Graph-based analysis of metabolic networks

JACQUES VAN HELDEN[1,2], LORENZ WERNISCH[1,3], DAVID GILBERT[4], AND SHOSHANA WODAK[1,2]

[1] European Bioinformatics Institute (EBI). Genome Campus - Hinxton Cambridge CB10 1SD - UK.

[2] Unité de Conformation des Macromolécules Biologiques. Université Libre de Bruxelles.50 av. F.D. Roosevelt. B-1050 Bruxelles. Belgium.

[3] School of Crystallography. Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

[4] Department of Computing, City University, Northampton Square, London ***EC1V*** *0HB, UK.*

E-mail: jvanheld@ucmb.ulb.ac.be, l.wernisch@mail.cryst.bbk.ac.uk, drg@cs.city.ac.uk, shosh@ucmb.ulb.ac.be
Tel: +32 2 650 20 13
Fax: +32 2 648 8954

Contents

Summary

This paper describes applications of graph theory to the analysis of metabolic networks. First it compares different ways of mapping metabolic networks onto a graph. Next, it discusses several crucial aspects that need to be taken into account for building the graphs in order to reflect the specificity of metabolic pathways. Those aspects are the treatment of ubiquitous compounds and of reversible reactions. This is followed by a summary of results obtained with different approaches, such as analysis of network structure, shortest path finding, path enumeration, and reaction clustering. Several of these approaches are based on the combination of information on metabolism and gene regulation, and have useful applications for the interpretation of gene expression data.

Introduction

Metabolic pathways have been characterised for many decades by biochemists on the basis of analyses of a few selected model organisms (Cohen, 1994). Enzymes-coding genes have been identified by genetic approaches, cloned, and sequenced; thousands of enzymes have been purified, and their mode of action, including substrates and inhibitors, have been investigated experimentally (Fersht, 1985). Until relatively recently, this enormous body of knowledge has been archived primarily in the scientific literature. In the last 10 years however, efforts have been made to store it in electronic form in a variety of databases. Those include some of the main resources for information on protein function such as SWISS-PROT (Bairoch & Apweiler, 2000) and more specialised databases such as KEGG (Goto *et al*., 2000), EMP/WIT (Overbeek *et al*., 2000), Ecocyc (Karp *et al*., 1996; Karp *et al*., 2000), and the enzyme resource BRENDA (Schomburg *et al*., 1990-1995).

These databases have in turn been used to predict enzyme functions for newly sequenced genes, by using criteria based on sequence similarity alone, or by applying more global analyses of the complement of enzymes found in the complete genome. One such method termed 'metabolic reconstruction' (Overbeek et al., 2000), assigns function to a gene product based on its likelihood of catalysing a reaction in a known metabolic pathway. It has now been applied to infer the metabolism of more than 40 completely sequenced micro-organisms, with the results stored in the WIT database.

The availability of all these data offers the unprecedented opportunity of obtaining a global view of the metabolic network of an organism and for analysing this network using objective quantitative approaches.

Several types of approaches have been proposed for mapping and analysing biochemical networks in absence of detailed kinetic information (Schuster *et al.*, 2000). Earlier methods constructed chemical transformation pathways leading from one compound to another by the successive addition of reaction steps (Mavrovouniotis *et al.*, 1990; Seressiotis A & J.E., 1986). Subsequently, various procedures based on flux analysis have been introduced and shown to be useful for analyzing metabolic networks and predicting their response to various perturbations (Fell, 1994; Schilling & Palsson, 1998; Schuster *et al.*, 1999; Schuster & Hilgetag, 1994; Simpson *et al.*, 1995).

This paper focuses on a different category of approaches, which rely on graph-based representations of networks (Fell & Wagner, 2000; Jeong *et al.*, 2000; Kuffner *et al.*, 1999; Kuffner *et al.*, 2000; Ogata *et al.*, 2000; van Helden *et al.*, 2001a; van Helden *et al.*, 2000; Zien *et al.*, 2000). In particular, it discusses several applications of graph theory to the analysis of metabolic networks, and to the mapping of information on gene expression data (van Helden et al., 2001a; van Helden et al., 2000; Zien et al., 2000) onto these networks. Specific examples of these applications are taken from work performed in our laboratory. Clearly, these applications are presently at their initial stages, with much room left for improvements, for which some directions are outlined in the Concluding Remarks.

What is presented here is thus not a comprehensive review of the field but merely an attempt to provide the reader with a glimpse of some of the challenges that we face in trying to establish correlations between the linear world of genome sequences and the non-linear world of cellular processes.

Sources of data on metabolic networks

A first crucial prerequisite for any genome-scale analysis of a metabolic network is to collect a consistent and (as far as possible) comprehensive set of data on the metabolic pathways. This includes information on all the enzymes, catalysed and non-catalysed chemical reactions, and the corresponding small molecule compounds.

One group of databases contains a comprehensive set of such data, with high information content (catalytic constants, inhibitors, activators, …) and numerous literature references. This includes BRENDA (Schomburg et al., 1990-1995) and EMP/WIT (Overbeek et al., 2000), representing some of the earliest initiatives to store these data. However, the information stored in these databases is, currently at least, not sufficiently structured to enable easy application of algorithmic approaches.

