



Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms

Hongwu Ma^{1,2} and An-Ping Zeng^{1,*}

¹GBF—German Research Center for Biotechnology, Microbial Systems, Mascheroder Weg 1, 38124 Braunschweig, Germany and ²Department of Bioengineering, School of Chemical Engineering & Technology, Tianjin University, 300072 Tianjin, P.R.China

Received on June 24, 2002; revised on August 14, 2002; accepted on August 28, 2002

ABSTRACT

Motivation: Information from fully sequenced genomes makes it possible to reconstruct strain-specific global metabolic network for structural and functional studies. These networks are often very large and complex. To properly understand and analyze the global properties of metabolic networks, methods for rationally representing and quantitatively analyzing their structure are needed.

Results: In this work, the metabolic networks of 80 fully sequenced organisms are *in silico* reconstructed from genome data and an extensively revised bioreaction database. The networks are represented as directed graphs and analyzed by using the 'breadth first searching algorithm to identify the shortest pathway (path length) between any pair of the metabolites. The average path length of the networks are then calculated and compared for all the organisms. Different from previous studies the connections through current metabolites and cofactors are deleted to make the path length analysis physiologically more meaningful. The distribution of the connection degree of these networks is shown to follow the power law, indicating that the overall structure of all the metabolic networks has the characteristics of a small world network. However, clear differences exist in the network structure of the three domains of organisms. Eukaryotes and archaea have a longer average path length than bacteria.

Availability: The reaction database in excel format and the programs in VBA (Visual Basic for Applications) are available upon request.

Supplementary Material: For Supplementary Material refer to *Bioinformatics* Online.

Contact: aze@gbf.de; hwm@gbf.de

INTRODUCTION

The fast development in genome sequencing provides a large amount of genome data. Up to now, about 90

organisms, including bacteria, archaea and eukaryotes, have been fully sequenced. The exploitation of this information for understanding the organization principle of genetic and metabolic networks is one of the important research areas in the post-genome era. From gene annotation information, the genes are classified into different functional groups. A great number of gene products are enzymes that catalyze cellular reactions forming a complex metabolic network. As a dynamically regulated, complex interactive nonlinear system, the study of metabolic networks has gained increasing attentions in recent years (Schuster *et al.*, 2000; Kuffner *et al.*, 2000; Price *et al.*, 2002; Schuster *et al.*, 2002). Several methods and metabolic databases are available to reconstruct an organism specific metabolic network from genome information, such as KEGG (<http://www.genome.ad.jp/kegg>), WIT (wit.mcs.anl.gov/WIT) and EcoCyc (biocyc.org/ecocyc/) (Ogata *et al.*, 1999; Overbeek *et al.*, 2000; Karp *et al.*, 2002). These databases can serve as an easy-to-use research platform for further analysis of the global metabolic network. Efforts have also been made to analyze the structure of whole metabolic networks. For example, Schilling *et al.* (1999) and Edwards *et al.* (2002) proposed to use flux balance analysis for predicting the flux distribution in a whole metabolic network under certain physiological conditions. However, this method gives little information about the network overall structure. Several methods such as elementary flux mode analysis and extreme pathway analysis (Schilling *et al.*, 2000; Schilling and Palsson, 2000; Schuster *et al.*, 1999, 2000) have been developed for analyzing the pathway structure of metabolic networks. However, these methods are primarily suitable to analyze relatively small networks. For large-scale networks reconstructed from genome information, decomposition methods should be first used to divide the whole network into small subsystems. The pathway structure of these subsystems may then

*To whom correspondence should be addressed.

