Structural analysis of metabolic networks based on flux centrality

Dirk Koschützki, B.H. Björn H. Junker, J. Jörg Schwender, Falk Schreiber

Structural Analysis of Metabolic Networks based on Flux Centrality

Authors: Dirk Koschützki[a,*], Björn H. Junker[b], Jörg Schwender[c], and Falk Schreiber[b,d]

[a]Department of Computer and Electrical Engineering, Furtwangen University of Applied Sciences, Robert-Gerwig-Platz 1, 78120 Furtwangen, Germany;

[b]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany;

[c]Biology Department, Brookhaven National Laboratory, 50 Bell Avenue, Upton, NY 11973, USA;

[d]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle, Germany;

[*]Corresponding author (Dirk.Koschuetzki@hs-furtwangen.de, Tel: +49 7723 9202327, Fax: +49 7723 9201109)

Abstract

Metabolic reactions are fundamental to living organisms, and a large number of reactions simultaneously occur at a given time in living cells transforming diverse metabolites into each other. There has been an ongoing debate on how to classify metabolites with respect to their importance for metabolic performance, usually based on the analysis of topological properties of genome scale metabolic networks. However, none of these studies have accounted quantitatively for flux in metabolic networks, thus lacking an important component of a cell's biochemistry.

We therefore analyzed a genome scale metabolic network of *Escherichia coli* by comparing growth under 19 different growth conditions, using flux balance analysis and weighted network centrality investigation. With this novel concept of *flux centrality* we generated metabolite rankings for each particular growth condition. In contrast to the results of conventional analysis of genome scale metabolic networks, different metabolites were top-ranking dependent on the growth condition. At the same time, several metabolites were consistently among the high ranking ones. Those are associated with pathways that have been described by biochemists as the most central part of metabolism, such as glycolysis, tricarboxylic acid cycle and pentose phosphate pathway. The values for the average path length of the analyzed metabolite networks were between 10.5 and 12.6, supporting recent findings that the metabolic network of *E. coli* is not a small-world network.

Keywords

Network centralities, Metabolism, Network analysis, Flux Balance Analysis

Introduction

Since the initial discovery of common design principles in various kinds of networks (Barabási and Albert, 1999; Watts and Strogatz, 1998), many studies have addressed the issue how topology relates to network function (Ekman *et al.*, 2006; Jeong *et al.*, 2001; Milo *et al.*, 2002). This especially holds true for metabolic networks, which have been reconstructed for different organisms from genome information (Duarte *et al.*, 2004; Edwards and Palsson, 2000). By graph theoretical analysis of the topology of genome scale metabolic networks it is possible to infer hypotheses about the functionality of metabolism. An example is the identification of common design principles of metabolic networks, which are thought to be responsible for robustness and error tolerance based on the comparison of parameters such as average path length (APL, which is the average of the shortest path length over all pairs of nodes in the network), connectivity distribution and substrate ranking between different organisms (Jeong *et al.*, 2000). Typically, such studies mainly analyze the effect of the presence or absence of a connection on viability or optimal performance. It has been shown that highly connected metabolites (hubs) are more important for the operation of a network than metabolites with low connectivity (Jeong *et al.*, 2000). In several studies metabolites have been ranked based on their place/position within the network to determine their importance, a process called *centrality analysis*, leading to significantly different results ((Arita, 2004; Fell and Wagner, 2000; Jeong *et al.*, 2000; Ma and Zeng, 2003a); see Discussion for details). All these studies describe biochemical reactions in metabolic networks as uniform links connecting metabolites and do not take into account stoichiometric constraints or any quantitative flux information about a particular connection. This may be an oversimplification considering that the real performance of a metabolic network is not a connectivity property but the conversion of matter from substrates towards end products, which furthermore occurs at different rates. One of several possible classifications of substrates of a metabolic network might be consisting of three categories: a) metabolites such as pyruvate which constitute the flux of carbon from metabolic substrates to end products, b) cofactors (e.g. ATP, NADH, NADPH) which are constantly detached from and attached to metabolic intermediates, and c) inorganic substances (e.g. H_2O, P*i*) which are mostly not transformed into metabolic end products. Cofactors and inorganic substances may be

summarized under the term currency metabolites (Ma and Zeng, 2003a). They are generally highly connected as a result of their role in metabolism. However, rather than simply drawing conclusions on the importance of a metabolite from the topology of the network, a classification that accounts for the role of a metabolite in the flow of carbon from substrates to end products would be more valuable. Recognizing this problem Arita used the concept of 'carbon atomic traces' to add the aspect of material transport into network analysis by following the fate of individual carbon atoms through the network (Arita, 2004). In consequence, the top hubs in his metabolic networks are less occupied by currency metabolites, but more by metabolites of carbon metabolism. However, as the author did not consider the quantity of carbon flux through the network, the mentioned study is still of qualitative nature.

Here we further develop the concept of metabolic network analysis by introducing the concept of *carbon flux centralities* (or short *flux centrality*) which combines centrality analysis with flux balance analysis (FBA) under various growth conditions (see Methods section for details). Based on this concept it was possible to a) rank metabolites according to their importance for the metabolic processes, b) derive a meaningful core metabolism, and c), for central metabolism, automatically cluster metabolites into pathways.

Results

Flux balance analysis

We used the *Escherichia coli* genome scale metabolic network iJR904 ((Reed *et al.*, 2003); see Methods) to simulate optimal growth under 19 different growth conditions (aerobic and anaerobic) with 15 different single carbon sources including sugars, sugar alcohols, organic acids and amino acids. The resulting flux vectors are given in SI Table 2. Simulation of aerobic growth on glucose, glycerol and acetate predicted a biomass yield of 0.51, 0.58 and 0.37 g dry weight (DW) per g substrate consumed, respectively, which is in the range of experimental values reported for *E. coli* MG1655 grown in chemostat culture (Weikert *et al.*, 1997). Simulation of anaerobic growth on glucose, ribose, sorbitol or gluconate resulted in excretion of the

fermentation products acetate, ethanol, formate, glycolate and succinate, and therefore much reduced biomass yield (see Methods section for details about the 19 growth conditions).

Carbon-balanced metabolite network and flux centrality

Within the multitude of biochemical reactions converting substrates to biomass during organism growth, carbon chains can be seen as the essential structures being transformed through the network, while protons and oxygen are often exchanged with free H_2O. In a similar manner, phosphate groups are mainly transiently attached to the transformed carbon chains. Therefore, a meaningful way to analyze metabolic networks is to describe the links connecting metabolites based only on carbon transitions. Hence, for calculating a *flux centrality* value we define that each metabolic connection is weighted according to the carbon flux passing through it. Consequently, all non-carbon connections have zero weight, which results in a considerable reduction of network size. The flux centrality value of a certain carbon metabolite defined in this way is a measure for the maximum rate at which other reachable metabolites can be produced from this metabolite, i.e. it indicates the importance of this metabolite for biomass formation.

To calculate the flux centrality, each of the 19 growth condition specific bipartite metabolite-enzyme networks has to be transformed into a unipartite metabolite network. As all metabolite networks should be carbon-balanced, we need to incorporate the information how many carbon atoms are transferred from one metabolite to another in any reaction. Each of the 19 growth conditions therefore resulted in a weighted metabolite network representing the carbon fluxes (see Methods section for details). The *flux centralities* calculated for these 19 weighted networks are shown in SI Table 3.