Another very important database with undoubtedly one of the richest information content on protein and enzyme function is SWISS-PROT (Bairoch & Apweiler, 2000). But here too this information is not readily amenable to algorithmic analyses, although recently, Object-Oriented parsing libraries have been distributed (Hermjakob *et al*., 1999), which greatly facilitates the conversion of the information stored in SWISSPROT into a more structured format. A very relevant resource for data on metabolism is ENZYME, a satellite database of SWISS-PROT, which contains the Enzyme Classification (EC), now used as reference world-wide, together with information on the function of the classified enzymes (Bairoch, 1993; Bairoch, 2000).

A second class of databases, of which LIGAND/KEGG (Goto et al., 2000; Kanehisa & Goto, 2000) is a typical example, contains well structured data but the information coverage is less extensive. KEGG is in fact primarily a genome database, and contains information only on proteins and enzymes whose gene sequence has been determined. Information on the many enzymes described in the literature, which have only been characterised biochemically or genetically is not stored.

Küffner and co-workers (Kuffner et al., 2000) compared the number of chemical reactions stored in BRENDA, KEGG and ENZYME and showed that BRENDA contains three times as many reactions as the latter two databases.

Another important resource is EcoCyc, the Encyclopaedia of *Escherichia coli* genes and metabolism (Karp & Paley, 1996; Karp et al., 2000). This database is both highly structured (Karp, 2000) and has a good information coverage (Riley, 1993; Riley, 1997; Riley, 1998). Initially restricted to *E. coli*, EcoCyc is currently being extended to other organisms (Karp et al., 2000).

Our group is presently developing a database containing information on protein function and biochemical pathways in different organisms, called aMAZE. One important aim of this database is to organise the information in ways that allow answering complex queries about biochemical networks using automatic procedures. To that end, we developed a flexible and powerful data model (van Helden *et al.*, 2001b; van Helden et al., 2000), that enables the representation of a variety of processes, including metabolic pathways, regulation of gene expression and enzyme function, signal transduction and transport. Presently, the database is populated with chemical reactions imported from KEGG and information on polypeptides imported from SWISS-PROT. Incorporation of information in BRENDA is under way, and we are adding custom annotation on metabolic regulation. The database and query tools will become available on the Web shortly (C. Lemer *et al.*, *in prep*).

Metabolic network graphs: representation and navigation

Mapping the metabolic network onto a graph

The second key issue that needs to be addressed in order to be able to perform systematic analyses of metabolic network, is the choice of an appropriate model for representing them. The metabolic network first needs to be mapped onto a graph. There are many ways for performing this mapping, and the choice among them may partly depend on the purpose of the analysis and on the algorithms used.

Figure 1A shows part of a metabolic network, represented according to the conventional way used by the biochemist. This network contains two alternative pathways for the conversion of L-homoserine into L-methionine, as observed in *E.coli* (left side) and *S.cerevisiae* (right) respectively. Figures 1B-D depicts three alternative ways of mapping this small network onto a graph. One is to associate one node to each compound, and to represent reactions as arcs connecting substrates to products (Figure 1B). Nodes are labelled with the compound name, and arcs with the identifier of the reaction (or its EC number). In this representation, a reaction with *n* substrates and *m* products appears as the label of *n*m* distinct arcs. Such a representation directly provides information about relationships between compounds, but it is not straightforward to use it for path finding, because any path finding algorithm would have to be adapted in order to avoid using two arcs with the same label in a path.

Another possibility would be to associate one node to each reaction, and to link them by an arc when they share an intermediate compound (the product of one reaction is the substrate of the next), as illustrated in Figure 1C. This time a compound used as substrate in *n* reactions and produced by *m* reactions appears as the label of *n*m* arcs, giving rise to the same complications for path finding as described above.

A more elaborate representation is to define a graph having two types of nodes, one for compounds and one for reactions (Figure 1D), with the arcs representing the input/output relationships between reactions and compounds. In this representation, arcs always link compounds to reactions (for substrates) or reactions to compounds (for products), but never reactions to reactions or compounds to compounds. This type of graph, called a *bipartite* graph, has been used in several approaches reviewed here, and will therefore be the representation of choice throughout the paper.

A particular case of bipartite graphs is the Petri nets, which have been used to represent metabolic pathways and to perform various analyses, such as path finding, network comparison, and simulation (Kuffner et al., 2000). In this representation, reactions are mapped onto transition nodes, and compounds onto place nodes.

We developed a general data model for representing biochemical pathways and their regulation (van Helden et al., 2001b; van Helden et al., 2000), based on of two main classes of objects: *BiochemicalEntities* and *BiochemicalInteractions*. The data model can be mapped onto a graph by assigning one node per BiochemicalEntity and one node per BiochemicalInteraction, and using arcs to connect BiochemicalIntreactions to their inputs and outputs. *BiochemicalInteractions* are further subdivided into *Transformations* and

Controls. *Transformations* behave like the transitions of a Petri net, in that they have *BiochemicalEntities* both as inputs and outputs. The difference between our model and Petri nets resides in the other subclass of interactions, *Controls*, which have the particularity of having another *BiochemicalInteraction* as output. *Controls* are themselves a subclass of *BiochemicalInteractions*. They represent regulatory interactions (inhibition, up-regulation, down-regulation), which most other databases do not represent, and can be regulated in turn by other *Controls*. Note that when *Control* interactions are taken into consideration, the underlying graph is not bipartite anymore.

