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

Stobbe, M.D.

**Publication date**
2012

# Chapter 3

## Knowledge representation in metabolic pathway databases

Miranda D. Stobbe[1,3], Gerbert A. Jansen[1,3], Perry D. Moerland[1,3], Antoine H.C. van Kampen[1,2,3,4]



[1] Bioinformatics Laboratory, Academic Medical Center, University of Amsterdam, PO 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] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, the Netherlands
[4] Netherlands Consortium for Systems Biology, University of Amsterdam, PO Box 94215, 1090 GE, Amsterdam, the Netherlands

Accepted for publication in: *Briefings in Bioinformatics*

# Abstract

The accurate representation of all aspects of a metabolic network in a structured format, such that it can be used for a wide variety of computational analyses, is a challenge faced by a growing number of researchers. Analysis of five major metabolic pathway databases reveals that each database has made widely different choices to address this challenge, including how to deal with knowledge that is uncertain or missing. In concise overviews we show how concepts such as compartments, enzymatic complexes, and the direction of reactions are represented in each database. Importantly, also concepts which a database does not represent are described. Which aspects of the metabolic network need to be available in a structured format and to what detail differs per application. For example, for *in silico* phenotype prediction a detailed representation of gene-protein-reaction relations and the compartmentalization of the network is essential. Our analysis also shows that current databases are still limited in capturing all details of the biology of the metabolic network, further illustrated with a detailed analysis of three metabolic processes. Finally, we conclude that the conceptual differences between the databases, which make knowledge exchange and integration a challenge, have not been resolved, so far, by the exchange formats in which knowledge representation is standardized.

## Introduction

Our understanding of metabolism is ever expanding, as evidenced by the increasing amount of bibliomic data. Pathway databases have been built to collect and capture this knowledge. Besides serving as knowledge repositories, the databases aim to represent the metabolic network in a digital format in such a way that it can be used for computational analyses. This has enabled numerous analyses ranging from the prediction of phenotypes (Jerby *et al*, 2010), studying evolution (Tanaka *et al*, 2006), to the analysis and interpretation of high-throughput data (Antonov *et al*, 2008). The number of pathway databases describing the metabolic network for one or more organisms continues to grow (Karp and Caspi, 2011; Oberhardt *et al*, 2009).

From the perspective of a researcher used to the compact representation of biological knowledge on metabolism in a pathway, it may seem trivial to represent the metabolic network in an electronic form. However, the biology of the metabolic network is complex and the terminology used by biologists changes over time and varies among biologists (Karp and Mavrovouniotis, 1994). Furthermore, a pathway database needs to accommodate a wide range of users with different requirements. Numerous choices need to be made by database developers and curators on how to represent and relate each component of the metabolic network (Figure 1). Important considerations hereby are what needs to be described in a structured and standardized form to enable computational analyses and what can be described as background information in unstructured text fields.

We selected five, frequently used, pathway databases and compared their approach to represent the human metabolic network in a digital format. Furthermore, we discuss how the databases deal with knowledge that is uncertain or missing. To illustrate the challenges faced in making biological knowledge amenable to computational analyses, we give a detailed description of how each database represents three complex metabolic processes: fatty acid beta oxidation, oxidative phosphorylation, and the pyruvate dehydrogenase reaction. Finally, we discuss the challenges posed by the differences in knowledge representation for analyses across pathway databases and exchange of knowledge between databases.

With this paper we intend to increase the awareness on the complexity of representing the current knowledge on metabolism. A detailed understanding of how knowledge is represented is crucial for users of pathway databases, as differences in representation can affect the outcome of computational analyses. As argued by Green and Karp (Green and Karp, 2006), the pathway definition alone

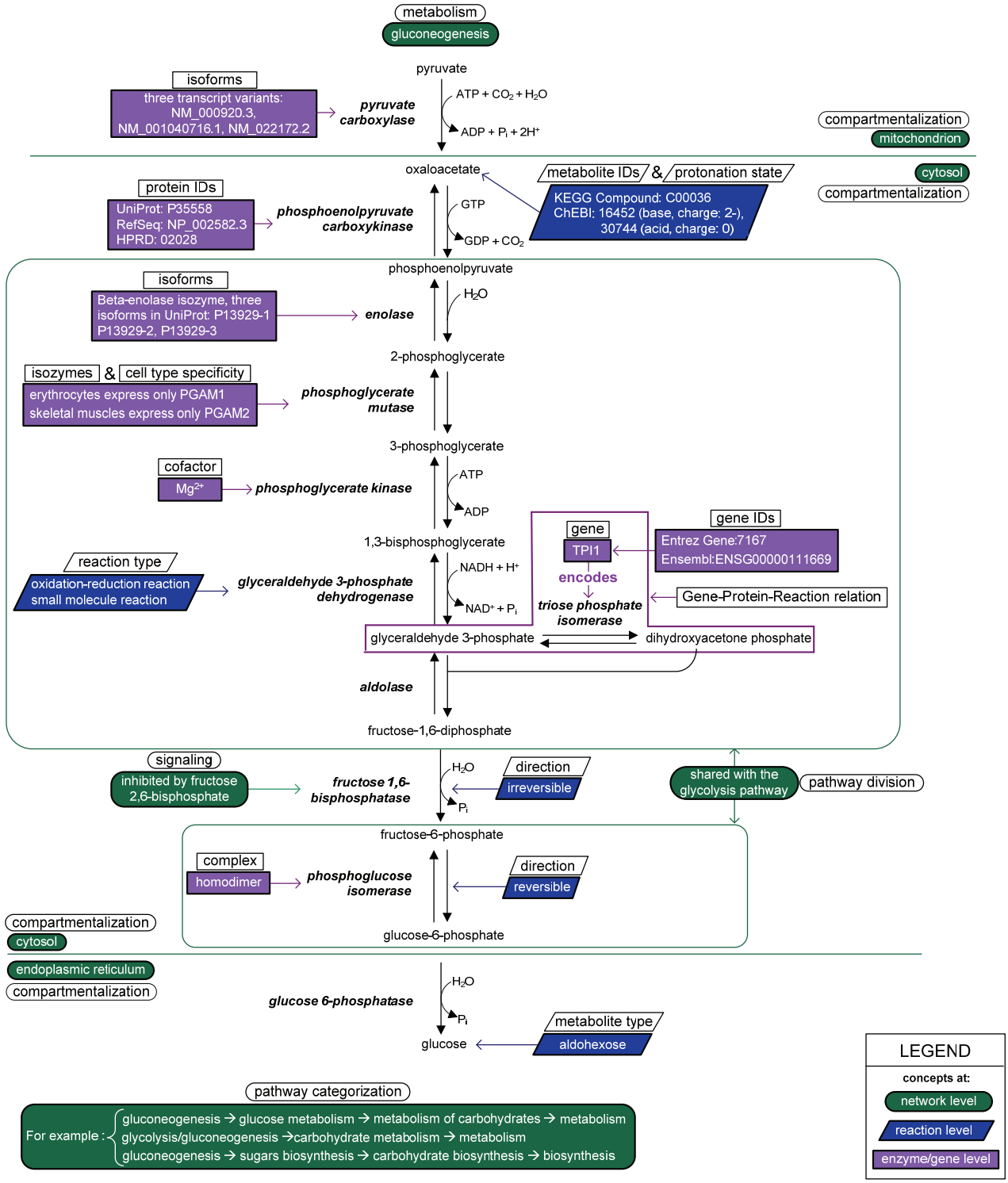



**Figure 1 - Gluconeogenesis.** Selection of the information that needs to be stored to accurately represent the gluconeogenesis pathway (Berg *et al*, 2012, pp. 469-513).

may already influence analysis results. Moreover, the choices made in how to represent the metabolic network affect the ability of the database to capture every detail. By pointing out the current limitations our research will also aid (future) database developers, knowledge curators, and domain experts in their quest to further improve knowledge representation.

| Database | # of organisms[1] | Name human network | Version | File formats and computational access |
|---|---|---|---|---|
| BiGG | 8 | *H. sapiens* Recon 1 | 1 | Tab-delimited files, SBML |
| BioCyc | 1700 | HumanCyc | 15.5 | Pathway Tools (export flat files), API, BioPAX, SBML |
| EHMN | 1 | EHMN | 2 | Excel file, SBML |
| KEGG | 1646 | --- | 61 | KGML, API, dbget, flat files |
| Reactome | 46 | --- | 39 | MySQL dump, API, BioPAX, SBML |

**Table 1 – Metabolic pathway databases compared.** [1]Note that numbers should not be directly compared, since the level of curation may differ per organism, both within and between databases.

