

# Finding relationships among gene ontology terms in biological documents using Association Rule mining and GO annotations

Dr.B.LShivakumar<sup>1</sup>

<sup>1</sup>Professor and Head, Department of Computer Applications,  
SNR Sons College  
Coimbatore, India  
blshiva@yahoo.com

R.Porkodi<sup>2</sup>

Assistant Professor, Department of Computer Science,  
Bharathiar University  
Coimbatore, India  
porkodi\_r76@yahoo.co.in

**Abstract**— The gene ontology is an important biological ontology that provides information about gene products which are represented as a controlled vocabulary for annotating gene products with three terms such as cellular components, molecular functions, and biological processes. Many approaches like clustering and association rule mining have been developed to find out relationships between gene products. The clustering approach is the traditional method to find out the interactions between genes, but this method fails to prove in bring out the useful and in depth relationships than association rule mining method. The association rule mining using apriori algorithm gives better relationships among the different gene products in different genes. This paper provides the implementation of apriori algorithm to find out efficient relationships among gene products of different genes given in biological documents.

**Keywords**- XQuery, XML data, Aprior, Association Rule mining, Gene Ontology, Go Terms.

## I. INTRODUCTION

The biological databases generate huge volumes of genomics and proteomics data after the draft of human genome sequences in 2001. The researchers use the existing sequence information to find similar patterns of genes, proteins and derive other sequence information. The National Center for Biotechnology Information (NCBI) is one major resource that maintains public biomedical annotation databases, which are represented in different useful formats that includes Extended Markup Language (XML) format. The XML format of databases is very useful, because XML is one of the powerful languages for representing the biological data in semi structured form and also the extraction of biological entities from XML format of databases are very easy at any extent. The dataset for this work consists of biological XML documents which are downloaded from Swissprot. The recent researches focus to identify the interactions between gene products of same gene or different genes using either

clustering [1] or association rule mining. The clustering can help researchers to discover the genes that may be involved in the same biological process. Association rule mining [2 & 3] can be used to find out relationships among genes using Gene Ontology (GO) annotations, which bring deeper observation and implications among genes.

The Gene Ontology [4] consortium provides a set of structured vocabularies (GO annotations) organized in a rooted directed acyclic graph (DAG), describing the roles of genes and gene products in any organism. The GO annotations of gene ontology represents every GO term associated with one of the three ontology term, cellular component, molecular function, or biological process. Every GO term in a gene is associated with the unique ontology term. The GO's molecular function describes the biochemical activity of the gene product, for example: *ATP binding, protein binding or electron carrier activity*. The biological process describes the biological goals to which the gene product contributes, for example: *oxidation reduction, ethanol oxidation, or glucose metabolic process*. The GO's cellular component refers to the location in the cell where the gene product exerts its activity, for example: *cytoplasm, membrane, or mitochondrion*.

In our work, we downloaded GO annotations from the GO website (<http://www.geneontology.org/>). The GO terms specified in every gene in the biological document are annotated using the downloaded GO annotation structure. The Table 1 shows the sample GO annotations along with equivalent GO terms for the gene 'ALDH9A1'.

TABLE 1. SAMPLE GO TERMS ANNOTATIONS FOR 'ALDH9A1'

GO TERMS ANNOTATIONS AND ONTOLOGY FOR 'ALDH9A1'		
GENE NAME		
GO TERM	ANNOTATION	ONTOLOGY
5737	'cytoplasm'	Cellular Component
5856	'cytoskeleton'	Cellular Component
5634	'nucleus'	Cellular Component

4029	'aldehyde dehydrogenase (NAD) activity'	Molecular function
19145	'aminobutyraldehyde dehydrogenase activity'	Molecular function
6081	'cellular aldehyde metabolic process'	Biological process
42445	'hormone metabolic process'	Biological process
55114	'oxidation reduction'	Biological process
42136	'neurotransmitter biosynthetic process'	Biological process

The relationships between the three ontologies of a gene or group of genes are derived using association rules between the entities present within the separated ontologies of the gene ontology. Such rules will discover how much possible a particular molecular function is associated with a particular biological process? and also how much possible a particular molecular function is associated with a particular cellular component?, for example: the molecular function *ATP binding* is closely associated with *chromosome segregation* biological process and both are take place in side the nucleus cellular component. This example focuses the deep association between nucleus, ATP binding and chromosome segregation.

Association rule mining [5] was first introduced by Agrawal et al. (1993). This is used to find interesting relationships or correlation among attributes in a large database. The results of the association rule mining are similar to that of clustering, which are groups of data with similar relations. For example, in the market basket analysis, association rule mining is used to discover itemsets which are frequently purchased together. The frequent itemsets provide useful knowledge for shop keeper about the purchasing habit of the consumers. The association rules in frequent itemsets may be large in numbers, which may include good as well as bad rules. These rules can be filtered so that only good rules can be taken for consideration to take decisions. This is possible with two factors: support and confidence. This paper extracts the different association rules between three ontologies associated with every gene or gene product to identify the better relations among gene products.

This paper is organized as follows: In section 2, related research work is described. Section 3 provides the description of proposed framework of this work. Section 4 describes results and discussions and finally the paper is concluded in section 5.