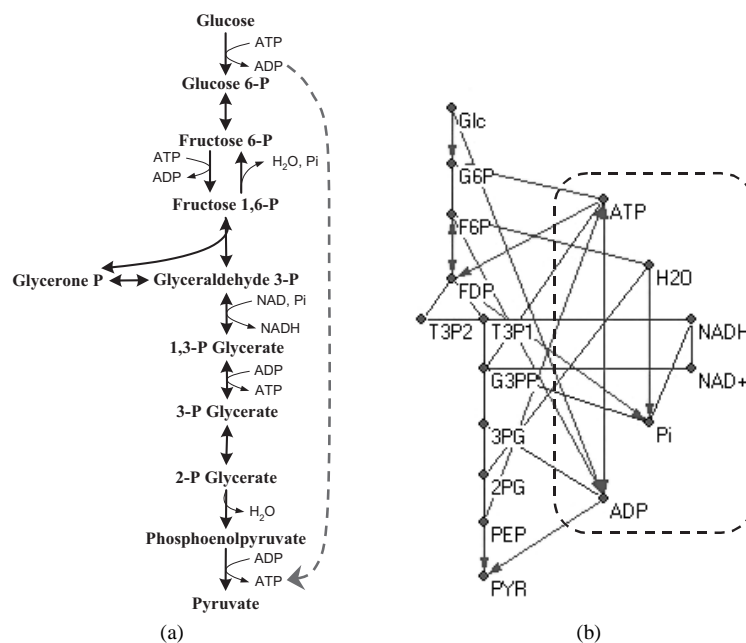


Fig. 1. The glycolysis pathway as a part of metabolic network. (a) The conventional way of presentation in biochemistry; (b) the connection structure in a graphic representation with current metabolites also as connections.

be properly analyzed by these methods (Schilling and Palsson, 2000; Schuster *et al.*, 2002).

Methods based on graph theory were shown to be useful for network structure analysis (Albert *et al.*, 2000; Jeong *et al.*, 2000). In these methods the metabolites correspond to nodes in the graph, and reactions correspond to connections between these nodes. Using such a graphic representation, Barabasi and his coworkers have studied the structure of metabolic networks using methods adopted from studies of the world wide web (Albert *et al.*, 2000; Jeong *et al.*, 2000). They found that like all the other networks studied, metabolic networks exhibited typical characteristics of small world networks. This means that most of the nodes have a low connection degree, while few nodes have a very high connection degree. The connection degree distribution follows a power law. The high degree nodes dominate the network structure and are called hubs of the network. Most of the nodes are connected through them by a relatively short path (Strogatz, 2001; Albert and Barabasi, 2002). For metabolic network, Jeong *et al.* (2000) calculated the average path length (AL), which is defined as the shortest path length averaged for every pair of metabolites in the whole network, for 43 organisms and found that AL is almost the same (about 3.2) for all the organisms. This means that most of the metabolites can be converted to each other in about only 3 steps. These results are surprising and in fact unexpected in view of the often long pathways for the synthesis of many metabolites. It is no-

ticed that in the study of Jeong *et al.* (2000), ATP, ADP and other current metabolites were also regarded as nodes in the network. This resulted in an unrealistic definition of the path length in many cases as illustrated with a part of the glycolysis pathway in Figure 1. It is obvious that the path length (number of reaction steps in the pathway) from glucose to pyruvate should be nine in terms of biochemistry. However, having accounted ATP and ADP as nodes in the network and due to the functions of ATP and ADP as cofactors in many reaction steps the path length between glucose and pyruvate was accounted as 2 in the work of Jeong *et al.* (2000) (dashed line in Figure 1a). This calculation of path length is obviously biochemically not meaningful. Similar problems can be also shown with the use of other current metabolites such as NADH, NAD⁺ etc. as nodes in Figure 1b.

Another characteristic of metabolic pathways is the irreversibility of many reactions as shown in Figure 1. Information about reaction reversibility is important in network analysis. However, until now there is no metabolic reaction database available that gives clear and enough information about it.

In this work, a more extensive and carefully revised bioreaction database is used to reconstruct the metabolic networks of 80 fully sequenced organisms from genome data. By distinguishing the major current metabolites from normal metabolites and by accounting for the effects of reaction irreversibility we show that quantitative

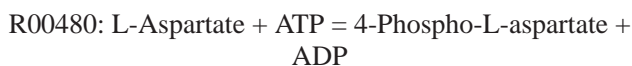
differences exist in the network structure of the three domains of organisms.

DATABASE FOR METABOLIC NETWORK RECONSTRUCTION

Reaction database

Our reaction database is based on the KEGG LIGAND database (Goto *et al.*, 1998, 2002). The LIGAND database includes three sections: COMPOUND, REACTION and ENZYME. Reaction information is included in the files 'reaction_name.lst' and 'reaction.lst'.

