"Insight Box":

INSIGHT: We present an upgraded bioreaction database useful for the reconstruction of metabolic networks. Apart from necessary updates due to progress in research and error correction, the database incorporates structural improvements and revised criterions, such as currency metabolites, reversibility information, reactant pairs, non-enzymatic spontaneous reactions, balanced stoichiometry, and glycans. We combine an automatic approach with manual curation in order to make the reconstructed metabolic networks more accurate and more reliable.
For evaluating this database, we encountered the problem of finding biologically feasible shortest paths, which turned out to be hard to compute.

INNOVATION: We present new methods to increase the quality of bioreaction databases.

INTEGRATION: We use methods from computer science to find biologically feasible paths in reconstructed metabolic networks.

An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks

Michael Stelzer[1], Jibin Sun[2], An-Ping Zeng[3], Tom Kamphans[4], and Sándor Fekete[4]

[1]Department of Bioinformatics and Biochemistry, Braunschweig University of Technology, Langer Kamp 19B, 38106 Braunschweig, Germany

[2]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, 32XiQiDao, Tianjin Airport Economic Park, Tianjin, China

[3]Institute of Bioprocess and Biosystems Engineering, Hamburg University of Technology, Denickestr. 15, 21073 Hamburg, Germany

[4]Department for Computer Science, Algorithms Group, Braunschweig University of Technology, Mühlenpfordtstr. 23, 38106 Braunschweig, Germany

ABSTRACT

The bioreaction database established by Ma and Zeng (Bioinformatics 2003. 19, 270-277) for *in silico* reconstruction of genome-scale metabolic networks has been widely used. Based on more recent information in the reference databases *KEGG LIGAND* and *BRENDA*, we upgrade the bioreaction database in this work by almost doubling the number of reactions from 3565 to 6851. Over 70 % of the reactions have been manually updated/revised in terms of reversibility, reactant pairs, currency metabolites and error correction. For the first time, 41 spontaneous sugar mutarotation reactions are introduced into the biochemical database. The upgrade significantly improves the reconstruction of genome scale metabolic networks. Many gaps or missing biochemical links can be recovered, as exemplified with three model organisms *Homo sapiens*, *Aspergillus niger,* and *Escherichia coli*. The topological parameters of the constructed networks were also largely affected, however, the overall network structure remains scale-free.
Furthermore, we consider the problem of computing biologically feasible shortest paths in reconstructed metabolic networks. We show that these paths are hard to compute and present solutions to find such paths in networks of small and medium size.

Availability:The upgraded version of the bioreaction database and supporting tools and materials are available from our website: http://www.tuhh.de/ibb.

KEYWORDS: Bioreaction database, network reconstruction and analysis, currency metabolite, metabolic network, reactant pair, shortest path

Contact: aze@tu-harburg.de, s.fekete@tu-bs.de

1 INTRODUCTION

Reconstruction of genome-scale metabolic networks has become a powerful tool for biological studies in recent years (Francke *et al.*, 2005), mainly due to the rapid development of genome sequencing technology (Margulies *et al.*, 2005). Network-based approaches can help to improve annotation of genome sequence, to interpret high-throughput omics data, to understand biological processes of pathogenesis or industrial production at a system level, and to rationally control or design biological systems (Duarte *et al.*, 2004; Famili *et al.*, 2003; Junker *et al.*, 2006; Klukas *et al.*, 2006; Nacher *et al.*, 2006; Rahman and Schomburg, 2006; Sun *et al.*, 2007). A prerequisite for a reliable reconstruction of metabolic networks is the availability of a bioreaction database that should cover as much as possible information on biochemical reactions. Such a database should also allow for unambiguous and biochemically or physiologically meaningful connections among the reactant pairs for functional analysis.

There are many bioreaction databases which can be helpful for metabolic reconstruction and analysis, including the well-known *Roche* wall chart of *Biochemical Pathways* (Michal and Schomburg, 2010), the *BioCyc* database and its *Pathway Tool* (Caspi *et al.*, 2010), the *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) databases including *KEGG PATHWAY* for maps of biological processes and *KEGG LIGAND* for chemical compounds, drugs, glycans and reactions (Kanehisa *et al.*, 2010), and the *BRaunschweig ENzyme DAtabase BRENDA* (Scheer *et al*., 2010). It is also possible to combine databases in an integrative manner for the reconstruction of metabolic networks to increase reliability (Radrich *et al.*, 2010).

However, the results by applying these databases must be carefully evaluated. For example, the path from glucose to pyruvate *via* the glycolysis pathway was once calculated as two steps by considering ADP as a conversion hub (Jeong *et al.*, 2000), which is obviously physiologically not meaningful (Ma and Zeng, 2003 a).

Ma and Zeng (2003 a) developed a bioreaction database based on the *KEGG LIGAND* database, and demonstrated its usefulness by reconstruction of high-resolution metabolic networks for over 80 organisms. In this database, Ma and Zeng defined the reversibility of the reactions according to literature data and biochemistry knowledge and introduced the concept of reactant pairs by considering the currency metabolites. The latter feature was also adopted by recent versions of *KEGG LIGAND* (Kotera *et al.*, 2004). All these features are important to reconstruct the metabolic network and allow a more proper analysis of network properties such as connectivity, shortest pathway length and modularity. The reaction database of Ma and Zeng has been widely used by different authors.

On the other hand, since the release of the bioreaction database of Ma and Zeng (2003 a), the *KEGG LIGAND* database has been improved significantly. For example, the total number of reactions in the *LIGAND* database increased from 3565 (status of Sept 2003, Release 27.0) to 6851 (status of Dec 2006, Release 40.0), 73 % of reactions were updated in terms of the reaction equations or corresponding enzymes. Since 2004, *KEGG* introduced reactant pairs into the *LIGAND* database as well and categorized them as five types: *main*, *cofac*, *trans*, *ligase,* and *leave* (Kotera *et al.*, 2004). However, there are still shortcomings in the *LIGAND* database. For example, the reactant pairs are not comparable to the pairs we provided by considering the currency metabolites which was demonstrated to be very useful for the analysis of the overall network property; the reversibility information of the reactions is still incomplete; many spontaneous reactions, in particular the mutarotation reactions of sugars which take place under real physiological conditions, are missing. Because the diastereomers (*α*- and *β*-anomer) are generally distinguished as different compounds in the *LIGAND* database, the missing of spontaneous mutarotation reactions between them may lead to the breakage of the natural utilization pathways of these sugars. The aforementioned points (number of reactions, definition of reactant pairs, reversibility) make it necessary for us to continue to upgrade our bioreaction database accordingly.

Several parameters can be used for the evaluation of reconstructed metabolic networks based on the data of the mentioned reaction databases representing different quality standards. One important feature is the shortest path between two nodes/metabolites in a network. Because this parameter is based on pure graph theoretical facts, it is in general difficult to use this one to determine biochemically feasible paths. To compute such feasible paths, the database must fulfill two requirements:

1. The data of the reaction database must show a certain quality, *i.e.* reasonable reactant pairs by removing currency metabolites from the reactions.
2. Reversibility information based on experimental data or, if not available, biochemical/thermodynamical rules.

Additionally, a shortest path must meet an important constraint: One edge between two nodes may represent more than one reaction, and therefore more alternative paths are possible. Pathways that use a reaction type twice are not meaningful. Thus, we are interested in (shortest) paths with distinct reaction types. Note that classical shortest-path algorithms such as *Breadth-first search* or Dijkstra's algorithm are not capable of finding such paths. Computing biologically feasible shortest paths under these conditions is considerably harder. That is, the computation needs much more memory capacity and calculation time. Within this work we introduce an implementation of a shortest path algorithm that is able to solve this problem. The benefit of considering these aspects is that the chance for finding biologically feasible shortest paths will increase significantly.

There are some other approaches to find biologically feasible shortest paths and to avoid irrelevant shortcuts between metabolites: In their work based on the *KEGG RPAIR* database, Faust *et al.* (2009) take the chemical structure of reactants into account to differentiate between side and main compounds of a reaction. Croes *et al.,* (2005) use shortest paths in networks where a compound is assigned a weight equal to its number of incident edges. Finding pathways based on transfers of atoms between chemical compounds (atom tracking) was presented by Boyer and Viari (2003) and Heath *et al.* (2010). McShan *et al.* (2003) use heuristic search methods for finding metabolic pathways by reasoning over transformations using chemical and biological information (*PathMiner*). Kaleta *et al.* (2009) compute elementary flux patterns to detect pathways. Independently from our work, Chakraborty *et al.* (2009) show that it is NP-hard to find the minimum of labels such that for any two vertices there is a *rainbow path, i.e.*, a path with pairwise distinct edge labels.

2 MATERIALS AND METHODS

2.1 Software

The software listed here was used for database upgrade, reconstruction, and analysis of metabolic networks:

MS Office Excel 2003 and its built-in programming language *V*isual *B*asic for *A*pplications (*VBA*) and freeware: *Cytoscape* (Christmas *et al.*, 2005), together with diverse plugins: *NetworkAnalyzer Version 1.0* (Albrecht, Assenov and Lengauer, 2006, *Max Planck Institute for Informatics*), *ShortestPath Version 0.3* (Rosa da Silva, 2006). *Pajek Version 1.17* (Batagelj and Mrvar, 1998), *ncluster* (Python script, Rosa da Silva, 2006) for estimating the network modularity.

Furthermore, we used a shortest path tool by Scheer and Stelzer (Java program, 2008, *Department of Bioinformatics and Biochemistry*, *Technische Universität Braunschweig*) and SPUL (Kamphans and Stelzer, 2008).

2.2 Databases

For the upgrade of the bioreaction database the following databases were used as reference:

KEGG: *Kyoto Encyclopedia of Genes and Genomes* (Kanehisa *et al.*, 2010), particularly the *LIGAND* database (release 44.0) as well as *BRENDA*: *BRaunschweig ENzyme DAtabase* (Scheer *et al.*, 2010, release 2007.2). The *KEGG* database builds the backbone of our biochemical reaction database. Primarily the *BRENDA* database was used to get experimental data about the reactions, for example the reversibility information.

2.3 Database upgrade

Information from the *KEGG LIGAND* database was extracted into an *Excel* spreadsheet and compared with the previous version of our database (Ma and Zeng, 2003 a). The information on reversibility and reactant pairs for those unchanged reactions were adopted from our previous database. All reactions were manually processed to keep consistent with the rules defined below.