# Results

To illustrate the differences in representation of the metabolic network, we selected the following five databases: H. sapiens Recon 1 (Duarte *et al*, 2007) from BiGG (Schellenberger *et al*, 2010) (referred to as Recon 1 in the rest of the paper), HumanCyc (Romero *et al*, 2004) from BioCyc (Karp *et al*, 2005), EHMN (Hao *et al*, 2010), KEGG (Kanehisa *et al*, 2012), and Reactome (Croft *et al*, 2011) (Table 1). Our analysis of how knowledge is represented in these databases was based on the descriptions given by the pathway database curators themselves in articles and online manuals (if available); when necessary we contacted the database curators for additional details. Moreover, for a more detailed insight we also examined the data files provided by each database, which contain the actual representation of the metabolic network (Supplementary Table S1). Note that we did not consider concepts that were only represented on the website of the database or knowledge that was only provided indirectly by references to other databases, *e.g.*, metabolite databases. For each database we analyzed how and to what detail knowledge is represented on the level of (i) the entire network, (ii) its reactions, and (iii) the enzymes and their encoding genes. On all three levels the databases have made different choices (Table 2). We focused in our review on how knowledge is represented rather than how it is collected, although both issues are intimately linked to how accurately the metabolic network is captured.

## *Representation of the network*

### *Types of networks*

An important difference on network level is whether a database only describes metabolic processes (EHMN, Recon 1) or also other types of biological processes such as signaling and genetic information processing (HumanCyc, KEGG, Reactome; Table 2 and 3). Since in Reactome the different types of processes are intertwined it is non-trivial to only retrieve the metabolic network, needed for instance to carry out flux balance analyses (Latendresse *et al*, 2012; Orth *et al*, 2010). Note that all five

| | | | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|---|---|
| network | type | metabolism | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | signaling | ✗ | ✗ | – | ✓ | ✓ |
| | | genetic information processing | ✗ | ✗ | – | ✓ | ✓ |
| | compartmentalization | | ✓ | ✓ | – | ✗ | ✓ |
| | division into pathways | | ✓ | ✓ | ✓ | ✓ | ✓ |
| reaction | reaction type | | ✗ | ✗ | ✓ | ✓ | ✗ |
| | linking of reactions | | ✗ | ✗ | ✓ | ✓ | ✓ |
| | physiological direction | | ✗ | ✓ | ✓ | ✓ | ✓ |
| | metabolite | type | ✗ | ✗ | ✓ | ✓ | ✗ |
| | | protonation state | ✗ | ✓ | ✓ | ✗ | ✗ |
| enzyme | isozymes | | ✗ | ✓ | ✓ | ✗ | ✓ |
| | isoforms | | ✗ | ✓ | ✗ | ✗ | ✓ |
| | complexes | heteromers | ✗ | ✓ | ✓ | – | ✓ |
| | | homomers | ✗ | ✗ | ✓ | ✗ | ✓ |
| | prosthetic groups/cofactors | | ✗ | ✗ | – | ✓ | ✓ |

**Table 2 – Representation of concepts in metabolic pathway databases.** Check mark: concept is represented. Cross: concept is not represented. Bar: database is able to represent the concept, but this is only done to a limited extent.

databases describe what is referred to as the global human metabolic network in which all possible reactions are combined, despite that they may not take place in every tissue or cell type. Defining tissue-specific models is left to algorithms like the one designed by Jerby *et al.* (2010) or more specialized (manually) curated networks like HepatoNet1 (Gille *et al*, 2010).

### *Compartmentalization*

Different cellular compartments have distinct metabolic functions. KEGG does not provide any information on compartments. In HumanCyc compartmentalization is work in progress, but their cell component ontology (Zhang *et al*, 2005) does allow for a detailed representation. EHMN, Recon 1, and Reactome provide a fully compartmentalized network. Reactome has the most fine-grained compartmentalization of the latter three databases and thereby conveys the most detailed knowledge (Table 3). Moreover, Reactome not only indicates the compartment for each reaction, but also for each enzyme and even for some pathways. Recon 1 and EHMN account for the same set of eight compartments, but handle subcellular locations not included in this set differently. In Recon 1 the intermembrane space of the mitochondrion, for example, is merged with the cytosol (Duarte *et al*, 2007). EHMN uses the hierarchy of the Gene Ontology (GO) to determine which of their eight compartments is the ancestor of the subcellular

| | | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|---|
| type | multiple types of processes[a]? | no, only metabolic reactions | no, only metabolic reactions | yes, all types of reactions represented in the same way | yes, different types of reactions represented differently | yes, all types of reactions represented in the same way and intertwined |
| compartmentalization | # compartments[b] | 8 | 8 | 27[c] | not applicable | 47 |
| | ontology | GO cellular component | no ontology used | Cell Component Ontology | not applicable | GO cellular component |
| | specified on the level of the | metabolites and/or reaction[d] | metabolites or reaction[e] | metabolites, reaction and enzyme | not applicable | metabolites, reaction, enzyme and pathway |
| | if compartment is unknown | 'uncertain', but for practical purposes also provided as 'cytosol' | cytosol | 'NIL' (few cases) or nothing is specified in which case 'cytosol' is the default | not applicable | not applicable (reaction not included in database if compartment is unknown) |
| pathway | definition | re-division of the pathways of KEGG and EMP:<br>- less overlap between pathways<br>- small functionally related pathways are grouped<br>- human-specific | definition of KEGG, but human-specific | guidelines used:<br>- a single biological process<br>- evolutionary conserved<br>- regulated as a unit<br>- boundaries at stable and high-connectivity metabolites | centered on the synthesis and/or degradation of one or more related substrates | a series of reactions, connected by their participants, leading to a biological outcome |
| | categorization | categories of KEGG | categories of KEGG | hierarchy, first classified according to type of process (*e.g.* degradation) and next on the type of metabolites | 11 categories (*e.g.* amino acid metabolism), based on the type of metabolites involved | hierarchy, based on the type of metabolites involved |
| | reactions without pathway? | yes, labeled 'isolated' | yes, labeled 'other' or 'miscellaneous' | yes, no link to instance of the pathway frame | yes | no |

**Table 3 – Representation differences on network level.** [a] Metabolic, signaling and gene information processing. [b] Includes extracellular space. [c] Includes generic compartments like 'in' and 'membrane' and the non-human compartment 'inner membrane (sensu Gram-negative Bacteria)'. [d] If it is a transport reaction, the compartment is only indicated at the metabolite level. [e] If it is a transport reaction, the compartment is indicated at the metabolite level, otherwise only for the entire reaction. GO: Gene Ontology, EMP: Enzymes and Metabolic Pathways database.

| | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|
| number of pathways | 69 | 96 | 257 | 84 | 171 |
| average number of reactions per pathway | 52 | 30 | 4 | 23 | 8 |
| % of reactions occurring in >1 pathway[a] | 1% | 9% | 13% | 10% | 12% |

**Table 4 – Pathway statistics.** Statistics based upon the same data we used previously (Stobbe *et al*, 2011) for a comparison of the content of the five databases. Only the metabolic pathways of HumanCyc, KEGG and Reactome are considered. For Reactome and HumanCyc the lowest level in the hierarchy was used when counting the number of pathways. If reactions only differ in direction and/or compartments they are counted as one. [a] With respect to the total number of reactions assigned to at least one pathway.

location in question. In the example above the ancestor is the mitochondrion. Such a more coarse-grained compartmentalization will result in a less accurate representation of, for example, oxidative phosphorylation (Supplementary Text S1).

### *Division into pathways*

All five databases divided their network into pathways to provide insight into the functional organization of the metabolic network. Although this division into pathways is not arbitrary and is based on biological criteria in each of the databases, there is no generally accepted definition of a pathway. Consequently, each database defines the boundaries of its pathways differently (Table 3). This results in a large difference in the number of pathways, the average number of reactions per pathway, and the overlap between pathways (Table 4). As described by Green and Karp (Green and Karp, 2006) for BioCyc and KEGG, the pathway definition might influence the outcome of pathway-based analyses.

Reactome defines a pathway as a series of reactions, connected by their participants, leading to a biological outcome (Reactome Glossary, 2012). In HumanCyc more strict guidelines are used, *i.e.*, a pathway is a single biological process that should be evolutionary conserved and regulated as a unit (Green and Karp, 2006). This partly explains the low average size of a pathway in HumanCyc. Moreover, variants of the same metabolic process are considered as separate pathways, thereby increasing the overlap between pathways. In contrast, in EHMN the emphasis is on the functional relationships between reactions and overlapping metabolic processes are merged into a single pathway (Ma *et al*, 2007). The pathways of KEGG are a mosaic of reactions that take place in any of the organisms included in KEGG and are substrate-centric (Green and Karp, 2006). The organism-specific version of a KEGG pathway consists of those reactions to which a gene of the organism of interest has been linked. This approach can result in artifacts. For example, 'lysine biosynthesis' is part of the human metabolic network in KEGG, although our metabolism lacks the ability to synthesize lysine. Recon 1 uses the same pathway definitions and

categories as KEGG, however only human-specific pathways are included and, *e.g.*, lysine biosynthesis is not included in Recon 1. Ultimately, pathways cannot be studied in isolation as the entire network is connected.

