BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 351–361

Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae

S. Schuster[1], T. Pfeiffer[2], F. Moldenhauer[1], I. Koch[3] and T. Dandekar[4, 5]

[1]Max Delbrück Center for Molecular Medicine, Department of Bioinformatics, D-13092 Berlin-Buch, Germany, [2]Ecology and Evolution, ETH Zentrum, CH-8092 Zurich, Switzerland, [3]Max Planck Institute for Molecular Genetics, Department of Bioinformatics, D-14195 Berlin-Dahlem, Germany, [4]Biocomputing & Structures Program, EMBL, D-69012 Heidelberg, Germany and [5]University of Freiburg, Institute of Molecular Medicine, D-79106 Freiburg, Germany

Received on March 2, 2001; revised and accepted on September 14, 2001

ABSTRACT

Motivation: Reconstructing and analyzing the metabolic map of microorganisms is an important challenge in bioinformatics. Pathway analysis of large metabolic networks meets with the problem of combinatorial explosion of pathways. Therefore, appropriate algorithms for an automated decomposition of these networks into smaller subsystems are needed.

Results: A decomposition algorithm for metabolic networks based on the local connectivity of metabolites is presented. Interrelations of this algorithm with alternative methods proposed in the literature and the theory of small world networks are discussed. The applicability of our method is illustrated by an analysis of the metabolism of *Mycoplasma pneumoniae*, which is an organism of considerable medical interest. The decomposition gives rise to 19 subnetworks. Three of these are here discussed in biochemical terms: arginine degradation, the tetrahydrofolate system, and nucleotide metabolism. The interrelations of pathway analysis of biochemical networks with Petri net theory are outlined.

Availability: METATOOL is available from ftp://mudshark.brookes.ac.uk/pub/software/ibmpc or http://www.bioinf.mdc-berlin.de/metabolic/. The program SEPARATOR for decomposing metabolic networks is available from http://www.bioinf.mdc-berlin.de/metabolic/.

Supplementary information: http://www.bioinf.mdc-berlin.de/metabolic/metatool/ http://www.bork.embl-heidelberg.de/Annot/MP/ (re-annotation of *M. pneumoniae* genome)

Contact: dandekar@embl-heidelberg.de;
koch_i@molgen.mpg.de;
pfeiffer@eco.umnw.ethz.ch;
stschust@mdc-berlin.de

INTRODUCTION

Modern genome research is producing a huge body of information that has to be organized, classified and interpreted. In functional genomics, it is promising to include the proteomics level and consider the gene products in the functional context, for example, in the context of metabolic networks (cf. Bork *et al.*, 1998). It is common practice in biochemistry to discuss metabolism in terms of distinct biochemical pathways. In the light of recent challenges in bioinformatics, it has turned out to be instrumental to use the pathway concept (Selkov *et al.*, 1997; Bork *et al.*, 1998; Karp *et al.*, 1999) and phrase it in a formalized way (Mavrovouniotis *et al.*, 1990; Schuster and Hilgetag, 1994; Wagg and Sellers, 1997; Schuster *et al.*, 2000a; Schilling *et al.*, 2000). In particular, this is of importance for establishing and curating metabolic databases, such as KEGG (Kanehisa and Goto, 2000; http://www.genome.ad.jp/kegg/), WIT (http://wit.mcs.anl.gov/WIT2/), EcoCyc (Karp *et al.*, 2000; http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html) and EMP (Selkov *et al.*, 1996; http://wit.mcs.anl.gov/EMP).

It is a challenge to apply pathway analysis to the virtually complete metabolism of a microorganism the genome of which is completely known. Here, we present a structural analysis of the metabolism of *Mycoplasma pneumoniae*. This is an interesting organism from a medical point of view. It is a parasitic bacterium infesting the human respiratory tract and one of the causative agents of tracheobronchitis and atypical pneumonia (cf. Jacobs, 1997). Moreover, involvement in other conditions such as neurological infectious diseases (Greenlee and Rose, 2000) and thrombotic thrombocytopenic purpura (Bar Meir *et al.*, 2000) has been discussed. Besides these

medical reasons, we have chosen this species because it is one of the smallest self-replicating organisms. Its genome has been completely sequenced (Himmelreich *et al.*, 1996) and includes 688 protein ORFs, as has recently been shown by re-annotation (Dandekar *et al.*, 2000). *Mycoplasma genitalium* has an even smaller genome and all its genes are a subset of the *M. pneumoniae* set of genes. Thus, many of the conclusions of our analysis can be a basis for a comparison of the two organisms, by 'deleting' the enzymes absent in *M. genitalium*.

In contrast to the simulation of the metabolism of *M. genitalium* by the modelling package E-cell (Tomita *et al.*, 1999), it is not our aim to perform a dynamic simulation. We think that the knowledge of kinetic parameters is, at the present time, too limited to achieve this with reasonable accuracy. Rather, we follow a purely structural approach and analyze the topology of metabolic pathways. As opposed to dynamic simulation of metabolic networks, pathway analysis has the advantage that only the stoichiometric structure and preferably the directionality of reactions need to be known, rather than the kinetic parameters of enzymes.

Earlier, we introduced the concept of elementary flux mode (Schuster and Hilgetag, 1994; Pfeiffer *et al.*, 1999; Schuster *et al.*, 2000a). An elementary flux mode is a minimal set of enzymes that can operate at steady state with all irreversible reactions proceeding in the appropriate direction. In complex and dense networks, the computation of elementary modes often meets with the problem of combinatorial explosion. It is difficult to say in a general way how the number of modes changes as the number of reactions increases. For example, when in an unbranched reaction sequence, a reaction removing the product is added and the former product is now considered as an intermediate, the pathway is simply enlarged by one reaction. On the other hand, one can easily design hypothetical systems for which the number of pathways grows exponentially with increasing number of reactions (cf. Mavrovouniotis *et al.*, 1990).

As it is very difficult to interpret a huge number of elementary modes, it is often convenient to decompose large metabolic networks into smaller subsystems. Usually, this decomposition is performed on an intuitive basis. Biochemists discern distinct 'metabolisms': sugar metabolism, nucleotide metabolism, lipid metabolism, etc. For purposes of biocomputing, it would be helpful if this decomposition could be done in an automated way. Schilling and Palsson (2000) proposed to subdivide the network into relatively isolated clusters of reactions, according to intuitive biological criteria. Here, we will present an easy-to-implement decomposition algorithm based on network topology.

