ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Identifying Gene Disease using Genomic and Proteomic Knowledge Base

Sabana Yasmin.B[1], Sindhiya.S[2] , Prema.V[3]

Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, SRM Nagar, Kattankulathur, Tamilnadu, India [1],[2]

Assistant Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, SRM Nagar, Kattankulathur, Tamilnadu, India[3]

ABSTRACT: To predict the hereditary disease in an individual we are using genomic and proteomic data. We are using the cross ontology to manipulate the protein value to identify the gene. Based on cellular component, molecular function and biological process values intrinsic and extrinsic calculation would be manipulated. For each proteomic analysis for every gene disease, we analyse OMIM id, disease caused by, associated genes. We done the Co-Regulatory modules between miRNA, TF and gene on function level with multiple genomic data.. We compare the regulations between miRNA-TF interaction, TF-gene interactions and gene-miRNA interaction with the help of integration technique. For optimising the solve we used Multiplicative update algorithm. Bayesian rose tree will be generated based on the values obtained from regulatory modules and protein. The final result will be encrypted and stored in cloud for future reference.

KEYWORDS: Cross Ontology, Collaborative Filter, Depth First Search, Multiplicative Update Algorithm, Bayesian Rose Tree, ABE encryption technique.

I.INTRODUCTION

Ontologies are specifications of a relational vocabulary. The large set of experimental data has been produced in molecular biology from the introduction of high-throughput technology. The gene and proteins used for the role of investigated molecule. To systemize such knowledge, formal instruments are used to manage the used terms

One of the most widely used GO is employed here. It includes three main sub ontologies such as: Biological Processes(BP), Cellular Component(CC) and Molecular Function(MF).Each ontology stores and organizes the biological concepts called GO terms which is used for describing functions and a textual description is also available. The relationship between the biomolecular entities and controlled terms that describes the biomolecular entity function.

Those data support scientist in several terminologies which describes the structural, functional and phenotypic biological features. These semantic annotations can effectively support the interpretation of genomics and proteomics test results and the extraction of biomolecular information, which can be used to formulate and validate biological hypotheses and possibly discover new biological knowledge.

The MOAL algorithm is used to mine the cross ontology. We are proposing cross ontology to manipulate the protein values from three sub ontologies for identifying the gene attacked disease.[1] Based on Cellular Component, Molecular Function and Biological process values, it will be split into two partitions. If the values obtained from these are high it will be keep in Extrinsic and if it is low, it will be in intrinsic.[2]

For each proteomic analysis for every gene disease, we analyse OMIM id, disease caused by, associated genes, medicine if available, and images of that particular gene disorder. Thus a common man also would be able to understand the membranes and enzymes associated for his / her gene disorder and able to identify intrinsic and extrinsic factors.

Copyright to IJIRCCE DOI: 10.15680/IJIRCCE.2019. 0702020

We done the Co-Regulatory modules between miRNA (microRNA),TF (Transcription Factor) and gene on function level with multiple genomic data.. We compare the regulations between miRNA-TF interaction, TF-gene interactions and gene-miRNA interaction with the help of integration technique. These interaction could be taken the genetic disease like breast cancer, etc.**Multiplicative Updating Algorithm** is used in our project to solve the optimization module function for the above interactions. After that interactions, we compare the regulatory modules and protein value for gene and generate Bayesian rose tree for efficiency of our result[3].ABE encryption technique is used to encrypted the result and it will be stored in cloud[4].

