The road to knowledge: from biology to databases and back again

Stobbe, M.D.

Publication date
2012

Chapter 4

Improving the description of metabolic networks: the TCA cycle as example

Miranda D. Stobbe[1,4,*], Sander M. Houten[3,*], Antoine H.C. van Kampen[1,2,4,5], Ronald J.A. Wanders[3], Perry D. Moerland[1,4]

[1] Bioinformatics Laboratory, Academic Medical Center, University of Amsterdam, P.O. Box 22700, 1100 DE, Amsterdam, the Netherlands
[2] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands
[3] Laboratory Genetic Metabolic Diseases, Departments of Clinical Chemistry and Pediatrics, Academic Medical Center, University of Amsterdam, P.O. Box 22700, 1100 DE, Amsterdam, the Netherlands
[4] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, the Netherlands
[5] Netherlands Consortium for Systems Biology, University of Amsterdam, P.O. Box 94215, 1090 GE, Amsterdam, the Netherlands

*These authors contributed equally.

Accepted for publication in: *FASEB Journal* (May 21, 2012)

Abstract

To collect the ever increasing, yet scattered knowledge on metabolism, multiple pathway databases, like the Kyoto Encyclopedia of Genes and Genomes, have been created. A complete and accurate description of the metabolic network for human and other organisms is essential to foster new biological discoveries. Previous research has shown, however, that the level of agreement between pathway databases is surprisingly low. We investigated if the lack of consensus between databases can be explained by an inaccurate representation of the knowledge described in scientific literature. As an example, we focus on the well-known tricarboxylic acid (TCA) cycle and evaluated the description of this pathway as found in a comprehensive selection of ten human metabolic pathway databases. Remarkably, none of the descriptions given by these databases is entirely correct. Moreover, there is consensus on only three reactions. Mistakes in pathway databases might lead to the propagation of incorrect knowledge, misinterpretation of high-throughput molecular data, and poorly designed follow-up experiments. We provide an improved description of the TCA cycle via the community-curated database WikiPathways. We review various initiatives that aim to improve the description of the human metabolic network and discuss the importance of the active involvement of biological experts in these.

Introduction

Metabolism has been studied for decades already and the interest in this topic is going through a marked revival (DeBerardinis and Thompson, 2012; Hanahan and Weinberg, 2011). To collect our ever increasing, but scattered knowledge on metabolism, pathway databases, like the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*, 2012), have been created. The number of pathway databases describing the metabolic network of one or more organisms is growing rapidly (Karp and Caspi, 2011; Oberhardt *et al*, 2009). For many organisms there are even multiple databases available describing their metabolic network. These databases provide a holistic view of the metabolic network (Oberhardt *et al*, 2009) and are routinely used to provide context for the analysis and interpretation of high-throughput molecular data (Rey *et al*, 2011). *In silico* models of the metabolic network can also be used to generate experimentally verifiable hypotheses, such as potential drug targets, or to simulate the effect of network perturbations, such as loss of function. Recent research has shown, however, that the level of agreement between the metabolic network descriptions of the same organism given by the various pathway databases is surprisingly low (Herrgård *et al*, 2008; Radrich *et al*, 2010; Stobbe *et al*, 2011). For example, five pathway databases that each describe the human metabolic network were shown to agree on only 199 (about 3%) of the close to 7,000 reactions they have combined (Stobbe *et al*, 2011). Databases differ in the way they retrieved information to build the metabolic network and in the way the network is curated. For example, *Homo sapiens* Recon 1 (Duarte *et al*, 2007) was built by first automatically generating a preliminary network based on, genome annotation and the reactions from KEGG (Kanehisa *et al*, 2012). Next, the network was manually refined using literature and computer simulations. In contrast, Reactome (Croft *et al*, 2011) takes an incremental approach, regularly adding new parts to its network, which are curated by selected experts and peer-reviewed.

Here, we discuss to what extent this lack of consensus between the databases is caused by an inaccurate representation of the knowledge described in scientific literature. To answer this question, we choose, as an example, the well-known tricarboxylic acid (TCA) cycle, a well-studied pathway that has been a subject of extensive research ever since its discovery by Hans Krebs in 1937 (Krebs and Johnson, 1937). Furthermore, this pathway can be found in virtually every student text book about biochemistry. One would expect the description of the TCA cycle in pathway databases to be highly accurate and, hence, a high level of agreement between these databases.

We evaluated the description of this pathway as found in a comprehensive selection of ten public human metabolic pathway databases (Table 1). Although pathway databases have certainly proven their value in a wide range of applications, we show that none of the selected ten descriptions is entirely correct based on a thorough review of the literature. Using the knowledge contained in the ten databases and additional scientific literature, we provide an improved description of the TCA cycle. The observations made for the TCA cycle are not unique to this pathway, but are also valid for the entire metabolic network. To further improve upon the description of the entire (human) metabolic network, various initiatives have been implemented to which the community of biological experts can contribute. We conclude by outlining some of these initiatives and discuss the challenges ahead.

Results

We retrieved the descriptions of the TCA cycle from a comprehensive set of ten databases (Table 1) and compared the descriptions to each other to identify differences. Figure 1 displays the union of all reactions from the ten databases and shows that there is consensus on only three reactions and two of the corresponding catalysts. The number of steps in which a conversion is described, such as the formation of D-*threo*-isocitrate from citrate, is one explanation of a difference between the databases. Next, two experts in the field of metabolism (SH and RW) compared the knowledge described in the literature with the descriptions from the ten pathway databases. Relevant literature was extracted from Medline based on MeSH terms and keywords related to the TCA cycle. We observed that many inconsistencies in the databases are explained by an inaccurate representation of the knowledge described in scientific literature (Table 2 and 3). In some cases, conclusive evidence in literature was lacking, referred to henceforth as 'unconfirmed'.

