BIOINFORMATICS

Vol. 19 no. 11 2003, pages 1423–1430
DOI: 10.1093/bioinformatics/btg177

The connectivity structure, giant strong component and centrality of metabolic networks

*Hong-Wu Ma and An-Ping Zeng**

Department of Genome Analysis,GBF - German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany

Received on October 24, 2002; revised on February 7, 2003; accepted on February 18, 2003

ABSTRACT

Motivation: Structural and functional analysis of genome-based large-scale metabolic networks is important for understanding the design principles and regulation of the metabolism at a system level. The metabolic network is conventionally considered to be highly integrated and very complex. A rational reduction of the metabolic network to its core structure and a deeper understanding of its functional modules are important.

Results: In this work, we show that the metabolites in a metabolic network are far from fully connected. A connectivity structure consisting of four major subsets of metabolites and reactions, i.e. a fully connected sub-network, a substrate subset, a product subset and an isolated subset is found to exist in metabolic networks of 65 fully sequenced organisms. The largest fully connected part of a metabolic network, called 'the giant strong component (GSC)', represents the most complicated part and the core of the network and has the feature of scale-free networks. The average path length of the whole network is primarily determined by that of the GSC. For most of the organisms, GSC normally contains less than one-third of the nodes of the network. This connectivity structure is very similar to the 'bow-tie' structure of World Wide Web. Our results indicate that the bow-tie structure may be common for large-scale directed networks. More importantly, the uncovered structure feature makes a structural and functional analysis of large-scale metabolic network more amenable. As shown in this work, comparing the closeness centrality of the nodes in the GSC can identify the most central metabolites of a metabolic network. To quantitatively characterize the overall connection structure of the GSC we introduced the term 'overall closeness centralization index (OCCI)'. OCCI correlates well with the average path length of the GSC and is a useful parameter for a system-level comparison of metabolic networks of different organisms.

Contact: aze@gbf.de

Supplementary Information: http://genome.gbf.de/bioinformatics/

*To whom correspondence should be addressed.

INTRODUCTION

Up to now, about 100 organisms have been fully sequenced. The use of the large amounts of sequence data for the understanding of structure and functionality of complex cellular networks such as gene regulation, protein interaction and metabolic networks is one of the most important issues in the post-genome research (Kitano, 2002; Noble, 2002). Among others, metabolic networks have gained much attention because they represent a key component on the way from genome sequences to the cellular metabolism and phenotypes and have important implications for many biological research areas such as metabolic engineering and biomedicine (Ideker *et al.*, 2001; Saqi and Sternberg, 2001; Burgard and Maranas, 2001; Price *et al.*, 2002). Furthermore, metabolic networks share some common features of many complex biological and non-biological systems, a deeper understanding of which may thus reveal unifying principles of the structure and function of natural and social networks (Strogatz, 2001; Redner, 2002).

Organism-specific metabolic networks can be reconstructed from genome information using databases such as KEGG, WIT and Ecocyc (Ogata *et al.*, 1999; Overbeek *et al.*, 2000; Wittig and De Beuckelaer, 2001; Karp *et al.*, 2002). Genome-based metabolic networks are normally very large and complex. Efforts are being made to analyze and understand the structure of these large-scale networks (Schilling and Palsson, 2000; Schuster *et al.*, 2002). In particular, methods from graph theories are shown to be useful for obtaining global structure properties of a metabolic network (Jeong *et al.*, 2000; Bilke and Peterson, 2001; Ma and Zeng, 2003; Wolf *et al.*, 2002). In this approach, the metabolites of a metabolic network are represented as nodes in the graph, and the reactions are expressed as connections between the nodes. Jeong *et al.* (2000) studied the large-scale organization of metabolic networks of 43 different organisms by graph theories. These authors found that metabolic networks have the feature of a scale-free network and almost

the same average path length (AL). The scale-free network is one class of the so-called small-world networks that are characterized by a short average path length, a high cluster coefficient and a power law (or similar) connection degree distribution (Amaral *et al.*, 2000; Strogatz, 2001). Fell and Wagner (2000) also found that the metabolic network of *Escherichia coli* is a small-world network by using different graph representation methods. The small-world property of metabolic networks is shown to be similar to other robust and error-tolerant networks such as computer, neural and certain social networks. More recently, Ravasz *et al.* (2002) showed that metabolic networks of organisms are organized as many small, but highly connected modules that combine in a hierarchical manner to larger, less cohesive units.

