ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

A Study On Semantic Similarity Of Gene Ontology Using Clustering

K.Umamaheswari [1], S.Niraimathi [2]

Research Scholar, Computer Science, NGM College, Coimbatore, India [1]

Assistant Professor, Computer Science, NGM College, Coimbatore, India [2]

Abstract: Gene Ontology is structured as a directed acyclic graph, and each expression has distinct interaction to one or more other terms in the same domain and sometimes to other domains. High throughput techniques have become a primary approach to gathering biological data. These data can be used to explore relationships between genes and to identify disease. Clustering is a common methodology for the analysis of array data and many research laboratories are generating array data with repeated measurement. Cluster analysis seeks to division a given dataset into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. The gene ontology is a gene(gene products)using terms from three structured vocabularies: Biological process, cellular component and molecular function. For measuring the semantic similarity on GO terms using novel method, namely shortest path (SP) algorithm.

Keywords: Clustering, Gene Ontology, Shortest path, Semantic similarity, Novel method

I. INTRODUCTION

Gene Ontology (GO) represents an important knowledge resource to describe the function of genes[2].It is a structured and controlled vocabulary, which characterizes the functional properties of gene/proteins using standardized terms. Semantic similarity is an important type of information derived from GO, the concept of which is originally used in the field of linguistics. In linguistics, two words are considered to be of similar meanings[1].It may be based on statistical and topological information about GO terms and/or their inter relationships in the topology.And it aims to expand our understanding of the relationships between GO-driven gene similarity and expression correlation .Such an assessment may allow one to justify the design of annotation-based predictive models and their integration with expression data models.It may provide the basis for novel methods to assess the predictive quality and reliability of functional genomics analyses involving gene expression or other types of related data[4].Although many methods were proposed for gene selection,most of them focused on selecting a set of relevant genes and improving the classification accuracy by the selected relevant genes .To enhance the usefulness of proposed method,the novel method is introduced,which is called two staged weighted sampling strategy (TSWS strategy).TSWS strategy can not only be used to select a set of relevant genes with higher classification accuracy but also be used to assist researchers to detect the relation between the expression levels of genes and the classes of a cancer for diagnosing[2].In GO,Molecular function represents information on the role played by a gene product.Biological process refers to a biological objective to which a gene product contributes.Cellular component represents the cellular localization of the gene product,including cellular structures and complexes[4].

II. LITERATURE SURVEY

Clustering is a useful exploratory technique for gene-expression data as it groups similar objects together and allows the biologist to identify potentially meaningful relationships between the objects[5].A GO term is a word used to describe a certain functional property. Every GO term has corresponding GO ID in the form of "GO:*******",.For example, the GO ID for the term "cellular component" is GO:0005575.The second concept is the relationship defined between GO terms[1].There are two kinds of relationships: 'is-a' relationship and 'part-of' relationship. One of the contributions of this paper is to exploit term-term similarity in GO hierarchies for computing gene-gene similarity[4]. Edge-based methods are intuitive,suppose t1 and t2 are two terms,and t is their lowest common ancestor.The distance method is counts the number of edges connecting the root with t,and the number of edges connecting t with t1 and t2.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

