Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Mining Disease Integrated Ontology

Taysir Hassan A. Soliman
Associate Professor,
Information Systems Dept.,
Faculty of Computers & Information, Assiut University, Egypt
taysirhs@yahoo.com

Marwa Hussein
Assistant Lecturer
Information Systems Dept.,
Faculty of Computers & Information, Assiut University, Egypt
marwa.hussien@gmail.com

Mohamed El-Sharkawi
Professor,
Information Systems Dept.,
Faculty of Computers & Information,
Cairo University, Egypt
m.elsharkawi@fci-cu.edu.eg

***Abstract*- Ontology has become a very vital issue to solve important issues regarding human diseases through data integration of chemical and biological data. Mining such data discovers highly important knowledge about diseases can give an important insight to arrive to new drug targets and assist in personalized medicine. In the current paper, a mining technique for diseases is developed based on integrated ontology and association rule mining algorithm. To perform mining, the semantic web, as a knowledge representation methodology is used to integrate data. In addition, an Ontology Association Rule Mining algorithm (OARM) is developed since existing algorithms cannot be applied because of the ontology nature of data containing several types of relations. To test our performance, prostate cancer data is obtained from NCI, which is related to 279 genes and 89 genes (from prostate cancer pathway).**

***Keywords*- Semantic Web, Gene Ontology, Disease-related information ontology, Association Rule Mining.**

I. INTRODUCTION

Human Diseases are often caused by mutations in gene networks or their products that are working together to keep a cell in a healthy state. These networks are often referred to as disease pathway. Information about diseases, their pathways and related genes are considered a vital potential for disease mechanism understanding and could also help in development of better treatments for these diseases. Fortunately, relevant data sources about diseases are currently distributed across a wide range of disparate, large-scale, publicly available databases, web sites and repositories and are described using a wide range of taxonomies and ontologies. Efficient integration and analysis of these data sets are required in order to reveal previously undiscovered interactions and pathways that will lead to the discovery of new drugs. Representation of disease information via utilizing the semantic web as a knowledge representation technology can help discover useful hidden information.

A main objective of the semantic web is to add semantics to the current web, by designing ontologies which explicitly describe and relate objects using formal and logic-based representations [2]. Ontologies are used to facilitate knowledge sharing and reusing by explicitly defined a finite set of concepts (classes) and relations between them for a specific domain [1]. The Open Biomedical Ontology [9] is a shared portal, which contains a large number of biomedical/ biological ontologies, including ontologies for managing medical terms (i.e. MeSh[1], and UMLS[2]). One of the early developed ontologies and can be considered the most important one is the Gene Ontology (GO). It can be divided into three sub-ontologies that describe gene products from different views, which are their associated biological processes, cellular components, and molecular functions in a species-independent manner. Another important ontology in the field of biochemistry is the CHEBI ontology, which is considered as being an annotated controlled vocabulary for biochemical compounds to promote the correct and consistent use of unambiguous biochemical terminology in molecular biology [6].

Recently, ontologies have been widely used as semantic web knowledge bases for sharing and querying. This can be achieved by first developing a hierarchy of classes, adding object relations between them (Ontology design). Second, creating concrete instances, also named individuals, of those classes (Ontology population) is performed. Text mining,

[1] http://www.ncbi.nlm.nih.gov/mesh
[2] www.nlm.nih.gov/research/umls/

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

ontology alignment, ontology merging and ontology mapping, relational databases' files, or combination of them are different methods that can be used in building semantic knowledge bases [4]. The main purpose of those semantic knowledge bases is that they can facilitate searching those knowledge bases, using expected relations rather than a set of keywords [1,4].

In the current paper, a disease mining technique is developed based on integrated ontology and association rule mining. This technique can be considered as a semantic web knowledge base containing a large amount of information about human diseases from a biological point of view, such as their associated genes, chemical compounds, enzymes, etc. Then, we use GO as a concept hierarchy for Ontology Association Rules Mining (OARM) Algorithm to discover interesting relations among diseases' related genes. The paper is organized as follows: section two describes related work; section three explains the proposed technique. Section four clarifies the results and section five gives the conclusions and future work.