Database	**Version**	**URL for TCA cycle pathway**
BioCarta	March 2001	http://www.biocarta.com/pathfiles/krebPathway.asp
EHMN	2	http://www.ehmn.bioinformatics.ed.ac.uk/?se=rea&sefor=rea&seterm=&pa=17
H. sapiens Recon 1	1	http://bigg.ucsd.edu/
HumanCyc	15.1	http://humancyc.org/human/new-image?type=PATHWAY&object=PWY-5690
INOH	4.0	http://www.inoh.org/download.html#MetabolicPathway
KEGG	59	http://www.genome.jp/kegg-bin/show_pathway?org_name=hsa&mapno=00020
Panther	2.1	http://www.pantherdb.org/pathway/pathwayDiagram.jsp?catAccession=P00051
Reactome	37	http://www.reactome.org/entitylevelview/PathwayBrowser.html#DB=test_reactome_37&FOCUS_SPECIES_ID=48887&FOCUS_PATHWAY_ID=71406&ID=71403&
SMPDB	1.0	http://pathman.smpdb.ca/pathways/SMP00057/pathway
UniPathway	2010_05	http://www.grenoble.prabi.fr/obiwarehouse/unipathway/upa?upid=UPA00223&oscode=HUMAN

Table 1 – Pathway databases. Ten pathway databases from which we retrieved their description of the TCA cycle, if possible a direct link to the TCA cycle is provided.

Below we discuss the outcome of our literature study and the inconsistencies observed in the databases in more detail (Figure 2, capital letters below refer to the different panels). The resulting improved description of the TCA cycle is illustrated in the blue boxes. This description is based on literature, the knowledge captured by the ten databases combined and our own expertise. In contrast to some databases, all reactions in our description are mass and charge balanced. In addition, we determined the protonation state of the metabolites at the pH level of the mitochondrion, which is estimated to be between 7.8 and 8.0 (Hoek *et al*, 1980; Llopis *et al*, 1998; Porcelli *et al*, 2005).

Citrate synthase (A)

The TCA cycle starts with the condensation of oxaloacetate and acetyl-CoA by the enzyme citrate synthase. Overall, the different databases have a high degree of agreement with respect to this reaction. Some disagreement exists over the reversibility of this reaction. Evidence in the literature shows that in rat liver mitochondria, radiolabeled citrate did not label products of mitochondrial acetyl-CoA metabolism. This shows that the citrate synthase reaction is irreversible in vivo (Greksak *et al*, 1982) .

Aconitase (B)

In this reversible reaction citrate is converted into D-*threo*-isocitrate. The reaction proceeds via the intermediate cis-aconitate. Since cis-aconitate is not the final product of this reaction, the necessity of adding it to the model could be questioned. Indeed some databases do not include this intermediate step. We argue that an intermediate should be included in a model when there is evidence available that the metabolite is a true intermediate that can be accurately measured enabling the use of the concentration of this metabolite in mathematical models (Hoppe *et al*, 2007). A second reason to include the intermediate is to be able to model the accumulation of the metabolite under pathologic conditions such as an inborn error of metabolism. Cis-aconitate meets the first criterion, as it is readily measured as a product of the aconitase reaction (Krebs and Holzach, 1952), but also in body fluids using organic acid analysis (Lawson *et al*, 1976). Therefore, we decided to include cis-aconitate in our description.

Isocitrate dehydrogenase (C and D)

The biochemistry associated with the reactions catalyzed by the two mitochondrial isocitrate dehydrogenase enzymes IDH2 and IDH3 is complex, which may explain some of the discrepancies between databases. The main difference between IDH2 and IDH3 enzyme is at the level of the electron acceptor, with the latter using NAD

Databases	enzyme and encoding gene(s)							
	missing[a]		incorrect		unconfirmed[a]		complex not indicated[a]	
	n	Fig. 2	*n*	Fig. 2	*n*	Fig. 2	*n*	Fig. 2
BioCarta	**13**	*c (3x), d*, e (2x), f (2x), g, h (3x), j*	**3**	*c, f, j*	**0**	-	**5**	*c, e, f, g, h*
EHMN	**1**	*g*	**3**	*b, d, j*	**1**	*e*	**5**	*c, e, f, g, h*
H. sapiens Recon 1	**0**	-	**5**	*b (2x), d, e, j*	**1**	*j*	**0**	-
HumanCyc	**3**	*d*, f * (2x)*	**3**	*a, b, j*	**2**	*e (2x)*	**1**	*f**
INOH	**2**	*d*, g*	**2**	*b, j*	**0**	-	**4**	*c, e, f, g*
KEGG	**0**	-	**5**	*b, d, f, g, j*	**1**	*e*	**5**	*c, e, f, g, h*
Panther	**12**	*c (3x), d*, e (2x), f, g* (2x), h (2x), j*	**3**	*c (2x), j*	**3**	*e (2x), j*	**5**	*c, e, f, g*, h*
Reactome	**0**	-	**2**	*f, g*	**0**	-	**0**	-
SMPDB	**3**	*d*, g* (2x)*	**0**	-	**0**	-	**1**	*g**
Unipathway	**13**	*c (3x), d, e (3x), f* (2x), g, h (2x), j*	**0**	-	**0**	-	**5**	*c, e, f*, g, h*

Table 2 – Overview of inconsistencies per pathway database: enzymes and encoding genes. For each database the number of times (*n*) a specific inconsistency was found is indicated and, whenever possible, linked to a specific panel of Figure 2 via letters a-j. [a] Excluding reactions not considered to be part of the TCA cycle in our description. * Enzyme (complex) is missing because the indicated reaction is not described in the database.

Databases	Reaction						
	not part of the TCA cycle	missing		direction incorrect[a]		inclusion of enzyme-bound intermediates	
		n	Fig. 2	*n*	Fig. 2	*n*	Fig. 2
BioCarta	**0**	**1**	*d*	**0**	-	**1**	*h*
EHMN	**16**	**0**	-	**4**	*a, d, e, h*	**1**	*e*
H. sapiens Recon 1	**2**	**0**	-	**0**	-	**1**	*h*
HumanCyc	**0**	**2**	*d, f*	**2**	*a, b*	**0**	-
INOH	**3**	**1**	*d*	**2**	*c, h*	**1**	*e*
KEGG	**8**	**0**	-	**3**	*a, c, h*	**1**	*e*
Panther	**1**	**2**	*d, g*	**4**	*b, f, i, j*	**1**	*c*
Reactome	**1**	**0**	-	**3**	*d, f, g*	**1**	*h*
SMPDB	**2**	**2**	*d, g*	**0**	-	**1**	*h*
Unipathway	**7**	**1**	*f*	**5**	*b, d, g, i, j*	**0**	-

Table 3 – Overview of mistakes per pathway database: reactions. For each database the number of times (*n*) a specific inconsistency was found is indicated and, whenever possible, linked to a specific panel of Figure 2. [a] Excluding reactions not considered to be part of the TCA cycle in our description.