In the studies of Jeong *et al.* (2000) and Ravasz *et al.* (2002) the same bioreaction database of 43 organisms was used. Recently, we extensively extended and revised the KEGG LIGAND reaction database (Goto *et al.*, 1998) by considering reaction reversibility and correcting many obvious errors (Ma and Zeng, 2003). The metabolic networks of 80 fully sequenced organisms were *in silico* reconstructed from the genome data and the revised reaction database and represented as directed graphs. In this graphic representation of metabolic network the substrate(s) of a reaction are connected with the product(s) by directed links if the reaction is irreversible, or by undirected links if it is reversible. The connections through currency metabolites such as ATP and NADH (see Table 2 in Supplementary materials for a list of currency metabolites) are removed in order to have a physiologically meaningful definition of the path length. With this improved definition of path length, different values of average path length of metabolic networks were obtained for the three different domains of organisms, revealing quantitative differences in the global structure of metabolic networks of different organisms. This is in contrast to the results of Jeong *et al.* (2000). Furthermore, the studies of Ravasz *et al.* (2002) and Ma and Zeng (2003) revealed that the connectivity degree distribution is not enough to describe the structural difference in metabolic networks. Other quantitative parameters are needed to better represent the network structure feature.

In this work, we present a comprehensive analysis of the connectivity structure of genome-based metabolic networks of 65 fully sequenced organisms. A bow-tie type structure with a giant strong component (GSC) is identified. A new parameter is introduced to describe the compactness of connectivity and the centrality distribution of metabolites in the GSC.

CONNECTIVITY STRUCTURE OF METABOLIC NETWORKS

For metabolic networks reconstructed from genome information as described in detail in Ma and Zeng (2003) we used the breadth first searching method (Broder *et al.*, 2000) to find all the connected metabolites for any metabolite and thus perform a connectivity analysis on the whole network. We found that in most of the metabolic networks about half of the metabolites can be only converted to a very limited number (usually less than 10) of metabolites. Although the number of metabolites reachable by the other metabolites is much higher, it is still not more than half of the metabolites. For a randomly chosen substrate and product pair, the probability that a path exists between them is less than 20% (10% for the metabolic network of certain organisms). This indicates that a metabolic network is far from a fully connected network. At the same time, we also found that there exist several fully connected metabolic sub-networks in which the metabolites can be converted to each other. These fully connected sub-networks may be called strong components of the metabolic network. In graph theory, a strong component of a network is defined as a subset of nodes such that for any pair of nodes u and v in the subset there is a path from u to v (Batagelj and Mrvar, 1998). In the following, the metabolic network of *E.coli* is used as an example for illustration.

All the strong components in the metabolic network of *E.coli* were identified by using the network analysis software Pajek (Batagelj and Mrvar, 1998). There are 29 strong components altogether that include no less than three metabolites. The largest strong component, also called the GSC, is a 274-node sub-network, whereas other strong components are much smaller and have a node number less than 15 (most of them only contain three or four nodes). The connection structure of the GSC is complex as shown in Figure 1. In general, there are several routes between any pair of nodes. This pathway redundancy obviously makes the GSC robust to removal of reactions by gene knock-out or mutation.

The output domains (defined as a subset of nodes which are reachable from a specific node) of the nodes in the GSC were calculated and compared. We found that these output domains are a same subset with 435 nodes that includes all the nodes in the GSC. The 161 metabolites that are in the output domain but not in the GSC ($435 - 274 = 161$) can be produced from the metabolites in the GSC, but cannot be converted to them. Thus, these metabolites form a product subset (P). In a similar way, the input domains (a subset of nodes from which a specific node can be reached) of the GSC were calculated, leading to the identification of a substrate subset (S) with 93 nodes. The metabolites in the substrate subset S can convert to