### *Representation of metabolic reactions*

A metabolic reaction can be defined as the synthesis or degradation of chemical compounds, which may or may not be a reversible process. The type of reaction, *e.g.*, an 'oxidation-reduction' reaction, is indicated by an Enzyme Commission (EC) number in all databases, although in Reactome a link to GO is preferred. KEGG and HumanCyc also have their own reaction ontology (Table 5). The ontology of HumanCyc enables selecting, for instance, only small molecule reactions. The exact representation of a metabolic reaction differs per database. For example, each database uses a different terminology in its data model to refer to the metabolites before the arrow, *e.g.*, 'substrates' or 'input', and after the arrow, *e.g.* 'products' or 'output' (Figure 2). In addition, the level of detail in which a conversion is described varies within and between databases (Case Study and Supplementary Text S2). Differences in detail between databases can reflect a disagreement on the number of steps required for the conversion. Intermediate steps may, however, also have been left out because of a lack of evidence, or to simplify the description of a process. For these reasons, an apparent disagreement on the underlying biology between multiple descriptions of the metabolic network could also be caused by different decisions on how to represent the same knowledge. If intermediate steps have been left out for simplification only, a mechanism to retrieve these steps should be provided to allow users to determine themselves the level of detail that is required for the application at hand. Only HumanCyc enables its curators to indicate 'subreactions', but this option has not been used yet (release 15.5). The reasons for leaving out intermediate steps are not indicated in a structured way in any of the databases.

#### *Linking of reactions*

For various types of network analysis, first the network needs to be constructed from the individual reactions in a pathway database (Lacroix *et al*, 2008). In HumanCyc, KEGG, and Reactome reactions are explicitly linked to the preceding and/or following steps both within pathways as well as across pathways (Table 5). For each reaction KEGG also stores its main compounds, which connect consecutive reactions. In HumanCyc the main compounds are deduced automatically per pathway for most reactions (Karp and Paley, 1994); in Reactome main compounds are only captured in the graphical representation of the pathway. Note that in Reactome a metabolic reaction can be preceded by the activation of an enzyme catalyzing this reaction. This

| | | | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|---|---|
| **reaction** | type | | indirectly (link to NC-IUBMB) | indirectly (link to NC-IUBMB) | 2 parallel reaction ontologies, *i.e.*, classification by conversion type[a] or by substrate | classified in KEGG BRITE [a], a collection of functional hierarchies | indirectly (link to GO Biological Process) |
| | linked to next/previous reaction? | | no | no | yes, to previous reaction(s) within a pathway<br>Links to other pathways via a metabolite. | yes, successive steps within a pathway are linked<br>Links to other pathways via a metabolite. | yes, to preceding reaction(s) and/or pathway(s) |
| | direction of writing | **irreversible** reactions | as stored in NC-IUBMB | physiological direction | as stored in NC-IUBMB | as stored in NC-IUBMB | direction it has in the pathway |
| | | **reversible** reactions | as stored in NC-IUBMB | direction it has in the pathway | as stored in NC-IUBMB | as stored in NC-IUBMB | both directions, stored separately |
| **metabolites** | type | | indirectly (links to metabolite databases) | indirectly (links to metabolite databases) | classified in their own compound ontology | classified in KEGG BRITE, a collection of functional hierarchies | indirectly (links to metabolite databases) |
| | protonation state | | always the neutral form | most common state at pH level 7.2 | most common state at pH level 7.3 | always the neutral form | always the neutral form |

**Table 5 – Representation differences on reaction level.** [a] Includes the NC-IUBMB classification. NC-IUBMB: Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.

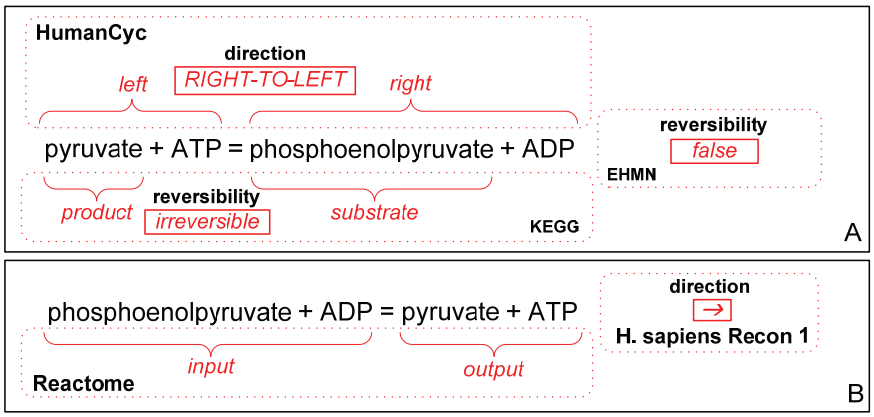



**Figure 2 – Last, irreversible, step of glycolysis**. Words in *italic* indicate reserved terms in a database. The term used to indicate the side of the metabolite in the reaction is shown using braces. The term used to describe the reversibility of the reaction is shown in a solid-lined box. Each database indicates the direction of the reaction differently as shown in the dotted-lined boxes. For example, KEGG indicates for the main metabolites of a reaction whether it is a substrate or product, thereby implying the direction of the reaction. HumanCyc, on the other hand, explicitly indicates the direction with, in this example, 'RIGHT-TO-LEFT'. (A) EHMN, KEGG, and HumanCyc store the reaction in the direction defined by NC-IUBMB. (B) Recon 1 and Reactome store the reaction in the physiological direction.

gives a more complete view of a biological process, compared to only describing its metabolic component. However, as stated above, retrieving solely the metabolic processes from Reactome is difficult. How to link reactions to each other is not explicitly indicated in EHMN and Recon 1. In this case reactions are generally linked based on the substrates or products they have in common. This strategy makes network construction more difficult due to 'currency' metabolites (ATP, $H^+$, etc) connecting unrelated reactions (Huss and Holme, 2007). The number of possible connections can be restricted by only linking reactions via metabolites that are assigned to the same compartment.

### *Physiological direction*

Knowing the physiological direction of reactions is crucial when, for example, building an *in silico* model to predict phenotypes. All databases indicate the direction in slightly different ways (Figure 2, Table 5). In general, reactions in EHMN, HumanCyc, and KEGG are stored in the direction defined by NC-IUBMB, which is not necessarily the physiological direction in which the reaction takes place in human. For example, the last step of glycolysis, the formation of pyruvate, is given in the direction opposite to the one in which it takes place (Figure 2A). EHMN only indicates that the reaction is irreversible and thus does not provide the correct physiological direction. In KEGG (ir)reversibility of a reaction is indicated independently of the specific organism, hereby ignoring that whether a reaction is reversible or not varies among species. Note that in Reactome reversibility is only

indicated by providing a link to the reaction in the opposite direction if the reaction is reversible. Finally, only in HumanCyc the direction can be defined in the context of (i) a specific enzyme that catalyzes the reaction, (ii) the pathway or (iii) only the reaction itself. Only for 4% of the reactions, mainly transport reactions, the direction is specified in the context of the enzyme. Furthermore, for around 40% of the reactions the direction is not specified at all. These reactions are also not included in any pathway.

*Metabolites*

The identity of a metabolite is determined by its name and by identifiers from specialized metabolite databases such as ChEBI (de Matos *et al*, 2010). An identifier is meant to unambiguously designate a metabolite across multiple resources. Identifiers enable, for example, mapping of experimental data onto the network. All five databases try to link a metabolite to at least one specialized metabolite database. The number of metabolite databases referred to differs (Supplementary Table S2). To minimize ambiguity, it is advisable that pathway databases provide a link to a single, common metabolite database for every metabolite. However, in practice this is not possible yet, as metabolite databases are far from complete.

Metabolite databases use different criteria for assigning IDs to metabolites. Consequently, the type of ID chosen affects the characterization of a metabolite in a pathway database and the level of distinction that can be made. For example, in ChEBI, both a base and its conjugate acid are assigned separate IDs, whereas in KEGG Compound no distinction between these two is made, and they are combined in a single entry with one ID. KEGG, EHMN, and Reactome prefer to state the neutral form of the metabolite. This choice may result in a different ChEBI ID compared to the one assigned to a metabolite by Recon 1 and HumanCyc, which specify the most common protonation state of a metabolite at a pre-defined and fixed pH level. Indicating the correct protonation state is important to be able to build a charge-balanced network. As the pH level varies between compartments, metabolites may have multiple protonation states. Ideally, databases should also be able to indicate these multiple states, which is currently not the case in any of the databases.