1. Reactant pairs. The reactions are used as linkage to build connection pairs (reactant pairs). For group transfer reactions, the transfer of carbon group is considered as a linkage, as suggested by Arita (2003 a,b), while the transfer of amino-, phospho-, and sulfo- groups are not. For example, the transfer of CoA will lead to a reactant pair (*e.g.* R_1-CoA – R_2-CoA). For methyl group transfer reactions mediated by *S*-adenosyl-*L*-methionine (SAM, C00019), SAM and the methyl-acceptor complex is a reactant pair. The reactant pairs defined in the *KEGG LIGAND* database can help this process only limitedly, because the carbon flow criterion and the currency metabolites (see below) were not consistently considered there. The *main*-pairs of *LIGAND* only mean that these pairs are present in the *KEGG PATHWAY* maps, which is different from the definition of reactant pairs in this work.

2. Currency metabolites. Currency metabolites (CM) usually refer to metabolites which take part in many reactions to transfer energy, electrons or certain functional groups (phosphoryl-, amino-, methyl- group, one carbon unit *etc.*; Huss and Holme, 2007; Neidhardt *et al.*, 1990). During the transferring process, the core structure of these metabolites remains unchanged. The function of these metabolites resembles currency (money) in the commercial world, therefore called currency metabolites. The typical examples are ATP and NADH. Besides the transferring function, most of the inorganic substances, such as H_2O, CO_2, O_2, Pi *etc*., involves in a huge number of reactions, and are frequently consumed and regenerated, having the feature of currency, too. Therefore, they are also treated as CMs. CMs are not used to build reactant pairs because otherwise the pathways deduced by graph theory from the metabolic network are physiologically meaningless due to the huge number of connections *via* these currency-like compounds (Ma and Zeng, 2003 a).

However, CMs cannot be defined *per se* (Ma and Zeng, 2003 a). For example, ATP is no more CM when used for adenylylation reactions or mRNA synthesis. So the recognition of CM has to be done manually for each reaction. The rules developed in the work of Ma and Zeng (2003 a) are generally followed with a few extensions. In the end, 153 metabolites are regarded as CM in this work (refer to the supplementary Tab. A1).

Here are some examples:

1) biuret + H_2O = urea-1-carboxylate + NH_3

→ One reactant pair was derived, ammonium was recognized as CM.

2) (*S*)-3-hydroxyisobutyryl-CoA + H_2O = (*S*)-3-hydroxyisobutyrate + CoA

→ Two reactant pairs, CoA is not CM.

3) ATP + adenylylselenate = ADP + 3’-phosphoadenylylselenate

→ One reactant pair, ATP and ADP are CM.

4) ATP + selenate = adenylylselenate + pyrophosphate

→ Two reactant pairs, ATP is not CM.

5) 2-C-methyl-*D*-erythritol 4-phosphate + CTP = 4-(cytidine 5’-diphospho) -2-C-methyl-erythritol + pyrophosphate

→ Two reactant pairs, CTP is not CM.

6) *L*-Alanine + 2-Oxoglutarate = Pyruvate + *L*-Glutamate

3. Reversibility of biochemical reactions. Strictly speaking, all reactions are reversible in principle. However, for the purpose of reconstruction and analysis of the metabolic network in a physiologically meaningful sense, it is important to define the reversibility according to real physiological situation. Information concerning reversibility and direction of reactions is shown in the *KEGG PATHWAY* maps (reflected in the file *reaction_mapformula.lst* of the *LIGAND* database) for some reactions, and the experimental proofs are collected in the *BRENDA* database. The criteria for irreversible reactions described by Ma and Zeng (2003 a) have been generally adopted with only one exception: the reactions, where a sugar unit was transferred *via* an activated sugar (UDP-, NDP- and dTDP-sugars), are treated as reversible because the high energy bond is preserved.

A few new rules are introduced:

- If both directions are possible depending on different biological situations, this reaction is treated as reversible.
- If the reversibility is uncertain, the reaction is treated as reversible.
- If ΔG of the reaction is close to 0, the reaction is treated as reversible. For example:

$$\text{CTP} + \text{R-Pi} \leftrightarrow \text{CDP-R} + \text{PPi}$$
$$\text{UDP-sugar} + \text{A} \leftrightarrow \text{UDP} + \text{B}$$

However, almost all the rules have exceptions. For example:

- Hydrolysis reactions are usually irreversible. But if the hydrolysis happens within a molecule (ring cleavage and without products left), then it can be reversible.
- Hydrolyase reactions are usually irreversible, but sometimes reversible according to the experimental data collected by *BRENDA*.

4. Non-enzymatic spontaneous reactions. Biochemical reactions are not always catalyzed by enzymes. 35 non-enzymatic spontaneous reactions were defined previously in the *LIGAND* database. However, we found that the spontaneous mutarotation reactions for sugars were not described in the *LIGAND* database unless the reaction was also known to be catalyzed by enzymes. It is known that sugars, esp. pentose, hexose and its derivates, may have three interconvertible forms in aqueous solution: *α*-, *β*-anomer and the open-chain isoform (Bailey *et al.*, 1970). The isoforms are distinguished as different compounds in the *LIGAND* database, and some enzymes exhibit strong optical specificity to only one of the isoforms. These spontaneous reactions are indispensable for reconstruction of a functioning metabolic network. For example, starch may enter the pentose phosphate pathway *via* a pathway shown in Figure 1. The second step can be catalyzed by the enzyme EC 5.3.1.9, but also takes place spontaneously with a halftime of 1.5 seconds (Bailey *et al.*, 1968). If the spontaneous reaction is not considered, the missing of EC 5.3.1.9 will make it impossible for starch to enter the pentose phosphate pathway in this case. It should be mentioned that not all the spontaneous mutarotation reactions take place as fast as the conversion between *α*- and *β*-*D*-glucose-6-phosphate. For example, the mutarotational half-time of *α*- or *β*-*D*-glucose is approximately seven minutes under physiological conditions (Bailey *et al.*, 1968).

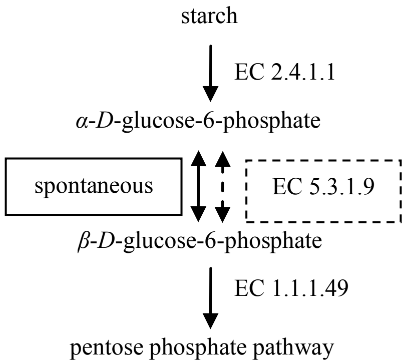


Figure 1: A pathway for starch to enter the pentose phosphate pathway.

Fourty-one spontaneous reactions for sugar mutarotation reactions were introduced in this work. Together with those spontaneous reactions already defined by *LIGAND*, the overall number of spontaneous reactions reached 76. A spontaneous reaction is included during reconstruction of an organism-specific metabolic network, only if one of its reactants is involved in the enzymatic reactions known for this organism.

5. *Balance* of COHNSP. In some cases the reactions of the database show incompleteness concerning the stoichiometry. This means that the number of the elements C, O, H, N, S or P for example is not the same on both sides of a reaction equation, *i.e.* educts and products. In order to get stoichiometric *balanced* reactions it was necessary in such cases to introduce or add molecules like CO_2, O_2, N_2, H_2O *etc.* in a suitable factor.

6. Error correction. Although the *LIGAND* database was improved continuously, errors and inconsistencies besides the problem of reversibility still exist (Tab. A2 of supplementary material). These errors were corrected manually according to our biological knowledge or the literature information. Some examples are:

The reaction R00011: 2 manganese + 2 H_2O = H_2O_2 + 2 manganese + 2 H^+, was corrected as: 2 Mn(II) + 2 H^+ + H_2O_2 = 2 Mn(III) + 2 H_2O. Manganese was distinguished as Mn(II) and Mn(III), and the reaction is treated as irreversible according to *BRENDA*.

In the reaction R00329: NDP + H_2O = nucleotide + orthophosphate, the product *nucleotide* was designated as *NMP* to avoid ambiguity.

Altogether 107 reactions have been corrected in this way.

2.4 Network reconstruction and analysis

The organism-specific metabolic network was reconstructed based on the list of Enzyme Commission numbers (EC) defined for each model organism using the following sources:

- *KEGG* (Kanehisa *et al.*, 2010) for *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast), *Ureaplasma urealyticum*, *Mycoplasma pneumoniae*, *Mycoplasma genitalium* (low GC content gram$^+$ bacteria) and *Borrelia burgdorferi* (spirochete)
- *BRENDA* (Scheer *et al.*, 2010) and (Ma and Zeng, 2003 a) for *Escherichia coli* K-12 MG1655 (*γ*-subdivision of proteobacteria)
- (Sun *et al.*, 2007) for a filamentous fungus *Aspergillus niger*.

The relevant reactions were extracted from the reaction database concerning their reversibility as well as reactant pairs by using *VBA* scripts and transferred into a special network format, which could then be interpreted by programs such as *Cytoscape* (Christmas *et al.*, 2005) or *Pajek* (Batagelj and Mrvar, 1998), typically as a metabolic or reaction graph. In the metabolic graph, nodes or vertices represent metabolites. Irreversible reactions are shown as arcs and reversible ones as edges or bi-directed arcs. In the reaction graph the nodes are the reactions while the arcs/edges represent the metabolites (Wagner and Fell, 2001).

In order to evaluate the impact of the database upgrade, the metabolic networks reconstructed from the former and upgraded databases were analyzed and compared in terms of functional and structural parameters of the networks, such as modularity-based network decomposition, shortest path, network diameter, average path length, and centrality *etc*.

Network decomposition is necessary for functional analysis of large-scale, genome-wide networks that are often hampered by the problem of combinatorial explosion due to the complexity of networks. Network decomposition is to break the network into biologically meaningful modules so that the connections among the nodes within a module are maximal and those between putative modules are minimal. Our group implemented a modularity-coefficient based program called *ncluster*, which was proved to be able to deduce biological meaningful modules from a reaction graph (Rosa da Silva, 2006 and personal communication; Ma *et al.*, 2004). In this study, both the genome-wide organism-specific network reconstructed from the former and upgraded reaction databases (only the largest connected part) and its giant strong components (GSC) (Kumar *et al.*, 2002; Ma and Zeng, 2003 b) were tested for decomposition using this program. The GSC was extracted from the genome-wide network using the program package *Pajek* (Batagelj and Mrvar, 1998).

Further network parameters like shortest path (SP; Cormen *et al.*, 2001), network diameter (*i.e.* the longest of all SP), average path length (AL; Batagelj and Mrvar, 1998), centrality (betweenness and closeness; Brandes, 2000, 2001; Freeman, 1977) and the node degree distribution (Newman, 2003) were analyzed by using *Pajek*. The AL was calculated both for the whole network (ALW) and for the GSC (ALG). Parameters like the SP, the number of nodes (*i.e.* network size) and edges, the average number of neighbors and connected pairs were calculated with the program *Cytoscape* (Christmas *et al.*, 2005). This program was also used for visualization of the networks.

