International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 25 Issue 6 – MAY 2018.

MiRNA & TRANSCRIPTION FACTOR ON GENOMIC DATA USING CROSS ONTOLOGY

Dr.C.Dhaya [#1] and V.Pooja [*2]

[#] *Assistant Professor/CSE, Adhiparasakthi Engineering College, Melmaruvathur.India*

[*] *PG Scholar, Dept of CSE, Adhiparasakthi Engineering College, Melmaruvathur.India*

***Abstract*— Gene Ontology (GO) is a structured repository of concepts that are associated to one or more gene products through a process referred to as annotation(the practice of capturing data about a gene) to understand how the ontologies are structured and avoid errors in interpretation. There are different approaches of analysis to get bio information(biological information like DNA characteristics of a person to identify the defect). One of the analysis is the use of Association Rules (AR) which discovers biologically relevant associations between terms of GO. In existing work GO-WAR (Gene Ontology-based Weighted Association Rules) is used for extracting Weighted Association Rules with high level of Information Content from ontology-based annotated datasets without the use of post-processing strategies. The MOAL algorithm is adapted to mine cross-ontology association rules, i.e. rules that involve GO terms present in the three sub-ontologies of GO. In this paper cross ontology is proposed to manipulate the Protein values from three sub ontologies for identifying the gene attacked disease. Also proposed system, focus on identifying the type of gene into intrinsic and extrinsic gene. Based on cellular component, molecular function and biological process values intrinsic and extrinsic calculation would be manipulated.**

***Index Terms*—Ontology, Gene Ontology, cross ontology**

I. INTRODUCTION

Ontologies are specifications of a relational vocabulary. Gene ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Different ontologies have been proposed to elucidate different fields. For instance, the Gene Ontology (GO) is one of the frameworks that are largely used. Gene Ontology includes three main sub-ontologies: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). The Cellular component refers the parts of a cell or its extracellular environment. Molecular function denotes the elemental activities of a gene product at the molecular level, such as binding or catalysis. Biological process is the operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each ontology stores and organizes biological concepts, called GO Terms, used for describing functions, processes and localization of biological molecules. Each GO term is uniquely identified by a code, it belongs to only one ontology, and for each GO Term a textual description is also available. For instance GO: 0006915 represents the apoptosis process.

II. RELATED WORKS

A. *Mining Spatial Gene Expression*

In this method association rule mining is used to identify relationships among up and down regulated genes in gene expression studies. These studies do not make use of the GO and its hierarchical structure. Previous research applying association rule mining to the GO includes studies mining single level, multi-level and cross-ontology association rules. Scheduling is a method to effectively distribute the available computing resources to incoming jobs.

B. *Automatic Content Specified Generation*

Automated Content Specified Generation generalizes the GO by calculating the information content of a node using both the ontology structure and the annotation dataset as a metric for generalization. Non-traditional definition of information content of a concept x as Ix = Px−Ox, where Px is the information gained by not generalizing concept x and Ox is the information lost if all the child terms of x are generalized to x. Px and Ox are calculated using information from the annotation dataset and the ontology structure. This approach is used to generate automatic slim sets from the GO, but it is unclear that how this approach will work for mining associations from multiple ontologies.

C. *GO-WAR for Mining Cross Ontology*

GO-WAR, i.e. Gene Ontology-based Weighted Association Rules Mining, a novel data-mining approach is developed to extract weighted association rules starting from

an annotated dataset of genes or gene products. The proposed approach is based on the following steps: (i) initially we calculate the information content for each GO term; (ii) then, we extract weighted association rules by using a modified FP-Tree like algorithm which is able to deal with the dimension of classical biological datasets. Publicly available GO annotation data is used to demonstrate proposed method.

III. SYSTEM ARCHITECTURE

Describing the overall features of the software is concerned with defining the requirements and establishing the high level of the system. During architectural design, the various web pages and their interconnections are identified and designed. The major software components are identified and decomposed into processing modules and conceptual data structures and the interconnections between the modules are identified. The following modules are identified in the proposed system.

