An Overview of Data Models for the Analysis of Biochemical Pathways

Yves Deville
Computing Science and Engineering Department
Université catholique de Louvain
Place Saint-Barbe 2
B-1348 Louvain-la-Neuve, Belgium
deville@info.ucl.ac.be

David Gilbert
Bioinformatics Research Centre
Department of Computing Science
University of Glasgow
17 Lilybank Gardens
Glasgow G12 8QQ, Scotland, UK
drg@brc.dcs.gla.ac.uk

Jacques van Helden, Shoshana J. Wodak
Service de Conformation de Macromolécules Biologiques et Bioinformatique
CP263, Université Libre de Bruxelles
Blvd du Triomphe
B-1050 Bruxelles, Belgium
{jvanheld,shosh}@ucmb.ulb.ac.be

12 June 2003

Abstract

Biochemical pathways such as metabolic, regulatory or signal tranduction pathways can be viewed as interconnected processes forming an intricate network of functional and physical interactions between molecular species in the cell. The amount of information available on such pathways for different organisms is increasing very rapidly. This is offering the possibility of performing various analyses on the structure of the full network of pathways for one organism as well as across different organisms, and has therefore generated interest in developing databases for storing and managing this information.

Analysing these networks remains however far from straight forward due to the nature of the databases, which are often heterogeneous, incomplete, or inconsistent. Pathway analysis is hence a challenging problem in systems biology and in bioinformatics.

Various forms of data models have been devised for the analysis of biochemical pathways. This paper presents an overview of the types of models used for this purpose, concentrating on those concerned with the structural aspects of biochemical networks. In particular, we classify the different types of data models found in the literature using a unified framework. In addition we describe how these models

have been used in the analysis of biochemical networks. This enables us to underline the strengths and weaknesses of the different approaches, as well as to highlight relevant future research directions.
Keywords biochemical networks, data analysis, graphs, modelling

1 Introduction

Modelling is an important research area in biology and in bioinformatics. Theorists are designing models in order to investigate a particular hypothesis about biological function that could explain experimental observations. On the other hand, experimentalists are building models in order to store experimental data on biological molecules and processes in databases, and to be able to analyse these data.

An increasing body of data is rapidly becoming available on the network of interactions between genes and proteins for whole organisms. This has prompted many groups to develop models for representing and analysing these networks, giving rise to various forms of biochemical network models, with different objectives.

Database models aim at representing experimental data in order to store them in databases, so as to take advantage of the data storage, management and retrieval facilities offered by database systems. However, database models are often unsuitable for analysing the structure of the stored networks, and thus specialised *data models*, such as graph-based models, have to be designed to facilitate the *analysis of biochemical networks*. These data models are usually implemented as in-memory data structures, containing information extracted from a database. Such models generally contain few quantitative data and are primarily aimed at qualitative analysis. Data models used for analysing networks are thus different and independent from database models. *Computational modelling* [4] is a more theoretical and mathematical approach, where the models aim at explaining biological systems and are generally geared towards the simulation of the dynamic properties of the system. These models usually enable quantitative analyses, and contribute, for example, to metabolic engineering, aimed at the development of methods for improving the metabolic capabilities of industrially relevant micro organisms [64].

In this paper, we concentrate on *data models for the analysis of networks*. The analysis of biochemical networks is part of *systems biology*, and seeks to define theories that explain the complex manner in which the components of a biological system give rise to the behaviour of that system [27].

Biochemical pathways, also called biochemical networks, is a general term that can be applied to several families of pathways. Metabolism can be viewed as a network of chemical reactions catalysed by enzymes, and connected via their substrates and products; a *metabolic pathway* is then a coordinated series of reactions. Metabolic regulation refers to the regulatory actions that operate on the level of the genes and enzymes involved in the pathway, thereby modulating the pathway output; a *regulatory pathway* is a coordinated series of reactions and molecular interactions regulating the expression and/or activity of enzymes

and transporters. Finally, signal transduction is a term describing the transfer of information (called signals) from one cellular location (often the extracellular medium) to another (often the cell nucleus); a *signal transduction pathway* is coordinated series of reactions and interactions realizing a signal transduction. Note that regulatory and signalling pathways are often closely interconnected. A separation between them can therefore be very difficult. This will have consequences on the data models.

Pathway databases hold data on biochemical pathways and their components (e.g. enzymes, substrates, products) and on the correponding interactions and chemical reactions. They are encyclopedic references for pathway information; they can be queried for information retrieval, and can be analysed through computer programs. Some existing databases focus on specific types of interactions : e.g. BRENDA (enzymatic catalysis) [51], DIP (protein-protein interactions) [67], Transfac (protein-DNA interactions [65] , and RegulonDB (protein-DNA interactions) [48]. There are several databases on metabolic pathways, such as KEGG (genes, enzymes, metabolic reactions) [21], EMP (enzymes, pathways) and WIT (metabolic pathway reconstruction) [41], EcoCyc (metabolic pathways, E.coli) and MetaCyc (metabolic pathways of other organisms) [25], aMAZE [60, 59], CSNDB [57], PathDB [49], UM-BBD [13], SHARKdb [44], etc. The BIND database [1] contains information on interactions that take part in signal transduction pathways. An analysis and comparison of these databases can be found in [66]. In most databases the information is represented in a (simple) relational form. The quality of the underlying relational data model is important for the extraction of suitable information for analysing the networks.

Why analyse pathways ? The quantity of available information on biochemical pathways for different organisms is increasing very rapidly. It has now become possible to perform detailed analyses of metabolic pathway structures for entire organisms. However, such analyses face various difficulties. The existing databases are very heterogeneous; data can be incomplete, inconsistent or approximate; our knowledge of the mechanism of gene regulation is today poorly structured and partial. The potential size of the pathways for analysis can also be very large, leading to problems of spatial and temporal computational complexity. This makes pathway analysis a challenging problem in systems biology and in bioinformatics.