Finally, in our data model, networks of biochemical pathways are represented, using the *Process/Pathway* class, which consists of a graph whose nodes are *BiochemicalEntities*, *BiochemicalInteractions* or even other *Processes/Pathways.* The latter case allows the representation of super-pathways, or 'pathways of pathways' (van Helden et al., 2000).

As shown elsewhere (van Helden et al., 2001b) this data model allows the representation of some complex regulations (e.g. the regulation of one specific catalytic activity of a multifunctional enzyme), which cannot be treated properly with other models. It furthermore has the advantage of lending itself naturally to the application of graph theory algorithms for network analysis.

Navigation through metabolic network graphs

Path finding (finding all possible paths between compounds A and B, or all paths leading to compound C etc.) is one of the most common tasks in graph analysis, and several algorithms have been developed for this purpose (Gross & Yellen, 1999). The choice of the appropriate method depends on the question to be addressed and on the type of network analysed (e.g. metabolism, signal transduction, protein-protein interactions).

This notwithstanding, classical path finding algorithms need some adaptation in order to yield meaningful results for metabolic networks. In particular it is crucial to take into account two specific properties of these networks. Namely, the reversibility of chemical reactions and the necessity for being selective when navigating through intermediate compounds. An improper treatment of these issues may produce perfectly acceptable solutions from the mathematical view point, but those solutions will be devoid of biochemical meaning.

Treatment of ubiquitous compounds

Although all the substrates of a given reaction are necessary for this reaction to take place, they cannot be considered as equivalent for path finding purposes. This is illustrated with a trivial example in Figure 2A: water (H_2O) is produced by reaction EC4.2.1.52 and consumed by reaction EC3.5.1.18. This is however not sufficient for connecting these two reactions as successive steps of a pathway, and it would make no biochemical sense to claim that L-aspartic semialdehyde can be converted to LL-diaminopimelic acid in two steps, with H_2O as intermediate compound (bottom of Fig. 2A). The problem is particularly evident for water, but also holds for a few other ubiquitous compounds, which are involved in hundreds of reactions but cannot be considered as valid intermediates for path finding, or for that matter, for establishing biologically meaningful network connections.

One simple and drastic solution would be to delete the corresponding nodes from the graph. But which compounds should be considered as invalid intermediates? Table 1 provides a list of the most connected compounds, as compiled from the reactions in KEGG. Compounds such as H_2O, NAD, NADH, orthophosphate are at the top of the list. Those should clearly not be considered as valid intermediates between two reactions. But the status of compounds such as pyruvate, or Acetyl-CoA is less clear. Another illustrative example of this difficulty is ATP, which is involved in hundreds of reactions, where it is used as energy source and can hence not be considered as a valid intermediate. But ATP is a perfectly valid intermediate between reactions in the pathways of nucleotide biosynthesis. Coming up with objective rules for defining valid intermediates in metabolic graphs would be very useful. Such rules could for example consider the transfer of matter (atoms) from one compound to the next.

Treatment of reaction reversibility

Another aspect specific to metabolic networks is reaction reversibility. This is the fact that the chemical reactions in these networks must be considered as reversible, unless specific information is given to the contrary. Indeed even chemical reactions with a strong directional preferences can be made to go in the reverse direction as a result of mass action because they take place in the context of other on-going processes in the cell. Information on the directional preferences (reversibility/irreversibility) of reactions in physiological conditions is annotated in some databases (BRENDA, WIT, EcoCyc).

One is faced with several choices for representing reaction reversibility in the metabolic graph. One possibility is to link reversible reactions and their reactants with undirected arcs, i.e. arcs that can be traversed in both directions during path finding (Figure 2B). This however would make it cumbersome to distinguish between substrates and products. Indeed, straightforward navigation through the graph would result in connecting 2 substrates (or 2 products) of the same chemical reactions to each other. In the context of path finding, this would mean that 2 substrates of the same reaction can be inter-converted in one step, thereby violating the laws of chemistry.

Another solution is to represent a reversible reaction as two separate nodes in the network, one for each direction, as illustrated in Figure 2C. With this solution however, path finding algorithms would have to be prevented from including both the direct and reverse reactions in the same path. Inclusion of both would indeed lead to the chemically meaningless situation where two substrates of a reversible reaction could be transformed into each other in two steps (Figure 2C).

Choice of graph analysis algorithms

In summary, several alternative representations could be suitable for mapping the metabolic network onto a graph. But whichever mapping is used, it is crucial to adapt path finding algorithms, in order to ensure that they return consistent and relevant answers. In particular, each time a reaction is added to a path, it is essential to check whether the corresponding reverse reaction has not already been included previously. This can be done easily with depth-first search (DFS) methods, which potentially have exponential running times, and not, or at least not in an obvious way, with breath-first-search (BFS) methods. For this reason, classical algorithms of counting paths based on powers of matrices (Gross & Yellen, 1999), which are the most efficient methods for graph analysis, are thus less useful in the case of metabolic networks since they can not be readily applied to them.

Structural properties of metabolic networks

Once a metabolic network is mapped onto a graph, it becomes possible to evaluate the global structural properties of the network.

