERINALDOL-RXN: dihydroneopterin aldolase	4.1.2.25	CT614	–	ECK3048	–	SACOL0559
H2NEOPTERINP3PYROPHOSPHOHYDRO-RXN: 3.6.1.-_H2NEOPTERINP3PYROPHOSPHOHYDRO-RXN	3.6.1.-	–	–	ECK1866	–	–
H2PTERIDINEPYROPHOSPHOKIN-RXN: 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase FolK	2.7.6.3	CT613	CBO0830	ECK0141	–	SACOL0560

"*6-hydroxymethyl-dihydropterin diphosphate biosynthesis I*" MicroCyc Cross-species comparison
"*6-hydroxymethyl-dihydropterin diphosphate biosynthesis I*" MetaCyc pathway

Reactions in « *tetrahydrofolate biosynthesis* »

Reactions	EC Number(s)	Chlamydia trachomatis D/UW-3/CX predicted	Clostridium botulinum A ATCC 3502 predicted	Escherichia coli K12 predicted	Halobacterium sp. NRC-1 not-predicted	Staphylococcus aureus subsp aureus COL predicted
DIHYDROFOLATEREDUCT-RXN: dihydrofolate reductase	1.5.1.3	CT612	–	ECK1601 ECK0049	–	SACOL1461
DIHYDROFOLATESYNTH-RXN: dihydrofolate synthase	6.3.2.12	–	CBO3163	ECK2309	–	SACOL1709
H2PTEROATESYNTH-RXN: dihydropteroate synthase FolP	2.5.1.15	CT613	CBO0828	ECK3166	–	SACOL0558

"*tetrahydrofolate biosynthesis*" MicroCyc Cross-species comparison
"*tetrahydrofolate biosynthesis*" MetaCyc pathway

Pathway Reactions

Reaction Order	Reaction ID	IMG Terms	Reaction Definition	Genome Count
1.	1064	□ 1271 GTP cyclohydrolase I (EC 3.5.4.16)	GTP + 2 H2O <=> Formate + 2-Amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl)dihydropteridine triphosphate	3675
2.	2368	□ 5393 dihydroneopterin triphosphate pyrophosphatase	7,8-dihydroneopterin triphosphate + H2O => 7,8-dihydroneopterin phosphate + pyrophosphate + H+	1076
3.	2369		7,8-dihydroneopterin phosphate + H2O => 7,8-dihydroneopterin + orthophosphate + H+	0
4.	2370	□ 7328 dihydroneopterin aldolase/dihydroneopterin epimerase (EC 4.1.2.25; EC 5.1.-.-)	7,8-dihydroneopterin <=> glycolaldehyde + 6-hydroxymethyl-7,8-dihydropterin	0
5.	2371	□ 7329 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase	6-hydroxymethyl-7,8-dihydropterin + ATP <=> 6-hydroxymethyl-dihydropterin diphosphate + AMP + H+	0

Fig. 4. (A) Microscope pathway species comparisons. (B) Description of IMG pathway and linkage of reactions to genes.

noted as soon as whole genome sequencing data allowed to branch away from classical models [8] and is now starting to be formalized in "phylometabolic" analyses [60]. The difficulty, as shown here, is that a great amount of manual curation is currently required to capture these variations that are not currently captured in annotation databases in any robust fashion. That said, it seems all the pieces are in place to capture these metabolic variants in the near future if increased collaboration and integration between the platforms occurs. A database that could encode IMG or BioCyc pathways, capture the Uniprot HAMAP based annotations, as well as the expert/Subsystem based SEED annotations, and make sure the link between the functional roles and the pathway reactions is made would be close to the level required for accurate metabolic reconstructions. Until the databases improve, a "naïve user" wanting to know if a specific pathway is found in a specific organism should: 1) always check several databases, not just one; 2) identify the signature gene(s) of a given pathway and check for their presence/absence in the genome; and 3) systematically check the recent literature on that pathway in all organisms, as any new enzyme of the pathway published in the last year will certainly have been missed.

