Journal of Computer Science 6 (11): 1276-1282, 2010
ISSN 1549-3636

Knowledge Discovery in Biochemical Pathways Using Minepathways

Ford Lumban Gaol
Faculty of Computer Science, Bina Nusantara University, Indonesia

Abstract: Problem statement: The advancement of the biochemical research gives profound effect to the collection of biochemical data. **Approach:** In the recent years, data and networks in biochemical pathways are abundant that allow to do process mining in order to obtain useful information. By using graph theory as a tool to model these interactions, it can be formally find the solution. **Results:** The core of the problem of mining patterns is a subgraph isomorphism which until now has been in the NP-class problems. Early identification showed that in the context biochemical pathways has unique node labeling that result simplifying pattern mining problem radically. **Conclusion:** Process will be more efficient because the end result that is needed is maximum pattern that could reduce redundant patterns. The algorithm that used is a modification of the maximum item set patterns that are empirically most efficiently at this time.

Key words: Biochemical pathways, graph theory, subgraph isomorphism, NP problems, maximum itemset pattern

INTRODUCTION

The advance developments in biochemical pathways have impact to the increasing availability of molecular data that allows analyzing the connectivity and interaction between biochemical pathways (Hartwell *et al*., 1999; Oltvai and Barabasi, 2002; Inokuchi *et al*., 2001; Cook and Holder, 2000; Rives and Galitski, 2003). The fast escalation of molecular data was suspected by successfully developed of BLAST and CLUSTAL (Altschul *et al*., 1997; Thompson *et al*., 1994) that have contributed to the availability of data in various processes biochemical pathways.

Biochemical pathways interaction data, in general, lead to biological or cellular networks (Krishnamurthy *et al*., 2003). They are often abstracted using graph modeling. Despite the availability of adequate data, the analysis of the efficiency of data generated by BLAST and CLUSTAL are not yet available.

The complexity of biochemical pathways can be understood by using a variety of concepts in graph theory. For example, biochemical pathways, is a model which uses hyperGraph, where the vertices are expressed as molecules and hyperedge expressed enzymes (reaction). It is possible to reduce the model into a more general connected graph with nodes as enzymes and the edge directed from an enzyme to the other states as the process of being consumed by the first enzyme reaction catalyzed by the others (Hartwell *et al*., 1999; Oltvai and Barabasi, 2002).

Two main issues in the graph that arise in the context of biochemical pathways datasets are aligning multiple graphs and find the frequent subgraph from the collection of graphs. Analysis of biochemical pathways in the context of both these problems will respond to various issues including the analysis of differences between the biochemical pathways structure among different organisms, as well as patterns of gene regulation (Akutsu *et al*., 1998; Olken, 2003; Ito *et al*., 2001; Ho *et al*., 2002).

In this study, we formulate the problem of frequent pattern that discovered from a collection of graphs is called biochemical pathways mining.

Frequent pattern mining is still an open problem because the core of graph mining is subgraph isomorphism problem which is classified as NP-hard class. However, modeling of biochemical pathways network can simplify the problem of graph mining. In the context of biochemical pathways that has a unique vertex labeling. This has resulted simplification pattern mining problem significantly. Algorithm for the extraction process to find the frequent pattern among biochemical pathways datasets using the model that proposed by (Karp and Mavrovouniotis, 1994; Gaol and Widjaja, 2008).

In this study, we will discuss the use of graph theory for the formalization of biochemical pathways. It will be discussed modeling to biochemical pathways and how to analyze it. In the discussion, will be discussed in depth the results of observation. The final part, will be concluded and the various aspects of advanced research that can be done.

MATERIALS AND METHODS

Approach: Graph mining is a very challenging issue because it deals with issues isomorphisma subgraph which is a NP-hard. Graph mining is more exciting because a lot of modeling graphs appear in many state of the art applications, industrial and scientific. Existence of graph mining algorithms in general is based on frequent pattern mining, which had already appeared in much literature (Patel *et al*., 2005; Razali and Ali, 2009). Definition and complication of the problem of graph mining is significantly dependent on the target application.

