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

Stobbe, M.D.

Publication date
2012

Chapter 6

General discussion

"Not all those who wander are lost."

J.R.R. Tolkien (The Fellowship of the Ring)

The interest in metabolism as a research topic is going through a marked revival (DeBerardinis and Thompson, 2012) as it has become clear that several of the most prevalent diseases, such as cancer, cardiovascular disease, diabetes, and obesity have a strong metabolic component. Cancer research groups, for example, are exploring the possibilities to develop drugs targeting the metabolic pathways involved (Hanahan and Weinberg, 2011). How to capture the increasing amount of knowledge on metabolism has become a firmly established topic of research in the relatively short history of bioinformatics and has led to the development of a multitude of metabolic pathway databases (see also: www.pathguide.org). The use of these databases in various types of analysis has become a common mainstay and examples of successful applications are plentiful. It remains, however, a challenge to gather all knowledge on metabolism and keep up with new discoveries. Moreover, due to the complexity of these networks, it is still a challenge to capture every detail of the metabolic network in a digital format that is suitable for a wide range of computational analyses. An accurate and complete description of metabolism is crucial to reach the ultimate goal of constructing *in silico* models that are capable of generating experimentally verifiable hypotheses, such as potential drug targets, or to simulate the effect of network perturbations such as loss of function. The systematic analyses described in this thesis provide an overview of the current status of human metabolic pathway databases regarding (i) how well they agree on the description of the metabolic network, (ii) how the knowledge is represented *in silico*. We further uncovered some of the obstacles that need to be overcome to further refine the description of the human metabolic network and keep it in sync with new discoveries.

Differences in content

Given the extensive research efforts in the field of metabolism, in past and present, one would expect that pathway databases agree at least on the core metabolic processes, like carbohydrate, nucleotide, and amino acid metabolism. We indeed confirmed that for these core processes, a higher level of agreement exists between the five human pathway databases than for other parts of the network. However, the consensus for this core is still only 4% of around 3,700 reactions that these databases jointly contain. We identified several other explanations for the differences in content (*Chapters 2 and 4*). First of all, differences are caused by disagreements on the biology underlying the metabolic network, possibly because there is controversy in literature as well. Our analysis of the TCA cycle also showed that the descriptions are not always in agreement with literature (*Chapter 4*). Secondly, another important explanation for the differences in content is that the databases differ in the breadth

and depth of their coverage of the metabolic network (*Chapter 2*). For example, lipid metabolism is described in greater detail in EHMN than in the other four databases. In general, each database has a particular focus and its curators have specific fields of expertise. Furthermore, each database is work in progress, which is to be expected given that it is a time-consuming challenge for curators to cover the huge volume of articles. This is a daunting task even for a single pathway (*Chapter 4*). Thirdly, the comparison is hampered by the difficulty of relating metabolites between the different databases and, consequently, of establishing whether the databases describe the same reaction (Chindelevitch *et al*, 2012; Herrgård *et al*, 2008; Radrich *et al*, 2010). However, we have shown that this issue certainly does not explain all observed differences. In the comparison of the core metabolic processes the problem of matching metabolites is less pronounced, but the consensus is still small (*Chapter 2*). Finally, we revealed several conceptual differences, which partly cloud the true disagreements on the underlying biology of the metabolic network. Examples include the number of steps in which a process is described and the use of generic substrates, like 'an amino acid', in reactions versus describing every specific instance (*Chapters 2 and 3*).

In our comparison we focused on reactions, EC numbers and genes as the main components of the metabolic network. There are, however, even more aspects to consider: whether there is agreement on the direction of the reactions, on the compartments in which the reactions take place and whether a catalyst is a complex or not, and so forth. Even when we arrive at a consensus network for all these aspects, further details still have to be worked out. Ott and Vriend (2006) have shown that in KEGG mistakes are made in describing the structure of metabolites. This type of information is important for drug development. Furthermore, the metabolic network differs per tissue and even per cell type, while the five databases we analyzed all aim to describe what is referred to as the global human metabolic network. As argued by Khatri *et al* (2012), tissue information is also essential to improve the accuracy and relevance of pathway analyses. There are already some examples of tissue-specific networks deduced from a global reconstruction, such a HepatoNet1 (Gille *et al*, 2010) for liver metabolism. Agren *et al* (2012) recently published 69 cell type specific models of the human metabolic network, as a first step towards a Human Metabolic Atlas. This resource could in the future be used in the field of personalized medicine to enable a systems-level approach for analyzing patient data, such as gene expression profiles and metabolomics measurements.