Examples of questions that could be solved by pathway analysis are :

- What are the possible paths from compound A to compound B ? How many paths, and how many steps within each path, lead from A to B ?
- What is the distribution of path lengths between 2 compounds ? What is the average path distance between compounds ?
- Give all paths traversing a set of specified compounds or reactions (e.g. given a set of co-regulated genes, find a path or pathway that could be formed with the catalysed reactions).
- Find all genes whose expression is directly or indirectly affected by a given compound.
- Show which pathways may be affected when one or more gene/proteins are turned off or missing.

- Compare biochemical pathways from different organisms and tissues, or at different stages of annotation; highlight common features and differences; predict missing elements.

The design of sophisticated tools for pathway analysis has great potential and practical value for research in various fields such as the search for antimicrobial agents, drug design, human disease analysis, bioremediation, metabolic engineering for bioprocesses and therapeutics, etc.

Objectives and results of our survey Various forms (types) of data models can be used for the analysis of biochemical pathways. We review and classify the different types of data models found in the literature using a unified framework, and describe how these models have been used in the analysis of biochemical pathways. This enables us to underline the strengths and weaknesses of the different approaches, while at the same time pointing to relevant future research directions.

In this overview the actual database models, algorithmic issues and visualization aspects are not covered. Although some simulation models will briefly be presented, computational modelling of biochemical network [4] will not be considered.

Related work There are several works which are complementary to this overview; they cover other classes of models (databases, computational) while we focus on data models for the analysis of networks. For example, in [64], the authors survey the modelling approaches in metabolic engineering; a literature review of modelling and simulation of genetic regulatory systems is described in [8]; an analysis and comparison of existing metabolic pathway databases is proposed in [66].

Structure of the paper Section 2 introduces the concepts of models and analysis. Sections 3 to 5 present different data models based on graphs, and Section 6 describes object-oriented data models. These models are analysed and compared; practical applications of each model are annotated by numerous literature references. Data models for simulation are briefly presented in Section 7. Finally, conclusions ar givenn and some research directions are presented in Section 8.

2 Models and analysis

When the objective is the creation of a data repository, the tool employed is a database. The underlying representational model is a database model (e.g. a relational data model), and simple analysis can be performed directly on the database through some query language, such as SQL. For more sophisticated analyses, specific algorithms are required. Most of the time such algorithms do not rely on a database model; they usually have their own data model (implemented in some data structure), as depicted in Figure 1. Database models are often unsuitable for algorithmic use. Hence, various data models have been designed and used for the analysis of biochemical pathways. Note that there are no fixed boundaries between database queries and specific algorithms, nor between database models and data models or data structures.

There is obviously some relationship between the analysis to be performed and the chosen data model. The choice of a data model drives the possible analyses, and vice versa (see Figure 2). Each model provides its own view of reality, taking into account particular aspects while neglecting others. A given model can thus be suitable for several analyses, and one analysis can be performed on different models.

It is thus impossible to determine which model is *best*; each of the models presented in the following sections has its own advantages and enables specific types of analysis. Models differ either by the chosen view, the coverage (different types of interactions that they can represent), their precision or the granularity (resolution of the basic information: atomic, molecular, supramulecular).

Models will be described using a graph framework. A graph $G(V, E)$ is a mathematical object where V is the set of nodes (vertices), and E is the set of edges, connecting pairs of nodes. An edge is an ordered pair of nodes (directed or oriented graph) or an unordered pair of nodes (undirected graph). Object-oriented models can be seen as a natural extension of graphs, where the nodes are typed, and different relations are defined between specific types of nodes. Objects also allow inheritance.

3 Compound graphs

The objective of employing a compound graph is to model a set of chemical reactions. Nodes are the chemical compounds. Edges between compounds can be defined in two ways. In the approach, an undirected edge connects two compounds if they occur in the same reaction (as substrates or products) [15]. The second approach is more classical; a directed edge connects compound A to compound B if A occurs as a substrate and B as a product in the same reaction.

As an example, let us consider the following simple reaction :

Reaction 1 : glutamate + ATP $\rightarrow$ gamma-glutamyl phosphate + ADP
EC 2.7.2.11, catalysed by gamma-glutamyl kinase

The compound graph of *reaction 1* is shown in Figure 3.

The use of graph theory, and in particular compound graphs, is a well established representation technique in biochemistry and in chemical engineering [47]. Compound graphs have recently been used in [15, 62] for the analysis of topological properties (connectivity, length, statistical properties, ...). The authors stress the small-world character of metabolic networks: their compound graphs are sparse, but much more highly clustered than an equally sparse random graph. Compound graphs are also used in [33].

The equivalent of compound graphs can be defined for signal transduction networks as well as for transcriptional regulation networks. In a transcriptional regulation graph, nodes represent genes, and a directed arc between gene A and gene B means that gene A codes for a transcription factor which regulates gene B.