## II. RELATED WORK

Many Algorithms have been developed so far in data mining such as clustering, classification and association analysis [6, 7, 8 & 9] and applied in many fields such as bioinformatics, physics and engineering to extract essential information from large number of data. Association rule mining proved that this is the important method in information extraction to extract relationships between various entities in biomedical documents of XML format. An algorithm to

construct a frequent treeby finding common sub trees embedded in the heterogeneous XML data proposed in paper [10]. The XMINE operator [16] has been introduced for extracting association rules from XML documents, where mapping the XML data to a relational structure is required before mining is performed. However, the recent works reported that the association rules can be mined directly from XML data without converting it into relational form.

The paper [11] proposed the mining of association rules from XML data using XQuery. In paper [12], usage of XQuery for mining association rules from MusicXML data is presented. In paper [13], association rule mining for XML data is implemented using java based apriori algorithm. In paper [14], the dependence relationships between gene ontology terms based on TIGR gene product annotations were identified. In paper [2], mining gene expression databases for association rules was implemented.

Recently the researchers started to work in microarray data, in paper [1], researchers developed an interactive gene clustering for cancer microarray data. In paper [16] researchers focused on mining multilevel gene association rules from micro array data set and they proved that the clustering fails to bring out the useful and in depth relationships than association rule mining method. This is the big motive for us to carry out this work. In this paper, the proposed framework use XML format of biological documents is given as input and processed using XQuery. All the facts that are specified in above literatures are studied and finally produced all possible association rules among the three separate ontologies.

## III. PROPOSED FRAMEWORK

The framework for finding relationships among gene ontology terms in biological documents using association rule mining and GO annotations is shown in Figure 1 Consists of four phases: DB2 XQuery, PreProcessor, Data encoder and Association rule generator.

### A. DB2 XQuery

This phase is used to store the actual XML biological data files into DB2 data base, in which each XML transaction data file as shown in Figure 2 is uniquely identified with transaction identifiers. The features are extracted from the DB2 database using XQuery features which offers easy access to the entire XML data or part of XML data. The key filter interface developed in DB2 XQuery provides the way to extract the necessary fields such as gene name and associated go terms in every biological XML file for mining association rules.