The file 'reaction_name.lst' lists all reactions appearing in the ENZYME section and KEGG/PAHTWAY database (5166 reactions in the release version 20.0). Each reaction is given a specific ID number. The file 'reaction.lst' is the same as 'reaction_name.lst' except that the compound names are converted to compound indices. For example, the following reaction in the 'reaction_name.lst':



is converted in the 'reaction.lst' into



The index representation is proper for program processing, while the name representation is more convenient for the analysis of results. So both equations were imported into our database in Excel format.

There exist mistakes in the original reaction database such as inconsistencies in compound names and mistakes in the reaction equations. We made efforts to correct these mistakes. Polymerization reactions and reactions with macromolecule participation were not included in our database. The final version of our database contained 4772 reactions.

Reversibility information of bioreactions

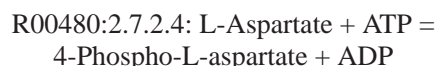
The reaction direction is shown in KEGG metabolic maps (direction inconsistencies in different maps exist). However, this information is not included in any database file. The reversibility information was checked or added manually according to rules stated below. The following kinds of reactions were regarded as irreversible in metabolic network:

- (1) Oxygen consumption reactions,
- (2) Most of the carbon dioxide production reaction (CO_2 is only used as substrate when a high energy substrate such as PEP (phosphoenolpyruvate) and ATP is consumed at the same time),
- (3) Most of the NH_3 production reaction (NH_3 is only used in two NH_3 assimilation reactions and carbamoyl phosphate production),
- (4) Most of the phosphate production reaction (the reaction is regarded as reversible only when phosphate reacts with another high energy substrate such as AcCoA)
- (5) Reactions in which S-Adenosyl-L-methionine is converted to S-Adenosyl-L-homocystine for providing a methyl group,
- (6) Reactions in which tetrahydrofolate (THF) is produced for transferring one carbon unit,
- (7) Most of the ATP (or other high energy metabolites) consumption reactions, except for reactions with another high energy metabolite such as GTP, CAP, acetyl phosphate, Acyl-CoA,
- (8) UDP-sugar consumption reactions for transferring sugar units,
- (9) CDP-diacylglycerol consumption reactions for phosphatidyl group transfer,
- (10) reactions like: 3'-Phosphoadenylylsulfate (PAPS) + A = adenosine 3',5'-bisphosphate (PAP) + B
- (11) several hydrolyzation reactions such as Acyl-R + H_2O = Acid (or fatty acid) + ROH, sugar-R + H_2O = sugar + ROH, Acyl-CoA + H_2O = Acid + CoA

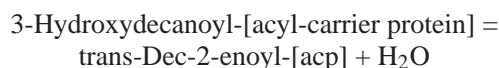
In the KEGG reaction database, some irreversible reactions have a wrong reaction direction. We have corrected this kind of mistakes. The number of irreversible reactions in our database is 1969.

Reaction-enzyme relation

The file 'reaction.tar.Z' contains the reactions that appear in the KEGG/PATHWAY database (3498 reactions). The data format in the file reaction.tar.Z is as follows:



This file was used to obtain the reaction-enzyme and enzyme-reaction relations. Note that these relations are not one to one. One enzyme may catalyze different reactions and the same reaction may be catalyzed by different enzymes. For example, the enzyme fatty-acid synthase (2.3.1.85) catalyzes 31 reactions in fatty acid synthesis pathway, while the reaction R04535:



is catalyzed by five different enzymes (2.3.1.85, 2.3.1.86, 4.2.1.58, 4.2.1.60, 4.2.1.61). In most enzyme databases, only the main reaction is listed for each enzyme. A complete metabolic network could not be reconstructed from such enzyme databases; this is a main reason why we choose the LIGAND database.

Enzyme–gene relation

Having constructed the overall reaction database, we should look at which enzyme is coded in a specific organism's genome so that we could reconstruct an organism specific metabolic network. There are different methods to obtain the enzyme–gene relation from KEGG database. One method is directly from the ENZYME section of the LIGAND database. An enzyme record example can be seen from http://www.genome.ad.jp/dbget-bin/www_bget?enzyme+1.1.1.81.