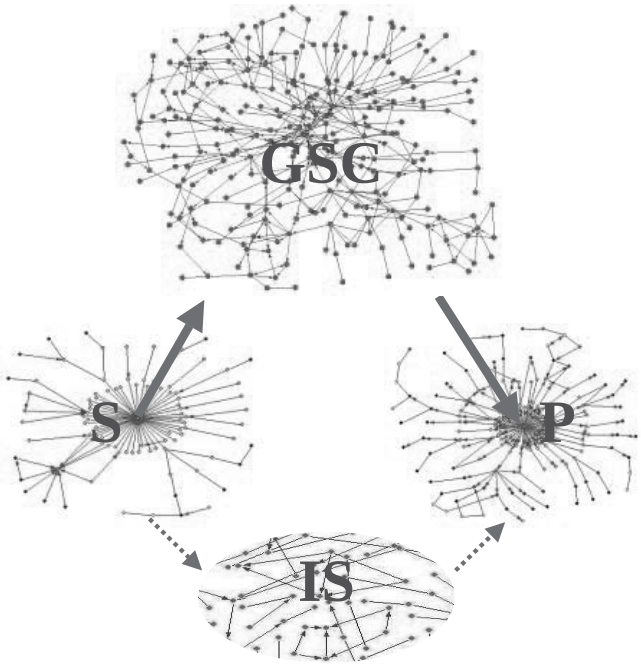

Fig. 1. Connectivity structure of metabolic networks: the network of *E.coli* as an example. **GSC**, giant strong component; **S**, substrate subset; **P**, product subset; **IS**, isolated subset. The central node in the substrate and product subsets represents the giant strong component. The arrow indicates the irreversibility of a reaction. The software package Pajek (Batagelj and Mrvar, 1998) was used to draw the graph.

any metabolites in the GSC and the product subset P, but cannot be produced from them. The connection structures of the subsets GSC, S and P are shown in Figure 1. Compared with the GSC structure, the structures of S and P are relatively simple; most of the pathways are only short linear pathways. All the other 283 metabolites that are not in the GSC, S and P form an isolated subset (IS). Certain metabolites that can be produced from metabolites in S or be converted to metabolites in P are also included in the subset IS. The metabolites in IS cannot be converted to or from the metabolites in the GSC.

The metabolic networks of other organisms show similar connection structure as that of *E.coli*. The number of metabolites in the GSC, substrate and product subsets for all the 65 organisms is given in Table 1 in the Supplementary materials. The number of metabolites in GSC is less than 300 for all organisms. For some small-scale networks, the GSC scale is only about 50. The scale of the substrate subset and product subset is smaller than the GSC for most organisms.

The macroscopic structure of a metabolic network as shown in Figure 1 is very similar to the so-called 'bow-tie' structure of the world wide web described by Broder *et al.* (2000). They found that the fully connected part (SCC) of World Wide Web is no more than 30% of the whole network. There are an IN subset that consists of web pages that can reach the SCC and an OUT subset which consists of pages that are accessible from the SCC. Our results indicate that the bow-tie connectivity structure may be common for large-scale directed networks. The scale-free property revealed by the power law connection degree distribution only reflects partial structure properties of the metabolic network. An important basic structure feature is uncovered in this work through a more detailed analysis. It is worth mentioning that the removal of connections through currency metabolites in the metabolic network is important for revealing this structure feature because these currency metabolites participate in so many reactions that most metabolites can be connected through them. But these connections do not represent real metabolic pathways. In fact, the use of this biologically more meaningful definition of the bioreaction path is also important for revealing the quantitative differences in metabolic networks of different organisms as shown by Ma and Zeng (2003).

STRUCTURE AND FUNCTIONAL ANALYSIS OF GSC

The connectivity structure as shown in Figure 1 indicates that the GSC is the most complex and core part of a metabolic network. It therefore deserves more detailed analysis. The cumulative input and output connection degree distributions of the GSC in *E.coli* and *Homo sapiens* are shown in Figures 2a and b respectively, wherein $P(k)$ is the fraction of nodes that have a degree of output (or input) larger than k. $P(k)$ was calculated by dividing the number of metabolites that had output (or input) connections larger than k by the total number of metabolites in the organism. A truncated power law distribution (linear relation in the logarithmic scale coordinates with the first point being scattered) can be ascertained. GSCs of other organisms studied also have similar power law degree distribution. This suggests that GSCs of all organisms are scale-free networks.

The average path length of GSC (ALG) and that of the whole network (ALW) were also calculated. The relationship between them is shown in Figure 3. ALG is somehow smaller than ALW. This is because the GSC is more tightly connected than the whole network. There is a nearly linear relation between ALG and ALW for most organisms, especially for those with relatively large strong components (node number greater than 40). These results indicate that the average path length of the whole network is primarily determined by that of the GSC.

Because of the large scale, it is often difficult to straightly obtain a comprehensive understanding of the

Table 1. The most central metabolites in the metabolic network of *E.coli*

Output center Metabolite	Mean distance	Input center Metabolite	Mean distance	Overall Center Metabolite	Mean distance
Pyruvate	4.2198	Pyruvate	4.663	Pyruvate	4.4414
2KD6PG	4.6007	Acetyl-CoA	4.9011	Acetyl-CoA	4.7582
Acetyl-CoA	4.6154	Malate	4.9011	Malate	4.8864
G3P	4.696	Acetate	4.9194	2KD6PG	4.9286
Serine	4.7473	Formate	4.9853	Acetate	4.978
Acetaldehyde	4.7729	Fumarate	5.1978	Acetaldehyde	5.0311
DR5P	4.8608	2KD6PG	5.2564	G3P	5.0641
Cystine	4.8645	Citrate	5.2821	PEP	5.2106
Malate	4.8718	Acetaldehyde	5.2894	HOAKG	5.2491
PEP	4.8938	Methylglyoxal	5.3516	Methylglyoxal	5.2766

2KD6PG, 2-Dehydro-3-deoxy-6-phospho-D-gluconate; DR5P, 2-Deoxy-D-ribose 5-phosphate; G3P, Glyceraldehyde 3-phosphate; HOAKG, D-4-Hydroxy-2-oxoglutarate; PEP, Phosphoenolpyruvate.