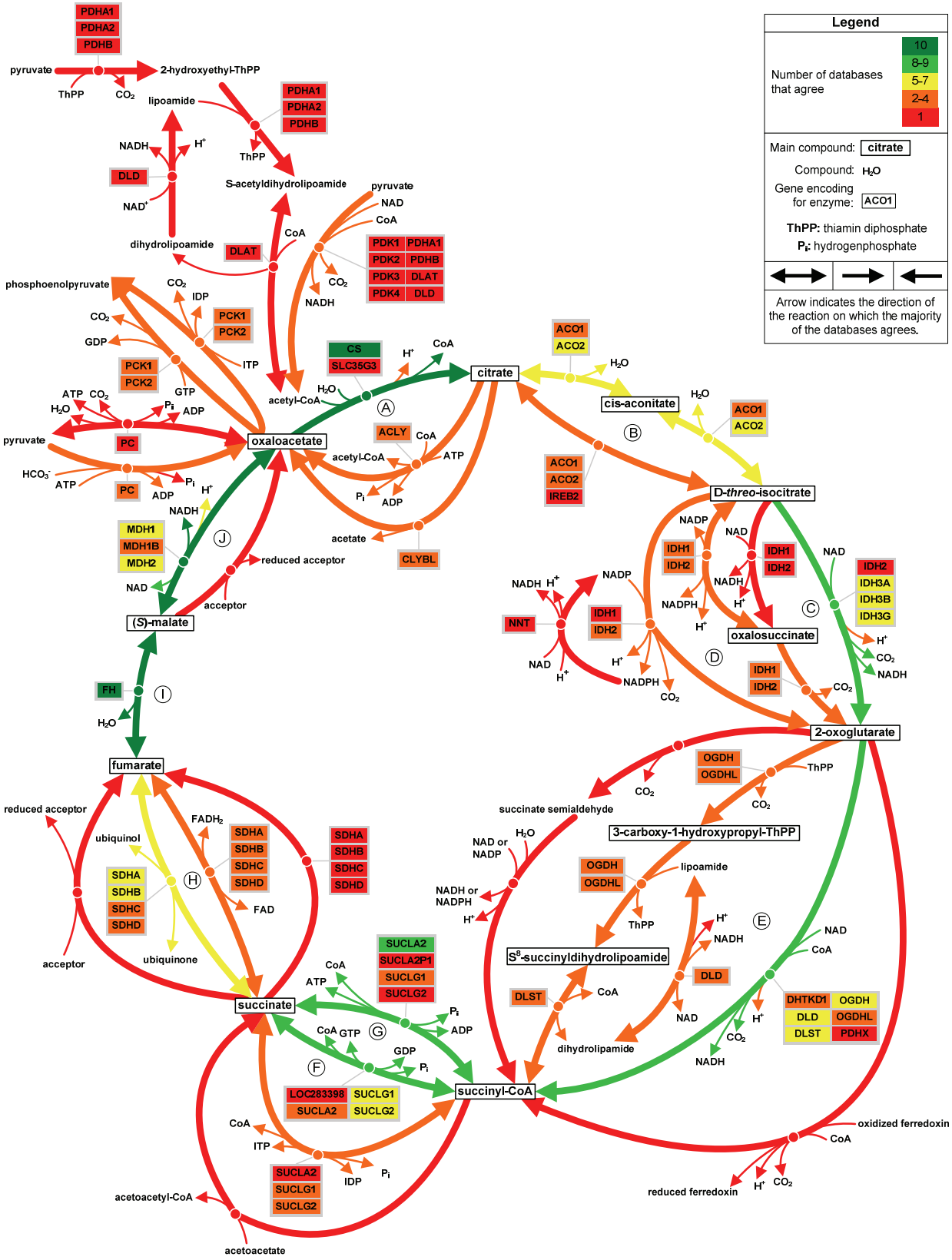


Figure 1 – Union of the descriptions of the TCA cycle given by ten pathway databases. Overview showing the reactions and genes annotated to play a role in the TCA cycle by the ten databases. The transport reactions found in the EHMN database (Hao *et al*, 2010) were excluded. The main metabolites of the TCA cycle are indicated by white rectangles. All genes linked to a reaction are combined in a box with grey borders. Colors indicate the level of agreement on a reaction and on the gene(s). The direction of an arrow is determined by what is indicated by the majority of the databases. In case of a tie irreversible was chosen. The letters (A-J) refer to the reactions described in Figure 2.