The gene name in a specific organism corresponding to that enzyme was listed in the item 'GENES'. From this information the enzyme-gene relation can be obtained. The results were represented as a binary value matrix R , such that $R_{ij} = 1$ if enzyme i was coded in the genome of organism j , and $R_{ij} = 0$ otherwise.

METABOLIC NETWORK RECONSTRUCTION AND REPRESENTATION

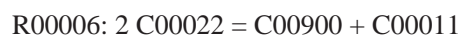
Metabolic network reconstruction and updating

The organism specific metabolic networks were reconstructed from the enzyme–gene relation and the reaction–enzyme information. The results were saved in a binary matrix R , where $R_{ij} = 1$ if reaction i exists in the metabolic network of organism j . The number of enzymes and reactions in the metabolic networks of 80 fully sequenced organisms in KEGG database can be seen in the supplementary material.

It should be mentioned that the genome annotation information is being continually updated. For most of these sequenced organisms, only about half of the ORFs have been assigned functions and errors may exist in certain annotated genes. However, most of the unknown genes or uncertain annotations are related to regulatory and signal transduction pathways or other complex cellular functions. For genes that code for metabolic enzymes, most of them can be identified accurately because the main part of metabolic network have been discovered and a large amount of metabolic enzymes identified experimentally. This makes it possible to construct a fairly completed metabolic network for the fully annotated organisms (Schilling *et al.*, 1999). The KEGG databases are continually updated by incorporating the most up-to-date gene annotation information from various databases (Kanehisa *et al.*, 2002). We can update the enzyme–gene and reaction–enzyme relations from the newest KEGG database files to incorporate new information for reconstructing strain-specific metabolic networks. The updating process can be carried out automatically by programs integrated in our database. For newly sequenced and annotated organisms, the corresponding metabolic networks can also be constructed through these updating programs.

Graph representation of metabolic networks

The metabolic network structure can be represented by a directed graph. In this representation the directed connections (corresponding to irreversible reactions in metabolic network) are called arcs, and undirected connections (reversible reactions) are called edges. For example, for the reversible reaction:

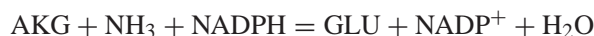


the corresponding edges are C00022–C00900 and C00022–C00011. Where C00022 is the compound index of the substrate, C00900 and C00011 are the indices of the products. In such a way, we can write a list of arcs and edges that represent its metabolic network structure for every organism from its reaction constituent. This is the basis for further structure analysis.

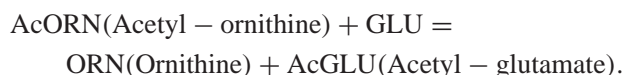
Removal of connections through current metabolites

The possible connection structure for a part of the glycolysis pathway is schematically shown in Figure 1b. The current metabolites (ATP, ADP, NADH, NAD⁺, H₂O and Pi) and the possible connections through them are shown in the dashed line area in Figure 1b. Current metabolites are normally used as carriers for transferring electrons and certain functional groups (phosphate group, amino group, one carbon unit, methyl group etc.) (Neidhardt *et al.*, 1990). As mentioned in the Introduction part, the connections through current metabolites should be avoided in calculating the path length from one metabolite to another. Therefore, these connections have been deleted from the list of arcs and edges obtained from the reaction database.

It should be mentioned that current metabolites cannot be defined *per se* by compounds but should be defined according to the reaction. For example, glutamate (GLU) and 2-oxoglutarate (AKG) are current metabolites for transferring amino groups in many reactions, but in the following reaction:



they are primary metabolites. The connections through them should be considered. The same situations are for NADH, NAD⁺ and ATP etc. Another problem is for the kind of reactions like:



The acetyl group is transferred between GLU and ORN in this reaction. Only the connections AcORN-ORN, GLU-AcGLU are included, but AcORN-AcGLU and GLU-ORN are excluded. If the latter two connections are