CONCEPTS AND TOOLS IN STRUCTURAL ANALYSIS

Besides the stoichiometry of reactions, information about the net direction of biochemical reactions is usually available, from the literature or metabolic databases. Thus, we can distinguish between irreversible and reversible reactions. Irreversibility is here not meant to exclude a reverse step, but this step should always have a lower rate than the forward reaction. Relative flux distributions fulfilling both a steady-state condition for intermediates and the sign restriction for irreversible reactions are called flux modes (Leiser and Blum, 1987; Schuster and Hilgetag, 1994). In order that a flux mode can be interpreted as a route through the network, it has to satisfy, in addition, a simplicity condition. This can be phrased in the following way: a flux mode is called elementary if a proper subset of the set of enzymes involved in the mode in question is not able to realize a flux mode. For a mathematical definition, see Heinrich and Schuster (1996) or Pfeiffer *et al.* (1999). The elementary modes correspond to the different basic functions the biochemical system is able to fulfil. This concept has been applied successfully in functional genomics, bioengineering and medicine (Liao *et al.*, 1996; Dandekar *et al.*, 1999; Schuster *et al.*, 1999, 2000a; Cornish-Bowden and Cárdenas, 2000; Rohwer and Hofmeyr, 2000).

The distinction between reversible and irreversible reactions is not always clear-cut because the directionality of a reaction depends on the concentrations of the reactants and products. This reasoning is important, for example, in the context of drug design, because inhibition of a reaction by a drug can lead to accumulation of the substrate and, hence, to a back-pressure effect on the preceding step. In such cases, the default option should be to take the reaction as reversible. Considering an irreversible reaction as reversible can only imply occurrence of additional elementary modes. The reason why all reactions which we certainly know are irreversible should be treated as irreversible is to reduce the number of modes.

Another classification necessary in pathway analysis concerns metabolites. A substance is called external if it can be considered to be present in large excess so that its concentration is virtually unaffected by the reactions under study. An internal metabolite (intermediate), however, must fulfil a balance equation implying that production equals consumption. This classification is not always clear-cut either. Fortunately, this ambiguity can be used for reducing combinatorial explosion, as will be outlined below.

An algorithm for computing elementary modes on the basis of methods from convex geometry was sketched in Schuster and Hilgetag (1994) and Pfeiffer *et al.* (1999) and given completely in Schuster *et al.* (2000a) (see http://bms-mudshark.brookes.ac.uk/algorithm.pdf for

exact formulas). This algorithm is performed by the program METATOOL written in C (Pfeiffer *et al.*, 1999).

A comparison between the concept of elementary modes and alternative approaches to defining the concept of biochemical pathway (e.g. Mavrovouniotis *et al.*, 1990) as well as related approaches in chemistry (e.g. Clarke, 1980) has been given in Schuster *et al.* (1999, 2000a,b). Recently, Schilling *et al.* (2000) proposed the concept of 'extreme pathway,' which is basically similar to the convex basis discussed in Pfeiffer *et al.* (1999). The convex basis includes a minimal set of flux modes from which, by non-negative linear combination, all admissible flux distributions at steady state can be obtained. The set of elementary flux modes comprises all modes of the convex basis and may include further modes that fulfil the above-mentioned simplicity condition. If all reactions are irreversible, then the convex basis coincides with the set of elementary modes. A special feature of the approach by Schilling *et al.* (2000) is that source and sink metabolites are formally considered as internal and exchange fluxes linking these substances with the surroundings are introduced. A distinction is made between these fluxes and the internal reactions, which are always taken to be irreversible.

Another property that can be derived from the structure of the metabolic network is the set of conservation relations (such as ATP + ADP = const.). These can be derived by determining the left-hand side nullspace of the stoichiometry matrix (Gavalas, 1968; cf. Érdi and Tóth, 1989; Heinrich and Schuster, 1996). This is performed by METATOOL, GEPASI (Mendes, 1997) and other biochemical simulation packages. As these relations are to reflect the conservation of chemical units, it is justified to invoke that all coefficients involved be non-negative. An algorithm for determining all non-negative conservation relations has been given (Schuster and Höfer, 1991). If all substances are involved in such relations, the system is called conservative (Horn and Jackson, 1972; Érdi and Tóth, 1989). This implies that a positive linear combination of all substance concentrations (e.g. total mass) is constant in time. If there is a positive linear combination that increases (decreases) in time, the system is called superconservative (subconservative; Érdi and Tóth, 1989). As biochemical networks are open systems which may, depending on conditions, have a positive or negative mass balance, they usually belong to none of these classes.

It has repeatedly been proposed to use the theory of Petri nets for modelling metabolism (e.g. Hofestädt, 1994; Reddy *et al.*, 1996; Heiner *et al.*, 2000, 2001; Küffner *et al.*, 2000). This is a method for modelling systems with concurrent processes. Petri nets are graphs with two different types of vertices: places and transitions (cf. Reisig, 1985; Starke, 1990), which, as for biochemical networks, correspond to metabolites and reactions, respectively. There are different types of tools which are able to describe systems either with discrete and continuous processes. Moreover, besides the so-called low-level Petri nets also high-level Petri nets were developed in order to be able to analyze and simulate more complex systems. The usual graphical representation of Petri nets is, however, uncommon to a biochemist's eye.

Table 1. Concepts used in the modelling of biochemical systems and their counterparts in Petri net theory

Modelling of biochemical systems	Petri net theory
Conservation relations	P-invariants
Semi-positive (non-negative) conservation relations	Semi-positive P-invariants
Conservative, subconservative and superconservative systems	Conservative, subconservative and su(pe)rconservative nets
Steady-state flux distributions	T-invariants
Elementary flux modes	Minimal T-invariants (less general since reversible transitions not allowed)

Interestingly, several concepts in the modelling of biochemical systems have counterparts in Petri net theory (Table 1). For example, there is a correspondence between elementary modes (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000a,b) and minimal T-invariants (Starke, 1990; Colom and Silva, 1990). However, transitions in Petri nets have hitherto been restricted to be unidirectional (irreversible). Thus, the minimal T-invariants are similar to the 'extreme currents' introduced by Clarke (1980). Reversible reactions would have to be described by two transitions in opposite directions. This has the drawback that all reversible reactions give rise to spurious minimal T-invariants consisting of the forward and reverse steps within one reaction. These invariants have to be discarded.

Colom and Silva (1990) presented several algorithms for computing the minimal T-invariants. One of these is used in the simulation package Integrated Net Analyser, (INA; http://www.informatik.hu-berlin.de/~starke/ina.html). In view of the above comparison of the two approaches (Table 1), the methods of Colom and Silva (1990) can also be employed for metabolic network analysis. The method for computing elementary modes (Schuster *et al.*, 2000a) is related to the algorithm implemented in INA. The former involves the important extension that reversible reactions need not be split into forward and reverse steps.

DECOMPOSITION PROCEDURE

As mentioned in the Section **Introduction**, in complex networks, often an enormous number of elementary modes

arises. Pathway analysis is cumbersome when the number of elementary modes is larger than the number of reactions (as sometimes happens in complex networks). A possible way of coping with this problem is by suitably classifying the metabolites according to whether or not they should be balanced with respect to production and consumption. A reasonable criterion for this classification is to minimize the number of elementary modes. This criterion is related to Kolmogorov complexity, which is defined as the length of the shortest program (algorithm) describing the system or process (cf. Varre *et al.*, 1999).