For example, one class of algorithms defined the issue a problem to find the sub-pattern isomorphism (not dependent labeling) in the database of graphs or a graph is large (Mamitsuka *et al*., 2003; Goto *et al*., 1997; Tohsato *et al*., 2000).

This approach is suited to applications that intend to focus on the relationship between entities. However, this contributes to the large computing time because the root of the problem of subgraph isomorphism is NP-problem that must solve in every step of the algorithm. As a result of research in graph mining is focused on sorting the node and several optimization techniques that simplify the problem of subgraph isomorphism (Kuramochi and Karypis, 2001; Mamitsuka *et al*., 2003).

A discussion of mining in the context graph will begin with the following definition (Olken, 2003).

Definition 1: Biochemical pathways of P (M, Z, R) is a collection of metabolites M, Z enzymes and reactions R, where each $r \in R$ reactions are enzymes associated with $Z(r) \subseteq Z$, a set of substrates $S(r) \subseteq M$ and a set of products $T(r) \subseteq M$.

The goal of mining is to find the biochemical pathways of certain patterns from the interaction of enzymes related to one another. The modeling of biochemical pathways with a simple digraph will be modeled the interaction of information efficiently. Each enzyme is symbolized by a unique vertex in a collection of pathway. This effectively simplifies the problem of graph mining significantly. Simplification process is certainly does not omit information but increasingly facilitate the process of information extraction. This will be increasingly facilitating the process of information extraction.

Definition 2: Biochemical pathways of P(M,Z,R), graph G(V, E) P can be constructed with the following: for enzyme $z_i \in Z$, there is node $v_i \in V$. There is an edge from v_i to v_j, i.e $(v_i, v_j) \in E$. Iff $\exists r_1, r_2 \in R$, such that $z_i \in Z(r_1)$, $z_j \in Z(r_2)$ and $T(r_1) \cap S(r_2) \neq \varnothing$.

Biochemical pathways can be modeled using graph modeling. There is a connation from one enzyme to other enzymes in the two enzymes If and only if graph part is of a product from one to another. A biochemical pathway was illustrated in Fig. 1. In the pathways, enzymes expressed as square, oval while metabolites otherwise. Each vertex represents a single enzyme in the oval-shaped graphs. The direction of the edges can be ignored, but shows the interaction between the enzymes.

Algorithm for mining biochemical pathways: Biochemical pathways for the mining algorithm based on (Gaol and Widjaja, 2008) using graph modeling in the previous section which facilitates the process of mining frequent pattern.

Definition 3: Given collection of graphs, $G_1, G_2, \ldots, G_n$ and support threshold ε, Maximum Frequent subgraph search process is the process of finding all maximum connected subgraph containing at least εn from the input graph.

This definition states that the support of a subgraph contained in n' a collection of graphs is $\frac{n'}{n}$.

A subgraph is frequent if its support is greater than or equal $\frac{n'}{n}$. The connectivity requires frequent interaction that must relate to one another. Formally, a graph is connected if there is at least one path that connecting any two vertex in a graph. The use of the connectivity is to reduce maximum subgraph that is redundant. Frequent subgraph is maximum frequent if not contained in another subgraph.

Gaol and Widjaja (2008) mentioned that although the graph mining is a difficult issue but by using the framework above, has simplified the problem. It was due to the above model that has a unique labeling of nodes resulting edges are also unique.

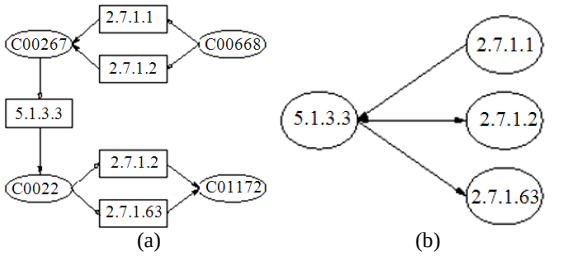


Fig. 1: Digraph modeling to biochemical pathways (a) part of the glycolysis pathway and (b) representation of the digraph

It makes possible to reduce problem to frequent itemset mining using edges as data. To frequent itemset mining algorithms will be selected among the existing algorithms.

