A Statistical Analysis of Relation Degree of Compound Pair on Online Biological Pathway Databases

Myungha Jang,
Department of Computer Science, Ewha Womans University, Seoul 120-75, Korea
myunghajang@gmail.com

Arang Rhie
Department of Computer Science, Ewha Womans University, Seoul 120-75, Korea
arrhie@gmail.com

Jiyoung Whang
Department of Computer Science, Ewha Womans University, Seoul 120-75, Korea
Jiyoung88@ewhain.net

Sanduk Yang
Department of Computer Science, Ewha Womans University, Seoul 120-75, Korea
sandukyang@gmail.com

Hyun S. Park,
Institute of Bioinformatics, Ewha Womans University, Seoul 120-75, Korea & Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea
82-2-3277-2831
neo@ewha.ac.kr

ABSTRACT

The basic graph layout technique, one of the visualization techniques, deals with the problem of positioning the vertices in a way to maximize understandability and usability in a graph. This technique is becoming a vital part for further development in the field of systems biology. However, applying the appropriate automatic graph layout techniques to the genome scale flow of metabolism requires understanding of the characteristics of metabolites and reactions, which suggest valuable information to bioinformatics software developers for better visualization of automatic graph layout. In this paper, we define the term *relation degree of a compound pair* to provide a reasonable way to visualize metabolic pathway atlas, based on the parsing result of the publicly available XML files. It is a preliminary step for future research in the area of automatic layout techniques in large-scale biological pathway domain.

Categories and Subject Descriptors

J.3.1 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms, Measurement

Keywords

drawing algorithm, XML, metabolic pathway, parsing, edge crossing, Relation Degree

1. INTRODUCTION

Automatic graph layout algorithms take an abstract graph structure as input and produce a visual representation graph. A graph is represented using nodes and edges. A node indicates vertices, as arcs feature edges connecting vertices. Arrows show the orientation of directed edges [1]. Graphical diagrams are intuitively helpful to understand biochemical reaction networks. However, NP-hard problems are still left to achieve optimal solutions for automatic layout of biological pathways under respect of aesthetics. There have been initiative such as KEGG(http://www.genome.ad.jp/kegg) and EcoCyc(http://biocyc.org) for biopathway databases. Afterwards, there had been numerous attempts to use heuristic algorithms to reach approximate solutions while accomplishing computational efficiency [2-7].

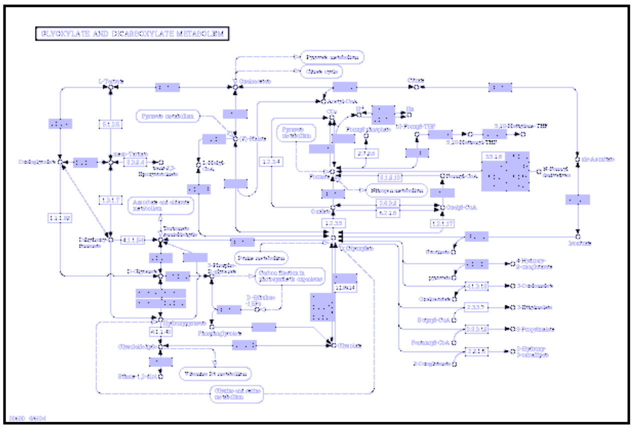


Figure 1. Map00630 reference pathway shows the manual layout of KEGG metabolic pathway of Glyoxylate and dicarboxylate metabolism (Source:http://www.genome.jp/dbgetbin/www_bget?pathway+map00630)

Unfortunately, a current state-of-the-art survey in this field reveals that research and development in automatic layout of biological pathway maps are still in its infancy from an aesthetic perspective. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database is a valuable information resource [8]. It

ICUIMC-09, January 15-16, 2009, Suwon, S. Korea
Copyright 2009 ACM 978-1-60558-405-8...$5.00.

