Genomics & Informatics Vol. 4(3) 118-124, September 2006

A Metabolic Pathway Drawing Algorithm for Reducing the Number of Edge Crossings

Eun-Ha Song[1]*, Min Kyung Kim[2] and Sang-Ho Lee[1]

[1]Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-750, Korea, [2]UC Irvine Institute for Genomics and Bioinformatics, Irvine, California 92697-3445, USA

Abstract

For the direct understanding of flow, pathway data are usually represented as directed graphs in biological journals and texts. Databases of metabolic pathways or signal transduction pathways inevitably contain these kinds of graphs to show the flow. KEGG, one of the representative pathway databases, uses the manually drawn figure which can not be easily maintained. Graph layout algorithms are applied for visualizing metabolic pathways in some databases, such as EcoCyc. Although these can express any changes of data in the real time, it exponentially increases the edge crossings according to the increase of nodes. For the understanding of genome scale flow of metabolism, it is very important to reduce the unnecessary edge crossings which exist in the automatic graph layout.

We propose a metabolic pathway drawing algorithm for reducing the number of edge crossings by considering the fact that metabolic pathway graph is scale-free network. The experimental results show that the number of edge crossings is reduced about 37~40% by the consideration of scale-free network in contrast with non-considering scale-free network. And also we found that the increase of nodes do not always mean that there is an increase of edge crossings.

Keywords: drawing algorithm, edge crossings, metabolic pathway, scale-free network

Introduction

The information of pathways and interactions are one of the important topics after the genome sequencing. The dynamic assembly or interaction among the proteins regulates the function of cells. The traditional representation of pathway flow, such as arrow and circle, has advantages in the familiarity and easiness to the biologist. For that reason, many kinds of pathway databases usually handle this information.

Metabolic pathways characterize the process of chemical reactions that perform a particular metabolic function. Because metabolic pathways have very complex structures, the graphical representations of these networks are necessary to understand functions of the networks easily.

Many pathway databases also contain the information of flows, such as KEGG (Kyoto Encyclopedia of Gene and Genomes), Biocarta, and so on (Becker and Rojas, 2001; BioCarta Team, 2001; Kanehisa *et al.*, 2002). In these databases, pathways are visualized in a static way. Pathway diagrams are manually drawn and storedas image files. Whenever the data has been updated, the corresponding images of pathways have to be edited manually to reflect the changes. Furthermore, static approach offers no flexibility in the level of detail to be displayed. Therefore, to solve these problems, new methods are required to visualize topological architecture of pathways automatically and dynamically.

Several specific dynamic layout algorithms for metabolic pathways have been developed. The pathway drawing algorithm in EcoCyc and the algorithm by Becker *et al.* are such examples (BioCarta Team, 2001; Karp *et al.*, 2002). These algorithms visualize metabolic pathways based on the fact that metabolic pathways are composed of circular and hierarchic components. However, these algorithms have the problem that they generate edge crossings exponentially according to the increase of nodes. Edge crossings may reduce the readability and cause the misconception.

In this paper, we describe a metabolic pathway drawing algorithm for reducing the number of edge crossings by considering the fact that metabolic pathway graph is scale-free network (Lee *et al.*, 2004). Also, we analyze our algorithm from the viewpoint of the number of edge crossings.

Related works

Pathway layout algorithm in EcoCyc

EcoCyc is one of the databases that describes the metabolic and signal transduction pathways of Escherichia *coli*, and also it contains the information of enzymes, transport

*Corresponding author: E-mail ehsong@ewhain.net, Tel +82-2-3277-3504, Fax +82-2-3277-2306
Accepted 20 June 2006

proteins and mechanisms of transcriptional control of gene expression. The graph layout algorithm that has been implemented in the EcoCyc is drawn by topological structure of the graph. It decomposes the graph into cyclic, linear and tree-structured components and then applies different layout methods individually. Because it does not consider the position of each node but only the position of each component (Karp and Paley, 1994). Therefore, it is impossible to predict and reduce the number of edge crossings which can exist in layout result.