Prior to detailed analysis (see below), the flux centrality values of all metabolites were summed over all growth conditions and the metabolites ranked accordingly. From the 395 ranked metabolites the Top-30 shown in Table 1 constitute most of central metabolism, which traditionally is divided into glycolysis, the tricarboxic acid (TCA) cycle and the pentose phosphate pathway (PPP). The colors of the metabolites in Table 1 match their occurrence in the visualization of the pathways shown in Figure 1. Virtually all glycolytic intermediates, TCA cycle intermediates and intermediates of the PPP are found within the top 30 ranked metabolites. In addition to these, there are several other metabolites present in this group: coenzyme

A (CoA), part of several anabolic and catabolic routes; malonyl-CoA, which is an intermediate in fatty acid synthesis; aspartate (ASP) and 5-Phospho alpha-D-ribose 1-diphosphate (PRPP), which are precursors of several amino acids and nucleotides; CO_2, which is integrated into organic compounds, e.g. via anapleurotic reactions.

Clustering of flux centrality values reveals modularity of metabolism

For further analysis of the data obtained in this study, the original flux centrality values for all metabolites under all growth conditions were visualized as a heatmap (Figure 2), with 395 metabolites and 19 growth conditions clustered by similarity. For presentation purposes, in Figure 2 we divided the metabolites into 6 groups (1-6), and the growth conditions into 3 groups (A-C).

Within most growth conditions, the metabolites of group 1 (Figure 3a) have outstanding high flux centrality values. Although under anaerobic growth (group C) most centrality values are smaller than under aerobic growth, most of group 1 metabolites still have relative high centrality values. A close-up into the hierarchical clustering tree (see Methods) of metabolite group 1 (Figure 3a) reveals that the respective clusters are almost exclusively constituted of metabolites from central metabolism, i.e. glycolysis, the TCA cycle, and the PPP, essentially consistent with the metabolites found in the top-30 list in Table 1. Metabolites of these three pathways are grouped into distinct sub-clusters (Figure 3a). The clustering reveals that under different conditions the intermediates between fructose 6-phosphate (F6P) and phosphoenol pyruvate (PEP) generally behave similar with respect to flux centrality, supporting the perception of glycolysis as one functional unit within the central metabolic network. Similar findings apply for the TCA cycle and the PPP (compare Figure 1 with Figure 3). Some findings are possibly different to the classical pathway definition: pyruvate (PYR) and acetyl-CoA (AcCoA) are assigned closer to the TCA cycle than to glycolysis. Furthermore, glucose 6-phosphate (G6P), 6-phospho-D-glucono 1,5-lactone (6PGL), and 6-phospho-D-gluconate (6PGC) form a cluster that is distant from all other central metabolic pathways.

In Figure 2 the metabolite groups 2 and 4 refer to the 15 growth substrates. For a particular growth condition all substrates which are not supplied have centrality values of zero while the supplied growth substrates and intermediates of metabolic routes connecting to central metabolism have high centrality values. This is

exemplified by the clustering tree in Figure 3b (magnified from Figure 2, group 2), which shows some metabolites that only play an important role when the cells are growing on lactose as a substrate.

Groups 3 and 5 consist of metabolites which belong to biosynthesis of a multitude of biomass precursors including several amino acids, purines, pyrimidines and different cofactors. The centralities of these metabolites are relatively independent of the growth condition, resulting in diffuse clustering. This reflects the fact that 1) the total flux into biomass is spread among more than 20 major biomass constituents and 2) for all simulations the biomass composition does not change and therefore the anabolic pathways in clusters 3 and 5 have largely the same activity regardless of the growth substrate supplied. Finally, group 6 consists of metabolites that have a flux centrality near zero for all growth conditions.

With respect to the clustering of growth conditions (Figure 2, groups A-C) different metabolic states can be differentiated. It appears that group A in Figure 2 unifies aerobic growth on different sugars and related derivatives (sorbitol, gluconate). The substrates of group A need to be metabolized through glycolysis and/or the pentose phosphate pathway (PPP) to be used as carbon source for the majority of biomass components. Accordingly, intermediates of glycolysis and PPP have higher centrality values than intermediates of the TCA cycle. In contrast, group B in Figure 2 relates mainly to organic acids as substrates. These are accessible for the synthesis of many biomass components without passing through glycolysis or the PPP, such as amino acids derived from pyruvate, oxaloacetate and ketoglutarate. Here TCA cycle metabolites have higher centrality values than those of intermediates of glycolysis and PPP. Under anaerobic conditions (group C), TCA cycle intermediates are ranked lower than glycolysis and PPP, which is expected since in the absence of oxygen oxidative phosphorylation and thus cyclic flux through the TCA cycle is suppressed and the TCA cycle can only be used for anabolic functions. The clustering tree reveals that the centrality distribution upon growth on glycerol (GLYC) and on lactose (LCTS) results in outcomes different from the other groups. For example, lactose is different because as the only disaccharide substrate in this study, it requires unique conversions at the carbohydrate level.

Discussion

Centrality and path length in metabolic networks

Former studies on the topology of metabolic networks lead to valuable conclusions about the design principles of metabolism. Especially centrality analysis has drawn particular attention over the past few years. Jeong *et al.* investigated the topology in metabolic networks from 43 organisms (Jeong *et al.*, 2000). They found that the metabolic networks have small-world properties, with the degree distribution following a power-law function and the average path length of 3.2. However, this number mainly reflects the extraordinarily high connectivity of a few metabolites such as water and ATP. In the *E. coli* iJR904 metabolic reconstruction used in this study, already 19% of the reactions include ATP, and accordingly in the study of Jeong *et al.* many metabolites are directly connected via ATP (Jeong *et al.*, 2000). Consequently, the Top-10 list of most important metabolites for *E. coli* obtained by Jeong *et al.* contains only inorganic metabolites and cofactors like ATP, with the exception of glutamate, which is acting as amino group donor in numerous transaminase reactions (see SI Table 5). Wagner and Fell (2000; 2001) found for *E. coli* an APL of 3.8 using a similar approach, but omitting ADP, ATP and NAD(P)(H). Both studies suggest that any metabolite can be made out of any other one by an average of less than 4 enzymatic steps. According to the Top-10 lists of connectivity in both studies (SI Table 5), group donors or acceptors (cofactors) dominate among the most highly connected metabolites.

This dominance can be related to some ambiguity observed in the way multi-substrate/multi-product enzyme reactions are decomposed into metabolite - metabolite interactions by unipartite graphs. As an illustrative example, a hypergraph representing the transketolase reaction (Figure 4a) can be transformed into a unipartite graph where each substrate is connected to each of the products (Figure 4b). However, although the transketolase uses ribose 5-phosphate (R5P) and produces glyceraldehyde 3-phosphate (G3P) in equal amounts, there is no mass transfer between the two. It appears that to define connectivity, mass transfer is a more rigorous criterion than stoichiometry. Leaving out this connection as shown in Figure 4c is a more accurate representation of the transketolase reaction, showing that mass transfer defines connection.

In order to discriminate minor mass transfers against the main transfers in a reaction, Ma and Zeng (2003a) introduced *current* or *currency metabolites*, which are inorganic molecules, cofactors, functional group donors or acceptors, that play only the 'second role' in a reaction. By manual classification, they did not allow connections through these currency metabolites, thus leading to an increased value for the APL of 8.2. In a second analysis of the reconstructed metabolic networks Ma and Zeng used the overall closeness centrality to identify central metabolites (Ma and Zeng, 2003b). Applying this method to the metabolic network of *E. coli* resulted in a Top-10 list of metabolites, which is dominated by intermediates of glycolysis and the TCA cycle (see SI Table 5).