II.RELATED WORK

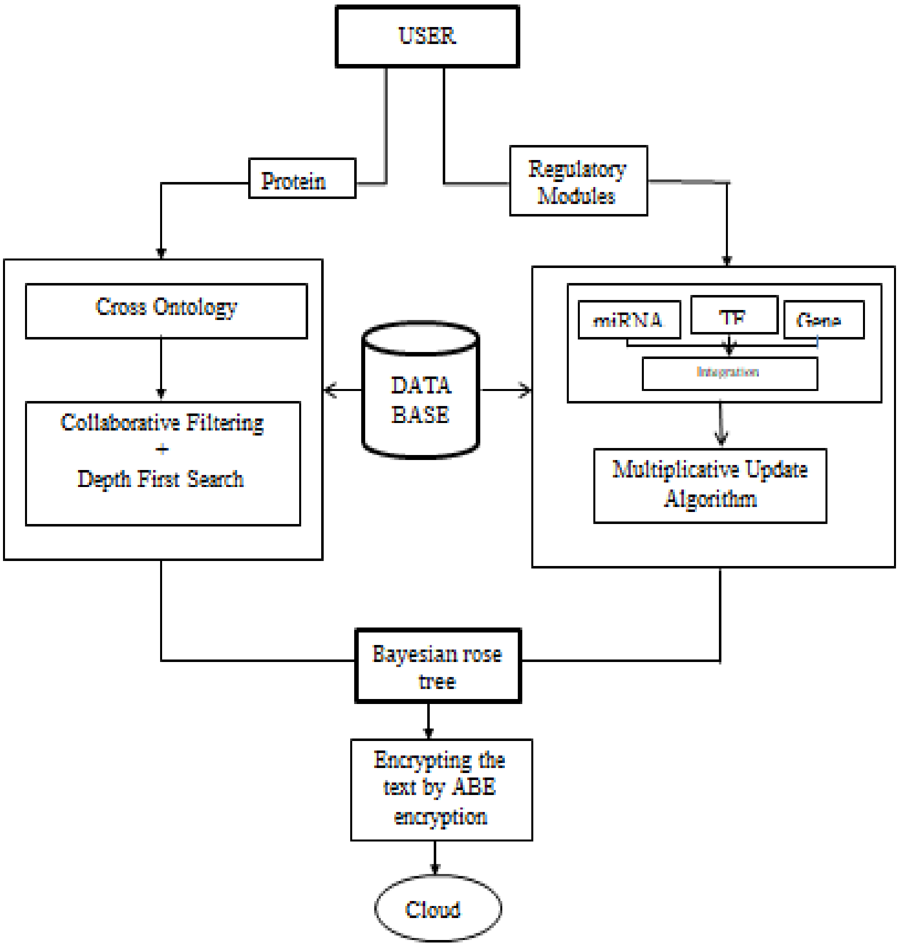
Hemert JV and Baldock R have stated that the association rule mining to identify relationships among up and down regulated genes in gene expression studies. These studies do not make use of the GO and its hierarchical structure[5].
Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O and Carazo JM, et al describes mine single level associations between GO annotations and expressed genes from microarray data integrated with GO annotation information. The approach does not utilize the inherent information provided by the GO structure thereby limiting the knowledge discovered [6].
Davis MJ, Sehgal MS and Ragan MA describe an approach for generalizing in the GO by calculating the information content of a node using both the ontology structure and the annotation dataset as a metric for generalization. They use a non-traditional definition of information content of a concept x as $I_x = P_x - O_x$, where P_x is the information gained by not generalizing concept x and O_x is the information lost if all the child terms of x are generalized to x. P_x and O_x are calculated using information from the annotation dataset and the ontology structure. They use this approach to generate automatic slim sets from the GO, but it is unclear how this approach will work for mining associations from multiple ontologies [7].
M. Hahsler, B. Grün and K. Hornikelaborates the use of AR presents two main issues due to the Number and the Nature of Annotations. The number of annotation is for each protein or gene is highly variable within the same GO taxonomy and over different species. The variability is caused by two main facts: (i) The presence of different methods of annotations of data; and (ii) the use of different data sources [8].
AnuraagVikram Kate and Harish Balakrishnan stated thatthe research in the field of Semantic web and Data Mining has advanced so much, that today their applications are tremendous in the domains of BioMedical, personalized e-Learning, Bioinformatics etc. In this paper a thorough investigation is done in the areas of Web mining and Semantic web, and the mixture of the techniques in these areas are used to simplify the complications in a specific domain. The domain that is focused here is of Molecular Biology (Lac Operon mechanism). Ontology mining is done, inorder to improvise the prevailing ontology functionally by the inclusion of DNA and RNA components, and also structurally by expanding the ontology in a different perspective. The analysis of these functional and structural modifications is done to overcome the present shortcomings in this ontology [9].