Becker *et al.* algorithm

Becker *et al.* developed a divide and conquer method on the basis of circular and hierarchical structural characteristics of pathway similar to EcoCyc (Becker and Rojas, 2001). Unlike EcoCyc, the whole graph is considered as a sum of subgraph. Each subgraph has different types of graphs, such as cyclic and hierarchical subgraphs. These subgraphs are drawn according to their topology and are reassembled to a whole graph.

In order to apply the above approach, metabolic pathways are represented as compound graphs. Reactants and products (the compounds) are represented as nodes and the reactions are represented as edges of the graph. Some chemical reactions have more than one main reactant or product. In these cases, the reaction has to be represented as a hyper-edge, i.e. an edge with multiple source and target nodes. But these hyper-edges lead to the occurrence of edge crossings in layout results and reduction in readability (Becker and Rojas, 2001).

Methods

Scale-free network

Scale-free network satisfies the following common features. The first one is that the network continuously expands by the addition of new vertices that are connected to the vertices which are already in the system. Secondly, there is a high probability that it will be linked to a vertex that already has a large number of connections (Barabasi and Albert, 1999; Goh *et al.*, 2001).

As shown the Fig. 1, scale-free network is composed of a few nodes with high degree and many nodes with low degree which are connected to highly connected nodes.

As shown in the plot of Fig. 2, the probability $p(k)$ that each node in the network has k links decays as a power law in scale-free network. This result indicates that we can determine that the network is a scale-free network by plotting the relation of k and $p(k)$ for a given network (Barabasi and Albert, 1999; Goh *et al.*, 2001).

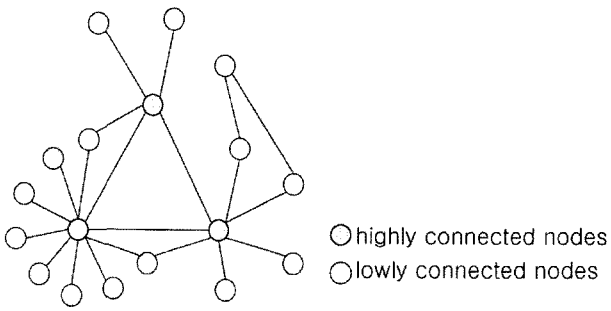


Fig. 1. An example of the scale-free network

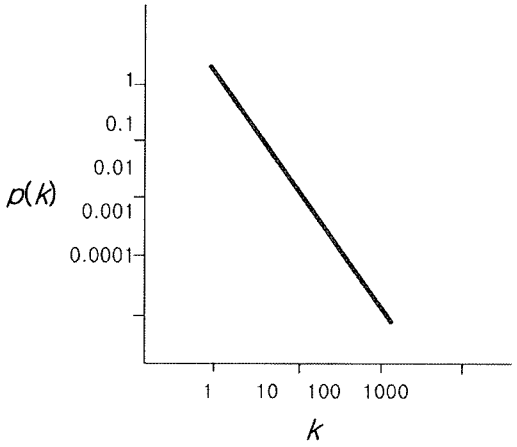


Fig. 2. The relation between k and p(k) of scale-free networks

Specially, Jeong *et al.* examined connectivity distribution $p(k)$ for the substrates k in *A. fulgidus* (Archae), *E. coli* (Bacterium) and *C. elegans* (Eukaryote) that have been obtained for all 43 organisms investigated in WIT database, respectively (Jeong *et al.*, 2000).

They showed that the connectivity distribution $p(k)$ for k in each domain follows the power law. Most metabolic pathways are scale-free networks (Jeong *et al.*, 2000). Consequently, it means that proteins with high connectivity are considered in the drawing of metabolic pathways.

Therefore, it is possible to reduce the edge crossings by placing protein (node) with high connectivity to the center of its connected components (Holme *et al.*, 2003).