II. RELATED WORK

Ontology usage research is still an area that needs to be explored, as semantic web technology, for biological data representation and sharing. Most of the work in this area is concerned with integrating different types of data from various sources in a specific domain in one adequate form (i.e. OWL) and then exploit this knowledge, using RDF queries. yOWL [11] describes a first approach to logically describe, integrate and query yeast biological data from the Saccharomyces Genome Database (SGD database) using the OWL-DL ontology language. By mapping the KEGG pathway database and GO annotations there was an attempt to classify diseases and identify taxonomic relations between them [5]

Mining diseases information research mainly depends on text mining medical publications [6]. While other attempts to uncover gene-disease relationships that are not directly stated [7], was based on a computational method that combines data extracted from the literature with data from public databases. Previous research in concept level mining research can be found at [8,9], concerning mining association rules from traditional relational databases with the use of ontologies to refine and improve the resulting rules. However, Tseng *et al*. [10] proposed a new algorithm called, MAGO for discovering the multilevel gene association rules from the gene microarray data and the concept hierarchy of Gene Ontology (GO). It can efficiently find out the relations between GO terms by analyzing the gene expressions with the hierarchy of GO. For example, with the biological process in GO, some rules like Process A (up) → Process B (up) can be discovered, which indicates that genes involved in Process B of GO are likely to be up-regulated whenever those involved in Process A are upregulated.

III. DESCRIPTION OF THE PROPOSED MINING APPROACH

The proposed work is composed of two main phases: 1) data integration and ontology development phase and 2) data mining phase, as shown in Fig. 1. The objective of the first phase is to integrate data related to human diseases from heterogeneous resources and develop the Disease-Related Information Ontology (DRI-Ontology). The objective of the second phase is to mine the developed ontology to discover hidden knowledge regarding diseases. In the next sections, each phase is explained in details.

1. Data Integration and Ontology Development

The objective of this phase is to integrate different types of data sources to develop our semantic knowledge base about human diseases. The DRI-Ontology was designed following the theoretical ontology design methodologies and tips [3]. There is no correct method for the development of ontology, but a variety of these methodologies exists. It is an iterative process, and may vary according to the application being developed. In the following sub-sections, we will discuss the steps for developing DRI-Ontology.

A. Data Selection

In order to cover the scope of the DRI-Ontology, various data sources concerning human diseases, pathways, genes, chemical compounds, proteins and enzymes are selected. This is an arbitrary choice intended to take the advantage of the variety of available data and the way these data are processed. Selected data sources are a collection of biological relational databases and a number of available biological ontologies in OWL format. Table I contains a list of the biological databases that are selected to be integrated in the DRI-Ontology and

their format. KEGG[3] (Kyoto Encyclopedia of Genes and Genomes) is an integrated database resource consisting of 16 main databases, broadly categorized into systems information, genomic information and chemical information. Also, the KEGG API web service facilitates accessing the desired data, without the need for developing flat files parsers. KEGG has unique advantages among other bioinformatics databases. For example, KEGG Disease contains information about molecular networks, also called disease pathway. Unlike other disease databases (i.e. OMIM) that only contain descriptive information for humans to read and understand. In addition, the KEGG Pathway database is the only database that contains information about disease pathways. Other pathways database (i.e. Reactome) only is concerned with metabolic and signal transduction pathways. Entrez Gene[4] and ChEBI databases are considered to be one of the largest repositories of genes and chemical compounds information, respectively. They both annotate genes and chemical compound with GO and ChEBI ontology, respectively, which is not in KEGG Gene or KEGG Compound databases. In addition, UniProt the largest protein database, also annotate proteins with GO ontology terms.