Objectives	Tools	Models
Data repository	Data Base	DB Model
Simple data analysis	DB query	DB Model
Data analysis	*Algorithms*	*Data model / data structure*

Figure 1: models and analysis

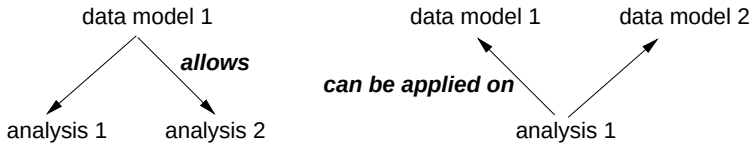

Figure 2: data model vs analysis

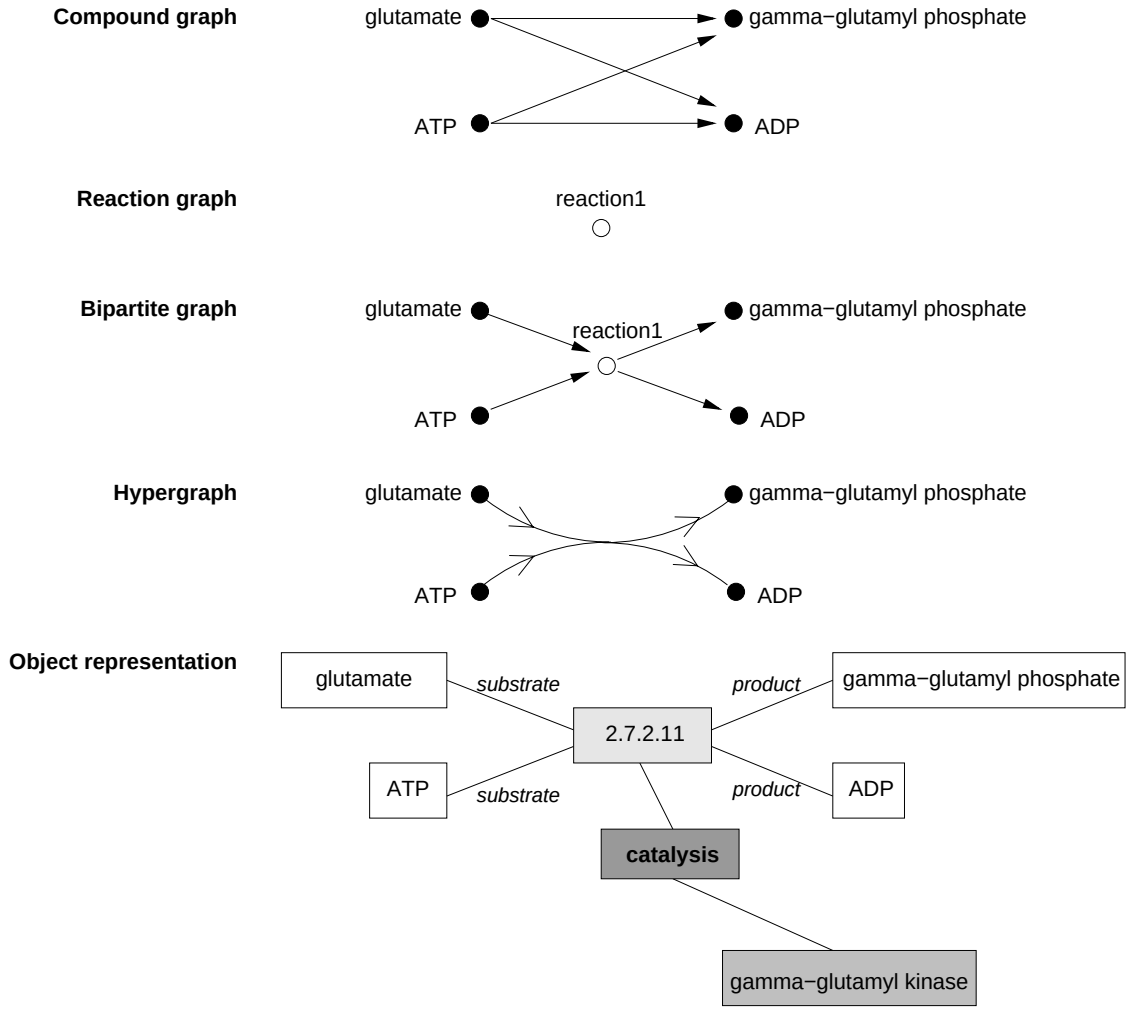

Figure 3: Reaction 1 : glutamate + ATP $\rightarrow$ gamma-glutamyl phosphate + ADP (EC 2.7.2.11, catalysed by gamma-glutamyl kinase

In their recent work [37, 54], Alon and co-authors analyse transcriptional regulation networks of *Escherichia coli* in order to uncover its underlying structural design, by means of the discovery of network motifs. In this work, a network motif is a pattern of interconnections occurring in networks with a significantly higher frequency than what would be expected in random networks. This analysis relies on sophisticated algorithms for the generation random networks that have been applied to other networks in neurobiology, ecology and engineering.

In the signal transduction graph used in [35], nodes are signalling molecules, and an edge represents a process relating two signalling molecules. Such a representation is used for path searching. A similar representation is used in [23].

Although compound graphs or their equivalent can be used to represent and analyse metabolic, regulatory or signalling pathways, this data model cannot combine these different pathways. Such a combination requires for instance a distinction to be made between nodes representing compounds and nodes representing genes, and to distinguish arcs representing a reaction from arcs representing the regulation of some signalling process. Compound graphs also have obvious limitations in their coverage since they represent only reactions within pathways and contain no information about the enzymes catalysing these reactions. The coverage is even more limited for regulatory and signalling pathways because of the large number of different types of interactions that occur in these pathways (assembly, transcriptional regulation, protein-protein interaction, translocation). The descriptive power of compound graph is also very poor because the structure of the reaction is lost in compound graphs. One can no longer distinguish if two substrates or two products are involved in the same reaction. As a consequence, different sets of reactions can lead to the same compound graph [17], thus introducing ambiguity (see Figure 4). The major characteristics of compound graphs are summarized in Figure 5.