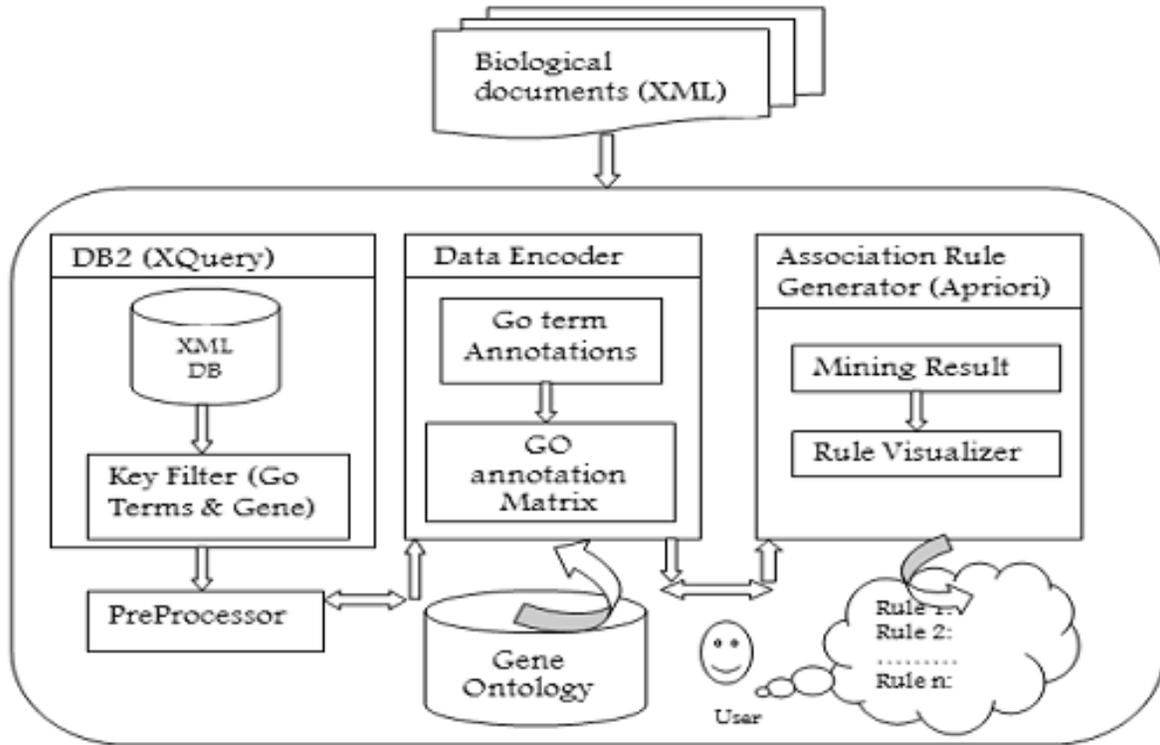


FIGURE 1. THE FRAMEWORK OF ASSOCIATION RULE MINING USING GENE ONTOLOGY AND APRIORI ALGORITHM

```

<?xml version="1.0" ?>
<Bioseq-set>
<Bioseq-set_seq-set>
<Seq-entry>
<Seq-entry_seq>
<Bioseq>
<Bioseq_id>
<Seq-id>
<Seq-id_swissprot>
<Textseq-id>
<Textseq-id_name>AL9A1_HUMAN</Textseq-
id_name>
<Textseq-id_accession>P49189</Textseq-
id_accession>
<Textseq-id_release>reviewed</Textseq-id_release>
<Textseq-id_version>3</Textseq-id_version>
</Textseq-id>
</Seq-id_swissprot>
</Seq-id>
<Seq-id>
<Seq-id_gi>62511242</Seq-id_gi>
</Seq-id>
.....
.....
    
```

FIGURE 2. A SAMPLE XML BIOLOGICALFILE

### B. Preprocessor

The preprocessor phase takes all biological XML file stored in database and apply series of processes to extract gene names and its associated GO annotations or terms along with ontology and names as shown in Table 2. The ontology terms and annotations for the gene 'ALDH9A1' is shown in Table 1. And then compute overall distinct in go terms as well as gene names which are the primary sources used for identifying the correct relationships and also to avoid redundancy among the data. Finally all distinct data is represented in a table as an input for the next phase.

### C. Data Encoder

The data encoder is used to encode the data with respective GO ontology description. The encoder computes the encoded binary array for every gene in such a way that the occurrence of go terms present in each file is coded with 1 against the corresponding go term in total distinct go terms as shown in Table 3. Before computing encoded array, first the Go term annotations must be computed for each Go term. The Go term annotations are computed for the available Go terms in each file by using the Gene Ontology structure which is available in online. For example, the gene 'ALDH9A1' has 9 Go terms in its document and the Go term annotations for the

above Go Terms are computed using Gene Ontology structure as shown in Table 1.

**TABLE 2: THE STRUCTURE OF DATA PROCESSED FROM DATASET**

Did	Gene name	GoTerms	Ontology	Go Desc.
1	'ALDH9A1'	[5737;5856;5634;4029;	<9x1 cell>	<9x1 cell
2	'ADH7'	<11x1 double>	<11x1 cell	<11x1 cell
3	'ACADSB'	[5739;3995;6631]	<3x1 cell>	<3x1 cell
4	'ACADSB'	[5739;3995;6631]	<3x1 cell>	<3x1 cell
5	'hcaB'	[18498;19439;55114]	<3x1 cell>	<3x1 cell
6	'hcaB'	[18498;19439;55114]	<3x1 cell>	<3x1 cell
7	'HADH'	[5759;3857;50662;6631]	<5x1 cell>	<5x1 cell
8	'ALDH5A1'	<17x1 double>	<17x1 cell	<17x1 cell
9	'ADH5'	[5737;4022;9055;5504;	<8x1 cell>	<8x1 cell
12	'poxB'	[5829;5886;8289;287;5	<9x1 cell>	<9x1 cell
13	'adhE'	[5829;5739;8774;4022;	<9x1 cell>	<9x1 cell
18	'TGFB1'	[5615;5178;8283;7162;	<5x1 cell>	<5x1 cell

After completing Go annotations for every gene in biological XML document using Gene Ontology, the encoded Go annotation array is constructed for the occurrence of Go term annotation against with the total distinct Go Term annotations collected from all biological documents as shown in Table 3.

Gene Names	Go Terms				
	Go ann1	Go ann2	Go ann3	...	Go ann <sub>m</sub>
G1	0	0	1	...	1
G2	0	1	0	...	0
G3	1	0	1	...	0
...	...	...	...	...	...
G <sub>n</sub>	1	1	0	...	0

**D. Association Rule Generator**

The association rule generator phase consists of two important components: mining result and rule visualizer. The mining of association rules is done using the implementation of apriori algorithm and finally the generated association rules are visualized in XML format.

After completing Go annotations for every gene in biological XML document using Gene Ontology, the encoded Go annotation array is constructed for the occurrence of Go term annotation against with the total distinct Go Term annotations collected from all biological documents as shown in Table 3.

**E. Association Rule Generator**

The association rule generator phase consists of two important components: mining result and rule visualizer. The mining of association rules is done using the implementation of apriori algorithm and finally the generated association rules are visualized in XML format.

1)

**Association Rule mining**

Association rule mining was first introduced by Agrawal et al. [5] for market basket analysis. The first step in association rule mining is to identify frequent sets, the sets of items that occur together often to investigate further. Consider a database D contains a set of transactions T that includes gene name and Go terms in each biological XML file, and each transaction consists of one or more items called itemsets, which is shown in Table 3. The itemset I=*i*<sub>1</sub>, *i*<sub>2</sub>, *i*<sub>3</sub>... *i*<sub>n</sub>, where I is a set of n distinct items, and a set of items such that T ⊆ I. An association rule is of the form A ⇒ B where A ⊆ I, B ⊆ I, and A ∩ B = ∅. The set of items A is called antecedent and the set B the consequent. The rules are considered to be good if they satisfy some additional properties, which are most important properties used in association rule mining: support and confidence.