The pathway databases have made different choices with respect to how much information about metabolites is contained in the pathway database itself. EHMN, Recon 1, and Reactome contain the least information on metabolites and refer to specialized metabolite databases for additional information. Aside from referring to metabolite databases, KEGG and HumanCyc themselves also provide detailed information about metabolites (Table 6). KEGG even has its own metabolite database

| | | EHMN | Recon 1 | Humancyc | KEGG | Reactome |
|---|---|---|---|---|---|---|
| formula | | ✗ | ✓ | ✓ | ✓ | ✓ |
| charge | | ✗ | ✓ | ✓ | ✗ | ✗ |
| mass | | ✗ | ✗ | ✓ | ✓ | ✗ |
| Gibbs free energy of formation | | ✗ | ✗ | ✓ | ✗ | ✗ |
| structure | InChI | ✗ | ✗ | ✓ | ✓ | ✗ |
| | SMILES | ✗ | ✗ | ✓ | ✗ | ✗ |
| | other[a] | ✗ | ✗ | ✓ | ✓ | ✓ |

**Table 6 – Characteristics metabolites.** Check mark: information is present. Cross: information is not available. [a] Other structure formats: mol file (KEGG), structure atoms and bonds (HumanCyc), atomicConnectivity and chemical fingerprint (Reactome)

(KEGG Compound), which is often referred to by many other pathway databases. HumanCyc and KEGG also provide their own hierarchical classification of the metabolites (Table 5 and Supplementary Figure S1), although not as extensively as ChEBI. These ontologies are a powerful way to provide some level of abstraction, and at the same time explicitly define what is meant by the abstract term. HumanCyc uses compound classes, *e.g.* 'an alcohol', in reactions as a level of abstraction. Such generic metabolites are, in general, linked to specific metabolites, and used to represent the broad substrate specificity of an enzyme (see Case Study). If for a generic metabolite no specific metabolites are provided at all, the substrate specificity of the enzyme is likely to be undetermined (BioCyc, 2012). Especially when constructing a computational model, it is important to be able to instantiate such generic metabolites and derive specific reactions. For HumanCyc a mechanism for instantiation of generic reactions has recently been added to their Pathway Tools software for the purpose of building flux balance analysis models (Latendresse *et al*, 2012). Instantiation is done using the compound ontology by selecting those combinations of specific instances of the generic metabolites that lead to a mass balanced reaction. However, not for every generic reaction appropriate instances exist which fulfill this requirement. Moreover, part of the generic reactions cannot be instantiated because multiple products are possible for a given substrate (and vice versa). Recon 1 was specifically built to serve as an *in silico* model capable of predicting phenotypes. For this purpose very generic reactions do not provide enough information and were therefore not included in Recon 1. The (broad) substrate specificity of an enzyme is, consequently, not explicitly captured in this database. In Reactome some level of abstraction is given by grouping metabolites that undergo the same conversion into a set (Supplementary Figure S2).

### *Representation of enzymes and their encoding genes*

A metabolic reaction is nearly always catalyzed by an enzyme, which in turn is encoded by one or more genes (Figure 1). An enzyme may be a single protein or a complex consisting of multiple copies of the same protein (homomer) or of multiple different proteins (heteromer). The concepts of an enzyme and a gene are represented differently in each database or, in some cases, even not at all (Figure 3). The same holds for the gene-protein-reaction relationship.

#### *Enzymes*

The concept of an enzyme, as defined above, is not explicitly represented in KEGG. Information on a protein, *e.g.*, its sequence, is indicated at the gene level. Furthermore, aside from a few exceptions, complexes are not indicated in KEGG. EHMN merely represents an enzyme by the Uniprot ID(s) assigned to the protein(s) constituting the enzyme; also in EHMN complexes are not represented. In Recon 1 heteromeric complexes are represented using a Boolean expression. However, this is not done on protein level, as one would expect, but at gene level. Homomeric complexes are not represented at all. HumanCyc and Reactome do have a separate enzyme level and both types of complexes, *i.e.*, heteromers and homomers, are represented.

#### *Genes*

A gene is defined by its Entrez Gene ID in EHMN, KEGG and Recon 1, while HumanCyc has its own definition which is closer to the definition of Ensembl. Reactome focuses more on proteins. Genes are not represented as a single entity in their MySQL database, but as a collection of identifiers from different genome databases (Figure 3). The various identifiers provided by Reactome are only united through their link to the same protein entry. As for metabolite databases, also genome databases use different criteria for assigning an identifier. The answer to the seemingly simple question of how many genes are involved in the human metabolic network according to each pathway database depends on which type of identifier one counts or whether one follows the convention of the pathway database itself.

#### *Gene-Protein-Reaction relationship*

There is not necessarily a one-to-one relation between a reaction and the catalyst, *e.g.*, multiple enzymes may catalyze the same reaction (isozymes) (Karp and Riley, 1993). EHMN and KEGG do not specify whether the products of multiple genes linked to the same reaction are isozymes, which can separately catalyze the reaction, or that the products together form a complex. This could result in incorrect conclusions with respect to the feasibility of a reaction when studying the effect of a protein deficiency

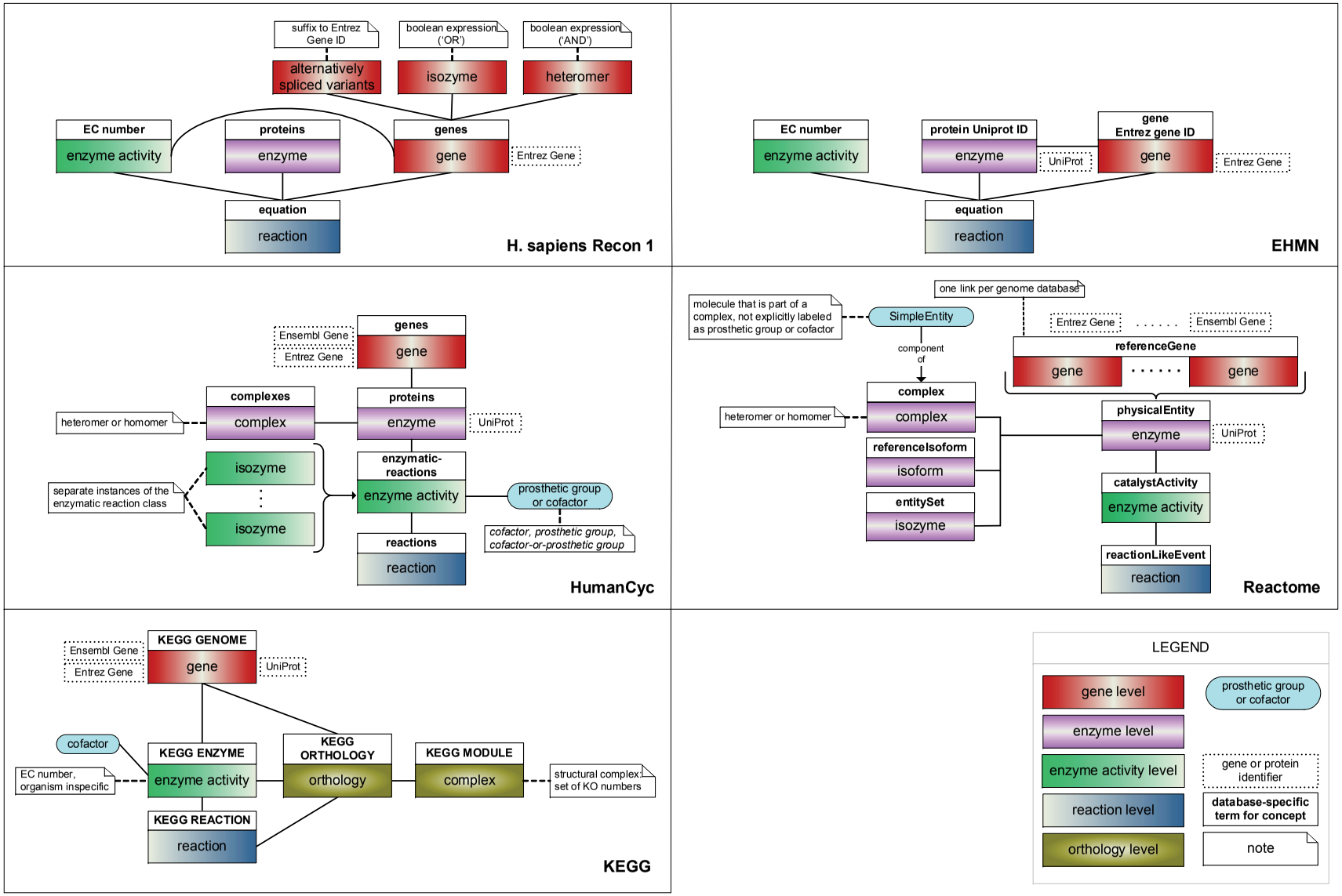



**Figure 3 – Representation differences on the level of enzymes and encoding genes.** Boxes with a dotted line indicate where in the data model the specific identifier is provided. KEGG: The KEGG MODULE database describes four types of modules, among which structural complexes. Only the complexes of the electron transport chain and oligosaccharyltransferase are represented in this way.