The global network built from all the chemical reactions stored in KEGG (catalysed by enzymes from all organisms) contains 5,871 compounds and 5,223 reactions, each represented as one node in the graph. Connections between reactions and their substrates are represented by 21,194 arcs, of which 7,116 involve ubiquitous compounds (14 compounds selected from Table 1). After filtering out these trivial connections, the bipartite graph contains 11,094 nodes and 14,078 arcs in our representation.

Using a different approach for mapping metabolic networks into a graph, Jeong and co-workers (Jeong et al., 2000) recently evaluated several global parameters of these networks, such as the network connectivity. Their first conclusion is that metabolic networks belong to the class of scale-free networks, where a small number of compounds are involved in many reactions, and are hence denoted as 'hubs', because they are heavily connected, whereas the majority of the other compounds are involved in only few reactions, and are hence weakly connected.

Figure 3, shows the results of a similar analysis, which we performed on the basis of the entire set of compounds and reactions from KEGG, mentioned above. Fig 3A shows in abscissa the number of connections (i.e. reactions in which a compound is involved), and in the ordinate the number of compounds with this number of connections/reactions. In this analysis substrates and products are pooled together since most reactions are reversible, and the direction found in the databases is often arbitrary. Details of the connectivity are provided in Table 1 for the 30 compounds with highest connectivity (the rightmost points on Figure 3A).

While the paper of Jeong and co-workers never mentions the identity of their so-called "hubs", our Table 1 shows clearly that these are none else than the ubiquitous compounds like H_2O, NAD, NADH, ATP, S-adenosyl-L-methionine. Biochemists know that these compounds are the actors of basic metabolic operations like oxido-reduction, phosphorylation, methyl transfer, and it is probably not a big surprise to them to see that these compounds are involved in significantly more reactions than other compounds.

Another interesting structural parameter is the average connectivity of the reactions in the network. Counting the number of substrates and products for the entire set of KEGG reactions, we find that on average, reactions have 2 substrates and 2 products (Figure 3B), but when ubiquitous compounds are discarded, these numbers drop to 1.4 and 1.3, respectively (Figure 3C).

The second conclusion of Jeong and co-workers (Jeong et al., 2000) is that metabolic networks have a small-world character, i.e. that on average, any two nodes can be connected by rather short paths. To reach this conclusion they calculate the average length (or average number of steps), of the shortest path between any pair of compounds, which is defined as the network diameter. It is important to remember however, that their conclusion crucially depends on the way in which the metabolic networks were mapped into the graph. In particular, on how ubiquitous compounds are dealt with. In their paper, all the compounds, including the heavily connected H_2O, NAD molecules, were considered as valid nodes. As a result, most of their shortest paths include shortcuts involving these molecules (the problem illustrated in Figure 2A), and are hence not chemically acceptable.

This explains the surprising observation made by these authors that the network diameter remains constant irrespective of the number of enzymes in the considered organisms. The number of connections involving ubiquitous compounds is indeed so large (e.g. over one thousand for water, in Fig. 3A) that adding or deleting a set of enzymes makes little differences.

These authors show on the other hand that the network diameter increases significantly from 3 to 10 steps when the 25 most ubiquitous compounds are removed as illustrated in their Figure 3e. Unfortunately, this larger value, which probably reflects much better what biochemists would consider as the average distance between two compounds in a metabolic network, is not used in their analysis. Using it would have indicated that the network diameter is affected by adding or deleting enzymes and that the metabolic world is not as small, as Jeong and colleagues conclude it to be.

In a more recent paper (Fell & Wagner, 2000), the diameter of a network representing the core metabolism in *Escherichia coli* is computed after removal of ubiquitous compounds. In this paper, two compounds are considered as connected irrespective of the side of the reaction equation in which they occur. The distance separating two substrates of a same reaction is thus 1, as in our Figure 2B. This is a deliberate choice, justified by the fact that the authors are interested in the capability of two compounds to interact, i.e. the fact that each substrate will influence the rate of the reaction, which will affect the concentration of the other substrate (D.A. Fell, personal communication). Even though perfectly justified in this case, it is important to bear in mind that the distances calculated by Wagner and Fell have nothing to do with the physical pathways describing the inter-conversion of compounds.

In summary, it appears that the global structural properties of metabolic networks published so far do not really reflect the biochemical notion of pathways, and should thus be interpreted with caution.

Graph analysis methods applied to metabolic networks

Pathway enumeration

With the metabolic network graph and some basic path navigation rules in place, a number of useful graph analysis operations can be performed in order to answer some typical queries. These include, finding all possible paths between a given pair of source and target nodes (two-end path finding), finding all possible paths starting from a given source node (one-end path finding), and finding all possible paths ending at a given target node (one-end reverse path finding). In performing these analyses, a limit on path length (number of intervening reactions) can be imposed to limit calculation time.

Two-ends path finding can be used with the aim of discovering alternative pathways between two compounds A and B in the global network built from all known reactions and compounds. One application of pathway enumeration might be to discover alternatives to the classical pathways. Indeed, it is well known that different organisms often use distinct pathways to synthesise or degrade the same metabolites. For example, biosynthesis of methionine from homoserine requires 3 steps in the yeast *Saccharomyces cerevisiae*, and 4 steps in the bacteria *Escherichia coli*, with only one reaction in common (Figure 1A). In other cases, such as lysine biosynthesis, completely different pathways are used in *E.coli* and *S.cerevisiae*, with not a single reaction in common. Since most biochemical knowledge has been derived from a few model organisms, one might expect that many more alternative pathways will be discovered, particularly as our knowledge of enzymes and the reactions they catalyse in different organisms expands.