Differences in knowledge representation

In our first comparison of the databases, we already observed several differences in how databases represent knowledge on metabolism (*Chapter 2*). Further analysis showed that widely different choices were made by the five databases in how and to what detail to represent the network in a structured way (*Chapter 3*). The choices made are often determined by the intended application domain of a database. For instance, *H. sapiens* Recon 1 is geared towards serving as a basis for mathematical models. For this purpose, it is important to accurately describe gene-protein-reaction relations and compartmentalization, and to ensure that the network is charge and mass balanced. KEGG chooses not to represent these aspects of the metabolic network and puts more emphasis on functional hierarchies of the components of the metabolic network and providing context for the analysis and interpretation of high-throughput data. The different requirements researchers have, may be part of the reason why so many pathway databases have been developed.

Our analysis also revealed that not every detail of the metabolic network can yet be captured in a structured way. One example is the representation of fatty acid beta oxidation. Different solutions were chosen by the databases, but none captures the complete process for all fatty acids. Furthermore, unstructured text fields, which cannot be easily interpreted by computer programs, often contain additional information such as the tissue-specificity of enzymes, which could be used to (automatically) derive the metabolic network for a particular tissue. Data provenance, indicating the type of evidence supporting a piece of information, can also be improved in most databases. Information on evidence is important as it can be used to guide further experiments and allows users to retrieve only that part of the network for which there is a high degree of confidence. Only HumanCyc and *H. sapiens* Recon 1 provide the type of supporting evidence and the level of confidence for a piece of knowledge. In addition, also complete lack of knowledge needs to be indicated explicitly. For example, for the 'missing gene' problem (*Chapter 1*) it is important to know whether the catalyst of a reaction is really unknown or that the reaction takes place spontaneously. Only in HumanCyc this difference is explicitly annotated. Finally, the more elaborate a data model is, the more of a challenge it will be to acquire all necessary details. This is problematic in practice as describing a metabolic process in full detail is a very time-consuming process and requires extensive knowledge that may not even be available yet. At the same time, there is not always a clear cut answer to the question which level of detail is required to be able to perform a wide range of possible computational analyses.

Integration of databases

The results of the comparisons outlined above illustrate that differences between the databases are large both with respect to content and their representation of the network. It is therefore advisable that users carefully weigh their decision when selecting one of the databases, as this choice may affect data analysis results (Lee *et al*, 2008; Zelezniak *et al*, 2010). If possible, users should compare and contrast the outcome of their analysis using different databases to ensure robustness of the results. In retrospect, for our initial quest to find candidates for missing genes, the best option might have been to apply the algorithm to the network of each database and combine the results, ranking the candidates predicted by multiple networks higher. This would, however, have been a far from optimal solution. Instead, having a single, complete, and accurate description of the human metabolic network is to be preferred.

Integration of the multiple descriptions of the (human) metabolic network and, importantly, the reconciliation of the differences between them will lead to a more complete and more accurate description. Moreover, by integrating the individual databases we can profit to the fullest extent from all the knowledge, time, effort, and money that has already been put into these pathway databases. The same is true for other types of networks, including signaling and gene regulatory networks. Also for these networks, comparisons have shown that large differences exist between the various databases available (Bauer-Mehren *et al*, 2009; Kirouac *et al*, 2012). Similar issues will play a role when integrating these types of network, including problems in matching of network components and differences in representation. Moreover, the various cellular processes are not isolated, but are inherently intertwined. Therefore, also integration of databases describing *different* cellular processes is necessary. Integration on this level is further complicated by the heterogeneity of the data.

We now discuss three approaches to integrate multiple metabolic network descriptions, *i.e.*, automatic, semi-automatic, and manual integration, and the strengths and weaknesses of each of these approaches.