In general, it is very difficult to determine whether the number of elementary modes decreases or increases when an intermediate is changed to external metabolite status (that is, when it is considered as a source or sink). To show the intricacy of the problem, we will discuss the following situation. Suppose the number of modes increases if an intermediate S_1 is hypothetically considered external, and the same holds for an intermediate S_2. One might assume intuitively that, in this situation, the number of modes would certainly not decrease if both S_1 and S_2 were considered external. However, this is not generally true, as the counterexample depicted in Figure 1 demonstrates. This system, which is unlikely as a set of simple enzymatic reactions, but not unreasonable as a 'summary' of part of metabolism, gives rise to the following ten modes with S_1 and S_2 considered internal (the various reactions are denoted by R_j): $\{R_4, R_6\}$, $\{R_4, R_7\}$, $\{R_5, R_6\}$, $\{R_5, R_7\}$, $\{-R_2, 2R_3, R_6\}$, $\{-R_2, 2R_3, R_7\}$, $\{R_1, R_3, R_6\}$, $\{R_1, R_3, R_7\}$, $\{R_2, 2R_1, R_6\}$, $\{R_2, 2R_1, R_7\}$.

When S_1 is made external, 12 modes arise: $\{R_1\}$, $\{-R_2, R_3\}$, $\{-R_2, R_4\}$, $\{-R_2, R_5\}$, $\{R_2, R_6\}$, $\{R_2, R_7\}$, $\{R_3, R_6\}$, $\{R_3, R_7\}$, $\{R_4, R_6\}$, $\{R_4, R_7\}$, $\{R_5, R_6\}$, $\{R_5, R_7\}$. For symmetry reasons, 12 modes arise also in the situation where S_2 (but not S_1) is considered external. With both S_1 and S_2 considered external, each reaction represents an elementary mode on its own, so that we obtain seven modes and, hence, a smaller number than in the original system. This example shows that it would not be a good searching strategy to test particular metabolites and leave exactly those internal that increased the number of modes if made external.

If a sufficient number of metabolites are all considered as external in addition to the initial substrates and final products, the system 'disintegrates' into subsystems (see Figure 2). These are then delimited by the metabolites hypothetically considered external. Each internal metabolite belongs to one and only one of the subsystems. As no general algorithm for a convenient classification of internal and external metabolites is known, we established the following operational definition and decomposition procedure. In addition to the substances taken up and excreted by the cell, those metabolites that take part in more than a threshold number of reactions are considered external.

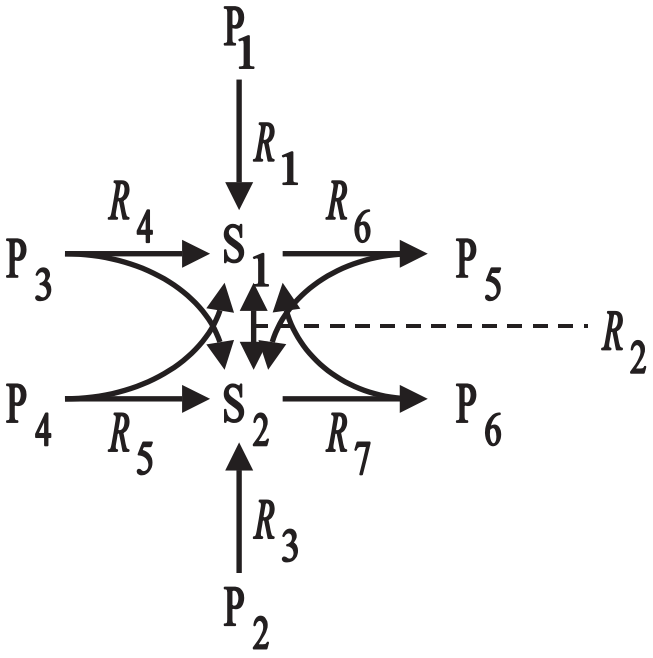

Fig. 1. Example system in which the number of modes depends in a counter-intuitive way on the choice of internal and external metabolites. P_i: external metabolites. S_i: optionally internal or external metabolites.

These branch-point metabolites are convenient points to cut a complex metabolic network into subnets which can be examined independently. For simplicity's sake, these metabolites are operationally regarded as buffered as they participate in many reactions and in several subnets. It is reasonable to fix the threshold to be four (Figure 2), but in some cases other values have turned out more useful in order not to obtain too large or too small subnetworks.

The motivation for the above-mentioned rule is illustrated in Figure 2. If the substance S shown in this figure is, for example, ATP, we could say that several ATP-producing pathways and ATP-consuming pathways can be combined in various ways. If S is involved in more than five reactions, the reduction of the number of pathways by treating S as external is even more sizeable. Of course, this heuristic reasoning has its limitations. First, the metabolite may be produced by one reaction and consumed by four reactions. On the other hand, if all reactions are reversible, then a reduction would be obtained even if the metabolite takes part in only four reaction sequences, because, with S being internal, six modes ($3 + 2 + 1 = 6$) would arise. Second, the number of routes depends not only on the local topology but also on the connectivity at more distant parts of the network.

Determination of the connectivity, that is, the number of reactions in which a metabolite participates, and the

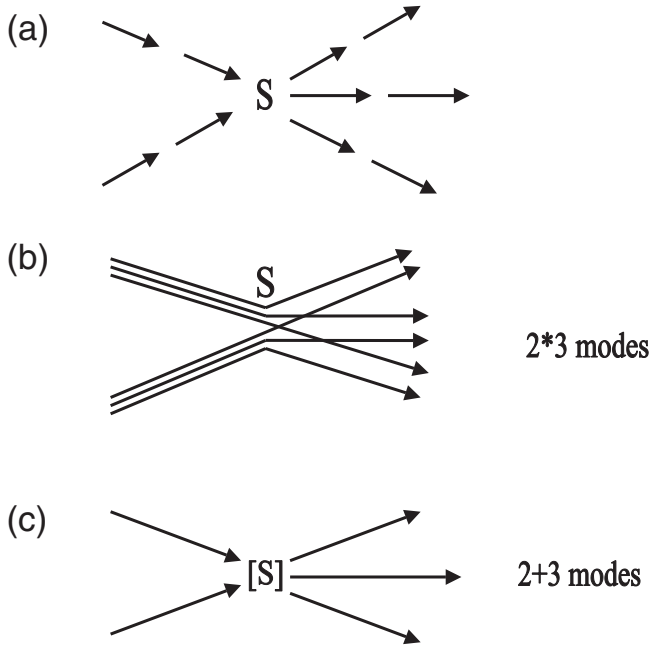

Fig. 2. (a) Schematic picture of a system in which a metabolite participates in two reaction sequences producing it and in three reaction sequences consuming it. (b) There are $2 \times 3 = 6$ possible routes formed by these reactions. (c) If, in contrast, the metabolite S is hypothetically considered as external, each reaction sequence forms a mode on its own, giving $2 + 3 = 5$ possibilities. Thus, to reduce the number of elementary modes, it is sensible to consider S as external.