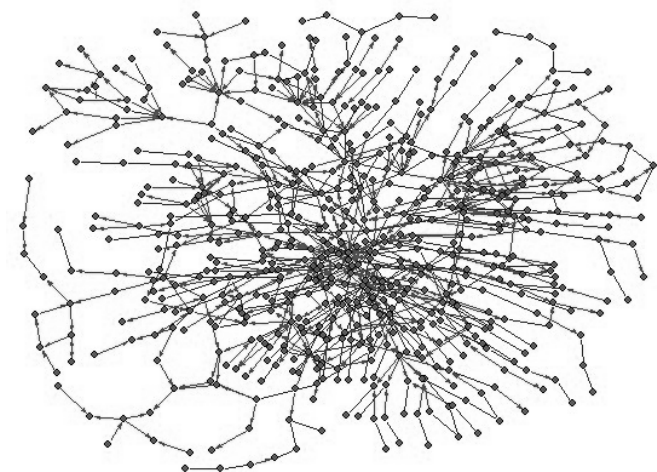


Fig. 2. Metabolic network structure of *E. coli* reconstructed from genome data. The lines with arrows correspond to irreversible reactions and the lines without arrow to reversible reactions. The picture was drawn using the Pajek program (Batagelj and Mrvar, 1998). (See supplementary data for colour version of this figure.)

considered, the path length from GLU to ORN will be one, and this is not in accordance with the pathway in real biochemistry.

The small molecules such as H_2O , NH_3 , O_2 , CO_2 and phosphate are also considered as current metabolites in this work (Neidhardt *et al.*, 1990). Our classification of current metabolites is somewhat similar to the classification of internal metabolites and external metabolites by Schuster *et al.* (2002). The current metabolite is like the external metabolite that participates in many reactions and is not in pseudo steady state in a sub-network. However, the external metabolites in the work of Schuster *et al.* (2002) was defined for the small sub-network. If considering the whole network, there will be no external metabolites except the input and output.

From the above discussion we can see that it is difficult to remove the connections through current metabolites automatically by program. We have to remove them manually. We have checked about 3000 reactions that appear in the KEGG metabolic maps and added corresponding connections one by one. In this way, a reaction–connection relation database was built, from which we can get the corresponding graph for any subset of reactions. This manual process does not lead to any difficulty in updating the metabolic networks with the newest annotation information. When the genome information is updated or a new genome is fully sequenced, we only need to update the reaction–enzyme and enzyme–gene relation databases; the reaction–connection relation is not affected.

More than half of the connections were removed by

deleting the connections through current metabolites as defined above. The structure of the resulted network was therefore more realistic and more amendable for analysis. As an example, the structure of the reconstructed metabolic network of *E. coli* is graphically shown in Figure 2.

GLOBAL PATHWAY STRUCTURE OF METABOLIC NETWORKS

Connection degree distribution

As mentioned in the Introduction part, Jeong *et al.* (2000) showed that metabolic networks were small world networks when considering the connections through current metabolites. We first examined if the structure of metabolic networks still have the characteristics of a small world network after deleting the connections through most current metabolites. One important feature of small world network is the power law distribution of connection degree among the nodes (Strogatz, 2001). In contrast to the small world network, a random network has a Poisson distribution of the connection degree. Here the connection degree is defined as the number of connections linked with that metabolite. Considering the direction, the number of connections starting from the metabolite is called output degree and the number of connections ending at the metabolite is called input degree. The output degree distributions of four typical organisms (*hsa*, *eco*, *bsu* and *ape* for eukaryotes, proteobacteria, gram positive bacteria and archaea, respectively) are shown in Figure 3. In this figure, $P(k)$ is the fraction of nodes which have a k -degree of outputs. It was calculated by dividing the number of metabolites which had k output connections with the total number of metabolites in the organism. Except for the first point with $k = 1$, a clear power law distribution (linear relations in the logarithmic scale coordinates) can be ascertained (Figure 3). The input degree distributions for these organisms also have similar power law relations (data not shown). The power law degree distribution exists in networks of all the organisms studied. This indicates that the metabolic network without connections through current metabolites is still a small world network.

