Mining Biological Network and genomes-A Systematic Review

Rohini M.
Assistant Professor,
Dept. of Computer Engineering,
Sri Krishna College of Engineering and Technology,
Coimbatore

D. Surendran, PhD
Professor,
Dept. of Computer Engineering,
Sri Krishna College of Engineering and Technology,
Coimbatore

ABSTRACT
The research community is inundated with data such as the genome sequences of various organisms, microarray data and so on, of biological origin. This data-volume is rapidly increasing and the process of understanding the data is lagging behind the process of acquiring it. The sheer enormity calls for a systematic approach to understanding this using computational method. The rapid progress of biotechnology and bio-data analysis methods has led to the emergence and fast growth of a promising new field: bioinformatics. It is a field having a tremendous amount of bio-data which needs in-depth analysis. Bio-data is available as, Nucleotide sequences (DNA and RNA sequences), Protein sequences, Genomes and structures in the form of Biological networks (metabolic pathways, gene regulatory network, and protein interaction network).

A framework to discover frequent patterns and modules from biological networks is presented. From the study of different Biological networks, it can be concluded that the best way to analyze and extract the information (frequent functional module) from the biological network is through graph mining since these networks can be modeled into different types of graphs according to the information needs to be extracted. But this graph-based mining approach often leads to the computationally hard problem due to their relation with subgraph isomorphism. Graph simplification technique is used that is suitable to biological networks, which makes the graph mining problem computationally tractable and scalable to large numbers of networks. So the detection of frequently occurring patterns and modules will be a computationally simpler task since the reduction in the effective graph size significantly.

Keywords
Data mining, Biological networks, graph mining, metabolic pathways.

1. INTRODUCTION
Molecular interaction data plays an important role in understanding biological processes at a modular level by providing a framework for understanding a cellular organization, functional hierarchy, and evolutionary conservation. As the quality and quantity of network and interaction data increase rapidly, the problem of effectively analyzing this data becomes significant.

The recent development of high-throughput technologies provides a range of opportunities to systematically characterize diverse types of biological networks. "Network Biology" has been an emerging field in biology. The variety of biological networks can be classified into two categories: (1) Physical networks, which represent physical interactions among molecules, e.g., protein-interaction, protein-DNA interaction, and metabolic pathways; and (2) Conceptual networks, which represent functional associations of molecules derived from genomic data, e.g., co-expression relationships extracted from microarray data, and genetic interactions obtained from synthetic lethality experiments. This large amount of data in the form of a biological network provides us the valuable information to study the functions and the dynamics of biological systems.

Due to the noisy nature of high throughput data, a significant number of spurious edges exist in biological networks, which may lead to the discovery of false patterns. Since biological modules are expected to be active across multiple conditions, it can be easily filtered out spurious edges by mining frequent patterns in multiple biological networks simultaneously. A straightforward approach is to aggregate these networks together and identify dense subgraphs in the aggregated graph. However, it could result in false dense subgraphs that may not occur frequently in the original networks. Figure 1 illustrates such an example with six graphs. If these graphs are added together to construct a summary graph, a dense subgraph comprising vertices a, b, c, and d is derived. Unfortunately, this subgraph is neither dense nor frequent in the original graphs.

A potential solution to the false pattern problem is mining frequent subgraphs directly. A subgraph is frequent if it occurs multiple times in a set of graphs. Frequent subgraph discovery, in general, is considered a hard problem. However, biological networks can often be modeled as a special class of graph where each gene occurs once and only once in a graph. That means, graph has distinct node labels, and there is no "subgraph isomorphism problem" which is NP-hard and so far constitutes the bottleneck of subgraph frequency counting.

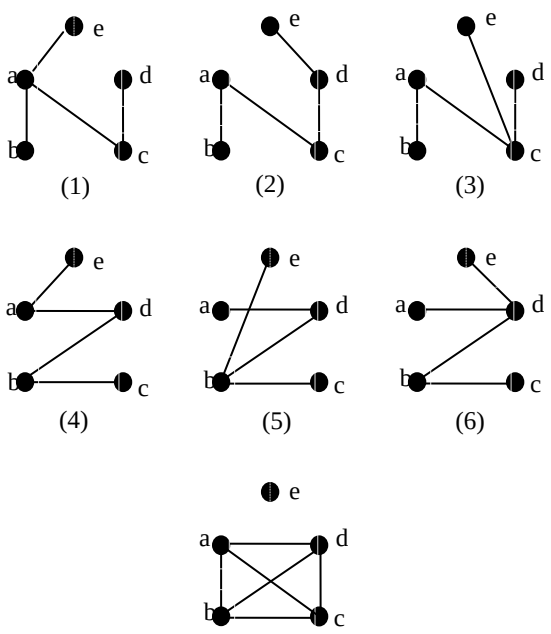

Figure 1: Given six graphs with the same vertex set but different edges sets, summary graph is constructed by adding these six graphs together and by deleting edges that occur less than three times in the graphs. The dense subgraph in the summary graph {a, b, c, d} does not occur in any original graph.