4 Reaction graphs

A reaction graph is a dual form of the compound graph. Here, the nodes are the reactions. There is an edge between reactions $R1$ and $R2$ if a compound is both a product of $R1$ and a substrate of reaction $R2$. The graph can be directed or undirected, depending on whether the reactions are considered as reversible or not. It is also possible to extend the definition of an edge by considering edges between two reactions when they share a compound [62]. The reaction graph of our example is reduced to a single node (Figure 3) as it only involves one reaction. A more detailed example is given on the left-hand side of Figure 6.

In [62] reaction and compound graphs are used for the analysis of topological properties of metabolic networks. Reaction graphs are also used in [40] for the detection of functionally related enzyme clusters. Reaction graphs are compared with genome graphs, where the nodes are the genes and edges join adjacent genes. The objective is to detect correlated clusters or local similarities, given a list of corresponding nodes

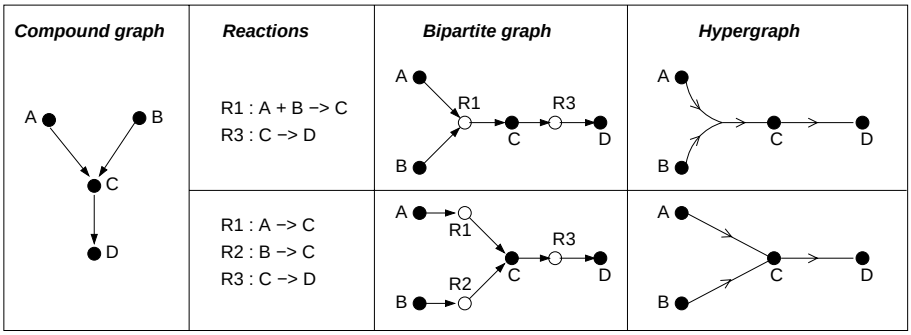

Figure 4: Compound Graph

	Compound graphs	**Reaction graphs**	**Bipartite graphs Hypergraphs**	**OO models**
Precision	ambiguous	ambiguous	non ambiguous (for reactions)	non ambiguous
Coverage	limited	limited	medium	large
Integration of different networks	no	no	limited	yes
Simplicity	++	++	+	+/–

Figure 5: Summary of main characteristics of data models

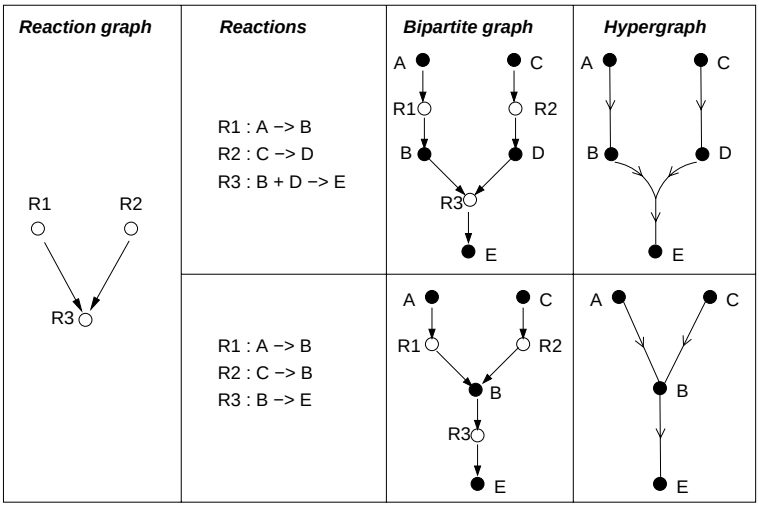

Figure 6: Reaction Graph

from the two graphs. A reaction R corresponds to gene G if the enzyme encoded by G catalyses reaction R. The result is a set of enzymes whose genes have closely related chromosomal positions, and catalysing successive reactions in the metabolic pathway.

Reaction graphs are similar in their limitations to compound graphs. Their coverage and precision are also very limited given that this time, compounds are not represented. It is therefore impossible to determine if products generated by two reactions and consumed as substrate by another reaction are identical or not. As a consequence, different sets of reactions can also lead to the same reaction graph [17], thus introducing ambiguity (see Figure 6). Reaction graphs are mostly limited to a (partial) representation of metabolic pathways. It is therefore inappropriate for modelling different types of networks. The major characteristics of reaction graphs are summarized in Figure 5.

Nonetheless, although compound and reaction graphs only offer a partial and sometimes ambiguous view of biochemical networks, such representations turn out to be sufficient and useful for some simple analyses such as topological and statistical properties, or the discovery of basic patterns. Such representations can also be helpful in some specific applications, such as the detection of functionally related enzyme clusters in [40].

The models described in the next sections extend the above basic graph approaches, and overcome some of their limitations.

5 Bipartite graphs and hypergraphs

In a bipartite graph, there are two classes of nodes, and no edges can relate nodes from the same set. In the context of biochemical networks, there are *compound nodes* and *reaction nodes*; an edge thus necessarily relates a compound node and a reaction node. Edges can be undirected or directed. A directed edge from a compound node to a reaction node denotes a substrate, while an edge from a reaction node to a compound node denotes a product of the reaction. Bipartite graphs can represent reactions without any ambiguity, as illustrated in Figures 4 and 6. The bipartite graph representation of our example is provided in Figure 3.

Instead of bipartite graphs, one can also use directed or undirected hypergraphs, a generalization of compound graph, where a hyperedge relates now a set of substrates to a set of products. An hypergraph can easily be transformed into a bipartite graph, and vice versa. They thus offer equivalent representations. The equivalence between bipartite graphs and hypergraphs is clearly illustrated in Figure 3, where there is a single hyperedge relating the two substrates and the two products of the reaction. Other illustration are provided in Figures 4 and 6. Hypergraphs have been used to model metabolic pathways [32].