Furthermore, since the main structure of organic compounds is given by carbon, the mass transfer of carbon in enzyme reactions may be considered as a meaningful criterion for connectivity in metabolite networks. To mediate the conservation of structural carbon moieties throughout a pathway, Arita (2004) followed the path of individual carbon atoms in metabolic reactions annotated for *Escherichia coli*. Using the measure of degree centrality, this approach resulted in a Top-10 list of important metabolites that is significantly different to the ones reported previously (see SI Table 5) and an APL of 8.4 for the network, thus similar to the value calculated by Ma and Zeng (2003a).

More recently, Rahman and Schomburg (2006) introduced a new centrality measure for the analysis of metabolic networks. The computation of their "load points" centrality is based on the number of k-shortest-paths passing through a given node. It can be applied either to the metabolite network, to rank metabolites, or to the enzyme network, to rank enzymes. They applied this centrality to rank metabolites of two *Bacillus* species of which one is a pathogen (see SI Table 5).

While Ma and Zeng (2003a), Arita (2004) as well as Rahman and Schomburg (2006) consider connectivity qualitatively, our present study defines connectivity quantitatively by carbon flux between metabolites. In summary, we derived metabolite networks in which only those metabolites are connected which exhibit carbon transfer in a biochemical reaction, and in which edges are weighted by flux as derived from a the simulated flux distribution (FBA) and the number of carbon atoms that are transferred in a reaction. These networks were analyzed with the novel concept of *flux centrality*, based on weighted network centrality investigation. As a result, it could be shown here for the first time, how the centrality of a metabolite in the

network is dependent on the flux distribution and hence the physiological state (growth condition). In addition, some metabolites consistently have high centrality values. In fact, summing the centrality values over all different growth conditions identified metabolites of central metabolism as having general importance in *E. coli* (see Table 1). Furthermore, the Top-30 list of important metabolites does not contain metabolites that principally act as group donors, such as ATP, NADPH, and glutamate, which are donors of phosphate, hydrogen, and amino groups, respectively.

Finally, using the flux weighted metabolite network the APL for each growth condition was computed considering the presence or absence of connections under each condition. The APL over all growth conditions is 11.3 with a range from 10.7 (Proline, aerobic) to 12.5 (Ribose, anaerobic). These values are larger than any value reported before. For example, for directed networks in *E. coli* Ma and Zeng (2003a) and Arita (2004) reported an APL of 8.2 and 8.4, respectively. Defining connectivity based on carbon transfer and flux as in this study therefore leads to the so far highest APL values for *E. coli*. While our metabolite networks nevertheless are scale-free (data not shown), the high APL shows that they do not have the small-world property. This is consistent with the observation of Arita (2004).

In a study similar to ours, Kim *et al.* (2007) simulate *E. coli* under different growth conditions and analyze network performance concerning metabolites. However, they characterize functionality of metabolites in a network by introducing severe network perturbations (removal of metabolites or reactions). The property of essentiality they describe relates to functional redundancy. In our study flux and network topology are integrated in the unperturbed system and important metabolites are those that are the most heavily used.

Defining a core metabolism

In addition to deriving the importance of a metabolite for metabolism, several studies aimed at defining the core of metabolism by computational means. Thereby ‘metabolic core’ can be defined as a central, highly connected part of metabolism, which provides key precursors for many biosynthetic reactions and is probably evolutionary highly conserved.

For example, Ma *et al.* (2004) used a network clustering algorithm to decompose the metabolic network of *E. coli* into functional modules, the most central of which are pyruvate metabolism, glyoxylate metabolism, valine, leucine and isoleucine synthesis.

Investigating the *E. coli* MG1655 genome scale reconstruction iJE660a (Edwards and Palsson, 2000) based on flux balance analysis, Almaas *et al.* (2004) define a High Flux Backbone by removing from each metabolite all reactions but the largest incoming and outgoing flux. This way they found that the overall activity of metabolism is dominated by several reactions with very high fluxes, and that *E. coli* responds to changes in growth conditions by reorganizing the fluxes predominantly within this High Flux Backbone (Almaas *et al.*, 2004). In a follow-up paper, Almaas and coauthors performed flux balance analysis under a variety of different simulated growth conditions generated by defining growth on combinations of some of the 89 potential input substrates in addition to a minimal uptake basis (Almaas *et al.*, 2005). They identified a set of reactions that are connected and carry non-zero fluxes under all growth conditions, which they named the *metabolic core*. However, in the first study (Almaas *et al.*, 2004), several 'key' metabolites and enzymes are absent from the metabolic core (e.g. citrate and isocitrate are not connected). In the second study (Almaas *et al.*, 2005) only few glycolytic reactions and none of the TCA cycle are part of the metabolic core. These findings are not consistent with the idea that central metabolism should be a coherent (functional) sub-network and the often expressed notion that glycolysis or the TCA cycle is considered as a part of central metabolism in *E. coli* (Edwards and Palsson, 2000; Szyperski, 1995).

In this study we identified one group of metabolites with consistently high centrality values and with highly diverse distribution of flux centrality values dependent on growth condition (see Figure 2, metabolite group 1). This group is essentially identical to the metabolites of the 'classical' central metabolism (glycolysis, pentose phosphate cycle, tricarboxylic acid cycle). It characterizes a coherent piece of metabolism providing precursors for most of the remaining metabolism. Our approach of classification of metabolites based on flux centrality can therefore help to define core metabolism in an unsupervised and systematic way in different organisms.

Cluster metabolites into pathways

A number of studies discuss algorithmic methods to cluster metabolites into functional units called modules or pathways. In metabolic networks of 43 organisms, Ravasz *et al.* (2002) calculated the common number of neighbors between any pair of metabolites. Applying a clustering approach to the results they were able to divide the networks into distinct modules. Similarly, Ma *et al.* (2004) proposed a decomposition method that uses the path length between any two pairs of reactions as the dissimilarity measure, resulting in 11 subnetworks with defined biological functions. Holme *et al.* (2003) divided bipartite metabolic networks into components by successively removing reaction nodes with high betweenness centrality. They conclude that biochemical networks consist of outer shells encapsulating a core of the most connected substances.

In addition to the ranking of metabolites and defining the core of metabolism, our study allows also the identification of parts of the metabolic network that can be recognized as functional units usually described as pathways such as glycolysis, PPP and the TCA cycle (compare Figure 1, Figure 3a). The recognition of metabolites being part of these pathways by our analysis is for the largest part without gaps. All metabolites that are recognized as part of the oxidative and non-oxidative pentose phosphate pathway are found in the cluster shown in Figure 3a and are therefore a functional metabolic unit. In addition, metabolites between fructose 6-phosphate (F6P) and PEP are found in one sub-cluster (Figure 3a) marking off glycolysis. Pyruvate and acetyl-CoA are usually classified as glycolytic metabolites. However, the clustered heatmap (Figure 2) reveals that when the carbon source enters the central part of metabolism downstream of PEP, the TCA cycle intermediates, AcCoA and PYR exhibit the largest flux centrality. In contrast, in the case that the carbon source enters upstream of PEP, the other glycolytic intermediates exhibit the largest flux centrality. These results suggest that these putative glycolytic intermediates rather form a metabolic unit with the metabolites of the TCA cycle. In eukaryotes this classification is even more obvious since pyruvate is the metabolite entering the mitochondria to supply respiration.