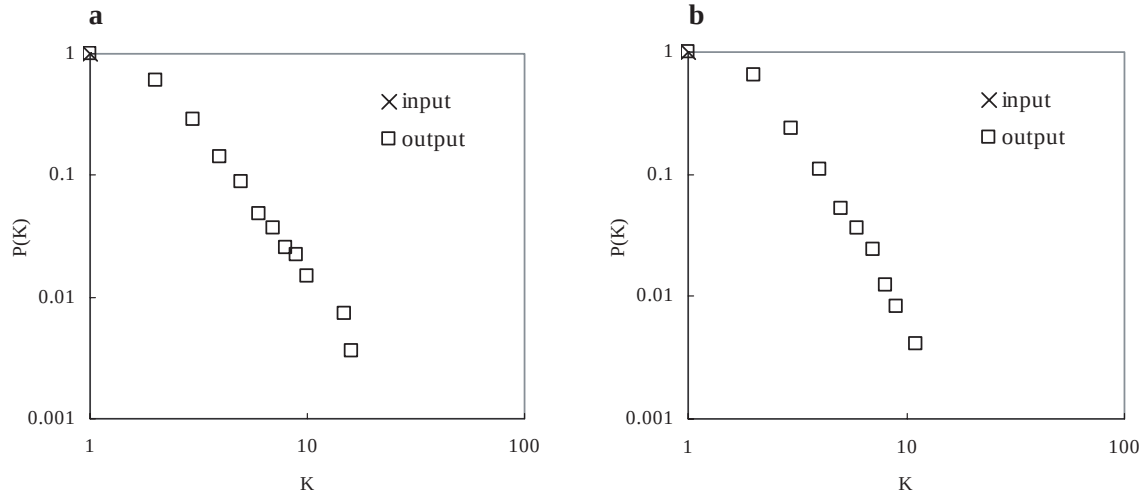

Fig. 2. Cumulative connection degree distribution of the GSCs in the metabolic networks of *E.coli* (**a**) and *Homo sapiens* (**b**).

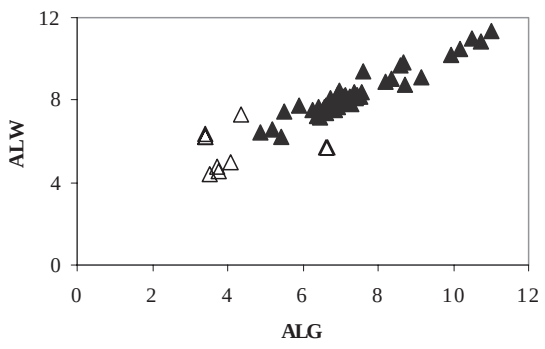

Fig. 3. Relationship between the average path length of the GSCs (ALG) and that of the whole network (ALW) for 65 organisms. The open triangles represent organisms that have a small giant strong component (node number less than 40).

biological features of genome-based metabolic network. A certain form of reduction or classification of the whole network is desired to make the network amenable to functional analysis. The connectivity structure of the metabolic network revealed in this work represents a step forward in this direction. The most important part of the network, the GSC, normally contains less than one-third of the nodes of the whole network, making it more amenable to functional analysis. Here, we use *Streptococcus pneumoniae*, an important Gram-positive pathogen, as an example to analyze the functional feature of its network structure. The whole network of *S.pneumoniae* consists of 486 metabolites, while its GSC contains only 87 metabolites. *S.pneumoniae* has a moderate metabolic network scale and GSC. To further reduce the complexity of the GSC, we suggest the following way to reduce the