3 RESULTS AND DISCUSSION

3.1 Statistical comparison of the two versions of databases

In total, the upgraded database contains 6851 reactions. This number is almost doubled in comparison to the previous database (3565 reactions having connection pairs; Ma and Zeng, 2003 a). The upgraded database has 3525 different EC numbers (2943 complete ones) while the previous database had 3115 EC numbers (2643 complete ones). It should be noticed that one enzyme might be able to catalyze more than one reaction and many different enzymes might catalyze the same reaction. The number of reactions in the upgraded database includes 76 non-enzymatic spontaneous ones, of which 41 are spontaneous sugar mutarotation reactions newly introduced by us (see Materials and Methods). At least 3286 reactions, or 48 % of the upgraded database, represent new entries in comparison to the former version of the database (Ma and Zeng, 2003 a, Tab. 1).

In the upgraded database, there are 4304 (1789 in the former version) irreversible reactions and 2547 (1776 in the former version) reversible ones including the 41 spontaneous mutarotation reactions. 235 irreversible reactions of the former database were changed as reversible while 731 reversible ones were changed to irreversible (Tab. 1). Totally, the reversibility of 27 % of the reactions was changed according to experimental data from literature.

For many pathways, especially for those involved in the secondary metabolism, which are of increasing biological and biotechnological interest, the upgraded database contains significantly more reactions (Tab. 2).

3.2 Impact of the database upgrade

To demonstrate the impact of the database upgrade, the metabolic networks for three model organisms, *Homo sapiens* (hsa), *Aspergillus niger* (anig), and *Escherichia coli* K-12 MG1655 (eco), were reconstructed using the different versions of bioreaction databases and compared in details (see Materials and Methods).

3.2.1 Number of nodes.

The networks based on the upgraded database include more reactions and metabolites than those based on the previous database (Tab. 3, Fig. 2 and supplementary Fig. A1). The number of metabolites (nodes) increased in all three examined organisms by more than 50 % while the number of reactions increased by more than 80 %. In the giant strong components (GSC) the number of metabolites and reactions was even doubled. Simultaneously, the number of isolated reactions and metabolites decreased because many in the previous network isolated metabolites and reactions were now reconnected to the major network *via* the newly introduced reactions (Fig. 2 and supplementary A1). For instance, in the human metabolic network based on the previous database (Fig. 2 A) there was a disconnected subnetwork consisting of 45 reactions (arcs/edges) and 33 metabolites (yellow nodes). As can be seen from Figure 2, this part is now integrated into the main metabolic network reconstructed from the upgraded database.

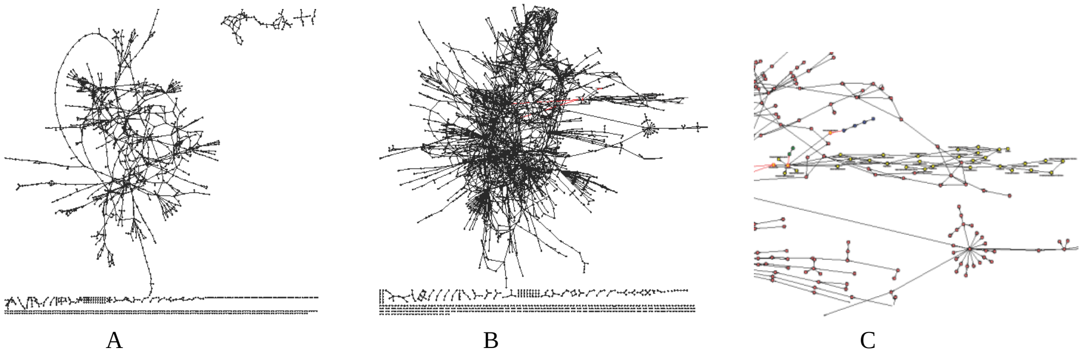


Figure 2: Comparison of the organism-specific metabolic networks for *H. sapiens* (hsa), reconstructed from the former [A] and upgraded [B and C for details] bioreaction databases. The network based on the upgraded data material shows a higher complexity and less disconnected parts than those based on the former database (Ma and Zeng, 2003 a). The example (yellow nodes) shows the integration of metabolites and reactions of the bile acid and steroid hormone biosynthesis into the major network.

A closer analysis shows that 28 metabolites belong to the steroid hormone biosynthesis pathway, two to the bile acid biosynthesis, and one to both pathways. For the left two metabolites there is so far no information available in the *KEGG* database concerning the pathways involved. It is known that human can synthesize bile acid and steroid hormone *de novo*. It is logic that their synthesis pathways should be connected to the main metabolic network. Now with the application of the upgraded database the missing links were identified (Fig. 2 B and C for details). Responsible for the integration of this part into the main network are two reactions which belong to both the bile acid and steroid hormone metabolism (R01461 and R01462). Another two reactions, R03310 and R07215, belonging to the steroid hormone metabolism, connect another formerly separated 4-node part (consisting of four metabolites, C01189, C03845, C05110, and C05111, blue nodes) and two new metabolites (C13550 and C15518, that are not defined in the old database, green nodes) into the main network.

Of the four connecting reactions (Fig. 2 C), R07215 only occurs in the upgraded database whereas the other three occur in both versions. The EC number for R03310 has been updated in the new database. Therefore only with the new database, this reaction can be extracted for *H. sapiens*. Integration of the four-metabolite subnetwork is mediated by the reactant pair (cholesterol and cholesta-5,7-dien-3-*beta*-ol) built from any of the two reactions R03310 and R07215. For the last two closely related reactions (R01461 and R01462), only one reactant pair has been described in the former database while two reactant pairs are defined now in the upgraded database based on the rule of following the C-flow. Exactly the second additional reactant pair finally links the isolated 33-node subnetwork to the main network *via* the reactant pair of cholesterol ester and acyl-CoA built from R01461 or *via* the pair of cholesterol ester and fatty acid from R01462 (red arcs/edges). The connecting metabolites for the integration of the mentioned subnetworks or parts are labeled by a red border.

3.2.2 Node degree.

Table 4 shows the ***number of the first neighbors*** for 23 potential hub metabolites (*i.e.* nodes with a relatively high connection degree) in the organism-specific metabolic networks. Most of the hub metabolites have significantly more first neighbors for each tested organism with the upgraded datasets. For example, acetyl-CoA has 51 first neighbors in the metabolic network of *A. niger* reconstructed with the updated database, while only 12 first neighbors with the former database. This is mainly due to the increase of reaction number but also because of modified rules for potential currency metabolites and reactant pairs. Moreover, glycans (usually termed as *G* number instead of *C* number) have now been recognized in this work but not in the former database (Ma and Zeng, 2003 a), where these substances have been ignored.

The estimated ***average number of first neighbors***, (*NetworkAnalyzer*, undirected graph) is for all organism-specific networks based on the former dataset smaller than for those based on the upgraded data material (Tab. 5). It is remarkable, that the values for *A. niger* are a little bit smaller than those for *E. coli*, although the network of *A. niger* has more nodes and reactions (Tab. 3). This indicates that the network size alone (*i.e.* the node quantity) tells nothing about how dense the nodes (metabolites, substances) are actually connected.

The ***average of node degree*** (*input*, *output* and *all*) for each organism-specific network (whole network and also the GSC) were estimated by *Pajek* (Batagelj and Mrvar, 1998) using a directed graph. The number of neighbors and the node degree are not mandatory the same because it is possible that two nodes are connected by more than only one arc/edge, *e.g.* in the case of two irreversible reactions with opposite directions. Therefore the node degree or connectivity has to be ≥ the number of first or direct neighbors found. The average of node degree slightly increased with the upgraded database (Tab. 5).

Another parameter, ***number of connected pairs*** (amount of two nodes respectively regarded as a pair of nodes which are connected by edges), shows dramatic differences between the networks (Tab. 5), reconstructed using the former and upgraded datasets. With the upgraded database, the number of connected pairs is 3 to 4 times higher than that with the former database, for all compared organisms. More importantly, the percentage of connected pairs is 53 %-59 % with the upgraded database while 31 % to 38 % with the former database. This means that the network constructed with the upgraded database not only quantitatively has more connected pairs, but is also qualitatively much better inter-connected, which can also be seen from the network pictures intuitively (Fig. 2 and supplementary Fig. A1).

As we mentioned in the Methods section, the currency metabolites (CMs) cannot be defined *per se*. In some cases, the potential CMs are treated as normal metabolites. Table A1 (supplementary material) listed the most important potential CMs (sorted by their occurrences in the upgraded reaction database) and their direct neighbors in the organism-specific metabolic networks according to the former and upgraded datasets. In most cases the number of first neighbors of special substances (not necessarily the node degree) was stable or only slightly changed between the metabolic networks constructed from the two databases. However, in some cases, *e.g.* ATP, one of the most frequent substances in both databases, shows a significant higher number of neighbors in the organism-specific networks based on the upgraded reaction database. This phenomenon can also be recognized for similar substances like CTP, GTP, UTP, and AMP, and as well as glutathione. Changes for NAD^+/NADH and $NADP^+$/NADPH were small, although they are among the most frequent substances in the list. The modification of the rules for CMs (see Materials and Methods) certainly has influence on the first neighbors of some substances. As an example, *L*-glutamate has a higher number of neighbors in the new organism-specific networks compared to the older ones (data not shown) but this substance explicitly is not treated as a potential CM in the upgraded reaction database.

3.2.3 Modularity.

The modularity (Ma *et al.*, 2004; Clauset *et al.*, 2004; Newman, 2004 a, b, 2006; Newman and Girvan, 2004) of both the larger connected part of the organism-specific metabolic networks of the three organisms and also their GSC components was estimated and analyzed (see Materials and Methods). The reaction graph was used instead of the metabolic graph (see above). In comparison to the former dataset, for the GSC, of the networks based on the upgraded database, the number of reactions almost doubled while the number of arcs/edges nearly quadrupled (Tab. 6). Consequently, the total number of pathways involved was greatly increased by up to 100 %. The best modularity of the GSC went down, from 0.67 to 0.55 for *E. coli*, from 0.75 to 0.52 for *A. niger* and from 0.79 to 0.65 for *H. sapiens*. The number of modules remained unchanged for *E. coli*, but much reduced for *A. niger* and *H. sapiens*.

The increased network complexity (node number and arc/edge number) reasonably have impact on modularities: the nodes are so densely connected that the breakdown of the network is costly. The question is whether the new decomposition is still biologically meaningful. Ideally, all successive reactions conducting a complete metabolic pathway are confined in the same module. However, because one reaction can involve in many metabolic pathways and the metabolic pathways are interconnected, the reactions conducting a pathway may be decomposed into several modules according to the optimal modularity algorithms. The less number of modules the specific pathway-belonging reactions are distributed into, the more biologically meaningful the decomposition is. We compared the distribution of metabolic pathway belonging reactions among the modules. Indeed, for most of the metabolic pathways, the reactions are less distributive when the upgraded database is applied. For example, the GSC network of *H. sapiens* based on the upgraded database has 31 reactions belonging to the valine biosynthesis/degradation pathway, which are partitioned into two modules, while the GSC based on the old database has 10 reactions which are partitioned into three modules.