Support *s* for a rule A ⇒ B, denoted by *s* (A ⇒ B), is the ratio of the number of transactions in D that contain all the items in A ∪ B to the total number of transactions in D. That is,

$$s(A \Rightarrow B) = \frac{\sigma(A \cup B)}{|D|}$$

Where the function σ of a set of items A denotes the number of transactions in D that contain all the items in A. σ(A) is also called the support count of A. Confidence *c* for a rule A ⇒ B, denoted by *c* (A ⇒ B), is the ratio of the support count of A ∪ B to that of the antecedent A.

$$c(A \Rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

TABLE 3. ENCODED ARRAY FOR GENE NAMES AND GO ANNOTATION TERMS

For a user-specified minimum support *s*<sub>min</sub> and minimum confidence *c*<sub>min</sub>, the task of association rule mining is to extract the association rules that have support and confidence greater than or equal to the *s*<sub>min</sub> and *c*<sub>min</sub> values from the given dataset D.

A set of items is referred to as an itemset. An itemset that contains *k* items is a *k*-itemset. A *k*-itemset L<sub>k</sub> is called frequent if L<sub>k</sub> ≥ *s*<sub>min</sub> × |D|. Such a *k*-itemset is also referred to as a frequent *k*-itemset. A frequent 1-itemset is simply called frequent items. The task of mining association rules from a collection of data is divided into two steps:

1. To find all frequent itemsets satisfying *s*<sub>min</sub>.

2. Generate association rules from the frequent itemsets satisfying  $s_{min}$  and  $c_{min}$ .

The generation of frequent itemsets is implemented using apriori algorithm. Apriori algorithm starts from large 1-itemsets and then extends one level up in every pass until all large itemsets are found. There are three operations to be performed for each pass, say pass k are given below.

1. Append the large (k-1) item
- 2.
3. sets to L.
4. Generated the potential large k-itemsets using the (k-1) itemsets. Such potential large itemsets are called candidate itemsets. The candidate generation procedure consists of two steps.
  - a. Join step is used to generate k-itemsets by joining  $l_{k-1}$  with itself.
  - b. Prune step is used to remove the itemsets X generated from the join step, if any of the subsets of X is not large. Since any subset of a large itemsets must be large.
5. Select the itemsets X from the candidate itemsets where  $support(X) \geq c_{min}$ .

### 2) Mining Result

Apriori algorithm is used to compute large itemsets that describes relationships between the cellular component, molecular function and biological process in all genes given in the documents and all association rules are generated for the large itemsets that satisfies the minimum support and minimum confidence.

### 3) Rule Visualizer

. This provides the visualization part of the generated association rules. Here the rules are visualized in XML format,

## IV. RESULTS AND DISCUSSION

This framework is tested on a data set that includes 150 biological XML documents, constructed from the human data set which is downloaded from the swissprot.

Here we present some of our association rules that describes relationships between three ontologies associated with genes which are extracted from biological XML documents. The detail of the data set is described in table 4 and also shown in Figure 3.

TABLE 4: DETAILS OF BIOLOGICAL DATASET USED IN THIS WORK

DATASET DETAILS	COUNT
No. of Documents in experimental dataset	150
Avg. No. of Go terms in each document	3
Maximum No. of Go terms in a document	21
Minimum No. of Go terms in a document	2
No. of Go terms in experimental dataset	634
No. of distinct Go Terms in experimental dataset	238
No. of Cellular components in experimental dataset	24
No. of Molecular functions in experimental dataset	103
No. of Biological processes in experimental dataset	111

The test data set contains 634 Go terms in which only 238 Go terms are unique, that are taken in to consideration for this work. The ratio of the three ontologies such as cellular components, molecular functions and biological processes out of 238 unique Go terms in the data set is represented in Figure 4.

The relationships between the 238 Go terms are identified as well as the relationships among the different molecular functions and the different biological processes are separately identified. We present some of our extracted association rules that describe the relationships between Go terms which are extracted from our data set.

```

<Item>magnesium ion binding, isocitrate dehydrogenase (NAD+)
activity NAD or NADH binding</Item>
<Rule1>magnesium ion binding->0.42, NAD or NADH binding->0.42
</Rule1>
<Item>ethanol binding, receptor antagonist activity</Item>
<Rule2>ethanol binding->0.5, receptor antagonist activity->0.5
</Rule2>
    
```

TABLE 5. RELATIONSHIPS AMONG MOLECULAR FUNCTIONS

The rules in table 5 describe the correlation among molecular functions. The Rule1 specifies relationships between *magnesium ion binding* and *isocitrate dehydrogenase(NAD+) activity*. The *magnesium ion binding* molecular function and *isocitrate dehydrogenase(NAD+) activity* has 42% chance to come with each other. The Rule2 specifies the relationship between *ethanol binding* and *receptor antagonist activity*. Each function has 50% chance to come with each other.