In summary, the cluster analysis offers an unbiased modularization of the metabolism that recognizes sub-networks that have been described as pathways long ago by biochemists as universal catabolic routes that deliver cofactors, energy and key substrates for a multitude of biosynthetic pathways.

Conclusion

In this study, flux centrality was presented as a new concept integrating flux analysis and topological analysis of metabolic networks. From the perspective of biochemists, flux centralities give most meaningful insights into the potential of metabolites in *E. coli*. Clustering of carbon flux centralities for the different growth conditions recognized well-known central metabolism pathways as functional units. While our current study considers different possible substrates for *E. coli*, this method could be applied to any organism, and an analogous study in other organisms by variation in biomass composition could be highly useful in revealing the importance of different metabolites and the modularity in anabolism as related to the biosynthesis of different biomass compounds.

Materials and Methods

Flux Balance Analysis of E. coli *genome scale metabolic network*

In order to study the flux distribution in a genome scale metabolic network of *E. coli* we used the metabolic network of *E. coli* K-12 as defined by Reed *et al.* (2003). This model includes transmembrane transport reactions, carbon source utilization pathways, central carbon metabolism as well as the metabolic pathways responsible for the synthesis and degradation of amino acids, lipids, nucleic acids, vitamins and cofactors.

The model configurations for optimization were basically as described in (Reed *et al.*, 2003). The non-growth associated ATP maintenance reaction was fixed to 7.6 mmol h^{-1} g DW^{-1}. Carbon dioxide, ammonia, sulfate, sodium, potassium, phosphate, protons, water and iron (II) were allowed to freely enter and leave the system. All other metabolites marked as extracellular were allowed to freely leave the system. Aerobic growth on 15 substrates (acetate, alanine, glucose, gluconate, glycerol, ketoglutarate, lactate, lactose, malate, octadecanoate, proline, pyruvate, ribose, sorbitol, and succinate) was simulated as well as anaerobic growth on glucose, ribose, sorbitol, and gluconate. For all carbon sources the uptake rates were limited to the equivalent of 60 mmol carbon h^{-1} g DW^{-1}. In the case of growth on glucose this means the uptake was limited to 10 mmol glucose h^{-1} g DW^{-1} as reported before (Reed *et al.*, 2003). For all conditions with aerobic

growth, oxygen uptake was non-limiting, while anaerobic growth was simulated by constraining oxygen uptake to zero.

All computations were performed on an SBML level 1 file of the model retrieved in June 2007 from http://www-bioeng.ucsd.edu/research/research_groups/gcrg/organisms/ecoli/ecoli_sbml.html. Simulation and optimization of the *in silico* model under the selected growth conditions were performed using the COBRA toolbox (Becker *et al.*, 2007). In all cases the biomass flux was first maximized by linear programming using the COIN-OR Linear Program Solver (CLP solver) (http://www.coin-or.org/projects/Clp.xml and Lougee-Heimer, 2003) and afterwards an additional quadratic optimization, again with the CLP solver, was performed to handle the problem of multiple alternate optimal solutions. During the second optimization step the biomass flux was fixed to the value of the first optimization step and all other flux values were minimized according to the Euclidean distance (Mahadevan *et al.*, 2003).

Selection of the growth conditions

In the *E. coli* network iJR904 we count 128 unique carbon substrates that can be imported into the model. Most of them can be classified by the chemical categories of carbohydrates, amino acids, organic mono- and dicarboxylates as well as purines, pyrimidines and co-enzymes. For the model simulations substrates were selected from these different groups. We also considered which of the *in silico* substrates are known from literature to support growth of *E. coli* cells if provided as sole carbon source. Accordingly we selected sugars and sugar alcohols (glucose, ribose, lactose, sorbitol, glycerol), organic acids (gluconate, acetate, ketoglutarate, pyruvate, lactate, malate, succinate, octadecanoate) and amino acids (l-alanine, l-proline) as substrates. We did not include purines, pyrimidines or co-enzymes since we did not find support that *E. coli* can grow on them. Based on this selection of substrates we claim that we have simulated a set of conditions that are of biological relevance and representative for the broad spectrum of *E. coli* metabolic capabilities. Supporting literature is*:* acetate (Andersen *et al.*, 1980, Fong *et al.*, 2003, Hempfling *et al.*, 1975, Liu *et al.*, 2005), a-ketoglutarate (Fong *et al.*, 2003), alanine (Liu *et al.*, 2005), d-gluconate (Alam *et al.*, 1989, Lin, 1987), d-glucose (Alam *et al.*, 1989, Andersen *et al.*, 1980, Fong *et al.*, 2003, Lin, 1987, Liu *et al.*, 2005), d-lactate (Andersen *et al.*, 1980, Fong *et al.*, 2003), d-ribose (Fong *et al.*, 2003, Lin, 1987), d-sorbitol (Alam *et*

al., 1989, Lin, 1987), glycerol (Andersen *et al.*, 1980, Fong *et al.*, 2003, Hempfling *et al.*, 1975, Liu *et al.*, 2005), lactose (Burstein *et al.*, 1965, Lin, 1987), l-proline (Liu *et al.*, 2005), malate (Fong *et al.*, 2003, Lin, 1987), octadecanoate (Nunn, 1987, Overath *et al.*, 1969), pyruvate (Andersen *et al.*, 1980, Fong *et al.*, 2003), succinate (Andersen *et al.*, 1980, Fong *et al.*, 2003, Hempfling *et al.*, 1975, Lin 1987, Liu *et al.*, 2005).

Transformation of flux distribution and network structure into weighted metabolite networks

For each growth condition the optimization of biomass flux resulted in a vector of steady state fluxes (SI Table 2). Each flux distribution vector was used to construct a flux-weighted metabolite graph representing the same metabolic model. In these graphs two metabolites were connected if they participate in the same reaction and at least one carbon atom was transported between the metabolites by the reaction under consideration. Accordingly, for all reactions in the metabolic network metabolite pairs were assigned using the carbon transitions described in the RPAIR database in KEGG (Oh *et al.*, 2007). The mapping towards RPAIR was mostly one-to-one and in all other cases additional textbook information was used. Edge directionality is given by the sign of the flux for the corresponding reaction. Edge weights are computed by multiplying the flux value with the amount of carbon atoms transferred between two metabolites (SI Table 4). For example, for the transketolase reaction (Figure 4) the flux value is multiplied by 3 for the edge connecting D-xylulose 5-phosphate and glyceraldehyde 3-phosphate. If by existence of isoenzymes, two metabolites would be connected by parallel edges, one edge was created instead and assigned the sum of the individual weights (i.e., fluxes). Similarly, as a final step to the completion of the metabolite network, all pairs of anti-parallel edges were replaced with a single edge having weights and directionality according to the difference of the individual weights.

For the 19 simulated growth conditions the whole procedure always resulted in one large metabolite network component and a few disconnected metabolites and/or very small network fragments containing only two metabolites. These isolates and fragments were removed. As a result we obtained 19 connected metabolite networks with a size of 283-331 metabolites. It should be noted that the resulting networks are mass

balanced, i.e. carbon flows into a metabolite and out are balanced. All 19 networks are scale-free for both, the in- and the out-degree as the corresponding distributions follows a power-law (data not shown).

