Knowledge Discovery in Metabolic Pathways

Muhammad Naeem[1], Misbah Naeem and Sohail Asghar[2]

[1]*Dept. of Computer Science, Mohammad Ali Jinnah University Islamabad, Pakistan*
[2]*University Institute of IT, PMAS-Arid Agriculture University, Rawalpindi Pakistan*

naeems.naeem@gmail.com, sohail.asghar@uaar.edu.pk

Abstract

Graph mining is a dynamic and active research area. In recent years, there is a remarkable boost in graph-structured data resulting graph mining a serious topic in research community. Graph clustering is the process of identifying similar structures in a large set of graphs. Graph clustering is also known as graph partitioning or grouping. This problem plays an important role in various data mining applications. Traditional approaches are centric towards optimization of graph clustering objectives such as ratio association or normalized cut. Spectral methods are also introduced which required Eigen-Vector *computation. However these techniques are slow. We have presented a novel algorithm for detecting closely related groups of graph structures in KEGG metabolic pathways. The technique is based on structural similarity of connected fragments in graph-structured data. The technique is scalable to directed as well as undirected graphs. Preliminary experiments with synthesized data collected from KEGG were performed and their results are reported. The second contribution of this study is the modeling and analysis of combined metabolic reaction networks and relation network and showing their behavior towards scale free network.*

Keywords: *Graph clustering, Biological metabolic pathways, Enzyme, Substrate, Relation network, Reaction network, Scale free Network, Cluster coefficient*

1. Introduction

With the advent of rapid increase in data from heterogeneous sources, it has been conceived that the conventional item-set and attribute-value representations are inadequate for many practical applications in domains such as bioinformatics, chemistry, social network analysis etc. This has motivated data mining research community to ignite learning within alternative and more elaborative representation formalisms such as relational algebra, computational logic, trees, sequences and graphs. All the way through the field of computer science, traditional data structures employed include graphs, trees and sequences. The stimulus for using such representations is that firstly they are more expressive in comparison to flat representations. It makes them more broadly and extensively applicable. Secondly they are potentially well-organized than multi-relational learning in mining techniques. Thirdly, the data structures of graphs, trees and sequences are among the most broadly applied representations and the best understood within various domains of computer science. Thus these representations theoretically as well as practically spur idyllic opportunities for developing appealing contributions in machine learning and data mining. Graph mining is targeted towards finding useful patterns within data representing novel knowledge. Various applications result in graphs of unusual sizes and complexities. Likewise, the applications have different requirements for the underlying mining algorithms. The vital step toward a systems-level understanding of biology was aimed to move away from reductionist to wholist approaches. Such approaches sometimes are also known in the name of bottom-up and top-down approaches, respectively. Conventionally, reductionists look at one component of the system to discover the connections to neighbors, roles in every process that the element is implicated in with its underlying mechanisms of action. In contrast, the wholist approach is aimed towards drawing a snapshot of all elements at a specific level like metabolic pathways, genes translation, transcription, DNA replication and proteins synthesis. Keeping in view of this philosophy, we in this research study have introduced a graph clustering technique *gMean* based on graph density measure using K-Mean clustering technique. The technique has been applied in the domain of bioinformatics. Finally, we shall also discuss imperative avenues of future research in this area. The focus of the article is on data preprocessing from the KEGG followed by clustering performed in the novel technique *gMean*. In the biological metabolic network two kinds of networks can be modeled. In first kind of network, chemical compounds (reactions) are treated as vertices and enzymes as edges. In the second type the edge and vertices are swapped with each other. The first network can also be understood as a bipartite graph. The network generated by only chemical compounds (reactions) is said to be compound (reaction) projection whereas the other network can be called enzyme projection.

2. Graph Theory Preliminary

To analyze a complex network, a clustering coefficient is an important property of a graph and its vertices. It is a global measure for the graph topology. This property is measured for vertices of a graph and then clustering coefficient of while graph is measured. In directed or undirected graph, cluster coefficient for a single vertex is measured as the ratio between number of connected pairs of neighbors of a vertex to all possible pairs of neighbor connections. The average of cluster coefficient of all of the vertices in a graph gives this value for graph. Researchers have investigated this property for the natural graphs as well as man made networks. Its value ranges between 0 and 1. A value close to one indicates that the graph is a highly organized graph. An absolute value of 1 shows that it is a clique graph. A value which is close to 0 indicated that the graph is a random graph. Random graph is highly prone to fracture over loss of some nodes whereas an organized graph is also known as scale free network exhibit its resilient characteristic at elimination of some nodes. Watts *et al.*, [17] described the clustering coefficient Cn of a node n as $Cn = 2en/(kn(kn\text{-}1))$, where kn denotes the number of neighbors of n and en corresponds to the number of connected pairs between all neighbors of n. In directed networks, the definition is slightly modified cause the number of possible pairs between neighbor are double. It results in the equation: $Cn = en/(kn(kn\text{-}1))$ for the directed network. This property is also an indication of cliquness of a graph.

3. Graph Modeling for Metabolic Pathways

We have devoted this section to elaborate some basic concepts related to networks in system biology. In the literature numerous architecture have been proposed showing that biological, cellular and metabolic pathways can be graphically modeled. Such interactions in the multilayered organization of organisms lay out the foundation of the links between individual molecules and large-scale organization of the cell via functional modules.

Biological networks representing cellular interactions are found in the form of signal transduction pathways, metabolic pathways, gene regulatory networks and protein interaction networks. Nontrivial advances in the understanding of genomics have motivated the efforts targeted towards detecting apposite models for such networks. It results in focusing the significant research attention. Among all of the above biological networks, metabolic pathways have a relatively longer history. Particular metabolic function is characterized by the process of chemical reactions. With the recent advances in application of computational methods to cellular biology, there have been successful attempts at synthesizing, modeling and organizing metabolic pathways into public databases, such as KEGG (2012). Metabolic pathways can be defined as the chains of reactions coupled to each other by chemical compounds (metabolites) through product substrate interaction. A directed hypergraph with each node representing a compound, and each hyperedge corresponding to a reaction or an enzyme is a natural mathematical model for metabolic pathways. The direction of a hyperedge indicates whether the compound is a substrate or product of the reaction. However there is possibility for the development of a much simpler model of directed graph, which is related to only relationships between enzymes. In such a model, enzymes are related 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. In fact, various binary relations can be show from KEGG (2012) metabolic pathways. Edges are usually labeled by the compounds relating the two corresponding enzymes. The structure of these networks reflecting temporal information is an interesting property of metabolic pathways. Specifically, an enzyme may appear more than once in the same pathway, indicating that this enzyme participate in the entire process at diverse time instances. The inference of this fact on graph modeling results in that any node of the graph (pathway) will not have the unique label and this label (enzyme) will be repeated on other node. Our interest in this study can be either in this temporal relationship or general relationship between pairs of enzymes. In such scenario it is possible to merge nodes in the graph having same label. It simplifies the graph modeling significantly. Moreover, extracting the existing temporal patterns from generic patterns is also simplified in this way. Design of graph modeling of biological networks delivers a framework for the solution of various problems targeted towards understanding the enzymatic interactions in pathways. Among these problems in the literature, researchers have mostly exploited the techniques related to graph matching, clustering, shortest-path computation, graph alignment, graph mining and subgraph homeomorphism. Graph clustering provides a suitable framework for discovery of functional modules, which can be defined as a substructure of a biological network distinguishable from other modules in terms of functionality. One approach to the detection of functional modules is graph clustering, or in our case it can be said as the discovery of dense graphs based on the assumption that a group of functionally related entities are likely to interact densely with each other while being somewhat separated from the rest of the network. The other approach is characterized by multiple alignments of graphs for the purpose of discovery of common substructures in the network. Its foundation lies in the fact that functional modules can be projected to recur among several pathways and/or organisms. In addition to this, graph alignment and graph mining is also aimed towards provision of other opportunities for analyzing biological networks. The prime focus of our study is on graph clustering in biological networks for finding functional modules defining substructure of a biological network distinguishable from other modules in terms of functionality. In order to investigate our study it is convenient to define the biological functionalities in mathematical equation as below. We shall define these mathematical equations in top down approach means first metabolic pathway defined formally followed by other equations giving fine detail over its component.