(Schellenberger *et al*, 2010). Furthermore, the EC number and KEGG Orthology (KO) number are used instead of a specific enzyme to connect a gene to the corresponding reaction in KEGG. Moreover, only the gene level is organism-specific and the relation between 'enzyme activity' (EC number) and reaction is not. To retrieve species-specific reactions, the gene coding for the enzyme that catalyzes a reaction needs to be known. In Recon 1 the relation between a protein and the encoding gene is only available via its website. Isozymes are represented using a Boolean expression and are defined on gene level in Recon 1. Reactome groups isozymes on enzyme level into a set, each member of which can catalyze the reaction it is linked to. In HumanCyc isozymes are implied when multiple proteins are separately linked to the same reaction (Figure 3). None of the five databases indicates tissue-specificity of isozymes and isoforms, aside from statements in unstructured comment fields, which cannot easily be used in computational analyses.

### *Representation of uncertain and missing knowledge*

Our current knowledge on human metabolism is incomplete and based on different types of evidence such as biochemical, genetic, or sequence data and studies on other organisms. It is, therefore, important that databases explicitly indicate the source of a piece of knowledge and the degree of confidence associated with it. The databases, except EHMN, commonly cite scientific articles as evidence source. HumanCyc and Recon 1 also indicate the type of evidence available (Table 7). HumanCyc uses an evidence ontology (Karp *et al*, 2004) with 160 terms such as 'Inferred from experiment', which can be combined with a probability that the evidence is correct. However, evidence codes are available for less than 18% of the reactions and their catalyzing enzymes. Evidence codes are available for each pathway as a whole. In Recon 1 the type of evidence is indicated for each reaction. Five types of evidence are discerned which are assigned a confidence score (Table 7) (Thiele and Palsson, 2010b).

It is also important to explicitly indicate a complete lack of information, which is, however, rarely done. In Recon 1, for instance, 'cytosol' is used as the default value if the compartment is unknown, instead of explicitly indicating that there is a lack of knowledge as done in EHMN (Table 7). Similarly, of the five databases only HumanCyc explicitly distinguishes a spontaneous reaction from a reaction for which the enzyme is unknown. In KEGG, spontaneous reactions are only indicated in unstructured comment fields. Moreover, for an organism-specific network spontaneous reactions cannot be retrieved as this requires the presence of a gene. The same holds for reactions for which the corresponding gene is unknown for the organism of interest.

| Database | Type of evidence | Degree of confidence |
|---|---|---|
| EHMN | ▪ source of transport reactions is indicated:<br>- dead-end analysis (Hao *et al*, 2010),<br>- *H. sapiens* Recon 1,<br>- another database | ▪ if the compartment of a protein is unknown it is set to uncertain |
| Recon 1 | ▪ five types of evidence are discerned:<br>- biochemical<br>- genetic<br>- sequence homology<br>- physiological<br>- modeling[a]<br>▪ literature references | ▪ five confidence score, ranging from 0 (low) to 4 (high), reflecting the information and evidence currently available:<br>4 – biochemical data<br>3 – genetic data<br>2 – sequence homology or physiological data<br>1 – modeling data[a]<br>0 – unevaluated<br>If multiple types of evidence are available scores are added.<br>▪ remarks in comment fields |
| HumanCyc | ▪ evidence ontology containing 160 terms, main evidence types:<br>- inferred from computation<br>- inferred from experiment<br>- inferred by curator<br>- author statement<br>▪ basis for assignment of a protein to a reaction, *e.g.*, EC number or decision of curator<br>▪ literature references | ▪ probability that the evidence is correct<br>▪ orphan reaction:<br>reaction for which no enzyme that catalyzes the reaction has been sequenced<br>▪ unbalanced reactions are labeled<br>▪ reaction may be labeled as hypothetical[b]<br>▪ remarks in comment fields |
| KEGG | ▪ literature references | ▪ remarks in comment fields |
| Reactome | ▪ data supported by evidence from other organisms is indicated<br>▪ literature references | ▪ CandidateSet: enzymes hypothesized to catalyze the reaction<br>▪ BlackBoxEvents used for:<br>- reactions that have imbalances for various reasons<br>- complex processes of which not all details are known<br>- summarizing a complex process in a single step for which each step is known<br>▪ OtherEntity:<br>entities that curators are unable or unwilling to describe in chemical detail<br>▪ GenomeEncodedEntity: polypeptide or polynucleotide whose sequence is unknown<br>▪ remarks in comment fields |

**Table 7 – Evidence description.** [a] Reaction is included, because it improved the performance of the *in silico* model. [b] The presence of the substrates, products or catalyst of the reaction have not yet been demonstrated.

Finally, it is not always indicated, or only in a comment field, that intermediate steps have been left out. In Reactome a 'BlackBoxEvent' can be used for this, which, however, does not necessarily mean that the intermediary steps are unknown.

## Case study

We selected three complex metabolic processes to further illustrate the different challenges faced by the pathway databases in accurately representing the metabolic network *in silico*. Here we discuss fatty acid beta oxidation (Berg *et al*, 2012, pp. 663-696) in more detail. Two more examples, oxidative phosphorylation (Berg *et al*, 2012, pp. 543-584) and the pyruvate dehydrogenase reaction (Berg *et al*, 2012, pp. 515-542), are discussed in the Supplementary Text S1 and S2, respectively. These three case studies show the implications of different design decisions on the ability of the pathway databases to represent a biological process in full detail.

### *Fatty acid beta oxidation*

We focused on the beta oxidation of saturated fatty acids with a straight chain of even length. One particular challenge in representing this pathway is the repetitive nature of this process. The chain of an activated fatty acid is shortened by two carbons via four subsequent reactions, yielding one unit of acetyl-CoA. Several chain-length specific isozymes are available for each cycle. For the complete oxidation of a fatty acid, this cycle is repeated until only acetyl-CoA is left. The number of cycles needed depends on the chain length of the fatty acid. There is a wide range of fatty acids of which the majority, *i.e.*, short-chain, medium-chain and long-chain fatty acids, are degraded in the mitochondrion. In mammals and many fungi, very-long-chain fatty acids are first shortened in the peroxisome after which they may be transported to the mitochondrion for further oxidation (Cornell *et al*, 2007). The exact enzymes involved differ in the two compartments, but the reactions are the same except for the co-substrates of the first step of a cycle.

### *Representation*

Each database has its own strengths and limitations in representing fatty acid beta oxidation (Table 8). KEGG and HumanCyc make no distinction between the peroxisomal and mitochondrial pathway, which emphasizes the similarities, but disregards the differences. For KEGG it is difficult to separate the two pathways, since KEGG does not provide information on compartments. Moreover, as mentioned above, pathways in KEGG are not species-specific and the distinction between the peroxisomal and the mitochondrial pathway does not hold for the majority of organisms described in KEGG. In HumanCyc the repetitive nature of this metabolic process is captured by describing a single cycle, using generic metabolites

| | | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|---|
| repetitive process | the four steps of a cycle | ✓ | ✗ | ✓ | ✓ | ✓ |
| | cycle described multiple times | ✓ | ✗ | ✗ | ✓ | ✓ |
| | degradation described for the complete range of fatty acids | ✗ | ✓ | –[a] | ✗ | ✗ |
| | every step of the complete degradation of fatty acids included in the database | ✓ | ✗ | –[b] | ✓ | –[c] |
| mitochondrial versus peroxisomal pathway | differences | ✓ | ✓ | ✗ | ✗ | ✓ |
| | similarities | ✗ | ✗ | ✓ | ✓ | ✗ |
| chain length specificity of enzymes | | ✓ | ✗ | ✗ | ✓ | –[c] |

**Table 8 – Representational challenges fatty acid beta oxidation**. Check mark: concept is represented. Cross: concept is not represented. Bar: not for all cases the concept is represented. [a] Not all possible fatty acids are specified as instances for each of the generic metabolites. [b] By using generic metabolites to represent one cycle all possibilities are captured, but not all possible specific instances can be deduced (Figure 4). [c] For the mitochondrial pathway each step of all cycles is described. For the peroxisomal pathway eight of the nine cycles are summarized in a single reaction ('BlackBoxEvent').

