ISSN:2320-0790

An International Journal of Advanced Computer Technology

GENE ONTOLOGY SIMILARITY METRIC BASED ON DAG USING DIABETIC GENE

S. Booma Shanthini[1], Dr. V. Bhuvaneswari[2]

M.Phil Research Scholar[1], Assistant Professor[2]
Department of Computer Applications
Bharathiar University
Coimbatore, India
boomashanthini@gmail.com[1], bhuvanes_v@yahoo.com[2]

***Abstract*— Bioinformatics and Data Mining provide exciting and challenging researches in several application areas especially in computer science. The association between gene and diseases are analyzed using data mining techniques. The objective of the paper is to study the various similarity metrics for analyzing the diabetic gene using data mining technique. This paper provides with an overview of different similarity metrics for gene clustering. A similarity metric is proposed to cluster diabetic gees based on DAG structure of gene ontology. The experimental verification is analyzed for evaluating the cluster with biological validation. The current OMIM dataset is used for the proposed work.**

Index Terms*— *Bioinformatics, Gene Ontology, Similarity Metrics, GO terms.

I. INTRODUCTION

The genomic data is represented in various data repositories for different biological entities. Data mining techniques are used in bioinformatics research for classifying and grouping the biological entnies like gene, protein and diseases. More researches is done related in associating genes and diseases. Diabetic Mellitus is a group of metabolic diseases in which a person has high blood sugar either because the body does not produce enough insulin or because cells do not respond to the insulin that is produced. Diabetic Mellitus is a disorder caused by the total (or relatives) absences of insulin, which manifests clinically as an elevated blood glucose. Diabetic Mellitus is characterized by abnormally high levels of sugar (glucose) in the blood, when the amount of glucose in the blood increases, example after a meal, it triggers the release of the hormone insulin from the pancreas. The Genetic Landscape of Diabetes introduces some of the gene that has been suggested to play a role in the development of diabetics.

Data Mining is the process of discovering meaningful, new correlation patterns and trends by shifting through large amount of data stored in repositories, using patterns recognition, statistical and mathematical techniques [3]. Data mining is also known as Knowledge Discovery in Databases (KDD). Bioinformatics is a branch of biological science which deals with the study of methods for storing, retrieving and analyzing biological data. It has been defined as a discipline that generates computational tools, databases and methods to support genomic and post genomic research. In bioinformatics, data mining is used to find patterns in sequences to calculate folding patterns, to determine genetic mechanism underlying a disease, to design ontology for multiple DNA or protein/gene sequence, and so on [3].

The Gene Ontology (GO) is the most generally used controlled terms in bioinformatics resources. The term ontology is original from Greek and is used in philosophy to mean ‘a description of what exists’. There are many definitions of the word and for the idea of work; ontology is ‘a specification of entities and their relationships [5]. The keyword ‘specification’ implies a formal organization. Ontologies for computing applications are schemas for metadata.

The ontology covers three domains Biological Process, Molecular Function and Cellular Component. A biological process is sequence of events accomplished by one or more ordered assemblies of molecular functions. It is a process of a

living organism and is made up of any number of chemical reactions or other events that result in a transformation. Molecular function describes activities, such as catalytic or binding activities, that take place at the molecular level. Cellular component is the element of a cell or its extracellular environment. The cellular component ontology explains locations, at the levels of sub cellular structures and macromolecular complexes [6].

Gene Ontology together with a set of individual instances of the kinds of entities it specifies constitutes a knowledge base. It may be difficult to distinguish between the knowledge contained in a knowledge base. Furthermore, ontologies can be used for reasoning and inference. The advantages of using ontologies have been argued extensively, but the main reason is that ontologies are attempting to capture the precise meaning of terms [2]. The word "Ontology" has been recognized in philosophy as the subject of existence. Gene Ontology is a structured and controlled vocabulary which characterizes the functional properties of gene using standardized terms. Ontologies are meant to provide an understanding of the static domain knowledge that facilitates knowledge sharing and reuse. This paper is categorized into six sections. Section 2 describes the literature review for Gene Ontology and Semantic Similarity Metrics. Section 3 shows the overview in Semantic Similarity Metric. Section 4 defines the framework and methodology to analyze the Diabetic Gene based on Directed Acyclic Graph (DAG) structure. Section 5 explains the implemented approaches with its respective results. Section 6 concludes with the comparison of the proposed approaches to cluster and biologically validated.