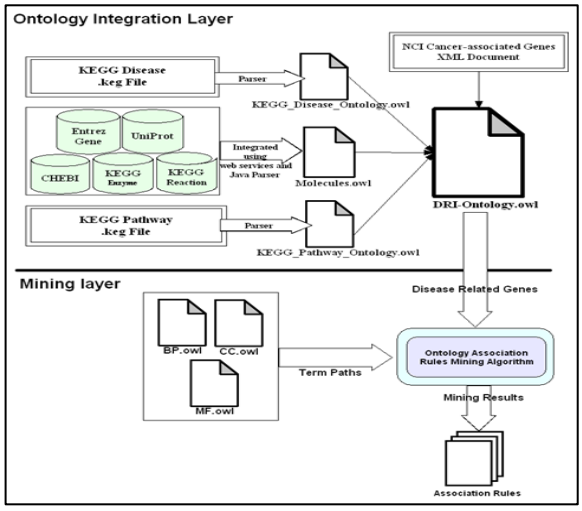

Figure 1. Block diagram describing proposed work

A. Defining Classes, Identifying Class Hierarchy, and Properties

The classes of DRI-Ontology were initially extracted from the attributes in the data sources (shown in Table I), and later on they are revised to reflect knowledge about human disease and related pathways. For instance, some of the diseases in the KEGG disease database are related to disease pathways. KEGG pathway database represents the biological pathway as an interacting network of genes, chemical compounds, and enzymes. From this, the following classes and class hierarchy are created, as shown in Fig. 2:

TABLE I: DATA SOURCES

Database Name	Data Format	Generated Ontology
KEGG Disease Hierarchy	.keg file	KEGG_Disease_Ontology.owl
KEGG Pathway Hierarchy	.keg file	KEGG_Pathway_Ontology.owl
KEGG Reactions	KEGG API web service	
Entrez Gene Database	Tab delimited file	Molecules Ontology.owl
NCI cancer associated genes	XML file	
Gene2GO	Tab delimited file	
CHEBI Database	CHEBI web service	
KEGG Enzymes Database	KEGG API web service	
UniProt Database	Uniprotjapi web service	

1. The Human_Disease class and its subclasses describing the hierarchical classification of human disease.
2. Biological_Pathway class and its subclasses describing also the different types of biological pathways (metabolic pathways, signal transduction pathways, disease pathways, etc)
3. Gene, Chemcial_Compound, Protein, and Enzyme classes are all grouped as subclasses of the Molecule class.

Object properties are also introduced to describe relations between ontology classes. DRI-Ontology classes and object properties are shown in Fig. 3. Also, several data properties are introduced to describe information associated with each class objects such as their ids, name, synonyms, etc.

A. Ontology Population

Ontology population is the process of assigning individuals to the ontology classes. A Java-based parser is developed, using the Protégé-OWL API with KEGG, UniProt, and ChEBI web services, to automatically populate individuals of the ontology classes form the relational databases listed above. An Intel Centrino computer with 4GB RAM, we are able to load the entire data-instantiated ontology using Protegé 3.4. The resulting number of instances is taken 16929 instances. In this work, a focus on Prostate cancer only is considered because of issues on reasoning performance and resource restrictions.