Because vertices labels are unique, then every edge that there are uniquely expressed by the labels of neighboring nodes to each other. This study leads to the idea to declare a connected subgraph with the set of edges, because of the uniqueness of each edge lead to the uniqueness of the subgraph that represent by the collection of edges.

Definition 4: An edge e which is a collection of labels of two vertices v_i, v_j. A set of unique edges ES = (e_1, e_2,. . . , E_k) is called the set of connected edge (edge sets connected) iff all the edges in these sets are connected, ie every subset of ES '⊆ES together using at least one node left in the edges ES\ES.

In the context of frequent motifs mining, a transaction is a set of motifs. In data mining we can made relationship between maximum frequent pattern can set up with frequent motifs mining, in which pathways relate with the relationship and a collection of linked edge corresponding motifs. The problem is how to find all frequent motifs contained in transactions that meet the threshold of support.

The basic idea of mining frequent item sets will be adapted for subgraph mining with a focus only for the connected set of edge that will pass through the search process. Besides it is necessary for elimination of redundancy in the sense that the same set of edges over more than one order. Subgraph mining is a modified algorithm based on (Gouda and Zaki, 2001). This algorithm in accordance with experimental results for the maximum pattern mining algorithm with the computational time is the most efficient compared to other algorithms maximum itemset pattern. This is due to its depth first backtracking enumeration based principles, which expand each subgraph with only the edges of the candidate set of edges. To ensure connectivity, the addition of edges to the subgraph is connected and to avoid redundancy by keeping track that you have visited.

Based on the research of Mamitsuka *et al*. (2003), one of the reasons for selecting the depth-first procedure is a limitation of memory size becomes very problematic when the database graph bigger. Algorithm for frequent pattern mining is expressed in Fig. 2, developed by (Mamitsuka *et al*., 2003) that the most efficient experimentally compared with other algorithms.

The algorithm will expand the edge sets of all edges in a collection of candidates one by one. If collection of an expanded edge is frequent, then the procedure will be called back. The algorithm terminate when an edge sets cannot be *e* again. In the algorithm, C (e_i) of the edge e_i neighboring states. Figure 3 is the example of execution for Algorithm Depth-first enumeration for mining frequent subgraph.

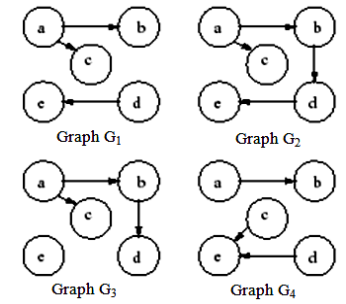
D is a set of edges that have been visited by the algorithm. Mining Path ways procedures (MFS, (e_i), C(e_i), (e_1, e_2, ..., e_{i-1})), for each edge e_i are frequent in the collection of graphs.

Procedure Mining Path ways MFS, E_k, C_k, D):

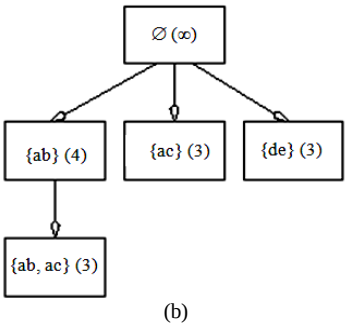
MFS: Collection of maximum frequent subgraph
E_k: Frequent subgraph with k edges

```
Maximum← true
For all of edges e_i ∈ C_k do
    D ← D ∪ {e_i}
    E_k+1 ← E_k ∪ {e_i}
    If E_k+1 are frequent then
        ismaximum← false
        C_k+1 ← (C_k ∪ N (e_i))\D
        MinePathways (MFS, E_k+1, C_k+1, D)
    If is maximumthen
        If E_k do not have superset in MFS then
            MFS ← MFS ∪ E_k.
```

Fig. 2: Algorithm Depth-first enumeration for mining frequent subgraph



(a)



(b)

Fig. 3: Example of execution from mining frequent subgraph. (a) Input from graph collection; (b) result from tree enumeration of frequent edges