contains metabolic pathway database in a form of wiring diagrams. Figure 1 shows the manually drawn pathway map for Glyoxylate and Dicarboxylate metabolism in KEGG. However, KEGG visualizes pathways in a static way. Pathway diagrams are manually drawn and stored as bitmap image files. These diagrams are displayed as interactive image maps with links to additional information on enzymes and to adjacent pathways [9]. While this visualization style offers a good pathway presentation, it does not provide the facilities to create and visualize dynamic pathways. Figure 2 is generated by an automatic layout scheme in KEGG to offer user flexibility for the same diagram in Figure 1. It is clear that Figure 2 differs in many aspects from Figure 1 or the conventional drawings in biochemistry textbooks; the arrangement of these vertices and edges impact understandability, based on aesthetics. For obvious reason, the strategy of automatic layout scheme is preferred to manually generated layout scheme for its flexibility.

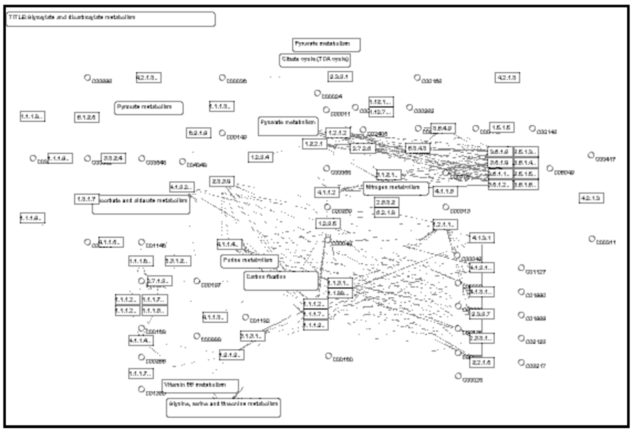

TITLE:Glyoxylate and dicarboxylate metabolism
Pyruvate metabolism
Citrate cycle (TCA cycle)
Pyruvate metabolism
Pyruvate metabolism
Nitrogen metabolism
Ascorbate and aldarate metabolism
Folate metabolism
Carbon fixation
Vitamin B6 metabolism
Glycine, serine and threonine metabolism

Figure 2. An automatic layout version of KEGG reference Map00630 of Glyoxylate and dicarboxylate metabolism

(Source:http://www.genome.jp/kegg-bin/xml/PathwayViewer?-v+0.6.1+map+00630)

Regardless of the drawing mechanisms, these different graphical representations in Figure 1 and Figure 2, should not be confused with the graph itself, i.e., the abstract, non-graphical structure. Different layouts can correspond to the same graph. At the level of software, a different format is needed for quantifying a model to the point where it can be simulated. KEGG offers a machine-readable format for representing models. The KEGG Markup Language (KGML: http://www.genome.jp/kegg/xml/) is an XML(Extensible Markup Language) representation of KEGG pathway maps, an exchange format of the KEGG graph objects. XML is classified as an extensible language because it allows users to define their own elements. XML is a flexible way to create common information formats and share both the format and the data on the World Wide Web [10],[11]. KGML enables automatic drawing of KEGG pathways and provides facilities for computational analysis and modeling of biological networks.