Definition 8: *An instance of metabolic pathway P can be described as formally.* $P = \{p \in R \mid L\}$ *where p is a metabolic pathway, R is reaction and L is relation .*
Definition 1: *A Reaction R(r) is expressed as .* $R = \{ r , r = (Z(z) \cap D(d)) \mid (Z(z) \cap S(s)) \}$*. Every reaction is a representation of two vertices and one edge such that* $V = Z(z)$, $E = VxV$ *with* $label(E) = D(d)$ *or* $label(E) = S(s)$.*W*here $Z(z) \in Z$ is set of enzymes, $D(d) \in D$ is set of products, $S(s) \in S$ is set of substrates.
Definition 2: *A Relation L(l) can be described as .* $L = \{l = (N(n) \cap D) \mid (N(n) \cap S)\}$*. Every relation* $l \in L$ *is associated with a set of Gene or Maps* $N(n) \subseteq N$*, such that* $V=N(n)$ *and* $label(E) = D(d)$ *or* $S(s)$.

Each metabolic pathway can be described as a set of chemical reactions which are catalyzed by enzymes. *Reactant*s are a set of compounds involved in every *reaction*. Those *reactants* which are input of the *reaction* are called *substrates* and the output *reactants* are termed as *products*. We took data of various organism metabolic pathway. These were included Bacteria, Mammals, S.aureaus, Eukaryotes and Plants. There are thousands of chemical reactions normally occurring in a cell. Metabolic pathways are one of the notable sets of the series of chemical reactions. In each pathway, chemical reactions are responsible to modify a principal component. Enzymes are proteins which expedite the whole process in a way that these enzymes are reclaimed at the end of the series of reactions. Their existence in the pathway is to act like catalyze. During this process these enzymes usually require vitamins, dietary minerals and other cofactors for their correct functionality. As during whole of the process, a large number of metabolites are involved, this make metabolic pathway elaborative and rich in information. Moreover, a lot of distinct pathways also co-exist inside a cell. We can term this large collection of pathway as metabolic networks which are essential part in the process of maintenance of homeostasis within a cell or organism. Metabolism is mainly divided into two categories, catabolic pathway and anabolic pathways. Former is characterized by the synthesis process. However, the later is related to breaking down of large compounds. A metabolic pathway is composed of the step-by-step alteration of an initial molecule to make a new product. The resulting product can eventually be used in either the end-product of metabolic pathway on immediate basis. In other cases these released products are used to initiate another metabolic pathway, or in the third case they are stored by the cell for any other possible usage. Substrate is a molecule which participates in a metabolic pathway depending on the availability of the substrate or otherwise at the demands of the cell. A metabolic pathway rate is characterized by an increase or decrease in concentration of catabolic or anabolic intermediates and/or end-products.

4. Clustering graph-structured Data

In statistical data analysis, clustering in its meaning is more close to automatic classification, community or numerical taxonomy analysis. The underlying distance measure in clustering calculates the *similarity* of two elements in all of the sets. It is evident that some elements may be close to one another according to one distance and farther away according to another, so this distance measure may influence the shape and population of the clusters. The common distance measures are the euclidean distance, manhattan distance, mahalanobis distance and hamming distance. In graph mining, some components of graphs may share structural similarities in the collection of large number of graphs. For example, organic compounds with a benzene ring have similar chemical properties because they are observed with the same benzene ring. Such observations motivate a research branch of computational chemistry called (quantitative) structure activity relationships. In Quantitative structure-activity relationship, the association between the property or activity of chemical compounds and their structure has been investigated. Another example in support of mining structural similarity mining is analysis of "small world". Such networks share the property that any node can be reached locally at a few steps how dense the graph may be because the overall average path length between two nodes in the network is relatively short due to the existence of some edges connecting distant nodes. In keeping view of this investigation that similar structures within graphs may result in common characteristics, in the proposed framework we targeted at clustering graph-structured data based on their structural similarity. When graphs are categorized into clusters, a small number of graphs can be selected from each cluster as representatives. As in the research on small-world networks, structural properties of a graph in terms of connectivity (degree) of nodes in the graph were considered and transform the graph representation into a corresponding spectrum. Such transformation motivates us to consider a graph into a spectrum behaving like dictionary of hash function. Thus, if a user can specify the desirable resource (pathway) combinations as a graph, our technique can be helpful in discriminating the metabolic pathway network of resource combinations into similar and dissimilar ones in terms of their grouping based on spectrum. We also proposed a method of sub clustering based on our notion of spectra within a cluster, to determine the appropriate number of clusters to be created. This spectrum is characterized by the density of a graph. These clusters are then subjected to sub clustering with respect to the transformed spectra by applying a novel clustering measure using K Mean. K-means is based on a simple iterative method. In K-means algorithm, a user specified number of clusters, k is provided to partition a given dataset. This algorithm was investigated by several researchers across different disciplines. The input to the algorithm is a set of d-dimensional vectors, $D = \{\mathbf{xi} \mid i = 1. . . N\}$, where $\mathbf{xi} \in \mathrm{D}$ denotes the ith data point. The algorithm is initialized by picking k points in D as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

4.1. Step 1: Data Assignment

Each data point is allocated to its *closest* centroid. This broke ties of each point arbitrarily while resulting in a partitioning of the data.