node number in the GSC. From Figure 1 it can be seen that there are many linear branches (the endpoint of which has only one connection and no branch point in the path) in the GSC. They are in the GSC only because these nodes are connected through a series of reversible reactions. These linear branches may be omitted for functional analysis. By removing the linear branches, the GSC of *S.pneumoniae* is reduced to its core network as shown in Figure 4a. Five major metabolic pathways (not necessarily complete due to omission of linear branches) can be identified in the core network of *S.pneumoniae*. These include the glycolysis pathway, the pentosephosphate pathway, the aromatic amino acid synthesis pathway, the glycerol metabolism and the pyrimidine synthesis pathway. There are also parts of lysine synthesis, valine synthesis, oxaloacetate anapleurotic and Entner–Doudoroff (ED) pathways in the core. These pathways are integrated into a network through certain metabolites such as pyruvate (PYR), 5-Phosphoribosyl diphosphate (PRPP) and glyceraldehydes phosphate (G3P). All these metabolites belong to the so-called hub metabolites identified in our previous work (Ma and Zeng, 2003). As links between the different functional systems, these metabolites may play a key role in metabolic regulation. In consistency with the discovery from genome analysis, there is no TCA cycle reactions in the core network (Tettelin *et al.*, 2001).

Figure 4b shows the core metabolic network of *Pyrococcus furiosus*, an archaeum that has approximately the same node number in its GSC as *S.pneumoniae*. The core network of *P.furiosus* contains 49 metabolites, 37 of which are also in the core of *S.pneumoniae*. These two core networks have similar metabolite composition and functional units, except that *P.furiosus* contains the major part of the TCA cycle in its core network structure. By a direct inspection of Figures 4a and b it is obvious that the core networks of *S.pneumoniae* and *P.furiosus* have different connection structures. The core network of *S.pneumoniae* is somewhat more densely connected than that of *P.furiosus*. The compactness of GSCs of different organisms is discussed in the next section.

Figure 4 shows that the GSC may include the most important pathways and metabolites for the metabolism of organisms and reflect their evolutionary history and nutritional requirement for growth. The biological meaning of these pathways and metabolites deserves more detailed analysis, in particular through comparison of different organisms and in combination with experimental functional studies.

The uncovering of the fundamental structure of the metabolic network has important implications for biotechnology and biomedicine. For example, understanding and manipulating the distribution and control of metabolic fluxes over the metabolic network is key for metabolic engineering of organisms and the therapy of certain

