International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

DISEASE IDENTIFICATION USING PROTEINS VALUES AND REGULATORY MODULES

Mrs. J. Hemavathy[1], B. Jaya[2], T. Ananthi[3], Rachel Thomas[4]

[1] Assistant Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India.
[2,3,4] Students, Final Year, Department of Information Technology, Panimalar Engineering College, Chennai, India.

Abstract- *Gene analysis has a huge scope in identifying the genetic disorders early and perform the respective diagnosis. Gene regulatory modules microRNA (miRNA) and transcription factor (TF) play a very important role in gene regulation. Clustering is a main challenge in gene analysis. Thus in existing system the multiple genomic and proteomic analysis are scattered in multiple distributed system. In our proposed architecture, we try to develop a common knowledge base for genomic and proteomic analysis using collaborative filtering (CF) and depth first search (DFS). Clustering process is being used to group the gene ontology and regulatory modules for each and every gene expressions. Finally the challenge of deriving taxonomy for a particular gene id is resolved using Bayesian rose tree (BRT).*

Key words: **Gene ontology, Regulatory Modules, collaborative Filtering, Depth First Search, Bayesian Rose Tree**

1. INTRODUCTION

Usage of computer and technology, huge amount of medical data's creates a huge scope of data mining techniques. Data mining techniques are widely used and popular among the medical research groups. Data mining technique are applied to obtain the solution from large amount of knowledge base, relationship/association among the variables, predict a specific disease based on historical datasets, assigning weight age to the variable etc. The objective of this research article is to develop a common knowledge base for genomic and protein patterns to identify the genetic disorders invoking regulatory modules as well. Also integration of collaborative filtering, graph based clustering, depth first search and Bayesian rose tree representation would provide an efficient and easy solution for representing the gene terms and identifying the associated diseases for a particular gene ID.

Increasing huge amount of bimolecular valuable data and information in life sciences there is large scope for gene analysis. The DNA compromises of gene and proteins. Gene analysis focus on identifying the association between the biomolecular entities. Thus in existing system the multiple genomic and proteomic analysis are scattered in multiple distributed systems. In our proposed architecture, we try to develop a common knowledge base for genomic and proteomic analysis, which can be accessed by doctors, scientists, researchers and others to provide solution for more genetic disorders. Thus analysis of gene and protein data provides vital opportunity for bioinformatics domain which produces biologically meaningful data and solutions.

Our proposed architecture help in understanding complex biological patterns and associations. For grouping of same gene information for a particular gene disorder graph clustering (for clustering regulatory modules like miRNA, Transcription Factor (TF) and gene), Collaborative filtering and depth first search (for gene ontology - Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)) approaches are been used. Clustering is defined as a group of similar data elements or data elements somewhere interconnected.

Gene ontology is being used to find the gene related disease. The process of collaboration started with three main Go sub ontologies and they are Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Each ontology has separate storing and organizing biological concepts called the GO Terms which are mainly used for describing the functions. In this project weighted-association rule mining was introduced to identify the gene associations. Thus our proposed system with mining and identifying the associations for a particular gene disorder among large set of gene ID would provide us an effective knowledge base for genetic disorders. Also proposing a integration or fusion technique of both gene ontology and regulatory modules would provide more accurate results. The microRNA (miRNA), gene and transcription factor (TF) are considered for gene regulation. In this for a particular genetic disorder the relationship between miRNA and TF are identified. Finally the results are represented as a tree integrating all the diseases associated with a particular gene ID using Bayesian Rose Tree (BRT).

2. PROPOSED SYSTEM

In our Project, we proposed co-regulatory modules 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 Multiplicative update algorithm is used to check the optimization function between the regulatory modules. We get the expression or some value from this algorithm then compare to protein values. The protein value get from Biological Process(BP), Molecular Function(MF) and Cellular Component(CC) with the help of cross ontology technique. At last we generate a bayesian rose tree structure for the relation between regulatory modules and protein values of our gene. By this structure we know our disease which was affected in our chromosome and also know how to cure. Also we can identify the symptoms applicable for our gene by our proposed system.