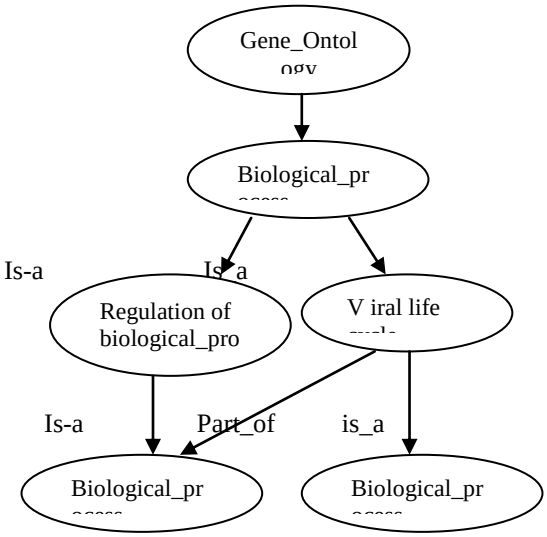

Fig 1: Example of DAG

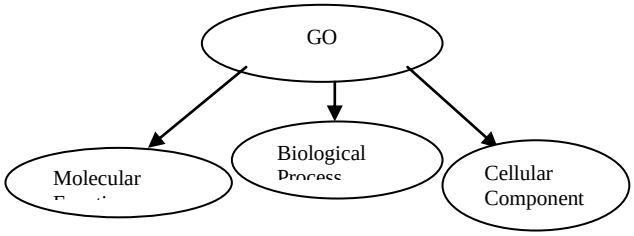

Fig 2: Example of GO

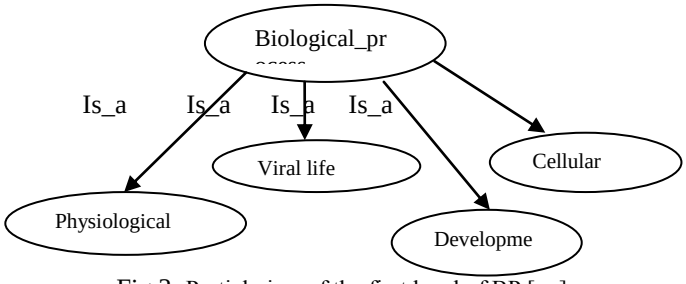

Fig 3: Partial view of the first level of BP.[…]

Semantic similarity is an important type of information derived from GO,the concept of which is originally used in the field of linguistics.In linguistics,two words are considered to be similar if they have similar meanings.Semantic similarity can be defined for both the GO terms and gene products.The state-of-the-art methods for specifying sematic similarity over the GO terms can be divided into three groups:edge-based,node-based,and a hybrid of the above two.For the edge-based approaches,they mainly consider the lengths of the paths connecting terms[8].In addition to the edge-based and node-based methods,there are also a number of hybrid methods proposed[7].Edge-based methods are intuitive,suppose t1 and t2 are two terms,and t is their lowest common ancestor.The distance method is counts the number of edges connecting the root with t,and the number of edges connecting t with t1 and t2.The distance between t1 and t2 is calculated using.eq.(1) below and can be easily converted to a similarity value:

Len(root,t)
Dist(t1,t2)=

Len(root,t)+len(t,t1)+len(t,t2)

Where len(x,y) is the length of the path connecting the nodes x and y,represented by the number of edges on the path.The distance method assumes that the weight of each edge is always 1[8].The disadvantage of the edge-based methods is that the weights of the edges at the same level are assumed to be the same.Node-based methodsfocus mainly on the specificity of the terms,which is expressed using the concept of information content(IC).The IC value for a term t is defined as

$$IC(t)=-\log p(t)$$

One of the first methods using IC values to measure the semantic similarity for GO terms.In this method,the semantic similarity for terms t1 and t2 is defined as[6].

$$Sim(t1,t2)=\max IC(t)$$

III. METHODOLOGY

The shortest path problem can be defined for graphs whether undirected, directed, or mixed. It is defined here for undirected graphs; for directed graphs the definition of path requires that consecutive vertices be connected by an appropriate directed edge. The shortest path problem is the problem of finding a path between two vertices (or nodes) in a graph such that the sum of the weights of its constituent edges is minimized. Two vertices are adjacent when they are both incident to a common edge. A path in an undirected graph is a sequence of rtices.P= (v1,v2,……..vn) VxVx…x such that v_i is adjacent to Vi+1 for 1<i<n. Such a path P is called a path of length n from V1to Vn.(The Vi are variables; their numbering here relates to their position in the sequence and needs not to relate to any canonical labeling of the vertices)[10].Again, we have a random variable Y defining the clustering results, and the random variables {X1,...,Xp,...,XP } corresponding to the annotation vectors of group of terms selected above. The measure in based on the approximation of the joint mutual information MIapp(X,Y) as proposed in [9].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