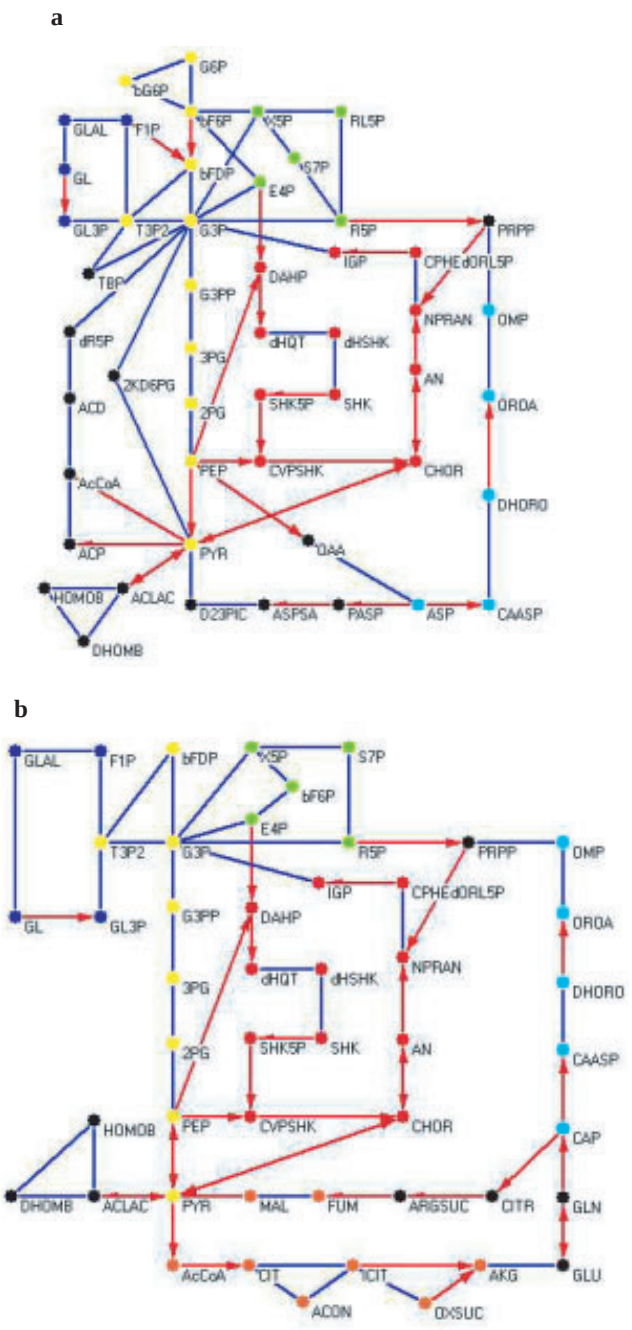

Fig. 4. Core of the metabolic networks of *Streptococcus pneumoniae* (**a**) and *Pyrococcus furiosus* (**b**). The node colors show metabolites of different functional pathways. Yellow: glycolysis pathway; green: pentosephosphate pathway; orange: TCA cycle; red: aromatic amino acid synthesis pathway; blue: glycerol metabolism; cyan: pyrimidine synthesis pathway; black: other functional systems. Blue lines without an arrow represent reversible reactions, red lines irreversible reactions. Arrows on both sides mean that there are two irreversible reactions in the opposite directions. The software Pajek (Batagelj and Mrvar, 1998) was used to draw the graphs. See Table 3 in Supplementary materials for abbreviations.

metabolic diseases (Bailey, 2001). However, for a large-scale metabolic network the estimation of metabolic flux and control can be very difficult or even impossible. A reduction of the metabolic network is almost always necessary. The GSCs and particularly the core networks of organisms contain a much smaller number of but key metabolites. They are more feasible for analysis of flux distribution and identification of all the possible elementary flux modes or extreme pathways (Schuster *et al.*, 2002; Schilling and Palsson, 2000). For other parts of the network, most of them are linear pathways and easy to analyze. The distribution of metabolic fluxes is mainly controlled by regulating the flux ratio at the branch-points. Most of the branch-points are in the core. Therefore one may focus on the core network when studying the flux distribution and its regulation in the metabolic network. This can largely simplify the analysis process.

CENTRALITY ANALYSIS OF THE GIANT STRONG COMPONENT

Besides the connection degree distribution and average path length, network centrality is another important structure parameter of large-scale networks by which the central metabolites of metabolic network can be identified (Batagelj and Mrvar, 1998). There are different definitions of centrality. The degree of centrality is defined by the connection degree of unit and is related with the above degree distribution. Sabidussi (1966) defined the term 'closeness centrality' of node $x(C(x))$ as follows:

$$C(x) = \frac{n-1}{\sum_{y \in U, y \neq x} d(x,y)} = \frac{1}{\bar{d}} \qquad (1)$$

where $d(x, y)$ is the distance between node x and node y; U is the set of all nodes; $\bar{d}$ is the average distance between x and the other nodes. For directed networks, the centrality is called output closeness centrality when $d(x, y)$ is defined as the path length from x to y. It is called input closeness centrality if $d(x, y)$ is defined as the path length from y to x. Here we define the overall closeness centrality as the reciprocal of the average of the mean input distance and the mean output distance. The most central metabolite is the metabolite that has the largest centrality value. The 10 most central metabolites in the *E.coli* metabolic network and the average path length from and to them are given in Table 1. Pyruvate is both the input center and the output center of the network. This result differs from that of Fell and Wagner (2000). They showed that the center metabolite is glutamate, with a mean path length of 2.46, followed by pyruvate (2.59). There are two reasons for the difference: (1) Fell and Wagner did not consider the reaction direction, thus their network is an undirected network; (2) glutamate participates in many reactions as an amino acid group carrier. In this instance, it was regarded as a currency metabolite in our study. Furthermore, only ATP, ADP, NADH and NAD were regarded as currency metabolites in the study of Fell and Wagner (2000).

Eight of these central metabolites (pyruvate, acetyl-CoA, phosphonolpyruvate, glyceraldehyde 3-phosphate, 2-dehydro-3-deoxy-6-phospho-D-gluconate, malate, fumarate and citrate) are in the central metabolism, namely the glycolysis and citrate acid cycle pathway. All the other central metabolites are directly connected with one or more of these eight metabolites. For example, serine and cysteine can be directly converted to pyruvate by irreversible reactions, they are thus the output center, but not the input center. These central metabolites may be in the central place of metabolic regulation because through these metabolites the environmental perturbation can be propagated to the whole network in a short time.