Jeong *et al.* (2000) ranked the metabolites according to their connection degree. They found that the hub metabolites were almost the same among all the organisms. Actually, most of the hub metabolites they identified are current metabolites as defined in our analysis (e.g. H_2O , ATP, ADP). Having excluded these current metabolites we identified 20 primary metabolites with the highest degree of connection for every organism and ranked these metabolites by the number of organisms in which they appeared as hubs. The results are shown in Table 1. It can be seen that the first ten metabolites

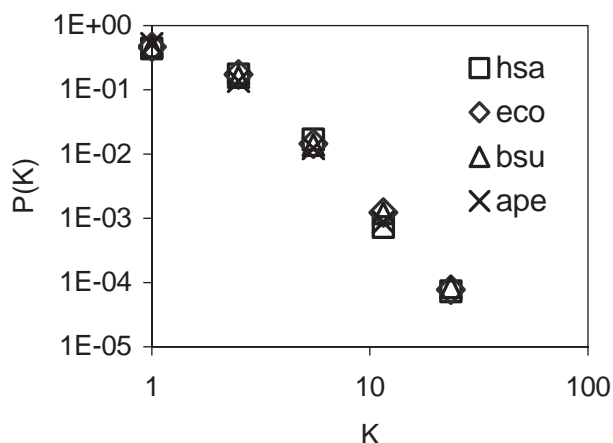


Fig. 3. Output degree distribution in four typical organisms. $P(k)$ is the fraction of nodes which have a k -degree of output connections. The original data has been logarithmically binned according to Huynen and van Nimwegen (1998). hsa: *Homo sapiens* (eukaryote), eco: *Escherichia coli* (gram negative bacteria), bsu: *Bacillus subtilis* (gram positive bacteria), ape: *Aeropyrum pernix* (archaea).

are hubs in most organisms. These include glycerate-3-phosphate, pyruvate, D-fructose-6-phosphate and D-glyceraldehyde-3-phosphate which are intermediates in the glycolysis pathway; D-ribose-5-phosphate and D-xylulose-5-phosphate which are intermediates in the pentosephosphate pathway and acetyl-CoA which is the metabolite linking glycolysis pathway, citric acid cycle and fatty acid synthesis pathway. 5-Phospho-D-ribose 1-diphosphate, the precursor for purine and histidine synthesis, is also within the first ten hub metabolites. L-Glutamate and L-aspartate, two important amino acids which are directly produced from precursors in citrate acid cycle and can be converted to many other amino acids, also have high connection degrees in most organisms. The universality of these high connection degree metabolites indicates their importance in metabolic network. They dominate the network structure in different organisms.

Average path length and network diameter

The primary function of metabolic network is to convert certain substrates to products (building blocks), and at the same time to acquire the energy and reducing power for growth and maintenance. It is one of the main tasks of metabolic network analysis to find possible conversion pathways between two metabolites. There may be many different pathways between two metabolites, among them the shortest pathway is of particular interest for network analysis. The shortest pathways from one metabolite to all other reachable metabolites were identified by the 'breadth first searching' method (Broder *et al.*, 2000). In

Table 1. The first 20 hub metabolites of metabolic networks ranked by the number of organisms in which the metabolite is a hub metabolite

Metabolite name	Abbreviation	Number of organisms
Glycerate-3-phosphate	3PG	80
D-Ribose-5-phosphate	R5P	80
Acetyl-CoA	AcCoA	78
Pyruvate	PYR	77
D-Xylulose 5-phosphate	X5P	75
D-Fructose 6-phosphate	F6P	73
5-Phospho-D-ribose 1-diphosphate	PRPP	69
L-Glutamate	GLU	69
D-Glyceraldehyde 3-phosphate	T3P	67
L-Aspartate	ASP	61
Propanoyl-CoA	PPCoA	46
Malonyl-ACP	Mal-ACP	45
Succinate	SUC	43
Acetate	ACT	41
Isocitrate	ICIT	39
Fumarate	FUM	39
Inosine 5'-phosphate	IMP	38
Oxaloacetate	OAA	34
Phosphoenolpyruvate	PEP	34
D-Glucose 6-phosphate	G6P	33

this method, the searching begins from one metabolite, M , in the network and proceeds to build up the set of nodes reachable from M in a series of layers. All the metabolites that are directly connected from M are in layer 1. All the metabolites which are directly connected from metabolites in layer k , but not in any earlier layers, are in layer $k + 1$. The layer number is the path length from M to the metabolites in that layer. In such a way, all the reachable metabolites and the path length from any metabolite can be found. Considering the reaction direction, the path length from metabolite A to metabolite B is not necessarily the same as that from B to A .