(Figure 2C) and the former NADP (Figure 2D). Most databases agree on this except two that incorrectly assign IDH2 to the NAD-dependent variant. More confusion exists over the direction of the reactions catalyzed by the IDH enzymes, *i.e.*, forward (oxidative decarboxylation) and/or backward (reductive carboxylation). IDH3 is allosterically regulated by positive (Ca^{2+}, ADP, citrate) and negative modifiers (NADH, NADPH, ATP) (Gabriel *et al*, 1986), which is consistent with its activity in the oxidative direction of the TCA cycle. Most databases agree on this direction. Much more controversy exists on the NADP-dependent reaction catalyzed by the IDH2 enzyme. Only half of the databases include the NADP-dependent reaction. Furthermore, only one database, *H. sapiens* Recon 1, mentions that this reaction is reversible, while the others claim that the NADP-dependent reaction operates in the oxidative direction of the TCA cycle. Biochemical evidence, however, indicates that the IDH2 enzyme operates in the (reverse) reductive direction, synthesizing D-*threo*-isocitrate. This is facilitated by the virtually fully reduced mitochondrial NADPH/NADP-redox state caused by the action of the nicotinamide nucleotide transhydrogenase, which is driven by the proton gradient across the mitochondrial membrane (Hoek and Rydström, 1988). Only under abnormal conditions, such as limiting substrate supply or hypoxia, which are characterized by a low proton electrochemical gradient, this reaction could theoretically proceed in the (forward) oxidative direction, but conclusive biochemical evidence is lacking. In contrast, it has been shown that in normoxic, but also hypoxic cell lines, the IDH2 enzyme is crucial for the reductive reaction to convert glutamine via 2-oxoglutarate into D-*threo*-isocitrate that is subsequently converted into citrate, which is exported to the cytosol where it is used for fatty acid synthesis (Mullen *et al*, 2012; Wise *et al*, 2011). With IDH2 and IDH3 operating in opposite directions, they form a substrate cycle that has been speculated to contribute to the fine regulation of the TCA cycle (Sazanov and Jackson, 1994). Indeed, multiple in vivo studies have established that this substrate cycle takes place in liver (Des Rosiers *et al*, 1994) and heart (Comte *et al*, 2002), but although biochemically plausible, it is not possible to conclusively attribute the reverse reaction to the IDH2 enzyme in the type of experiment that was done. In spite of this biochemical evidence, it was recently suggested that IDH2 serves as the main enzyme in the oxidative direction (Hartong *et al*, 2008). This was based on an observation in two patients with mutations in the β-subunit of the IDH3 complex presenting with retinitis pigmentosa and no other disease phenotypes that point to general TCA cycle dysfunction. Although interesting, this genetic finding does not prove that IDH2 enzyme functions in the oxidative direction. Further biochemical studies to address the role of IDH2 enzyme were not reported, nor studies addressing a more likely compensating role for the cytosolic IDH1 enzyme. In the

latter scenario, D-*threo*-isocitrate would be transported to the cytosol, converted to 2-oxoglutarate using NADP, which is readily available in the cytosol. Next, 2-oxoglutarate is transported back to the mitochondrion. For our description, we have decided to include only the reductive direction for IDH2 reaction because, as explained above, the oxidative direction does not take place under normal conditions. The role of IDH2 in the TCA cycle under pathophysiological conditions is unclear and should be further investigated. We therefore indicated the possibility that the IDH2 reaction is reversible as 'unconfirmed' in Figure 2D.

Databases also differ on including oxalosuccinate as an intermediate in the IDH reactions. In vitro studies have shown that the IDH2 enzyme can use oxalosuccinate as a substrate for reduction to D-*threo*-isocitrate as well as decarboxylation to 2-oxoglutarate. Although the catalytic mechanism that was inferred from the crystal structure of IDH2 enzyme shows that oxalosuccinate is an intermediate in the dehydrogenation of D-*threo*-isocitrate (Ceccarelli *et al*, 2002), there is strong evidence that oxalosuccinate is not a free intermediate (Siebert *et al*, 1957). Moreover, enzyme-bound oxalosuccinate was also not detected (Ramakrishna and Krishnaswamy, 1966). Most likely, the decarboxylation reaction proceeds very rapidly. Although oxalosuccinate is probably also a catalytic intermediate for the IDH3 reaction, the IDH3 enzyme complex does not accept oxalosuccinate as a substrate (Plaut and Sung, 1954). Since oxalosuccinate does not fulfill the two criteria we set (see section on aconitase), we have decided not to include oxalosuccinate as a TCA cycle intermediate in the reactions performed by IDH2 and IDH3.

2-Oxoglutarate dehydrogenase (E)

The 2-oxoglutarate dehydrogenase complex performs a series of complicated reactions including oxidative decarboxylation, formation of CoA ester and reoxidation of a lipoamide cofactor for which three different subunits are necessary, commonly referred to as the E1, E2 and E3 subunits. The overall reaction is irreversible, which is driven by the decrease in free energy and the removal of the CO_2 generated in the first, E1-catalyzed step (Sheu and Blass, 1999). Some databases describe this as a single reaction, while others use multiple steps. Representing it as a single reaction has the disadvantage that it will be more difficult to indicate, which reaction step is catalyzed by the different subunits. Indeed, a specific inherited defect has been described in only the E3 subunit (Liu *et al*, 1993). Theoretically, this would not affect the E1 and E2 enzymes that initiate the 2-oxoglutarate dehydrogenase reaction. However, in practice the entire complex operates as one functional unit with all intermediary products being enzyme–bound. Moreover, the complete cycle of steps has to be completed before a new reaction can start, which probably explains

