31 PathDB: a second generation metabolic database

P. Mendes, D.L. Bulmore, A.D. Farmer, P.A. Steadman, M.E. Waugh and S.T. Wlodek

National Center for Genome Resources 1800A Old Pecos Trail, Santa Fé, NM 87505, USA.

Introduction

Digital computers were first used in biochemistry circa 1960 [1]. However, during the period since then they have been used mostly in numerical applications such as parameter estimation or simulation of dynamics. It was only with the advent of sequence databases in the 1980s that metabolic databases started appearing. Most of these have been essentially electronic encyclopedias, with little more features than what could be found in a printed book. Here we introduce PathDB, a metabolic database that combines storage of specific and detailed information about metabolism with powerful software for data query, navigation and visualization. This combination makes PathDB more than just a simple electronic encyclopedia, enabling discovery of facts that were already known but have been hidden in the complexity of metabolic data.

Databases are computerized archives of information that usually have powerful means of indexing data. Databases can be organized in various ways. Some simply store information on disk files relying on the computer operating system for management (a popular variant is to keep the whole database in a single large text file where special characters or strings are used to separate the records). Relational databases organize data in tables where each row represents one single entity. Relational database management systems (DBMS) are currently the most efficient to manage large amounts of data, partly due to the powerful Structured Query Language (SQL), used to query them. Object-oriented databases are becoming fashionable. Such databases organize data in a hierarchy of classes that can inherit properties from each other. Although object-oriented databases are arguably the easiest to design, their performance for large volumes of data is still

far from relational databases.

ENZYME [2] is a database that contains the classification of enzymes by the Nomenclature Committee of the IUBMB. This has become the most used database of metabolism, perhaps because annotators of genomes like to classify enzymatic gene products according to their EC number. Apart from allowing quick searches of enzymes by EC number or vice-versa, ENZYME provides little more benefit. The pioneering work of Selkov and co-workers since the late 1980s [3] resulted in a collection of electronic information about enzymes and pathways derived from the published literature. These data are now available on the Internet in EMP [4] and MPW [5]. EMP contains quite a lot of detailed information about specific enzymes, contrasting with ENZYME that only contains classes of enzymes. EMP is a relational database that can be queried using SQL but it lacks a user-friendly mechanism for making complex queries and has limited visualization capabilities. KEGG [6] is a database of metabolic pathways that contains nice diagrams of pathways. These are static images that are updated when new steps are added to the pathway. Unlike EMP, KEGG has very little detail about the enzymes, basically not much more than the generic information that is contained in ENZYME. EcoCyc [7] is a database of *E. coli* genes and metabolism currently under control of Pangea Systems Inc. (though still free for academics). This database is based on a frame knowledge representation system, similar to an object-oriented database. Unlike the other metabolic databases, EcoCyc is specific for a single species and has benefited from curation by a domain expert. EcoCyc is original in that its graphical user-interface (GUI) includes automatic layout of metabolic pathways [8]. This is an advanced feature that we think is required if metabolic databases are to be more than simple electronic books. However the benefits of using such technology in EcoCyc are unclear, at least in the academic version, since pathways in this system are static objects and the same diagram is always drawn for any specific one. For the user there is no difference between this or the case where the pathway diagrams are stored as static images, like KEGG. Ochs and co-workers have described [9, 10] their development of a metabolic database based on the relational model. Much in the same line that we take here, they argued that computerized metabolic maps should be useful to uncover known, but not noticed, relations between the data. However, these features are not present in the database that they have made available on the Internet, Pathways+. Finally, a collaboration between the European Bioinformatics Institute and the University of Köln has made public the enzyme data collected since 1987 by Schomburg, which is essentially an electronic form of the “Enzyme Handbook” [11]. It contains an extensive amount of data but does not have any means for making complex queries or any type of graphical visualization.

Each of the databases described above contain at least a feature that makes them unique. Unfortunately there is no single one combining all the desirable features. Furthermore all are rather limited in terms of the concept of pathways they adopt: invariably they describe pathways as well defined sequences of steps,

usually the same as the textbook pathways. And none provide a user-friendly interface to complex queries. As an example, the simple query "Find all reactions that take ATP as a substrate and AMP as a product" cannot be executed by any of the user interfaces provided (with the exception of SQL in EMP, but this is far from a user-friendly).