1.1 Problems with Graph Mining

Most graph mining algorithms in the literature are based on the well-studied association rule mining, or more generally, the frequent itemset problem. This problem can be defined as follows. Given a set of items S= {i1, i2, ….., in} and a set of transactions T = { T1, T2, ….., Tm} over S, i.e. Ti $\subset$ S for all i, find all subsets t of S such that $\sigma(t) = |\{Ti \in T: t \subset Ti\}| / |T| \geq \sigma^*$. Here, $\sigma(t)$ is the support of an itemset t and σ^* is the prescribed threshold on support, signifying the desired frequency of patterns to be mined. Frequent itemset mining algorithms are generally based on the lattice or downward closure property of support. This property states that an itemset mining algorithms enumerate all potentially frequent itemsets by effectively pruning the search space.

So in terms of graph mining, downward closure translates to the fact that a subgraph is frequent only if all of its subgraphs are frequent. But most existing graph mining algorithms generalize, frequent itemset mining algorithm to structured data. However, this generalization poses significant challenges for the following reasons:

1.1.1 Subgraph Isomorphism

While counting frequencies of the subgraph in the graph database, one must verify whether a given structure is a subgraph of a graph in the database. This requires the solution of the NP-complete subgraph isomorphism problem at all explored point of the solution space.

1.1.2 Canonical Labeling

Frequent itemset mining algorithms dictate a lexicographic order on items and represent itemsets as ordered sets to ensure that no itemset is considered more than once. However, such an ordering of nodes and /or edges in graphs is not trivial and computing canonical labels for graphs to sort them in a unique and deterministic manner is equivalent to a testing isomorphism between graphs.

1.1.3 Connectivity

While taking advantage of the downward closure property in frequent itemset mining, candidate is generated in a bottom-up fashion by extending itemsets with the addition of items one by one. In the case of graph mining, an extension of subgraphs is not trivial since it is necessary to maintain the connectivity of candidate subgraphs, since the target frequent pattern is desired to be connected, in general.

2. BIOLOGICAL NETWORKS

In a multi-layered organization of living organisms, cellular interactions form the bridge between individual molecules (e.g., genes, mRNA, proteins and metabolites) and large-scale organization of the cell through functional modules. Common abstraction for cellular interaction includes protein interaction networks, gene regulatory networks, metabolic pathways, and signaling pathways.

2.1. Protein Interaction Networks

Protein interaction networks are comprised of groups of interacting proteins. These networks provide the experimental basis for understanding the modular organization of cells, as well as useful information for predicting the biological function of individual proteins. Recently, there have been several efforts aimed at organizing protein interaction networks into databases such as BIND and DIP. This experimental data reveals either pairwise interactions, as in two-hybrid experiments, or multi-way interactions between a set of proteins, as in mass spectrometry experiments. Pairwise interactions are conveniently modeled by simple undirected graphs in which nodes represent proteins and an edge between two nodes represents the interaction between the corresponding proteins. Multi-way interactions are modeled using hyperedges that represent interactions between various proteins in a hypergraph.

2.2 Gene Regulatory Networks

Gene regulatory networks, also referred to as genetic networks, represent regulatory interactions between pairs of genes and are generally inferred from gene expression data through microarray experiments. A simple and frequently used mathematical model for gene regulatory networks is a Boolean network model. In this model, nodes correspond to genes and a directed edge from one gene to the other represents the regulatory effect of the first gene on the second. The edge is labeled by either a + or - sign to represent up or down-regulation, respectively.

2.3 Metabolic Pathways

Metabolic pathways characterize the process of chemical reaction that, together, performs a particular metabolic function. Metabolic Pathways are chains of reactions linked to each other by chemical compounds (metabolites) through product-substrate relationships. Metabolic pathways are thus network of biochemical reactions transform one or several substrates (metabolites) into one or more products (metabolites, as well). A natural mathematical model for metabolic pathways is a directed hypergraph in which each node corresponds to a compound, and each hypergraph corresponds to a reaction (or equivalently enzyme). The direction of a pin of a hyperedge indicates whether the compound is a substrate or a product of the reaction.