Metabolic pathway drawing algorithm by considering scale-free network structure

Metabolic pathway drawing steps are as follows (Fig. 3): At first, the connectivity of nodes is considered for a given metabolic pathway graph. Namely, we should check if the input graph has scale-free network structure. In KEGG, the number of the nodes that their degree is more than 6 is few

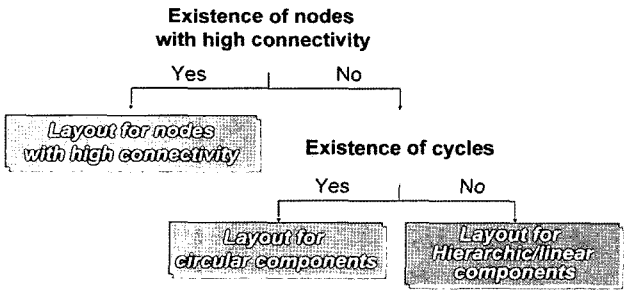


Fig. 3. Flowchart of metabolic pathway drawing algorithm

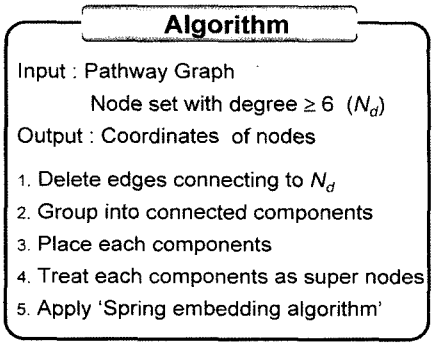


Fig. 4. Layout for nodes with high connectivity

in contrast with that of the nodes that their degree is less than 5. Therefore, in the first step, we consider that nodes which their degree is more than 6 are as nodes with high connectivity. Next, edges that are connected to the selected node with high connectivity are deleted. Then the remaining graph is grouped into connected components and the connected components are placed by using graph layout algorithms according to their topologies. After that, each component is collapsed into a super node. 'Spring embedding algorithm' provided by the graph library is applied to super nodes and the selected node with high connectivity. Spring embedding algorithm models the graph as a system of particles with forces acting on them and attempt to find minimum energy configuration of this system. Each edge acts as a spring with a preferred length and exerts a repulsive or attractive force on the nodes connected by it. Each spring length, i.e. edge length that connects each super node and the node with high connectivity is set to twice the longest one between the width and the length of the bounding box of the corresponding super node. Selection of long enough spring lengths is done in an attempt to avoid overlapping among super nodes. Finally, each connected component in each super node is mirrored properly. Each connected component is mirrored on x-axis, y-axis and xy-axis. By comparing with current layout and these mirrored layouts, we select a layout that has minimum sum of the length of edges between nodes in the connected component and the node with high connectivity as a final layout. This is a heuristic attempt to further reduce the number of edge crossings.

Fig. 4 shows the layout algorithm for nodes with high connectivity.

If any node with high connectivity does not exist in the given input graph, it is checked whether a cycle exists or not. If any cycle exists, then the longest cycle is found and nodes that do not belong to the cycle are grouped into connected components. Then the circular layout algorithm is applied to the cycle and each connected component is placed inside or outside the cycle by considering its topology. Actually, searching for the longest cycle is NP-complete problem. However, we can find a cycle component that is regarded as the longest cycle by using heuristic method.

If no circular components exist, then layout algorithms for linear or hierarchic components are applied to the graph. Then if the component is linear structure, linear/snake layout algorithm is applied and in case of tree-structured components, tree layout algorithm is applied. Otherwise, hierarchic layout algorithm is applied. Finally we can obtain a layout result for the whole graph.

Results

Experimental environment

Experiments for our algorithm were performed on Windows-XP environment and this algorithm was implemented in Java and the Java-based graph library yFiles which was used to create and manipulate the graph (Wiese *et al.*, 2000).