Here we describe PathDB, a metabolic database developed and hosted at the National Center for Genome Resources. PathDB has been designed to store a wide range of data in very great detail, including kinetic information; locations ranging from sub-cellular to whole organism level; taxonomic information; and thermodynamic properties of reactions. PathDB is a relational database with a non-redundant, hierarchical design. A Query Tool allows construction of complex queries and a Pathway Viewer generates pathway diagrams automatically. PathDB is available on the web at http://www.ncgr.org/software/pathdb .

Design and implementation

At the lowest level PathDB uses a relational database management system (currently Sybase v. 11.9.2). This allows the database to grow without concerns for performance issues and, more importantly, to allow powerful queries to discover relations between the data. On top of this relational database we have constructed an intermediary layer that masks it to an object-oriented view such that the user interface software sees the database as if it was object-oriented. This provides flexibility and insulates most of the software from the particularities of the DBMS used, thus minimizing the changes required if the DBMS was to change. The user-interface for PathDB consists of a suite of programs written in the Java language. This choice was made partly to ensure that the software would work on as many hardware configurations as possible. This software runs under Sun Microsystems' JRE 1.1.7 or above.

The Query Tool (QT) is the front-end for searching the database. It connects directly to a server at NCGR which provides the interface to the Sybase DBMS. The QT allows simple and wild-card queries for each of the basic data types. Results are returned as a list of objects which can then be "transformed" into other types of data. These transformations consist of retrieving objects of different nature but which are related to the original ones. For example, one can query for all compounds whose name starts with the string "Glucos%" (the '%' character is a wild-card meaning "any other characters after this position"). The result would be a list of compounds including Glucose and Glucose 6-phosphate. With that list the user has a choice of several transformations, for example "Reactions in which these compounds take part" or "Pathways in which these compounds participate", among others. Transformations allow one to navigate the complex web of relationships between metabolic entities. The QT also allows one to combine several query results using the set operations union, intersection and difference.

Using this software one can thus ask questions as complex as "from what species is there evidence about enzymes that catalyze reactions involving medicarpin?" or even more.

Pathway Viewer (PV) is a component for graphical visualization of pathways. We have constructed it such that it displays pathways in a manner very similar to what biochemists have been using in publications. Steps are represented by arrows connecting compounds. We allow for some of the co-substrates or co-products to be classified as "secondary" and represented in a smaller font on the side of the step arrow. The "primary" compounds are displayed in a single location in the graph with all steps that produce and consume them connecting to that location. "Secondary" compounds are represented several times, as many as steps in which they take part. PV is capable of laying out pathway diagrams automatically. This has been achieved by adding, where needed, extra nodes representing the step. This transformation makes a pathway become a graph (in fact a special type called a Petri net). Efficient graph layout algorithms exist and we have used a number of these in PV, plus a new one that we developed ourselves (manuscript in preparation) to layout cyclical pathway structures. The user has the possibility of using any of these algorithms to draw the pathway. This important feature allows the user to visualize the pathway in different ways. Some of these may highlight a property of the pathway that would be difficult to discover by inspection in other diagrams. Although PV normally produces pathways arguably as good as many in publications, the user has the possibility of rearranging the diagram by dragging the compounds on the screen. The program takes care of "carrying" the steps connected to the compound being moved with it so that the pathway diagram is never destroyed.

Because some users may not want to download and install a program just to query PathDB a few occasional times we have also developed a simple WWW interface. This is based on a PERL-CGI program that queries the DBMS directly and can be accessed from a Web browser. This also provides a means for adding links in web pages to specific records in PathDB. This interface, though, does not provide the advanced query capabilities that the QT does, but it is a fast way for answering simple queries (like those that can be made to the other databases mentioned earlier).

Discussion

PathDB is the result of our intention to make a database that is more than a simple electronic reference of metabolism. To carry out useful data mining of metabolic information, we believe that a metabolic database must be based on a rich data model allowing for storage of quantitative data, such as kinetics and thermodynamics of reactions, including error estimates if available. Furthermore, the objects represented in the database must be specific single biological entities.