From the study of different Biological networks, it can be concluded that the best way to analyze and exact the information (frequent functional module) from biological

network is through graph mining since these networks can be modeled into different types of graphs according to the information to be extracted.

3. PROBLEM STATEMENT

A framework is developed to discover frequent patterns and modules from biological networks efficiently. Modeling of Biological network can be done with the Graphs (directed or undirected), that should be capable of capturing all the required information uniquely and efficiently. Some of the interesting biological networks are metabolic pathways, gene regulatory network and protein interaction network. For example metabolic pathways can be formally defined as:

Definition 1: A metabolic pathway P(M, Z, R) is a collection of metabolites M, enzymes Z, and reactions R, where each reaction r Є R is associated with a set of enzymes Z(r) $\subseteq$ Z, a set of substrates S(r) $\subseteq$ M, and a set of products T(r) $\subseteq$ M.

An appropriate method is required to transform it into graphs uniquely and efficiently. After transforming the biological network in the form of graphs its found to have the frequent subgraph to find the frequent functional modules, for that algorithm is to be designed to find frequent subgraph without leading it to graph isomorphism (NP complete) problem and should be scalable to large number of network.

3.1 Frequent subgraph discovery problem

Definition 2: Given a collection of graph G1, G2… Gn and support threshold ε, the Maximal Frequent Subgraph Discovery problem is one of finding all maximal connected subgraphs that are contained in at least ε n of the input graphs. For example, consider the graph collection of figure 1. it has five edges in all, ab, ac, bd, ce and de. Here the edge set {ab, ac} is maximal frequent subgraph as support threshold value is 3.

4. SOLUTION METHODOLOGY

The Biological network is selected to discover frequent pattern is metabolic pathways, as the dataset is easily available, provided by Kyoto Encyclopedia of genes & Genomes (KEGG) [20].

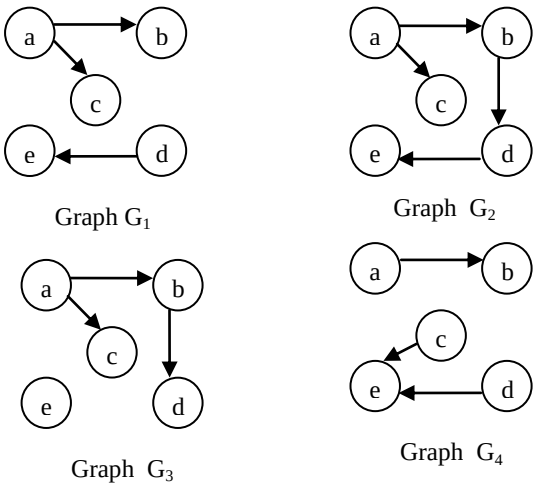

Figure 2: Example for frequent subgraph mining

The KEGG pathway maps are graphical image maps representing networks of interacting molecules responsible for specific cellular functions. KEGG has provided the metabolic pathways maps in the form of xml format (KGML). In KGML the pathway element specifies one graph object with the entry elements as its nodes and the relation and reaction elements as its edges. The relation and reaction elements indicate the connection patterns of rectangles (gene products) and the connection patterns of circles (chemical compounds), respectively, in the KEGG pathways.

4.1 Modeling of metabolic pathways in the form of graph

A natural mathematical model for metabolic pathways leads to a hypergraph as discussed above, and will lead to the problem of graph isomorphism, canonical labeling and connectivity. Since the main goal in mining metabolic pathways is to discover common motifs of enzymes interaction that are related to each other, it is possible to replace this hypergraph by a simpler directed graphs, that are capable of capturing the interaction information efficiently. It further simplifies the graph mining problem specially for biological network, by representing each enzyme by a unique node, independent of the number of times the enzyme appears in the underlying pathway.

The approach for graph formalism is to draw the directed edge from one enzyme to another in the graph if and only if the second enzyme consumes a product of the first one. Figure 3 illustrate the directed graph model for metabolic pathways. In the pathway, enzymes are shown by rectangular boxes while metabolites are shown by ovals. Nodes, each corresponding to exactly one enzyme, are shown by ovals in the graph. In such a model, enzymes correspond to nodes of the graph and a directed edge from one enzyme to another indicates that a product of the first enzyme is a substrate of the second.

An enzyme may show up more than once in the same pathway, implying that this enzyme takes part in the whole process at different time instants. The implication of this fact is that more than one node in the graph (pathway) might have the same label (enzyme). One might either be interested in preserving these temporal relationships or only in general relationships between pairs of enzymes. In the latter scenario, one may merge nodes in the graph with identical labels. By merging nodes with identical labels, simplifies graph analysis problem substantially.