II. Literature Review

This section provides with an overview of Gene Ontology related to Semantic Similarity Metric for clustering Biological data.

Haixuan Yang et al., [4] (2012) have proposed the functional similarity between gene products to well-structured controlled vocabularies where biological terms are organized in a tree or in a Directed Acyclic Graph (DAG) structure. Their result shows the use of downward random leads to more reliable similarity measures. Ying Shen et al., [15] (2011) proposed the shortest path algorithm searched for the shortest path that connect two terms and uses the sum of weights on the path to estimate the semantic similarity between Gene Ontology terms. C. Tatsiopoulos et al., [11] (2009) presented a new ontology mapping technique to map concepts in one ontology without any user intervention. It is based on association rule mining concept hierarchies of the input ontologies.

Toralf Kirsten et al., [12] (2007) proposed a methodology for instance-based ontology matching which utilized the associations between molecular-biological objects and ontologies. They provided a comparison with metadata-based ontology matching. Monica Chagoyen et al., [9] (2006) used the scientific literature to establish potential relationships among cellular processes. They used a document based similarity method to compute pair wise similarities of the biological processes described in the Gene Ontology.

Stefano Bianchi et al., [10] (2009) depicted a biomedical data integration process and demonstrated for integrating various data sources containing clinical within Hyper genes. They proposed semantic data integration based on information models serving as a common language to represent the data semantics. Xiang-hua Xu et al., [14] (2008) presented the methods for measuring concepts semantic similarity. They have compared ontology based to the distance based calculation model. This method provided an effective quantification for the semantic relationships and calculates semantic similarity more precisely.

Hisham Al-Mubaid et al., [5] (2008) compared four semantic similarity measures to compute the similarity between genes using GO annotations within a Gene Clustering application. Semantic similarity measures used in many applications long time before used in the gene functional analysis applications. Jin Gou et al., [6] (2007) represented an object nodes with a hierarchical logical model using ontology technology. They clustered the nodes to a new node based on semantic relationships. It is used to describe an improved knowledge space structure where distance between nodes is quantified with structural and semantic correlation using a hierarchical structure.

III. Semantic Similarity Metric – An Overview

Semantic similarity is a concept whereby a set of documents or terms within term lists are assigned a metric of their meaning or semantic content. This can be achieved for instance by defining a topological similarity by using Ontologies to define a distance between words or using statistical means such as a Vector Space Model to correlate words and textual contexts from a suitable text corpus [15]. The various similarity measures used for Gene biological relationship is discussed.

A. Node-based (Information Content) Approach

Node based approach is to determine the conceptual similarity is called the information content approach. In this approach, the term is used for annotation, the lower its semantic value, and this may lead to different semantic values of the Gene Ontology terms for Gene Ontology (GO) annotation data derived from different sources [16]. This method contains the following measures such as, Resnik, Lin's, Jiang and Conrath measure.

B. Vector-based Approach

Vector-based methods embed ontological terms in a vector space by associating each term with a dimension. Usually a vector is binary consisting of 0's and 1's where 0 denotes the absence of a term in an annotation's.

$$sim_{\cos}(g_1, g_2) = \frac{v_1, v_2}{|v_1||v_2|} \qquad (1)$$

Where vi represents a vector of terms constructed from an annotation Gi. The source of descriptiveness, commonality and difference is the same as the situation for set-based approaches.

C. Edge-based (Distance) Approach

The edge based approach is a more natural and direct way of evaluating semantic similarity in a Gene Ontology. It estimates the distance (e.g. edge length) between nodes which corresponds to the terms/genes being compared. In a more realistic scenario, the distances between any two adjacent nodes are not necessarily equal. It is therefore necessary to consider the edge connecting the two nodes should be weighted.

