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

Stobbe, M.D.

Publication date
2012

Summary

Metabolic processes are constantly taking place in our body no matter whether we eat, sleep or exercise. Food is broken down to generate energy, for example, to enable our muscles to contract and to synthesize building blocks needed to maintain our body's cells. In the 19th century researchers first realized that metabolism can be viewed as a network of connected biochemical reactions. Enzymes catalyze these reactions, that is, they enable them and accelerate the rate at which they take place. Several of the most prevalent diseases in modern society, including diabetes, cardiovascular disease and obesity involve disruptions of metabolic processes. In patients with diabetes mellitus, for example, this results in an excessive amount of glucose in the blood. Metabolism also plays an important role in cancer, as a tumor requires energy to grow.

The study of metabolism has a long history, with the first scientific article on a metabolic process dating back to the 17th century[1]. Currently, nearly six million articles on metabolism have been published according to Medline, the largest repository of biomedical articles. New insights continue to be published, which all provide pieces of the puzzle about the mechanisms of metabolic processes in human and many other organisms. We can, however, not fully understand metabolism if we only study the individual parts of the network of reactions. This would be similar to trying to understand the principles of powered flight and working details of a modern aircraft by only considering the components of the airplane laid out on a hanger floor[2]. In the case of metabolism, we therefore need to know how the reactions and the enzymes catalyzing them fit together and act as a whole. To this end databases have been developed to collect and organize the current knowledge on metabolism scattered across a multitude of scientific articles. These databases, referred to as 'metabolic pathway databases', are like a digital encyclopedia and can often be freely accessed via a standard web browser. An important long-term goal in the field of bioinformatics is to use the knowledge gathered in such databases to build an accurate computer model of (human) metabolism. Such a model could be used to run simulations to determine the effect of disruptions of metabolic processes in a certain disease and predict possible drug targets for treating the disease.

Gathering all knowledge on metabolism to build an accurate model and keeping track of new discoveries is a huge challenge. The literature on metabolism is extensive while at the same time not for every piece of the (human) metabolic network conclusive evidence is available. A second challenge is that the metabolic

[1] Sanctorius S (1614) Ars de statica medicina.
[2] Vastrik I *et al* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biology* **8**: R39.

network needs to be represented in an electronic format that a computer is able to understand and work with. Notwithstanding these challenges, various initiatives have already resulted in more than ten human metabolic pathway databases. One would expect these databases to contain largely the same information. In this thesis we show that the contrary is true. The differences between the databases are extensive, which may influence computational analyses based on these databases. We also propose ways to resolve these differences and arrive at a more complete and more accurate description of human metabolism.

We compared five often used human metabolic pathway databases and showed that of the nearly 7,000 reactions the databases jointly contain only 199 reactions are common to all five databases (*Chapter 2*). We also zoomed in on a single well-known metabolic process, the tricarboxylic acid (TCA) cycle, which serves as an example in nearly every biology and chemistry curriculum. This process plays a key role in generating energy. The entire process was described for the first time in 1937 by Hans Krebs[3], for which he was awarded the Nobel Prize. Our comparison showed that even for this well-studied process there is considerable disagreement between the five databases. We identified several explanations for the lack of consensus between the five descriptions of the metabolic network. One important explanation is that the databases complement each other, *i.e.*, to some extent they provide different pieces of the puzzle of the complete description of human metabolism. To profit from all the time, effort and also the money that have already been put into the development of these different databases, we should strive towards integrating the knowledge gathered. In this way we can further improve the digital description of human metabolism.

Integrating the knowledge contained in the various databases is, however, easier said than done. The databases have made widely different choices on how to represent the metabolic network in a digital format (*Chapter 3*). Several formats have been proposed to standardize how knowledge should be stored and enable integration and exchange of data between different databases. However, we show in our analysis that even such a standardized format is not enough to resolve all differences in representation of the metabolic network. Another issue that makes the comparison and integration of databases difficult is that different names are used to refer to the same biochemical compounds and enzymes. Multiple naming standards have been proposed in the literature, but these are not always used or different conventions are

[3] Krebs HA, Johnson WA (1937) The role of citric acid in intermediate metabolism in animal tissues. *Enzymologia* **4**: 148-156.

used by the databases. Another important challenge is that if we would simply combine the different descriptions of human metabolism provided by the databases, we do not resolve conflicting information nor filter out mistakes. For these reasons a completely automated integration is impossible.

For the TCA cycle we manually integrated the descriptions as given by ten different human metabolic pathway databases (*Chapter 4*). We identified and resolved the differences between the descriptions using literature and the knowledge of two experts in the field of metabolism. In this way, we were able to propose an improved description of the TCA cycle. This endeavor illustrates the importance of going back to the biology as described in the literature and the crucial role of experts in resolving the differences between the databases. It is, however, quite a time-consuming endeavor, even for a metabolic process with a relatively small number of steps. To do so for all metabolic processes will require a combined effort of a large group of experts. Moreover, it needs to be an ongoing process as new facts on metabolism continue to be discovered.

We built a web application called 'Consensus and Conflict Cards' (C_2Cards) that can be used in the quest to further improve the description of the human metabolic network (*Chapter 5*). The application provides concise and easy-to-retrieve overviews of the differences between five metabolic pathway databases. Experts can use the C_2Cards to try to resolve the disagreements on, for example, which enzyme catalyzes a specific reaction. By showing the different views the databases have on a piece of the metabolic network both controversial as well as complementary biological knowledge may be revealed. A number of case studies illustrate that in some cases additional biochemical experiments are required to resolve the differences observed. By making the C_2Card application available as a webpage the threshold for experts to use the tool is low.

In conclusion, the results described in this thesis give a clear signal to the scientific community that it is of great importance to integrate the knowledge captured by multiple databases. Our analyses and the C_2Cards tool provide a starting point for such an endeavor. The contribution and continuing commitment of a broad community of experts will be necessary to bring us closer to reaching the ultimate goal of a highly accurate model of (human) metabolism.