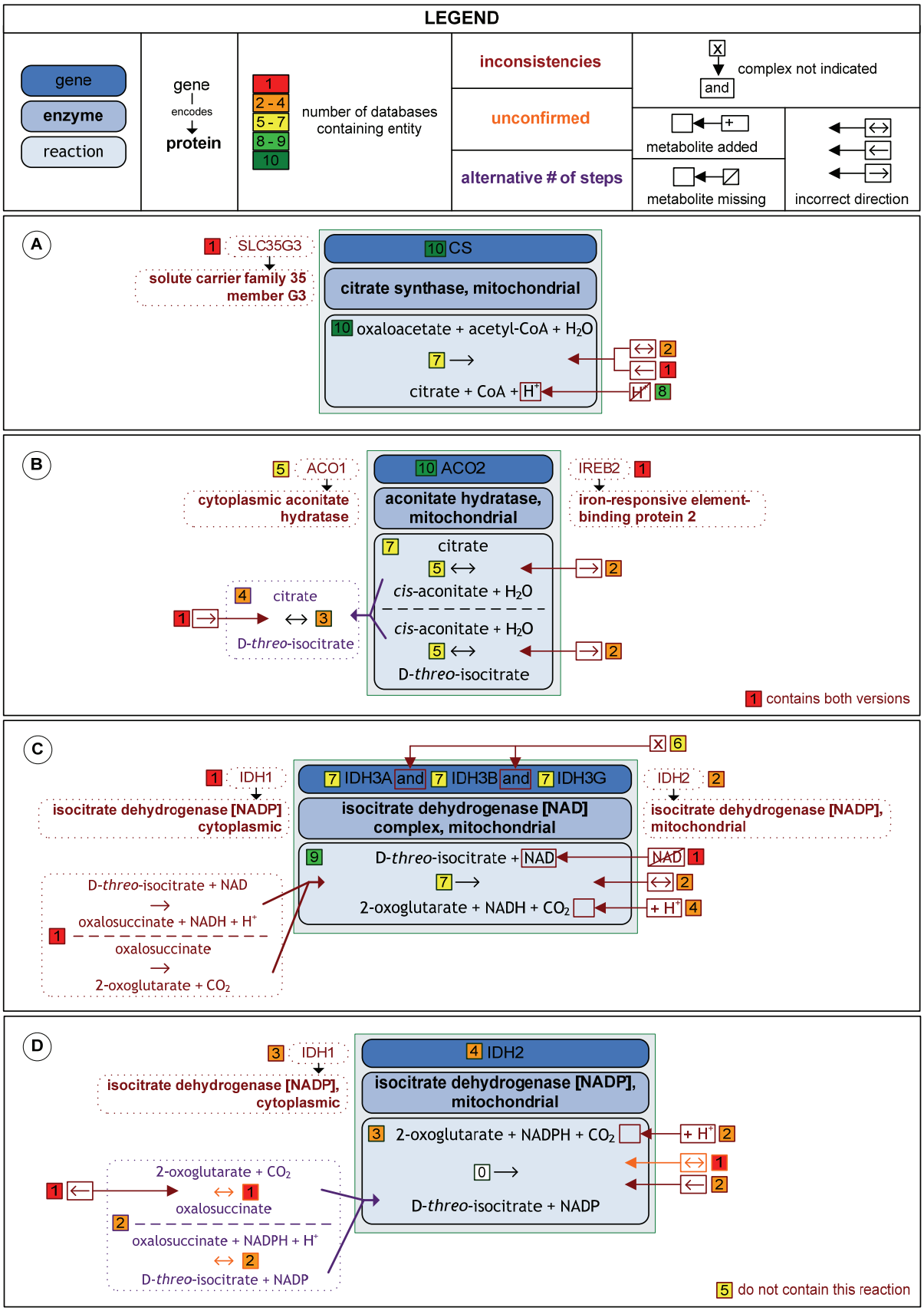
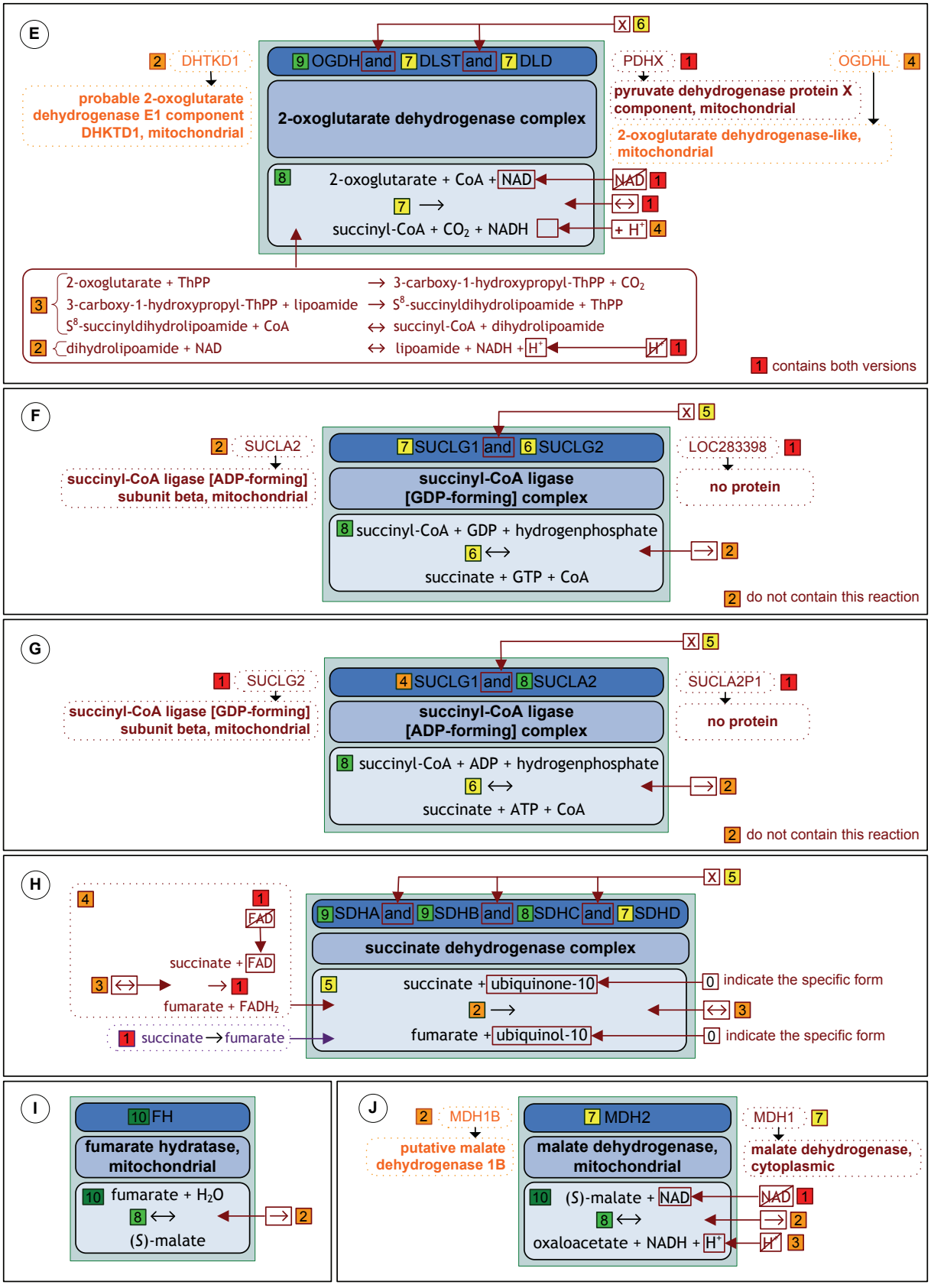