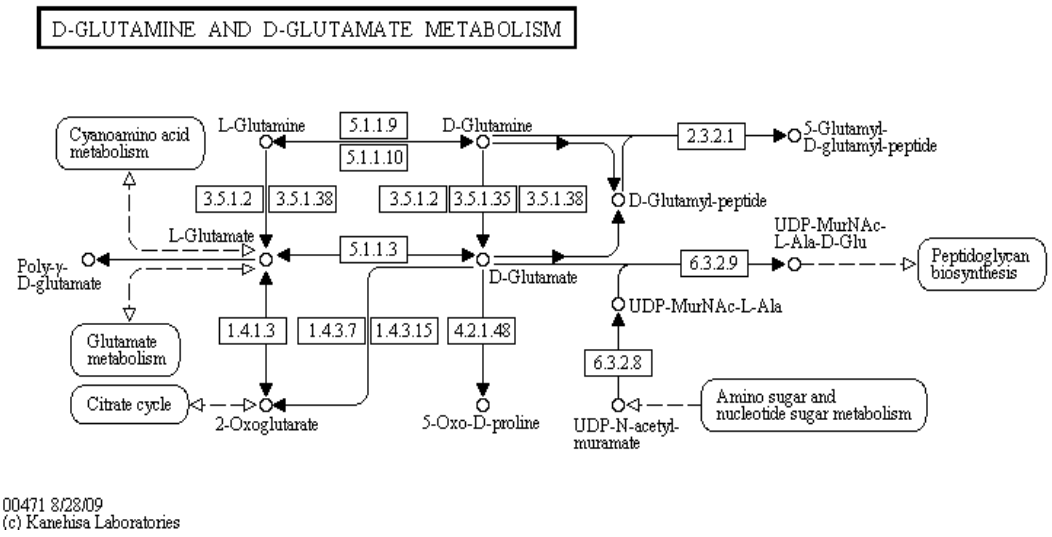


Fig. 4: Mining pathways that are repeated for various values of support threshold in glutamine among 255 organisms

C_k: Collection from candidate of edges
D: Collection of visited edges

RESULTS AND DISCUSSION

Experiment: By using the algorithm MiningPathways, mining process of collection, which is extracted from Biochemical pathway databases KEGG. Current KEGG biochemical pathways have a complete database which is also a basic reference path ways that can be displayed as networks enzyme formed manually. Pathway-pathway organisms formed automatically with the help enzyme genes identification. At the end of the year 2009, KEGG contains pathway maps from several processes including carbohydrate metabolism, energy, lipid, nucleotide and amino acid metabolism to 257 organisms.

Mining was conducted in several pathways that are part of metabolism for several different organisms. Samples from the sub-sub-pathways are repeated in this collection. They are part of the pathway that metabolizes glutamate metabolism, alanine-aspartate and pyrimidine as stated in Fig. 4 s/d Fig. 6 respectively. Base on KEGG ID, vertices in graphs are labeled by an enzyme, which can be queried from the KEGG website for more detailed information.

We successfully found some repeated sub-pathways. For example, a pathway of metabolism glutamine containing 6-vertices and eight edges that appears in 50 from 255 organisms. Sub-pathway that is represented by the node-edge thickness and edge in Fig. 4. Consisting of enzyme-enzyme, Citrate cycle, Cyanoamino acid metabolism, amino sugar and nucleotide sugar, Peptidoglycan biosynthesis.

We will try with some different support threshold in various metabolic pathways. For example, when the support threshold was reduced to 2.99% (35 organisms) for the metabolism of glutamate, the largest a sub-pathways that can be found consisting of 3 vertices and 12 edges. Sub-pathway was expressed by vertex thickness and edges. As stated in the figure, this pathway contains Citrate cycle, which is also related to another enzyme with L-glutamine. When the reduction of support threshold to 3.97% (35 organisms), in this research managed to find a sub-pathway of 6-vertices and 13 edges, which is expressed in the whole graph in the picture. Loop for Cyanoamino acid implies that these enzymes involved in two successive reactions, which was part of a sub-pathway-subpathway repeats.