4.2. Step 2: Relocation of "Means"

The center or mean of all data points allocated denotes each cluster while this centroid may be relocated in every scan. If the data points appears with a probability measure that is weights, this leads to conclustion that the relocation has been achieved to the expectations (weighted mean) of the data partitions. When the assignments (and hence the **cj** values) no longer alters, it indicates that the algorithm converges. Note that each iteration needs $N \times k$ comparisons, which determines the time complexity of one iteration. The number of iterations needed for convergence is not fixed. Its variation usually depends on value of N, but as a first cut, this algorithm may be considered linear in the dataset size. A resolvable issue in this step is how to quantify "closest" in the said assignment step. Typically euclidean distance is considered as the default measure of closeness, in which case one can readily demonstrate that the non-negative cost function $\sum_{i=1}^{N}(\arg\min \| Xi - Cj \|_2^2)$ will be reduced whenever there is a alteration in the assignment or in the relocation steps, resulting which convergence is assured in a finite number of iterations. The greedy-descent nature of k-means on a non-convex cost also entailed that the convergence is only to a local optimum, and infact the algorithm is usually quite sensitive to the initial centroid locations.

5. Literature Review

In this section, we shall investigate literature review of clustering in general and then converging to graph clustering in particular. Hattori *et al.*, [9] proposed the approximation of the similarity of two compounds based on graph representation. According to this technique, each chemical structure can be modeled by a two-dimensional network where the vertices are related to the atoms while the edges corresponds to the bonds between them. The similarity of the two compounds is calculated by detecting their common subgraphs, followed by their alignment accordingly. They provide an online web service SIMCOMP Search to calculate the similarity score between two compounds by means of the graph representation [16]. However, the web-server only provides similarity scores greater than 0.4.

We can categorize algorithms in two classes. One is exact approach employing exhaustive approach whereas the other techniques know as approximate algorithms are characterized by provision of nondeterministic output. This method yields different results on different execution even if the output is kept same. The objective of approximate algorithms is to come up with a solution which is very close to real solution. This technique is usually far more efficient whereas its efficiency is achieved at the cost of minor and negligible difference in comparison to the exhaustive or brute force algorithms. Although, no global definition for a graph clustering has been defined; however Schaeffer [6] has defined some desirable cluster properties on which a graph clustering can be defined. These properties are including each vertex of every edge in a cluster must be reachable to any other vertex of the same cluster through its own cluster only. It means the path from each vertex of cluster should be internal. However these characteristic are limited to only a single large graph. Then some other properties must be defined. In literature, graph clustering measures have been defined in terms of vertex similarity which is based on distance measure. It means the degree of a vertex plays an important role. The existing global techniques are proficient in dealing with large graphs that is graphs having millions of vertices sparsely [11]. In a global clustering, every node of the input graph is associated to a cluster in the output of the method, however in a *local* clustering; the cluster assignments are achieved for a specific subset of vertices, usually only one vertex. Dhillon *et al*., [10] proposed a general algorithm based on multilevel methods. This approach originally based on METIS [7] and works for a wide class of graph clustering objectives. The technique was divided into three phases each of which was specialized for each graph clustering objective including Ratio Association, Ratio Cut, K-L Objective and Normalized Cut. Murty *et al.*, [1] provided the taxonomy of the clustering algorithms in a hierarchical structure. Majority of the clustering algorithms can be classified either hierarchical or partitioning clustering algorithms. The classifications of clustering algorithms are based on the several crosscutting aspects [1].

- Agglomerative vs. Divisive: This division is related to algorithmic structure and operation, as an agglomerative approach is a bottom-up construction while divisive is top-down approach.
- Monothetic vs. Polythetic: This aspect refers to the different use of features in the process, either sequential or simultaneous. While majority of algorithms are polythetic, a simple monopthetic algorithm was reported in [15].

- Hard vs. Fuzzy: This refers to the membership of the data. A hard clustering allocates one data point to only one cluster wheras in a fuzzy clustering algorithm, a single point is allocated to multiple clusters. Some of applications actually demand the fuzziness.
- Deterministic vs. Stochastic: The clustering algorithm is using to the optimal techniques in this category. Most of partitioning clustering algorithms require either a random search technique for optimization or a deterministic objective function.
- Incremental vs. Non-incremental: The algorithm in which the size of data can be amplified is regarded as as incremental otherwise non-incremental.

The Categorization of existing clustering heuristics are also diversified as can be observed in the survey papers of clustering algorithms [1, 4]. Bjvrn *et al.*, [4] reviewed some of graph-based clustering algorithms for bioinformatics applications. Clustering algorithms are very useful for biological networks such as Protein-Protein Interaction (PPI), Transcriptional Regulatory Network and Metabolic Networks. Clique-based and Center-based clustering techniques developed for small data sets whereas for large data sets, some of the known techniques are distance k-neighborhood, k-cores and quasi-cliques as well. Various existing research has addressed similarity measures and clustering methods for graph-structured data. Related work includes Topological Fragment Spectra (TFS) [18] which characterizes the properties of chemical compounds in terms of fragments (subgraphs) within the compounds. ANF [5] is an approach to the fast calculation of similarity for large-scale graph-structured data. Our method is based on density measures which in term represent the specific relationship between enzyme and compound reactions. Zantema *et al.*, [8] described an algorithmic improvement for detection of frequently occurring patterns as well as modules in biological networks. They showed that this improvement is based on fact that finding frequent sub-network problem is reducible to the problem of finding maximal frequent item sets. They performed their experiments in metabolic pathways obtained from the KEGG database. Jose *et al.*, [12] proposed a metabolic pathway alignment technique. They showed conserved reactions in different groups of organism. They also suggested that yet to be specified biological role of these conserved reactions. Chen *et al.*, [14] developed a mathematical model for prediction of substrate-enzyme-product triads network. They introduced a molecular graph to calculate the similarity between substrate compounds and the product compound. However, this comparison was not between complete reaction and relation network. Barabási *et al.*, [2] argued that if a graph shows any of the structural properties then it is expected that it will also show the other characteristic measures. Based on these characteristic it can be inferred in what kind of network type the graph belong to. Is it a regular network, scale free network, random network or small world network? Cluster coefficient plays an important role in placing any network among these categories. Complex networks are the backbone of complex systems. Barabási *et al.*, [2] discussed structural metrics for complex networks. These are average path length, degree distribution, cluster coefficient. On the basis of these metrics, four network models have been identified in the literature. These literatures include regular networks, random networks, small-world networks, scale-free networks. Each model of network is believed to exist in specific range of these metrics. For example, clustering coefficient value for regular network, small world network and random network is usually in range of 1.0 to 0.75, 0.75 and below 0.75 to 0. In the next section, we shall discuss the proposed framework which is based on the highlights investigated in this literature review.