Table 4
IMG THF-related pathway assertions. None of the pathways were correctly predicted, including IMG-1005 and IMG-1006 in *E. coli*. Analysis was performed on IMG/JGI: (https://img.jgi.doe.gov/cgi-bin/w/main.cgi) by adding the five chosen genomes (abbreviations given in Table 1) and the selected IMG pathways (listed with IMG numbering) in the cart, and then conducting the IMG pathway distribution analysis in these genomes.

IMG pathway	Ct	Cb	Ec	Hs	Sa
1005 — 6-hydroxymethyl-dihydropterin diphosphate biosynthesis	a(0/5)	a(1/5)	a(2/5)	a(0/5)	a(1/5)
1006 — Tetrahydrofolate biosynthesis	a(0/3)	a(1/3)	a(2/3)	a(1/3)	a(2/3)
1032 — Folate precursors biosynthesis in Archaea	a(0/2)	a(0/2)	a(0/2)	a(1/2)	a(0/2)
1034 — Tetrahydromonapterin biosynthesis *via* folX diversion	a(0/3)	a(0/3)	a(0/3)	a(0/3)	a(0/3)
1037 — 6-hydroxymethyl-dihydropterin diphosphate biosynthesis *via* PTPS-III bypass reaction	a(0/3)	a(1/3)	a(1/3)	a(0/3)	a(1/3)

Assertion: a — absent or not asserted; p — present or asserted; u — unknown; N/A — no data available.
Evidence level (g/R): g — number of reactions with associated genes; R — total number of reactions in pathway.

Table 5
Prediction of the state of the folate biosynthesis pathway for specific organisms in CMR.

Organism	State	FolE (O)	FolB (R)	FolK (R)	FolP (R)	FolC (R)	FolA (R)
Staphylococcus aureus subsp. aureus COL	Yes	N	Y	Y	Y	Y	Y
Escherichia coli K12-MG1655	Yes	Y	Y	Y	Y	Y	Y
Clostridium botulinum A Hall	Some evidence	Y	N	Y	Y	Y	N
Halobacterium sp. NRC-1	Not supported	N	N	N	Y	Y	N
Chlamydia trachomatis serovar D	Some evidence	N	Y	Y	Y	N	Y
Evidence		TIGR00063	TIGR00525	TIGR01498	TIGR01496	TIGR01499	PF00186

R — required, signature gene; O — not required, not signature gene.

Acknowledgments

This work was funded by the US National Science Foundation (Grant MCB-1153413). I thank Andrew Hanson for providing a template for Fig. 1 and for the comments on this work. I thank Jennifer Thiaville for editing the manuscript and Svetlana Gerdes for her general input. I am grateful to Ross Overbeek for his vision on genome annotations and for building the SEED infrastructure, without which none of this work would have been possible.