The exhaustive enumeration of paths between the source and target compounds, though straightforward, may however not always yield a practical solution. Often, the number of computed paths is so large that the result is devoid of any biochemical meaning. This is clearly illustrated by Küffner and co-workers, who computed more than 500,000 possible paths of at most 9 steps from glucose to pyruvate (Kuffner et al., 2000) using the set of reactions and compounds in databases such as KEGG and BRENDA.

One way of reducing the number of computed paths is to impose constraints on the derived solutions. For example to compute only closed paths, defined as sub graphs in which the net production and consumption of compounds is zero, except for the starting and end compounds and a predefined list of ubiquitous molecules (Kuffner et al., 2000). Another constraint can be to set limits on path width (see (Kuffner et al., 2000) for the formal definition of closed pathway and path width). This was shown to reduce the number of paths in the above mentioned case from 500,000 to 541 (with a width limit of 2) or 170 (width limit of 1) (Kuffner et al., 2000). However, restricting path width too severely, bears the danger of missing a sub branch of a branched pathway, and can therefore not be applied indiscriminately.

Yet another option is to restrict the metabolic network to the subset of reactions known to be catalysed in the organism of interest. However, while this strongly reduces the size of the metabolic graph, and consequently the number of computed paths, it also limits the ability of finding alternative pathways, and this for two main reasons. It precludes the consideration of spontaneous reactions, those not catalysed by any enzyme (e.g. the third reaction of proline biosynthesis: glutamate gamma-semialdehyde <=> 1-pyrroline carboxylate), as well as the reactions for which no enzyme has as yet been annotated in the corresponding genome. Imposing constraints on the graph itself is thus not a satisfactory solution for reducing the number of computed paths when the aim is to predict new pathways or to reconstruct the pathways of an organisms whose metabolism is incompletely characterised.

Scoring enumerated pathways with gene expression data

An alternative to imposing constraints a priori, is to use data from other sources for guiding the selection of likely paths. In one such approach pertinent pathways are selected from among all the solutions provided by the path enumeration procedure on the basis of gene expression data (Zien et al., 2000). The underlying rationale is the observation that enzymes participating in a given pathway tend to be co-regulated at the gene expression level. In the described approach, Zien and coll (Zien et al., 2000) rank the 541 glycolysis pathways computed by Kueffner et al (2000), according to 3 criteria derived from experimental data on the transcriptional response of *S.cerevisiae* to diauxic shift (DeRisi *et al*., 1997). One is a measure of the level of transcriptional response of the genes coding for the enzymes catalysing all the reactions in a given pathway. The second is a measure of the level of correlation between this response in all the enzyme coding genes in the pathway, and the third is a single criterion combining both measures. This scoring results in the top ranking of the yeast glycolytic pathway, suggesting that this approach may have useful applications for the prediction of metabolic pathways in organisms in which these pathways have not been previously characterised.

Pathway reconstitution from clusters of reactions

Instead of using gene expression data to score *a posteriori* the paths produced by exhaustive enumeration (Zien et al., 2000), one can use this information as the starting point for building pathways (van Helden et al., 2001a; van Helden et al., 2000). The problem at hand is thus formulated differently and the goal becomes to identify the set of processes which would link together, in a biologically meaningful way, most or all of the activities carried out by a cluster of functionally related genes.

An algorithm which performs such operation has recently been developed in our laboratory (van Helden et al., 2001a; van Helden et al., 2000). It works roughly as follows. Starting from a cluster of functionally related genes, it first identifies the reactions catalysed by their products from available information in databases such as KEGG, BRENDA and SWISS-PROT. These reactions are used as starting nodes or 'seeds' which the procedure then endeavours to interconnect via the shortest paths. In the simplest case, two seed reactions share an intermediate compound (the product of one reaction is a substrate of the other). But this is not always so and the program therefore allows to intercalate additional reactions, which were not part of the initial set of seeds.

The underlying algorithm is a single linkage clustering over the metabolic graph, with the distance metric between two reaction clusters being the length of the shortest path which links these clusters. The returned result is a subgraph containing all the seed nodes, as well as the arcs and intermediate nodes (reactions and compounds) traversed during the linkage process.

It was shown (van Helden et al., 2001a; van Helden et al., 2000) that this algorithm is cable of reconstituting the methionine pathway of *S. cerevisiae*, starting from a cluster of co-regulated genes identified from DNA micro-array experiments (Spellman *et al.*, 1998). We are currently systematically evaluating the performance of the program by testing its ability to reconstitute known metabolic pathways starting each time from different subsets of reactions. Figure 4 illustrates such tests performed with the *E. coli* lysine pathway. In this pathway, shown in Fig. 4A, L-aspartate is converted into L-lysine in 9 steps, each catalysed by a specific enzyme. Figure 4B displays the reconstructed pathway when all 9 reactions (defined only by their EC numbers) are provided to the clustering algorithm, in any order. We see that the reactions are placed in the appropriate order to form the lysine pathway as biochemists know it, and that the reactions could be ordered correctly, solely on the basis of the intermediate compounds. Figure 4C, displays the reconstructed pathway when only half of the EC numbers are provided as seeds. We see that the algorithm inferred the missing reactions, with however several alternatives for the intercalated reactions.