Glucose is the commonly used carbon source for many organisms. The shortest pathway structure from glucose to all the reachable metabolites in *E. coli* network is shown as an example in Figure 4. The number of reachable metabolites from glucose is 386. The average path length from glucose to all the reachable metabolites is calculated as 7.68. Repeating the above calculation process for all the metabolites in the network, the average path length (AL) for the whole network can be calculated. The AL value is 8.20 for *E. coli*. Another structure parameter is network diameter. It is defined as the path length of the longest pathway among all the shortest pathways (Batagelj and Mrvar, 1998). As shown in Figure 4, the longest pathway from glucose is the pathway to 6-pyruvoyltetrahydropterin, an intermediate in folate biosynthesis. The length of this pathway is 15. The network diameter can be calculated by comparing the

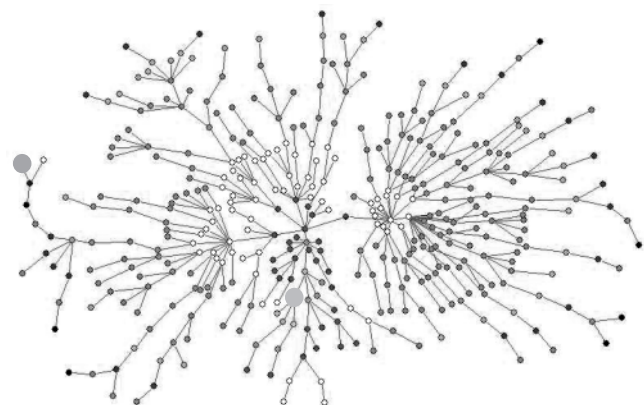


Fig. 4. The pathway structure from glucose to all reachable metabolites in *E. coli* metabolic network established by the 'breadth first searching' method. The big node near the center corresponds to glucose. The longest pathway is the pathway to 6-pyruvoyltetrahydropterin (big node at the left side of the graph). (See supplementary data for colour version of this figure.)

longest pathway for all the metabolites. It is 23 for *E. coli*.

The values of AL and network diameter for the 80 organisms were calculated (see supplementary material). Figure 5 shows the relations of AL with network scale that is represented by the node number. AL has a trend to increase with the network scale, especially for small-scale networks (e.g. node number less than 300). Similar trend can also be shown for the network diameter. We have checked the organisms with small-scale networks and found that they were all parasites. Further analysis revealed that the metabolic networks of these organisms were not well connected and contained many separated small networks or pathways. This resulted in a relatively short average path length. These results are consistent with the fact that the parasitic organisms have lost a great number of metabolic genes in the evolutionary process to adapt to the environments of host cells (Podani *et al.*, 2001). For the relatively large networks (i.e. node number greater than 300), the relation between AL and network scale (node number) is not very clear. Even for networks with a similar scale, the AL values vary greatly. For example, *Rattus norvegicus* (rno) and *Vibrio cholerae* (vch) have a similar scale, while their AL values are 10.99 and 7.64 respectively. These results clearly differ from the results of Jeong *et al.* (2000). They found a nearly constant and much shorter average path length for different kinds of organisms when using current metabolites as connections.

It can also be seen from Figure 5 that eukaryotes and archaea have a longer AL than bacteria. The average AL values for these three domains of organisms are 9.57, 8.50 and 7.22 (7.73 for bacteria without considering parasites),

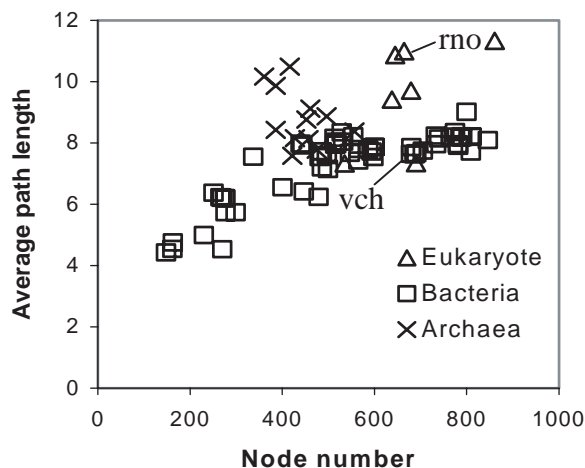


Fig. 5. Relation of average path length with network scale (node number).

respectively. Average diameter values are 33.1, 23.4 and 20.6, respectively. This indicates that, although the fundamental structure of metabolic network is similar for all the organisms, they do exhibit quantitative differences in the metabolic network structure as described by the parameters average path length and network diameter.