[3] http://www.genome.jp/kegg/
[4] http://www.ncbi.nlm.nih.gov/gene

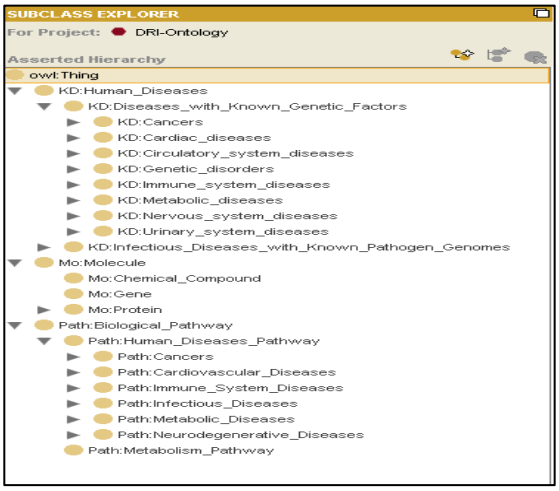

Figure 2. DRI-Ontology classes and class hierarchy

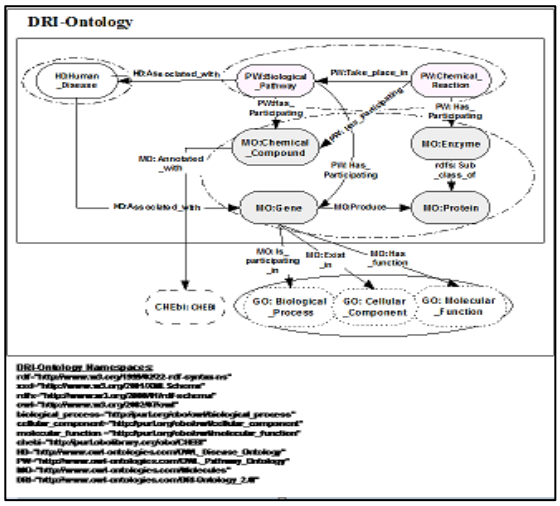

Figure 3. DRI-Ontology classes and properties diagram

2. MINING LAYER

A. Transaction Generation

To generate transactions, we have to query genes that are associated to a specific disease in DRI-Ontology. Genes may have a direct relation to disease through (KD: associated_with relation) or indirect relation through the diseases' pathways. Each gene can be annotated with terms from any of the three GO ontologies or terms from all of them. To apply the association rules for mining disease related genes, each gene is considered related to a given disease as a transaction. In a transaction, each gene is annotated by the GO terms of the three GOs ontologies. Thus, three types of transactions are obtained. The three GO sub-ontologies terms are organized as nodes in directed acyclic graphs (DAG), which are directed graphs with no path starting and ending at the same node. Is a, part of, has part, regulates, +ve regulation, and –ve regulations are the types of relations that are used in the Gene ontology. Moreover, a child node in the graph may have more than one parent node. In general, a gene could be annotated by one or more GO terms from any of the three GO ontologies. Low level terms at GO DAG contain more details. For example, given a disease name, say colorectal cancer, in the DRI-Ontology colorectal cancer has 62 associated genes. Then, for each GO term that is annotated by any of the colorectal cancer genes get its path(s) to the root. For example, consider APC gene annotated with the term (GO:0001822) as a biological process term, which named (Kidney development), the term itself has two parents (GO:0072001, GO:0048513), producing two different paths. As other terms in the path(s) have more than one parent, the number of paths may grow exponentially. This situation resulted in more than one path for the term. Fig. 4 illustrates GO:0008122 location in GO Ontology and Fig.5 contains the (GO:0008122) possible paths, respectively.

In association rules mining from Gene Ontology, GO term's path(s) to the root are used to replace the original transaction table. We implement a recursive function that given a GO term gets all possible paths to the root. These paths serve as the input of the mining algorithm. Note here that the number of transactions will depend on the number of genes, the number of annotated GO terms for each gene, and the number of paths that one term may produce.

B. Ontology Association Rule Mining Algorithm

Having the transactions generated, they are considered as the input for our ontology-based association rules mining algorithm based on FP-growth approach. The main goal of the proposed algorithm is to find association rules between GO terms. Then, GO2Chebi and GO2EC mappings (found at OBO) to find hidden relations between diseases known genes and chemical compounds or enzymes.

```
Input:
T: Set of Transaction (terms's path(s) to the
root)
Min_sup: minimum support for finding frequent
1-itemSet
Min_Conf: minimum confidence value for
generating association rules
    1. Scan transactions (paths), find terms
       with frequency greater than or equal
       to a Min_sup value
    2. Order the frequent terms in
       decreasing order
    3. Construct a tree which has only the
       root
    4. Scan Transactions again; for each
       path:
            a. add the terms from the path to
               the existing tree, using only
```

the frequent terms (i.e. terms discovered in step 1.)

b. repeat (a) until all paths have been processed

5. Enumerate all co-occurred itemsets by examining the tree: the co-occurred itemsets are present in those paths for which every node is represented with the frequency ≥ 1
6. Generate Association rules that satisfied min_conf
7. Scan terms in association rules, find if they've mapping with ChEBI Ontology, if exist get the gene(s) having this transaction as one of its paths