decomposition into subnetworks (based on a connected component analysis algorithm) is performed automatically by the C program SEPARATOR written by one of the authors (T.P.). A routine for determining the connectivity in the complete network is also included in the most recent version of METATOOL. The program SEPARATOR parses a file with the syntax of METATOOL input files. It then generates the input files for METATOOL for each subnetwork. The threshold value for the connectivity can be specified by the user. Alternatively, the metabolites to be taken as externals can be defined. Before the program terminates, the user is asked whether the program is to be run again with other specifications or the result is to be saved. The running time for computing elementary modes (e.g. by METATOOL) depends so strongly on system size that it is easier to analyze several small systems than one big system. Accordingly, this time is considerably reduced when SEPARATOR is run first.

Considerable support to our heuristic decomposition rule is lent by the theory of scale-free networks (Watts and Strogatz, 1998; Jeong *et al.*, 2000). These are characterized by robustness and error-tolerance and their small-world properties, that is, any node (metabolite) is linked with any other node by a relatively short path. A method to distinguish them from randomly generated networks is by counting the number of reactions in which each metabolite participates and comparing the distribution of these numbers. The metabolites involved in a relatively large number of steps can be considered as 'hubs,' which mediate connections between remote parts of the network. These are usually cofactors such as ATP, ADP, or NAD or central branching points such as pyruvate or glutamate (Jeong *et al.*, 2000; Fell and Wagner, 2000), which are indeed usually regarded as externally buffered substances rather than as intermediates. A similar statistic has been presented by Ouzounis and Karp (2000).

To our eyes, the analysis of Jeong *et al.* (2000) is unnecessarily complicated because they introduce, for each enzymatic reaction, a temporary educt–educt complex in order to be able to tackle the resulting elementary reaction steps by graph theory. In fact, this is related to the Petri-net approach, in which two types of nodes are used. However, even for bimolecular reactions, the connectivity distribution can readily be determined by just counting the number of reactions in which each metabolite participates.

PATHWAY ANALYSIS OF *MYCOPLASMA PNEUMONIAE* METABOLISM

M. pneumoniae is a model organism for the determination of minimal genetic requirements of an autonomously reproducing cell because it has a very restricted metabolism. It does not involve any *de novo* synthesis of amino acids or nucleotides. For constructing the metabolic map of *M. pneumoniae*, data from the sequence and metabolism database KEGG (see Section Introduction) were downloaded by anonymous ftp. The list of enzymes obtained was cross-checked by data from the literature (e.g. Schimke, 1967; Maniloff *et al.*, 1992; Himmelreich *et al.*, 1996; Pollack *et al.*, 1997) and sequence analysis methods, including results from a systematic re-annotation effort of the *M. pneumoniae* genome (Dandekar *et al.*, 1999, 2000). We have neglected the synthesis and degradation of macromolecules because little information is available on this part of metabolism.

Two files were obtained from KEGG: one containing all EC numbers with the associated reactions and one with the EC numbers of all enzymes detected in *M. pneumoniae* so far. A problem arises with multifunctional enzymes, for which the former file contains, attached to the respective EC number, several reactions. As these were compiled by using information from many species, it is not clear whether all of them really occur in *M. pneumoniae*. We have used the following operational criterion, also readily applicable to new genomes. If a reaction contains a metabolite that does not occur in any other reaction

Table 2. List of acronyms

Acronym	Full name
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
C1	One carbon unit
CDP	Cytidine diphosphate
CMP	Cytidine monophosphate
CTP	Cytidine triphosphate
dCTP	Deoxycytidine triphosphate
DNA	Deoxyribonucleic acid
dTDP	Thymidine diphosphate
dTMP	Thymidine monophosphate
dUMP	Deoxyuridine monophosphate
dUTP	Deoxyuridine triphosphate
F6P	Fructose 6-phosphate
(f-)Met	(Formyl-)methionine
GA3P	Glyceraldehyde-3-phosphate
NAD(H)	Oxidized (reduced) form of nicotinamide adenine dinucleotide
NADP(H)	Oxidized (reduced) form of nicotinamide adenine dinucleotide phosphate
P_i	Inorganic phosphate
PP_i	Inorganic pyrophosphate
PPP	Pentose phosphate pathway
PRPP	Phosphoribosylpyrophosphate
R5P	Ribose 5-phosphate
UDP	Uridine diphosphate
UMP	Uridine monophosphate
UTP	Uridine triphosphate

Table 3. Connectivity of the most frequent metabolites in *M. pneumoniae* and comparison with results of Jeong *et al.* (2000)[a]

Substance	Connectivity	Hub(in)	Hub(out)
ATP	22	+	+
ADP	21	+	+
H_2O	19	+	+
P_i	18	+	+
CMP	10	+	−
PP_i	10	+	+
NADP	9	−	−
NADPH	9	−	−
H^+	9	−	−
UMP	9	+	−
GA3P	8	−	−
NAD	8	−	−
Uridine	8	−	+
Cytidine	8	−	+
NADH	7	−	−
5,6,7,8-tetrahydrofolate	6	−	−
7,8-dihydrofolate	6	−	−
PRPP	6	−	−

[a]We define connectivity as the number of enzymatic reactions in which a metabolite participates. The columns entitled Hub(in) and Hub(out) indicate whether the metabolites have been classified by Jeong *et al.* (2000) as such (+) or not (−) for *M. pneumoniae*. A metabolite is classified as a hub(in) or hub(out) if it belongs to the ten substrates with the largest number of incoming or outgoing links, respectively. In addition to the indicated metabolites, the following hubs(in) have been given by Jeong *et al.* (2000): pyruvate, phosphoenolpyruvate, NH_4^+, and the following hubs(out): pyruvate, phosphoenolpyruvate, glucose (which participate in less than six reactions according to the reaction table used in our analysis).

potentially occurring in *M. pneumoniae*, and is not taken up or excreted by the cell, this reaction is deleted.