Interestingly, if we look at the distribution pattern of the belonging pathways across the modules, more pathways tend to be co-present or co-absent when applying the updated database. For example, reactions belonging to the pathways of glutamate and glutamine metabolism, urea cycle, arginine, and proline synthesis are always found also in three modules in case of the *H. sapiens* GSC based on the upgraded data, whereas with the former database, reactions belonging to the glutamate and glutamine metabolism are also found in other modules additionally. The same co-distribution pattern are also found for the pathways threonine, methionine, and lysine synthesis, as well as the pathways for pentose phosphate metabolism, purine, folate, and riboflavin synthesis. The co-distribution pattern of these pathways are consistent to their tight biological relations.

3.2.4 Changes in further network parameters.

Shortest path. The shortest path (SP; Cormen *et al.*, 2001) between two selected nodes in a graph is the path having the lowest costs (in a weighted graph) or simply the smallest number of steps (unweighted graph). In the case of a metabolic network, the SP describes the number of necessary reactions to convert one metabolite into another. It should be noticed that the shortest path is not necessarily the same as the metabolic pathway defined in the biochemistry textbook (Rosa da Silva, 2006). We analyzed the SPs calculated by the software *Cytoscape* (Christmas *et al.*, 2005) and by *SPUL* (Kamphans and Stelzer, 2008), which yields biologically more feasible paths, but cannot handle large networks due to the computational complexity. The SP from *D*-glucose to pyruvate based on the upgraded database was shorter than the one based on the former database in all the three organisms studied (Tab. 7). The natural biochemical pathway is often different from the SP calculated. The natural glycolysis pathway is composed of nine steps (*i.e.* ten nodes/metabolites) from *D*-glucose to pyruvate (Ma and Zeng, 2003 a). In the network of *H. sapiens* based on the upgraded database, four glycans which do not belong to the glycolysis were involved in the estimated SP from glucose to pyruvate. The higher connectivity degree of the networks based on the upgraded database (which results in alternative *short cuts*) may be one of the reasons causing shorter SPs.

A biochemical pathway may be longer than a graph-theoretical shortest path because a network contains reversible reactions that lead to shortest paths that are not meaningful in a biochemical sense: Given a reaction that leads from a substrate to two different products and back — such as from 2-dehydro-3-deoxy-*D*-galactonate 6-phosphate (*KEGG* compound ID: C01286) to glyceraldehyde 3-P (C00118), and pyruvate (C00022) —, it is possible to find a shortest path that leads from one product *via* the substrate to the other one. To avoid this effect, it is necessary to store the reaction types as labels on the edges of the network, and assure that a shortest path passes no label twice (Rosa da Silva, 2006). Unfortunately, such paths are hard to compute, see Chapter 4.

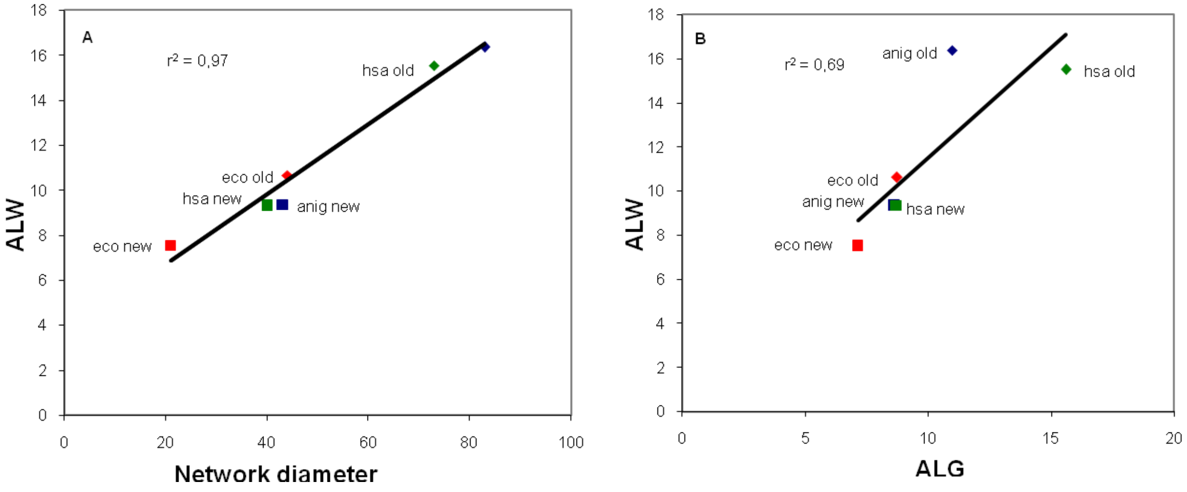
Centrality indices. The network betweenness centrality expresses how many SP go through a node. A higher value of this parameter stands for a greater importance of the node in the whole network (Brandes, 2000, 2001; Freeman, 1977). The overall network betweenness centrality is calculated based on the centrality value of each node (*Pajek*; Batagelj and Mrvar, 1998). For the GSC based on the former and upgraded databases, the measured centrality is shown in Table 7. Astonishingly only the centrality value for *A. niger* increased with the usage of the upgraded data material whereas the values for *E. coli* and also *H. sapiens* decreased.

Another estimated centrality index, called the closeness centrality, describes how close the position of a node to all other nodes in a network is. The closeness centrality is estimated as the reciprocal of the average distance from a node to all other nodes (Brandes, 2000, 2001; Rosa da Silva, 2006; Rosa da Silva *et al.*, 2007). For the network GSC based on the former and upgraded data the three parameters *input*, *output*, and *all* closeness centrality can be calculated as described previously (*Pajek*; Batagelj and Mrvar, 1998). For the three model organisms tested here, all three parameters increased when the upgraded data material was applied (Tab. 7), indicating that in general the nodes in the network based on the upgraded data material are more centralized and therefore less peripheral.

Average path length and network diameter. The impact of the database update on the average path length of the whole networks (ALW) and of the GSC (ALG) as well as the network diameter (Batagelj and Mrvar, 1998) are shown in Figure 3 and Table 7 for all organism-specific metabolic networks. The values of both parameters for the network based on the updated database are smaller. The ALW is proportional to the network diameter (the correlation coefficient $r^2 = 0.97$; Fig. 3 A), revealing that the correlation between these two parameters found previously (Ma and Zeng, 2003 a) is still conserved. And the ALG is also roughly proportional to ALW ($r^2 = 0.69$, Fig. 3 B), that is consistent to the previous finding (Ma and Zeng, 2003 b). Though the network properties are altered significantly, such as a decrease in the AL and network diameter and an increase in the average node degree, the new organism-specific networks still belong to scale-free networks (Barabási and Bonabeau, 2003) in general.

3.2.5 Glucose subnetwork.

The glucose subnetwork is a subset of the whole network, where all the metabolites can be converted (reachable) from glucose. The glucose subnetworks were extracted from the whole network (directed graph) by using the *BFS*-algorithm (*breadth first search*; Broder *et al.*, 2000; Cormen *et al.*, 2001; *Pajek*, Batagelj and Mrvar, 1998), and its features are shown in Table 8. Similar to the whole network and the GSC, the glucose subnetwork based on the upgraded database shows a higher complexity as indicated by the immense increased number of nodes, the smaller AL value and the smaller distance between glucose and its farthest reachable metabolites. Remarkable is the finding that the estimated AL for the glucose subnetworks based on the upgraded data always shows a value close to eight whereas this value is higher but diverse for those based on the former database (Tab. 8). Further analysis of other organism-specific glucose subnetworks based on the upgraded data material confirmed that many of them also show values for the AL around eight (data not shown).



A B

Figure 3: Correlation between the average path length (AL) of the whole network (ALW) and the network diameter [A], the AL of the whole network (ALW) and that of the giant strong component (GSC, ALG) [B] of the organism-specific metabolic networks. The reconstruction of the networks is based on both the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases. It is obvious that by using the upgraded database for the reconstruction of metabolic networks, the relation between the parameters shown has been conserved.

3.3 Comparison with literature-based metabolic networks.

The automatically reconstructed metabolic networks based on the upgraded database were compared with the recent human-curated literature-based metabolic network for *H. sapiens* and for the yeast *S. cerevisiae* (sce). In the recent work of Ma *et al.* (2007), the reconstructed *Edinburgh human metabolic network* contains 2823 reactions, only slightly less reactions than our metabolic network for human (2897 reactions, Tab. 9). The human metabolic network reconstructed by Duarte *et al.* (2007), called *Recon 1*, contains 1982 unique metabolic reactions (excluding transport reaction and not considering the compartment specificity) and 1509 unique metabolites if ignoring the compartment information. These numbers are significantly less than our human network, indicating that the network automatically reconstructed from the upgraded bioreaction database can supply many additional candidate reactions and therefore be helpful for the manual metabolic network reconstruction.

Our metabolic network for *S. cerevisiae* based on the upgraded database consists of 1482 metabolites and 1800 reactions. In the work of Förster *et al.* (2003) the metabolic network for *S. cerevisiae* contains only 584 metabolites and 1175 reactions which are much less than the size of the network we constructed. For this comparison it should be noticed, that the work of Förster *et al.* (2003) was done a few years ago and the availability of reaction information at that time was not as much as we have now.

Generally speaking, the reconstruction of a metabolic network from literature data costs a lot of time and efforts. Though such network is very reliable due to the direct experimental proofs, it is incomplete due to the limited availability of experimental results. Based on the upgraded reaction database and the easy-to-obtain genome annotation information, our method can construct the theoretical metabolic network in minutes. By comparing the automatic reconstruction and the literature-based manual reconstruction as shown in the human and yeast metabolic network above, the automatic construction usually compasses most of the information included in the manual reconstruction, indicating the high accuracy of the automatic reconstruction. And in addition, the automatic reconstruction usually predicts much more reactions and pathways than reported in the literature, therefore forming many genome-scale hypothesis concerning the new metabolic reactions and metabolic potentials, and finally promoting the organism-specific systems biology studies.

4 COMPUTING SHORTEST PATHS IN METABOLIC NETWORKS

In this section, we discuss a difficulty that arises when computing shortest paths in metabolic networks (for example, to evaluate the underlying database): In general, shortest paths computed using a graph-theoretical point of view may not be biologically meaningful, as mentioned in the introduction. If the database is carefully adjusted (*i.e.*, currency metabolites that lead to infeasible shortcuts are removed and reversibility of reactions is modeled accurately), certain shortest paths can be used. These paths have to meet an additional constraint, as described in the following.