Automatic integration

A fully automatic integration of the multiple metabolic pathway databases would clearly be the fastest approach. However, it is not possible to integrate them in a fully automated way (*Chapter 2*), even though this is commonly assumed to be the case. From a technical perspective, automatic integration is already quite challenging and with respect to reaching consensus on the underlying biology it is virtually impossible. From a technical perspective there are three main challenges. Firstly,

retrieving the content of databases is cumbersome to automate. Each database requires a different approach that may even have to be adapted with each new release of a database, since underlying data models are subject to change. Some databases offer an Application Programming Interface (API), which should shield the user from such changes. However, also the API may be subject to change between subsequent releases. In addition, APIs do not always provide access to the entire content of a pathway database. Secondly, databases use different representations and definitions. The different pathway definitions used by each database, for example, make it impossible to integrate networks in a modular way, *i.e.*, per pathway. It is important to realize that even the smallest difference in terminology and the definition of a concept needs to be accounted for when integrating databases. Humans may easily tell when different terms or concepts are equivalent, but computers need to be programmed to do so. The results of our comparison described in *Chapter 3* provide guidance on how to translate the different representations in a single format. Thirdly, there is a lack of a common ground to compare metabolites. As yet, the problem of matching metabolites has not been resolved by naming standards (*e.g.*, IUPAC) and small molecule databases (*e.g.*, ChEBI) that aim to provide an unambiguous way to specify metabolites (*Chapter 2*). It has been suggested to use identifiers that depend on the structure of a metabolite, such as SMILES (Weininger, 1988) and InChI (McNaught, 2006), instead of identifiers from a particular metabolite database. However, in this case a difference in the level of detail with which a structure is described may prevent a straightforward match. The same holds for differences in protonation state. The question then is to what detail the structure of the metabolites needs to be the same to consider them a match.

Standard representation formats for molecular pathways, *e.g.*, BioPAX (Demir *et al*, 2010) and SBML (Hucka *et al*, 2003), have been proposed to facilitate the information exchange between different resources (Strömbäck and Lambrix, 2005). We investigated the use of BioPAX for our comparison, but the BioPAX files turned out to be insufficient as information about genes encoding for the metabolic enzymes was not represented. BioPAX also did not resolve the issues related to matching of metabolites. Furthermore, although BioPAX reduces differences in terminology, most of the conceptual differences between the databases that prevent their integration remain (*Chapter 3*). For example, differences between databases in the representation of gene-protein-reaction relationships are still present in the BioPAX files. Consequently, even if all databases offer their data in the same exchange format - which is currently not the case - we would still need tailored scripts to get the data needed for our comparison from the different databases. In summary, curators will

need to adhere to strict guidelines to make the BioPAX files more easily comparable.

The representation differences outlined above complicate an automatic approach for determining whether databases do not agree on the underlying biology or only made a different choice on how to represent the biology. For example, a difference in the number of steps in which a process is described could either point to an alternative route or to a difference in representation. Furthermore, the question remains how to combine the networks in an automated fashion. Simply taking their union will neither resolve conflicting information nor filter out erroneous information. Only including those parts on which the majority of the pathway databases agree is a better approach. However, this does not take into account that databases are not independent as they often share the same knowledge resources and may also have copied data from each other. Consequently, even if the majority agrees on a piece of knowledge it may still be an error that has been propagated through the databases. Moreover, it cannot be excluded that even the majority can be wrong and a single database with an opposing statement may be right. Manual curation is, therefore, crucial if one wants to combine all knowledge captured by these databases and to decide what is correct and what is not.

Semi-automatic integration

Algorithms have recently been developed to integrate two networks in a semi-automatic manner to overcome some of the difficulties of a fully automated approach. One example is MetaMerge, which starts by matching metabolites and reactions (Chindelevitch *et al*, 2012). Next, users are asked to confirm the matched reactions and their metabolites. This manual step can be skipped, but this will give less reliable results. The reactions that match exactly form a core set, which is expanded by adding reactions of which almost all metabolites match. These new reaction matches are again shown to the user for approval. These two steps are repeated until no reactions are a good enough match. Next, the core network is complemented with reactions that could not be matched. In the resulting merged description conflicting reactions may have been included, which also need to be resolved manually. Another example of a semi-automatic integration procedure is the algorithm proposed by Radrich *et al* (2010), which they used to integrate two descriptions of the metabolic network of *A. thaliana*. The algorithm results in three merged descriptions with a decreasing degree of confidence. The core description contains only reactions that are found in both networks with a high degree of confidence. In the intermediate description reactions are included that are likely to match. The third level simply contains all remaining reactions from the original two

descriptions. Also in this algorithm, part of the metabolite matches has to be checked manually. Furthermore, conflicting information on the second and third level still needs to be resolved manually. Due to the inability to match the metabolites automatically, Radrich *et al* could not integrate two other descriptions that are available for *A. thaliana*. Note that in both approaches when integrating more than two network descriptions, the size of the initial core description will be significantly smaller (*Chapter 2*), which limits its utility.