The quantitative differences in the metabolic network structure of different organisms reflect their different evolutionary history as shown in an evolutionary analysis of cellular metabolism based on the constructed metabolic networks (Ma and Zeng, unpublished results). The different average path length for the three different domains also indicates varied tightness and centrality of metabolic pathways/networks (Ma and Zeng, 2002). A more detailed comparison of the reactions and pathways of individual organisms may help to further understand the biological meaning of the structure differences and give useful hints for future work. For example, it may be interesting to identify the short-cuts that lead to the lower AL for bacteria. The reactions (enzymes and genes) corresponding to these short-cuts can then be determined. These enzymes may be specific for bacteria and important for metabolic conversion or pathogenic functions. This kind of knowledge can be used for strain improvement through metabolic engineering or for selecting proper pathway targets to develop drug against pathogenic bacteria.

CONCLUSION

By considering the reaction reversibility and deleting the connections through the current metabolites, a physiologically more meaningful metabolic network can be constructed from the annotated genome information of an organism. The constructed network can be well

represented by a directed graph. In this way, the network topology structure can be studied purely mathematically without any biochemistry bias. The metabolic networks of 80 fully sequenced organisms have been studied. Different from previously reported in the literature the average path length was found to differ among these organisms. Eukaryotes and archaea have a longer path length and a larger network diameter than bacteria, indicating difference in the structure of metabolic networks.

ACKNOWLEDGEMENT

This work was financially supported by the Bundesministerium für Bildung und Forschung (BMBF), Germany (Grant No. 031U110A) and by the National Natural Science Foundation of China (Grant No: 20028607).

SUPPLEMENTARY DATA

For Supplementary data, please refer to *Bioinformatics* online.

REFERENCES

- Albert, R. and Barabasi, A.L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47–97.
- Albert, R., Barabasi, A.L. and Jeong, H. (2000) Power-law distribution of the World Wide Web. *Science*, **287**, 2115.
- Batagelj, V. and Mrvar, A. (1998) Pajek—program for large network analysis. *Connections*, **21**, 47–57.
- 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.
- Edwards, J.S., Covert, M. and Palsson, B. (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ. Microbiol.*, **4**, 133–140.
- Goto, S., Nishioka, T. and Kanehisa, M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.
- Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.
- Huynen, M.A. and van Nimwegen, E. (1998) The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, **15**, 583–589.
- 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.
- Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
- Kuffner, R., Zimmer, R. and Lengauer, T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825–836.
- Ma, H.W. and Zeng, A.P. (2002) The hierarchical structure giant strong component and centrality of metabolic networks. *Bioinformatics*, submitted.
- Neidhardt, F.C., Ingraham, J.L. and Schaechter, M. (1990) *Physiology of the Bacterial Cell: a Molecular Approach*. Sinauer Associates.
- 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.
- Podani, J., Oltvai, Z.N., Jeong, H., Tombor, B., Barabasi, A.L. and Szathmari, E. (2001) Comparable system-level organization of Archaea and Eukaryotes. *Nature Genet.*, **29**, 54–56.
- Price, N.D., Papin, J.A. and Palsson, B.B. (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.
- Schilling, C.H., Edwards, J.S. and Palsson, B.O. (1999) Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.*, **15**, 288–295.
- Schilling, C.H., Letscher, D. and Palsson, B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.
- Schilling, C.H. and Palsson, B.O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, **203**, 249–283.
- Schilling, C.H., Schuster, S., Palsson, B.O. and Heinrich, R. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, **15**, 296–303.
- Schuster, S., Dandekar, T. and Fell, D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.
- Schuster, S., Fell, D.A. and Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.
- 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.