6. gMean: Proposed Framework

In the literature review, it was highlighted that researchers have emphasized on vertex similarity and density measures. However in bioinformatics domain, KEGG (2012) data can be modeled in two kinds of graphs. In one kind it can be modeled as enzyme as vertex and Compound as edges. In the other network, enzyme represents edges and compound as node or vertex. Current KEGG (2012) biochemical pathways hold a complete database which is also a basic reference path ways that can be shown in form of networks enzyme formed manually. By means of enzyme genes identification, pathway organisms are formed automatically. KEGG (2012) database contains pathway maps from several Processes which include carbohydrate metabolism, lipid, energy, nucleotide and amino acid metabolism. The numerical figure of these was up to 257 organisms in 2009. Figure-1 represents our proposed architecture of gMean. The input database is taken from KEGG (2012) which are metabolic pathways. For the sake of simplicity, we have divided our architecture into five steps.

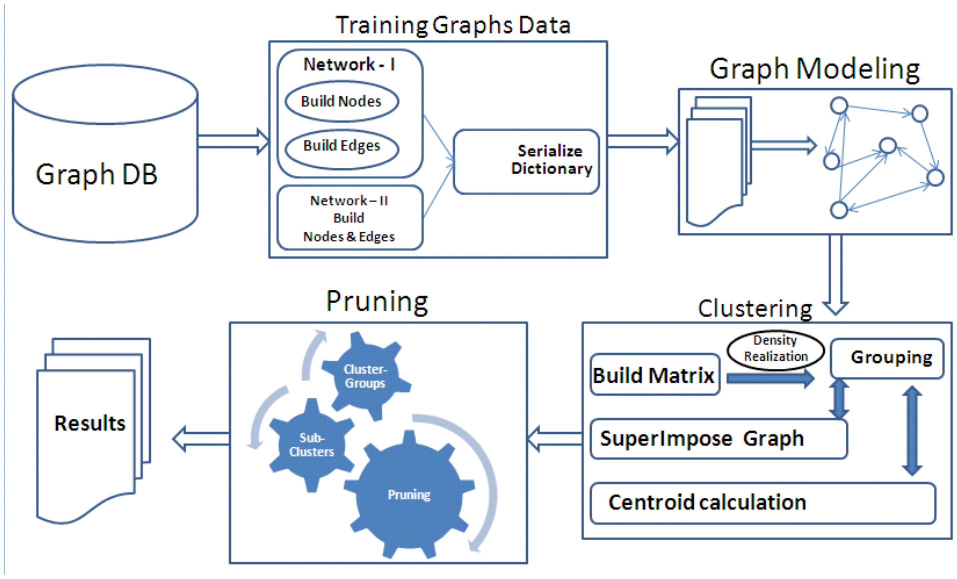
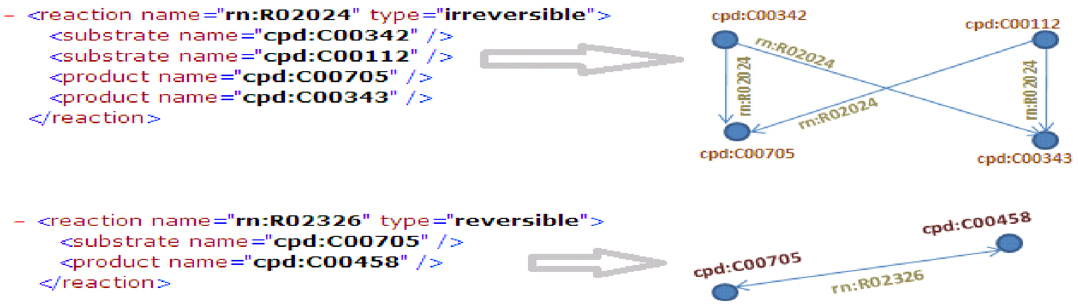


Figure 1. gMean: Proposed Framework

6.1. Step-1 Training Graphs Data

This steps deals with the issue of which preprocessing steps are required to be applied to ensure the data more suitable for data clustering. Data preprocessing is a broad area comprising of versatile strategies and techniques that are interrelated in complex ways. Some notable approaches for preprocessing data are including Aggregation, Sampling, Dimensionality reduction, Feature subset selection, Feature creation, discretization and binarization, Variable transformation. Among these approaches, we have adopted Sampling.

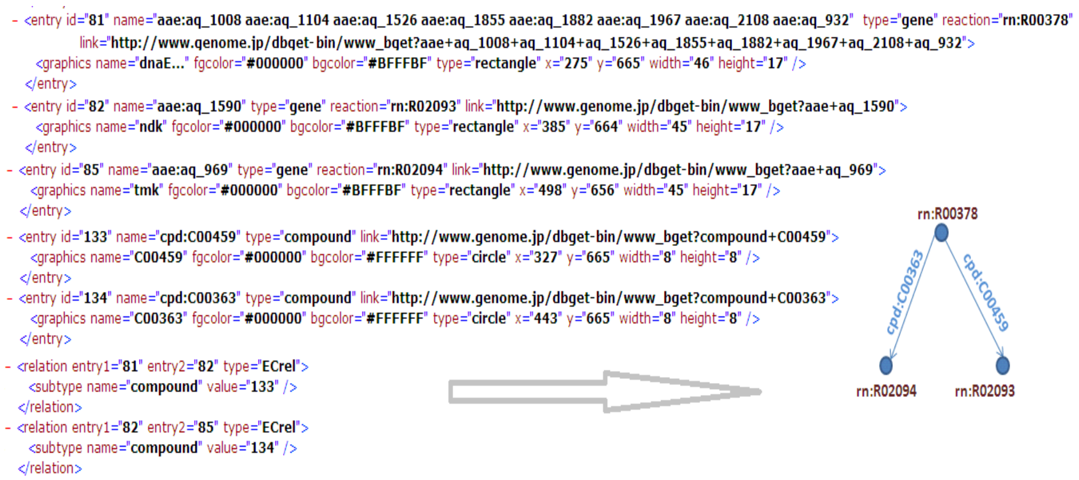


```
- <reaction name="rn:R02024" type="irreversible">
    <substrate name="cpd:C00342" />
    <substrate name="cpd:C00112" />
    <product name="cpd:C00705" />
    <product name="cpd:C00343" />
  </reaction>

- <reaction name="rn:R02326" type="reversible">
    <substrate name="cpd:C00705" />
    <product name="cpd:C00458" />
  </reaction>
```