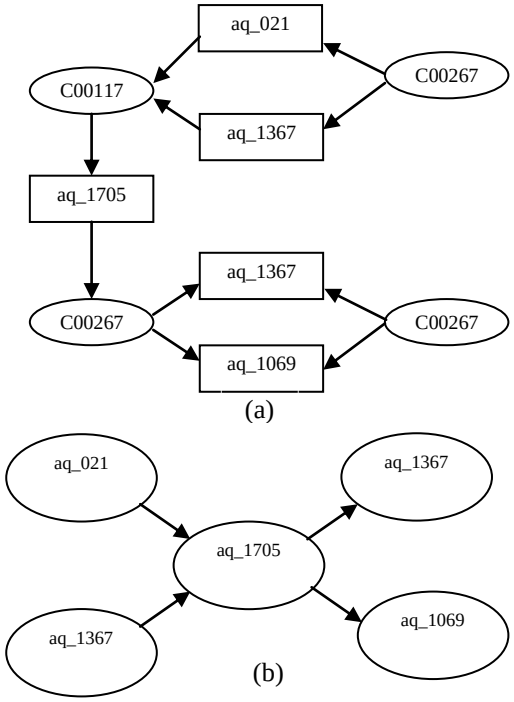

Figure 3: Graph model for metabolic pathway (a) Directed hypergraph representation (b) Directed graph representation

4.2 Mining metabolic pathways

The aim is to find the Maximal frequent subgraph from the simplified graphs which are formalized from the dataset. A subgraph is frequent if its support is greater than ε (support threshold), and it is maximal if it is not contained by another frequent subgraph. Although graph mining is a NP hard problem in general, but here subgraph isomorphism is no longer an issue as it is implicitly enforced by node labeling. Since in framework uniqueness of nodes implies the unique labeling of edges, provides us the opportunity of reducing the problem to frequent itemset mining by specifying edges as fundamental data units. Since frequent itemset mining problem is extensively studied, and there exist many effective and well tuned algorithms, these algorithms are adapt to graph mining problem.

A connected subgraph is represented by a set of edges, since the uniqueness of each edge implies uniqueness of a subgraph represented by a set of edges. Since a unique edge e is a set of two node labels v_i, v_j. A set of unique edges ES = $\{e_1, e_2, e_k\}$ is called a connected edgeset if and only if all edges in the set are connected.

The link between the maximal frequent connected subgraph discovery problem and frequent itemset mining problem is discovered, where graphs (pathways) correspond to transactions and connected edgesets correspond to itemsets. In frequent itemset mining, transactions are set of items and the problem is one of finding all frequent itemsets that exist in more than a specified number of transactions. The fundamental approach used by frequent itemset mining algorithms is to construct frequent itemset from smaller to larger sets based on the fact that any subset of a frequent itemset must be frequent. This is also true for edgeset in the problem. This provides efficient pruning of the search space, since most large set are eliminated without consideration.

In connected subgraphs, it is more efficient to consider only connected edgeset throughout the search process. While maintaining connectivity, it is also necessary to avoid redundancy, in terms of considering the same set of edges more than once in a different order. In order to handle these two issues efficiently, depth first enumeration algorithm based on backtracking is used, which extends each subgraph with only edges from a candidate edgeset. Connectivity is maintained by only adding edges that are connected to the current subgraph and avoid redundancy by keeping track of already visited edges.

4.3 Procedure for mining maximal frequent subgraphs

First find the candidate set edges and frequent edgeset from the threshold value, for the number of graph available. Upon each invocation extend the edgeset (subgraph) by all edges in the candidate set one by one. If the extended edgeset is frequent then the procedure is again invoked for the extended edgeset. When the edgeset can not be extended further this edgeset is recorded and the edgeset which is not contained by any other recorded edgeset will be the maximal frequent edgeset (subgraph). For example: for the input graph given in fig 2 the procedure will work as follows, This collection has five edges in all, ab, ac, bd, ce and de. Figure 4 shows the enumeration tree for mining subgraph that exists in at least three of the input graphs. Procedure is invoked for ab, ac and de, since these are the only frequent edges (threshold =3). Edges bd and ce are not considered since they are not contained in at least three graphs. The frequency of each edgeset is shown in parentheses. At the first invocation, the algorithm starts with edgeset {ab}, whose candidate set is {ac}, and extends it with edge ac as the edgeset {ab, ac} is frequent. Since no further extension is possible, this edgeset is recorded as a maximal frequent subgraph. Note that extension of the edgeset with edge de is not considered since this edge is not connected to the edgeset under consideration, so it never gets into the candidate edgeset. Furthermore, extension of the edgeset {ac} with edge ab is not considered since this edge has already been visited.