Sussna (1993) considered the first three factors in the edge weight determination scheme. The weight between two nodes c1 and c2 is calculated as follows:

$$Wt(t_1,t_2)=\frac{Wt(C_1\rightarrow_r C_2)Wt(C_1\rightarrow C_2)}{2d} \qquad (2)$$

Given

$$Wt(x\rightarrow y)=\max_r-\frac{\max_r-\min_r}{n_r(x)} \qquad (3)$$

where r is a relation of type r, r' is its reverse, d is the depth of the deeper one of the two, max and min are the maximum and minimum weights possible for a specific relation type r respectively, and n x r ( ) is the number of relations of type r leaving node x.

D. Set based approach

Set based methods for measuring the similarity of annotations are based on the Tversky ratio model of similarity, which is a general model of distance between sets of terms. It is represented by the formula

$$\frac{f(G_1\bigcap G_2)}{f(G_1\bigcap G_2)+\alpha^* f(G_1-G_2)+\beta^*(G_1-G_2)} \qquad (4)$$

Where G1 and G2 are sets of terms or annotations from the same ontology and f is an additive function on sets (usually set cardinality).

E. Graph based Approach

Consider the similarity between annotations in terms of the sub-graph that connects terms within each annotation. Annotation similarity is then measured in terms of similarity between two graphs. Graph matching has only a week correlation with similarity between terms.

$$Sim(w_1,w_3)=2d_{\max}\left[\frac{c_1 sen(w_1)\min c_2 sen(w_2)}{len(c_1,c_2)}\right] \qquad (5)$$

It is also computationally expensive to compute, graph matching being an NP-complete problem on general graphs.

F. Term-based Approaches

Term-based approaches depend on a function s (Ti, Tj) where Ti and Tj are terms from two annotations G1 and G2 respectively. s(Ti, Tj) provides a measure of distance/similarity between these two terms. Once distances has been measured between all possible pairs of terms they are then aggregated using an operation such as max or the average of all distances. For example:

$$S_{avg}(G_1,G_2)=\sum_{i=1}^{n}\sum_{j=1}^{m}s(T_i,T_j) \qquad (6)$$

More sophisticated term based approaches combine multiple measures of term similarity and aggregate similarity values using more complex functions. The paper explained DAG based semantic similarity for clustering Gene using Gene Ontology.

IV. Framework and Methodology

The objective of the proposed work is to analyse Diabetic Gene using Gene Ontology semantic similarity metric based on Directed Acyclic Graph (DAG) structure for Biological process. The framework of Diabetic Gene is shown in Figure 1.

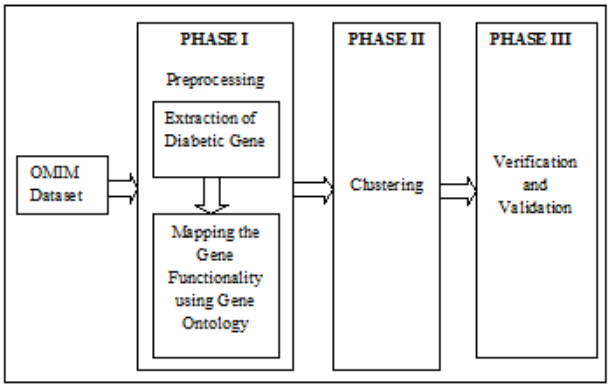

Fig. 1. Framework of Diabetic Gene using Hierarchical Clustering

A. Extraction of Diabetic Gene

The Diabetic Gene is extracted from OMIM dataset for human taxonomy. The gene is mapped to the gene-human dataset to retrieve their GO term. The GO term is used to describe the functions of gene involved. The Gene-GO matrix is constructed which consists of gene and their corresponding GO terms. A total of 100 gene is used for the proposed work.