Figure 2. Reaction Network

We shall not discuss other approaches in detail as they are beyond the scope of this research. However, we shall investigate Sampling in context of graph clustering. In sampling, usually a subset of the data objects are analyzed. KEGG (2012) provided more than 250 organism data, each of which was consisting of hundreds of metabolic pathways. Each metabolic pathway can be modeled into a single graph. Usually Sampling suffer from loss of information. However in our case, we just took three organism data set. So there was no issue of loss of information.



```
- <entry id="81" name="aae:aq_1008 aae:aq_1104 aae:aq_1526 aae:aq_1855 aae:aq_1882 aae:aq_1967 aae:aq_2108 aae:aq_932"  type="gene" reaction="rn:R00378"
        link="http://www.qenome.jp/dbqet-bin/www_bqet?aae+aq_1008+aq_1104+aq_1526+aq_1855+aq_1882+aq_1967+aq_2108+aq_932">
    <graphics name="dnaE..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle" x="275" y="665" width="46" height="17" />
  </entry>
- <entry id="82" name="aae:aq_1590" type="gene" reaction="rn:R02093" link="http://www.genome.jp/dbget-bin/www_bget?aae+aq_1590">
    <graphics name="ndk" fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle" x="385" y="664" width="45" height="17" />
  </entry>
- <entry id="85" name="aae:aq_969" type="gene" reaction="rn:R02094" link="http://www.genome.jp/dbget-bin/www_bget?aae+aq_969">
    <graphics name="tmk" fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle" x="498" y="656" width="45" height="17" />
  </entry>
- <entry id="133" name="cpd:C00459" type="compound" link="http://www.genome.jp/dbget-bin/www_bget?compound+C00459">
    <graphics name="C00459" fgcolor="#000000" bgcolor="#FFFFFF" type="circle" x="327" y="665" width="8" height="8" />
  </entry>
- <entry id="134" name="cpd:C00363" type="compound" link="http://www.genome.jp/dbget-bin/www_bget?compound+C00363">
    <graphics name="C00363" fgcolor="#000000" bgcolor="#FFFFFF" type="circle" x="443" y="665" width="8" height="8" />
  </entry>
- <relation entry1="81" entry2="82" type="ECrel">
    <subtype name="compound" value="133" />
  </relation>
- <relation entry1="82" entry2="85" type="ECrel">
    <subtype name="compound" value="134" />
  </relation>
```

Figure 3. Relation Network

6.2. Step 2: Graph Modeling

In graph modeling, metabolic pathways are shaped in terms of nodes and edges. Our goal in modeling metabolic pathways is to discover those common pathways which have same enzyme substrate or compound ratio interactions related to each other. That is why we have modeled metabolic pathways showing directed graphs proficient in capturing the interaction information effectively as shown in Figure 2 and 3. Furthermore, we model two kinds of networks. One of which was having enzyme as node and compound / substrate as link label whereas in the other kind of network, we represent enzyme as the label of the edge and compound / substrate as the node of the graph. In both of the cases the enzyme was termed a unique index, independent of the number of times the enzyme appears in the underlying pathway. Such restriction was imposed for the sake of achieving simplicity while providing biologically meaningful results. An important aspect of this modeling is that it does not incur any loss of information as the model was easily reversible to its original information of the pathways after clustering has been done. Formally it can be defined it as if a g*iven metabolic pathway P(M,Z,R,S), the associated directed graph G(V,E) of P is constructed as below: for any vertex V:(i, $z_i \in$ Z) for any Edge E: (z_i,z_j,s) where Z,M,R,S corresponds to enzyme, Metabolite, Reaction and substrate compound respectively*.

It indicates that a directed edge exist from one enzyme to another in the graph if and only if the second enzyme consumes a product of the first one.

6.3. Step-3: Clustering

We have employed graph clustering based on internal density of each metabolic pathway. Schaeffer [6] has defined the *local density* of a graph in *G* = (*V*, *E*) to be simply $\delta(G(S)) = \frac{|E(S)|}{\binom{|S|}{2}}$ where E(S) denotes number of edges and |S| denotes number of vertices. If every node is connected to every other node then such clique graph have local density equal to 1. If there is no edge found in the graph then its local density come out to be 0. By means of this definition of local density, it is possible to calibrate all of the metabolic pathways on the same scale of 0 to 1.

6.4. Step-4: Pruning

In this step, we grouped all of the graphs with same local density measure. This results in provision of a pruning measure. On the basis of this measure, observations in each cluster can be filtered out. In this way, only those clusters survive which have observation points with significant number of same local density.

6.5. Step-5: Results

At this stage, we get those clusters which have the points with maximum level of similarity in terms of density, In the first step, actual names of the enzymes, metabolites, substrates and compounds were serialized. In this step, we deserialize all of the numbers into these actual values. This results in forming tables as shown in table-1 and table-2

7. Results and Discussion

Type Literature investigated so far has shown that biological processes have high potential to be modeled into a graph. More than 71 thousand pathways were retrieved from KEGG website (2012). In first step, these were categorized according to the numeric nomenclature provided by KEGG (2012) such as all pathways ending in *00010* are grouped. In addition to it, pathways related to organism including *Bacteria*, *Archaea*, *Eukarya*, *Mammals*,*Plants*, *S. pyogenes*, *Escherichia coli* and *S. aureus* were also separated into their respective collection. Second grouping was made in the light of investigation found [12]. We performed our experiments on these groups in two modes. In first mode, we perform experiments on each individual pathway. In second mode, we treated all of the pathways as a single huge network followed by performing the experiment on it. First contribution of this study is that it can group those metabolic pathways which have the closest ratio between enzymes and substrate products. These groups may give some notions for further analysis of metabolic pathways. Metabolic pathways are a representation of metabolic substrates and products and enzymes modeling a digraph. In this digraph, a known metabolic reaction exist that acts on a given substrate yielding a desired product. In the literature a lot of statistical properties of these giant metabolic networks have been explored. Similarity characteristic of a fragment is characterized by the ratio of edges to vertices in a metabolic pathway. We shall divide our experiments into two categories. The first category was grouping of pathways using the above similarity measure whereas the second experiment is related to computations of clustering coefficient followed by analysis of the behavior of the biological network. First experiment was performed on the organism of pathway with names ending in numeral *00240*. Collection of metabolic pathways in this group was comprising of 587 networks. These networks were modeled with Enzyme projection, means compounds were vertices while enzyme were taken as the label of the edges.

Table 1. High Density Cluster, (Centroid = Centroid= 0.171631527)

Pathway	Name	Pathway	Name
maq00240	Marinobacter aquaeolei	ace00240	Acidothermus cellulolyticu
mfa00240	Methylobacillus flagellatu	aeh00240	Alkalilimnicola ehrlichei
net00240	Nitrosomonas eutropha	ajs00240	Acidovorax sp. JS42
neu00240	Nitrosomonas europaea	bmb00240	Brucella abortus 9-941
ngo00240	Neisseria gonorrhoeae FA 1	cjj00240	Campylobacter jejuni 81-17
nme00240	Neisseria meningitidis MC5	ezma00240	Zea mays
nmu00240	Nitrosospira multiformis	mav00240	Mycobacterium avium 104
pcr00240	Psychrobacter cryohalolent	mbo00240	Mycobacterium bovis AF2122
pvi00240	Chlorobium vibrioformis	mmc00240	Mycobacterium sp. MCS
aae00240	Aquifex aeolicus	mpa00240	Mycobacterium avium paratu
		pna00240	Polaromonas naphthalenivor

Clustering is a good choice to make a brief summary of actual data. This technique is able to produce a variety of results which can be treated as a summary for any other data mining technique. Table 1 & 2 both are describing those metabolic pathways which are most coherrent in terms of the graph density. In Table 1, density of all of the metabolic pathways is of 0.1681, which reveals the internal structure of these metabolic pathways is most similar from one aspect. This revealation may help biologist to further investigate into these metabolic pathways from biological perspective.