In [24] metabolic networks of 43 organisms are modelled as bipartite graphs, permitting a systematic comparative analysis showing that these metabolic pathways have the same topological scaling properties. Bipartite graphs have also been used in [58, 61] for the analysis of metabolic networks stored in KEGG. The

authors analyse global structural properties, and different graph analysis operations, such as path finding are proposed, using classical graph algorithms. A set of pool metabolites (H_2O, O_2, ...) were discarded from the graph, and a specific treatment was applied to prevent traversing reversible reactions from substrate to substrate or from product to product. Pathway reconstruction from a cluster of reactions is also considered. An extension of the path finding operation has been developed in [7], where weighted graphs were used to assign a specific score to each compound, according to its connectivity. This permits the reduction of the probability of traversing pool metabolites, without discarding them from the graph.

Algorithms for pathway synthesis have been developed and applied in [53, 14]. The objective here is to find pathways from initial compounds to final products. The approach, based on [17], uses axioms for feasible reaction pathways, such as *intermediate compounds are entirely produced by previous reactions and completely consumed by subsequent reactions*, and axioms for combinatorial feasible reactions, such as *a reaction is forward or backward, but not both*. The algorithms involve graph manipulation, branch and bound and integer programming. Note that stochoimetric information is handled here.

Although not explicitly described, the underlying data model in [16] is equivalent to bipartite graphs, and is used for comparing metabolic pathways. A global distance between topologically identical pathways is introduced; it is a combination of the individual distance between the substrates, and the distance between enzymes of the reactions (using alignment of the corresponding genes). Bipartite graphs have also been chosen as data model for the hierarchical analysis of dependences in metabolic networks [19]. The representation of compounds and reactions is also necessary for pathway prediction [41, 42, 44]. Given a gene and a metabolic pathway, the objective is to provide some kind of measure of the number of enzymes catalysing the reactions in the pathway which are encoded by the gene.

Another extension of bipartite graphs is proposed in [18] to model signal transduction pathways. In such graphs, each node can itself contain another graph. Interaction edges model interaction between biochemical entities, while decomposition edges reflect the hierarchical structure. Information is represented by Prolog predicates, and path search algorithms are implemented with HiLog programs [6] (Prolog programs with higher order facilities). This extension of graphs to handle hierarchical structures has also been used in [9] to model cellular processes. In that case however, these graphs are bipartite, with state nodes and transition nodes.

Bipartite graphs and hypergraphs offer an unambiguous representation of the reactions and compounds in biochemical networks. The coverage is limited as possible controls of reactions (catalysis, inhibition, ...) cannot be explicitly represented. This simple data model is appropriate when the analysis is limited to reactions and compounds. This however covers many applications such as topological properties, path finding, synthesis and prediction. Without extensions, bipartite graphs cannot simultaneously model metabolic, regulatory and signalling pathways. The major characteristics of reaction graphs are summarized in Figure 5.

6 Object models

More sophisticated data model are required when different aspects of metabolic pathways, such as catalysis and regulatory information, must be covered. Object-oriented models can be seen as a generalization of bipartite graphs, where the nodes are typed, and permit a much more detailed description. Object models also allow inheritance, providing a powerful tool to structure data. Usually the definition of an object-oriented model of a biochemical pathway requires different tasks: the definition of the objects and their attributes, the design of an hierarchy between the different objects, and the definition of all the relations between the objects. When modelling biochemical systems, activities objects should be distinct from those representing the biochemical entities. Attributes of an activity could be qualitative or quantitative.

In Figure 7 we present a simplified object model of metabolic regulation. It is described in UML, a widely used modelling language [3]. Only the classes (without attributes) and relations are shown. Our example is represented as an instance of this model in Figure 3. In this model, not only the biological entities are classes (compound, enzyme), but also the activities (reaction, catalysis). This allows the definition of attributes for activities, as well as more elaborate relationships such as enzymatic activation or inhibition. The symbols (1 and $*$) on the relations describe the cardinality of the relations, where $*$ stands for *any*. For instance, a compound can be related through the *substrate* relation to 0, 1 or more reactions, and vice versa. A reaction may have different catalysts, but a catalyst is always related to *one* reaction and *one* enzyme.

Object-oriented data models were already used in molecular biology a decade ago (e.g. [10]). Object models are currently used for the design of biological databases such as UM-BDD [12], EcoCyc [26], aMAZE [60, 59] and SHARKdb [44]. The BioMiner software system [55] is also based on an object data model; among the available tools, PathFinder predicts biochemical pathways by comparing groups of related organisms. The Intact project [45], aiming at representing and analysing protein-protein interactions, takes an object-oriented approach.

Like bipartite graphs and hypergraphs, a (non trivial) object-oriented data model can represent reactions unambiguously. However, the potential coverage of an object-oriented data model is larger. Such a model allows the representation of various aspects of metabolic (e.g. enzymatic reaction), regulatory (e.g. gene activation) or signalling (e.g. complex assembly, translocation) pathways. An interesting advantage of object-oriented models is their ability to integrate metabolic, regulatory *and* signalling networks in a single data model. This is especially important as a clear separation between regulatory and signalling pathways is difficult, and combined analyses would be useful. Object models are not, however, a panacea. The design of a suitable model is a difficult task and the resulting model can be complex and difficult to understand and to exploit. Object-oriented data models are especially well suited when a more detailed and structured representation than bipartite graph is required. The major characteristics of reaction graphs are summarized in Figure 5.

7 Simulation models

Data models for the analysis of networks are usually driven by the existing information, collected from experiments and stored in databases. A complementary approach is the computational modelling of biochemical networks [4]. Computational models are theories and mathematical models of processes, often metabolic or regulatory, explaining the behaviour of biological systems. Such models can be used for simulation and for engineering. When applied to metabolic networks, they often aim at the development of targeted methods for improving the metabolic capabilities of industrially relevant micro organisms [64]. The use of metabolic pathways and metabolic fluxes are at the core of metabolic engineering [56]. In this section, we briefly overview some of the existing data models oriented towards simulation.

