The road to knowledge: from biology to databases and back again
Stobbe, M.D.

Citation for published version (APA):
Stobbe, M. D. (2012). The road to knowledge: from biology to databases and back again
Chapter 1

General introduction

“There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.”

J.R.R. Tolkien (The Hobbit)
From biology …

The study of metabolism has a long history, the roots of which can be traced back to the year 1614 when Santorio Sanctorius published his results on body weight fluctuations during the course of a day (Sanctorius, 1614). The realization in the 19th century that the reactions within a cell are the same as those studied in chemistry marked a new era in the field of metabolism. Another important milestone was the discovery that enzymes catalyze the metabolic reactions (Buchner, 1897). Series of reactions leading to a particular biological outcome have traditionally been organized into pathways. The first complete metabolic pathways were described in the 1930s, among which the tricarboxylic acid (TCA) cycle. This classical pathway was discovered by Hans Krebs (Krebs and Johnson, 1937) for which he earned a Nobel Prize in 1953. From the 1930s onwards an increasing number of metabolic pathways have been unraveled. Although often studied in isolation, pathways interact and together constitute what is referred to as the metabolic network. This highly organized and complex network of reactions is able to adapt to a constantly changing environment. Nowadays, metabolic networks are studied on a genome-wide scale for a wide range of organisms, fueled by the tremendous progress in the assembly and functional annotation of whole genomes over the past 15 years.

… to pathway databases …

The knowledge on metabolism gained over the years is scattered across a multitude of resources, including scientific literature. To collect and organize this knowledge a growing number of (public) metabolic pathway databases have been created for many different organisms. Figure 1 provides a high-level overview of the main aspects of the metabolic network as described in a pathway database. One key objective of these databases is to accurately represent the metabolic network in a format suitable for computational processing and analyses. The databases also serve as a digital encyclopedia and are often accompanied with powerful visualization aids. One of the first databases to be made publicly available was the Kyoto Encyclopedia of Genes and Genomes (KEGG) in 1995 (Kanehisa et al., 2012). KEGG was initiated to depart from a list of parts, such as catalogs containing functions of individual genes, to pathways that show how these parts interact. A metabolic pathway database integrates the functional annotation of the genome and the metabolic network of an organism. Such an integrative approach is crucial, since the inner workings of the metabolic network cannot be unraveled by only considering the individual roles of its parts, like we cannot understand the principles of powered flight and working details of a modern aircraft by only considering the components of an airplane laid out on a hanger floor (Vastrik et al., 2007).
Figure 1 - The main aspects of the metabolic network as described in pathway databases. The description of the metabolic network in a pathway database can also be used as a basis for the formulation of a mathematical model.

Only by studying the metabolic network at the systems level one can begin to understand (human) metabolism in healthy and diseased states and for this pathway databases are instrumental.

**Metabolic network construction**

The great value and potential of pathway databases has inspired many research groups to construct their own database for their organism of interest. The metabolic networks of several key organisms, such as *S. cerevisiae* and *H. sapiens*, are even described in a multiple of databases. Different strategies have been used to build a metabolic network (Table 1). Frequently a genome-based approach is taken to
construct an initial draft using the functional annotation of the genome of a particular organism as a basis. Using gene annotation, reactions are retrieved from other databases such as the one of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, KEGG LIGAND (Goto et al., 2002) or MetaCyc (Caspi et al., 2012). The manner and degree of manual curation of the resulting genome-scale network differs among the various pathway databases. In most cases the genome-based draft is further refined based on expert knowledge, scientific literature, and other resources. A next step taken by research groups focusing on systems biology is to convert the network into a constraint-based mathematical model (Thiele and Palsson, 2010b). This model is iteratively adapted and fine-tuned by verifying its ability to simulate known metabolic functions. An alternative approach to the genome-based approach is to build the metabolic network incrementally. One particular example is Reactome in which independent researchers with expert knowledge of parts of the metabolic network are invited to curate a specific process (Croft et al., 2011). New information is peer-reviewed before being added to Reactome. WikiPathways is another example of a database using an incremental approach (Pico et al., 2008). This database enables contributions from the entire community and any user can either add a pathway or edit an existing one.

... and back again.