A. *Architecture model*

Co-regulatory modules are proposed between Transcription Factor, gene and MiRNA on functional level with genomic data. The integration technique is implemented between miRNA, Transcription Factor (TF) and gene. After integration, Iterative Multiplication update algorithm is used to check the optimization function between the regulatory modules. We get The expression or some value is got from this algorithm then compared to protein values. The protein value is got from Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) with the help of cross ontology technique. At last a Bayesian rose tree structure is generated for the relation between regulatory modules and protein values of gene. By this structure disease which was affected in our chromosome is identified and also how to cure? What are the symptoms are applicable for that gene by proposed system.

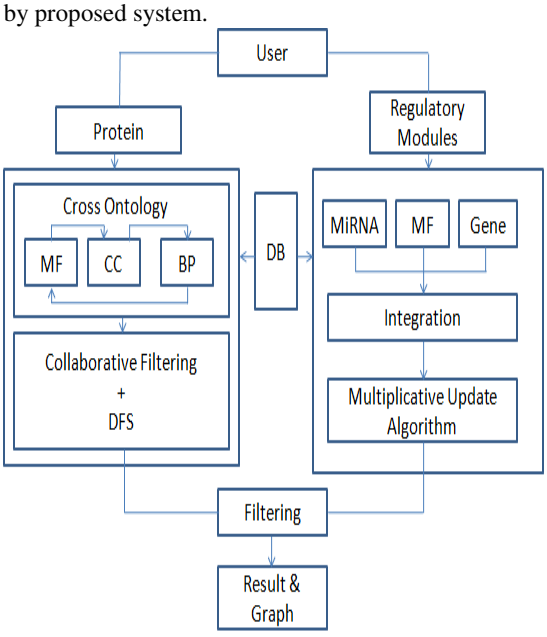

Figure 1. Architecture diagram

IV. SYSTEM MODULES

A. *Cross Ontology*

A cross ontology consists of the primary ontology and the secondary ontology or multiple ontology for the comparison process by providing a cross ontology value. Cross ontology is proposed to manipulate and compare the protein values compared from three sub ontology (Molecular Function, Cellular Component &Biological Process) for identifying the type of gene (intrinsic and extrinsic) attacked disease. The MOAL algorithm is adapted to mine cross-ontology association rules, i.e. rules that involve GO terms present in the three sub-ontology of GO. Retrieval of relevant materials from a domain can be easily implemented by using Cross Ontology.

B. *Collaborative Filtering*

Collaborative filtering is a method of making automatic predictions (filtering) collecting preferences and similarities. Semantic mining is used for logical analysis. User gets the disease details for Gene ID from Ontology base with help of Collaborative filtering. It is more versatile (applied to any domain & also provide cross-domain) also work best when the user space is large (insensitive to user size).

C. *Depth First Search (DFS) Algorithm*

The DFS algorithm is a recursive algorithm that uses the idea of backtracking. It involves exhaustive searches of all the nodes by going ahead, if possible, else by backtracking. Depth first search is used for searching relevant diseases with the cross ontology output. It is better to heuristic methods (choosing a likely-looking branch) and memory requirement is only linear with respect to the search graph.

D. *Regulatory Modules*

Due to clustering all anonymized data are not derived together. It will get cluster by cluster after anonymization. All data is gathered after anonymization to store the total anonymized data in the server and release this anonymized dataset for further use. By this approach the dataset can be efficiently used with privacy.

E. *Integration Technique*

In this study, two approaches are used to the integration of mRNA, miRNA, and protein expression data, in order to identify cancer-related miRNAs and investigate relationships between miRNAs and the regulatory networks in cancer. A new computational method is presented for the ranking of cancer-related miRNAs based on the number of identified correlated genes, using both mRNA and protein datasets. Ranking lists constructed for each miRNA may advance understanding of the cancer-related miRNAs. Additionally, a method is presented for the construction of modules containing mRNAs, miRNAs, and proteins. The modules were constructed based on the SAMBA bi-clustering algorithm and a Bayesian network model . To construct these modules, proposed system extended the approach proposed by Jin and Lee by adding a step in which the proteins are included into mRNA-sample modules prior to the inclusion of miRNAs. The identified modules represent subgroups of highly correlated mRNAs, miRNAs, and proteins, and may

explain regulatory networks between miRNAs and genes.

F. *Optimization using Multiplicative Updating*

In this module, the optimization model function is solved effectively by using the iterative multiplicative updating algorithm. It is an algorithmic technique which "maintains a distribution on a certain set of interest, and updates it iteratively by multiplying the probability mass of elements by suitably chosen factors. This method is popular due to its simplicity.