Several approaches exist for the mathematical modelling of metabolism [38, 11, 50]. A review of existing models can be found in [64], and a literature review of modelling and simulation of genetic regulatory systems is proposed in [8]. *Metabolic flux analysis* considers models that relate to the quantification of flux. It is based on the principle of mass conservation. It requires information on stoichiomtry; no data on enzyme kinetic is required. This includes metabolic balancing and isotopic balancing. Examples of metabolic balancing can be found in [28, 29], [52] and [36] (flux optimisation). Isotopic balancing can be achieved by isotopic labelling of ^{13}C; ^{13}C and analysing the balancing of the labelled species. Metabolic Flux Analysis, see [63] for review, is presently a important tool in metabolic engineering as it allows a detailed quantification of all intra-cellular fluxes.

Kinetic models detail the interaction between substrates and enzymes; they involve the solution of differential equations, for which several quantitative parameters have to be specified. Examples can be found in [20], [43] and [4]. Finally, *metabolic control analysis* is concerned with quantifying the control of flux among enzymes.

Petri nets have been widely used for the formalisation and the simulation of biochemical processes. Petri nets, an active research domain in computer science and mathematics, are a graph-oriented formalism allowing the modelling and analysis the concurrent behaviour of systems. Petri nets are special bipartite graphs with an associated semantics. The two types of nodes are *place nodes* and *transition nodes*; places may contain tokens. Depending on the tokens, transitions will be enabled, producing new configurations. A complete description of Petri nets is beyond the scope of this paper (see for instance [39]). In the context of biochemical models, Petri nets can be seen as an extension of the bipartite graph model. Place nodes are the compounds, and the transition nodes are the reactions. Among other things, Petri nets have been used for the representation and the simulation of metabolic knowledge [22, 31], for the simulation of *dynamic behaviour* [5, 30, 34], as well as for the analysis of gene expression [68]. Object-oriented simulation models have also been developed. Other approaches, originating from concurrency in computer science, have also been used to model biochemical processes. For example, pi-Calculus [46], a process algebra, and concurrent logic programming [2].

8 Conclusion and research directions

In this paper we have reviewed and classified the different forms of data models found in the literature for the analysis of biochemical pathways. We concentrated on models for analysing the structure of these networks, focusing on the analysis of existing information collected from experiments and stored in databases. We used a framework based on graph formalism when presenting the surveyed works.

Why have different data models been used for analysing biochemical networks? We observed that existing pathway analysis is mostly performed on simplified models, such as compound- reaction- or bipartite-graphs. Such models permit the use of existing path searching algorithms, and can be applied to large networks. However such models only permit simple analysis. More elaborate data models, such as object models, are now being used for pathway analysis. Object models are a powerful tool for modelling the complex structure and the intricate interactions of biochemical pathways. The sophisticated analysis of such pathways requires more elaborate search methods and algorithms. Object models can integrate metabolic, regulatory and signalling network in a single data model, removing the (artificial) barrier between these networks, and allowing analysis of the whole network. We believe that the following research directions would be beneficial to the field. A first research direction could be the design of unified object-oriented data models, combining metabolic, regulatory and signal transduction pathways, facilitating the development of new integrated search and analysis tools. Another research direction could be the design of elaborate search methods and analysis algorithms, that can operate on the combined networks.

Why use different data models for different applications (databases, analysis of networks, simulation, etc.)? Graph-based models, such as compound- reaction- or bipartite- graphs are mostly used for analysing networks, but are inappropriate for database modelling. Extension of graphs, such as Petri nets, are adapted for simulation purposes. Object-oriented models are adequate for database modelling and for analysing networks. Because object-oriented data models are mostly designed to model experimental data, they usually do not handle quantitative information. However, object models allow the description of dynamic properties (e.g. attributes in class diagrams, sequence diagrams, state charts) and can therefore integrate simulation information. Various mathematical models have however been developed to handle flux analysis or kinetics aspects; these computational models can be used for simulation and for metabolic engineering. However, computational models are usually not yet able to take into account the complex structure and the interactions involved in biochemical pathways. Existing computational models are therefore not adapted for the analysis of biochemical networks.

According to Bower and Bolouri, “recent advances in biology require a more direct connection between modelling and experiment. Instead of being a means to demonstrate a particular preconceived functional idea, modelling should be seen as a way to organize and formalize existing data on experimentally derived relationships between the components of a particular biological system”[4]. This raises the question of integrating data models. Is it conceivable to use one type of model for the modelling of databases, pathway

analysis and other computational uses? Today, the answer is no. However, object-oriented models are good candidates for such an integration. A possible research direction could be the extension of object-oriented data models with quantitative information, together with models of flux and dynamics. Simulation could then be performed by extracting suitable information from the model and using a specialized computational tool.

Acknowledgments

The two first authors acknowledge the support received from the UK Engineering and Physical Sciences Research Council (EPSRC, project GR/S07490/01). The authors wish also to thank Christian Lemer, Jean Richelle and Aik Choon Tan for useful discussions. We are grateful to the anonymous reviewers for their constructive comments.

References

[1] GD Bader, I Donaldson, C Wolting, BF Ouellette, T Pawson, and CW Hogue. BIND–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–5, 2001.

[2] Alexander Bockmayr and Arnaud Courtois. Using hybrid concurrent constraint programming to model dynamic biological systems. In *18th International Conference on Logic Programming*, volume LNCS 2401, pages 85–99. Springer, July 2002.