Figure 2 – Improved description of the TCA cycle. Reaction-wise overview of our description of the TCA cycle based on literature and the ten databases evaluated. Blue colored boxes contain the correct gene(s), enzyme(s), and reaction. The numbers in square boxes indicate how many of the ten databases are in agreement with this description and their color indicates the level of agreement. (*continued on next page*)



Inconsistencies with the literature as found in any of the ten databases are shown in dark red outside the blue boxes. Genes and their products for which a role in the TCA cycle could neither be confirmed nor refuted by evidence from the literature are indicated in orange font. The same holds for the direction of reactions. Reactions in purple font differ in the number of steps the conversion is described in compared to our description.

why E3-deficient patients accumulate only the substrate 2-oxoglutarate. In line with the criteria mentioned above, we therefore describe the reaction in a single step.

The majority of the databases attribute the OGDH, DLST and DLD proteins to this reaction, although four databases do not indicate that the proteins act as a complex. Some databases (also) assign the DHTKD1, OGDHL and PDHX proteins. The first two are similar to the E1 subunit (OGDH) of the 2-oxoglutarate dehydrogenase complex. Both proteins have recently been further characterized using sequence comparison. OGDHL most likely represents a previously unknown isoform of the OGDH protein. Thus, OGDHL might play a role in the TCA cycle, but its expression levels are much lower than that of OGDH. More biochemical evidence is needed to elucidate the role of OGDHL. DHTKD1 may accommodate more polar and/or bulkier structural analogs of the 2-oxoglutarate metabolite (Bunik and Degtyarev, 2008). The DHTKD1 gene most likely encodes a dehydrogenase with a new function. For the PDHX protein there is no biochemical evidence supporting that it is necessary in the formation of the 2-oxoglutarate dehydrogenase complex (McCartney *et al*, 1998). Moreover, patients with PDHX deficiency have a selective pyruvate dehydrogenase deficiency (Aral *et al*, 1997).

Succinyl-CoA synthetase (F and G)

In this reversible reaction, an energy-rich CoA ester bond is cleaved, which was classically thought to be coupled to the formation of GTP (Figure 2F). In 1998, it was, however, established that there is a second succinyl-CoA synthetase which produces ATP rather than GTP (Figure 2G) (Johnson *et al*, 1998). Some databases also give a third purine nucleotide as a product, ITP (Figure 1). Although in vitro IDP is a substrate for this enzyme, it is very unlikely to play a role in vivo. The concentrations of IDP and ITP are very low as compared to the other nucleotides and considered a byproduct of purine nucleotide metabolism (Bierau *et al*, 2007).

The two succinyl-CoA synthetase enzymes are dimers with one common alpha subunit (SUCLG1) and a beta subunit that confers nucleotide specificity: SUCLG2 for the GTP-specific isozyme and SUCLA2 for the ATP-specific isozyme. Although widely expressed, the relative amounts of these two subunits vary from tissue to tissue. SUCLA2 is highly expressed in testis, brain, heart, and kidney (Lambeth *et al*, 2004). SUCLG2 is expressed in liver, kidney, and heart, but barely detected in brain and testis (Lambeth *et al*, 2004). These two complexes are often incorrectly represented in the different databases. Three types of mistakes are made: (i) not all components are indicated, (ii) all three proteins mentioned are assigned to both reactions, (iii) it is not described that the proteins form a complex.

The correct representation of these complexes is important to be able to understand the effect of deficiencies in the SUCLA2 and/or SUCLG1 proteins, as the biochemical consequences differ. In a SUCLG1 deficiency, both the GTP- and ATP-specific isozymes are affected, whereas in a SUCLA2 deficiency only the ATP-specific isozyme is deficient (Ostergaard *et al*, 2007). Consequently, in the former both reactions would be affected, while in the latter only the ADP/ATP-dependent reaction is influenced.

Succinate dehydrogenase (H)

The succinate dehydrogenase enzyme oxidizes succinate into fumarate and is also known as complex 2 of the respiratory chain. The enzyme is membrane associated and contains four different subunits, which are not all included by each database and also not always indicated as forming a complex. Most biochemistry textbooks teach that electrons are transferred to FAD giving $FADH_2$, which may explain the choice made by four of the databases. Succinate dehydrogenase, however, is a covalent flavoprotein, therefore the FAD is an enzyme-bound prosthetic group that can not dissociate from the enzyme (Mewies *et al*, 1998). In fact, the electrons are contained in enzyme-bound $FADH_2$ and further transferred into the electron transfer chain via ubiquinone-10 forming ubiquinol-10. Since it is ubiquinol-10 that dissociates from succinate dehydrogenase, we decided to describe these cosubstrates in the reaction.

Fumarate hydratase (I)

Fumarate hydratase catalyzes the reversible hydration of fumarate into (*S*)-malate. The highest level of agreement between databases is on this reaction. The only disagreement concerns the reversibility of this reaction. Although undoubtedly reversible, two databases give this reaction as unidirectional. In these two databases, however, all reactions are unidirectional, while in the other eight databases the information on the reversibility of a reaction is provided.

Malate dehydrogenase (J)

Malate dehydrogenase completes the TCA cycle by converting (*S*)-malate into oxaloacetate. In the liver, malate dehydrogenase is shared between the TCA cycle and gluconeogenesis illustrating that this reaction is reversible (Des Rosiers *et al*, 1995; Fernandez and Des Rosiers, 1995). There are two malate dehydrogenase enzymes, the cytosolic MDH1 and the mitochondrial MDH2. Some of the databases associate the MDH1 enzyme with the TCA cycle, whereas only the MDH2 enzyme can perform this role because of its mitochondrial localization (see below). Two databases also associate the MDH1B enzyme with this reaction. There is, however, no supporting evidence for this. Furthermore, the protein is most likely not localized to mitochondria.