to describe the oxidation for the complete range of saturated fatty acids. Furthermore, a 'polymerization link' explicitly connects the fatty acid before and after the removal of the 2-carbon acetyl-CoA (Figure 4). Which combinations of instances of the generic metabolites together form the specific reactions is, however, not explicitly indicated. Moreover, in this particular example, several of the required metabolite instances for each of the four steps are lacking. The chain length specificity of the enzymes is also not captured in HumanCyc. Reactome chooses to represent the repeating cycles of the mitochondrial pathway as separate subpathways and to describe each step of every cycle. In this way, Reactome is able to represent the chain length specificity of each enzyme. Note though that all four steps in the peroxisomal pathway are described for one cycle only, the next eight cycles are lumped into a single step. No mechanism is provided that allows users to retrieve the intermediate steps of these eight cycles. Similarly, in Recon 1 the repetitive nature of this process is not captured as the conversion of, for example, palmitoyl-CoA into octanoyl-CoA is described in a single step instead of 16 steps. Also, the four steps of a single cycle are described for none of the fatty acids. This decision is indicated in the comment field, but the intermediate steps cannot be retrieved. Describing each step is, however, important to be able to simulate enzyme deficiencies that lead to the abnormal build up of intermediate products of a cycle of beta oxidation (Das *et al*, 2006; Molven *et al*, 2004; Wanders *et al*, 1992; Wanders *et al*, 1990). In EHMN and KEGG every step is described for fatty acids with a chain of length 16 or shorter, but not for those with a longer chain length. A disadvantage of describing every step is that, in contrast to the generic approach of HumanCyc, it

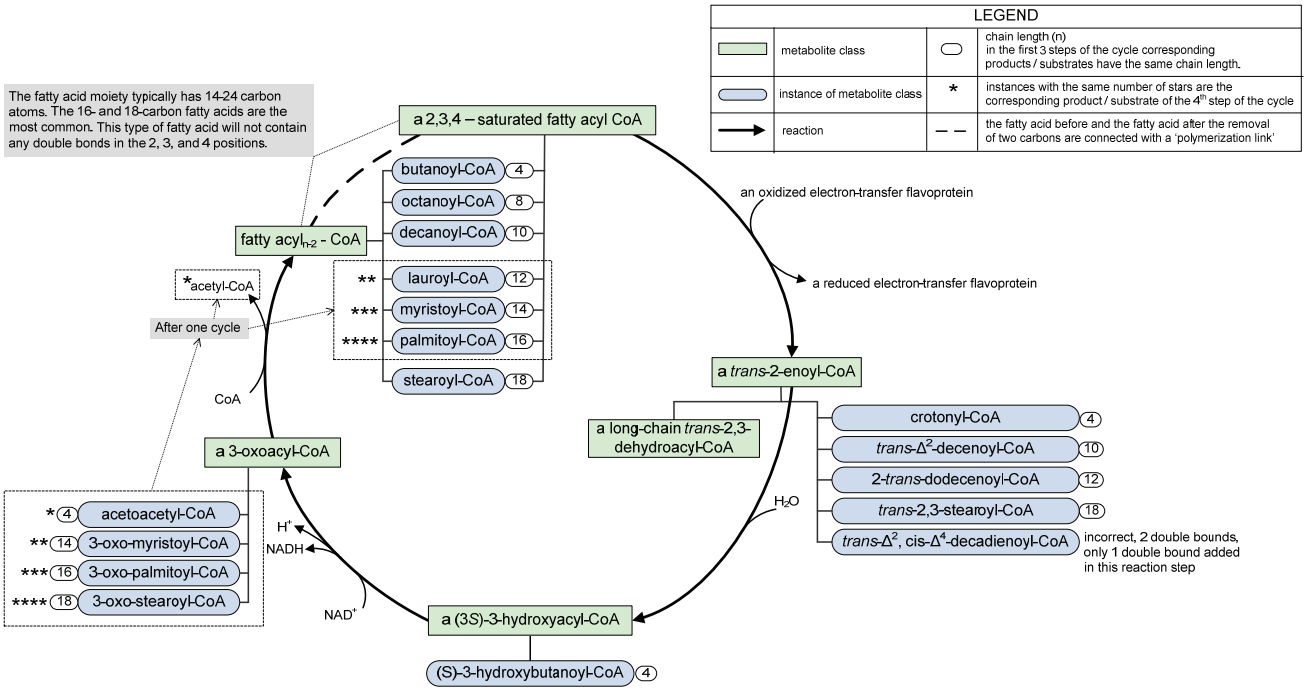



**Figure 4 – Fatty acid beta oxidation in HumanCyc.** Part of the 'fatty acid β-oxidation I' pathway in HumanCyc, only the reactions of the cycle itself are shown.

requires that each of the highly similar steps needs to be specified separately for the whole range of possible fatty acids. Moreover, not for all applications the intermediate steps are of interest. A mechanism that allows database users to switch between a high-level and a more detailed representation would be preferred.

# Challenges

### *Within a database*

It remains a challenge to capture every detail of the knowledge on the human metabolic network, both for relatively straightforward processes like gluconeogenesis (Figure 1) and the more complex processes illustrated by the three case studies. On the other hand, the question which level of detail is required to be able to perform a wide range of possible computational analyses does not always have a clear-cut answer. For example, to represent a (de)polymerization processes it is not always an option nor always strictly necessary to specify each step. The degradation of glycogen, for instance, consists of thousands of steps (Berg *et al*, 2012, pp. 637-661), but not each intermediate product might be of interest. Furthermore, it is also important for users to have some degree of abstraction, such as the ontologies provided by some databases, to see how everything fits in the bigger picture.

Unstructured text fields in the databases, which cannot be easily used in computational analyses, frequently contain more detailed information, such as the tissue-specificity of an enzyme. As argued by Khatri *et al*. (2012), tissue- and cell-specific information is essential to improve the accuracy and relevance of pathway

analyses. In the Biological Connection Markup Language format (Beltrame *et al*, 2011) that was recently proposed, this information can be stored, but this format has not yet been adopted by the major databases. Ultimately, there are even more factors to consider to accurately capture the complete (human) physiology in a digital format. This includes the inherent dynamics of metabolic processes and the multiple levels at which these processes are being controlled. The complexity of metabolism is further increased by its multi-scale nature ranging from cellular compartments, and cells to organs. These issues are not (yet) addressed by the pathway databases selected in this review. However, several large-scale projects have taken up this challenge. For example, the goal of the Virtual Liver project (http://www.virtual-liver.de) (Holzhütter *et al*, 2012) is to construct a multi-scale representation of liver physiology.

The ability to indicate that a piece of knowledge is missing is another desirable characteristic of a pathway database. This is, however, not yet done in each database. A further extension of the pathway databases is to not only provide affirmative evidence as done by HumanCyc and Recon 1, but to also indicate 'negative evidence' such as a statement that a reaction cannot take place in human. This would enable users to distinguish such cases from knowledge gaps. This information is highly valuable, especially given that the (human) metabolic network is not yet complete. Note that although we focussed on the human network, most observations also hold for the other organisms the five databases describe (when applicable).

### *Across databases*

Efforts to reconcile different descriptions of the metabolic network for a specific species are hampered by the representation differences discussed in this paper (Chindelevitch *et al*, 2012; Herrgård *et al*, 2008; Radrich *et al*, 2010; Thiele *et al*, 2011). Similar problems arise in analyses that require the metabolic network of multiple organisms from different databases, *e.g.*, to study evolution (Tanaka *et al*, 2006). For these purposes it is essential to be aware what differences could be caused by a difference in representation rather than a true difference in opinion on the underlying biology. One example is the number of steps a process is described in (see Case Study and Supplementary Text S2). Furthermore, it is important to realize that even the smallest difference in terminology and the definition of a concept needs to be accounted for.

To simplify the exchange of knowledge between databases several standards have been proposed such as SBML and BioPAX, each with their own advantages (Strömbäck and Lambrix, 2005). However, these do not resolve all representational

differences that we discussed. HumanCyc and Reactome provide their network in the BioPAX (level 3) format (Demir *et al*, 2010). They, however, followed their own representation when converting their data into this format. For example, in Reactome a reversible reaction is stored separately in both directions in their own data model and also in their BioPAX file. In HumanCyc's data model a reversible reaction is stored only once and its direction is indicated as 'reversible'. The same is done in their BioPAX file. Semantic standards like BioPAX can also not enforce the level of detail in which a process needs to be described by a curator or how a pathway is defined. Curators will, therefore, need to adhere to strict guidelines to make these exchange formats more easily comparable. Alternatively, rules could be formulated to translate one representation into the other. Based on Figure 3, for example, one could develop more precise rules to translate the different ways of representing the relations between gene products. The results of our comparison and the accompanying overviews provide useful insights in the road ahead to further simplify integration of the knowledge contained in pathway databases. Integration will also enable the construction of a more accurate *in silico* representation of metabolic networks. Endeavors in this direction have already been undertaken for multiple organisms (Herrgård *et al*, 2008; Thiele *et al*, 2011), including human (Thiele *et al*, submitted).