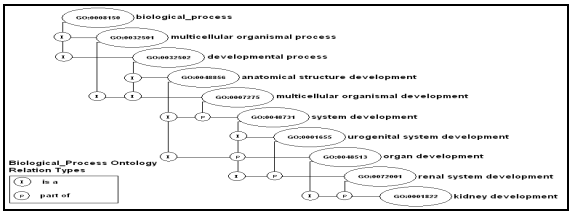

Figure 4. (GO: 001822) location in GO Ontology

Path(s) to the Root						
Path 1	GO:0001822	GO:0072001	GO:0001655	GO:0048731	GO:0007275	GO:0032502
Path 2	GO:0001822	GO:0072001	GO:0001655	GO:0048731	GO:0007275	GO:0032501
Path 3	GO:0001822	GO:0072001	GO:0001655	GO:0048731	GO:0048856	GO:0032502
Path 4	GO:0001822	GO:0072001	GO:0048731	GO:0007275	GO:0032502	
Path 5	GO:0001822	GO:0072001	GO:0048731	GO:0007275	GO:0032501	
Path 6	GO:0001822	GO:0072001	GO:0048731	GO:0048856	GO:0032502	

Figure 5. GO terms (GO: 0001822) paths to the root

IV. EXPERIMENTAL RESULTS

In the current paper, Prostate cancer disease information is used to illustrate the proposed mining technique. As illustrated in Fig. 6, 89 genes related to prostate cancer were annotated with 772 biological processes, 172 molecular functions, and 107 cellular component, non redundant GO terms. Those generated 19564, 268, and 2653 transactions, respectively. It is well known that GO ontology can be considered as a DAG with a maximum depth of 17 levels. Therefore, the maximum size of any transaction will not exceed 16 items (GO terms), as the root will be deleted because it is meaningless to be in the mining process. Thus, the number of terms in each transaction will vary according to the level that the original term, as illustrated in Fig. 7. Fig. 7 shows number of terms in the transactions for the three GO categories. We can see that in MF transactions the maximum number of terms in the transactions is 9 terms, and in CC transactions is 14 terms, while in BP transactions is 16 terms. The variation of the number of terms at each level in the three ontologies is due to the differences in their structures (depth and density).

A. Evaluation of Ontology Association Rules Mining Algorithm (OARM)

OARM algorithm is used to mine hidden relations between diseases' related genes and other enzymes or chemical compounds by using genes' annotated GO terms. OARM is implemented using a FP- Growth approach, that we are able to mine the complete set of co-ocurrent itemsets without candidate generation. It is obviously that the run time will be mainly influenced by the number of annotated terms for each gene, the location of the term in the GO Ontology, and consequently the transactions generated. As the number of genes in most of the cancer's pathways is usually a small number and the deepest level of the GO Ontology is 17, then run time is mainly influenced by the min_sup value that is used only in the first iteration to get the frequent 1-item set that will be used later for getting other co-ocurrent item sets > 1. For an experimental evaluation of the performance of OARM algorithm, different min_sup values are used, ranging from 20% to 5%, and measure amount of time, memory consumption, and number of rules generated at each value. Fig. 8 shows the mining results for using OARM algorithm with variations of the min_supp against the number of rules, running time, and memory.

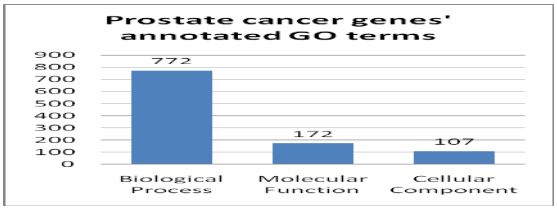

Figure 6. Prostate cancer genes' annotated with GO terms

B. Generated Association Rules

Here, we will enumerate a set of rules from mining molecular functions of the prostate cancer genes with min_sup=7.5% and min_conf=80%:

1. protein kinase binding & catalytic activity & kinase activity→transferase activity
2. transferase activity & kinase activity & transferase activity, transferring phosphorus-containing groups→catalytic activity.

Some of these terms were successfully mapped to CHEBI and EC (Enzyme committee) terms. For example, consider the following rule:

- regulation of macromolecule biosynthetic process & negative regulation of gene-specific transcription & cellular macromolecule metabolic process→regulation of cellular biosynthetic process.

The cellular macromolecule metabolic process was found to be mapped to the ChEBI term 'macromolecule'. Thus, we can conclude that because

the other three terms are in the same rule (i.e. occurred in the same path produced by some biological process that is annotated by genes associated with prostate cancers) can be related with the same ChEBI term. Thus, the following examples of prostate cancer genes are (CREBBP, ATF4, E2F1, TCF7L1, CREB3L4, RELA, AR, EP300, CREB1) that are annotated with biological processes having this rule as part of their paths to the root also can be related to CHEBI term "macromolecule".