Manual integration

Efforts are ongoing in the form of reconstruction annotation jamborees to manually integrate the different descriptions of the metabolic network. In a jamboree, experts from multiple disciplines, including biochemistry, molecular biology and systems biology, come and work together on refining the description of the metabolic network of a particular organism. Jamborees have been held for multiple organisms already, including human (Thiele et al, submitted), and have also resulted in consensus networks (Herrgård *et al*, 2008; Thiele *et al*, 2011). Our C_2Cards application can assist in such an endeavor by bringing the differences between descriptions of the metabolic network of the same organism to the attention of experts via concise overviews (*Chapter 5*). As we have shown, this may even lead to the conclusion that further biochemical characterization experiments are required. C_2Cards provides a good starting point to construct a consensus network. A manual approach will require the commitment of a large group of experts from various research fields.

In summary, although the (semi-) automatic approaches proposed are likely to speed up the integration process, they address only part of the challenges we discussed above. It will require more than only technical solutions to integrate the multitude of available descriptions of the (human) metabolic network. Manual curation will remain necessary to resolve the conflicts and filter out erroneous data. This is, however, a huge undertaking. To reduce the manual effort required adhering to standards or guidelines, like MIRIAM (Minimal Information Required In the Annotation of Models), should be a prerequisite for (pathway) databases. The standards need to be followed to the letter and coexistence of multiple interpretations as we observed for BioPAX (*Chapter 3)* should be prevented. Furthermore, it is crucial that the metabolites, proteins, and genes are unambiguously identified. The effort that all this requires will be worthwhile in the end, as it will prevent the knowledge to be lost to the community and allows other researchers to build upon the already discovered facts (Le Novère *et al*, 2005).

The road ahead: keeping up with new discoveries

The extent of our knowledge on cellular processes for different organisms varies, but in general it is still far from complete and new pieces of the puzzle continue to be discovered. Even for a classical metabolic pathway like the TCA cycle, discussion about this small but crucial part of metabolism continues (*Chapter 4*). Also in the latest version of 'Biochemistry' (Berg *et al*, 2012), one of the most popular student text books on the subject, a change was made to the description of this pathway. The importance of an accurate description of the cellular networks has been recognized by many research groups. Pathway databases are still being improved, traditionally by only few researchers (Finn *et al*, 2012). For some organisms, multiple research groups even work independently on improving their own description of the metabolic network. Maintaining and refining these databases requires not only a continuing commitment of the research group, but also long-term funding. Similarly, initiatives like reconstruction annotation jamborees cannot just be a one-time effort and need to be repeated regularly. Unfortunately, it is still a major challenge to acquire funding for setting up and maintaining public databases. For precisely this reason KEGG, one of the most widely used pathway databases, had to switch to paid subscriptions to their FTP site (http://www.genome.jp/kegg/docs/plea.html). Interestingly, performing the experiments to acquire data is much more expensive yet easier to fund. Failing to make the newly discovered facts broadly available through a database may result in loss of the knowledge gained. To quote Amos Bairoch: "*It's quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in often badly written text and then spend some more millions trying to second guess what the authors really did and found.*" (Bairoch, 2009). The key to keeping up with new discoveries will be to stimulate active contributions from the scientific community or what is referred to as 'social engineering' (Kitano *et al*, 2011).

Social engineering

To complement the curation done by small groups of curators or during jamborees, it has been suggested to mobilize a much larger part of the life sciences community to get involved. One of the strengths of a social engineering approach is that by directly involving the researchers that actually discovered a new scientific fact, misinterpretation of the article could be avoided. Furthermore, once a critical mass of people with the relevant expertise is reached, the curation process will be accelerated and conflicting information more easily resolved.