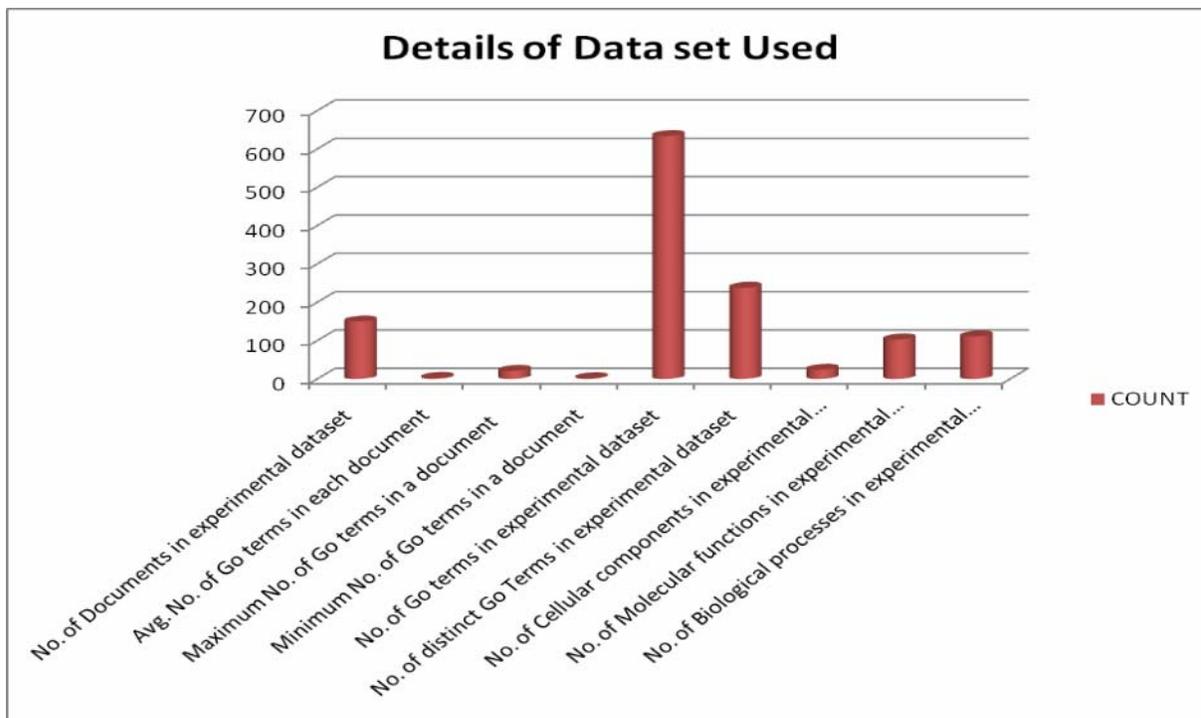


FIGURE 3. DETAILS OF DATA SET USED IN THIS WORK

The rules in table 6 describe relationships among several biological processes which are extracted from biological documents. The rule1 describes the relationship between *terpenoid biosynthetic process* and *phosphorylation process*, each process has 50% chance to come with each other, i.e the 50% possibility is predicted to have *terpenoid biosynthetic process* along with *phosphorylation process* and vice versa.

```

<Item>terpenoid biosynthetic process, phosphorylation</Item>
<Rule1>terpenoid biosynthetic process->0.5, phosphorylation->0.5
</Rule1>
<Item>respiratory system process, peptidyl-cysteine S-nitrosylation positive
regulation of blood pressure, formaldehyde catabolic process, oxidation
reduction</Item>
<Rule2>oxidation reduction->0.82</Rule2>
<Item>respiratory system process, peptidyl-cysteine S-nitrosylation positive
regulation of blood pressure, response to nitrosative stress, oxidation
reduction</Item>
<Rule3>oxidation reduction->0.82</Rule3>
<Item>respiratory system process, peptidyl-cysteine S-nitrosylation,
formaldehyde catabolic process, response to nitrosative stress, oxidation
reduction</Item>
<Rule4>oxidation reduction->0.82</Rule4>
    
```

TABLE 6. RELATIONSHIPS AMONG BIOLOGICAL PROCESSES

The remaining rules in table 6 describes the relationship between *oxidation reduction* biological process and the three other biological processes, in which we identified that *oxidation reduction* biological process is a dominating process which occurs in almost all genes. The association rules with oxidation reduction are described in rule2, rule3 and rule4.