B. Mapping Gene Functionality

The extracted gene is mapped to Gene Ontology to find the functionality of gene involved. The GO term for each gene is extracted and mapped for functionality using the recent Gene Ontology. The Gene Ontology is vocabulary consist of gene information classified based on their function as Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). The gene is retrieved based on the Biological Process (BP) for proposed work. The snapshot of Gene Ontology is shown in Figure 2.

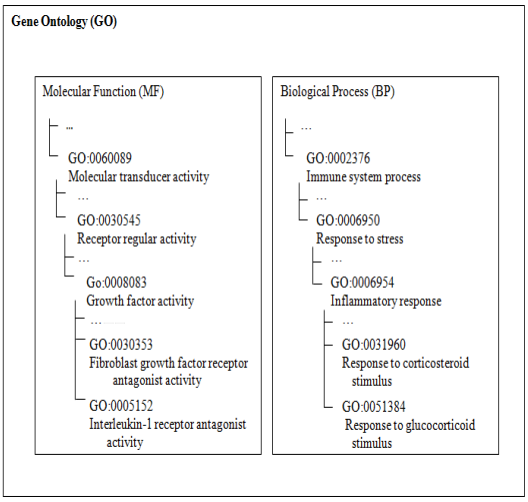

Fig. 2. View of Gene Ontology with functionality

C. Semantic Similarity using Weight (DAG)

The DAG-GO structure of the gene is used to assign the weight for each GO term for clustering gene. The top term in the DAG are generic, lower weight are assigned. The bottom nodes are specific and highly related and they are assigned higher weight. The DAG structures with bottom nodes are assigned more weight and the top nodes are assigned lower weight. The weight gets increased when moved to the bottom layers. The weight is assigned between (0 and 1). The equation given below is used to calculate the weight.

$$1-\alpha^{-x-y} \qquad (7)$$

The weights are represented as Global Weight and Local Weight. The sum of the weight of all the GO terms of the gene is called as Global Weight. The Local Weight is the weight of individual GO term for the gene. The weight is assigned based on the functionality using the DAG structure. The depth used for the weight between the GO terms from the higher level to the lower level is calculated is shown in the Figure 3.

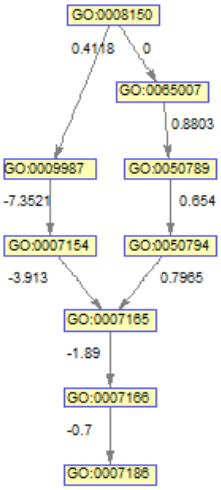

Fig. 3. Weight assigned for GO terms using DAG

D. Hierarchical Clustering

Hierarchical Clustering is used to analyze the Gene associations of Diabetic Gene based on DAG weight. The clusters are generated for Gene involved in Biological Process and the Gene are also analyzed separately for Biological Process.

V. Results and Discussion

The experimental result is implemented and tested for Diabetic Gene using hierarchical clusters to analyze biological results of Gene related to Biological Process.

A. OMIM Dataset

This work is used in Diabetic Gene for clustering is downloaded from OMIM dataset. Diabetic Mellitus is a group of metabolic diseases in which a person has high blood sugar either because the body does not produce enough insulin or because cells do not respond to the insulin that is produced. Online Mendelian Inheritance in Man (OMIM) focuses on the relationship between phenotype and genotype. The Diabetic Gene is downloaded from the OMIM dataset and the gene map is used to search in the Mendelian Inheritance in Man (MIM). The downloaded gene is mapped with their corresponding GO terms using Gene_Info for human gene. The gene information (Gene_Info) dataset contains 1048576 genes with their corresponding information (Tax_id, GeneID, Synonyms, Chromosome, Locus Tag, Full nomen authority etc).

B. Weight assignment using DAG

The weight is assigned based on the Gene functionality using the Directed Acyclic Graph (DAG) structure for clustering the Gene. The weight is assigned between 0 and 1. The weight is represented as Global weight and Local weight. The top node and bottom node are assigned based on their weight.

C. Hierarchical Clustering

The Gene clustered is used the proposed approach for analyzing Diabetic Gene based on DAG similarity metric is proposed.

1) Gene clusters for Biological Process: The Dendrogram view of clusters generated for Biological Process of Diabetic Gene is presented in Figure 4.