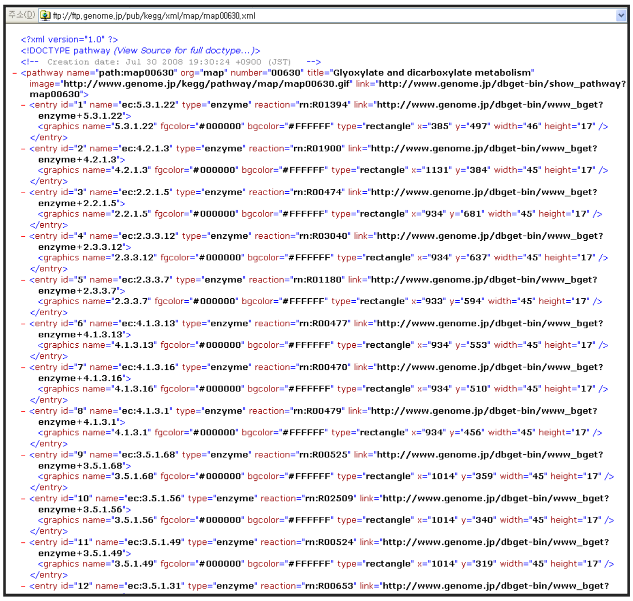
Regarding automatic graph layout, diverse aesthetic criteria have been developed in order to increase readability of the drawn graph. Such aesthetic criteria includes: minimizing the total number of edge crossings, minimizing the area of the drawing, displaying the symmetries of the graph, and minimizing the smallest angle between two edges adjacent to the same vertex. However, through several experiments, it has been proofed that reducing the number edge crossings is the prior aesthetic to consider [12].

Thus, to apply the automatic graph layout techniques to the genome scale flow of metabolism domain, understanding the characteristics of nodes and edges by parsing and analyzing KEGG XML files or KGML is crucial for software developers.

In this paper, we provide a parsing result of a publicly available database, KEGG, using our XML parser, to provide a statistic analysis of the characteristics of metabolites and reactions for automatic layout algorithms for global pathways in the area of systems biology.

2. Implementation of the KGML Parsing Module

The Extensible Markup Language (XML) is a general-purpose specification for creating custom markup languages. Its primary purpose is to help information systems share structured data. By adding semantic constraints, application languages can be implemented in XML.



```
  <?xml version="1.0" ?>
  <!DOCTYPE pathway (View Source for full doctype...)>
  <!--  Creation date: Jul 30 2008 19:30:24 +0900 (JST)   -->
- <pathway name="path:map00630" org="map" number="00630" title="Glyoxylate and dicarboxylate metabolism"
    image="http://www.genome.jp/kegg/pathway/map/map00630.gif" link="http://www.genome.jp/dbget-bin/show_pathway?
    map00630">
  - <entry id="1" name="ec:5.3.1.22" type="enzyme" reaction="rn:R01394" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+5.3.1.22">
      <graphics name="5.3.1.22" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="385" y="497" width="46" height="17" />
    </entry>
  - <entry id="2" name="ec:4.2.1.3" type="enzyme" reaction="rn:R01900" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+4.2.1.3">
      <graphics name="4.2.1.3" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="1131" y="384" width="45" height="17" />
    </entry>
  - <entry id="3" name="ec:2.2.1.5" type="enzyme" reaction="rn:R00474" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+2.2.1.5">
      <graphics name="2.2.1.5" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="934" y="681" width="45" height="17" />
    </entry>
  - <entry id="4" name="ec:2.3.3.12" type="enzyme" reaction="rn:R03040" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+2.3.3.12">
      <graphics name="2.3.3.12" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="934" y="637" width="45" height="17" />
    </entry>
  - <entry id="5" name="ec:2.3.3.7" type="enzyme" reaction="rn:R01180" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+2.3.3.7">
      <graphics name="2.3.3.7" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="933" y="594" width="45" height="17" />
    </entry>
  - <entry id="6" name="ec:4.1.3.13" type="enzyme" reaction="rn:R00477" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+4.1.3.13">
      <graphics name="4.1.3.13" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="934" y="553" width="45" height="17" />
    </entry>
  - <entry id="7" name="ec:4.1.3.16" type="enzyme" reaction="rn:R00470" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+4.1.3.16">
      <graphics name="4.1.3.16" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="934" y="510" width="45" height="17" />
    </entry>
  - <entry id="8" name="ec:4.1.3.1" type="enzyme" reaction="rn:R00479" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+4.1.3.1">
      <graphics name="4.1.3.1" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="934" y="456" width="45" height="17" />
    </entry>
  - <entry id="9" name="ec:3.5.1.68" type="enzyme" reaction="rn:R00525" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+3.5.1.68">
      <graphics name="3.5.1.68" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="1014" y="359" width="45" height="17" />
    </entry>
  - <entry id="10" name="ec:3.5.1.56" type="enzyme" reaction="rn:R02509" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+3.5.1.56">
      <graphics name="3.5.1.56" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="1014" y="340" width="45" height="17" />
    </entry>
  - <entry id="11" name="ec:3.5.1.49" type="enzyme" reaction="rn:R00524" link="http://www.genome.jp/dbget-bin/www_bget?
      enzyme+3.5.1.49">
      <graphics name="3.5.1.49" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="1014" y="319" width="45" height="17" />
    </entry>
  - <entry id="12" name="ec:3.5.1.31" type="enzyme" reaction="rn:R00653" link="http://www.genome.jp/dbget-bin/www_bget?
```