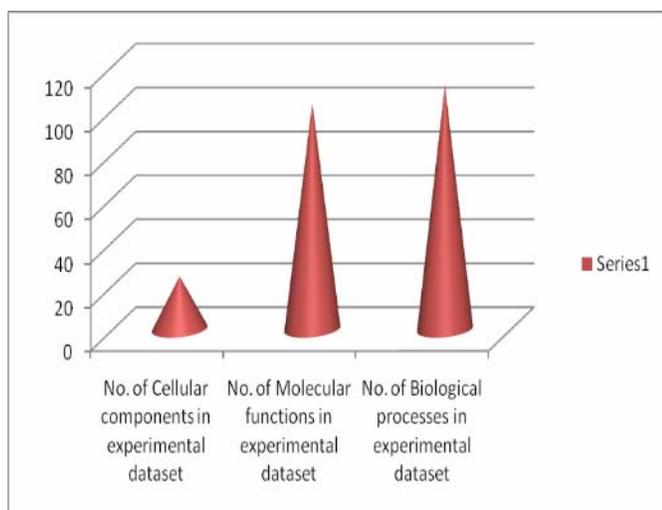


FIGURE 4. THE RATIO OF THREE ONTOLOGIES

```

<Item>retinol dehydrogenase activity[MF] ethanol oxidation[BP] retinol
binding[MF]</Item>
<Rule4>retinol dehydrogenase activity[MF]->0.3 ethanol oxidation[BP]->0.5
retinol binding[MF]->0.2</Rule4>
<Item>mitochondrial alpha-ketoglutarate dehydrogenase complex[CC] carboxy-
lyase activity[MF] oxidation reduction[BP]</Item>
<Rule5>oxidation reduction[BP]->0.9</Rule5>
    
```

TABLE 7a. RELATIONSHIPS AMONG MULTIPLE ENTITIES

The rules in table 7a and 7b describe the relationships among multiple entities such as cellular component, molecular function and biological process. This shows the possibility of occurrences like which are the entities can occur together to stimulate any biological activities in a human body. For example in rule1 of table 7a, *cytosol* is a cellular component in which molecular function *IML dehydrogenase activity* will take place, that is the possibility of occurrences of *IMP dehydrogenase activity* take place inside the cellular component *cytosol* is 78%, but at the same time this molecular function is stimulated also in some other cellular components. The rule3 of table 7a describes the relationship among cellular component, molecular function and biological process. This rule shows that all have equal chance of possibilities to come together.

```

<Item>IMP dehydrogenase activity[MF], cytosol[CC]</Item>
<Rule1>IMP dehydrogenase activity[MF]->0.22 cytosol[CC]> 0.78 </Rule1>
<Item>alcohol dehydrogenase (NAD) activity[MF], mitochondrion[CC]</Item>
<Rule2>alcohol dehydrogenase (NAD) activity[MF]->0.45 mitochondrion[CC]-
>0.55</Rule2>
<Item>3-methyl-2-oxobutanoate dehydrogenase (2-methylpropanoyl-
transferring) activity[MF], mitochondrial alpha-ketoglutarate
dehydrogenase complex[CC], branched chain family amino acid catabolic
process[BP]</Item>
<Rule3>3-methyl-2-oxobutanoate dehydrogenase (2-methylpropanoyl-
transferring) activity[MF]->0.33 mitochondrial alpha-ketoglutarate
dehydrogenase complex[CC]->0.33 branched chain family amino acid
catabolic process[BP]->0.33</Rule3>
    
```

TABLE 7B: RELATIONSHIPS AMONG MULTIPLE ENTITIES

The rule5 of table 7b shows that again the triangular relationship among the entities. This shows that the biological process *oxidation reduction* has 90% possibility to occur along with the cellular component *mitochondrial alpha - ketoglutarate de hydrogenase complex* and the molecular function *carboxylyase activity*.

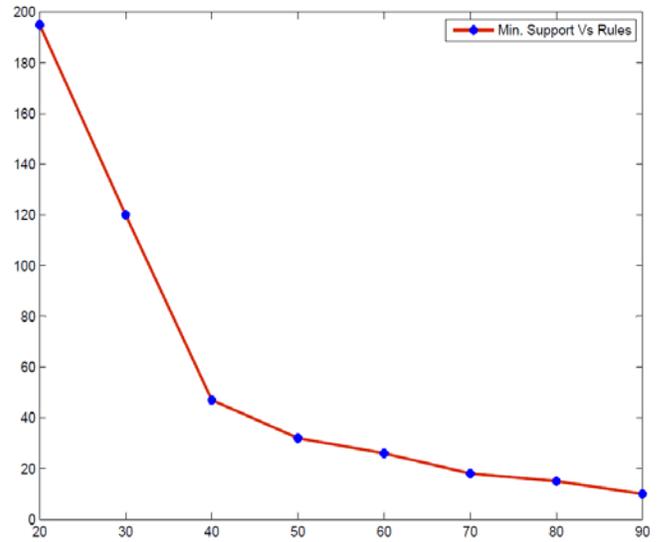


FIGURE 4. MIN. SUPPORT VS NUMBER OF RULES

The Figure 4 shows the number of rules generated for each minimum support and confidence threshold values. The minimum support and minimum confidence are used to filter out the good rules. Here both thresholds are varying and the graph shows that the number of itemsets produced is increased when the minimum support and confidence are decreases. When both thresholds are increased, it produces minimum number of rules in its itemsets, but it gives better results.

The Figure 5 shows that how much time is taken for producing rules in all itemsets. The number of rules produced is based on both thresholds and the time factor is either increased or decreased based on the number of rules. It is observed that the time taken to generate the large item set for XML data set is increased when the support count decreases.