2.1 GENE ONTOLOGY

Gene Ontology is a widely used framework for the model of biology. The GO defines the concepts used to describe gene function and its relationships between these concepts.Gene Ontology includes three main sub-ontologies namely Biological Process, Molecular Function, and Cellular Component. Each ontology stores and organizes biological concepts, called GO Terms, used for describing functions, processes and localization of biological molecules. Different ontologies are been proposed and used to analyze different type of fields. The gene ontology functions are classified into three aspects, they are molecular function of gene products, molecular activities of gene products and cellular component.

The association or relationship between various GO terms and biological concepts are processed using annotation technique. Annotations for each genetic disorders are stored in the common database, thus this database is termed as Gene Ontology Annotation (GOA) database. Thus this proposed system is a collaborative effort to obtain consistent descriptions of genes in various database or sources. For this approach, Gene Ontology-based Weighted Association Rules Mining (GO-WAR) is proposed to extract the association rules among the gene id's with high level information content.

The information content (IC) can be defined into two classes namely extrinsic and intrinsic techniques. The extrinsic information content involves the annotation data and whereas intrinsic information content involves structural information extracted from the GO terms. The gene ontology results are transmitted to collaborative filtering and depth first search techniques for providing solution for the genetic disorders.

For providing accurate solutions, our proposed system provides a fusion of both Gene ontology approach and gene regulatory modules. Thus the fusion of both gene ontology and gene regulatory modules provides tremendous accurate solution for genetic disorders. The co-regulation study between microRNA (miRNA) and transcription factor (TF) has become an important issue recently. In our proposed system, we identity and analyse miRNA and TF for genetic disorders by integrating various types of genomic data. Graph clustering technique is used to obtain the associations between miRNA and TF for a particular gene id.

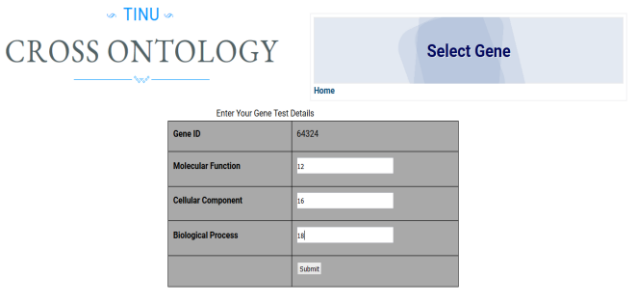


Figure1. Gene Ontology

2.2 COLLABORATIVE FILTERING

Each gene ID refers to a particular diseases. Collaborative filtering is applied to the results obtained using Gene Ontology-based Weighted Association Rules Mining algorithm for optimization. Accuracy is been increased by applying Collaborative filtering technique. The information contents (extrinsic and intrinsic) and gene ontology results are been compared to obtain more accurate recommendation of gene to the patients who has genetic disorders.

Intrinsic:

If the normal protein value of human is compare to lower than that of calculating cross ontology value (comparing BP&CC or MF&CC or MF&BP) is said to be intrinsic.

Extrinsic:

If the normal protein value of human is compare to higher than that of calculating cross ontology value (comparing BP&CC or MF&CC or MF&BP) is said to be extrinsic.

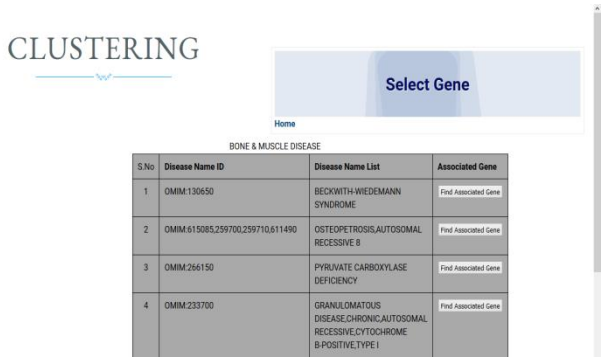


Figure2. Collaborative Filtering

2.3 DEPTH FIRST SEARCH

Depth first search (DFS) is an algorithm for searching tree or graph based data structure. DFS implemented for searching the most affected disease for a given gene ID. While using DFS we can able to identify the entire diseases for a particular gene ID's. DFS provides more efficient search results. The diseases are searched on the basis of chronological order. By using DFS search algorithm we can able to figure all associated diseases associated with each and every gene id.