The above definition of centrality is referred to each single node in the network and can be called unit centrality. The unit centrality distribution in a network can be described by the network closeness centralization index C which is defined according to (Freeman, 1979):

$$C = \frac{(2n-3)\sum_{x \in U}(C^* - C(x))}{(n-1)(n-2)} \qquad (2)$$

where n is the node number in the network, C^* is the highest value of closeness centrality, $C(x)$ is the closeness centrality for the node x. Here we use the overall closeness centrality to calculate the overall closeness centralization index (OCCI) with Equation (2). It should be noted that OCCI is not equal to the average of the input centralization index and the output centralization index. OCCI can be used as a parameter to show the structural difference of networks. For example, for the two kinds of networks shown in Figure 5, the centralities of all the nodes in the circle-type network are the same, thus the OCCI is 0, whereas in the star-type network the centrality of the central node is much higher than other nodes, and the OCCI is 1. The real network should have a value of OCCI between 0 and 1. It can also be seen that the star network is scale free, the average path length is close to 2 irrespective of the node number (when the node number is large enough). The average path length of the circle network is related with the scale and longer than that of the star network. The calculated OCCI of the GSC for the organisms which have relatively large GSCs (node number greater than 40) are calculated. Figures 6a and b show the relationships of the average path length of GSC with the network scale and the network centralization index respectively. A relatively large scattering in the correlation between ALG and the scale of GSC exists. However, an obvious relationship between ALG and the centralization index is observed: ALG decreases with

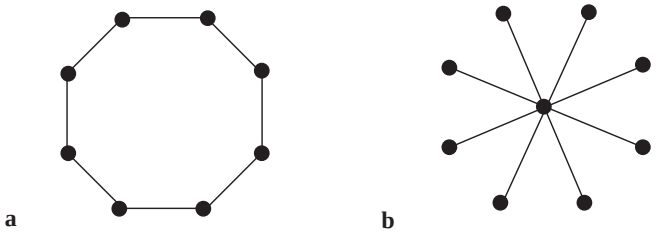

Fig. 5. Two different types of networks: circle network (**a**) and star network (**b**). The connections in these networks are undirected. If the node number is n, the value of the overall closeness centralization index (OCCI) is 0 for the circle network, its average path length is $(n+1)/4$; for the start network, the value of OCCI is 1, the average path length is close to 2.

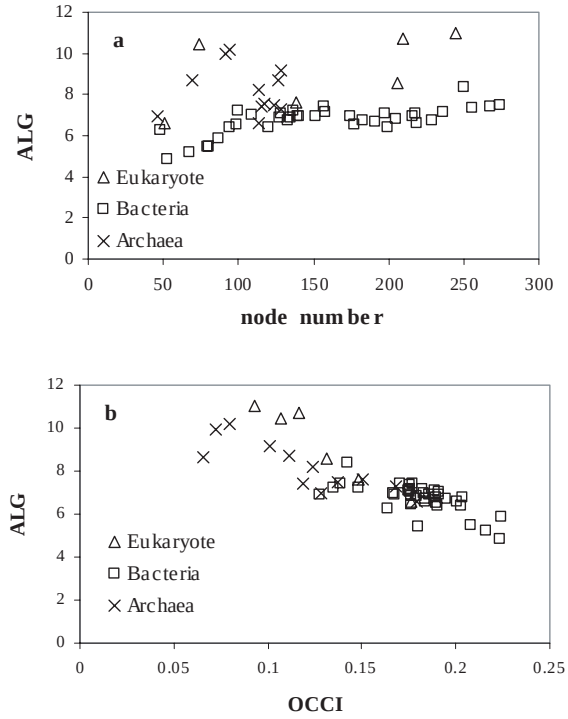

Fig. 6. Relationships between the average path length of the GSC (ALG) and the node number (network scale) (**a**) and the overall closeness centralization index OCCI (**b**) for various organisms.

OCCI. For the three domains of organisms, eukaryotes and archaea have longer ALG and smaller values of OCCI than bacteria. This reveals that the average path length is mainly determined by the network connection structure that can be quantitatively represented by OCCI. Thus OCCI can be used as a measure of small-worldness of the metabolic network.

The difference in the ALG and OCCI of different organisms is due mainly to the different connection structures as shown in Figure 4 for the core networks of *S.pneumoniae* and *P.furiosus*. The OCCI for the two core networks are calculated to be 0.2393 (spn) and 0.1027 (pfu) respectively. The average path length is also very different (5.56 and 7.89 respectively). In more detail, we may analyze the pathways from N-carbamoyl ASP (CAASP) to other metabolites as an example to show the difference between these two networks. N-carbamoyl ASP is the farthest node according to its average distance to other metabolites in both cores. There is a common six-step pathway from CAASP to glyceraldehyde phosphate (G3P), a key metabolite in the glycolysis pathway. Only through G3P CAASP can access all the other nodes. In the core of *S.pneumoniae*, G3P is highly connected and is the most central node in the network with a value of closeness centrality 0.307. However, in the core of *P.furiosus*, G3P is not the most central node and has a lower value of centrality (0.17). This leads to a higher value of average distance for CAASP in the core of *P.furiosus* (11.6, compared with 9.48 in the core of *S. pneumoniae*). From this example, we can see how the network centrality affects the average path length. Metabolites in metabolic network of bacteria are clustered compactly by certain highly connected central nodes, leading to a short average path length.