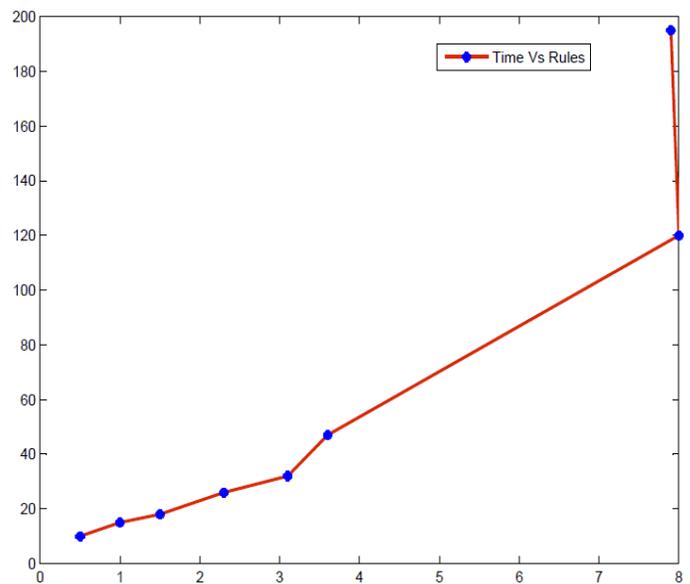


FIGURE 5. TIME VS NUMBER OF RULES

In figure 6, the number of itemsets produced is based on the minimum support threshold. It is observed that the number of itemsets produced is decreasing when the minimum support is increases.

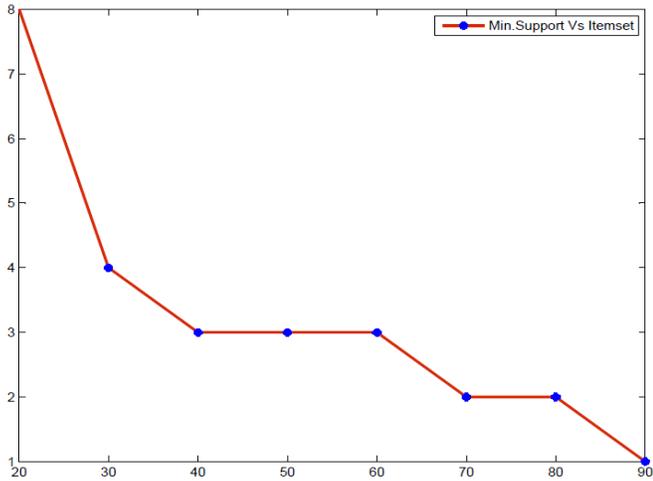


FIGURE 6. MIN. SUPPORT VS ITEMSETS

The Figure 7 is similar to Figure 4, but here the minimum support is kept constant and varying the minimum confidence. However, it shows that again the number of rules generated is decreases when the minimum confidence threshold is increases. The Figure 8 is similar to Figure 5 that show how the time factor is varying with respect to the number of rules produced when the minimum confidence is increasing. When both thresholds are used, it is observed that the better association rules

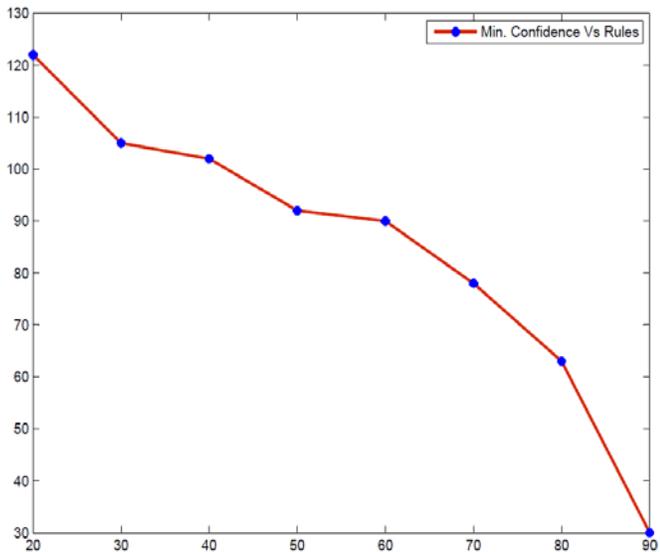


FIGURE7. MIN. CONFIDENCE VS NUMBER OF RULES

are produced than using only one threshold value as either minimum support or minimum confidence.

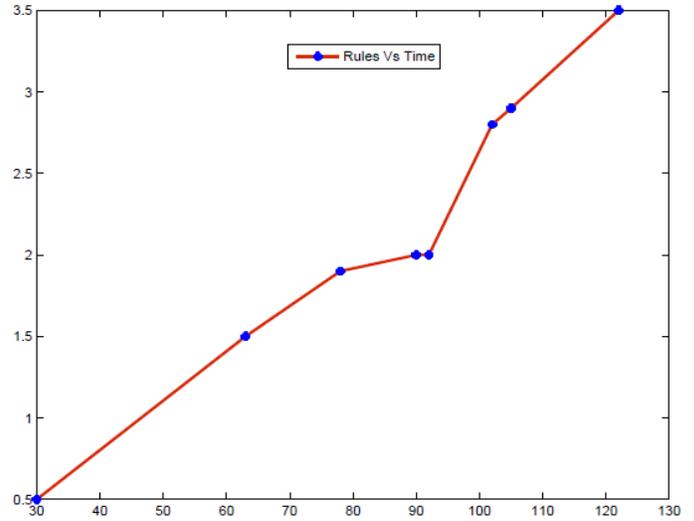


FIGURE 8. TIME VS NUMBER OF RULES

## V. CONCLUSION

This paper presents an efficient framework for extracting relationships between gene ontology terms in biological documents using association rule mining and GO annotations. This may be useful for the biologists to arrive any kind of decisions in their research in gene prediction and identification of diseases in their respective area. This provides set of possible rules for every gene product which may be useful at the time of predicting the gene expression patterns and extracting the relationships between gene products. The associations of various GO terms are grouped by prior biological knowledge which is organized in the form of GO annotations, that it proves that the association rules produced by our system are good and this may be referred by any future research in this area.