Flux centralities computed

Centralities are functions that assign to every vertex of a network a numerical value. By convention, the higher a centrality value, the more important (or central) is a vertex within the network under consideration. Based on centrality values vertices can be ranked, which highlights the most central metabolites. Different concepts of centralities are known and about 20 of them are discussed in a recent review (Koschützki *et al.*, 2005). A simple example is the out-degree centrality, which is calculated as the number of outgoing connections from a vertex. Some centralities can be generalized to account for the weights assigned to edges. For example, out-degree can be extended by summing all weights of the outgoing edges of the vertex under consideration.

In this study the general concept of centrality based on the qualitative property of connectivity was extended by considering the quantity of carbon-flux derived weights attributed to the edges. The centrality developed in this study is based on the concept of maximum flow. Considering any two vertices *s* and *t* in a network, the maximum flow between the two is defined by the largest flow that is observed for all possible paths connecting the two (Ahuja *et al.*, 1993). In simple terms, out of the many connecting paths we are looking for a set of paths forming the "strongest carbon flux" connecting *s* and *t*. We denote this maximum flow between *s* and *t* as *max_flow (s,t)*. In order to derive a centrality, we refer to the shortest-path closeness centrality (Koschützki *et al.*, 2005), which sums the lengths of the shortest-paths from a vertex *s* to all other vertices *t* ($t \in V$, *V* is the set of all vertices in the network). Accordingly, we define the maximum-flow closeness as:

$$mfc(s) = \sum_{t \in V} max\_flow(s,t)$$

This formula might be broken in the created PDF. It should read (in LaTeX-Notation):

mfc(s) = \sum_{t \in V} max_flox (s,t)

For each vertex the resulting value of the maximum-flow closeness centrality might be interpreted as a "metabolic potential" of the respective metabolite. The higher a *mfc* value of a metabolite, the more of it is

converted into other metabolites throughout the network. It should be noted that a more precise name of the defined centrality is out-maximum-flow closeness, because the flow leaving the vertex of interest is computed and a corresponding in-maximum-flow closeness, computing the flow entering the vertex, might be defined in a similar way.

The average path length (APL) was computed for the unweighted directed networks. Pairs of vertices without any connecting path were not considered in the calculation.

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF) under grant 0312706A (DK+FS) and by the Office of Basic Energy Sciences of the US Department of Energy (BHJ+JS). Additionally, we thank Eva Grafahrend-Belau for assisting us with the computation of the flux values.

References

Ahuja RK, Magnanti TM, Orlin JB (1993) *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall.

Alam KY, Clark DP (1989) Anaerobic Fermentation Balance of *Escherichia coli* as Observed by In Vivo Nuclear Magnetic Resonance Spectroscopy. *J Bacteriol.* 171:6213-6217.

Almaas E, Kovácz B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 427:839-843.

Almaas E, Oltvai ZN, Barabási AL (2005) The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* 1:e68.

Andersen KB, von Meyenburg K (1980) Are growth rates of *Escherichia coli* in batch cultures limited by respiration? *J Bacteriol*. 144: 114-123.

Arita M (2004) The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* 101:1543-1547.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509-512.

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols* 2:727-738

Burstein C, Cohn M, Kepes A, Monod J (1965) Role of lactose and its metabolic products in the induction of the lactose operon in *Escherichia coli. Biochem Biophys Acta*. 95:634-639.

Duarte NC, Herrgard MJ, Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 14:1298-1309.

Edwards JS, Palsson BØ (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* 97:5528-5533.

Ekman D, Light S, Björklund ÅK, Elofsson A (2006) What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol* 7:R45.

Fell DA, Wagner A (2000) The small world of metabolism. *Nat Biotechnol* 18:1121-1122.

Fong SS, Marciniak JY, Palsson BØ (2003) Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. *J Bacteriol*. 185:6400-6408.

Hempfling WP, Mainzer SE (1975) Effects of varying the carbon source limiting growth on yield and maintenance characteristics of *Escherichia coli* in continuous culture. *J Bacteriol*. 123:1076-1087.

Holme P, Huss M , Jeong H (2003) Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19:532-538.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* 407:651-654.

Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411:41-42.

Junker BH, Klukas C, Schreiber F (2006) VANTED: a system for advanced data analysis and visualization in the context of biological network. *BMC Bioinformatics* 7:e109.

Kim PJ, Lee DY, Kim TY, Lee KH, Jeong H *et al.* (2007) Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc Natl Acad Sci USA* 104:13638-13642.

Koschützki D, Lehmann KA, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O (2005) Centrality indices. in *Network Analysis: Methodological Foundations*, eds Brandes U, Erlebach T (Springer, Heidelberg), LNCS Tutorial 3418, pp 16-61.

Lin ECC (1987) Dissimilatory pathways for sugars, polyols and carbohydrates. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaschter M, Umbarger HE (eds) *Escherichia coli* and Salmonella: cellular and molecular biology, ASM Press, Washington DC, pp 244-284.

Liu M, Durfee T, Cabrera JE, Zhao K, Jin DJ, Blattner FR (2005) Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J Biol Chem.* 280:15921-15927.

Lougee-Heimer R (2003) The Common Optimization INterface for Operations Research. *IBM Journal of Research and Development*. 47(1):57-66.

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

Ma HW, Zeng AP (2003b) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 19:1423-1430.

Ma HW, Zhao XM, Yuan YJ, Zeng AP (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* 20:1870-1876.

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264–276.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824-827.

Nunn WD (1987) Two-carbon compounds and fatty acids as carbon sources. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaschter M, Umbarger HE (eds) *Escherichia coli* and Salmonella: cellular and molecular biology, ASM Press, Washington DC, pp 285-301.

Oh M, Yamada T, Hattori M, Goto S, Kanehisa M (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J Chem Inf Model* 47: 1702-1712.

Overath P, Pauli G, Schairer HU (1969) Fatty acid degradation in *Escherichia coli*. An inducible acyl-CoA synthetase, the mapping of old-mutations and the isolation of regulatory mutants. *Eur. J. Biochem.* 7:559-574.

Rahman SA, Schomburg D (2006) Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics* 22:1767-1774.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL(2002) Hierarchical organization of modularity in metabolic networks. *Science* 297:1551-1555.

Reed JL, Vo RD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 4:R54.

Szyperski T (1995) Biosynthetically directed fractional ^{13}C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism. *Eur J Biochem* 232: 433-448.

Wagner A, Fell DA (2001) The small world inside large metabolic networks. *Proc Roy Soc London B Biol* 268:1803-1810.

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440-442.

Weikert C, Sauer U, Bailey JE (1997) Use of a glycerol-limited, long-term chemostat for isolation of *Escherichia coli* mutants with improved physiological properties. *Microbiology* 143:1567–1574.

Figures

Figure 1. **Central metabolism of *E. coli.*** The colors differentiate the three major pathways of central metabolism: red, glycolysis; green, TCA cycle; blue, pentose phosphate pathway (PPP). Squares enclose the 15 metabolites used as sole carbon source for the FBA simulations. Additional paths between some pairs of metabolites are not shown here for simplicity. The figure was created with the software VANTED (Junker *et al.*, 2006). Abbreviations not given in Table 1: 6PGL, 6-phospho-D-glucono 1,5-lactone; AC, acetate; ALA, L-alanine; GLC, D-glucose; GLCN, D-gluconate, GLU, L-glutamate; GLYC, glycerol; LAC, D-lactate; LCTS, lactose; OCDCA, octadecanoate, PRO, L-proline; RIB, D-ribose; SBT, sorbitol.