In Fig. 5, the biggest pathways that is found in metabolism Metabolism of Terpenoids and Polyketides for the three levels of support threshold is different. The bold sub-pathway of 5-vertices and eight edges occurred in 50 of 196 organisms (35.1%), thick sections with a 5 vertices and 11 edges occurred in 30 organisms (17.4%) and overall graph of six-vertices and 16 edges occurred in 18 organism (11.5%). Note that Biosynthesis of type II polyketide backbone and Tetracyclines. The interaction with the Cyanoamino, expressed by the dashed line in Fig. 5, included in a repeated sub-pathway of the metabolism of alanine-aspartate but not including the sub-sub-pathways that is greater than the smaller frequency, which is an important result to be written.

Figure 6 show that the analysis of various levels from the threshold of sub-recurring pathway for metabolism Tetracycline biosynthesis. There are beta-Alanine biosynthesis, cytosine, Uridine, glutamine, Pyrimidine, of 4-vertek vertek and five edge-edge occurred in 43 of 276 organisms (23.6%), thick sections with a 5- vertices and 7 edges of the organisms occurred in 34 (21.8%) and overall from the graph of 7-vertices and 13 edges occurred in 23 of organisms (15.4%).

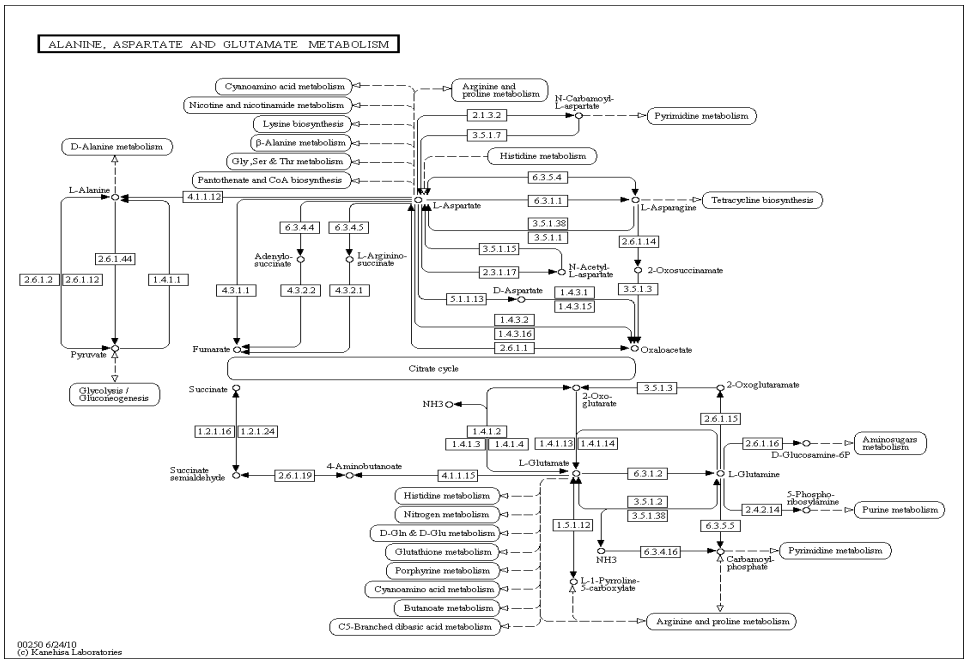


Fig. 5: The discovery of sub-pathway that repeated for a variety of different support values on alanine-aspartate metabolism among 157 organisms