References

[1] Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2012;40:D571–9.
[2] Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 2014;42:D206–14.
[3] Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 2014;42:D560–7.
[4] Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, et al. MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res 2013;41:D636–47.
[5] Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
[6] Bork P, Bairoch A. Go hunting in sequence databases but watch out for the traps. Trends Genet 1996;12:425–7.
[7] Frishman D. Protein annotation at genomic scale: the current status. Chem Rev 2007;107:3448–66.
[8] Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 2003;7:238–51.
[9] Plata G, Fuhrer T, Hsiao T-L, Sauer U, Vitkup D. Global probabilistic annotation of metabolic networks enables enzyme discovery. Nat Chem Biol 2012;8:848–54.
[10] Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 2010;11:40–79.
[11] Henry CS, Overbeek R, Xia F, Best AA, Glass E, Gilbert J, et al. Connecting genotype to phenotype in the era of high-throughput sequencing. Biochim Biophys Acta 2011;1810:967–77.
[12] Green M, Karp P. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinforma 2004;5:76.
[13] Rolfsson O, Palsson B, Thiele I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. BMC Syst Biol 2011;5:155.
[14] Hanson AD, Pribat A, Waller JC, de Crécy-Lagard V. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list—and how to find it. Biochem J 2009;425:1–11.
[15] Smith AAT, Belda E, Viari A, Medigue C, Vallenet D. The CanOE Strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comput Biol 2012;8:e1002540.
[16] Gerdes S, El Yacoubi B, Bailly M, Blaby IK, Blaby-Haas CE, Jeanguenin L, et al. Synergistic use of plant-prokaryote comparative genomics for functional annotations. BMC Genomics 2011;12(Suppl. 1:S2).
[17] Cossins EA, Chen L. Folates and one-carbon metabolism in plants and fungi. Phytochemistry 1997;45:437–52.
[18] Hanson AD, Gregory Iii JF. Synthesis and turnover of folates in plants. Curr Opin Plant Biol 2002;5:244–9.
[19] Nar H, Huber R, Auerbach G, Fischer M, Hosl C, Ritz H, et al. Active site topology and reaction mechanism of GTP cyclohydrolase I. Proc Natl Acad Sci U S A 1995;92:12120–5.
[20] El Yacoubi B, Bonnett S, Anderson JN, Swairjo MA, Iwata-Reuyl D, de Crécy-Lagard V. Discovery of a new prokaryotic type I GTP cyclohydrolase family. J Biol Chem 2006;281:37586–93.
[21] Graham DE, White RH. Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. Nat Prod Rep 2002;19:133–47.
[22] de Crécy-Lagard V, Phillips G, Grochowski LL, Yacoubi BE, Jenney F, Adams MWW, et al. Comparative genomics guided discovery of two missing archaeal enzyme families involved in the biosynthesis of the pterin moiety of tetrahydromethanopterin and tetrahydrofolate. ACS Chem Biol 2012;7:1807–16.
[23] de Crécy-Lagard V, El Yacoubi B, de la Garza R, Noiriel A, Hanson A. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. BMC Genomics 2007;8:245.
[24] Li H, Graupner M, Xu H, White RH. CofE catalyzes the addition of two glutamates to F420-0 in F420 coenzyme biosynthesis in *Methanococcus jannaschii*. Biochemistry 2003;42:9771–8.
[25] Rood JI, Laird AJ, Williams JW. Cloning of the *Escherichia coli* K-12 dihydrofolate reductase gene following mu-mediated transposition. Gene 1980;8:255–65.
[26] Narayana N, Matthews DA, Howell EE, Nguyen-huu X. A plasmid-encoded dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D2-symmetric active site. Nat Struct Biol 1995;2:1018–25.
[27] White PA, Rawlinson WD. Current status of the aadA and dfr gene cassette families. J Antimicrob Chemother 2001;47:495–6.
[28] Giladi M, Altman-Price N, Levin I, Levy L, Mevarech M. FolM, a new chromosomally encoded dihydrofolate reductase in *Escherichia coli*. J Bacteriol 2003;185:7015–8.
[29] Grochowski LL, Xu H, Leung K, White RH. Characterization of an Fe^{2+}-dependent Archaeal-specific GTP Cyclohydrolase, MptA, from *Methanocaldococcus jannaschii*. Biochemistry 2007;46:6658–67.
[30] Mashhadi Z, Xu H, White RH. An Fe^{2+}-dependent cyclic phosphodiesterase catalyzes the hydrolysis of 7,8-dihydro-D-neopterin 2′,3′-cyclic phosphate in methanopterin biosynthesis. Biochemistry 2009;48:9384–92.
[31] Adams NE, Thiaville JJ, Proestos J, Juárez-Vázquez AL, Barona-Gómez F, Dirk Iwata-Reuyl D, et al. Promiscuous and adaptable enzymes fill "holes" in the tetrahydrofolate pathway in Chlamydia. mBio 2014 [in press].
[32] Pribat A, Jeanguenin L, Lara-Nunez A, Ziemak MJ, Hyde JE, de Crécy-Lagard V, et al. 6-Pyruvoyltetrahydropterin synthase paralogs replace the folate synthesis enzyme dihydroneopterin aldolase in diverse bacteria. J Bacteriol 2009;191:4158–65.
[33] Phillips G, Grochowski LL, Bonnett S, Xu H, Bailly M, Blaby-Haas C, et al. Functional promiscuity of the COG0720 family. ACS Chem Biol 2012;7:197–209.
[34] Satoh Y, Kuratsu M, Kobayashi D, Dairi T, et al. New gene responsible for para-aminobenzoate biosynthesis. J Biosci Bioeng 2013.
[35] Levin I, Giladi M, Altman-Price N, Ortenberg R, Mevarech M. An alternative pathway for reduced folate biosynthesis in bacteria and halophilic archaea. Mol Microbiol 2004;54:1307–18.
[36] Levin I, Mevarech M, Palfey BA. Characterization of a novel bifunctional dihydropteroate synthase/dihydropteroate reductase enzyme from *Helicobacter pylori*. J Bacteriol 2007;189:4062–9.