The classification of reversible versus irreversible enzymes was mainly performed according to the information given in the metabolic maps in KEGG. However, for some enzymes, different directionalities are given on different maps. This need not be an inconsistency because the same enzyme can serve different functions in distinct parts of metabolism. For example, thymidylate synthase (EC 2.1.1.45) is indicated as being irreversible in the direction of dTMP synthesis on the map 'Pyrimidine metabolism', while it is indicated as being reversible on the map 'One carbon pool by folate' (for an explanation of acronyms such as dTMP, see Table 2). In cases where database information is inconsistent or insecure, one can consider the reaction as reversible without losing any pathway (cf. Section Concepts and Tools). In the system under study, however, we treat thymidylate synthase as irreversible because the biological function of this enzyme in *M. pneumoniae* is clearly dTMP synthesis for incorporation into DNA.

Although the metabolism of *M. pneumoniae* is relatively small in comparison to *Escherichia* or *Salmonella* species, it is so large that an enormous number of pathways arise. Therefore, it is sensible to decompose the metabolic map into smaller subsystems, as outlined in the previous section. Table 3 shows the ranking of the metabolites that participate in more than five reactions. It should be noted that this ranking depends on how multifunctional enzymes are treated. As our operational criterion for selecting the relevant reactions for such enzymes may fail in some cases, the ranking can certainly be improved as soon as further biochemical knowledge about *M. pneumoniae* becomes available. This also explains the deviations from the data given by Jeong *et al.* (2000) (see Table 3).

Our calculations confirm the general result of Jeong *et al.* (2000) in that the metabolism of *M. pneumoniae* can be regarded as a scale-free network. From Table 3, it can be seen that there is one metabolite adjacent to 22 links and one adjacent to 21 links. The other end of the distribution of connectivities is formed by 42 metabolites that participate in two reactions each and 60 metabolites that participate in one reaction each (not shown). Fitting the entire distribution in the log–log space by a linear function gives a correlation coefficient of $r = -0.9393$, so that the statistical correlation according to a power law is highly significant ($p < 1\%$). Translation back into original variables yields the power law $P(k) =$

$75.34k^{-1.49}$ with k and P denoting the number of links and frequency, respectively.

Applying the decomposition procedure outlined in the previous section and cancelling all subnetworks that only involve external metabolites, 19 subnetworks are obtained (Pfeiffer, 1999). It has turned out that the biochemical interpretability of results is considerably improved by slightly editing the automated classification, based on biochemical knowledge. For example, in some subsystems, it is useful to combine the reactions using NADP with those using NAD, in order to avoid combinatorial explosion. Moreover, for enzymes with broad specificity, the functions with low activities have been deleted. For example, uridine kinase (EC 2.7.1.48) can use a wide spectrum of nucleotide triphosphates. However, CTP, dCTP, UTP and dUTP are very poor substrates.

The six most important subnetworks correspond to sugar import, glycolysis plus pentose phosphate pathway plus fragmentary lipid metabolism, lower part of glycolysis, nucleotide interconversion, one-carbon unit pool and arginine degradation (Figure 3). For all 19 subnetworks, the elementary modes can be computed easily. Their number is less than 20 per subnetwork and, hence, small enough to be tractable. To illustrate the biochemical relevance of the method, we will here give an overview of the interpretation of the results for three subsystems. A more detailed discussion of all subsystems in terms of functional genomics will be given elsewhere.

C1 pool

The tetrahydrofolate system, which is capable of transferring one-carbon (C1) units, gives rise to five elementary modes, which we have discussed earlier (Pfeiffer *et al.*, 2000). Two modes transfer a C1 unit to dUMP, thus producing dTMP, which is (in Subsystem 3) phosphorylated to give dTDP. dUMP is produced from uracil. The C1 group comes either from formate or from serine. The function of these modes is to supply thymidine nucleotides for DNA synthesis. The redundancy implied by the occurrence of two modes (as well as the redundancy in the synthesis of formyl-methionine in this subsystem, cf. Pfeiffer *et al.* (2000)) has to be taken into account in drug design. Formate-tetrahydrofolate ligase (EC 6.3.4.3) would not be a suitable drug target because it can be bypassed via a mode involving glycine hydroxymethyltransferase (EC 2.1.2.1).

Arginine degradation

Although many *Mycoplasma* species take up arginine from the host and degrade it to produce ATP, it is a matter of debate whether *M. pneumoniae* does so as well. While Schimke (1967) did not find sufficient biochemical evidence for this pathway, the three enzymes necessary for it have been spotted in the genome: arginine deiminase (EC 3.5.3.6), ornithine carbamoyltransferase (EC 2.1.3.3) and carbamate kinase (EC 2.7.2.2) (Himmelreich *et al.*, 1996). We have found that, based on the sequence data, the arginine degradation system involves three elementary modes provided that a consumption reaction for ornithine is included. Two modes do indeed represent ATP synthesis from ADP by degrading arginine via carbamoyl phosphate, either to ornithine and carbamate or to ornithine, CO_2 and NH_3. The third mode represents the degradation of carbamate to CO_2 and NH_3 via carbamoyl phosphate.

Nucleotide metabolism

Although *M. pneumoniae* is not able to synthesize nucleotides *de novo*, it harbours a considerable number of enzymes converting nucleotides. Several of them can convert a wide spectrum of substrates. For example, uridine kinase (EC 2.7.1.48) catalyzes 18 reactions, for example ATP + uridine $\leftrightarrow$ ADP + UMP. In the subnetwork of nucleotide metabolism, several modes producing nucleotide diphosphates from purine and pyrimidine bases have been found. For example, UDP is formed by two different modes (Figure 4). The ATP consumed in these modes is regenerated by glycolysis or arginine degradation. The formation of nucleotide diphosphates is physiologically meaningful because the nucleotide bases can be taken up from the host while the nucleoside phosphates can probably not. These modes are related to the salvage pathways (cf. Stryer, 1995).

The modes in nucleotide metabolism exhibit considerable redundancy (Figure 4) and, consequently, allow faster replication of this intracellular parasite. The design of a bacteriostatic drug can exploit this analysis, because the three enzymes shared by the two modes shown in Figure 4 would be candidate drug targets, while adenine phosphoribosyltransferase would not because there is an alternative pathway circumventing it. Moreover, it can be seen that the molar yield with respect to ATP is different for the two pathways shown in Figure 4. While the first mode consumes 3 moles of ATP per mole of UDP produced, the second mode uses 1,5 moles of ATP. From a bioenergetic point of view, it is not, however, better than the first mode because one third of the ATP molecules used is degraded down to adenine rather than ADP.