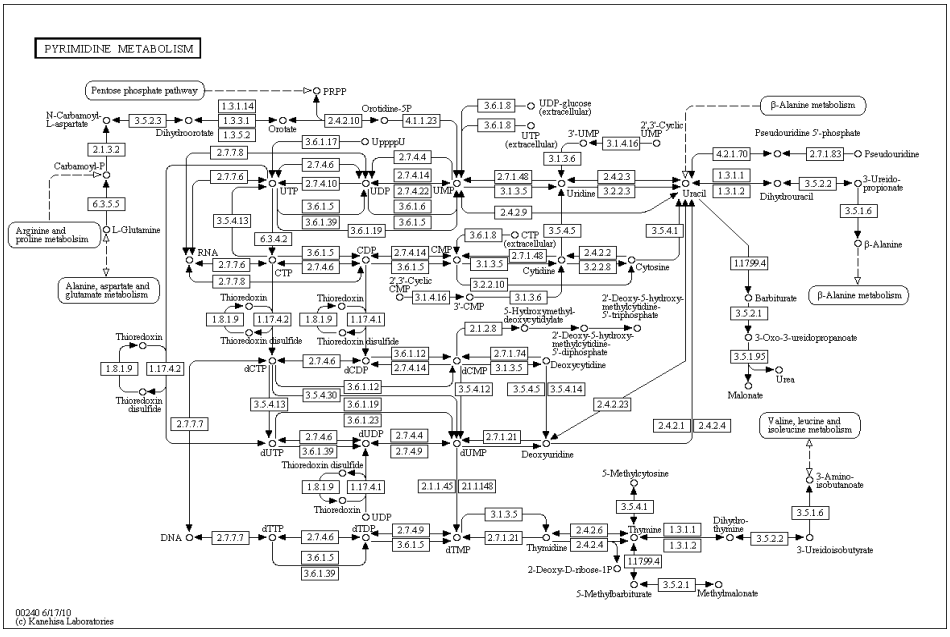


Fig. 6: The discovery of pathway-sub-sub-recurring pathway for a variety of different support values on Tetracycline biosynthesis metabolism among 156 organisms

Table 1: Time used in mining various different biochemical pathway for various different minimum support

Metabolism	Min support (%)	No. of subpathway frequent	No. of most of edges	Runtime (sec)
Glutamate				
	10.0	34	15	0.52
	12.5	39	13	0.17
	15.0	21	11	0.03
	20.0	12	9	0.00
	10.0	120	15	0.44
Metabolism of Terpenoids and Polyketides				
	12.5	78	15	0.19
	15.0	49	12	0.04
	20.0	23	7	0.00
Tetracycline biosynthesis				
	10.0	34	16	3.08
	12.5	25	16	1.84
	15.0	21	12	0.15
	20.0	15	11	0.02

In Table 1 the report base on results from mining collections biochemical pathway some minimum supports. The report is based on number of repeated maximum path, the number of edges in the discovery of path-the biggest path and time in seconds for executions of the metabolism. Metabolism of Terpenoids and Polyketides collection has a total of 2804 vertices and 11339 edges from 155 organisms. Collections of Tetracycline biosynthesis pathway have 2681 vertices and 8481 edges from collection of 156 organisms.

By using a Pentium Core Duo with 2 GB of memory can be mined collections of less than a second pathway to support the value of a relatively high threshold to obtain meaningful results in the sense that the size of the pathways repeatedly found. For lower values of support, many sub-sub-pathways that became recurrent and size of the pathways increased significantly. For this reason, need more time to produce all the pathway are frequent. We conclude that with the increasing of the pathways that will give an exponential effect with the time computation. The exponential effect with the time computation will give more challenging for finding the knowledge inside the Biochemical Pathways www.kegg.com.

CONCLUSION

With the increasing amount of data and interactions of bio-molecular networks, that will affect the problem of mining patterns, motifs and modules within the network biology become very interesting. This study provides a framework for mining is the process of biological networks using graph model that emphasizes the efficiency of computation time by using the method (Gaol and Widjaja, 2008) which has a highly efficient computing time to the present. For further research will be reviewed preprocessing techniques to data from www.kegg.com and empirical testing uses a model algorithm.

REFERENCES

Akutsu, T., S. Kuhara, O. Maruyama and S. Miyano, 1998. Identification of gene regulatory networks by strategic gene disruptions and gene over expressions. Proceeding of the 9th Annual ACMSIAM Symposium on Discrete Algorithms, Jan. 25-27, Society for Industrial and Applied Mathematics, San Francisco, California, United States, pp: 695-702. http://portal.acm.org/citation.cfm?id=314613.315050

Altschul, S.F., T.L. Madden, A.A. Scheffer, J. Zhang and Z. Zhang *et al*., 1997. Gapped BLAST and PSIBLAST: A new generation of protein database search programs. Nucleic Acids Res., 25: 3389-3402. DOI: 10.1093/nar/25.17.3389

Cook, D.J. and L.B. Holder, 2000. Graph-based data mining. IEEE Intell. Syst., 15: 32-41. DOI: 10.1109/5254.850825

Gaol, F.L. and B. Widjaja, 2008. Semistructured mining and its application. Int. J. Comput. Inform. Syst., 4: 97-109.