There would be serious limitations if the database only contained classes of enzymes as described in the EC classification. How many alcohol dehydrogenases are there classified with the number 1.1.1.1, and how different are their properties? Another requirement for accurate representation of metabolism is to include transport steps and spontaneous reactions. We have also provided a data model in which every enzyme, transport protein or pathway can be labelled to be located in a specific sub-cellular and/or organ location. Enzymes and transport proteins are further subdivided into subunits which may contain a link to a gene in a sequence database and all of the latter are linked to their own taxonomic species. Naturally, PathDB links data to bibliographic references to the primary literature. This is essential for data validation and for further research.

Given that the definition of pathways is still an open question which can easily generate disagreement among biochemists (where does a pathway begin and end?), we have decided to adopt the simplest concept for PathDB. A pathway is a set of metabolic steps (reaction or transport) connected by common intermediates, without any other requirements. This means that a single step can be seen as a pathway, as well as the whole of the database. More importantly this allows the "textbook pathways" to be represented in the database along with others that may have never been seen before. We are currently developing the software to allow users to search for routes between any two compounds in the database, to find the vicinity of compounds or steps and to visualize them. This is the main reason why we developed Pathway Viewer with automatic layout since the number of pathways contained in even a small subset of the database is enormous (all possible sets of connected reactions). This makes storing predefined pathway diagrams virtually impossible.

A metabolic database is an extremely useful tool to modelers of metabolism. The slowest step of constructing mathematical models of metabolism is chasing the parameter values in the literature. We plan to make PathDB export data in the Gepasi [12] file format.

The only critical issue that we have not yet addressed properly in PathDB is the amount of data stored. For this database to become really successful we think that it should contain a very large proportion of the information existent in the literature. Collecting all those data into electronic format is a costly exercise that requires expert supervision. Currently NCGR is seeking funding for this large scale endeavour. Until then PathDB is focused on plant metabolism, an area that is until now poorly addressed in electronic format. However we stress that there is nothing in the database system specific to plant metabolism and we hope to soon start adding data from other kingdoms of life.

References

1. Chance, B. (1960) Analogue and digital representations of enzyme kinetics, *J. Biol. Chem.* **235**, 2440–2443.

2. Bairoch, A. (1993) The enzyme data-bank, *Nucleic Acids Res.* **21**, 3155–3156.

3. Selkov, E.E., Goryanin, I.I., Kaimatchnikov, N.P., Shevelev, E.L. and Yunus, I.A. (1989) Factographic data bank on enzymes and metabolic pathways *Studia Biophysica* **129**, 155–164.

4. Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr. and Yunus, I. (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database *Nucleic Acids Res.* **24**, 26–8.

5. Selkov, E., Jr., Grechkin, Y., Mikhailova, N. and Selkov, E. (1998) MPW: the Metabolic Pathways Database *Nucleic Acids Res.* **26**, 43–5.

6. Goto, S., Nishioka, T. and Kanehisa, M. (1998) LIGAND: chemical database for enzyme reactions *Bioinformatics* **14**, 591–9.

7. Karp, P.D., Riley, M., Paley, S.M. and Pelligrinitoole, A. (1996) Ecocyc - an encyclopedia of *Escherichia coli* genes and metabolism, *Nucleic Acids Res.* **24**, 32–39.

8. Karp, P.D. and Paley, S. (1994) Automated drawing of metabolic pathways in *Proceedings of the Third International Conference on Bioinformatics and Genome Research* (Lim, H., Cantor, C. and Bobbins, R., eds).

9. Ochs, R.S., Qureschi, A., Sycz, A. and Vorbach, J. (1996) A computerized metabolic map. 2. Relational structure, extended modeling, and a graphical interface., *J. Chem. Inf. Comput. Sci.* **36**, 594–601.

10. Ochs, R.S. and Conrow, K. (1991) A computerized metabolic map, *J. Chem. Inf. Comput. Sci.* **31**, 132–7.

11. Schomburg, D., Salzmann, D. and Stephan, D. (1990-1995) *Enzyme Handbook, Classes 1-6*, Springer, .

12. Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**, 361–363.