The formation of CDP is less clear. There is CDP-diglyceride synthetase and phosphatidylglycerophosphate synthase in *M. pneumoniae*, but none of them would yield or use free CDP. Instead, they are involved in phospholipid synthesis. Uridine kinase (which can use, alternatively, cytidine) and cytidylate kinase (EC 2.7.4.14) could produce CDP from cytidine. However, it is unclear where the latter metabolite comes from. Himmelreich *et al.* (1996) have postulated a cytidine-producing pathway via UMP and uridine, involving uridine kinase and cytidine

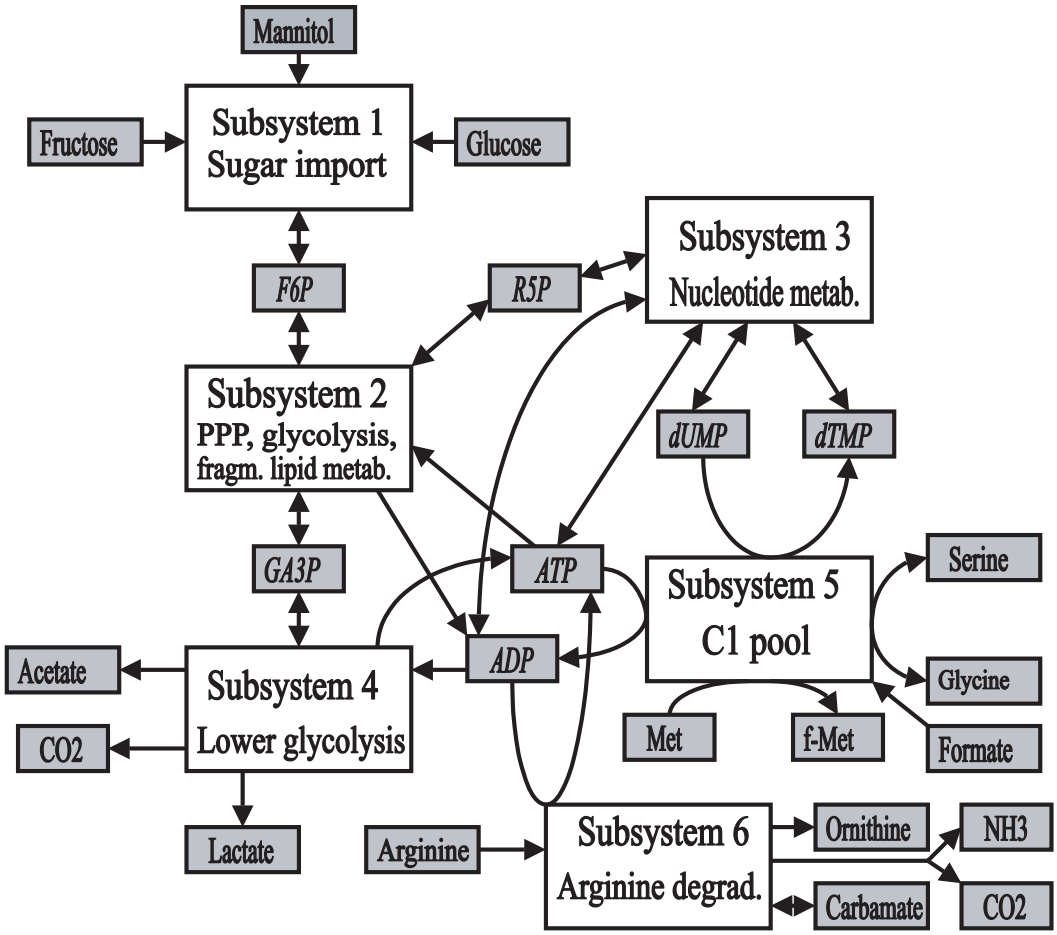

Fig. 3. Subnetwork structure of *M. pneumoniae* metabolism. Metabolites with italicized symbols have been declared external in the decomposition algorithm, all other metabolites shown are external because they are taken up or excreted by the cell or incorporated into macromolecules. Redox equivalents such as NADP are not shown.

deaminase (EC 3.5.4.5), both used in the reverse direction. However, these enzymes are given in the databases as irreversible. By declaring them (hypothetically) reversible in the input file to METATOOL, we obtain several modes producing CDP from uracil.

CONCLUSIONS

Pathway analysis helps elucidate the complex architecture of cell physiology by detecting the basic functional units (pathways). Although each of these units can be understood much more easily than the entire metabolic map of an organism, one is often faced with the problem that in complex, dense networks an enormous number of pathways exist. This is not a flaw of pathway analysis itself, but is rather due to the limited capacity of the human mind, which cannot survey too many items. To improve our understanding of the results of pathway analysis, it is therefore advantageous to decompose the metabolic map into smaller, manageable subsystems. Here, we have presented a decomposition procedure based on local connectivity of metabolites. An implementation in C (program SEPARATOR) is available from http://www.bioinf.mdc-berlin.de/metabolic/. By choosing the threshold value appropriately, the size of the subsystems can be regulated.

Our connectivity analysis is supported by an approach presented by Jeong *et al.* (2000), who were able to show that metabolic networks usually behave as scale-free networks. They differ from randomly generated networks or very regular networks (such as a fishermen's net) by the fact that the distribution of numbers of reactions in which a metabolite participates decays according to a power law. From an evolutionary viewpoint, this can be understood by assuming that newly recruited reactions preferably utilize or produce metabolites that have had a large connectivity already. We simplified the approach of Jeong *et al.*

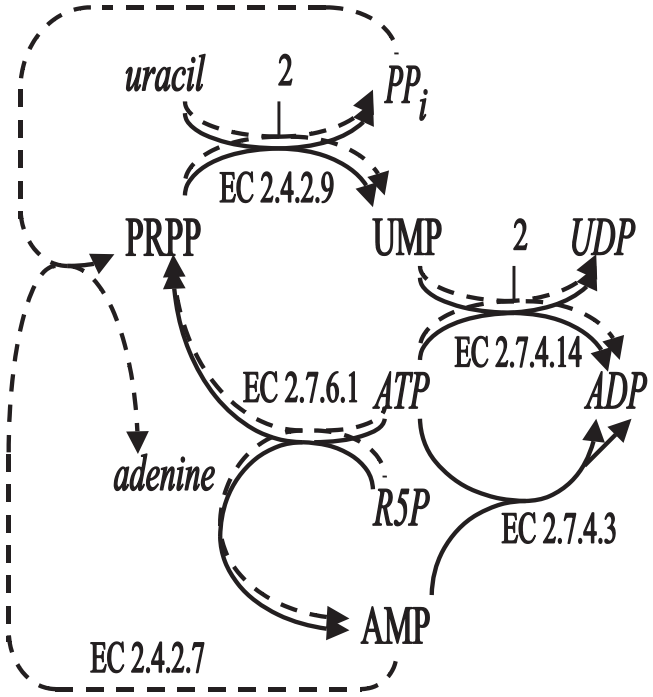

Fig. 4. Two elementary modes producing UDP from uracil. Italicized symbols stand for external metabolites. The lists of enzymes involved in these modes are {cytidylate kinase (EC 2.7.4.14, acting here on UMP), uracil phosphoribosyltransferase (EC 2.4.2.9), adenylate kinase (EC 2.7.4.3), ribose-phosphate pyrophosphokinase (EC 2.7.6.1)} (solid arrows), {2 cytidylate kinase (acting here on UMP), 2 uracil phosphoribosyltransferase, adenine phosphoribosyltransferase (EC 2.4.2.7), ribose-phosphate pyrophosphokinase} (dashed arrows). The factors 2 in these lists and in the figure signify that the flux carried by the respective enzymes is double the flux through the other reactions. The overall stoichiometries for the two modes read:
$3\mathrm{ATP} + \mathrm{R5P} + \mathrm{uracil} = 3\mathrm{ADP} + \mathrm{PP}_i + \mathrm{UDP}$;
$3\mathrm{ATP} + \mathrm{R5P} + 2\mathrm{uracil} = 2\mathrm{ADP} + \mathrm{PP}_i + 2\mathrm{UDP} + \mathrm{adenine}$.