Let G be a graph (in our case a metabolic network) that consists of a set of vertices (metabolites), V, and a set of edges (reactions), E, where an edge $e=(v_1, v_2)$ connects two vertices of the graph. Given two vertices, $s \in V$ and $t \in V$, a *shortest path* from s to t in G is a sequence of edges from E that connect s and t using as few edges as possible. A path that is shortest in the sense of graph theory may not be feasible from a biochemical point of view, because it may use the same reaction twice, as explained in Section 3.2.4.

Thus, we are interested only in *feasible* shortest paths, that is, shortest paths from s to t with distinct reaction types. Such a path may be longer than a shortest path, see Figure 4. To store the reactions in the graph, we use *labels* for the edges. Altogether, a mathematical model for our problem is the following.

Problem *Shortest Path with Unique Labels* (SPUL)

Given a graph $G=(V, E)$, a mapping $\ell: E \longrightarrow \mathbb{N}$ that assigns a label to every edge, and two vertices, $s \in V$ and $t \in V$, find a shortest path $P=(e_1=s, e_2, \ldots, e_k=t)$ with pairwise distinct edge labels; that is, for $1 \leq i < j \leq k$: $\ell(e_i) \neq \ell(e_j)$.

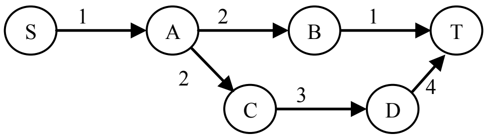


Figure 4: The shortest path from S to T is S→A→B→T, but this path is infeasible, because it passes the label '1' twice. The shortest feasible path is S→A→C→D →T.

4.1 Shortest Paths in Unlabeled Networks

Given a graph, $G=(V, E)$, it is quite simple to compute the shortest path between two vertices of the graph. The *Breadth-First Search* algorithm (*BFS*), see Algorithm 1, is known to every first-grade student of computer science. Even if the edges have different length, the problem is still easy to compute using Dijkstra's algorithm; see, for example, Cormen *et al.* (2001).

Note that *BFS* creates a tree of all shortest paths by storing for every vertex, v, its predecessor, *v.father*, on the shortest path to the start vertex. Thus, a shortest path from a given vertex to the start vertex can be found simply by following the *father* pointers.

```
Let $Q$ be a queue of vertices
insert start vertex into $Q$
while $Q$ is not empty do
    $v$ := first vertex from $Q$
    remove first vertex from $Q$
    for all vertices $v'$ adjacent to $v$ do
        if $v'$ was not visited before then
            $v'.father$ := $v$
            mark $v'$ as visited
            append $v'$ to $Q$
            report shortest path to $v'$
        end if
    end for
end while
```

Algorithm 1: BFS

4.2 Shortest Paths with Unique Labels

Things get *considerably* harder, if we add labels to the edges and require that no label is passed twice on a path between two vertices. Note that neither *BFS* nor Dijkstra's algorithm are able to find the feasible shortest path shown in Figure 4.

To express the hardness of a problem, it is very common in computer sciences to estimate the order of growth in the resources (*i.e.*, running time and memory requirements) needed by a program as a function in the input size. In our case, the input size is the number of vertices and edges in the graph. While the running time of *BFS* is linear in the input size (that is, for example, if we double the input size, the running time doubles also), we can show that the running time is most likely to be exponential for shortest paths with unique labels; that is, if we have a graph with *v* vertices and *e* edges, the running time is in the order of 2^{v+e}. More precisely, we can show that our problem belongs to the class of *NP-complete* problems (Garey and Johnson, 1979). It is a widely held belief that there is no sub-exponential solution for NP-complete problems. The impact of this running time is shown in Figure 5.

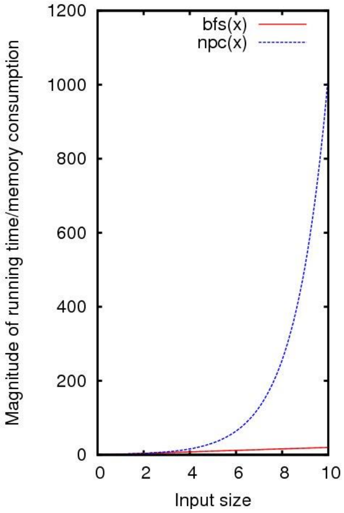


Figure 5: Magnitude of resource request for an NP-complete problem compared to *BFS*.

Theorem. Given a graph $G=(V, E)$ with a mapping $\ell: E \longrightarrow \mathbb{N}$ that assigns a label to every edge, it is NP-complete to determine if there is a path $P=(e_1, e_2, ..., e_k)$ that uses every label at most once; that is, for
$1 \leq i < j \leq k$: $\ell(e_i) \neq \ell(e_j)$.

Proof. We show our theorem using a common technique in computer science known as proof by *reduction*. That is, we take a well-known hard problem—in our case 3-SAT—and show that this problem would be easy to solve if the shortest path with unique labels problem (SPUL) would be easy to solve. This is done by describing how to translate an input to 3-SAT to SPUL such that a solution to SPUL yields a solution to 3-SAT.
Given a set of *binary variables*, x_1, ..., x_n, and a set of *clauses*, C_1, ..., C_m, consisting of three literals (*i.e.*, $C_i = L_{i1} \vee L_{i2} \vee\ L_{i3}$, where L_{ik} denotes a negated ($\overline{x_k}$) or unnegated variable (x_k)), the problem *3-SAT* asks if there is an assignment of x_1, ..., x_n to 0 or 1 such that all clauses are fulfilled (Garey and Johnson, 1979).
A 3-SAT instance can be transformed to a SPUL instance as follows: For every clause C_i we use a clause gadget that consists of three parallel edges labelled with L_{ik} (*i.e.*, with x_{ik} or $\overline{x_{ik}}$ for an unnegated or negated variable, respectively). The variable gadget for variable x_j consists of two parallel paths, one with all negated labels, one with all unnegated labels. For the whole input, we start with a vertex, *s*, and add all variable gadgets followed by all clause gadgets. The last vertex is labeled *t*; see Figure 6. To find a path from *s* to *t*, we have to pass either the negated or the unnegated branch for every variable. Thus, after passing the variable gadgets we have either all negated or all unnegated variables left to pass the clause gadgets without using a label twice. This is possible if and only if the given formula is satisfiable. □

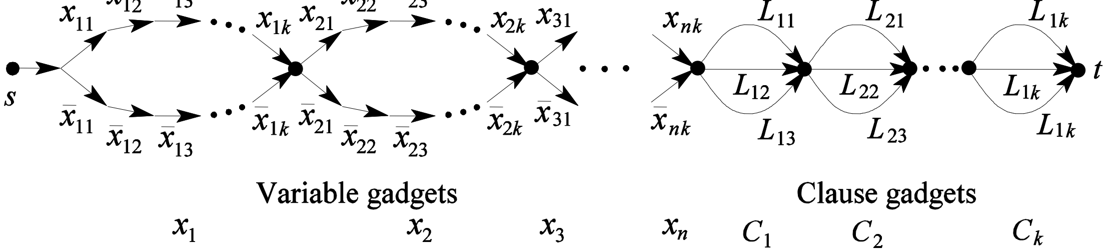


Figure 6: Transforming a 3-SAT instance to an SPUL instance.

```
Let $Q$ be a queue of edges
insert “dummy egde” to startnode into $Q$
while $Q$ is not empty do
    $e$ := first edge from $Q$
    remove first edge from $Q$
    for all edges $e'$ adjacent to $e.target$ do
        if $e'.label$ was not used on the shortest path from $s$ to $e$ then
            $e'.father$ := $e$
            append $e'$ to $Q$
            if $e'.target$ was not visited before then
                report shortest path to $e'.target$
            end if
        end if
    end for
end while
```

Algorithm 2: Shortest Path with Unique Labels

4.2.1 A Memory-Consuming Solution.

We use a modified version of *BFS* to solve our problem, see Algorithm 2. Similar to *BFS*, we store all paths found so far. But instead of storing a shortest-path tree for the vertices as in *BFS*, we construct a feasible-shortest-path tree on the edges of the graph, see Figure 7. That is, we store every possible feasible path leading to a vertex in the tree during the search. When the search reaches a vertex, *v*, *via* an edge, *e’*, we can determine if the path to *v via e’* is feasible (*i.e.*, no label occurs twice). By the *BFS*-manner of this algorithm, the first feasible path found to a vertex is also the feasible shortest path. The drawback is that the algorithm is quite memory consuming, because it stores all feasible paths to all vertices.

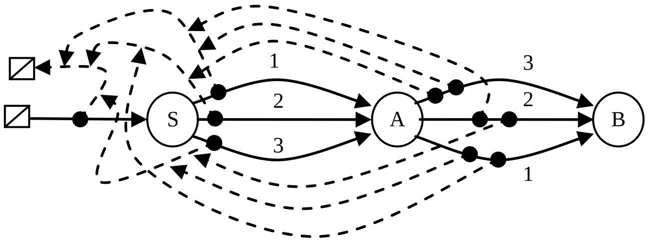


Figure 7: An example for a shortest feasible path tree constructed by Algorithm 2. The network consists of three nodes (S, A, and B), three edges from S to A, and three edges from A to B. The dashed lines show the shortest feasible path tree.

4.2.2 Balancing Time and Memory Requirements.

To save memory, we used a different solution by exploiting the fact that a metabolic network has many parallel edges. Thus, our search does not have to explore all edges incident to a vertex, but only those edges that lead to different vertices. Instead of storing one label per edge on a shortest path, we store a set of labels. This significantly decreases the number of shortest paths that we have to store. The drawback is that we have to find a feasible combination of labels when the search progresses (*i.e.*, when we want to add a new edge to a path). This can be solved by a simple backtracking; that is, we successively test combinations of labels until we either find a feasible combination or no more combination is possible. The idea is that in most cases this backtracking does not require much time, because a feasible combination is found quickly. Only if there is no feasible combination, we have to test all of them. Clearly, this heavily depends on the structure of the input network.

4.2.3 Preprocessing.

Before we start our algorithm, we perform a simple *BFS* to determine, which vertices can be reached at all (*i.e.*, there is a feasible or infeasible path). We store these vertices, and abort the search as soon as feasible paths to all of them have been found. In a second stage of the preprocessing, we perform a simple *BFS* again; this time checking if the found path is feasible and reporting feasible paths.

4.2.4 Comparison.

Table 10 shows results for large databases. There was not sufficient memory to compute all paths. Thus, we compare the number of paths found until the program was stopped because there was no memory left. It turned out that the memory consuming solution (Alg. A) is much faster than our second approach, but finds less paths. The preprocessing with *BFS* further improves the number of found paths. Smaller organisms are compared in Table 11: We compared the number of (graph theoretically) shortest paths (SP) to the number of shortest paths with unique labels (SPUL). Furthermore, we listed the number of biologically infeasible shortest paths that were found using *BFS* (*i.e.*, paths such as S→A→B→T in Figure 4).