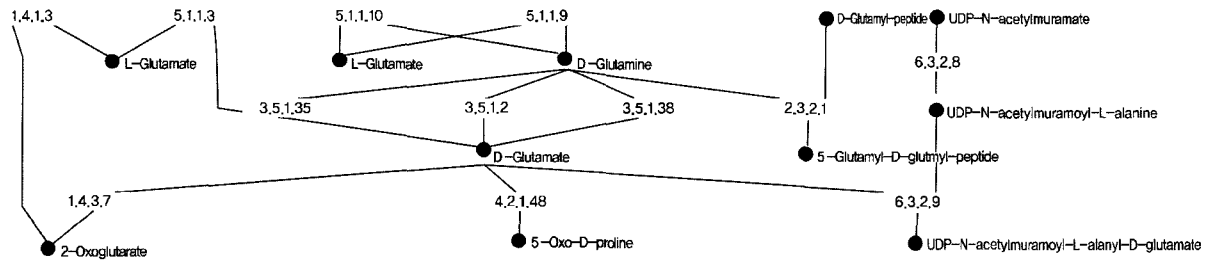
In order to experiment, we implemented two algorithms: one is the new algorithm i.e. an algorithm with considering scale-free network structure and the other is the existing algorithm which does not consider scale-free network structure, respectively.

Experiments were performed by two algorithms with metabolic pathways in KEGG as test data. Then we compared with thenumber of edge crossings for two layout results.

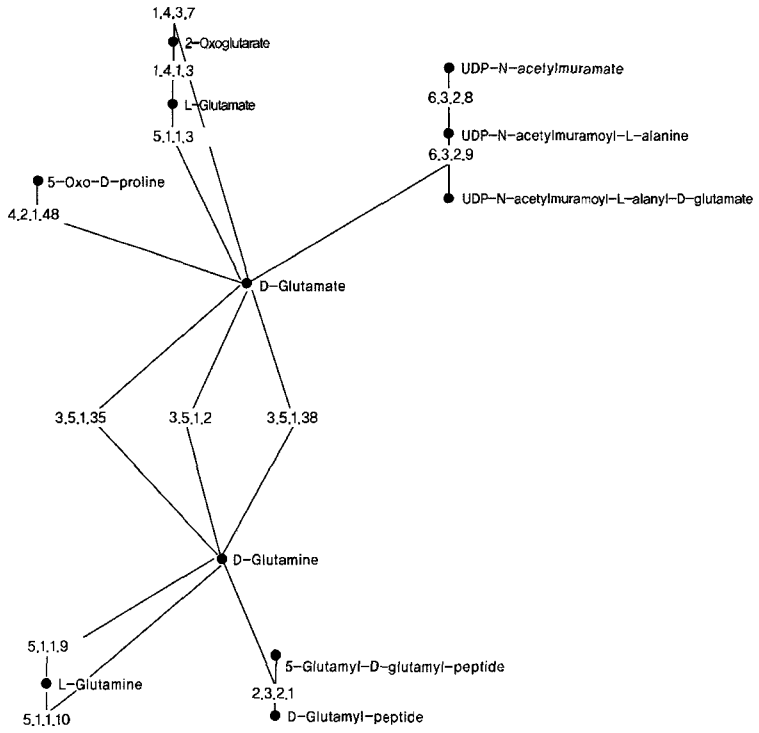
Experimental Results

A metabolic pathway in the Fig. 5 is a *D-Glutamine* and *D-Glutamate* metabolism, which is composed of two highly connected nodes and other linear components. The number of node is 12 and the number of edge is 26.

In both cases, we found only few edge crossings. Therefore, if the number of nodes and edges are small, the occurrence of edge crossings is also rare. Although



(a) A Layout with not considering scale-free network



(b) A layout with considering scale-free network

Fig. 5. *D-Glutamine* and *D-Glutamate* Metabolism

the number of edge crossing is not meaningful, the readability of this flow is increased by positioning the highly connected nodes in the scale-free network.