Table 2. Low Density Cluster, (Centroid = Centroid = 0.130770705)

Pathway	Name	Pathway	Name
ade00240	Anaeromyxobacter dehalogen	par00240	Psychrobacter arcticum
dme00240	Drosophila melanogaster	pfa00240	Plasmodium falciparum 3D7
dol00240	Desulfococcus oleovorans	rpd00240	Rhodopseudomonas palustris
efa00240	Enterococcus faecalis	mtc00240	Mycobacterium tuberculosis
estu00240	Pyrimidine metabolism	ape00240	Aeropyrum pernix
hpj00240	Helicobacter pylori J99	bas00240	Buchnera aphidicola Sg
lma00240	Leishmania major	ccv00240	Campylobacter curvus
mhy00240	Mycoplasma hyopneumoniae 2	cff00240	Campylobacter fetus
mkm00240	Mycobacterium sp. KMS	cgb00240	Corynebacterium
mmz00240	Methanococcus maripaludis	dpop00240	Pyrimidine metabolism
nfa00240	Nocardia farcinica	ebvl00240	Pyrimidine metabolism
nma00240	Neisseria meningitidis Z24	eli00240	Erythrobacter litoralis

The second Table-2 is showing the members of second cluster. The density of these metabolic pathways ranges from 0.1425 to 0.1448. It is particularly mentioned that these two tables are showing the result on network of type 1 in which compound were taken as vertices and enzymes as edges. If we swap both of these then a new type of network is built.

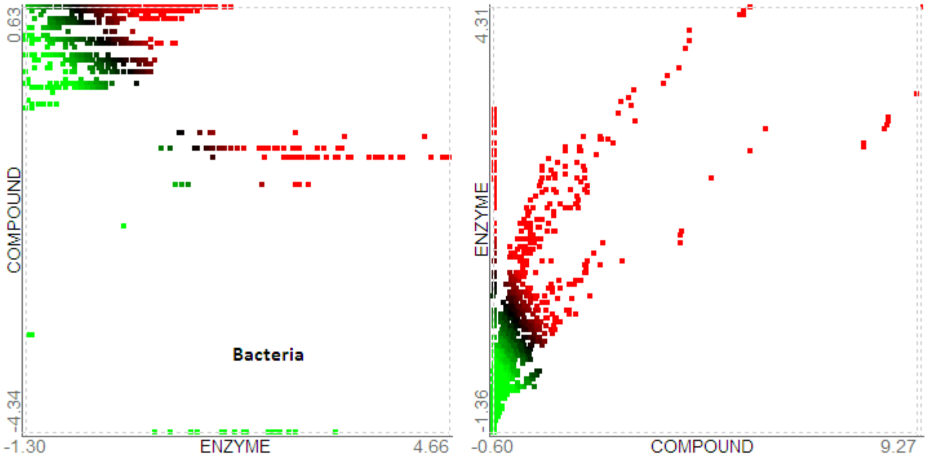


Figure 4. Bacteria Metabolic Pathways

The correlation between variables e.g. enzyme and compound is also significant for multivariate statistical analyses such as hierarchical cluster analysis and principal component analysis (PCA). We investigated similarities and dissimilarities in metabolic correlations in the taxon of *Bacteria*, *Eukarya*, *Mammals*, *Plants* and *S.aureaus*.

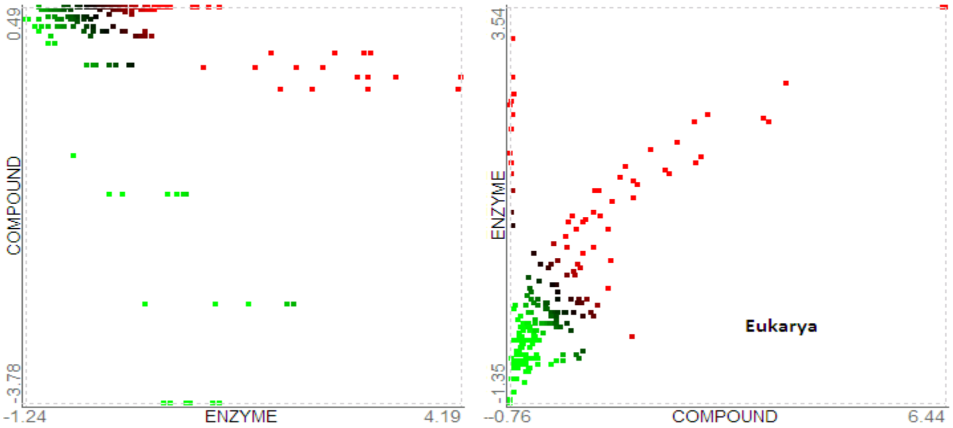


Figure 5. Eukarya Metabolic Pathways

From Figure 4 to 8, we have shown results of our prototype over various pathways. These pathways were first classified into different taxon. Some of these are making hyperbolic curves while others are projecting parabolic curves. In all of these scatter diagram, it is evident that some of the organism have high correlation to each other over their pathway structure. In each of the figure, we have made a comparison between both kinds of network. The observations shown in green color represent metabolic pathways with close structures. In network 2, the dense regions are more packed as compared to network of type 1. This clearly indicates that if enzymes are treated as independent variable, then ratio of compound to enzyme makes a sense with dense clustering.