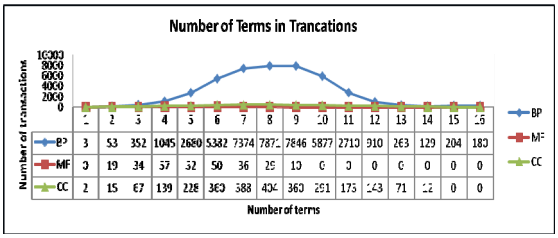

Fig. 7.Number of terms in transactions

Another example with mapping the resulted rules with EC terms, consider the following rule: phosphotransferase activity& transmembrane receptor protein kinase activity→catalytic activity & transferase activity & protein kinase activity & transmembrane receptor protein tyrosine kinase activity. Thus, examples of prostate cancer genes are (PDPK1, MAPK3, AKT1, CREBBP, BRAF, CHUK) that are annotated with biological processes which can be related to the mapped enzymes (EC:2.7.1: Phosphotransferases with an Alcohol Group as Acceptor & EC:2: transferases & EC:2.7.10.1: Receptor protein-tyrosine kinase).

C. CONCLUSIONS AND FUTURE WORK

Since most previous disease mining techniques were based on text mining, a new disease mining technique was presented in this paper. This technique is based on ontology development and association rule mining. It consists of two main phases: data integration and ontology development and association rule mining. The KEGG database was massively used because of its valuable information regarding genes, pathways, chemical compounds related to genes. Our created ontology was developed using the Protégé tool which was of great importance. Besides a new ontology mining algorithm OARM was developed based on FP-growth. Future work includes taking microarray gene expression experiments as part of our data enrichment for disease mining as well as analyzing biological networks related to diseases.

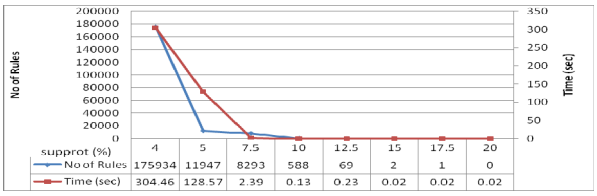

Fig. 8.Different minSupprot values in biological process

REFERENCES

[1] H. Al-Mubaid and R. K. Singh, "A text-mining technique for extracting gene-disease associations from the biomedical literature," *International Journal of Bioinformatics Research and Applications*, 6(3), pp. 270-286, 2010.

[2] G.Antoniou, and F. Harmelen. *A Semantic Web Primer* (Second ed.). The MIT Press, 2008.

[3] A. Bellandi, *et al*., "Pushing constraints in association rule mining: an ontology based approach," IADIS International Conference, IADIS, Spain, pp. 179-186, 2007.

[4] J. Chabalier, *et al.,* Integrating biological pathways in disease ontologies. *Medinfo , 129* (1), pp.791-795, 2007.

[5] O. Corcho, *et al.,* "Methodologies, tools and languages for building ontologies. Where is their meeting point?," *Data & Knowledge Engineering* , 46 (1), pp. 41–64, 2003.

[6] K. Degtyarenko, *et al*., "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Research ,* Database issue(36), 2008.

[7] G. Gonzalezy, *et al.,* "Mining gene-disease relationships from biomedical literature:weigthing protein-protein interactions and connectivity measures," *Pacific Symposium on Biocomputing*, pp. 28-39. 2007.

[8] C.Marinica and F.Guillet, "Improving post-mining of association rules with ontologies," *Applied Stochastic Models and Data Analysis: The XIII International Conference.* Vilnius, Lithuania, pp. 76-80, 2009.

[9] OBO Foundry, The Open Biomedical Ontology, www.obofoundry.org/2007.

[10] V. S. Tseng, *et al.*, "Efficient mining of multilevel gene association rules from microarray and gene ontology," *Information Systems Frontiers , 11* (4), 433-447, 2009.

[11] N.Villanueva-Rosales, and M. Dumontier, "yOWL: An ontology-driven knowledge base for yeast biologists," *Journal of Biomedical Informatics ,* 41(5), pp. 779–789, 2008.