5 CONCLUSION

In this study, we upgraded the bioreaction database which was first established in 2003 in our group. Together with modification of the rules for currency metabolites, revision of reversibility and reactant pairs as well as addition of more non-enzymatic spontaneous reactions, the updated database contains almost doubled number of reactions. Using *E. coli*, *A. niger,* and *H. sapiens* as examples, we could show that the organism-specific metabolic network, reconstructed automatically by using the upgraded database and genome annotation, is more complete and more reliable. The network parameters were also significally affected by the upgrade, however, the reconstructed network remains scale-free. The upgraded reaction database allowed a fast and accurate reconstruction of genome level metabolic networks. This will further facilitate the exploitation of genome and experimental information, and accelerate the network-based biological studies.

ACKNOWLEDGEMENTS

This work was financially supported by the German Research Foundation (DFG), project SFB 578 and by Federal Ministry of Education and Research (BMBF), project InterGenomics.

Tom Kamphans was supported by 7th Framework Programme contract 215270 (FRONTS).

The authors thank Maurice Scheer for his Java program and Sony Adhi Susanto for his support and discussion concerning the upgrade of the bioreaction database during his practical course.

REFERENCES

M. Arita (a), Representing metabolic networks by the substrate-product relationships. *Genome Informatics*, 2003, **14**, 300-301

M. Arita (b), *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Research*, 2003, **13**, 2455-2466.

J. M. Bailey, P. H. Fishman, P. G. Pentchev, Studies on Mutarotases, II. Investigations of possible rate-limiting anomerizations in glucose metabolism. *Journal of Biological Chemistry*, 1968, **243**, 4827-4831.

J. M. Bailey, P. H. Fishman, P. G. Pentchev, Anomalous mutarotation of glucose 6-phosphate. An example of intramolecular catalysis. *Biochemistry*, 1970, **9** (5), 1189-1194.

A. L. Barabási, E. Bonabeau, Scale-free networks. *Scientific American*, 2003, **288**, 50-59.

V. Batagelj, A. Mrvar, *Pajek* – program for large network analysis. *Connections*, 1998, **21**, 47-57.

F. Boyer, A. Viari. Ab initio reconstruction of metabolic pathways. *Bioinformatics,* 2003, **7** 19:ii26–ii34, DOI: 10.1093/bioinformatics/btg1055.

U. Brandes, Faster evaluation of shortest-path based centrality indices. *Konstanzer Schriften in Mathematik und Informatik* Nr. 120, ISSN 1430-3558, 2000.

U. Brandes, A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 2001, **25** (2), 163-177.

A. Broder, R. Kumar, F. Maghoul, R. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the Web. *Computer Networks*, 2000, **33**, 309-320.

R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research,* **38**:D473–D479, 2010. DOI:10.1093/nar/gkp875

S. Chakraborty, E. Fischer, A. Matsliah, R. Yuster. Hardness and algorithms for rainbow connectivity. In *Proc. 26th Internat. Sympos. Theor. Aspects Comput. Sci.*, pages 243-254, 2009. arXiv:0902.1255v2.

R. Christmas, I. Avila-Campillo, H. Bolouri, B. Schwikowski, M. Anderson, R. Kelley, N. Landys, C. Workman, T. Ideker, E. Cerami, R. Sheridan, G.D. Bader, and C. Sander. *Cytoscape*: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Am Assoc Cancer Res Educ Book*, pp. 12-16, 2005. http://www.cytoscape.org.

A. Clauset, M. E. J. Newman, C. Moore. Finding community structure in very large networks. *Physical Review E,* 2004, **70**, 066111.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to algorithms. *MIT Press*, 2nd edition, 2001.

D. Croes, F., Couche, S. J., Wodak, J. van Helden. Metabolic pathfinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Research*, **33**:W326–W330, 2005. DOI:10.1093/nar/gki437.

N. C. Duarte, M. J. Herrgard, B. Ø. Palsson, Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 2004, **14**, 1298-1309. DOI: 10.1101/gr.2250904.

N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. Ø. Palsson, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 2007 **104** (6), 1777-1782.

I. Famili, J. Förster, J; Nielsen, B. Ø. Palsson, *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences of the United States of America*, 2003 **100** (23), 13134-13139.

K. Faust, D. Croes, J. van Helden, Metabolic pathfinding using RPAIR annotation. *J. Mol. Biol.*, 2009, 388:390–414. DOI:10.1016/j.jmb.2009.03.006.

J. Förster, I. Famili, P. Fu, B. Ø. Palsson, J. Nielsen, Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 2003, **13**, 244-253. DOI: 10.1101/gr.234503.

C. Francke, R. J. Siezen, B. Teusink, Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 2005, **13** (11), 550-558. DOI:10.1016/j.tim.2005.09.001.

L. C. Freeman, A set of measures of centrality based on betweenness. *Sociometry*, 1977, **40**, 35-41.

M. R. Garey, D. S. Johnson, Computers and intractability; A guide to the theory of NP-completeness. W.H. Freeman, 1979.

A. P. Heath, G. N. Bennett, L. E. Kavraki. Finding metabolic pathways using atom tracking. *Bioinformatics*, 2010, **26**:1548–1555. DOI:10.1093/bioinformatics/btq223.

M. Huss, P. Holme, Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Systems Biology*, 2007, **1** (5), 280-285.

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

B. H. Junker, C. Klukas, F. Schreiber, *VANTED*: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 2006, **7** (109), DOI: 10.1186/1471-2105/7/109.

C. Kaleta, L. F. de Figueiredo, S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Research*, 2009, **19**:1872–1883.

T. Kamphans, M. Stelzer. SPUL: Shortest path with unique labels. C++ program, 2008. http://www.kamphans.de/spul.html.

M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 2010, **38**:D355–D360. DOI:10.1093/nar/gkp896.

C. Klukas, B. H. Junker, F. Schreiber, The *VANTED* software system for transcriptomics, proteomics and metabolomics analysis. *Journal of Pesticide Science*, 2006, **31** (3), 289-292.

M. Kotera, R. Yamamoto, K. Tonomura, M. Hattori, T. Komeno, S. Goto, M. Oh, J. Yabuzaki, M. Kanehisa, *RPAIR*: A reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics,* 2004, **15**, P062.

R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, The Web and Social Networks. *Computer*, 2002, **35** (11), 32-36.

H. Ma, A. P. Zeng (a), Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics,* 2003, **19** (2), 270-277.

H. Ma, A. P. Zeng (b) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics,* 2003, **19** (11), 1423-1430. DOI: 10.1093/bioinformatics/btg177.

H. Ma, X. M. Zhao, Y. J. Yuan, A. P. Zeng, Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, 2004, **20** (12), 1870-1876. DOI: 10.1093/bioinformatics/bth167.

H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, I. Goryanin, The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, 2007, **3**, DOI: 10.1038/msb4100177.

D. McShan, S. Rao, and I. Shah. *PathMiner*: Predicting metabolic pathways by heuristic search. *Bioinformatics*, 2003, **19**:1692–1698. DOI: 10.1093/bioinformatics/btg217.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg, 2005, Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380. DOI: 10.1038/nature03959.

G. Michal, D. Schomburg, Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology. Wiley-Interscience; 2 edition; 2011 (in print).

J. C. Nacher, J. M. Schwartz, M. Kanehisa, T. Akutsu, Identification of metabolic units induced by environmental signals. *Bioinformatics*, 2006, **22** (14), e375-e383. DOI: 10.1093/bioinformatics/btl202.

F. C. Neidhardt, J. L. Ingraham, M. Schaechter, Physiology of the bacterial cell: a molecular approach. *Sinauer Associates*, 1990.

M. E. J. Newman, The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, 2003, **45** (2), 167-256.

M. E. J. Newman (a), Detecting community structure in networks. *European Physics Journal B*, 2004, **38**, 321-330.

M. E. J. Newman (b), Fast algorithm for detecting community structure in networks. *Physical Review E,* 2004, **69**, 066133.

M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Physical Review E, Statistical, nonlinear, and soft matter physics*, 2004, **69**, 026113.

M. E. J. Newman, Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**, 8577-8582. DOI:10.1073/pnas.0601602103.

K. Radrich, Y. Tsuruoka, P. Dobson, A. Gevorgyan, N. Swainston, G. Baart, J.-M. Schwartz, Integration of metabolic databases for the reconstruction of genome-scale metabolic networks, *BMC Systems Biology*, 2010, **4**:114.

S. A. Rahman, D. Schomburg, Observing local and global properties of metabolic pathways: *load points* and *choke points* in the metabolic networks. *Bioinformatics*, 2006, **22** (14), 1767-1774. DOI: 10.1093/bioinformatics/btl181.

M. Rosa da Silva, Bioinformatics tools for the visualization and structural analysis of metabolic networks. PhD thesis, TU Braunschweig, 2006.
http://csresources.sourceforge.net/ShortestPath, http://www.helmholtz-hzi.de/systemsbiology

M. Rosa da Silva, J. Sun, H. Ma, F. He, A. P. Zeng, Metabolic networks. In: *Analysis of Biological Networks* (Junker, BH; Schreiber, F., editors), Wiley Series in Bioinformatics, Wiley & Sons, 233-253, 2007.

M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele, and D. Schomburg. *BRENDA*, the enzyme information system in 2011. *Nucleic Acids Research*, 2011, **39**, D670-D676, DOI:10.1093/nar/gkq1089.

J. Sun, X. Lu, U. Rinas, A. P. Zeng, Metabolic peculiarities of Aspergillus niger disclosed by comparative metabolic genomics. *Genome Biology*, 2007, **8**, R182.

A. Wagner, D. A. Fell, The small world inside large metabolic networks. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, 2001, **268**, 1803-1810.

Table 1: Statistical comparison of the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases.

	Former database	Upgraded database	Difference/change in reversibility
Reactions total	3565	6851	3286
Number of irreversible reactions (1)	1789	4304	2515/235
Number of reversible reactions (0)	1776	2547	771/731
Number of all (complete) EC numbers	3115 (2643)	3525 (2943)	410 (300)
Spontaneous reactions	35	76	41

Table 2: Comparison of the numbers of reactions involved in the secondary metabolism and the glycolysis pathway in the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases. Only the reactions having at least one reactant pair were counted.

	Number of reactions	
Pathway	Former database	Upgraded database
Glycolysis/Gluconeogenesis	39	47
Terpenoid biosynthesis	44	146
Diterpenoid biosynthesis	0	83
Monoterpenoid biosynthesis	0	24
Limonene and pinene degradation	0	59
Indole and ipecac alkaloid biosynthesis	0	68
Flavonoid biosynthesis	70	76
Alkaloid biosynthesis I	47	53
Alkaloid biosynthesis II	15	41
Penicillin and cephalosporin biosynthesis	4	17
Streptomycin biosynthesis	5	19
Tetracycline biosynthesis	1	12
Clavulanic acid biosynthesis	0	8
Puromycin biosynthesis	0	10
Novobiocin biosynthesis	0	35

Table 3: Comparison of organism-specific metabolic networks and their giant strong components (GSC) reconstructed based on the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases.