## Discussion

The five pathway databases each have made different decisions on how and in what detail to represent the metabolic network. Of these five databases EHMN provides the least detailed information and HumanCyc the most (Table 2). At the same time not every aspect of the data model of HumanCyc is used yet. Filling in every detail will likely require a lot of time and effort from the curators. Moreover, the lack of knowledge on human metabolism may currently preclude the usability of every feature of HumanCyc's data model. It also depends on the application at hand which aspects of the metabolic network are important and to what detail they need to be represented in a structured format. The overviews given in this paper provide insight into the differences between the databases and can help to make a well-informed decision on which database to use. The different choices the database developers have made each have their own advantages and disadvantages. For example, to perform *in silico* simulations and predicting metabolic phenotypes, Recon 1 may be preferred. This network is fully compartmentalized and fully mass and charge balanced. Also the relation between gene products is provided, which is important for simulating the effect of, for example, gene defects. On the other hand, for pathway enrichment analyses (Antonov *et al*, 2008) each of the five databases may

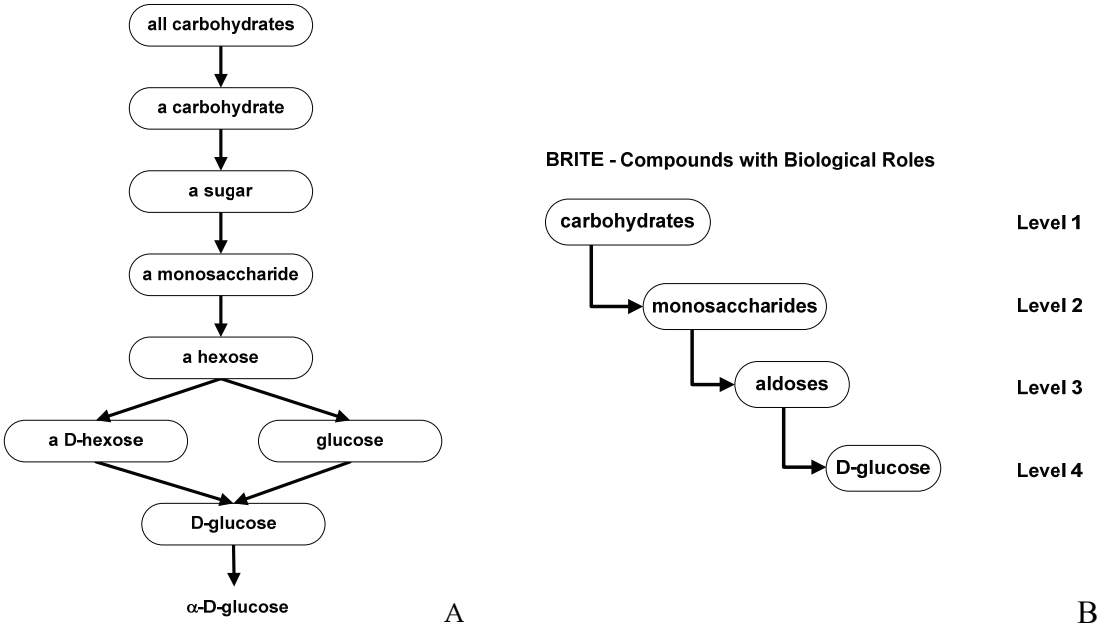
suffice and the determining factor is the pathway definition used. Other factors than the representation of the network will also play a role in selecting the database that best fits the application, such as the coverage of the metabolic network (Stobbe *et al*, 2011), *e.g.*, EHMN contains a more extensive description of lipid metabolism than other databases. Finally, retrieving and capturing every detail of the (human) metabolic network in a digital format is a huge challenge and will require a joint effort of a broad scientific community.

## Acknowledgements

We would like to thank the reviewers for their helpful comments and suggestions for improving the presentation and comprehensibility of the paper. 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.

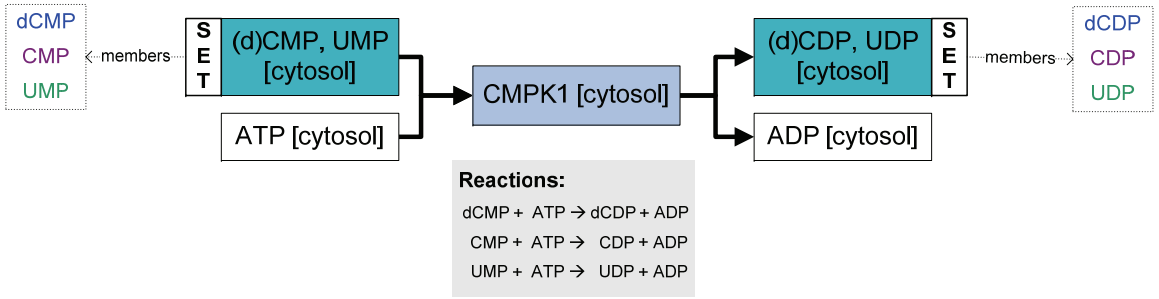
## Supplementary material

### Supplementary Figure S1 – Metabolite hierarchy HumanCyc and KEGG




A) HumanCyc: The different levels of the metabolite hierarchy ending at α-D-glucose.
B) KEGG: hierarchy has four levels and stops at D-glucose in contrast to HumanCyc (A).

## Supplementary Figure S2 – Sets used in Reactome




Sets are used to group metabolites undergoing the same conversion. The relation between the two sets is not explicitly indicated. Consequently, deducing the correct specific reactions is dependent on the members of the sets being in the same order.

## Supplementary Table S1 – Information sources used

| Database | Data files analyzed | Online manuals | Articles of interest |
|---|---|---|---|
| EHMN | Excel file | --- | (Hao *et al*, 2010; Ma *et al*, 2007) |
| Recon 1 | Tab-delimited files, SBML | --- | (Duarte *et al*, 2007; Schellenberger *et al*, 2010; Thiele and Palsson, 2010b) |
| HumanCyc | Pathway Tools (export flat files), BioPAX file | - User's guide available via the Pathway Tools program<br>- Curators's guide:<br>http://bioinformatics.ai.sri.com/ptools/curatorsguide.pdf | (Green and Karp, 2006; Karp *et al*, 2005; Karp *et al*, 2004; Latendresse *et al*, 2012; Romero *et al*, 2004; Zhang *et al*, 2005) |
| KEGG | KGML, flat files | http://www.kegg.jp/kegg/xml/docs/ | (Kanehisa *et al*, 2010; Kanehisa *et al*, 2012) |
| Reactome | MySQL dump, BioPAX file | http://wiki.reactome.org/index.php/Main_Page | (Croft *et al*, 2011; Vastrik *et al*, 2007) |

## Supplementary Table S2 – Identifier types

| | **Metabolite databases** |
|---|---|
| EHMN | KEGG Compound, KEGG Glycan |
| Recon 1 | KEGG Compound, CAS |
| HumanCyc | KEGG Compound, ChEBI, PubChem Compound, CAS, InChI, SMILES, NCI, UM-BBD-CPD, KNApSAck, Wikipedia |
| KEGG | KEGG Compound, KEGG Glycan, ChEBI, PubChem Substance, CAS, InChI, 3DMET, KNApSAcK, NIKKAJI, LIPIDMAPS, LIPIDBANK |
| Reactome | KEGG Compound, ChEBI, PubChem Substance |

| | **Protein databases** |
|---|---|
| EHMN | UniProt |
| Recon 1 | --- |
| HumanCyc | NCBI Protein, UniProt |
| KEGG | HPRD, NCBI-GI, UniProt |
| Reactome | RefSeq, UniProt |

| | **Gene databases** |
|---|---|
| EHMN | Entrez Gene |
| Recon 1 | Entrez Gene |
| HumanCyc | Entrez Gene, Ensembl Gene, UCSC, UniGene, Entrez, Genecards, RefSeq_NM |
| KEGG | KEGG Gene ID, NCBI-GeneID (=Entrez Gene), Ensembl Gene, HGNC |
| Reactome | Entrez Gene, Ensembl Gene, UCSC, KEGG Gene ID, BioGPS |

Note that some identifiers are only present for a few proteins/genes/metabolites.

## Supplementary Text S1 – Case Study: Oxidative phosphorylation

*Biological background*

Oxidative phosphorylation is a crucial biological process as it is the main source of ATP in human (Berg *et al*, 2012, pp. 543-584). The electrons from NADH and $FADH_2$ are transferred to oxygen via a number of electron-transfer reactions using multiple transmembrane complexes in the mitochondrion. Via prosthetic groups electrons flow through the complexes. This causes protons to be pumped from the matrix side of the inner mitochondrial membrane to the cytosolic side of the membrane, creating a pH gradient and membrane potential that drive the synthesis of ATP.