[37] Pribat A, Blaby IK, Lara-Nunez A, Gregory III JF, de Crécy-Lagard V, Hanson AD. FolX and FolM are essential for tetrahydromonapterin synthesis in *Escherichia coli* and *Pseudomonas aeruginosa*. J Bacteriol 2010;192:475–82.

[38] Gabelli SB, Bianchet MA, Xu W, Dunn CA, Niu ZD, Amzel LM, et al. Structure and function of the *E. coli* dihydroneopterin triphosphate pyrophosphatase: a Nudix enzyme involved in folate biosynthesis. Structure 2007;15:1014–22.

[39] Klaus SM, Wegkamp A, Sybesma W, Hugenholtz J, Gregory III JF, Hanson AD. A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. J Biol Chem 2005;280:5274–80.

[40] McLennan A. Substrate ambiguity among the nudix hydrolases: biologically significant, evolutionary remnant, or both? Cell Mol Life Sci 2013;70:373–85.

[41] Gengenbacher M, Xu T, Niyomrattanakit P, Spraggon G, Dick T. Biochemical and structural characterization of the putative dihydropteroate synthase ortholog Rv1207 of *Mycobacterium tuberculosis*. FEMS Microbiol Lett 2008;287:128–35.

[42] de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, et al. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. Mol Syst Biol 2008;4.

[43] Klaus SM, Kunji ER, Bozzo GG, Noiriel A, de la Garza RD, Basset GJ, et al. Higher plant plastids and cyanobacteria have folate carriers related to those of trypanosomatids. J Biol Chem 2005;280:38457–63.

[44] Eudes A, Erkens GB, Slotboom DJ, Rodionov DA, Naponelli V, Hanson AD. Identification of genes encoding the folate- and thiamine-binding membrane proteins in Firmicutes. J Bacteriol 2008;190:7591–4.

[45] Rodionov DA, Hebbeln P, Eudes A, ter Beek J, Rodionova IA, Erkens GB, et al. A novel class of modular transporters for vitamins in prokaryotes. J Bacteriol 2009;191:42–51.

[46] Green JC, Nichols BP, Matthews RG. Folate biosynthesis, reduction, and polyglutamylation. In: Neidhart FC, editor. *Escherichia coli* and *Salmonella*. , Cellular and molecular biologyWashington, DC: American Society for Microbiology; 1996. p. 665–73.

[47] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33:5691–702.

[48] Porat I, Sieprawska-Lupa M, Teng Q, Bohanon FJ, White RH, Whitman WB. Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*. Mol Microbiol 2006;62:1117–31.

[49] Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 2009;462:1056–60.

[50] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40:D109–14.

[51] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42:D199–205.

[52] Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 2014;42:D581–91.

[53] Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, et al. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res 2010;38:D396–400.

[54] Davidsen T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, et al. The comprehensive microbial resource. Nucleic Acids Res 2010;38:D340–5.

[55] Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2014;42:D459–71.

[56] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.

[57] Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, et al. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res 2012;40:D761–9.

[58] Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, et al. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res 2013;41:D584–9.

[59] Chen IM, Markowitz VM, Chu K, Anderson I, Mavromatis K, Kyrpides NC, et al. Improving microbial genome annotations in an integrated database context. PLoS ONE 2013;8:e54859.

[60] Braakman R, Smith E. The emergence and early evolution of biological carbon-fixation. PLoS Comput Biol 2012;8:e1002455.