## Acknowledgment

This work was performed as part of the Minor Research Project, which is supported and funded by University Grants Commission, New Delhi, India.

## REFERENCES

- [1] Gruzda A. Inhatowicz A. & Zak D. (2006). Interactive gene clustering – a case study or breast cancer microarray dat. Information systems frontiers.
- [2] Creighton C. & Hanash S. (2003). Mining gene expression databases for association rules. Bioinformatics, 19, 79-86.
- [3] Carmona-saez P. Chagoyen M. & Rodriguez A., Trelles O., Carazo J.M. & Pascual-Montano A. (2006). Integrated

analysis of gene expression by association rules discovery. *BMC Bioinformatics*, 7(54), 1-16.

[4] Gene Ontology Consortium. Creating the Gene Ontology. *Genome Research* 2001. 11:1425-1433.

[5] R. Agrawal, T. Imielinski & A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, D.C., USA, May 1993, pp. 207-216.

[6] Agrawal R. & Srikant R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20<sup>th</sup> International Conference on Very Large Databases*, Santiago, Chile, 487 – 499.

[7] Han J. & Fu Y. (1995). Discovery of multiple-level association rules from large databases. *Proceedings of the 21<sup>st</sup> International Conference on Very Large Databases*, 420 – 431.

[8] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey (2008). A Text Mining Technique using Association Rules Extraction, *International Journal of Computational Intelligence* 4: 1, 2008.

[9] Kotlyar M. & Jurisica I. (2006). Predicting protein-protein interactions by association mining. *Information systems frontiers*, 8(1), 37-47.

[10] A. Termier, M.-C. Rousset, and M. Sebag. Mining XML data with frequent trees. In *DB Fusion Workshop'02*, pages 87-96.

[11] Jacky W.W. Wan Gillian Dobbie, Mining Association Rules from XML Data using XQuery, The University of Auckland, Private Bag 92019, Auckland, New Zealand.

[12] Yu-Chih Shen, Jia-Lien Hsu and Shuk-Chun Chung, MF-tree: Extracting and Clustering the Structural Features from Music Object in MusicXML, Department of Computer Science and Information Engineering, Fu Jen Catholic University, Taiwan.

[13] Qin Ding and Gnanasekaran Sundarraj (2005). Association Rule Mining from XML data, Computer Science Program, The Pennsylvania State University at Harrisburg, Middletown, PA 17057, USA.

[14] Anand kumar, Barry smith & Christian borgelt (2004). Dependence relationships between Gene ontology terms based on TIGR Gene product annotations, *Computerm 2004-3rd International Workshop on Computational Terminology*, Geneva, Switzerland.

[15] Vincent S. Tseng, Hsieh-Hui Yu & Shih-chiang Yang (2009). Efficient mining of multilevel gene association rules from microarray and gene ontology, *Information System Frontiers*, Vol. 11(4), 433-447.

[16] D. Braga, A. Campi, M. Klemettinen, and P. L. Lanzi. Mining association rules from xml data. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2002)*, September 4-6, Aix-en-Provence, France 2002.

**B.L.Shivakumar** received Ph.D. in Computer Science from Bharathiar University, Coimbatore. M.Phil. in Computer Science from Manonmaniam Sundaranar University, in 2003 and M.Sc. in Computer Science from Bharathidasan University, in 1996. He also received Post Graduate Diploma in Business Administration (PGDBA), Co-operative Management (PGDCM) and Bachelor of Library and Information Science from Annamalai University. In 1997, he joined SNR Sons College as a Lecturer in the department of Computer Science, and currently is the Professor and Head of the department of Computer Applications. He has authored or co-authored over 15 Research Papers in journal and conferences. He is recipient of Bharat Jyoti award conferred by The India International Friendship Society, New Delhi and Best Programme Officer award by Bharathiar University. His interest includes Computer Forensic Science, Digital Image Processing and Cloud computing.

#### **Co-Author**

**Ms R. Porkodi** received the Bachelors Degree (B.Sc.) in Computer Science, Masters Degree (MCA) in Computer Applications in 1995, 1998 respectively from Madurai Kamarajar University, India. She has qualified UGC-NET, for Lectureship in the year 2000. She is pursuing her doctoral research at Bharathiar University in the area of Text mining and Natural Language Processing. Her research interest includes Text Mining, Information Retrieval and Bioinformatics and she is currently working in the Department of Computer Science and Engineering, Bharathiar University, India.