*Representation*

In HumanCyc oxidative phosphorylation is not represented as a single pathway, but only as isolated reactions. In EHMN it is also not represented as a separate pathway, moreover not all reactions are described. KEGG does represent oxidative phosphorylation as a pathway, but the reactions are only shown in the pathway diagram itself and cannot be retrieved in an automated way. All five databases only describe the details of oxidative phosphorylation, such as the role of the prosthetic groups and the change in pH gradient, in unstructured text or in a figure. The redox half-reactions are captured by the data model of HumanCyc, providing some more detail. For KEGG this process is one of the few cases in which the enzymes are indicated as complexes, while in HumanCyc, although possible in their data model, the complexes are not represented as such. Recon 1 lacks the ability to indicate the intermembrane space as the compartment for part of the metabolites. Instead, in Recon 1, the protons are indicated to be pumped into the cytosol. HumanCyc erroneously indicates the 'periplasmic space' and 'inner membrane (sensu Gram-negative Bacteria)' as compartments for some of the metabolites. In Reactome the compartments of the metabolites are correctly indicated. The enzymatic complexes are assigned to the mitochondrial inner membrane. Although true for components of the complexes, all of them are transmembrane enzymes with also parts in the mitochondrial matrix and mitochondrial intermembrane space.

**Table S1.1 - Representation challenges oxidative phosphorylation**

| | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|
| complexes (heteromer) | ✗ | ✓ | ✗ | ✓ | ✓ |
| redox half-reactions | ✗ | ✗ | ✓ | ✗ | ✗ |
| pH gradient | ✗ | ✗ | ✗ | ✗ | ✗ |
| role prosthetic group | ✗ | ✗ | ✗ | ✗ | ✗ |
| metabolites in intermembrane mitochondrial space | ✗ | ✗ | ✗ | ✗ | ✓ |
| transmembrane enzymes | ✗ | ✗ | ✗ | ✗ | ✗ |

Check mark: concept is represented. Cross: concept is not represented.

## Supplementary Text S2 – Case Study: Pyruvate dehydrogenase reaction

*Biological background*

Under aerobic conditions pyruvate is converted to acetyl-CoA in three reaction steps, catalyzed by the pyruvate dehydrogenase complex (Figure S2.1A) (Berg *et al*, 2012, pp. 515-542). The first two steps are catalyzed by the pyruvate dehydrogenase component (E1), and the third step by dihydrolipoyl transacetylase (E2). Two more steps are needed to reactivate lipoamide, a prosthetic group, to enable another cycle. Both these steps are catalyzed by the third component dihydrolipoyl dehydrogenase (E3). Only acetyl-CoA, $CO_2$ and NADH are released from the complex, all other (intermediate) products remain bound to the complex.

*Representation*

In each database, a different number of steps is used to describe the degradation of pyruvate (Figure S2.1). EHMN, HumanCyc and KEGG describe it in multiple steps and each component of the pyruvate dehydrogenase complex is linked separately to the reaction in which it is active. In EHMN the same process is even described three times in the same pathway with a different number of steps (Table S2.1, Figure S2.1C, Figure S2.1E, and Figure S2.1G). In HumanCyc there are also two alternative routes described, but these are not assigned to any pathway. Recon 1 and Reactome describe the degradation of pyruvate in one step and represent the catalyst as a complex. In the comment field of Recon 1 it is indicated that multiple reaction steps are lumped into one. In earlier releases (before version 32) of Reactome this process was presented in more detail, describing the process in five steps as in the student textbook 'Biochemistry' (Berg *et al*, 2012) and indicating the changes in the prosthetic groups of the complex. The active component of the complex, however, was also not indicated at that time. It is important to note that in this case the difference in number of steps in which the databases describe this process is not a disagreement on the underlying biology, but a difference in representation. Although one could argue that representing all intermediate products of the pyruvate dehydrogenase reaction as unbound is incorrect as most remain attached to the enzyme complex.

| | | EHMN | Recon 1 | HumanCyc | KEGG | Reactome |
|---|---|---|---|---|---|---|
| network | pathway assigned to | glycolysis and gluconeogenesis | glycolysis / gluconeogenesis | - acetyl-CoA biosynthesis (from pyruvate)<br>- isolated reactions | - glycolysis / gluconeogenesis<br>- citrate cycle (TCA cycle)<br>- pyruvate metabolism | pyruvate metabolism |
| reaction | number of steps process is described in | 4, 3 and 1 | 1 | within pathway: 4<br>isolated reactions: 3 and 1 | 4 | 1 |
| | intermediate products represented as | unbound metabolites | not applicable | proteins | unbound metabolites[a] | not applicable |
| enzyme | complex or separate components? | separate components | complex (heteromer) | separate components | separate components | complex (heteromer) |
| | indicated which protein catalyzes which step? | yes | no | partly, two steps are missing an enzyme | yes | no |
| | prosthetic groups indicated? | no | no | no | yes | yes |

**Table S2.1 - Differences in representation of the pyruvate dehydrogenase reaction** [a] The names of the metabolites and structure of three metabolites do suggest that they are attached to an enzyme.

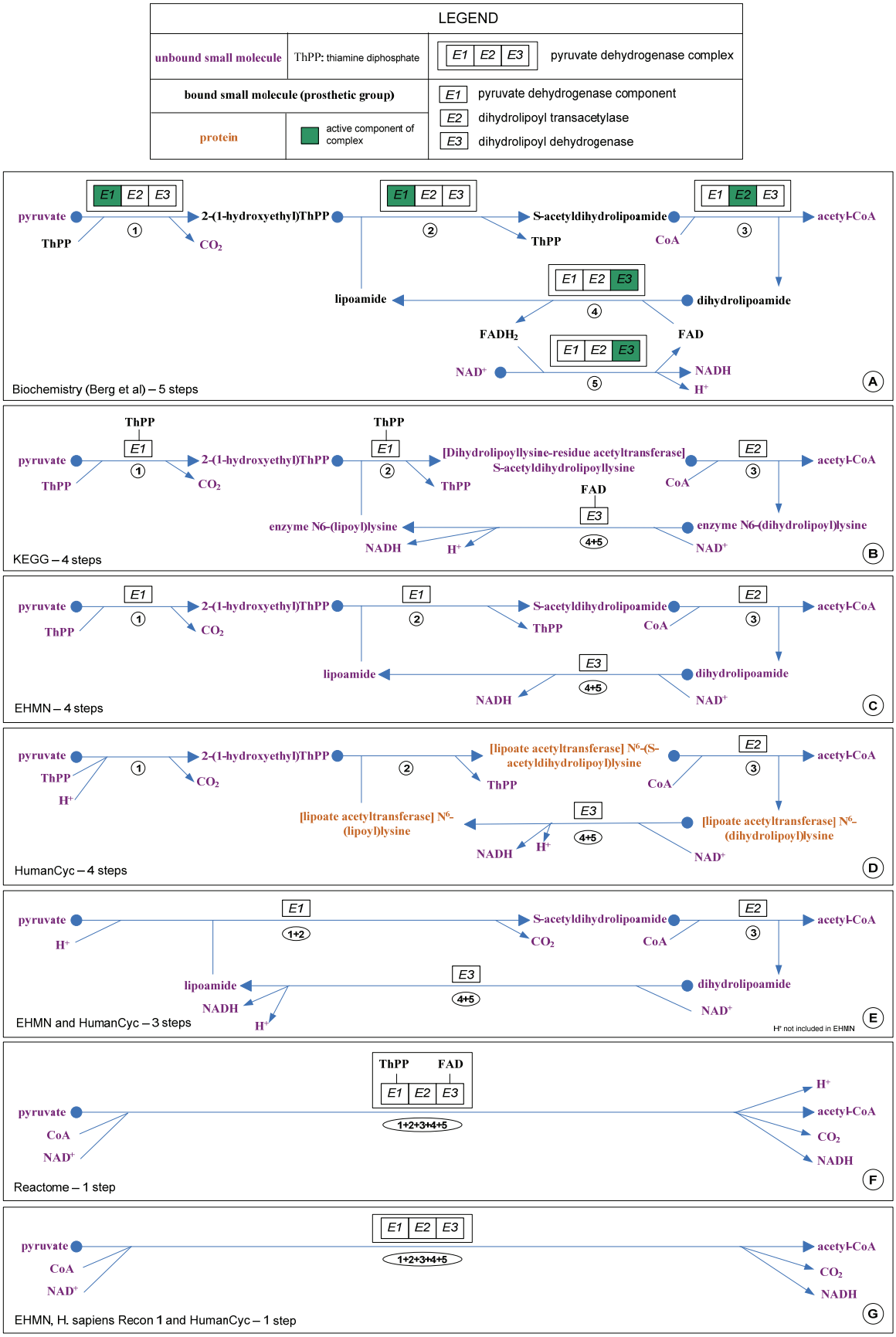



**Figure S2.1 - Pyruvate dehydrogenase reaction as represented in 'Biochemistry' (Berg *et al*, 2012, pp. 515-542) (A) and in each of the databases (B-G).**