Figure 3. KEGG xml file of Map00630: The root element can be preceded by an optional XML declaration. This element states what version of XML is in use (normally 1.0); it may also contain information about character encoding and external dependencies.

(Source: ftp://ftp.genome.jp/pub/kegg/xml/map/map00630.xml)

KEGG pathway, one of the representative pathway databases, adopted an XML representation of ~~the~~ metabolic pathways. KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks [8]. The molecular reaction network is the most unique data object in KEGG, which is stored as a collection of pathway maps in the PATHWAY database. As of KEGG Release 47.0+/07-01 of July 08, 94,068, KEGG pathways are generated from 372 reference pathways. Figure 3 is an example of well-formed XML document of the KEGG Map00500, corresponding to Figure 1. Each object is identified by the KEGG object identifier, consisting of a five-digit number prefixed by an upper-case alphabet, such as C00101 (Chemical compound: 00101), and R00623 (Reaction: 00623), or prefixed by a 2-4 letter code for PATHWAY such as map00010.

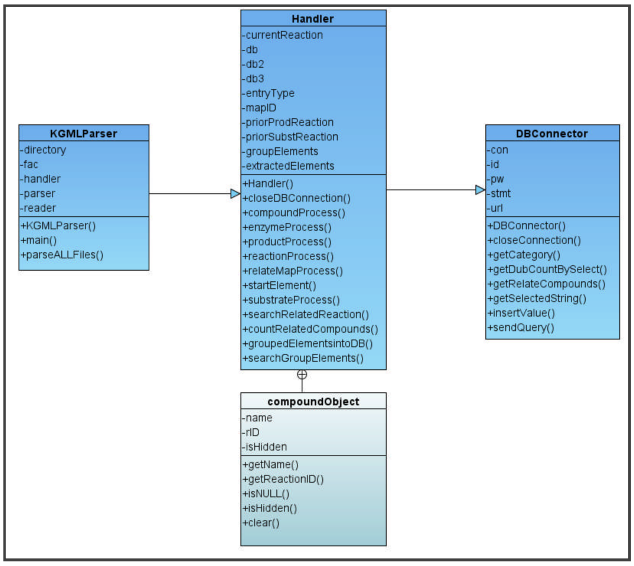


Figure 4. KGML Parser and Its Class Diagram: The system was developed in Eclipse platform, and it was implemented using the Java SDK 1.5, Java.xml package, and the MySQL database, with Tomcat application server for future use.

The first step to deal with an XML document programmatically is to take it and parse it. As the document is parsed, the data in the document becomes available to the application using the parser. The KGML Parser module is dedicated for parsing KEGG Markup Language (KGML), an XML representation of KEGG pathway maps. The parsing module of KEGG XML files has been implemented on Eclipse platform and has been deployed with Sun's J2SE Reference Implementation. Four classes, KGMLParser, Handler, compoundObject, and DBConnector represent the domain model of the system as in Figure 4, based on MVC (Model-View-Controller) model. One of the biggest challenges is to develop automated and tractable techniques to ensure static-type safety and optimize the program, which involves basic reasoning tasks involving complex constructions [13].