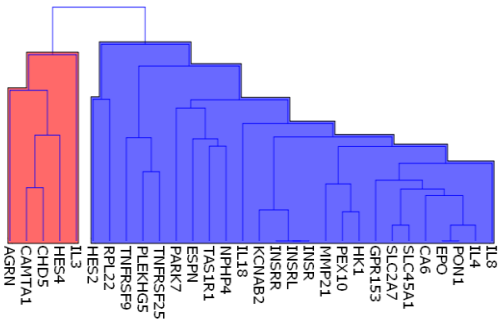

Fig. 4. Clusters for Biological Process

The Table I provides with the information related to the clusters with maximum grouping for the Figure 4.

TABLE I. CLUSTES FOR BIOLOGICAL PROCESS WITH 30 GENE

Cluster	Gene	No. of Gene
Cluster 0	*IL8, IL4, PON1, EPO, CA6, SLC45A1, SLC2A7, GPR153, HK1, PEX10, MMP21, INSR, INSRL, INSRR, KCNAB2, IL18, NPHP4, TASIR1, ESPN, PARK7, TNFRSF25, RPL22, PLEKHG5, TNFRSF9, HES2*	*25*
Cluster 1	*IL3, HES4, CHD5, CAMTA1, AGRN*	*5*

2) Clusters for Biological Process using Set based Approach:

The proposed work is compared with set based approach. The clusters generated for set based similarity metric for Biological Process is shown in Figure 5.

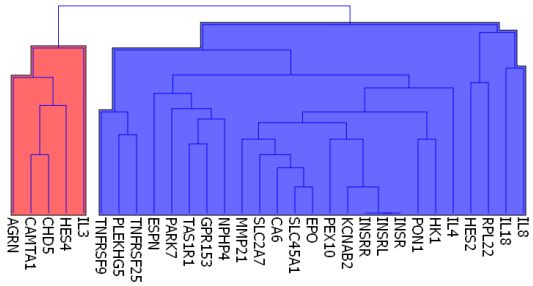

Fig. 5. Clusters for Biological Process using Set based Approach

The Gene is grouped into two clusters: Cluster 0 and Cluster 1 which is presented in Table II. It shows 25 Gene are clustered in Cluster 0 and 5 Gene are clustered in Cluster 1.

TABLE II. CLUSTERS FOR BIOLOGICAL PROCESS USING SET BASED APPROACH

Cluster	Gene	No. of Gene
Cluster 0	*IL8, IL18, RPL22, HES2, IL4, HK1, PON1,INSR, INSRL, INSRR,KCNAB2, PEX10, EPO, SLC45A1, CA6, SLC2A7, MMP21, NPHP4, GPR153, TASIR1, PARK7, ESPN, TNFRSF25, PLEKHG5, TNFRSF9*	*25*
Cluster 1	*IL3, HES4, CHD5, CAMTA1, AGRN*	*5*

D. Discussion and Analysis of Clusters

The experimental results of Diabetic Gene using the proposed approach for cluster are analyzed. On biological validation it is found that 42% of Genes PON1, TASIR1, IL18, GPR153, IL8, NPHP4, HK1 and PLEKHG3 of cluster 0 comes under protein binding and its sub activity and 19% under receptor activity, 9% under growth activity, 9% under hematopoietin receptor binding for Molecular Function and 100 % of Genes INSR, INSRR, INSRL and KCNAB2 of Cluster 1 comes under transportation activity using Gene Ontology Terms. The Cluster Gene based on biological functionality is shown in Table III.