[3] Grady Booch, Ivar Jacobson, and James Rumbaugh. *Unified Modeling Language User Guide*. Addison Wesley, 1998.

[4] James M. Bower and Hamid Bolouri (Editor), editors. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, Massachusetts, 2001.

[5] Ming Chen and Andreas Freier. Petri net based modelling and simulation of metabolic networks in the cell. In *Bioinformatics Research and Education Workshop*, Hinxton,UK, 2002. EMBL-EBI.

[6] Weidong Chen, Michael Kifer, and David Scott Warren. HILOG: A foundation for higher-order logic programming. *Journal of Logic Programming*, 15(3):187–230, 1993.

[7] F. Couche. Recherche de chemins sur les voies metaboliques. Memoire de fin d'etudes, licence en informatique, ULB, 2002.

[8] Hidde de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[9] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, G. Nisanci, R. Cetin-Atalay, and M. Ozturk. PATIKA: An integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 18(7):996–1003, 2002.

[10] F. Dorkeld, G. Perrière, and C. Gautier. Object-oriented modelling in molecular biology. In J.G. Ganascia, editor, *Proceedings of the Artificial Intelligence and Genome WorkshoI, JCAI*, pages 99–106, 1993.

[11] J.S. Edwards and B.O. Palsson. How will bioinformatics influence metabolic engineering? *Biotechnology and Bioengineering*, 58:162–169, 1997.

[12] LB Ellis, SM Speedie, and R McLeish. Representing metabolic pathway information: an object-oriented approach. *Bioinformatics*, 14(9):803–806, 1998.

[13] L.B.M. Ellis, C.D. Hershberger, E.M. Bryan, and L.P. Wackett. The university of minnesota biocatalysis/biodegradation database: Emphasizing enzymes. *Nucleic Acids Research*, 29:340–343, 2001.

[14] L. T. Fan, B. Bertok, and F. Friedler. A graph-theoretic method to identify candidate mechanisms for deriving the rate law of a catalytic reaction. *Computers and Chemistry*, 26:265–292, 2002.

[15] D.A. Fell and A. Wagner. Animating the cellular map. Animating the cellular map, chapter Structural properties of metabolic networks: implications for evolution and modeling of metabolism, pages 79–85. Stellenbosch University Press, Stellenbosch, 2000.

[16] Christian V. Forst and Klaus Schulten. Evolution of metabolisms: a new method for the comparison of metabolic pathways. In *Proceedings of the third annual international conference on Computational molecular biology (RECOMB99)*, pages 174–181. ACM Press, 1999.

[17] F. Friedler, K. Tarjan, Y. W. Huang, and L. T. Fan. Graph-theoretic approach to process synthesis: Axioms and theorems. *Chem. Engng Sci.*, 47(8):1973–1988, 1992.

[18] Ken-ichiro Fukuda and Toshihisa Takagi. Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837, 2001.

[19] Julien Gagneur, David B. Jackson, and Georg Casari. Hierarchical analysis of dependency in metabolic networks. *Bioinformatics*, 19(8):1027–1034, May 2003.

[20] I Goryanin, TC Hodgman, and E Selkov. Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics*, 15(9):749–758, 1999.

[21] S. Goto, T. Nishioka, and M. Kanehisa. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res*, 28:380–382, 2000.

[22] R. Hofestädt and S. Thelen. Quantitative modeling of biochemical networks. *In Silico Biol.*, 1(1):39–54, 1998.

[23] T. Igarashi and Kaminuma T. Development of a cell signaling networks database. In *Proc. PSB Pacific Symposium on Biocomputing 2*, pages 187–197, 1997.

[24] H. Jeong, Tombor B, Albert R, Oltvai ZN, and Barabasi AL. The large-scale organization of metabolic networks. *Nature*, 406:651–654, 2000.

[25] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, S. M. Paley, and A. Pellegrini-Toole. The ecocyc and metacyc databases. *Nucleic Acids Res*, 28(1):56–59, 2000.

[26] Peter D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3):269–285, 2000.

[27] Peter D. Karp, Suzanne Paley, and Pedro Romero. The Pathway Tools software. *Bioinformatics*, 18(90001):225S–232, 2002.

[28] S. Klamt and Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports*, 29(1-2):233–236, 2002.

[29] S. Klamt, Schuster S., and Gilles E.D. Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnol. Bioeng.*, 77:734–751, 2002.

[30] I. Koch, S. Schuster, and M. Heiner. Simulation and analysis of metabolic networks by timedependent petri nets. In E. Wingender, R. Hofestädt, and I. Liebich, editors, *Computer Science and Biology - Proceedings of the German Conference on Bioinformatics, ISBN 3-00-005121-X*, pages 208–210. GBF-Braunschweig, 1999.

[31] Robert Kuffner, Ralf Zimmer, and Thomas Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825–836, 2000.

[32] Krishnamurthy L, Nadeau J, Ozsoyoglu G, Ozsoyoglu M, Schaeffer G, Tasan M, and Xu W. Pathways database system: an integrated system for biological pathways. *Bioinformatics*, 19(8):930–7, May 2003.

[33] H. Ma and Zeng A-P. Reconstruction of metabolic network from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19:270–277, 2003.

[34] Hiroshi Matsuno, Hitoshi Aoshima, Atsushi Doi, Yukiko Tanaka1, Mika Matsui, and Satoru Miyano. Biopathways representation and simulation on hybrid functional petri net. *In Silico Biology*, 3(32), 2003.

[35] Gerhard H.W. May. A graph-based pathway searching system over a signal transduction database. Information technologies, University of Glasgow, 2002.