The pathway reconstructed using only four EC numbers as seeds (less than half the reactions in the known lysine pathway) is shown in Fig. 4D. We see that, in addition to identifying the missing reactions and correctly rebuilding the lysine pathway, the procedure detects two other paths. One is a shortcut converting 2,3,4,5-tetrahydrodipiclinate into meso-2,6-diaminoheptanedioate in 2 steps. The other is a degradation pathway, converting L-lysine into L-aspartate in two steps. Catalysts for this pathway are missing in *E.coli*, but are found in bacteria like Pseudomonas. The program was thus capable of reconstituting the *E.coli* pathway, but also identified pathways, which are not pertinent to this organism. These pathways can however be readily eliminated by scoring the pathways computed by our procedure according to the presence/absence of enzymes in the considered organism.

Finally, we observed that the algorithm fails to reconstruct the correct pathway (Fig. 4E) when only 3 of the 9 EC numbers are used as seeds, due to the fact that the remaining pairs of seed EC numbers can be linked by shorter paths.

Shortest path analysis of gene fusion data

Another way to obtain information about functional relatedness between genes is by gene fusion-fission analysis (Enright *et al.*, 1999; Marcotte *et al.*, 1999a; Marcotte *et al.*, 1999b; Tsoka & Ouzounis, 2000). This approach is based on the observation that 2 or more genes (termed "fusion partners" or "components") from one organism display significant sequence similarity with the sequences of non-overlapping segments of a single gene in another organism. Marcotte and coworkers (Marcotte et al., 1999a) showed that genes frequently involved in fusion events share a common keyword in their Swissprot annotation, indicating a functional relationship between them. Tsoka and Ouzounis (Tsoka & Ouzounis, 2000) showed that enzymes are more frequently involved in fusion events than a random selection of other proteins.

A direct way to assess the functional relationship between two fused genes is to calculate the length of the shortest path between the reactions they catalyse. The shorter the path, the greater the likelihood that the two enzymes functionally interact (for example, that they are involved in the same pathway). We calculated the shortest path length or distance between the two reactions catalysed by each pairs of fused enzyme from Tsoka and Ouzounis (Tsoka & Ouzounis, 2000), with the results displayed in Fig. 5A. A distance of 0 corresponds either to isofunctional enzymes, or to subunits of an enzymatic complex. A distance of 1 corresponds to enzymes catalysing successive steps in a metabolic pathway. We compared this result with the distribution of distances between pairs of randomly selected enzymes (Figure 5B). This analysis shows that fusions generally occur between enzymes involved in closely related reactions.

Concluding remarks

In this paper we discussed a number of approaches that use graph theory to represent metabolic networks, to analyse their global properties and to perform various operations on them. In particular we showed that several issues, such as reaction reversibility and ubiquitous compounds, need to be dealt with in an appropriate manner in order to obtain biologically meaningful results. But what is considered appropriate, may depend on the particular application that one has in mind. For instance, an adjacency matrix such as defined Fell & Wagner (2000), in which two substrates of the same reaction are directly connected, cannot be used to construct biochemically meaningful pathways, but is perfectly valid for determining if 2 compounds present some interaction at the level of mass action. The mapping onto a graph and the path finding algorithms should thus be reconsidered for each particular application.

Even when algorithms are adapted to take into account the particularities of metabolic networks, an analysis restricted to metabolic reactions is often poorly informative, due to the innumerable possible paths for converting any compound into any other. Biological processes however differ from chemical experiments in that cells are able to regulate the concentration and/or activity of their enzymes and transporters Thus, a comprehensive analysis should integrate metabolic, genetic and physiological information. With methods made available for measuring the level of expression of all the genes of an organism (Brown & Botstein, 1999; DeRisi et al., 1997), we are in a position to integrate metabolic and genetic information, as seen in one example discussed above. In cases where gene expression data are not available, metabolic network analysis can be combined with other sources of information, such as data from genome scale analyses of protein-protein interactions (Ito *et al.*, 2001; Uetz *et al.*, 2000; Uetz & Hughes, 2000). Alternatively, one could use clusters of genes or proteins predicted to be functionally related on the basis of theoretical analyses such as those recently described by Eisenberg and colleagues (Enright et al., 1999; Marcotte et al., 1999a; Marcotte et al., 1999b; Pellegrini *et al.*, 1999; Tsoka & Ouzounis, 2000). However, since the latter type of data is probably less reliable than those derived from experimental approaches, their effectiveness in helping the identification of relevant pathways may probably be reduced.

It is also clear that a set of objective rules are needed for building metabolic network graphs, which can be used to establish a biologically meaningful correlation between the information on genes, proteins and metabolism. Methods based on flux analysis, and particularly those based on the ‘elementary flux modes' (Clarke et al., Schuster et a. 2000) seem particularly relevant in this regard, since they are entirely based on stoicheometry and require no knowledge of kinetic parameters. Although these methods have so far been applied only to a limited set of metabolic processes, without a systematic incorporation of regulation, particularly that on the genetic level, combining some of the underlying concepts with graph analysis methods should be a useful way of going forward.