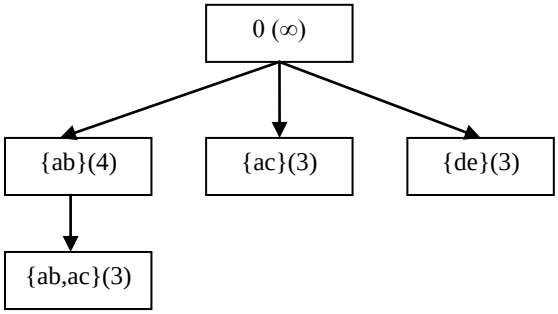

Figure 4: Resulting tree of frequent edgesets for input graphs shown in fig 2.

5. CONCLUSION AND FUTURE WORK

This paper proposes a framework for mining biological network especially for metabolic pathways, the implementation of the proposed methods is being carried out, the results will show the performance and efficiency of the proposed framework, which will be compared with other existing approach for graph mining for the similar dataset. Since graph simplification approach is used which is not leading to NP hard problem of subgraph isomorphism, better results are expected then the existing algorithms (gSpan and FSG). The framework should be extended for the other biological network as protein interaction network and gene regulatory networks, and a generalized framework should be designed and developed to extract frequent pattern and modules from any kind of biological network.

6. REFERENCES

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 487-499, Sept. 1994.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD (Washington D.C., USA), 1993.

[3] R. Agrawal and R. Srikant. "Mining sequential patterns", In ICDE, 1995.

[4] D.J. Cook and L.B. Holder, "Graph-Based Data Mining", IEEE Intelligent Systems, Volume. 15, no. 2, pp. 32-41, 2000.

[5] A. Inokuchi, T.Washio, and H. Motoda. "An apriori-based algorithm for mining frequent substructures from graph data", In PKDD'00, 2000.

[6] G. Cong, L. Yi, B. Liu, and K. Wang, "Discovering frequent substructures from hierarchical semi-structured data", Proc. Second SIAM Int'l Conf. Data Mining (SDM'02), 2002.

[7] Olken F, "Biopathways and protein interaction databases", A lecture in Bioinformatics Tools for Comparative Genomics, Berkeley, CA, feb 2003.

[8] Gouda,K. and Zaki,M.J. "Efficiently mining maximal frequent itemsets", IEEE International Conference on Data Mining (ICDM'01), San Jose, CA, November, pp. 163-170, 2001.

[9] W. A. Rives and T. Galitski, "Modular organization of cellular networks", Proc Natl Acad.Sci.. Usa, 100, 1128-1133, 2003.

[10] Y. Tohsato, H. Matsuda and A. Hashimoto, "A multiple alignment algorithm for metabolic pathways analysis using enzyme hierarchy", Eighth International Conference Intelligent Systems for Molecular Biology (ISMB'00), pp. 376-383, August-2000.

[11] P. D.Karp and M. L. Mavrovouniotis, "Representing, Analyzing and Synthesizing Biochemical Pathways", IEEE expert, 11-21, April 1994.

[12] N. Vanetik, E. Gudes, and E. Shimony, "Computing Frequent Graph Patterns From Semi-Structured Data." ICDM'02, 2002.

[13] Michihiro Kuramochi and George Karypis, "An Efiicient Algorithm for Discovering Frequent Subgraphs", IEEE Transaction on Knowledge and Data Engineering, Vol. 16, No.9, Sept 2004.

[14] Jiawei Han, "How can data mining help bio-data Analysis?" Workshop on data mining in Bioinformatics with SIGKDD02 Conferences" 2002.

[15] Mehmet Koyuturk, Yohan Kim, Shankar Subramaniam, "Detecting Conserved Interaction Patterns in Biological Networks".

[16] YanX, Han J: sSpan: Graph-based substructure pattern mining. In IEEE Intl. Conf. Data Mining, 721-724, 2002.

[17] Bioinformatics center and Institute of chemical research http://www.genome.ad.jp

[18] Protein Data Bank (PDB) http://www.rcsb.org/pdb/

[19] Biobase Biological database http://www.biobase.de/

[20] Metabolic pathways dataset: ftp://ftp.genome.jp/pub/kegg/xml/