Figure 2. **Heatmap of flux centrality values derived from simulations of 19 different growth conditions with *E. coli* iJR904 (Reed *et al.*, 2003).** High centrality values are marked as red spots, intermediate values as yellow spots and low to zero values are marked white. The centrality values were clustered using hierarchical clustering with Euclidean distance and the complete linkage method. Clustering was computed with the software system R (http://www.R-project.org). Of the 761 metabolites in the flux balance model the 395 metabolites with non-zero centrality value are mapped here with their centrality values. Figure 3a and 3b show details of the dendrogram. Groups 1-6 and A-C are discussed in the text. A large-scale figure including metabolite names is given as SI Figure 5. Growth conditions shown in the bottom are marked + for aerobic and – for anaerobic growth. Abbreviations for growth conditions, see Table 1, AC – Acetate, ALA – L-Alanine, GLC – D-Glucose, GLCN – D-Gluconate, GLYC – Glycerol, LAC – D-Lactate, LCTS – Lactose, OCDCA – Octadecanoate, PRO – L-Proline, RIB – D-Ribose, SBT – sorbitol.

Figure 3. **Magnifications of two metabolite clusters from Figure 2.** (a) is the cluster with the highest flux centralities, comprising of intermediates from glycolysis, the pentose phosphate pathway and the tricarboxic acid cycle. Metabolites were colored according to Figure 1. (b) representative cluster of a pathway only active under one growth condition. Abbreviations for metabolites, see Table 1, 6PGL – 6-phospho-D-glucono 1,5-lactone, GAL – D-Galactose, GAL1P – Galactitol 1-phosphate, GLC – D-Glucose, LCTS –

Lactose (cytosolic, extra cellular and boundary), UDPG –UDPglucose, UDPGAL –UDPgalactose. The addition of the letter ‘c’ after a metabolite means ‘cytosolic’, ‘e’ means ‘external’, and ‘b’ means ‘boundary’ (Reed *et al.*, 2003).

Figure 4. **Visualization of the construction process of the networks**. (a) representation of the reaction transketolase as a hypergraph; (b) representation of the same reaction as a substrate graph (metabolite graph) where all substrates are connected to all products; (c) representation used in this study. Metabolites are connected only if a transfer of carbon atoms occurs and edge weights are dependent on the number of carbon atoms. Abbreviations for metabolites, see Table 1, TKT1 – Transketolase.

Tables

Table 1. Top-30 metabolites based on the summed flux centrality value under 19 different growth conditions. Colors: red (glycolysis), green (TCA cycle), blue (pentose phosphate metabolism).

Supporting Information

Four supplemental tables, one supplemental figure and one supplemental document:

SI Table 2: FBA-based flux values for the 19 growth conditions.

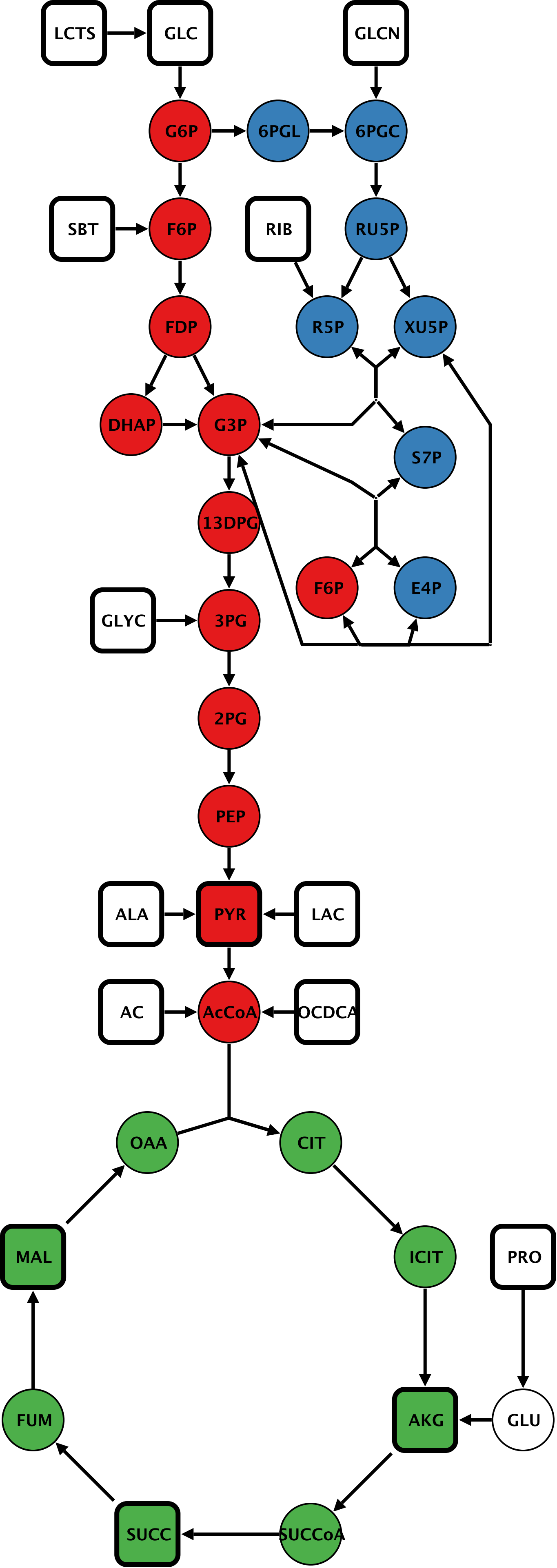
SI Table 3: Centrality values according to the flux centrality for the 19 growth conditions.

SI Table 4: Carbon transfer information for each reaction.

SI Table 5: Top 10 list of metabolites according to different centrality measures extracted from several publications (see discussion).

SI Figure 5: Figure 2 with labels for each metabolite on the left side.

SI Document 1: A description of the method to compute the flux centrality values.


LCTS
GLC
GLCN
G6P
6PGL
6PGC
SBT
F6P
RIB
RU5P
FDP
R5P
XU5P
DHAP
G3P
S7P
13DPG
F6P
E4P
GLYC
3PG
2PG
PEP
ALA
PYR
LAC
AC
AcCoA
OCDCA
OAA
CIT
MAL
ICIT
PRO
FUM
AKG
GLU
SUCC
SUCCoA