3. Parsing Results of Metabolic Pathway XML Documents

Analyzing KEGG XML documents states valuable information that can be applied to automatic layout algorithm aesthetics for global layout of metabolic pathway.

3.1 In/Outgoing Edges of Compounds

Metabolic studies, used to be dedicated to a single pathway, recently began to focus on the entire network [14]. For example, KEGG, the representative biological pathway database now provides the KEGG Metabolism Atlas [15], by manually combining existing metabolic pathway maps. In a global pathway, positioning vertices impact the number of edge crossings in a layout more than in a single pathway. Especially, compounds with a higher degree of in/outgoing edges have greater influence on applications based on automatic graph layout schemes. Therefore, a careful attention in positioning those compounds is needed in order to get a better readable visualization of the graph from a graph-theoretic perspective.

In single metabolic pathways, two compounds are connected on each other with an edge that expresses a single reaction. Each compound could be a substrate or product in several reactions of pathways in metabolism. In regards to drawing a global pathway, highly connected compounds with many reactions would have a great number of ingoing and outgoing edges. We counted the numbers of in/outgoing edges of each compound by searching reactions that the compound took part in the parsing data. In total, 4271 compounds that appear at least once in a single pathway (including 4089 compounds, 180 glycans and 2 drugs) were used in this analysis.

Through this analysis, we obtained a result that 3093 compounds, which takes up 72% of the total compounds, have less than two edges. Among those 3093 compounds, 493 compounds, which take up 15.9% of the whole, do not have a single edge, not related to any other compound. Figure 5 compares the number of compounds according to the range of the number of in/outgoing edges. The compound with the greatest number of edges is C00024, Acetyl-CoA with 55 reactions.

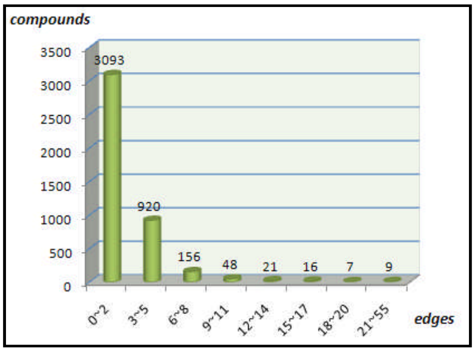


Figure 5. The number of compounds versus the range of the number of edges

Table 1 shows top 17 compounds with the greatest number of edges.

C00024	Acetyl-CoA	55
C00022	Pyruvate	51
C00014	NH3	38

C00025	L-Glutamate	29
C00011	CO2	28
C00037	Glycine	24
C00049	L-Aspartate	24
C00097	L-Cysteine	23
C00048	Glyoxylate	22
C00118	(2R)-2-Hydroxy-3-(phosphonooxy)-propanal	21
C00029	UDP-glucose	20
C00058	Acetyl-CoAFormate	20
C00100	Propanoyl-CoA	19
C00036	Oxaloacetate	18
C00083	Malonyl-CoA	18
C00047	L-Lysine	18
C00219	(5Z,8Z,11Z,14Z)-Icosatetraenoiacid	18

Table 1. Top 17 Compounds with the great number of edges

3.2 Relation Degrees of Compound Pairs

As ~~was~~ previously mentioned in Section 3.1, 72% compounds have less than two edges. To analyze the relations between these compounds, we define the term *relation degree of a compound pair* - the phrase *relation degree of a compound pair* is defined as the number of compounds that appear through existing reactions between two compounds, including the two compounds. For example, in Figure 6, there are 5 compounds connected between two compounds, C01291 and C16471. Those are C16475, C16466, C16468, C16470, C16469, and the numbers of their in/outgoing edge numbers are two, which means that they appear only once in this pathway, and do not appear in any other pathway. In this case, relation degree of the compound pair (C10291, C16471) is 7.