III.EXISTING SYSTEM

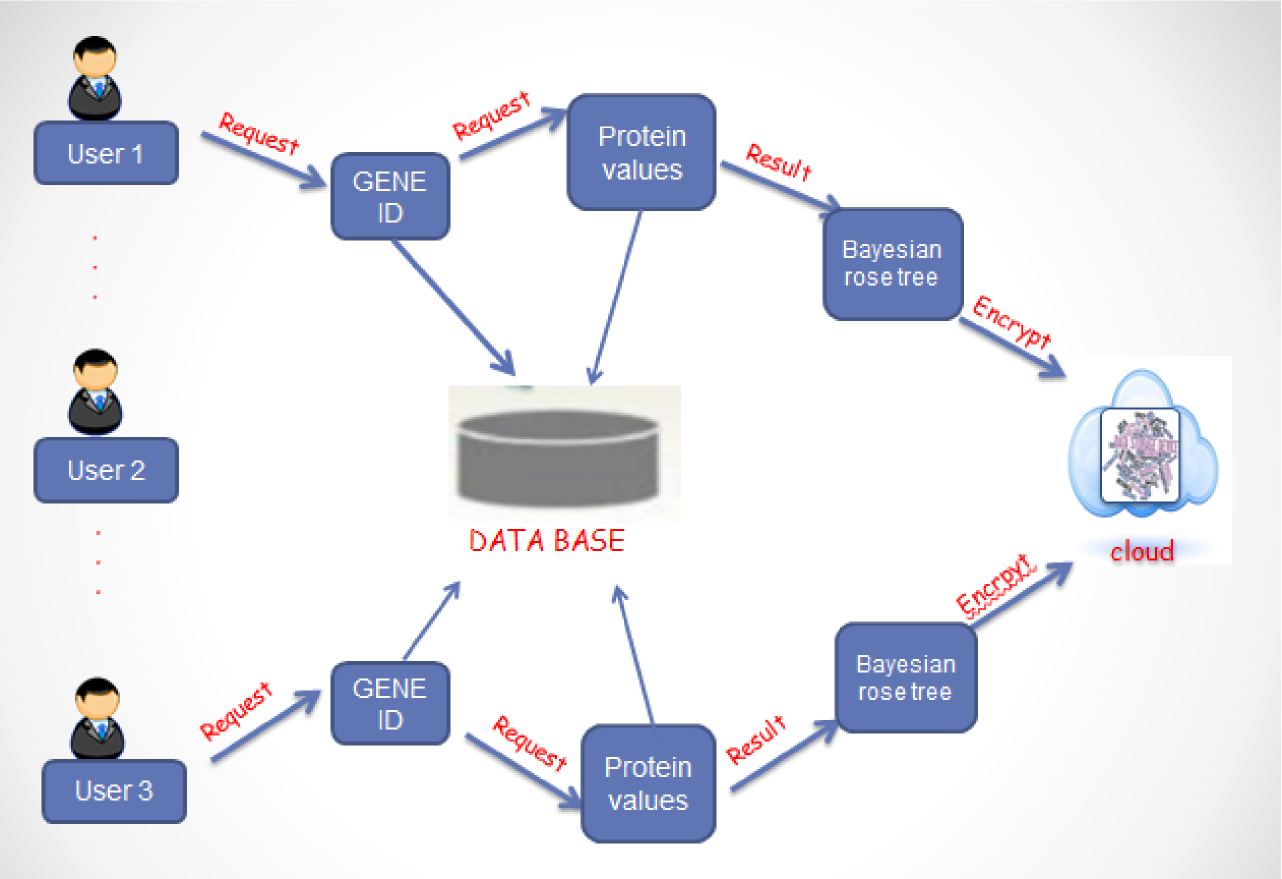
The existing system proposed association rules to support GO curators. It evaluates the annotation consistency in order to avoid possible inconsistent or redundant annotations. It uses the method called Classical association rules mining algorithms. It helps in understanding and answering the complex biological phenomena on multiple biomolecular information. The genomic and proteomic data is scattered, the biologists finds it difficult in accessing those data. There is no common knowledge base for gene analysis.

IV. PROPOSED SYSTEM

The co-regulatory modules is implemented between Transcription Factor, gene expression and MiRNA on functional level with genomic data. The integration technique is implemented between miRNA, Transcription Factor (TF) and gene. After integration, Iterative Multiplicating update algorithm is used to check the optimization function between the regulatory modules. The optimized result from algorithm will be given to the protein. With the help of cross ontology the protein value will be obtained from Biological Process, Molecular Function and Cellular Component. At last we generate a bayseian rose tree structure for the relation between regulatory modules and protein values of our gene. The tree structure helps in understanding the disease and the cure available for that disease. Also we provide food and available exercise for the particular disease.



4.1 Architectural diagram



4.2 System model

V.MODULES AND ITS DESCRIPTION

MODULES:

1. Gene Ontology
2. Collaborative Filtering
3. Depth First Search
4. Regulatory modules
5. Integration Technique
6. Multiplicative Update Algorithm
7. Tree Representation

1.GENE ONTOLOGY:

The most widely used framework in biology is Gene Ontology. The GO defines concepts used to describe gene function, and relationships between these concepts. The Gene Ontology classifies three aspects like molecular function and activities of gene product, cellular component in which the gene products are active and biological process pathways made up of the activities of multiple gene products.It provides the consistent description of gene products in heterogeneous database. In our project we are proposing gene ontology, User login and register their details and get the gene id from Ontology base with the help of KNN algorithm. Full details of overall project are maintained our database and ontology base. The cross ontology is used to manipulate the protein value to obtain the disease. Also our proposed system, focus on intrinsic and extrinsic. Based on *cellular component,* molecular *function*and*biological process*values intrinsic and extrinsic calculation would be manipulated.

2.COLLABORATIVE FILTERING:

In our Project we used semantic mining for logical analysis. User get the details from Ontology base with help of Collaborative filtering, also the gene disease and symptoms with the help of logical calculation for protein value of human and normal value for particular gene id, then cross ontology process we get the BP,CC&MF value for gene to identify the gene have Intrinsic or extrinsic.

I) 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.