In the Fig. 6, it shows a layout result for 'Lysine biosynthesis' metabolic pathway with 36 nodes. As you see in the Fig. 6, Lysine biosynthesesmetabolic pathway is a complex and linear structured network which has no cycle as a subgraph. The number of edge crossings is reduced remarkably by considering nodes with high connectivity. We can easily grasp that highly connected nodes exist in the metabolic pathway.

Fig. 7 shows a layout results for a β-Alanine metabolic pathway, which has more complex structure and is composed of 60 nodes. Here the number of edge crossings is reduced remarkably by applying a new algorithm.

The experimental results show that the number of edge crossings is reduced about 37~40% as the number of nodes increases in pathway graph in contrast with non-considering scale-free network. Therefore, it is possible to reduce unnecessary edge crossings by considering the facts that metabolic pathway graph is scale-free network. Then it enhances the readability and the understanding of the flow of the metabolic pathway.

Especially, the number of edge crossings is reducedremarkably in case of pathway graphs with linear structure without circular components.

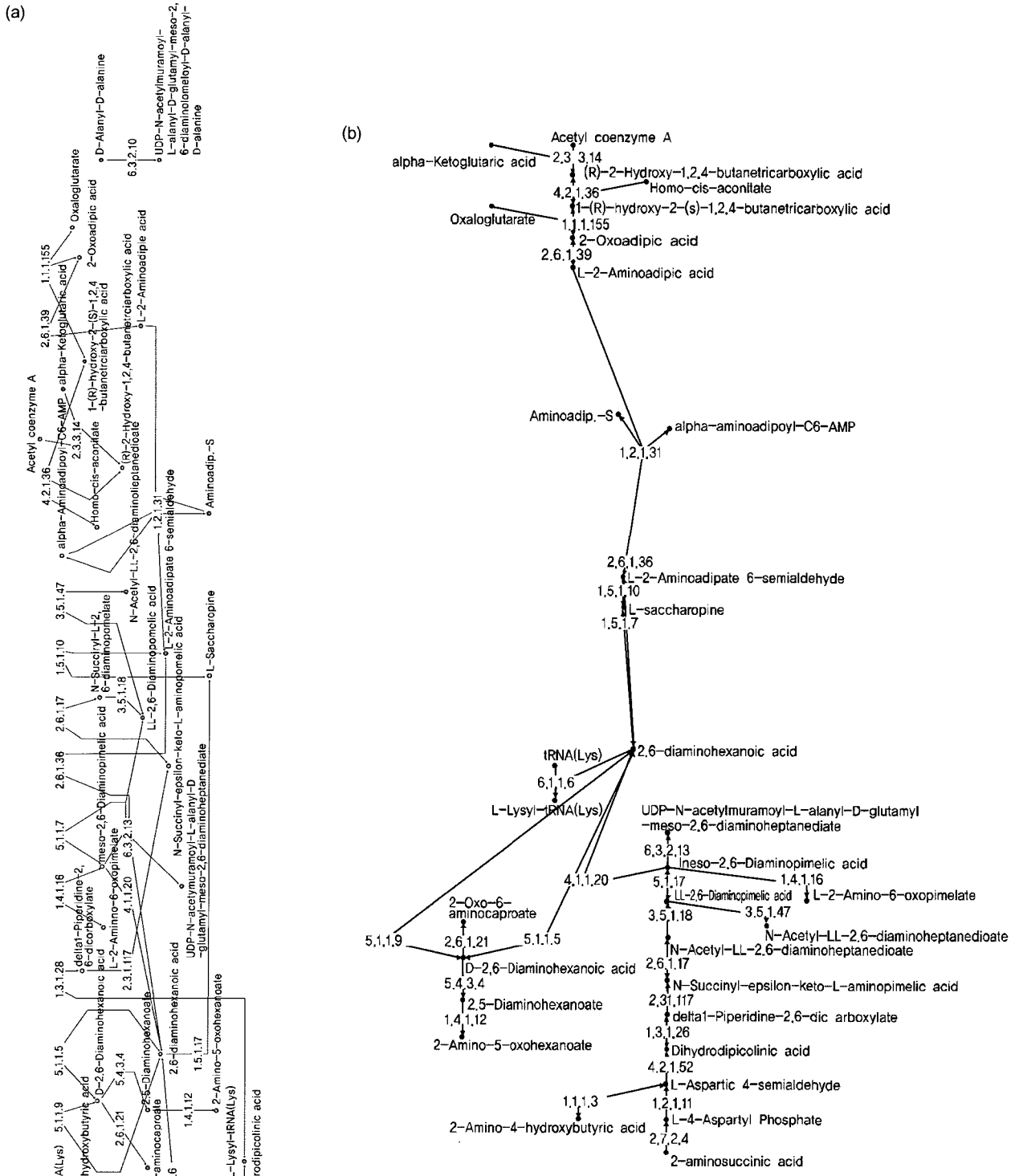
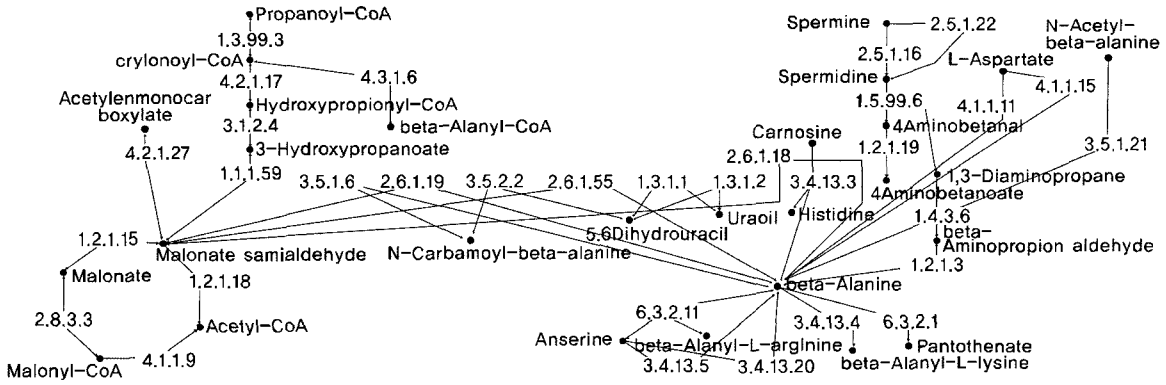
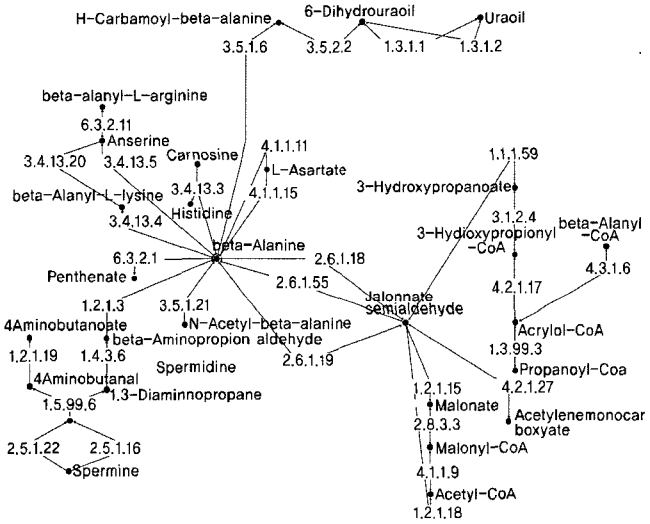


Fig. 6. Lysine biosynthesis metabolism. (a) A layout with not considering scale-free network (b) A layout with considering scale-free network