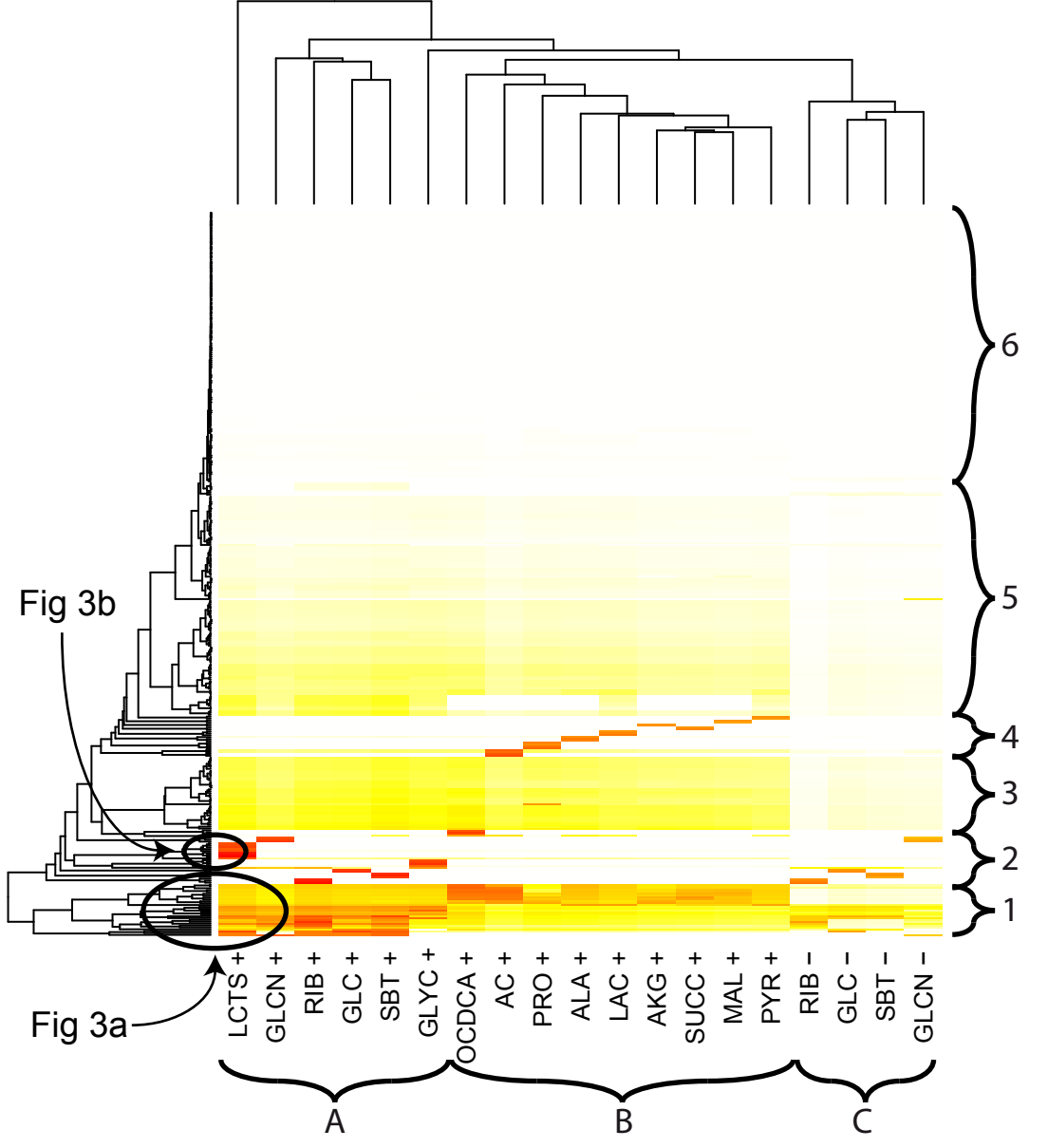
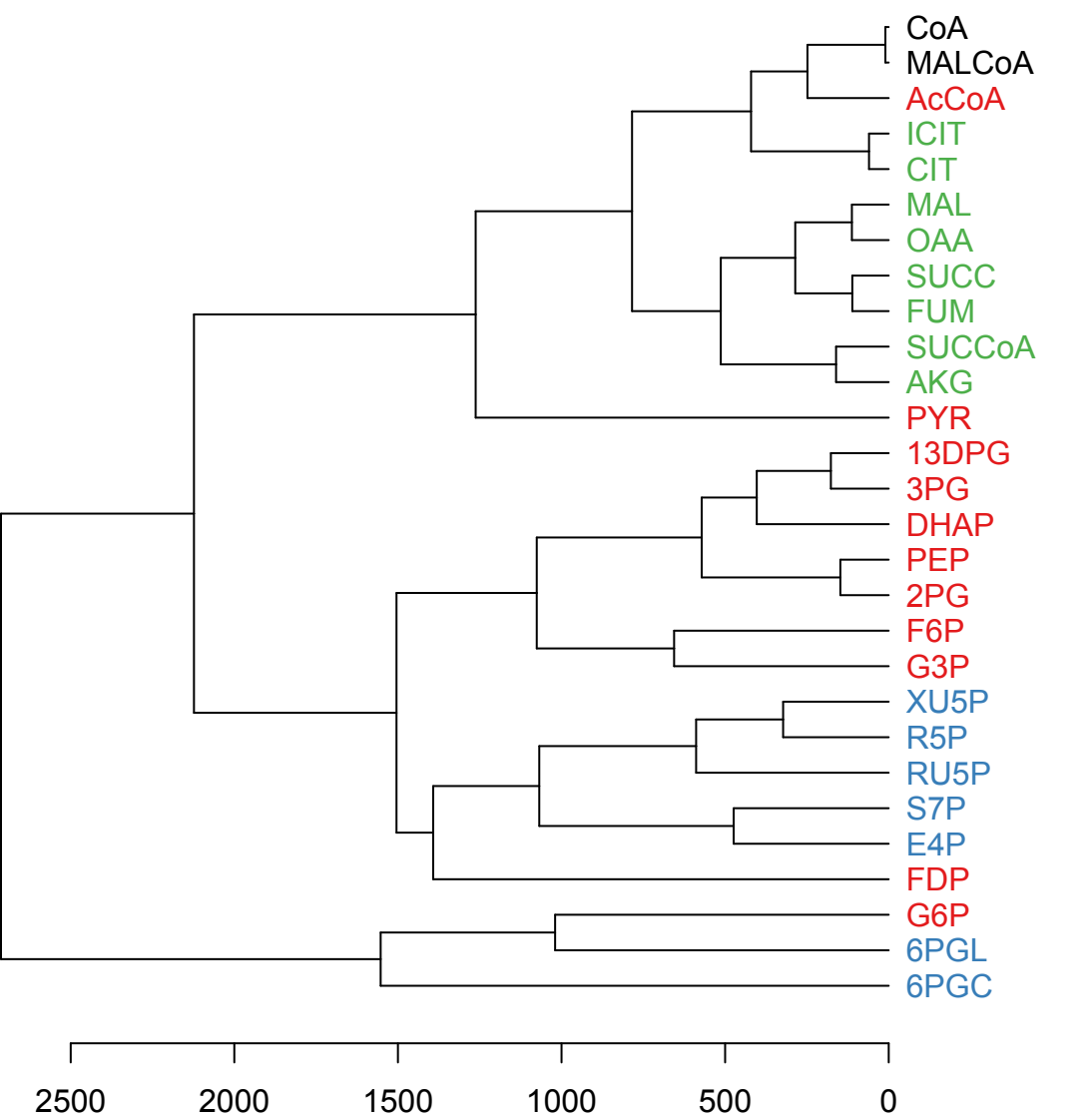
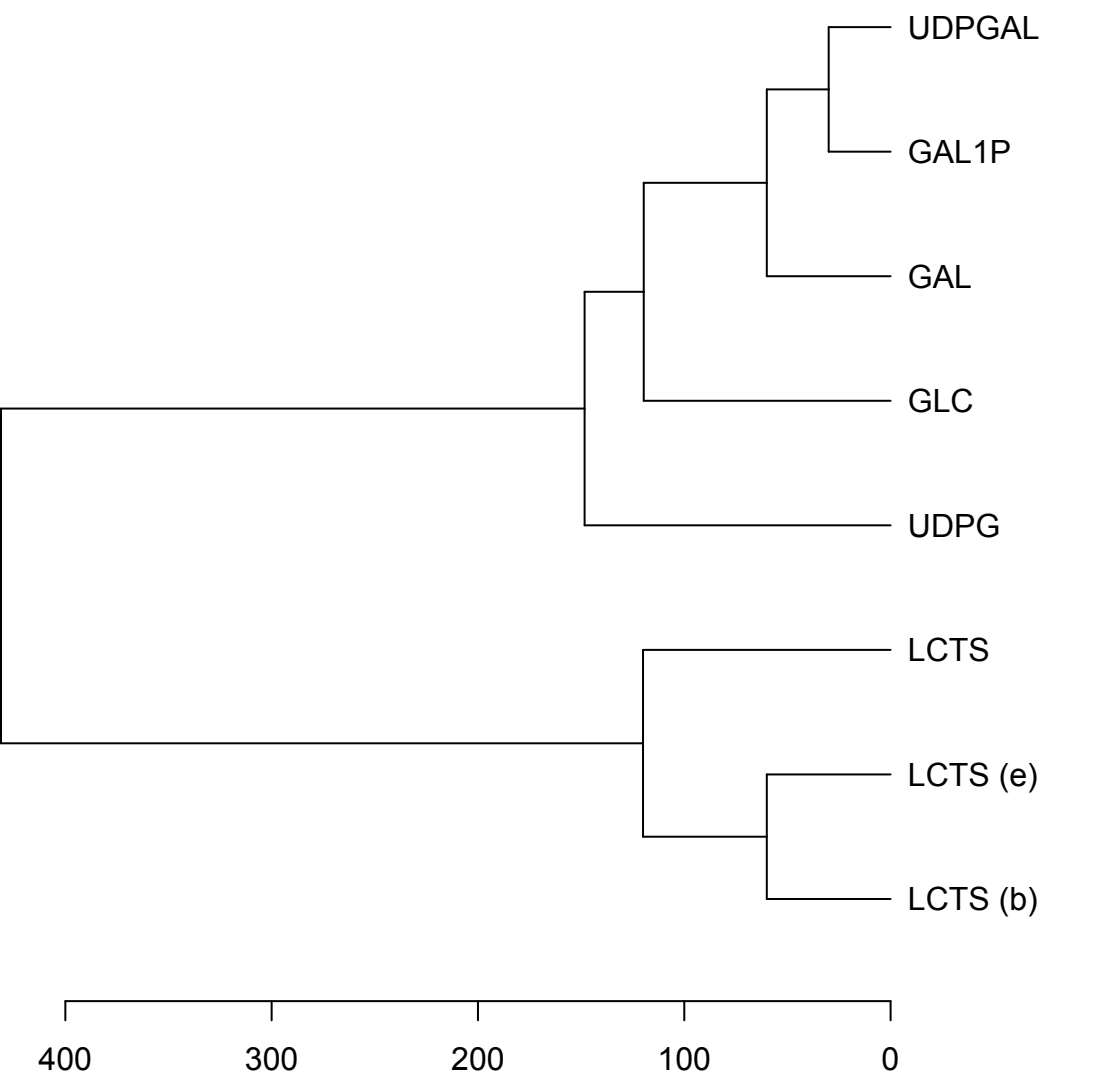
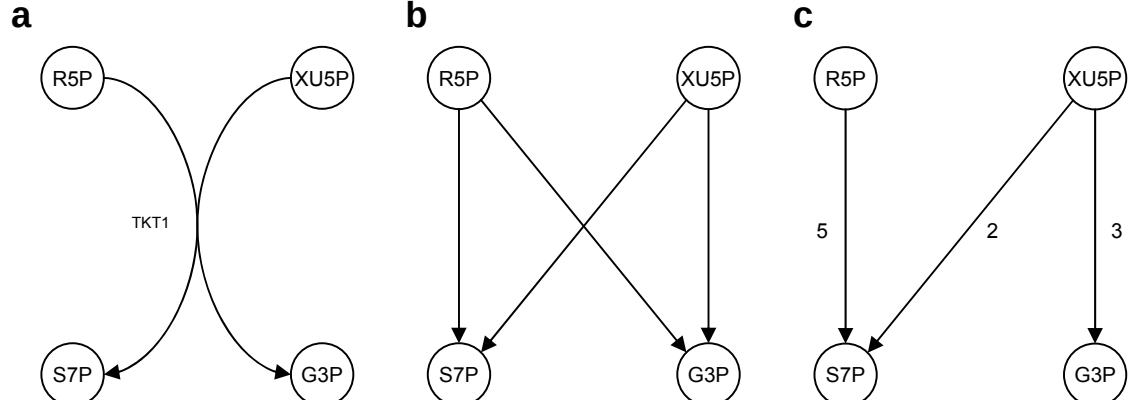