Various initiatives already exist in which the scientific community as a whole can assist in curating existing pathway databases. WikiPathways is a typical example (Pico *et al*, 2008), which follows the same strategy as Wikipedia. Everyone can contribute to this database by either improving the pathways provided by the database or adding new pathways using the embedded pathway editor PathVisio (van Iersel *et al*, 2008). The content of Reactome is also available in this format via a dedicated WikiPathways portal. In this way, improvements to pathways of the centralized database of Reactome can be proposed. We contributed to WikiPathways ourselves by improving the existing description of the TCA cycle (*Chapter 4*). Through its availability in WikiPathways, our description based on the knowledge of ten pathway databases, literature and expert knowledge can easily be adapted if new facts are discovered.

Kitano *et al*. (2011) proposed an open-flow model of knowledge aggregation, whereby both users and multiple pathway databases contribute their knowledge to a centralized forum. Refinements are fed back into the participating pathway databases. Our C_2Cards application would fit in such a model. The C_2Cards application could serve as the central forum in which the various available pathway databases are combined including a current state-of-the-art consensus network, such as Recon 2 for human (Thiele *et al*, submitted). The concise overviews provided by a C_2Card bring the differences between the consensus network and other descriptions to the attention of the scientific community (*Chapter 5*). Experts could then resolve these differences and add to the corresponding C_2Card a nanopublication as supporting evidence. A nanopublication is a traceable author statement, consisting out 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 (Groth *et al*, 2010). Based on the contributions of experts a team of curators will then decide to incorporate the necessary changes in the consensus network, if enough evidence supports this claim. Currently, it is already possible to add comments to a C_2Card, enabling a first discussion on the inconsistencies observed. Furthermore, we are planning to add the option to automatically alert the curators if there are updated or additional C_2Cards.

Lowering the barrier to contribute

For social engineering to be successful, a large community needs to be willing to spend the additional time and effort required. They may, however, feel that they do not directly benefit from this, although at the very least it provides an additional way to advertise one's research (Finn *et al*, 2012). To entice experts to participate,

contributions should be clearly acknowledged and the threshold to contribute should be kept low. In the C_2Cards application, for example, curation is done at the level of only a single reaction or the metabolic functions of a single gene product (*Chapter 5*). Furthermore, by requiring experts to contribute in the form of nanopublications, as mentioned above, their contribution will become traceable and citable (Mons *et al*, 2011). Journals could play a role by requiring authors to contribute their results to initiatives such as WikiPathways. Several publishers already recommend authors to deposit, for example, their computational model of a biological process into the BioModels Database, a repository for sharing peer-reviewed and published mathematical models (Li *et al*, 2010). In a way, this is similar to the current requirements to deposit microarray and sequencing data in public databases.

Another approach would be to let authors semantically annotate their own articles. This would enable the automated retrieval of knowledge that can be used to further refine the description of cellular processes (Jensen and Bork, 2010) and significantly reduce the workload for curators of pathway databases. To assist authors and to reduce the manual effort in annotating an article, well-designed tools are required. The question is though to what extent the authors themselves are the best candidates to do this, as it can be quite challenging to correctly annotate an article in a systematic way (Jensen and Bork, 2010). Again, a community approach in which readers correct and enhance the annotation may be beneficial.

The issue remains that for funding agencies and tenure track committees mainly the number of publications and their citation index truly counts. Appropriate recognition for other types of contributions to science will require a cultural shift in the scientific arena to a situation in which contributions are not only measured in terms of publications. The Scholar Factor, proposed by Bourne and Fink, is an example of a metric that takes into account the number of entries you have made in a public database (Bourne and Fink, 2008). In this way all scientific contributions of a researcher are acknowledged instead of only being based on the number of articles.

Concluding remarks

We would like to stress the importance of broad community efforts for unraveling the complete physiology of an organism. Improving currently available databases is a continuous effort, both with respect to their content and their data model. This, however, should not be used as a reason to justify the development of yet another database, but should instead encourage collaborations to further improve existing resources. The solution is not to continue to build more and more databases or to

design new exchange formats, but to convince the community to contribute to existing initiatives, to stick to the exchange formats, and to lower the threshold for doing so. Ultimately, by joining forces we will be more capable of eliciting knowledge from the huge amounts of data being generated and constructing an accurate model of the cellular processes taking place in human and many other organisms.