Annotation of gene function

Our analysis brought to light several genes, found in one or more pathway databases, which are suggested to be involved in the TCA cycle, but for which no evidence in literature exists. The availability of the complete human genome allowed for the identification of new genes, without any functional characterization. Based on homology the products of three genes, *i.e.*, MDH1B, DHTKD1 and OGDHL, were annotated to play a role in the TCA cycle in several databases, but without formal biochemical proof. We therefore indicated these enzymes as 'unconfirmed' (Figure 2 and Table 2). Furthermore, in one of the databases, HumanCyc (Romero *et al*, 2004), the protein encoded by SLC35G3 is associated with the citrate synthase reaction, but there is no evidence at all for such a role. In another database, Reactome (Croft *et al*, 2011), pseudogenes (SUCLA2P1 en LOC283398) are linked to the succinyl-CoA synthetase reactions. For the products of ACO1, IREB2, IDH1 and MDH1, there is currently no evidence that they can be localized in the mitochondrion where the TCA cycle takes place. They all catalyze a reaction also found in the TCA cycle, but the proteins are localized in the cytosol (and also peroxisome for the IDH1 protein). Although some of these databases mention that these reactions are compartmentalized to the cytosol, their annotated role in the TCA cycle is incorrect.

Links of the TCA cycle with other pathways

The TCA cycle is a hub in cellular metabolism connecting many different pathways. There are many associated transporters, and anaplerotic and cataplerotic reactions, supplying or removing the main metabolites of the TCA cycle. In defining the boundaries of this pathway for our description, we focused on the biochemical cycle itself, with no real starting substrate or end product (Berg *et al*, 2012, pp. 515-542). Some databases include reactions that transport TCA cycle intermediates and selected reactions associated with TCA cycle intermediates such as the pyruvate carboxylase, phosphoenolpyruvate carboxykinase and pyruvate dehydrogenase reaction. Pyruvate carboxylase is an example of an important anaplerotic reaction, which supplies the TCA cycle with oxaloacetate, and is therefore by definition not part of the cycle itself. The same holds for the cataplerotic phosphoenolpyruvate carboxykinase reaction. Finally, the pyruvate dehydrogenase reaction is often included because it generates the acetyl-CoA that is used for the synthesis of citrate in the first step of the TCA cycle, but it is not part of the cycle itself. Moreover, acetyl-CoA can also be produced from fatty acids and amino acids. Therefore we also left out the pyruvate dehydrogenase reaction including all its associated regulating kinases and phosphatases.

Dissemination

One database, WikiPathways (Pico *et al*, 2008), was excluded from the analyses described above as it only provided the main metabolites of each reaction. Instead we made use of an important characteristic of this database, namely, that WikiPathways enables researchers to adapt the description of a pathway. Based on the results of our comparison and literature study, we refined the description of the TCA cycle in WikiPathways (http://www.wikipathways.org/index.php?title=Pathway:WP78&oldid=47741) and also added literature references. Our corrections can be viewed in detail by looking at the differences with an earlier description in WikiPathways using the Pathway Difference Viewer (Pico *et al*, 2008) (http://wikipathways.org/index.php/Help:Viewing_Pathways). Our improved description of the TCA cycle can be downloaded in various formats to allow for a broad dissemination of our results. The original description of WikiPathways is also shown in the Wikipedia entry on this pathway and therefore we requested an update.

Given the extensive amount of literature on the TCA cycle spanning decades of research, fully capturing current knowledge on this biochemically complex process remains a huge challenge. Consequently, we cannot exclude that our description still contains some inconsistencies. Moreover, as shown above parts of the TCA cycle are subject of ongoing research, which might lead to new insights. In line with the philosophy of WikiPathways, we therefore encourage others to refine our description.

Discussion

Metabolic pathway databases have proven highly valuable in a broad range of applications ranging from the analysis and visualization of high-throughput data (Antonov *et al*, 2008) to *in silico* predictions of phenotypes (Jerby *et al*, 2010). At the same time it is important to be aware, however, of possible limitations inherent to pathway databases. Based on the evaluation of ten descriptions of the TCA cycle we conclude that none of the selected pathway databases accurately represent the knowledge available in the literature on this pathway. In the UniPathway database (Morgat *et al*, 2012), for example, 13 enzymes are missing, while in KEGG five enzymes are incorrectly linked to one of the reactions of the TCA cycle (Table 2). Furthermore, we also observe a difference in how the boundaries of the TCA cycle are defined. For example, following our definition, 16 reactions of EHMN and 8 of KEGG (Table 3) would not be part of this pathway.

Our detailed analysis confirms our hypothesis that part of the lack of consensus can be explained by these inconsistencies and a partial coverage of the literature. Also in the description in 'Biochemistry' (Berg *et al*, 2012), one of the most popular student textbooks, we identified similar inconsistencies. Two reactions are incorrectly indicated to be reversible, *i.e.*, the NAD-dependent IDH reaction and the 2-oxoglutarate dehydrogenase reaction. The GDP-dependent succinyl-CoA synthetase reaction is described as acting only in the opposite direction of the TCA cycle and not as reversible. Furthermore, the NADP-dependent IDH reaction is left out. Given that the TCA cycle is one of the most well-known pathways our results are surprising. However, our review also showed that the biochemistry of this pathway may not be as clear cut as one would expect, which explains some of the inconsistencies discovered. This is underscored by recent literature showing that this particular pathway is still actively studied. One example is the controversy surrounding the direction of the IDH reaction catalyzed by the IDH2 enzyme, despite the traditional biochemical evidence that is available. Lack of consensus may also be partly explained by a different judgment between curators on the strength of the evidence from the literature. For example, part of our evidence has been obtained in model organisms such as rat and mouse. Since core metabolic pathways such as the TCA cycle are broadly conserved across organisms, we considered evidence from mammals conclusive. Human-specific evidence is often lacking, but once it becomes available it can be added to the description of the TCA cycle on WikiPathways.

Our results are likely to extend to other pathways as well. For five of the eleven databases, we have previously shown that the lack of consensus translates to the entire human metabolic network (Stobbe *et al*, 2011). Moreover, in that analysis on network level various mistakes were found in other pathways as well. Overviews of all differences between these five databases can be retrieved via the web application called Consensus and Conflict Cards (C_2Cards) (http://www.c2cards.nl). A C_2Card provides a concise overview of what the databases do and do not agree on with respect to a single reaction, gene or EC number. This enables experts to more easily identify mistakes and to reconcile the descriptions for other pathways than the TCA cycle.