Figure3. Depth First Search

2.4 REGULATORY MODULES

The regulatory module is being used to identify the associations and clustering the gene regulatory modules microRNA, TF and gene for the respective gene ID. The microRNA and TF is the process mainly compare the about the cells of the cancer cells. The integration technique is used to integrate both gene ontology and regulatory modules. In the process of the regulatory module we process the Multiplicative algorithm. The multiplicative algorithm indicates about the process of the disease id, disease name, disease image and the symptoms of disease. To resolve the optimization, in the proposed system we propose this multiplicative updating algorithm in the project.

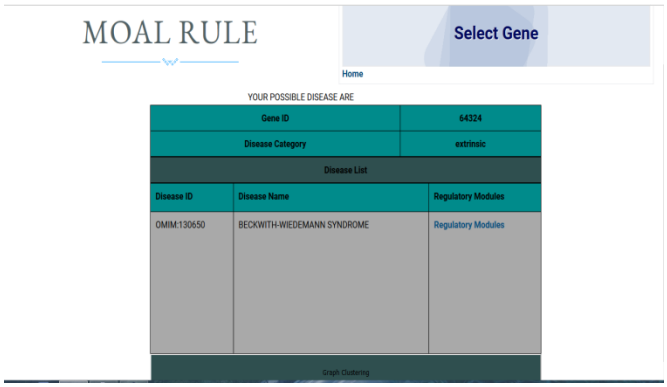


Figure4. Regulatory Modules

2.5 BAYESIAN ROSE TREE

In the Bayesian rose tree we compare the gene ontology and regulatory module. The process of the Bayesian rose tree is to compare the values from the gene ontology and regulatory modules to identify the precaution of the disease. The data's are being processed in the tree structure format. In this proposed system we convert the tree structure into the text format as a record in data base and the results are also being stored in the cloud as record/file.