By showing a power law (or similar) connection degree distribution several previous studies suggested that most real networks are small world networks. However, it should be mentioned that the connection degree distribution is merely a local structure property of the network. It only makes use of the information with respect to how many nodes are directly connected with a specific node. Matching degree distribution does not mean a match in the large-scale global properties. For example, the degree distribution of the whole metabolic network also follows a power law, but a significant part of it is not connected with each other at all (Fig. 1). In contrast to the connection degree distribution, the centrality distribution represented by the overall closeness centralization index can better reflect the large-scale structure properties, because it considers not only the nodes directly connected with the specific node, but also all the other connected nodes and how far they are. In the above, we have shown the relationship between the average path length and the overall closeness centralization index of metabolic network. This structure parameter may also be used for the study of other networks such as computer networks, protein networks and social networks.

CONCLUSION

The results presented in this work showed that the metabolic networks of various organisms are not fully connected. A ‘bow-tie’ similar connectivity structure that

was previously found for the web network connection structure also exists in metabolic networks. GSC, the so-called giant strong component, represents the most complex part and the core of a metabolic network. The GSC exhibits characteristics of a small world network. The average path length of GSC is nearly linearly related with that of the whole network. The uncovered connectivity structure represents a rational reduction of seemingly complicated metabolic networks. It renders the structural and functional analysis of these networks more amenable and focused.

The parameter 'unit closeness centrality' can be used to identify the most central metabolites in the GSC of an organism. To describe the distribution of unit closeness centrality in the GSC we introduced the term 'overall closeness centralization index (OCCI)' that turned out to correlate well with the average GSC path length. The average path length is mainly determined by OCCI, but not the network scale. The GSCs of organisms from the three domains of life showed clearly different values of OCCI and thus also different average path length. The structural and functional details for these differences deserve further investigation that may shed more lights on the design principles and evolution of metabolic networks.

ACKNOWLEDGEMENT

This work was financially supported through the project 'Intergenomics' of the Ministry for Education and Research (BMBF), Germany (Grant No. 031U110A) and the National Natural Science Foundation of China (Grant No: 20028607 and 20036010). The authors thank Dr Hanno Biebl for his critical reading of this paper.

REFERENCES

Amaral,L.A.N., Scala,A., Barthelemy,M. and Stanley,H.T. (2000) Classes of small-world networks. *Proc. Natl Acad. Sci. USA*, **97**, 11149–11152.

Bailey,J.E. (2001) Reflections on the scope and the future of metabolic engineering and its connections to functional genomics and drug discovery. *Metab. Eng.*, **3**, 111–114.

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

Bilke,S. and Peterson,C. (2001) Topological properties of citation and metabolic networks. *Phys. Rev. E.*, **64**, 036106.

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

Burgard,A.P. and Maranas,C.D. (2001) Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.*, **74**, 364–375.

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

Freeman,L.C. (1979) Centrality in social networks: Conceptual clarification. *Social Networks*, **1**, 215–239.

Goto,S., Nishioka,T. and Kanehisa,M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.

Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

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

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc database. *Nucleic Acids Res.*, **30**, 56–58.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

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

Noble,D. (2002) The rise of computational biology. *Nat. Rev. Mol. Cell Biol.*, **3**, 459–463.

Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

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

Price,N.D., Papin,J.A. and Palsson,B.O. (2002) Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.*, **12**, 760–769.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Redner,S. (2002) Networking comes of age. *Nature*, **418**, 127–128.

Sabidussi,G. (1966) The centrality index of a graph. *Psychometrika*, **31**, 58–603.

Saqi,M.A. and Sternberg,M.J. (2001) A structural census of metabolic networks for *E.coli*. *J. Mol. Biol.*, **313**, 1195–1206.

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

Schuster,S., Pfeiffer,T., Moldenhauer,F., Koch,I. and Dandekar,T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics*, **18**, 351–361.

Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Tettelin,H., Nelson,K.E., Paulsen,I.T., Eisen,J.A. *et al.* (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B*, **268**, 1803–1810.

Wittig,U. and De Beuckelaer,A. (2001) Analysis and comparison of metabolic pathway databases. *Brief. Bioinform.*, **2**, 126–142.

Wolf,Y.I., Karev,G. and Koonin,E.V. (2002) Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, **24**, 105–109.