II) EXTRINSIC:

If the normal protein value of human is high(comparing BP&CC or MF&CC or MF&BP) then it is said to be extrinsic.

MOAL (Multi ontology data mining at all levels)algorithm for mines the cross ontology relationship between the ontologies.MOAL algorithm to mine cross-ontology association rules.By using collaborative filtering, user get the details about the gene id for cross ontology technique we have to compare the protein value and getting BP& MF value, or MF&CC value or CC&BP value getting the gene disease and symptoms for user requirements.

3.DEPTH FIRST SEARCH:

Depth first search in relation to specific domains such as searching for solutions in artificial intelligence. The DFS will traverse to the entire nodes in the graph . In this cases search is performed to a depth due to limited resources. Such as memory or disk space one typically does not use data structures to keep track the set of all previous visited vertices.

When DF search is performed to a limited depth the time is still linear in terms to number of expanded vertices. Edges although this number is not the same as the capacity of the entire graph because some vertices may be searched more than once and others not at all. As a result is much smaller than the space needed for searching to the same depth using BFS. DF search also lends itself much better to heuristic methods for choosing a likely-looking branch.

4.REGULATORY MODULES:

The set of genes are responsible for regulating different conditions and it is organize in interacting modules network. For identifying regulatory modules from gene expression data we use probabilistic method. We identifies co-regulated genes modules, along with its conditions under which regulation occurs. A Saccharomyces cerevisiae method is used to expression data set which shows the ability to identify functionally coherent modules.

Multiple biological data are analysed to identify the affected cell in the gene. We considered among 856 human genes with putative roles in cell cycle regulation, we identified 56 transcription factors and 49 gene ontology groups. We reconstructed regulatory modules to infer the underlying regulatory relationships. Four regulatory network were used to identified from the interaction network. Each transcription factor and predicted target gene groups was examined by training a recurrent neural network to find the relationship between them.

5.INTEGRATION TECHNIQUE:

In this module, we use a fusion technique to integrate both gene ontology and regulatory modules. This is the first time we are proposing a fusion technique in gene analysis which produces increased accuracy.

6.MULTIPLICATIVE UPDATE ALGORITHM:

A novel approach to identify miRNAs and transcription factors co-regulatory modules (miRNA-TF-gene) is essential. To this end, an objective function is constructed by integrating the miRNA/TF/gene expression profiles, target site information (miRNA-gene and TF-gene regulations) as well as the protein-protein interactions. In order to obtain the optimal solution of the objective function, we solve the optimization model function effectively by iterative multiplicative updating algorithm.

7.TREE REPRESENTATION:

We briefly introduce the BRT algorithm. Bayesian hierarchical algorithm is used to produce the tree structure and each node in the tree is known as rose tree.BRT is a tree data structure and it is variable and unbounded number of branches per nodes.

VI.CONCLUSION

Relevant progresses in biotechnology and system biology are creating a remarkable amount of biomolecular data and semantic annotations; they increase in number and quality, but are dispersed and only partially connected. Integration a nd mining of thesedistributed and evolving data and information have the high potential of discovering hidden biomedi cal knowledge useful in understanding complex biological phenomena, normal or pathological, and ultimately of enhan cing diagnosis, prognosis and treatment; but such integration poses huge challenges. Our work has tackled them by dev eloping a novel and generalized way to define and easily maintain updated and extend an integration of many evolving and heterogeneous data sources; our approach proved useful to extract biomedical knowledge about complex biological processes and diseases.

REFERENCES

1.Giuseppe Agapito, Marianna Milano, Pietro H. Guzzi (2015)Improving annotation quality in gene ontology by mining cross-ontology weighted association rules.
2.Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi , Marianna Milano, Using GO-WAR for mining cross-ontology weighted association rules, Elsevier, 2015
3. Charles Blundell, Yee Whye The Gatsby Computational Neuroscience Unit, University College London, London, UK Katherine A. Heller Department of Engineering, University Cambridge, Cambridge, UK, Bayesian Rose Trees.
4.R.Nitya Lakshmi, R.Laavanya, M.Meenakshi, Dr.C.SureshGanaDhas, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Namakkal, India. Attribute Based Encryption Schemes.
5.Hemert JV, Baldock R (2007) Mining spatial gene expression data for association rules. Proceedings of the 1st international conference on Bioinformatics research and development. Berlin, Germany: Springer-Verlag. 66–76.
6.Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, et al. (2006) Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics 7: 54.
7.Davis MJ, Sehgal MS, Ragan MA (2010) Automatic, context-specific generation of Gene Ontology slims. BMC Bioinformatics 11: 498.
8.M. Hahsler, B. Grün, K. Hornik, *SIGKDD Explorations*, pp. 0-4.
9.AnuraagVikram Kate, Harish Balakrishnan (2014) Ontology mining in Molecular Biology domain (Lac Operon problem)