Fig 3b
Fig 3a
LCTS +
GLCN +
RIB +
GLC +
SBT +
GLYC +
OCDCA +
AC +
PRO +
ALA +
LAC +
AKG +
SUCC +
MAL +
PYR +
RIB –
GLC –
SBT –
GLCN –
6
5
4
3
2
1
A
B
C


CoA
MALCoA
AcCoA
ICIT
CIT
MAL
OAA
SUCC
FUM
SUCCoA
AKG
PYR
13DPG
3PG
DHAP
PEP
2PG
F6P
G3P
XU5P
R5P
RU5P
S7P
E4P
FDP
G6P
6PGL
6PGC
2500
2000
1500
1000
500
0


UDPGAL
GAL1P
GAL
GLC
UDPG
LCTS
LCTS (e)
LCTS (b)
400
300
200
100
0


a
R5P
XU5P
TKT1
S7P
G3P
b
R5P
XU5P
S7P
G3P
c
R5P
XU5P
5
2
3
S7P
G3P

Position	Metabolite ID	Metabolite Name	Summed Centrality Value	Pathway
1	G3P	Glyceraldehyde 3-phosphate	11584,21	Glycolysis
2	F6P	D-Fructose 6-phosphate	11564,81	Glycolysis
3	13DPG	3-Phospho D-glyceroyl-phosphate	10705,56	Glycolysis
4	AcCoA	Acetyl CoA	10619,46	Glycolysis
5	3PG	3-Phospho D-glycerate	10486,82	Glycolysis
6	PYR	Pyruvate	10204,73	Glycolysis
7	MAL	L-Malate	10157,91	TCA
8	FUM	Fumarate	10109,16	TCA
9	MALCoA	Malonyl CoA	10061,77	
10	CoA	Coenzyme A	10030,76	
11	OAA	Oxaloacetate	10009,09	TCA
12	SUCC	Succinate	9942,48	TCA
13	RU5P	D-Ribulose 5-phosphate	9878,29	PPP
14	CIT	Citrate	9873,87	TCA
15	2PG	D-Glycerate 2-phosphate	9872,83	Glycolysis
16	ICIT	Isocitrate	9694,15	TCA
17	DHAP	Dihydroxyacetone-phosphate	9652,38	Glycolysis
18	AKG	2-Oxoglutarate	9641,94	TCA
19	PEP	Phosphoenolpyruvate	9547,94	Glycolysis
20	SUCCoA	Succinyl CoA	9170,58	TCA
21	R5P	alpha-D-Ribose 5-phosphate	9064,74	PPP
22	XU5P	D-Xylulose 5-phosphate	8762,09	PPP
23	FDP	D-Fructose 1,6-bisphosphate	8130,85	Glycolysis
24	G6P	D-Glucose 6-phosphate	7173,67	Glycolysis
25	S7P	Sedoheptulose-7-phosphate	6975,80	PPP
26	E4P	D-Erythrose 4-phosphate	6389,81	PPP
27	ASP	L-Aspartate	6340,43	
28	CO2	CO2	6002,98	
29	6PGC	6-Phospho D-gluconate	5847,29	PPP
30	PRPP	5-Phospho D-ribose 1-diphosphate	5567,11	