(2000) in some instances. We do not consider temporary educt–educt complexes and do not decompose reactions into forward and reverse steps. Moreover, we do not distinguish between hubs with respect to incoming reactions and those with respect to outgoing reactions. While Jeong *et al.* (2000) classify the top ten metabolites as hubs, we determine the metabolites forming the boundaries of subnetworks by comparing their connectivity with a threshold value. Thus, each metabolite can be classified even if not all other metabolites have been checked yet.

Our method is based solely on network topology and does not, *a priori*, include any biological bias. As we have shown by applying the method to the metabolic map of *M. pneumoniae*, it results in a decomposition that is in agreement with biochemical intuition and practice. For example, subsystems such as nucleotide metabolism and arginine degradation are obtained. The elementary modes within the subsystems can readily be obtained, are easy to survey and can be interpreted in biochemical terms.

Schilling and Palsson (2000) have demonstrated earlier that a pathway analysis of a substantial part of microbial metabolism is feasible, studying *Haemophilus influenzae* (for a review of the merits of that work, see Schuster, 2000). They subdivided the metabolic network of *H. influenzae* into six subsystems motivated by biological reasoning.

Interestingly, several concepts developed in pathway analysis have been established independently in Petri net theory (cf. Reisig, 1985; Starke, 1990). As the dynamic modelling of biochemical systems usually involves continuous variables, hybrid and continuous Petri net models are required for this type of modelling (Alla and David, 1998; Matsuno *et al.*, 2000). For the computation of invariants, discreteness is not a problem, though, as long as the stoichiometric coefficients are integers, because the invariants can then be scaled so as to involve integers as well (Heiner *et al.*, 2000, 2001). It is worthwhile investigating the use of Petri net theory for 'classical', algebraic biochemical modelling and *vice versa*(!) in more detail in the future. This concerns, amongst others, routines for an automated decomposition of large networks. It is desirable that such a routine uses not only the 'local' information on the number of reactions in which a metabolite is involved but also 'global' information about far-reaching interactions in the network. For example, it might be helpful to adapt clustering methods from protein structure analysis (cf. Patra and Vishveshwara, 2000). Another interesting idea is to detect central points in metabolism which are connected with all other metabolites by a minimum number of links (Fell and Wagner, 2000).

It would be favourable if, upon re-composition of the subnetworks, the elementary modes can simply be combined with each other (cf. Schilling and Palsson, 2000). While this is feasible in the system shown in Figure 2, it is not in the system depicted in Figure 1. When S_1 and S_2 are treated as external, seven subsystems made up of the particular reactions arise. The modes in the original system are not, however, pair-wise combinations of these reactions.

The application of pathway analysis to *M. pneumoniae* demonstrates the versatility and usefulness of the elementary-modes approach in the analysis of larger systems. The metabolic capabilities of substrate–product conversion can be derived. Moreover, alternative routes producing the same product can be compared with respect to their molar yield, as in the example of UDP synthesis. (For a general discussion of this point, see Schuster *et al.*, 1999, 2000a,b.) The analysis is also helpful in the detection of appropriate (or inappropriate) drug targets.

When there are bypasses in the network, as in the case of formate tetrahydrofolate ligase, suppressing the enzyme by a drug will have a very limited effect. Importantly, glycine hydroxymethyltransferase cannot be considered as a simple substitute of tetrahydrofolate ligase because also other enzymes are involved (see Section Pathway Analysis of *Mycoplasma pneumoniae* Metabolism). Thus, pathway analysis is needed to decide which bypasses exist in the system.

In several of the subnetworks in *M. pneumoniae*, many enzymes do not enter any mode. This is indicative of missing links in the functional assignment of ORFs. Thus, gaps in the functional reconstruction of the metabolic map based on the annotated databases can be detected. For example, the route of synthesis of CDP proposed by Himmelreich *et al.* (1996) is very unlikely for thermodynamic reasons according to our analysis. Pathway analysis can be used as a guideline in filling these gaps (Dandekar *et al.*, 1999; Schuster *et al.*, 1999; Schilling and Palsson, 2000). This tool should be used in conjunction with other methods such as sequence comparison, biochemical assays, and metabolic control analysis.

ACKNOWLEDGEMENTS

We would like to thank Drs David Fell (Oxford), Peer Bork (Heidelberg) and Claus Hilgetag (Boston) for stimulating discussions and two anonymous referees for helpful comments. Research was funded by the Deutsche Forschungsgemeinschaft (SFB 544/B2, SPP 'Computer science methods for the analysis of large genomic data sets,' and Heisenberg Program).

REFERENCES

Alla,H. and David,R. (1998) Continuous and hybrid petri nets. *J. Circuits Syst. Comput.*, **8**, 159–188.

Bar Meir,E., Amital,H., Levy,Y., Kneller,A., Bar-Dayan,Y. and Shoenfeld,Y. (2000) *Mycoplasma-pneumoniae*-induced thrombotic thrombocytopenic purpura. *Acta Haematol.*, **103**, 112–115.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Clarke,B.L. (1980) Stability of complex reaction networks. *Adv. Chem. Phys.*, **43**, 1–216.

Colom,J.M. and Silva,M. (1990) Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computation of minimal P-semiflows. In Rozenberg,G. (ed.), *Advances in Petri Nets*. Springer, Berlin, pp. 79–112.

Cornish-Bowden,A. and Cárdenas,M.L. (2000) From genome to cellular phenotype-a role for metabolic flux analysis? *Nature Biotechnol.*, **18**, 267–268.

Dandekar,T., Schuster,S., Snel,B., Huynen,M. and Bork,P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, **343**, 115–124.

Dandekar,T., Huynen,M., Regula,J.T., Ueberle,B., Zimmermann,C.U., Andrade,M.A., Doerks,T., Sanchez-Pulido,L., Snel,B., Suyama,M., Yuan,Y.P., Herrmann,R. and Bork,P. (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res.*, **28**, 3278–3288.