G. *Tree Representation*

In this module, the tree consist of gene id as root element after that the leaf nodes contains its molecular function value, biological process value, cellular component values and also contains the symptoms ,diseases and curing possibilities of the related gene id. This tree representation is more useful to predict the details easily about the gene.

V. METHODOLOGY

A. *Multi Ontology data mining at All Levels (MOAL) Algorithm*

String-matching is a very important subject in the wider domain of text processing. String-matching algorithms are basic components used in implementations of practical software existing under most operating systems. By using this string pattern matching algorithm the preprocessing step is done easily. The downloaded Dataset is not in a clear form. To make those data in a clear form or in a table form we need to do preprocessing.

Moreover, they emphasize programming methods that serve as paradigms in other fields of computer science (system or software design). Finally, they also play an important role in theoretical computer science by providing challenging problems. Although data are memorized in various ways; text remains the main form to exchange information. This is particularly evident in literature or linguistics where data are composed of huge corpus and dictionaries. This is the reason why algorithms should be efficient even if the speed and capacity of storage of computers increase regularly.

B. *Iterative Multiplication Update Algorithm*

Optimization model function is solved effectively by iterative multiplicative updating algorithm. It is an algorithmic technique which "maintains a distribution on a certain set of interest, and updates it iteratively by multiplying the probability mass of elements by suitably chosen factors. This method is popular due to its simplicity.

VI. CONCLUSION

By proposed system everyone should know their disease which will affect them. More reliable to know the curing possibilities since our paper identifies the predictor accuracy and states decision tree data mining technique provides accuracy. This paper also illustrates the scope of data mining technique in medical domain. In future, Automatic decision tree based prediction with machine learning would be useful tool for medical research groups for predicting cancers.

REFERENCES

[1] M. Cannataro, P. H. Guzzi, and A. Sarica, "Data mining and-life sciences applications on the grid," Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, vol. 3, no. 3, pp. 216–238, 2013. [Online]. Available: http://dx.doi.org/10.1002/widm.1090

[2] P. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," Briefings in bioinformatics, vol. 13, no. 5, pp.569–585, 2012. [Online]. Available: http://bib.oxfordjournals.org/content/early/2011/12/02/bib.bbr066.short

[3] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, and et al, "The gene ontology (go) database and informatics resource." Nucleic Acids Res Nucleic Acids Res, vol. 32, no. Database issue, pp.258–61, January 2004.

[4] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology," Nucl. Acids Res., vol. 32, no. suppl 1, pp. D262–266, January 2004. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh021

[5] M. Masseroli, O. Galati, and F. Pinciroli, "GFINDer: Genetic disease and phenotype location statistical analysis and mining of dynamicall y annotated gene lists", Nucleic Acids Res., vol. 33, pp. W717-W723 , 2005.

[6] M. Masseroli, "Management and analysis of genomic functional and phenotypic controlled annotations to support biomedical investigation and practice", IEEE Trans. Inf. Technol. Biomed., vol. 11, 4, pp. 376 -385, 2007.

[7] T. Etzold, A. Ulyanov, and P. Argos, "SRS: Information Retrieval Sys tem for molecular biology data banks", Methods Enzymol., vol. 266, pp. 114-128, 1996.

[8] T.A. Tatusova, I. KarschMizrachi, and J.A. Ostell, "Complete genom es in WWW Entrez: data representation and analysis", Bioinformatics , vol. 15, pp. 536-543, 1999.

C.Dhaya received the **B.E.** degree in Computer science and Engineering from Adhiparasakthi Engineering College, Anna University, Chennai, India in 2000. Received **M.E.** degree in Computer Science and Engineering in Jerusalem College of Engineering, Anna University, Chennai, India in 2007. Received **Ph.D** degree in Computer Science and Engineering in Pondicherry Engineering College, Anna University, Chennai, India in 2014. Her research interest includes Software architecture, Software Engineering, Network Security, Quality Metrics and Information Security.

V.Pooja received the **B.E.** degree in Computer Science and Engineering from Adhiparasakthi Engineering College, Anna University, Chennai, India, in 2016. Currently doing **M.E.** in Computer Science and Engineering in Adhiparasakthi Engineering College, Anna University, Chennai, India. Her research interest includes Data mining and Ontology.