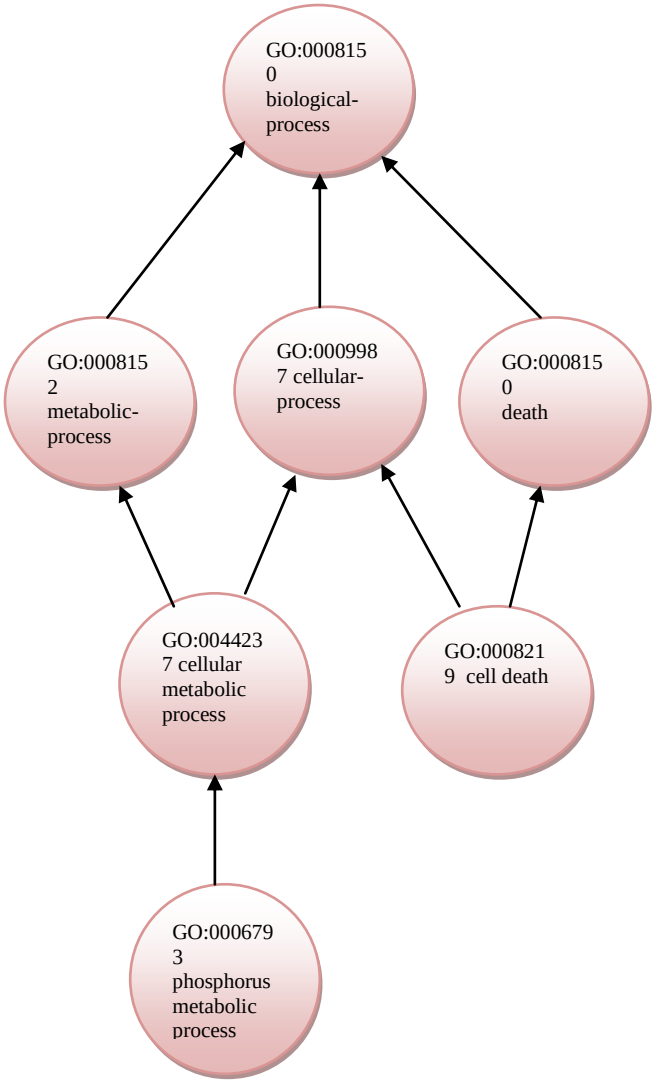

Fig 4: Gene ontology process

A. *Validation Index*

. We use the index proposed in [3] to obtain a global measure of fitness by comparing a clustering (partition) with the set of terms selected with MutSel. Again, we have a random variable Y defining the clustering results, and the random variables {X1,...,Xp,...,XP } corresponding to the annotation vectors of group of terms selected above. The measure in based on the approximation of the joint mutual information MIapp(X,Y) as proposed in [9].Several types of data can be used to assess the accuracy of existing methods for measuring semantic similarity.Now we propose a both protein-protein interaction(PPI) data and gene expression datasets [4] to evaluate the correctness of computed semantic similarities

III. CONCLUSIONS

A method has been proposed that provides a number of advantages over typical approaches to gene clustering.It intelligently weights similarity measures by their predictive power,allowing a number of statistics to be utilized regardless of their individual usefulness.In this paper,a new method for measuring the semantic similarity,namely the shortest path algorithm(SP).This algorithm has the advantage of less computation time due to fewer variables.

REFERENCES

[1]Cheng.J.,Cline,M.Martin,J.Finkelstein,D.,Awad,T.,Kulp,D.,andSiani-Rose,M.A.(2004)"A Knoledge-based clustering algorithm driven by gene ontology",Journal of Biopharmacentical Statistics,Vol.14,pp.687-700.

[2]Chih-Chung Yang and Wen-Shin Lin(2013),'Two stages weighted sampling strategy for detecting the relation between gene expression and disease',Int.J.Data Mining and Bioinformaitcs,Vol.x,xxxx.

[3] Francis D. Gibbons and Frederick P. Roth. Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. Genome Res., 12(10):1574–1581, 2002.

[4]Haiying Wang,Francisco Azuaje,(2004)'Gene Expression Correlation and Gene Ontology-Based Similarity:An Assessment of Quantitative Relationships',IEEE Symposium on Computational Intelligence inBioinformatics and Computational Biology(CIBCB-2004).

[5] Ka Yee Yeung, Mario Medvedovic and Roger E Bumgarner (2003),' clustering gene expression data with repeated measurements'

[6] Resnik,P.(1999)"semantic similarity in taxonomy:An information-based measure and its application to problems of ambiguity in natural language",journal of Artificial intelligence Research,Vol.11,pp.95-130.

[7] Wang,J..Du,Z,,Payattakool,R.,Yu,P.S.,and Chen,C.(2007)"A new method to measure the semantic similarity of GO terms",Bioinformatics,Vol.23,pp.1274-1281.

[8]Ying shen,Shaohong Zhang,Hau-san Wong, 'Characterization of Semantic Similarity on Gene Ontology based on a Shortest Path Appoarch'Int.J.,Vol.x,No.x,xxxx.

[9] The Gene Ontology Consortium,"Creating the gene ontology resource: Design and implementation," Genome Research, vol.11,pp.1425-1433,2001.

[10] K.Umamaheswari,and S.Niraimathi, " A Study on student Data analysis Using Data mining Techniques", International Journal Of Advanced Research In Computer Science And Software Engineering, Vol-3,issue-8,August,2013.