Érdi,P. and Tóth,J. (1989) *Mathematical Models of Chemical Reactions*. Manchester University Press, Manchester.

Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nature Biotechnol.*, **18**, 1121–1122.

Gavalas,G.R. (1968) *Nonlinear Differential Equations of Chemically Reacting Systems*. Springer, Berlin.

Greenlee,J.E. and Rose,J.W. (2000) Controversies in neurological infectious diseases. *Semin. Neurol.*, **20**, 375–386.

Heiner,M., Koch,I. and Schuster,S. (2000) Using time-dependent petri nets for the analysis of metabolic networks. In Hofestädt,R,, Lautenbach,K. and Lange,M. (eds), *Modellierung und Simulation Metabolischer Netzwerke*, Preprint no. 10, Faculty of Computer Science, University of Magdeburg, pp. 15–21.

Heiner,M., Koch,I. and Voss,K. (2001) Analysis and simulation of steady states in metabolic pathways with Petri nets. In Jensen,K. (ed.), *CPN '01—Third Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools*. University of Aarhus, Denmark, pp. 15–34.

Heinrich,R. and Schuster,S. (1996) *The Regulation of Cellular Systems*. Chapman and Hall, New York.

Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–4449.

Hofestädt,R. (1994) A petri net application to model metabolic processes. *Syst. Anal. Modelling Simul.*, **16**, 113–122.

Horn,F. and Jackson,R. (1972) General mass action kinetics. *Arch. Rational Mech. Anal.*, **47**, 81–116.

Jacobs,E. (1997) *Mycoplasma* infections of the human respiratory tract. *Wien. Klin. Wochenschr.*, **109**, 574–577.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Karp,P.D., Krummenacker,M., Paley,S. and Wagg,J. (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.*, **17**, 275–281.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Paley,S.M. and Pellegrini-Toole,A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, **28**, 56–59.

Küffner,R., Zimmer,R. and Lengauer,T. (2000) Pathway analysis in metabolic databases via Differential Metabolic Display (DMD). *Bioinformatics*, **16**, 825–836.

Leiser,J. and Blum,J.J. (1987) On the analysis of substrate cycles in large metabolic systems. *Cell Biophys.*, **11**, 123–138.

Liao,J.C., Hou,S.Y. and Chao,Y.P. (1996) Pathway analysis, engineering and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.*, **52**, 129–140.

Maniloff,J., McElhaney,R.N., Finch,L.R and Baseman,J.B. (1992) *Mycoplasmas. Molecular Biology and Pathogenesis*. ASM, Washington.

Matsuno,H., Doi,A., Nagasaki,M. and Miyano,S. (2000) Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomput.*, **5**, 338–349.

Mavrovouniotis,M.L., Stephanopoulos,G. and Stephanopoulos,G. (1990) Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.*, **36**, 1119–1132.

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

Ouzounis,C.A. and Karp,P.D. (2000) Global properties of the metabolic map of *Escherichia coli*. *Genome Res.*, **10**, 568–576.

Patra,S.M. and Vishveshwara,S. (2000) Backbone cluster identification in proteins by a graph theoretical method. *Biophys. Chem.*, **84**, 13–25.

Pfeiffer,T. (1999) *Diploma Thesis*, Humboldt University, Berlin.

Pfeiffer,T., Sánchez-Valdenebro,I., Nuño,J.C., Montero,F. and Schuster,S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics*, **15**, 251–257.

Pfeiffer,T., Dandekar,T., Moldenhauer,F. and Schuster,S. (2000) Topological analysis of metabolic networks. Application to the metabolism of *Mycoplasma pneumoniae*. In Hofmeyr,J.-H.S., Rohwer,J.M. and Snoep,J.L. (eds), *BioThermoKinetics 2000*, Animating the Cellular Map, University Press, Stellenbosch, pp. 229–234.

Pollack,J.D., Williams,M.V. and McElhaney,R.N. (1997) The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.*, **23**, 269–354.

Reddy,V.N., Liebmann,M.N. and Mavrovouniotis,M.L. (1996) Qualitative analysis of biochemical reaction systems. *Comput. Biol. Med.*, **26**, 9–24.

Reisig,W. (1985) *Petri Nets: An Introduction*. Springer, Berlin.

Rohwer,J.M. and Hofmeyr,J.H.S. (2000) An integrated approach to the analysis of the control and regulation of cellular systems. In Cornish-Bowden,A. and Cárdenas,M.L. (eds), *Technological and Medical Implications of Metabolic Control Analysis*. Kluwer, Dordrecht, pp. 73–79.

Schilling,C.H. and Palsson,B.O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, **203**, 249–283.

Schilling,C.H., Letscher,D. and Palsson,B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.

Schimke,R.T. (1967) Studies on the metabolism of arginine by *Mycoplasma*. *Ann. New York Acad. Sci.*, **143**, 543–547.

Schuster,S. (2000) Biotechnology *in silico*. Metabolic proteomics of *Haemophilusinfluenzae* (commentary). *Trends Biotechnol.*, **18**, 328.

Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.

Schuster,S. and Höfer,T. (1991) Determining all extreme semi-positive conservation relations in chemical reaction systems. A test criterion for conservativity. *J. Chem. Soc. Faraday Trans.*, **87**, 2561–2566.

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

Schuster,S., Fell,D.A. and Dandekar,T. (2000a) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.*, **18**, 326–332.

Schuster,S., Dandekar,T., Mauch,K., Reuss,M. and Fell,D. (2000b) Recent developments in metabolic pathway analysis and their potential implications for biotechnology and medicine. In Cornish-Bowden,A. and Cárdenas,M.L. (eds), *Technological and Medical Implications of Metabolic Control Analysis*. Kluwer, Dordrecht, pp. 57–66.

Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E., Pronevitch,L., Selkov,E. Jr and Yunus,I. (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Res.*, **24**, 26–28.

Selkov,E., Maltsev,N., Olsen,G.J., Overbeek,R. and Whitman,W.B. (1997) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene*, **197**, GC11–GC26.

Starke,P.H. (1990) *Analyse von Petri-Netz-Modellen*. Teubner, Stuttgart.

Stryer,L. (1995) *Biochemistry*. Freeman, New York.

Tomita,M., Hashimoto,K., Takahashi,K., Shimizu,T.S., Matsuzaki,Y., Miyoshi,F., Saito,K., Tanida,S., Yugi,K., Venter,J.C. and Hutchison,C.A. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics*, **15**, 72–84.

Varre,J.S., Delahaye,J.P. and Rivals,E. (1999) Transformation distances: a family of dissimilarity measures based on movements of segments. *Bioinformatics*, **15**, 194–202.

Wagg,J. and Sellers,P. (1997) Enumeration of flux routes through complex biochemical reactions. *Proc. Pac. Symp. Biocomput.*, 453–464.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.