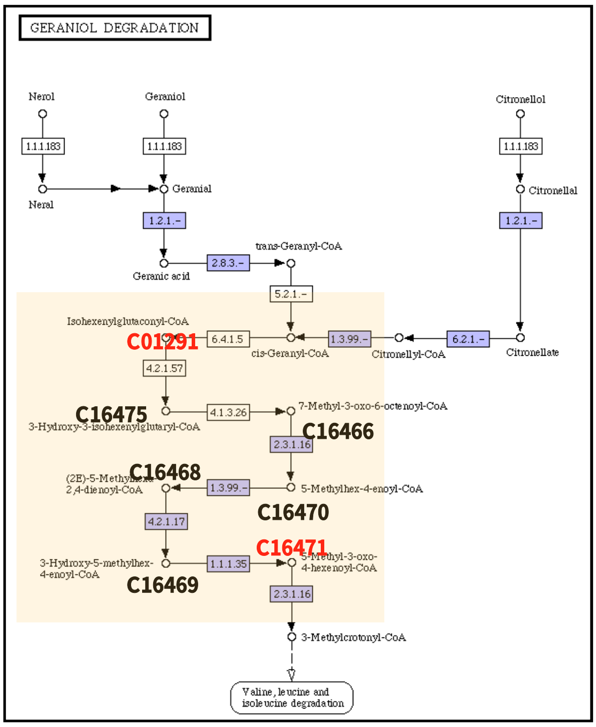


Figure 6. KEGG reference pathway of Geraniol degration in Xenobiotics Biodegradation and Metabolism shows an example of the relation degree of the compound pair (C01291 ,C16471).

While applying our devised algorithm, we obtained a result of relation degrees for 445 pairs of compounds with an average relation degree of 3.77. It means that 3 or 4 compounds, in average, can make 445 separate groups by themselves. Compounds found in these groups are 1246, in total.

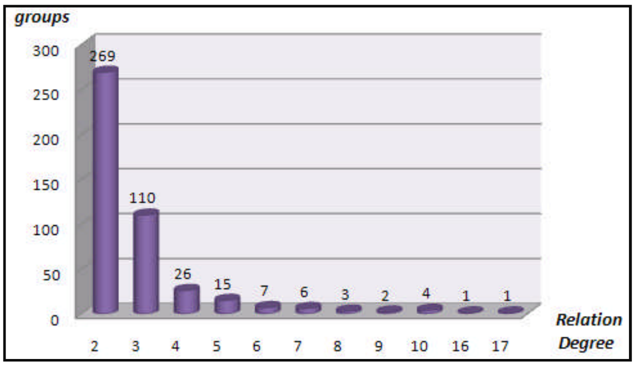


Figure 7. The number of searched groups per each Relation Degree.

Table 2 shows the compound pair (C01226, C16317), which has the highest relation degree of 17 among 480 found pairs.

[redacted]

map00592	***C01226***	17
map00592	C04780	
map00592	C04672	
map00592	C16327	
map00592	C16328	
map00592	C16329	
map00592	C16330	
map00592	C16331	
map00592	C16332	
map00592	C16333	
map00592	C16334	
map00592	C16335	
map00592	C16336	
map00592	C16337	
map00592	C16338	
map00592	C16339	
map00592	***C16317***	

Table 2. The compound pair with the highest relation degree among 480 pairs.

Mostly, each group consists of compounds that exist only in the same pathway.