Goto, S., H. Bono, H. Ogata, W. Fujibuchi and T. Nishioka *et al*., 1997. Organizing and computing metabolic pathway data in terms of binary relations. Pac. Symp. Biocomput., 1: 175-186. PMID: 9390290

Gouda, K. and M.J. Zaki, 2001. Efficiently mining maximum frequent itemsets. Proceeding of the IEEE International Conference on Data Mining, Nov. 29-Dec. 2, IEEE Computer Society, Washington DC., USA., pp: 163-170. http://portal.acm.org/citation.cfm?id=658047

Hartwell, L.H., J.J. Hopfield, S. Leibler and A.W. Murray, 1999. From molecular to modular cell biology. Nature, 402: C47-C51. DOI: 10.1038/35011540

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader and L. Moore *et al*., 2000. Systematic identification of protein complexes in *Saccharomyces cerevisae* by mass spectrometry. Nature, 415: 180-183. DOI: 10.1038/415180a

Inokuchi, A., T. Washio, T. Okada and H. Motoda, 2001. Applying the a priori-based Graph mining method to mutagenesis data analysis. J. Comput. Aided Chem., 2: 87-92.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida and M. Hattori *et al*., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci., 98: 4569-4574.PMID: 11283351

Karp, P.D. and M.L. Mavrovouniotis, 1994. Representing, analyzing and synthesizing biochemical pathways. IEEE Expert, 9: 11-21. DOI: 10.1109/64.294129

Krishnamurthy, L., J. Nadeau, G. Ozsoyocglu, M. Ozsoyocglu and G. Schaeffer *et al*., 2003. Pathways database system: An integrated system for biological pathways. Bioinformatics, 19: 930-937. DOI: 10.1093/bioinformatics/btg113

Kuramochi, M. and G. Karypis, 2001. Frequent subgraph discovery. Proceeding of the IEEE International Conference on Data Mining, Nov. 29-Dec. 2, IEEE Computer Society, Washington DC., USA., pp: 313-320. http://portal.acm.org/citation.cfm?id=658027

Mamitsuka, H., Y. Okuno and A. Yamaguchi, 2003. Mining biologically active patterns in metabolic pathways using microarray expression profiles. ACM SIGKDD Explorat. Newslette, 5: 113-121. DOI: 10.1145/980972.980986

Olken, F., 2003. Biopathways and Protein Interaction Databases. A Lecturer in Bioinformatics Tools for Comparative Genomics: A Short Course. 1st Edn., Lawrence Berkeley National Laboratory, Berkeley, CA., pp: 125.

Oltvai, Z.N. and A.L. Barabasi, 2002. Life's complexity pyramid. Science, 298: 763-764. DOI: 10.1126/science.1078563

Patel, R., S.S. Rana and K.R. Pardasani, 2005. Model for load balancing on processors in parallel mining of frequent itemsets. Am. J. Applied Sci., 2: 926-931. http://www.scipub.org/fulltext/ajas/ajas25926-931.pdf

Razali, A.M. and S. Ali, 2009. Generating treatment plan in medicine: A data mining approach. Am. J. Applied Sci., 6: 345-351. http://www.scipub.org/fulltext/ajas/ajas62345-351.pdf

Rives, A.W. and T. Galitski, 2003. Modular organization of cellular networks. Proc. Natl. Acad. Sci., 100: 1128-1133. DOI: 10.1073/pnas.0237338100

Thompson, J.D., D.G. Higgins and T.J. Gibson, 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22: 4673-4680. PMID: 7984417

Tohsato, Y., H. Matsuda and A. Hashimoto, 2000. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. Proceeding of the 8th International Conference Intelligent Systems for Molecular Biology, Aug. 19- 23, AAAI Press, La Jolla, CA., pp: 376-383.