	Organism-specific metabolic networks*					
	eco old	eco new	anig old	anig new	hsa old	hsa new
Number of metabolites (nodes)	1156	1718	1513	2280	1341	2174
Number of reactions (arcs/edges)	1217	2172	1593	2871	1498	2897
Number of metabolites GSC (nodes)	256	485	288	653	386	718
Number of reactions GSC (arcs/edges)	356	792	389	996	511	1173
Number of modules GSC (reaction graph)	9	9	8	5	12	8

*eco = *E. coli* , anig = *A. niger* , hsa = *H. sapiens*

Table 4: The 23 potential hub metabolites and their number of first neighbors in the organism-specific metabolic networks, ranked according to the metabolite/organism-specific network with the highest value (*i.e.* acetyl-CoA/anig new). Only for the substances isocitrate and glycerate 3-phosphate no differences in the number of neighbors occurred between the networks based on the former (Ma and Zeng, 2003 a) and upgraded bioreaction datasets for all three organisms tested.

	Organism-specific metabolic networks*					
Metabolite name	anig new	anig old	eco new	eco old	hsa new	hsa old
Acetyl-CoA	51	12	44	15	40	13
Pyruvate	36	17	50	24	16	11
Acetate	30	6	15	6	18	5
D-Galactose	24	11	23	10	25	12
L-Glutamate	23	9	21	9	25	11
D-Glucose	15	3	20	7	16	2
Carboxylate	15	6	6	-	16	-
D-Fructose 6-phosphate	14	9	13	7	12	8
Succinate	11	11	11	5	5	3
Propanoyl-CoA	11	11	11	8	11	10
D-Glucose 6-phosphate	11	5	11	6	7	4
5-Phospho-*alpha-D*-ribose 1-diphosphate	10	4	12	4	11	4
L-Aspartate	10	7	12	8	13	9
Oxaloacetate	9	6	8	8	8	7
D-Ribose 5-phosphate	9	8	10	8	8	7
D-Xylulose 5-phosphate	8	6	8	6	6	5
D-Glyceraldehyde-3-phosphate	8	9	13	13	7	9
Fumarate	7	4	7	4	6	4
Malonyl-(ACP) CoA	6	2	4	2	5	2
Isocitrate	6	6	6	6	4	4
Citrate	4	2	6	5	5	4
Phosphoenolpyruvate	4	3	7	4	6	3
Glycerate 3-phosphate	4	4	5	5	4	4

*anig = *A. niger* , eco = *E. coli* , hsa = *H. sapiens*

- = no occurrence

Table 5: Changes in node-specific parameters of organism-specific metabolic networks and their giant strong components (GSC) reconstructed based on the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases.

Number of connected pairs		517,346 (38 %)	1,700,176 (57 %)	716,030 (31 %)	2,788,696 (53 %)	663,238 (36 %)	2,801,572 (59 %)
Average node degree whole network	input/output	1.53	1.67	1.49	1.61	1.55	1.74
	all	2.11	2.53	2.11	2.52	2.24	2.67
GSC	input/output	2.43	2.73	2.33	2.61	2.27	2.81
	all	2.97	3.59	3.02	3.46	2.99	3.66

*eco = *E. coli* , anig = *A. niger* , hsa = *H. sapiens*

[%] = percentage of the number of connected pairs relating to the number of all possible pairs in the netw ork

Table 6: Estimated best modularity, number of modules and pathways found, number of nodes and arcs/edges for the giant strong component (GSC) of the organism-specific reaction graph based on the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases.

	Organism-specific GSC based on a reaction graph*					
	eco old	eco new	anig old	anig new	hsa old	hsa new
Number of reactions (nodes)	436	776	467	986	609	1108
Number of arcs/edges	1769	5620	1829	6880	2032	7309
Best modularity	0.67	0.55	0.75	0.52	0.79	0.65
Number of modules	9	9	8	5	12	8
Number of pathways	48	96	55	110	54	98

*eco = *E. coli* , anig = *A. niger* , hsa = *H. sapiens*

Table 7: Changes in further parameters of organism-specific metabolic networks and their giant strong components (GSC) reconstructed based on the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases.

	Organism-specific metabolic networks*					
	eco old	eco new	anig old	anig new	hsa old	hsa new
Shortest path between *D*-Glucose and Pyruvate	6	5	8	7	9	7
Network diameter	44	21	83	43	73	40
Average path length whole network (ALW)	10.64	7.54	16.37	9.37	15.53	9.34
Average path length GSC (ALG)	8.73	7.14	10.98	8.61	15.61	8.69
Betweenness centrality GSC	0.39	0.34	0.32	0.44	0.41	0.30
Closeness centrality (*all*) GSC	0.16	0.19	0.12	0.17	0.10	0.14

Table 8: Organism-specific glucose subnetworks with number of metabolites, the average distance (*i.e*. average path length, AL) among reachable pairs and the distance between glucose and its most distant metabolite. The data were generated using the *BFS*-algorithm (*breadth first search*; Broder *et al*., 2000; Cormen *et al*., 2001; Pajek, Batagelj and Mrvar, 1998) together with a directed graph. The organism-specific networks are based on the former (Ma and Zeng, 2003 a) and upgraded databases.

	Organism-specific glucose subnetworks*					
	eco old	eco new	anig old	anig new	hsa old	hsa new
Number of metabolites (nodes)	440	769	507	1044	578	1120
Average distance (*i.e.* AL) among reachable pairs	9.41	7.58	11.52	8.09	10.99	7.88
Distance of most distant vertices	28	23	28	19	35	21

*eco = *E. coli* , anig = *A. niger* , hsa = *H. sapiens*

Table 9: Comparison of reconstructed organism-specific metabolic networks based on the upgraded database with also human-curated literature-based networks.

	Organism-specific metabolic networks*							
	hsa					sce		
	Edinburgh	*Recon I*	Upgraded	Difference			Upgraded	Difference
	Ma *et al.*	Duarte *et al.*	database	to upgrade		Förster *et al.*	database	to upgrade
Number of metabolites (nodes)	n.m.	2712	2174	-	- 538	584	1482	+ 898
Number of reactions (arcs/edges)	2823	3311	2897	+ 74	- 414	1175	1800	+ 625

*hsa = *H. sapiens* , sce = *S. cerevisiae*

n.m. = not mentioned in the publication

35

Table 10: Examples of running time and found paths starting in vertex 1 (XEON CPU 3.0 GHz, 16 GB memory).

	Vertices	Edges	Vertices reachable from start	Paths found			Running time in min		
				Alg A	Alg B	B with preproc.	Alg A	Alg B	B with preproc.
A. niger	2547	7818	1488	1283	1369	1382	1.2	23.4	23.6
E. coli	1895	5525	1111	911	1012	1021	1.3	35.6	35.9
H. sapiens	2474	7873	1614	1347	1507	1526	1.2	20.2	20.5

Table 11: Comparison on the number of paths found in several organism-specific metabolic networks[1].

	uur -CM	uur +CM	mpn -CM	mpn +CM	bbu -CM	bbu +CM	mge -CM	mge +CM
Number of SP	2372	17038	2191	6652	6357	39552	6793	36246
SP false[2]	1113	7736	903	2723	3887	22688	3925	17786
SP correct	1259	9302	1288	3929	2470	16864	2868	18460
Number of SPUL	1308	13306	1601	4577	2513	22809	3061	22281

[1] eco = *E. coli*, anig = *A. niger*, hsa = *H. sapiens*, uur = *U. urealyticum*, mpn = *M. pneumoniae*, bbu = *B. burgdorferi*, mge = *M. genitalium,* +/- CM: with/without currency metabolites

[2] SP containing more than one arc/edge belonging to the same reaction, therefore biological not meaningful

40

Table A1: List of all potential currency metabolites (CM), their frequency in the former (Ma and Zeng, 2003 a) and upgraded bioreaction databases and their number of direct neighbors in the organism-specific metabolic networks reconstructed based on the data of both databases.

	Frequency of compound in reactions		Organism-specific metabolic networks*					
Compound	in former database	in upgraded database	eco old	eco new	anig old	anig new	hsa old	hsa new
H_2O	1255	2236	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
H^+	359	1269	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
O_2	379	858	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
$NADP^+$	387	725	1	2	1	2	1	4
NADPH	386	722	-	1	-	1	-	1
NAD^+	392	666	2	4	4	7	6	9
NADH	387	657	-	1	-	1	-	1
ATP	311	467	3	21	4	17	4	20
CO_2	250	430	2	4	4	5	2	2
Pi	253	395	-	1	-	1	-	1
CoA	196	371	3	4	3	4	1	3
ADP	234	333	4	5	4	4	5	6
NH_3	212	297	-	2	3	4	1	1
PPi	158	288	1	2	-	1	-	1
S-Adenosyl-*L*-homocysteine	91	236	2	2	4	4	2	2
UDP	95	224	3	6	3	4	3	6
Acceptor	73	189	-	-	-	-	-	-
Reduced acceptor	73	187	-	-	-	-	-	-
H_2O_2	98	162	-	2	-	2	-	2
AMP	89	160	4	7	9	13	12	17
Glutathione	27	65	6	28	4	27	4	28
CMP	48	64	2	4	2	2	3	2
dTDP	6	48	2	2	2	2	2	2
HCl	8	45	-	-	2	-	-	-
3'-Phosphoadenylyl sulfate	21	45	3	2	3	2	2	2
Adenosine 3',5'-bisphosphate	22	44	2	3	3	4	2	2
GDP	27	40	4	6	3	3	3	5
FAD	29	40	1	1	1	1	1	1
$FADH_2$	29	38	1	-	-	-	-	-
CTP	27	31	3	7	2	9	2	12
GTP	23	31	7	9	5	9	6	10
Oxidized ferredoxin	17	31	1	1	1	1	1	1
Reduced ferredoxin	15	31	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
H_2SO_4	20	29	1	-	2	1	2	1
UMP	23	28	4	7	4	6	5	8
H_2SO_3	19	26	3	1	4	2	2	1
UTP	23	26	3	8	2	6	4	7
Oxidized glutathione	12	26	1	1	1	1	1	1
HNO_2	17	25	2	3	2	2	-	-
Acyl-carrier protein	22	23	2	2	1	1	-	2
H_2S	10	20	1	1	1	1	-	-
ITP	20	20	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized thioredoxin	15	20	1	1	1	1	1	1
Reduced thioredoxin	15	20	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Cl^- (Chloride)	2	19	-	-	-	-	-	-
IDP	19	19	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome c	2	19	1	-	1	1	1	1
H_2	14	18	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome c	13	18	1	-	1	1	1	1
GMP	14	16	6	6	5	6	9	9
CO	2	13	-	1	-	-	-	-
NDP	2	13	-	3	-	1	-	3
CDP	11	13	3	4	3	3	3	4
dATP	11	13	2	3	1	2	1	2