4. Discussion and Future Work

In our first attempt, we tried to visualize all metabolic pathway information automatically in a single atlas map. However, it only resulted in a confusing diagram which was difficult to interpret. Through this work, we learned the necessity of analyzing the characteristics of metabolites and reactions [16]. However, we, then have focused only on shared and duplicate nodes between two pathways [13]. Therefore, by analyzing parsing results of KEGG XML documents, we could obtain statistical data related to the number of in/outgoing edges of compounds, and relation degrees of compound pairs as we defined in this paper. This research would be valuable to present novel algorithms for multiple pathway maps. Because the number of edge crossings does not increase exponentially according to the increase of the number of nodes, positioning vertices influences the number of edge crossings, which impact the overall graph layout. Hence, understanding characteristics of compounds and reaction from XML document parsing results is worth to take in consideration. This gives an issue of applying different strategies on compounds according to the degree of complexity of edge crossings.

In this paper, biological pathway analysis has been proposed based on a public metabolic database in an XML document parsing result. This gives a preliminary step to provide automatic layout algorithms for multiple pathways, in the future. Without the analysis of the edges of compounds in the pathway map, visualizing all the pathways globally in a single atlas map generally would result in a confusing diagram. A solid algorithm based on this statistic analysis for automatic layout in biological networks will be left as future work.

5. Acknowledgements

We appreciate Prof. Woong-Yang Park for valuable comments.

6. REFERENCES

[1] Kaufmann, M., and Wagner, D., (Eds.) (2001) Drawing Graphs: Methods and Models, LNCS 2025, Springer

[2] Moritz, Y.B. and Isabel, R. (2001). A graph layout algorithm for drawing metabolic pathways, Bioinformatics, 17(5):461-467.

[3] Yuan Wang (2008). Familiar Layouts Generation for Metabolic Pathway Graph Visualization, MS Thesis, Case Western Reserve University

[4] M. Y. Becker and I. Rojas (2001). A Graph Layout Algorithm for Drawing Metabolic Pathways, BIOINFORMATICS, Vol. 17, No. 5, pp.461-467.

[5] Karp, P.D. and Paley, S.M. (1994). Automated drawing of metabolic pathways, Proc. of the 3rd Intl. Conference on Bioinformatics and Genome Res., 225 – 238.

[6] E.H. Song, M.K. Kim, and S.H. Lee (2006). A Metabolic Pathway Drawing Algorithm for Reducing the Number of Edge.

[7] M. Kato et al. (2005). Automatic Drawing of Biological Networks Using Cross Cost and Subcommponent Data, Genome Informatics 16(2):22-31.

[8] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

[9] Moritz Y.Becker and Isabel Rojas. (2000) A graph layout algorithm for drawing metabolic pathways, BIOINFORMATICS, Vol. 17, No 5, pp 461-467.

[10] Bray, Tim; J. Paoli, C. M. Sperberg-McQueen, E. Maler, F. Yergeau (2006). <Extensible Markup Language (XML) 1.0 (Fourth Edition) - Origin and Goals>. World Wide Web Consortium.

[11] "XML Definition" SOA Online. 21 Jul 2008 < http://searchsoa.techtarget.com/sDefinition/0,,sid26_gci213404,00.html>

[12] H.C. Purchase, J. Allder, D.Carrington.(2002) Graph LayoutAesthetics in UML Diagrams : User Preference, Journal of Graph Algorithms and Applications, vol.6, no.3, pp. 255-279.

[13] S.H Kang, M.H Jang, J.Y Whang, and H.S Park(2008). Parsing KEGG XML Files to Find Shared and Duplicate Compounds Contained in Metabolic Pathway Maps: A Graph-Theoretical Perspective, Genomics & Informatics, vol. 6, no. 3 (in press)

[14] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabasi (2000). The Large-scale Organization of Metabolic Networks, NATURE, Vol. 407, pp.651-654.

[15] Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways, Nucleic Acids Res. May 13.

[16] E.H. Song, S.I. Ham, S.D. Yang, A. Rhie, H.S. Park, and S.H. Lee (2008). J2pathway: A Global Metabolic Pathway Viewer with Node Abstracting Features, Genomics & Informatics, Vol. 6, No. 2, 118-124.