One of the most popular applications of a pathway database is to provide context for the analysis and interpretation of high-throughput data such as obtained from next-generation sequencing, microarrays, proteomics or metabolomics. Typically, the differentially expressed genes, proteins or metabolites are mapped to the metabolic network to identify the affected metabolic processes. This gives more insight in the possible mechanisms underlying the condition of interest than solely looking at a long list of individual genes, proteins or metabolites (Khatri et al., 2012). Pathway databases also have been used in the study of evolution of the metabolic network across species (Tanaka et al., 2006) and disease-related studies, such as the relation between diseases frequently co-occurring in the same patient (Lee et al., 2008). Furthermore, the holistic view that the metabolic network offers, allows researchers to more easily identify gaps in our knowledge on metabolism. For instance, analyzing the network can reveal metabolites that are produced, but not consumed, indicating that the fate of the metabolites is unknown (Rolfsson et al., 2011). One of the ultimate goals is to construct the full metabolic network in all its detail as a mathematical model that can be used to generate experimentally verifiable hypotheses, to identify potential drug targets and to simulate the effect of network perturbations, such as loss of function (Oberhardt et al., 2009).
<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Construction strategy</th>
<th>Curation</th>
<th># of organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioCyc</td>
<td>Collection of Pathway/Genome Databases, each describing the genome and metabolic pathways of a single organism</td>
<td>genome-based</td>
<td>Three levels of curation: Tier 1 - highly curated, tens of man-years of literature-based curation, Tier 2 - moderately curated, ( i.e., ) partly curated and partly automatically derived, Tier 3 - automatically derived, not curated</td>
<td>&gt; 1700</td>
</tr>
<tr>
<td>BiGG</td>
<td>Biochemical Genetic and Genomic knowledgebase of large scale metabolic networks</td>
<td>genome-based</td>
<td>Curation based on primary literature articles, reviews, and biochemical textbooks and by performing functional validation tests</td>
<td>8*</td>
</tr>
<tr>
<td>EHMN</td>
<td>Edinburgh Human Metabolic Network</td>
<td>genome-based</td>
<td>Curated based on literature and the literature-based Enzymes and Metabolic Pathways database</td>
<td>1 (human)</td>
</tr>
<tr>
<td>INOH</td>
<td>Integrating Network Objects with Hierarchies</td>
<td>incremental</td>
<td>Constructed manually from text books</td>
<td>1 (human)</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes: resource for understanding high-level functions and utilities of the biological system from molecular-level information</td>
<td>genome-based</td>
<td>The main 15 databases KEGG consists of are manually curated. Several auxiliary databases containing genomic information, such as the one for draft genomes, are automatically generated.</td>
<td>&gt; 1700</td>
</tr>
<tr>
<td>Panther</td>
<td>Protein Analysis THrough Evolutionary Relationships: protein sequences are grouped into functional subfamilies, which allows for a more accurate association with biological pathways.</td>
<td>incremental</td>
<td>Manually curated by 20 external experts with an option for community pathway curation</td>
<td>48</td>
</tr>
<tr>
<td>Reactome</td>
<td>Open-source, open access, manually curated and peer-reviewed pathway database</td>
<td>incremental</td>
<td>Human network is curated, other organisms are derived automatically using orthology** Curation is done by invited biological experts and each module is peer-reviewed before it is added to Reactome</td>
<td>20</td>
</tr>
<tr>
<td>UniPathway</td>
<td>A resource for the exploration and annotation of metabolic pathways</td>
<td>starting from reactions, which are linked to proteins</td>
<td>Curated based on literature and existing metabolic resources such as KEGG and MetaCyc</td>
<td>&gt; 2800</td>
</tr>
<tr>
<td>WikiPathways</td>
<td>Open, public platform dedicated to the curation of biological pathways by and for the scientific community</td>
<td>incremental</td>
<td>Curated by the community</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 1 - Selection of publicly available pathway databases. *Ten networks are described, three of which are of E. coli. **There are 4 spin-offs, e.g., FlyReactome and Gallus Reactome, maintained by other research groups.
Outline

Our own interest in metabolic pathway databases started when developing algorithms for the identification of plausible candidates for ‘missing genes’ (Orth and Palsson, 2010). A gene is referred to as ‘missing’ if for a metabolic reaction that is known to take place the gene product (enzyme) catalyzing it is unidentified. When evaluating the performance of our algorithms to identify such missing genes we, however, observed that the outcome depended heavily on the metabolic pathway database used. The dependency of the outcome of computational analyses on the choice of a specific metabolic database has also been observed by others (Lee et al, 2008; Zelezniak et al, 2010). A single, complete, and accurate description of the metabolic network is therefore essential, not only for resolving the ‘missing genes’ problem, but in many other applications such as the analysis of high-throughput data, and the in silico prediction of phenotypes. This motivated us to continue on a different path and focus on the metabolic pathway databases themselves to unravel the reasons for the differences observed in the outcome of computational analyses. Some people considered (and still consider) this path to be short, straight and simple to travel, but it turned out to be a long, but rewarding journey.