Lastly one should add that generalising these approaches to other types of cellular processes is also very important. Among those, the highly connected networks of signal transduction pathways have been receiving increased attention (Takai-Igarashi & Kaminuma, 1998; Takai-Igarashi *et al*., 1998; Weng *et al*., 1999). The data model developed for the aMAZE database (van Helden et al., 2001b; van Helden et al., 2000) already includes a representation of these interactions, and provides a good basis for extending the algorithmic approaches presented here. However, one should beware that these different types of interactions cannot a priori be considered equivalent to chemical reactions. For example, transcriptional regulation is intrinsically directional (the transcription factor regulates expression of its target genes, but the opposite is not true), and there is no consumption of the input. It is thus likely that specific rules will have to be defined for different subclasses of interactions, to make sure the mapping and graph traversal are biologically meaningful .

Acknowledgements

Georges Cohen of the Insitut Pasteur, Paris is gratefully acknowledged for useful suggestions and for help in the interpretation of the results. We are indebted to the entire aMAZE team, and particularly to Christian Lemer for numerous discussions and important suggestions. This work has been supported in part by a consortium of companies (Astra-Zeneca, Aventis, Organon, and Roche) and by funds from the European Commission (QLRI-CT-1999-01333).

References

- Bairoch, A. (1993). The ENZYME data bank. *Nucleic Acids Res* **21**(13), 3155-6.

- Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res* **28**(1), 304-305.
- Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**(1), 45-48.
- Brown, P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**(1 Suppl), 33-7.
- Cohen, G. N. (1994). *Biosyntheses*, Chapman & Hall.
- DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338), 680-6.
- Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events [see comments]. *Nature* **402**(6757), 86-90.
- Fell, D. A. (1994). . In *Modern Trends in Biothermokinetics*. Plenum Pub Corp, New-York.
- Fell, D. A. & Wagner, A. (2000). The small world of metabolism. *Nat Biotechnol* **18**(11), 1121-2.
- Fersht, A. (1985). *Enzyme Structure and Mechanism*. 2nd edit, W H Freeman & Co.
- Goto, S., Nishioka, T. & Kanehisa, M. (2000). LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res* **28**(1), 380-382.
- Gross, J. & Yellen, J. (1999). Graph theory *and its applications*. Discrete mathematics and its applications (Rosen, K. H., Ed.), CRC press, London.
- Hermjakob, H., Fleischmann, W. & Apweiler, R. (1999). Swissknife - 'lazy parsing' of SWISS-PROT entries. *Bioinformatics* **15**(9), 771-2.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**(8), 4569-74.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**(6804), 651-4.
- Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**(1), 27-30.

- Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics* **16**(3), 269-85.
- Karp, P. D. & Paley, S. (1996). Integrated access to metabolic and genomic data. *J Comput Biol* **3**(1), 191-212.
- Karp, P. D., Riley, M., Paley, S. M. & Pelligrini-Toole, A. (1996). EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res* **24**(1), 32-9.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M. & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28**(1), 56-59.
- Kuffner, R., Zimmer, R. & Lengauer , T. (1999). Pathway Analysis in Metabolic Databases via Differential Metabolic Display (DMD). *In Silico Biology*.
- Kuffner, R., Zimmer, R. & Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* **16**(9), 825-836.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**(5428), 751-3.
- Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function [see comments]. *Nature* **402**(6757), 83-6.
- Mavrovouniotis, M. L., Stephanopoulos, G. & Stephanopoulos, G. (1990). Computer aided synthesis of biochmical pathways. *Biotechnol. Bioeng.* **36**, 1190-1192.
- Ogata, H., Fujibuchi, W., Goto, S. & Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters [In Process Citation]. *Nucleic Acids Res* **28**(20), 4021-8.
- Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Jr, E. S., Kyrpides, N., Fonstein, M., Maltsev, N. & Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28**(1), 123-125.

- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**(8), 4285-8.
- Riley, M. (1993). Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**(4), 862-952.
- Riley, M. (1997). Genes and proteins of Escherichia coli K-12 (GenProtEC). *Nucleic Acids Res* **25**(1), 51-2.
- Riley, M. (1998). Genes and proteins of Escherichia coli K-12. *Nucleic Acids Res* **26**(1), 54.
- Schilling, C. H. & Palsson, B. O. (1998). The underlying pathway structure of biochemical reaction networks. *Proc Natl Acad Sci U S A* **95**(8), 4193-8.
- Schomburg, D., Salzmann, D. & Stephan, D. (1990-1995). *Enzyme handbook*. 13 vols, Springer.
- Schuster, S., Dandekar, T. & Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* **17**(2), 53-60.
- Schuster, S., Fell, D. A. & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks [see comments]. *Nat Biotechnol* **18**(3), 326-32.
- Schuster, S. & Hilgetag, C. (1994). On elemntary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **2**, 165-182.
- Seressiotis A & J.E., B. (1986). MPS: an algorithm and database for metabolic pathway synthesis. *Biotech Lett.* **8**, 837-842.
- Simpson, T. W., Colon, G. E. & Stephanopoulos, G. (1995). Two paradigms of metabolic engineering applied to amino acid biosynthesis. *Biochem Soc Trans* **23**(2), 381-7.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**(12), 3273-97.