(a) A layout with not considering scale-free network



(b) A layout with considering scale-free network

Fig. 7. β- Alanine metabolism

Table 1. Experimental results

Metab olic pathway	# of nodes	The number of edge crossings		Existence of cycle
		A layoujt with not considering scale-free network	A layoujt with considering scale-free network	
D-Alanine	12	1	0	×
D-Glutamine and D-glutamate	23	1	0	×
Lysine biosynthesis	39	18	0	×
Preptidoglycan biosysnthesis	42	1	0	×
Toluene and xylene degradation	43	2	1	×
2-4-Dichlorobenzoate degradation	43	1	0	×
β-Alanine	60	13	1	○
Androgen and estrogen	63	46	36	○
Alanine aspartate	71	37	17	○
N-Glycans biosysnthesis	88	1	0	×
Pyruvate	94	170	106	○
Aspartate/Glutamate	122	160	96	○

Table 2. Experimental results

	KEGG	Ecocyc	Becker et.al algorithm	Our algorithm
node representation	compounds	compounds	compounds	compounds + enzymes
edge representation	EC number direction	direction	direction	direction
side-compound representation	○	○	×	×
usage of bipartite graph	×	×	×	○
visualization range	single pathway	single pathway	single pathway	multi pathways
graphical representation	static	dynamic	dynamic	dynamic

Table 1 shows experimental results for 12 metabolic pathways.

Discussion

In this paper, we described the metabolic pathway drawing algorithm based on ascale-free network structural characteristic and analyzed the algorithm from viewpoint of the number of edge crossings.

We experimented for 12 metabolisms in KEGG database and validated that the number of edge crossings was reduced according to the increase of nodes in case of considering scale-free network structure. In the case of linear or hierarchical graphs, we could obtain the best results.

The experimental results are very similar to common structure of metabolic pathway as compared with layout results in the previous systems it is possible to assist understanding of biological researchers. In contrast to visualize only a single pathway in previous layout system, our layout algorithm is possible to visualize multi pathways simultaneously. These results are suitable for research on genome scale flow of metabolism. Moreover, our algorithm enhances the readability by making use of the bipartite graph which represents not only compounds but also enzymes catalyzing the reaction as nodes in metabolic pathway graph structure.

However, 106 edge crossings exist in the complicated pathways, pyruvate, as shown in Table 1. Therefore new methods are needed for reducing edge crossing drastically or new module should be added to pathway layout algorithm for abstracting it in order to simplify the topology of the given graph. Hence, we now research a new method that finds highly connected subgraphs existing in a pathway graph, i.e. dique, and layouts by considering these subgraphs.

In the case of dynamic layout systems, there are weaknesses to represent side-compounds as shown in Table 2. It is because the side compounds can participate in many reactions in pathway. In the future, we need to discuss how to model and represent side-compounds.

References

Barabasi, A.L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science* 286, 509-512.

Becker, M.Y. and Rojas, I. (2001). A Graph Layout Algorithm for Drawing Metabolic Pathways. *Bioinformatics* 17, 461-467.

BioCarta Team. (2001). Biocarta: Charting Pathways of Life. *http://www.biocarta.com.*

Goh, K.I., Kahng, B., and Kim, D. (2001). Universal Behavior of Load Distribution in Scale-Free Networks. *Physical Review Letters* 87, 278701.

Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork Hierarchies of Biochemical Pathways. *Bioinformatics* 19, 532-538.

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.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG Databases at GenomeNet. *Nucleic Acids Res.* 30, 42-46.

Karp, P.D. and Paley, S. (1994). Automated Drawing of Metabolic Pathways. *Third International Conference on Bioinformatics and Genome Research*. 225-238.

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.

Lee, S.H.,Song, E.H., Lee, S.H., and Park, H.S. (2004). An Algorithm for Drawing Metabolic Pathways based on Structural Characteristics, *J. Korea Information Science Society* 31, 1266-1275.

Wiese, R., Eiglsperger, M., and Schabert, P. (2000). The Y-files Graph Library: Documentation and Code. *http://www-pr.informatik.uni-tuebingen.de/yfiles.*