	Frequency of compound in reactions		Organism-specific metabolic networks[*]					
Compound	in former database	in upgraded database	eco old	eco new	anig old	anig new	hsa old	hsa new
e^-	4	12	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
IMP	12	12	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dADP	12	12	3	3	3	3	3	3
PQQ	11	12	-	-	-	-	-	-
HNO_3	8	11	1	2	1	1	-	-
HS_2O_3	8	11	1	-	1	-	1	-
FMN	5	11	2	3	2	3	2	2
OH^-	1	10	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
NO	8	10	-	1	-	-	-	-
$PQQH_2$	10	10	-	-	-	-	-	-
Ubiquinone	9	10	3	3	2	3	2	3
Cl^- (Chloride ion)	1	9	-	-	-	-	-	-
dCMP	8	9	2	2	2	2	3	3
dGTP	7	9	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized rubredoxin	4	9	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced rubredoxin	4	9	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
PPPi	5	8	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dCTP	6	8	3	4	1	2	1	2
dTMP	5	8	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dUMP	6	8	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ubiquinol	7	8	1	1	1	1	1	1
S	5	7	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
H_2Se	7	7	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
SeO_3	6	7	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
HBr	no occurrence[+]	7	-	-	-	-	-	-
dTTP	6	7	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dUTP	7	7	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Fe^{2+}	no occurrence[+]	6	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dCDP	6	6	3	3	2	2	3	3
dGDP	6	6	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dUDP	6	6	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced FMN	0	6	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
HSO_3^-	1	5	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome b5	1	5	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome b5	1	5	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Electron-transferring flavoprotein	2	5	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Red. elec.-transf. flavo.	2	5	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
$O_2^{\cdot -}$	2	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Fe^{3+}	no occurrence[+]	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dGMP	4	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized adrenal ferredoxin	5	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced adrenal ferredoxin	5	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized flavoprotein	2	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced flavoprotein	2	4	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
I_2	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
I^-	2	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
H_2SeO_4	2	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Alkylphosphonate	1	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dAMP	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferredoxin	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome c2	2	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome c2	2	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Quinone	no occurrence[+]	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Hydroquinone	no occurrence[+]	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Coenzyme F420	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced coenzyme F420	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized dithiothreitol	0	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Dithiothreitol	0	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Amino group donor	0	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
RING	3	3	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Hg^{2+}	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
HI	2	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
H_3PSeO_3	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.

	Frequency of compound in reactions		Organism-specific metabolic networks*					
		in upgraded				anig	hsa	hsa
Compound	in former database	database	eco old	eco new	anig old	new	old	new
Mg^{2+}	2	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Fe	6	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized azurin	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced azurin	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Cytochrome c	2	2	n.e.	-	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome c3	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome c-553	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome c-553	1	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome b1	2	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome b1	2	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Apocytochrome c	2	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Donor	0	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized donor	0	2	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
$H_2S_2O_3$	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
SO_2	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Mn	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Br^-	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
NMP	no occurrence+	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Cytochrome c3	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome c3	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferricytochrome b-561	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Ferrocytochrome b-561	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Flavodoxin semiquinone	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Dihydroflavodoxin	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized flavodoxin	no occurrence+	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced flavodoxin	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Decylubiquinone	no occurrence+	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Decylubiquinol	no occurrence+	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized plastocyanin	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Reduced plastocyanin	1	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Oxidized putidaredoxin	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
Putidaredoxin	0	1	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
NTP	no occurrence+	no occurrence+	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dNTP	no occurrence+	no occurrence+	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dNDP	no occurrence+	no occurrence+	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
dNMP	no occurrence+	no occurrence+	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.

*eco = *E. coli*, anig = *A. niger*, hsa = *H. sapiens*
- = no occurence of substance in the organism-specific metabolic network
n.e. = not estimated
+ = no occurrence in the *KEGG* database
0 = no occurrence in the former database

Table A2: List of reactions of the upgraded bioreaction database containing a special error type or inconsistency. At all 107 affected reactions were identified and had to be improved or corrected based on actual biological knowledge and literature as far as possible.

Reaction index	Error type:	stoichiometrical not balanced	wrong direction in pathway map irreversible reaction	wrong com-pound/ not specified	wrong pathway map	error in chemical drawing	wrong reaction	wrong reactant pair
R00011				x				
R00329				x				
R00631			x					
R00632			x					
R00634			x					
R00778			x					
R00993		x						
R01303			x					
R01347		x						
R01348		x						
R01367			x					
R01427					x			
R01433			x					
R01679			x					
R01726					x			
R01827								x
R02116					x			
R02139							x	
R02222		x						
R02300				x				
R02442		x						
R02724		x						
R02764			x					
R03124		x						
R03376						x		
R03551			x					
R03643							x	
R03765			x					
R03933		x		x				
R04020			x					
R04044						x		
R04097			x					
R04131			x					
R04132			x					
R04224			x					
R04399			x					
R04461		x		x				
R04721								x
R04776				x				
R04809			x					
R04813			x					
R04895			x					
R04904		x		x			x	
R04908		x				x		
R04931				x			x	
R05083			x					
R05091					x			
R05226			x					
R05302		x						
R05380			x					
R05419		x						
R05465							x	
R05526					x			
R05527					x			
R05528					x			
R05534					x		x	
R05545								
R05552			x					
R05596			x					
R05599			x					

Reaction index	Error type:	stoichiometrical not balanced	wrong direction in pathway map irreversible reaction	wrong compound/ not specified	wrong pathway map	error in chemical drawing	wrong reaction	wrong reactant pair
R05600			x					
R05601			x					
R05602			x					
R05740		x						
R05771			x					
R05775			x					
R05821		x						
R05843			x					
R05864			x					
R06138				x				
R06348					x			
R06369			x					
R06449						x		
R06458		x						
R06459		x						
R06627			x					
R06635		x						
R06636		x						
R06637		x						
R06641		x						
R06643		x						
R06644		x						
R06645		x						
R06731		x						
R06759		x						
R06897			x					
R06942			x					
R06952					x			
R07291			x					
R07390					x			
R07475					x			
R07692					x			
R07693					x			
R07780			x					
R07848					x			
R07849					x			
R07890					x			
R07894					x			
R07898					x			
R07934					x			
R07935					x			
R07936					x			
R07937					x			
R07950					x			
R07951					x			
R07952					x			
R07953					x			

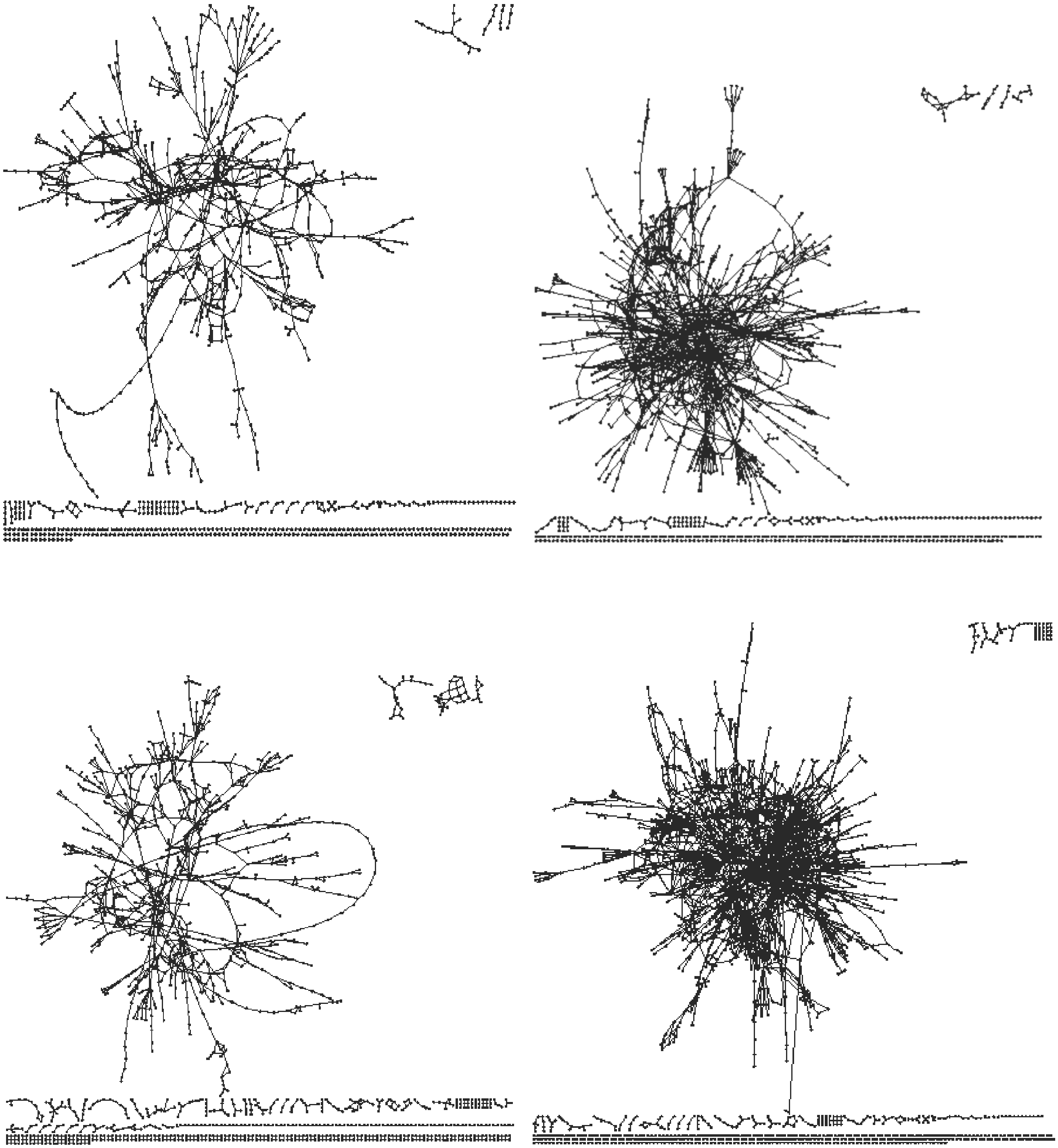


Figure A1: Comparison of the organism-specific metabolic networks for the two model organisms *E. coli* (eco) and *A. niger* (anig), top down, reconstructed from the former (left) and upgraded (right) bioreaction databases. The networks based on the upgraded data material show a higher complexity and less disconnected parts than those based on the former database (Ma and Zeng, 2003 a).