In Chapter 2 and 3 we describe the results of the analysis and comparison of the description of the human metabolic network as given by five frequently used pathway databases. We compared these databases on two levels. On the level of the biology underlying the metabolic network we wanted to see if different descriptions agree on which reactions take place in human and which genes are involved. It is to be expected that to some extent the descriptions differ since the networks were built and curated in different ways and by different research groups (Table 1). On the other hand, these databases aim to represent the metabolic capabilities of the same organism and one would therefore expect that at least the core of the metabolic network described by these databases would be similar, but we show that this is not the case. So far, however, the extent of the differences and the explanations for the differences had not been systematically analyzed for the human metabolic network, likely because the differences were assumed to be small. The results of our comparison are valuable in their own right but also provide further insight in the reasons why these databases differ in the first place. The second level on which we compared the pathway databases is the representation of the metabolic network in a digital format. Not only differences in content may influence the results of analyses, but also the varying definitions that are used by the databases to, for instance, define a pathway (Green and Karp, 2006). A detailed understanding of how a pathway database represents knowledge is, therefore, also important. These two comparisons
provide useful insights in the road ahead to integrate the knowledge contained in the pathway databases into a single metabolic network and the challenges to be met.

**From pathway database experts…**

In *Chapter 2* we describe the results of the comprehensive and systematic analysis of the descriptions of the human metabolic network as found in five of the major databases, *i.e.*, EHMN (Hao et al., 2010), *H. sapiens* Recon 1 (Duarte et al., 2007), HumanCyc (Romero et al., 2004), and the metabolic subsets of KEGG and Reactome. Reasons why the comparison of the databases is not as trivial as it may seem include the widely different ways in which data is stored and organized, and the different ways in which it needs to be retrieved. Another reason is the difficulty of establishing whether databases refer to the same metabolite and consequently the same reaction. We have successfully addressed these challenges. The results of our comparison show a surprisingly limited consensus, even for core metabolic processes like amino acid metabolism. Moreover, the databases differ in the breadth and depth of their coverage of the human metabolic network emphasizing the importance of integration efforts to further refine its description.

In *Chapter 3* we discuss how the five pathway databases solve the challenge of capturing the knowledge on human metabolism in a digital format. The choices made in how to represent knowledge affect the ability of a pathway database to capture the biological complexity of human metabolism. It depends on the application at hand which aspects of the metabolic network are important and to what detail they need to be represented. The differences in representation we identified are of interest for (future) database developers, knowledge curators, and domain experts and can help to further improve knowledge representation.

**… to biologists**

Gathering all up-to-date knowledge on metabolism to build an accurate metabolic network is a huge challenge as the relevant literature is extensive. This is further complicated by the changing nomenclature of enzymes and metabolites in the course of time. Moreover, not for every piece of the metabolic network conclusive evidence is available and some parts might still be subject to controversy. *Chapter 4 and 5* illustrate that the different views on the same biological system offered by the databases can reveal both controversial and complementary biological knowledge. By exploiting these different views we can further improve the description of the (human) metabolic network. These chapters also underline the importance of the involvement of experts on metabolism.
The aim of Chapter 4 is to increase the awareness of the scientific community of the existing differences and biological inaccuracies within the descriptions provided by pathway databases, and to convince experts to help resolve them. For this purpose, we use one of the most well-known pathways, the TCA cycle, as an example. We show that the lack of consensus between pathway databases can partly be explained by an inaccurate description of the knowledge found in scientific literature. We compared the descriptions of the TCA cycle as given by ten databases. None of these were entirely consistent with the literature. Based on the ten descriptions, additional literature research and the knowledge of two experts in the field of metabolism we propose an improved description of the TCA cycle. This proved to be quite a time-consuming challenge, even for this relatively small pathway. First of all, it is challenging to oversee the large volume of articles related to the TCA cycle from 1937 onwards and to cope with the changing nomenclature for enzymes and metabolites. Moreover, the biochemistry behind the TCA cycle turned out to be not as clear cut as one might expect and active involvement of experts proved to be crucial to resolve the conflicting information in the ten databases.

In Chapter 5 we present the web application 'Consensus and Conflict Cards' (C2Cards) that allows experts to more easily identify the differences between metabolic pathway databases mentioned above. Case studies illustrate that the concise overview a C2Card provides can reveal conflicting information that requires additional biochemical experiments to be resolved. Although built for *H. sapiens*, C2Cards can easily be constructed for other organisms as well. Identifying conflicts is essential for ongoing efforts to reconcile the various descriptions of the metabolic network available for a particular organism. The examples of conflicting information uncovered by the C2Cards application also illustrate the advantage of combining descriptions and the importance of going back to the literature.

We conclude by discussing the road ahead in further refining the description of the (human) metabolic network (Chapter 6).