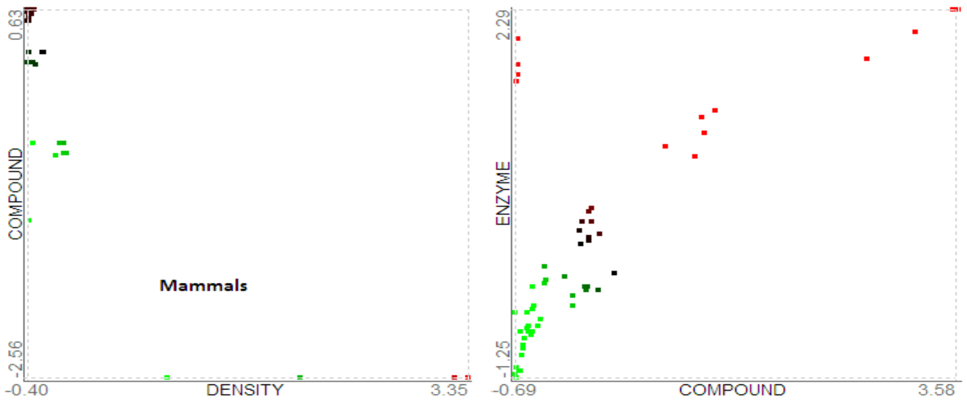


Figure 6. Mammals Metabolic Pathways

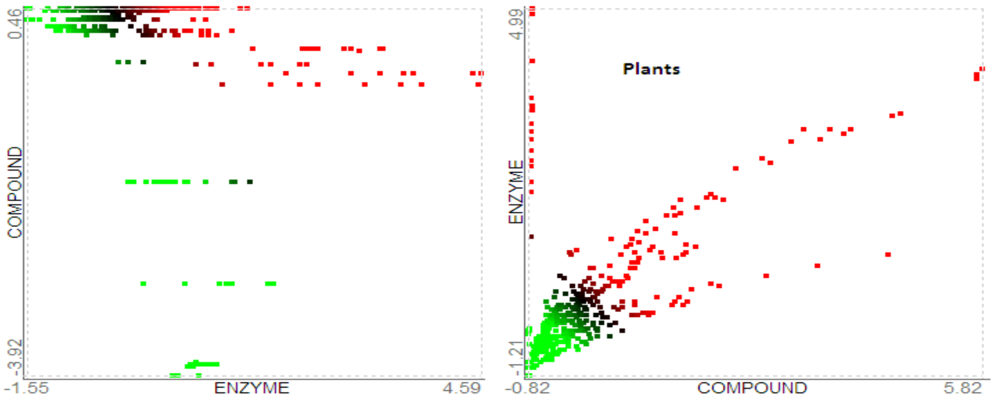


Figure 7. Plants Metabolic Pathways

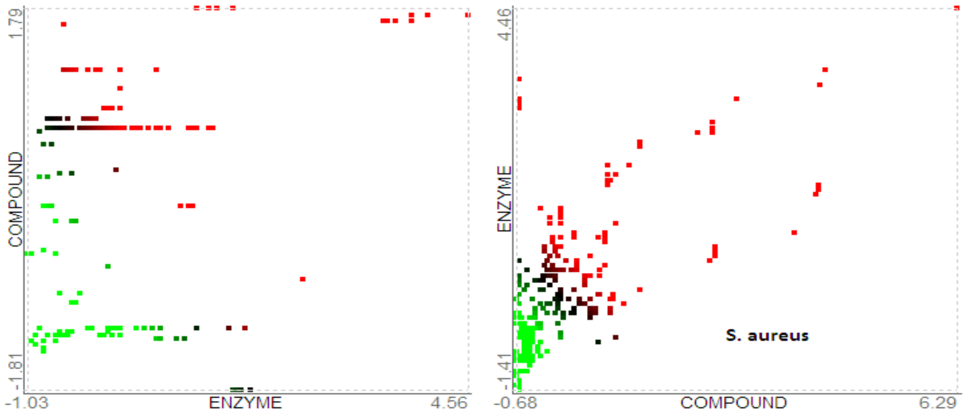


Figure 8. S.aureaus Metabolic Pathways

Figure 9 is depicting a brief overview of typical structural features of metabolic pathways considered in this study. Projections in this figure are showing that average edges, average vertices are much higher in case of network modeled with compound as vertices and enzymes as edges. We demonstrate that the graph-clustering approach identifies enzyme, substrate compound dependent metabolic clusters associated to the biochemical pathway. Metabolic correlations complement information about changes in metabolic network density. This can also helpful in elucidating the organization of metabolically functional modules.

This part of discussion is related to the experiment performed to analyze the network characteristics. It was discussed in literature review that cluster coefficient is an important metric based on which we can classify any complex network. Conventionally it was believed that all of the biological networks are small world networks. This study involves the computation of all of the metabolic pathways available till March 2011. We classified them into organism structures, numeric nomenclature and then in the final all of the pathways were combined to produce a single giant graph. Table 3 shows that organism wise, these networks can be classified close to random graph. This is true for both kinds of network. As shown in Table 3, eight organism of various domain share one common fact that all of them suffer from random network tinged with semi small world effect. However if we compare both kinds of network to each other then relation network is looking more organized as compared to reaction network which is prone to be more randomized.

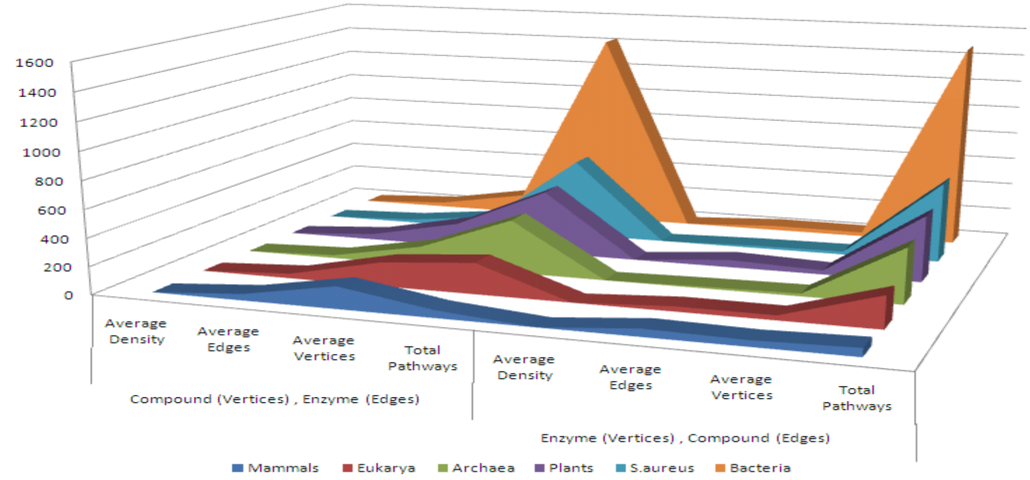


Figure 9. Typical Structure Features of Metabolic Pathways Considered in this Work

Table 3. Cluster Coefficient Values for Various Organism

Organism	Reaction Networks			Relation Network		
	Nodes	Edges	Cluster Coefficient	Nodes	Edges	Cluster Coefficient
Archaea	548	995	0.282945828	1253	5692	0.298068503
E.coli	1276	2298	0.23494575	2258	12121	0.317079891
Eukarya	616	1266	0.281683438	1297	6316	0.345615045
Mammals	243	513	0.332470885	567	2176	0.420365383
Plants	735	1563	0.267926251	1017	3930	0.417011967
S.aureus	818	1400	0.23596242	1356	7198	0.281645346
S.pyogenes	637	1074	0.280801931	1244	6395	0.31226332
Bacteria	811	1804	0.269044937	32867	241109	0.37067541

The investigation made in our proposed framework holds correct for the complete set of pathways as depicted by Table 4. Relation network is again found to be much small world as compared to reaction network. From Table 4, another interesting aspect observed is that relation network is more densely packed. If we examine this fact in the definition of clustering coefficient then we can conclude that it is due to comparatively high number of nodes and connecting edges.