We expect that the issues encountered in our curation effort extend to other organisms. In fact, various analyses have already shown that also for other organisms there is a lack of consensus between the multiple descriptions of the metabolic network that are available (Chindelevitch *et al*, 2012; Herrgård *et al*, 2008; Radrich *et al*, 2010; Thiele *et al*, 2011). A complete and accurate description of the metabolic network for human and other organisms is essential to foster new

biological discoveries. For example, the holistic view that a model of the metabolic network offers, allows for the identification of gaps in our knowledge on human metabolism for which further experiments are required (Rolfsson *et al*, 2011). Furthermore, pathway databases are used more and more often as the primary knowledge resources on metabolism by biologists. Reliance on databases will continue to increase since the complexity of the high-throughput datasets that are handled increases as well. Mistakes in pathway databases are also propagated to other types of resources such as the Gene Ontology (GO) (Ashburner *et al*, 2000), STRING (Szklarczyk *et al*, 2011) and Wikipedia. According to GO, for example, the cytosolic proteins that are encoded by ACO1, IDH1 and MDH1 play a role in the TCA cycle. Pathways contained in WikiPathways are featured by Wikipedia as interactive pathway maps and, therefore, contain the same inconsistencies. For the TCA cycle we requested an update of the current Wikipedia entry to our description in Wikipathways. Incorrect information may lead to misinterpretation of high-throughput molecular data and the design of poorly designed follow-up experiments. Furthermore, a consequence of the many differences between the descriptions is that one could arrive at different conclusions for an analysis depending on which database one uses (Lee *et al*, 2008; Zelezniak *et al*, 2010).

Various initiatives are in place that aim to improve upon the current state of affairs and for which the support of a broad community is essential. One example of an initiative to improve upon an already existing description of the human metabolic network is the Reactome Portal (http://wikipathways.org/index.php/Portal:Reactome). Pathways from the Reactome database have been incorporated into WikiPathways and thus can be edited by everyone, hereby following the same principle as Wikipedia. Periodically, curators of Reactome evaluate the changes made and decide whether or not to include it in their centralized database, which cannot be edited by the public. An important difference with Wikipedia and a bottleneck for initiatives like the Reactome Portal is that the pool of experts knowledgeable enough to contribute is much smaller. Therefore, a larger percentage of the community needs to contribute to reach the necessary momentum to improve upon the current descriptions of the (human) metabolic network.

A second example of an endeavor to refine and reconcile current descriptions of the (human) metabolic network is the organization of reconstruction annotation jamborees (Herrgård *et al*, 2008; Thiele and Palsson, 2010a; Thiele *et al*, 2011). In a jamboree, experts from multiple disciplines, including biochemistry, molecular description of the metabolic network. Our improved description of the TCA cycle

provides a nice example of the reconciliation of ten individual descriptions. The knowledge on the metabolic network will continue to expand and therefore the consensus network needs to be kept up-to-date, for example, by organizing subsequent jamborees. This requires the continuing commitment of experts.

Other initiatives focus on the accurate extraction of knowledge from the scientific literature. One of the explanations for the lack of consensus between the descriptions of the metabolic network is that pathway database curators have based their conclusions on a different set of articles and/or interpreted the literature differently (Mo and Palsson, 2009). Importantly, curators may not have interpreted the article as intended by the authors. The original authors of a novel scientific fact are generally not the ones that put their conclusions into a pathway database. For the TCA cycle, our literature study led to the reappraisal of the knowledge that the NAD/NADH and NADP/NADPH redox couples (Houtkooper *et al*, 2010) are widely different. It is, however, quite a challenge to oversee the huge volume of articles already available, which is further complicated by the changing nomenclature of enzymes and metabolites. Moreover, as the biomedical literature grows exponentially, it is impossible for curators of pathway databases to keep track of everything published on metabolism. To cope with this, various innovative ideas have been proposed. For example, authors could add semantic annotation to their article (Jensen and Bork, 2010), making the knowledge described more machine-readable. This would allow for an easier way to automatically retrieve and put new knowledge into a pathway database. Another approach is to provide the newly discovered facts also as nanopublications (Groth *et al*, 2010). A nanopublication is a traceable author statement, which consists of three parts: a *statement*, *e.g.*, protein X (subject) catalyzes (predicate) reaction Y (object), *conditions* under which the statement holds, *e.g.*, a specific compartment, and *provenance* of the statement, *e.g.*, author and literature. By staying close to the source of the newly discovered fact on metabolism, misinterpretation of the article can be prevented. Moreover, an additional advantage is that it will significantly reduce the workload for curators of metabolic pathway databases when retrieving new knowledge. It remains a challenge, however, to convince experts to spend time and effort into improving the description of a metabolic network as they may feel that they do not directly benefit from this. Therefore, the efforts required should be kept as minimal as possible and the contributions of the expert should be clearly acknowledged. In the C_2Cards application, for example, curation is done at the level of a single reaction or the metabolic functions of a single gene product. Nanopublications provide an explicit recognition of the knowledge contributed by the expert(s). Journals could also play a

role by, for example, requiring authors to contribute their results to one of the discussed initiatives.

The active involvement of a broad community, across multiple disciplines within the field of biology, is essential to further improve the current description of the metabolic network of human and other organisms (Kitano *et al*, 2011). Via this article, we would, therefore, like to urge biologists to donate their knowledge and actively contribute to reach the ultimate goal of a complete and biologically accurate description of the (human) metabolic network.

Acknowledgements

This research was carried out within the BioRange programme (project SP1.2.4) of The Netherlands Bioinformatics Centre (NBIC; http://www.nbic.nl), supported by a BSIK grant through The Netherlands Genomics Initiative (NGI) and within the research programme of the Netherlands Consortium for Systems Biology (NCSB), which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. Sander M. Houten was supported by the Netherlands Organization for Scientific Research (VIDI-grant No. 016.086.336).