[36] P Mendes and D Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[37] R. Milo, S Sen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[38] John A. Morgan and David Rhodes. Mathematical modeling of plant metabolic pathways. *Metabolic Engineering*, 4:80–89, 2002.

[39] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

[40] H Ogata, Fujibuchi W, Goto S, and Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, 28(20):4021–8, 2000.

[41] R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Jr Selkov, Kyrpides N., Fonstein M., Maltsev N., and Selkov E. Wit: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res*, 28(1):123–125, 2000.

[42] Suzanne M. Paley and Peter D. Karp. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, 18(5):715–724, 2002.

[43] T Pfeiffer, I Sanchez-Valdenebro, JC Nuno, F Montero, and S Schuster. METATOOL: for studying metabolic networks. *Bioinformatics*, 15(3):251–257, 1999.

[44] John Pinney. Metashark: a database and software toolkit for metabolic reconstruction from genomic DNA. Master of research in bioinformatics, The University of Leeds, School of Biochemistry and Molecular Biology, 2002.

[45] Intact Projet. http://www.ebi.ac.uk/intact/.

[46] A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Proceedings of the Pacific Symposium of Biocomputing*, PSB2001, pages 459–470, 2001.

[47] Mah R.S.H. Application of graph theory to process design and analysis. *Comput. Chem. Engin.*, 7:239–257, 1983.

[48] H. Salgado, A. Santos, U. Garza-Ramos, J. VanHelden, E. Díaz, and J. Collado-Vides. Regulondb (version 2.0): A database on transcriptional regulation in escherichia coli. *Nucleic Acids Res.*, 27:519–607, 1999.

[49] Faye Schilkey. PathDB : a pathway database. http://www.ncgr.org/pathdb.

[50] C.H. Schilling, Schuster S., Palsson B.O., and Heinrich R. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 15:296–303, 1999.

[51] D. Schomburg. The brenda database. http://www.brenda.uni-koeln.de/.

[52] S. Schuster, Dandekar T., and Fell D.A. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, 17:53–60, 1999.

[53] H. Seo, D.-Y. Lee, S. Park, L.T. Fan, S. Shafie, B. Bertok, and F. Friedler. Graph-theoretical identification of pathways for biochemical reactions. *Biotechnology Letters*, 23:1551–1557, 2001.

[54] S. Shen-Orr, R Milo, S Mangan, and U Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[55] M. Sirava, T. Schafer, M. Eiglsperger, M. Kaufmann, O. Kohlbacher, E. Bornberg-Bauer, and H. P. Lenhof. BioMiner–modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*, 18(90002):219S–230, 2002.

[56] Gregory Stephanopoulos. Metabolic fluxes and metabolic engineering. *Metabolic Engineering*, 1:1–11, 1999.

[57] Takai-Igarashi T, Nadaoka Y, and Kaminuma T. A database for cell signaling networks. *J.Comp.Biol.*, 5(4):747, 1998.

[58] J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, and S. Wodak. Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. In O. Gascuel and M.-F. Sagot, editors, *Computational Biology : First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000*, volume LNCS 2066, pages 155–172. Springer, 2001.

[59] Jacques van Helden, Avi Naim, Christian Lemer, Renato Mancuso, Matthew Eldridge, and Shoshana J. Wodak. From molecular activities and processes to biological function. *Brief Bioinform*, 2(1):81–93, 2001.

[60] Jacques van Helden, Avi Naim1, Renato Mancuso, Mattew Eldridge, Lorenz Wernisch, David Gilbert, and Shoshana J. Wodak. Representing and analysing molecular and cellular function in the computer. *Biol Chem*, 381(9-10):921–35, 2000.

[61] J. van Helden J, Wernisch L, Gilbert D, and Wodak S J. Graph-based analysis of metabolic networks. In Mewes H-W et al., editor, *Bioinformatics and genome analysis*, pages 245–274. Springer-Verlag, 2002.

[62] Andreas Wagner and David Fell. The small world inside large metabolic networks. In *Proc R Soc Lond B Biol Sci*, volume 268(1478), pages 1803–10, 2001.

[63] Wolfgang Wiechert. 13c metabolic flux analysis. *Metabolic Engineering*, 3(3):195–206, July 2001.

[64] Wolfgang Wiechert. Modeling and simulation: Tools for metabolic engineering. *Journal of Biotechnology*, 94(1):37–63, 2002.

[65] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüß, I. Reuter, and F. Schacherer. Transfac: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–319, 2000.

[66] U. Wittig and A. De Beuckelaer. Analysis and comparison of metabolic pathway databases. *Briefings in Bioinformatics*, 2(2):126–142, 2001.

[67] I Xenarios, L Salwinski, XJ Duan, P Higney, S Kim, and D Eisenberg. Dip: The database of interacting proteins. a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30:303–5, 2002.

[68] A. Zien, Kueffner R., Zimmer R., and Lengauer T. Analysis of gene expression data with pathway scores. *Ismb*, pages 407–417, 2000.

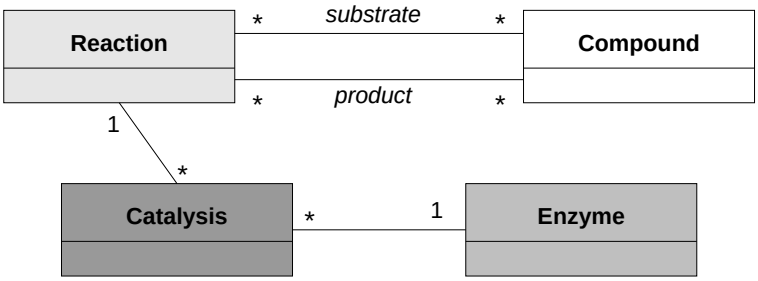

Figure 7: A simplified object model of metabolic regulation