Table 4. Network Characteristics of all Pathways

Pathway	Nodes	Edges	Cluster Coefficient
Reaction	3987	7293	0.1845
Relation	47646	392638	0.3244

Table 5. Cluster Coefficient Values for Individual Metabolic Pathways

Pathway	Reaction Network			Relation Network		
	Nodes	Edges	Cluster Coefficient	Nodes	Edges	Cluster Coefficient
aha00252	23	44	0.542199488	39	183	0.697790774
bad00252	16	24	0.526442308	23	96	0.721348941
cne00252	22	38	0.439669421	35	110	0.694110413
aac00010	23	44	0.450724638	26	43	0.439044289
aae00020	16	28	0.575	29	45	0.286939497
aae00052	5	12	0.866666667	4	4	0.583333333
aau00760	13	25	0.48974359	20	37	0.51
hch00252	23	40	0.476768968	37	159	0.680926388
aba00520	9	18	0.410582011	12	24	0.624170274
etae00252	26	49	0.419758673	43	198	0.765211115
dtni00252	23	41	0.357575758	37	122	0.718546601
esbi00252	24	43	0.380982906	38	140	0.719789056
fal00252	24	41	0.461538462	35	129	0.756848073
fra00252	23	39	0.438735178	34	114	0.746053874
aci00252	24	39	0.485119048	35	144	0.72769227
eosi00252	26	47	0.351465201	42	178	0.726513131

Table 5 represent a sample set from all of the calculations performed on more than seventy thousand metabolic networks. While analyzing this data, it was evident clearly that a very sharp variation in the cluster coefficient was found. Some of them were originated with zero values for cluster coefficients as shown in the attached dataset result sheet. However in the figure, we have shown only a few of them showing high value of cluster coefficient. Based on the this data, one can deduce that all of the metabolic pathways are not exhibiting regular or small world effect rather this property varies from one network to other network.

8. Conclusion

As the data and interactions of bio-molecular networks is on a rapid increase with the advent of modern biological techniques, it aimed towards bigger challenges for data mining community in the problem of mining patterns, motifs and modules turning graph mining more interesting and useful in biological domain. This study provides a framework targeted towards using graph mining which will focus on the efficiency of computation time by employing *gMean*. At the same time this study has provided a framework to provide summarized information about the relative relation between enzyme and substrate compound. The results shown in Table 1 and Table 2 clearly indicate that there are a certain group of metabolites having the same ration between enzyme and substrate compounds. This study can be useful further in the analysis of enzyme reaction in metabolic pathways in biological perspectives.

More than seventy thousand metabolic pathways belonging to various organisms of life were graphically modeled. We have shown that, even though considerable variation in their individual constituents and pathways, some of these metabolic pathways have found with the same network metrics depicting striking similarities among these complex biological systems. However, when these pathways were combined to form a single network, they illustrate more relatedness towards a random network. It is evident that not all but most of these pathway networks may represent a common blueprint of interactions for the large-scale organization in cellular constituents.

References

[1] K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review", ACM Comput. Surv., vol. 31, no. 3, **(1999)**, pp. 264-323**.**

[2] L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional Organization", Nat Rev Genet, vol. 5, **(2004)**, pp. 101-113**.**

[3] A. L. Barabási and R. Albert, "Emergence of scaling in random networks", Science, vol. 286, no. 5439, **(1999)**, pp. 509-512.

[4] H. J. Bjvrn and F. Schreiber, Analysis of Biological Networks, Wiley, **(2008).**

[5] C. R. Palmer, P.B. Gibbons and C. Faloutsos, "ANF: A fast and scalable tool for data mining in massive graphs", Proc. of the KDD-2002, **(2002)**.

[6] E. S. Schaeffer, "Survey Graph clustering, Sciencedirect: computer science review 1", **(2007)**, pp. 27-64.

[7] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs", SIAM J. Sci. Comput., vol. 20, no. 1, **(1999)**, pp. 359-392.

[8] H. Zantema, S. Wagemans, D. Bo sˇnacˇ ki, "Finding Frequent Subgraphs in Biological Networks Via Maximal Item Sets", Bioinformatics Research and Development BIRD, Communications in Computer and Information Science, Springer, vol. 13, **(2008)**, pp. 303-317.

[9] M. Hattori, Y. Okuno, S. Goto and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways", J Am Chem Soc, vol. 125, no. 39, **(2003)**, pp. 11853-11865.

[10] I. Dhillon, Y. Guan and B. Kulis, "A fast kernel-based multilevel algorithm for graph clustering", Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM. **(2005)**, pp. 629-634.

[11] J. E. Hopcroft, O. Khan, B. Kulis and B. Selman, "Natural communities in large linked networks", Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining, KDD, ACM, New York, NY, USA, **(2003)**.

[12] C. C. Jose, K. Satou and G. Valiente, "Finding Conserved and Non-Conserved Reactions Using a Metabolic Pathway Alignment Algorithm", Genome Informatics, vol. 17, no. 2, **(2006)**, pp. 46-56.

[13] KEGG. Kyoto Encyclopedia of Genes and Genome. Obtained through the Internet: http://www.kegg.com/, [accessed 13/10/2012], **(2012)**.

[14] L. Chen, K-Y. Feng, Y-D. C, K-C Chou and H-P Li, "Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition", BMC Bioinformatics, vol. 11, pp. 293, **(2010)**.

[15] M. R. Anderberg, "Cluster Analysis for Applications", Academic Press, Inc., New York, NY, **(1973)**.

[16] SIMCOMP Search (2011), http://www.genome.jp/ligand-bin/search_compound.

[17] D. J. Watts, S. H. Strogatz, "Collective dynamics of 'small-world' networks", Nature, vol. 393, pp. 440-442, **(1998)**.

[18] Y. Takahashi, H. Ohoka and Y. Ishiyama, "Structural similarity analysis based on topological fragement spectra", Adavances in Molecular Similarity, vol. 2, **(1998)**, pp. 93-104.

Authors

Muhammad Naeem is a Research scholar at department of computer science, M. A. Jinnah University Islamabad Pakistan. His research area includes machine learning, semantic computing, text retrieval, graph mining, classification and data mining.

Sohail Asghar is a Director/ Associate Professor at Arid-Agriculture University Rawalpindi Pakistan. He earns PhD in Computer Science from Monash University, Melbourne, Australia in 2006. Earlier he did his Bachelor of Computer Science (Hons) from University of Wales, United Kingdom in 1994. His research interest includes data mining, decision support system and machine learning.