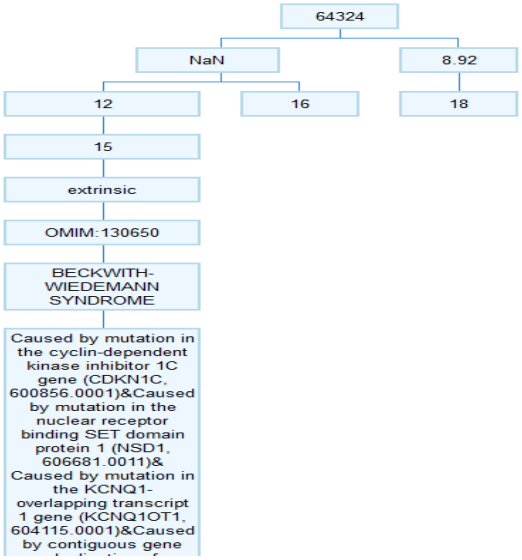


Figure5. Bayesian Rose Tree

3. HARDWARE AND SOFTWARE REQUIREMENTS

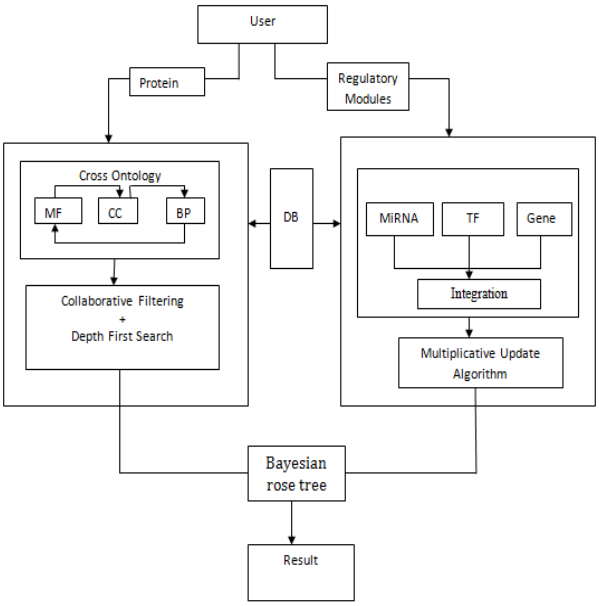
3.1Hardware Requirements

- Processor - Pentium –III
- Speed - 1.1 Ghz
- RAM - 256 MB(min)
- Hard Disk - 20 G
- Floppy Drive - 1.44 MB

3.2 Software Requirements

- Operating System :Windows 95/98/2000/XP
- Application Server : Tomcat 5.0/6.X
- Front End : HTML, Java, JSP, AJAX
- Scripts: JavaScript.
- Server side Script : Java Server Pages
- Database Connectivity: Mysql.

4. ARCHITECTURE DIAGRAM



5. CONCLUSION

The recent progress in biotechnology created a huge scope for data mining and clustering techniques. Fusion architecture of both gene ontology and gene regulatory modules may provide useful and complex information for a particular gene which leads to provide efficient and accurate diagnosis. But this is of huge challenge.

Our proposed work results found satisfactory in identifying and analysing complex biological structures using, collaborative filtering and depth first search. The intrinsic and extrinsic values are also calculated during gene ID analysis. The fusion results and its taxonomy are represented using BRT. To our knowledge, gene ontology and regulatory modules integration and addressing complex queries are not implemented and available in the existing system.

REFERENCE

[1] JiaweiLuo, Gen Xiang and Chu Pan, Discovery of MicroRNAs and Transcription Factors Co-Regulatory Modules by Integrating Multiple Types of Genomic Data, IEEE 2017.

[2] K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, A Graph Theory Algorithmic Approach to Data Clustering and its Application, IJSEAT, vol 3, issue 9, IEEE 2017.

[3] Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi and Marianna Milano, Extracting Cross-Ontology Weighted Association Rules from Gene Ontology Annotations, IEEE 2016.

[4] Yangqiu Song, Shixia Liu, Xueqing Liu and Haixun Wang, Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees, IEEE 2016.

[5] Marco Masseroli, ArifCanakoglu, and Stefano Ceri, Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction, IEEE 2016.

[6] Giuseppe Agapito, Marianna Milano, Pietro H. Guzzi "Improving Annotation Quality in Gene Ontology by Mining Cross-Ontology Weighted Association Rules", IEEE 2015.

[7] K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, Graph–Theoretic Clustering for Image Grouping and Retrieval, IEEE 2015.

[8] 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, IEEE 2013.

[9] R. Priscilla, C.N Prashantha and S. Swamynathan, Analysis of Gene Expression Data using MATLAB Software, IEEE 2013.

[10] ShwetaKharya, Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, IJCSEIT, Vol.2, No.2, IEEE April 2012.

[11] 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, IEEE 2012.

[12] Huseyin Kaya and S. GunduzOguducu, A New Approach for Mutation Analysis Using Data Mining Techniques, IEEE 2010.

[13] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, Efficient Graph-Based Image Segmentation, IEEE 2010.

[14] M.Masseroli, "Management and analysis of genomic functional and Phenotypic controller Annotation to Support Biomedical Investigation and Practice" , IEEE Trans. Inf. Technol. Biomed., vol. 11, 4, pp. 376-385, IEEE 2007.

[15] M.Masseroli, O. Galati, and F. Pinciroli, "GFINDer: Genetic Disease and Phenotype Location Statistical Analysis and Mining of Dynamically Annotated Gene Lists", Nucleic Acids Res., vol.33, pp. W717-W723, IEEE 2005.