- Takai-Igarashi, T. & Kaminuma, T. (1998). A pathway finding system for the cell signaling networks database. *in silico Biology*.
- Takai-Igarashi, T., Nadaoka, Y. & Kaminuma, T. (1998). A database for cell signaling networks. *J Comput Biol* **5**(4), 747-54.
- Tsoka, S. & Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion [In Process Citation]. *Nat Genet* **26**(2), 141-2.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**(6770), 623-7.
- Uetz, P. & Hughes, R. E. (2000). Systematic and large-scale two-hybrid screens. *Curr Opin Microbiol* **3**(3), 303-8.
- van Helden, J., Gilbert, D., Wernisch, L., Schroeder, M. & Wodak, S. (2001a). Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. *Lecture Notes in Computer Sciences* **2066**, 155-172.
- van Helden, J., Naim, A., Lemer, C., Mancuso, R., Eldridge, M. & Wodak, S. (2001b). From molecular activities and processes to biological function. *Briefings in Bioinformatics* **2**(1), 98-93.
- van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D. & Wodak, S. J. (2000). Representing and analysing molecular and cellular function using the computer. *Biol Chem* **381**(9-10), 921-35.
- Weng, G., Bhalla, U. & Iyengar, R. (1999). Complexity in Biological signalling systems. *Science* **284**, 92-96.
- Zien, A., Kuffner, R., Zimmer, R. & Lengauer, T. (2000). Analysis of gene expression data with pathway scores [In Process Citation]. *Ismb* **8**, 407-17.

Legends to figures

Figure 1. Alternative modes for mapping a metabolic network onto a graph. **A.** Typical drawing of a metabolic pathway, as can be found in biochemical

textbooks. Two alternative pathways for the conversion of L-homoserine into L-methionine. The number alongside each reaction represents the catalytic mechanism, and is the EC number. **B.** Each compound is represented as a node, and arcs indicate the conversions between compounds performed by reactions (as shown by the labels in italic). Note that each reaction is represented by a set of arcs. **C.** Each reaction is represented as a node, and arcs indicate the intermediate compounds (as shown by the arc labels in italic). Note that each compound is represented a set of arcs. **D.** A bipartite graph, where two distinct types of nodes are used to represent compounds and reactions (boxed labels) respectively.

Figure 2. Specific issues for path finding in metabolic graphs. **A.** Invalid intermediate compounds. **B-C.** Treatment of reaction reversibility. **B.** Using non-directed arcs would amount to considering that the substrates of a reaction can be transformed into each other in one step. **C.** Instantiating separate nodes for the direct and reverse reaction solves this problem, but direct and reverse nodes cannot be combined in the same path, because this would amount to transforming substrates into each other in two steps. See text for details.

Figure 3. Metabolic network connectivity as deduced from reactions in the KEGG database. **A.** Connectivity of compounds. The abscissa indicates the number of reactions in which a compound is involved as reactant, and the ordinate the number of compounds with this connectivity. **B.** Connectivity of reactions. The abscissa indicates the number of substrates, or products, of a reaction, and the ordinate the number of reactions with this connectivity. All reactants are counted, including ubiquitous compounds like H_2O. **C.** Reaction connectivity is strongly reduced when ubiquitous compounds are discarded from the count.

Figure 4. Metabolic pathway reconstitution by reaction clustering. We illustrate the concept by reconstructing *Escherichia coli* lysine pathway, starting from a subset of EC numbers. **A.** Lysine pathway in *Escherichia coli*. **B-D.** Pathways found by reaction clustering, starting from an unordered set of seed EC numbers (shaded boxes on the diagram). **B.** All EC numbers from pathway A are provided as seeds. **C.** Half of the EC numbers are provided as seeds. **D.** Four EC numbers are provided. **E.** Three EC numbers are provided.

Figure 5. Length of the shortest pathway between reactions catalyzed by **(A)** fusion partners and **(B)** random pairs of enzymes.

Tables

rank	description	product	substrate	sum	discarded
1	H2O	979	636	1615	1
2	NAD+	303	275	578	1
3	NADH	265	304	569	1
4	NADP+	314	250	564	1
5	NADPH	248	311	559	1
6	Oxygen	189	338	527	1
7	ATP	124	311	435	1
8	Orthophosphate	166	183	349	1
9	ADP	220	104	324	1
10	CO2	157	166	323	1
11	CoA	210	93	303	1
12	H+	130	142	272	
13	NH3	122	148	270	
14	Pyrophosphate	156	96	252	1
15	UDP	153	37	190	1
16	S-Adenosyl-L-methionine	29	145	174	
17	S-Adenosyl-L-homocysteine	134	31	165	
18	Pyruvate	33	117	150	
19	AMP	85	57	142	
20	H2O2	78	60	138	
21	L-Glutamate	56	76	132	
22	2-Oxoglutarate	49	80	129	
23	Acceptor	66	60	126	
24	Acetyl-CoA	24	98	122	
25	Reduced acceptor	56	66	122	1
26	Acetate	30	57	87	
27	UDPglucose	16	63	79	
28	D-Glucose	16	46	62	
29	Succinate	19	40	59	
30	CMP	31	23	54	

Table 1: The 30 compounds with the highest connectivity (calculated from KEGG data).