TABLE III. BIOLOGICAL FUNCTIONALITY OF CLUSTERED GENE

CLUSTER	GENES (%)	ACTIVITY
Cluster 0	*70*	*Biological Validation*
	42	*Protein Binding*
	19	*Receptor Activity*
	9	*Growth Activity*
	9	*Hematopoietin Receptor Binding*
Cluster 1	*100*	*Biological Validation*
	30	*Transporter Activity*

VI. CONCLUSION

Bioinformatics is the science of managing, mining, integrating and interpreting information from biological data. Gene Ontology is a collection of controlled vocabularies that describes the biology of a gene product. The work is used to find the Gene weight for Gene-GO matrix using the DAG structure based on Hierarchical Approach. The OMIM dataset is taken for processing. The experimental result it is found that 70% of genes is clustered in one group and 30% of clusters in other group. On comparison of the proposed approach with Set Based Approach it is found that the same set of genes are clustered which shows the biological relevance of the proposed similarity metric.

REFERENCES

[1] Bharati M. Ramageri , “Data Mining Techniques And Applications”, In Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305 , 2010.

[2] Gaston K. Mazandu and Nicola J. Mulder., “A Topology-Based Metric for Measuring Term Similarity in the Gene Ontology”, Published in the Journal of Advances in Bioinformatics, vol. 2012, pp.17, 2012.

[3] Gray .B, Fogel., David W.CorneYipan, “Computational Intelligence in Bioinformatics”, IEEE Press Services on Computional Intelligence, ISBN: 978-0470-10526-9, 2007.

[4] Haixuan Yang, Tamas Nepusz, et.al, “Improving GO Semantic Similarity Measures by Exploring the Ontology Beneath the Terms and Modelling Uncertainty”, Published in Oxford Journal, vol. 28, no. 10, 2012.

[5] Hisham Al-Mubaid and Anurag Nagar “Comparison of Four Similarity Measures Based on GO Annotations for Gene Clustering”, IEEE symposium, ISSN: 1530-1346, 2008.

[6] Jin Gou, Yangyang Wu et. al, “An Ontology Based Knowledge Clustering Method in Knowledge Space”, IEEE Conference Publications, vol. 1, pp. 4244-1035, 2007.

[7] Lin .D, “An Information-Theoretic Definition of Similarity”, In proceeding of the 15th International Conference on Machine Learning, pp.296-304, 1998.

[8] Li Liao, YuzhongQu, et.al, "A Software Process Ontology and Its Application", cited from: www.wenku.baidu.com, 2002.

[9] Monica Chagoyen, Pedro Carmona-Saez, et.al, “A Literature-based Similarity Metric for Biological Processes”, 2006, cited from: www.biomedcentral.com.

[10] Stefano Bianchi, Anna Burla et. al, “Biomedical Data Integration-Capturing Similarities while Preserving Disparities”, IEEE Conference Publications, id. 19963617, 2009.

[11] Tatsiopoulos .C and Boutsinas .B, "Ontology Mapping Based on Association rule Mining", Published in International Conference on Enterprise Information Systems, 2009.

[12] Toralf Kirsten, Andreas Thor, Erhard Rahm "Instance-Based Matching of Large Life Science Ontologies", 2007, cited from: citeseerx.ist.psu.edu.

[13] Wang .J. Z, Du .Z et. al, "A new method to measure the semantic similarity of GO Terms", Published in Bioinformatics Oxford Journals, vol. 23, no. 10, pp. 1274-1281, 2007.

[14] Xiang-hua Xu, Jia-lai Huang et. al, "A Method for Measuring Semantic Similarity of Concepts in the Same Ontology" International Multi-symposiums on Computer and Computational Sciences, IEEE Conference Publications, ISBN: 978-0-7695-3430-5, 2008.

[15] Ying Shen, Shaohong Zhang et. al, "Characterization of Semantic Similarity on Gene Ontology based on a Shortest Path Approach", Published in International Journal Data Mining and bioinformatics, 2011, cited from: www.tonji.edu.cn.

[16] Zhang .P, Zhang .J et.al, "Gene Functional Similarity Search Tool (GFSST)", BMC Bioinformatics, vol. 7, pp. 135, 2006.

S. Booma Shanthini M.Phil Research Scholar in Bharathiar University. Area of interest is Data Mining.

Dr.V.Bhuvaneswari Assistant Professor in Bharathiar University. She is guiding the M.Phil Research Scholars. Her area of interest is Data Mining and Bioinformatics.