Predicting MicroRNA-Disease Associations Using Transcription Factors Based On Multiple Types Of Genomic Data

P. Bakiyalakshmi¹, B. Thamaraiindhu, P. Balakumar M.E., Ph.D³, P. Veeralakshmi M.E., Ph.D⁴

¹,² Student, ³,⁴ Associate Professor,

¹,² Information Technology, ³ Computer Science and Engineering

¹,²,³,⁴ Prince Shri Venkateshwara Padmavathy Engineering College

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. One of the analysis is the use of Association Rules (AR) which discovers biologically relevant associations between terms of GO. GO-WAR (Gene Ontology-based Weighted Association Rules) for extracting Weighted Association Rules from ontology-based annotated datasets. Adapting the MOAL algorithm to mine cross-ontology association rules, i.e. rules that involve GO terms present in the three sub-ontologies of GO. We are proposing cross ontology to manipulate the protein values from three sub ontologies for identifying the gene attacked disease, focusing on intrinsic and extrinsic. Based on cellular component, molecular function and biological process values intrinsic and extrinsic calculation would be manipulated. 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 interactions could be taken the genetic disease like breast cancer, etc... Iterative 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.

Keywords—micro-RNA(miRNA), transcription factor, co-regulatory module, genomic data, Gene Ontology-based Weighted Association Rules.

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. The ontology covers three domains: cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and biological process, 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. The introduction of high-throughput technologies in molecular biology has produced the accumulation of a large set of experimental data. Such amount of experimental data has been integrated with additional information able to explain such data. For instance, genes and proteins have been accompanied by the storing of additional information used for the elucidation of the role of the investigated molecules. In order to systematize such knowledge, formal instruments such as controlled vocabularies and ontologies have been used to manage the used terms. 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). 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 [1]. For instance, GO: 0006915 represents the apoptosis process.

Increasingly large amounts of valuable, but heterogeneous and sparse, biomolecular data and information are characterizing life sciences. In particular, semantic controlled annotations of biomolecular entities, i.e. the associations between biomolecular entities (mainly genes and their protein products) and controlled terms that describe the biomolecular entity features or functions, are of great value; they support scientists with several terminologies and ontologies describing structural, functional and phenotypic biological features of such entities [2](e.g. their sequence polymorphisms, expression in different tissues, or involvement in biological processes, biochemical pathways and genetic disorders).

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 biomedical knowledge. Recently, researchers studied the co-regulation of miRNAs and TFs by finding out their shared downstream targets [3]. The method adopts probabilistic models and statistical tests to measure the significance of the shared targets between the regulators, and to remove the insignificant co-regulating interactions that occurred by chance. Gene enrichment analysis was used in [4] to identify significant co-regulation between the transcriptional and posttranscriptional layers. They found that some biological processes merged only in co-regulation and that the disruption of co-regulation may be closely related to cancers, suggesting the importance of the co-regulation of miRNAs and TFs. Tran et al. [5] proposed a rule-based method to discover the gene regulatory modules that consist of miRNAs, TFs, and their target genes based on the available predicted target binding information.

In this paper, we propose a novel approach called to identify miRNAs and transcription factors co-regulatory modules (miRNA-TF-gene). 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 [6]. In order to obtain the optimal solution of the objective function, we solve the optimization model function effectively by iterative multiplicative updating algorithm. We get the expression or some value from this algorithm then compare to protein values. Based the protein values, we are going to provide the details of possible diseases, symptoms and cure for the particular geneid.

II. MATERIALS AND METHODS

We performed a joint analysis of miRNA, TF and gene variables to identify miRNA-TF-gene co-regulatory modules, TF and gene variables to identify miRNA-TF-gene co-regulatory module. MicroRNAs (miRNAs) and transcription factors (TFs), as two vital gene regulatory molecules in multicellular organisms, share a common regulatory logic [4]. MicroRNAs are a family of small, non-coding RNAs that regulate gene expression in a sequence-specific manner, which participate in the regulation of numerous cellular process at the post-transcriptional level, such as cancer progression. We downloaded the disease-related miRNA sets from HDMM [7] and disease-related gene sets from SemFunSim [8] and DisGeNET(http://www.disgenet.org/) for the cancer-specific miRNAs and mRNAs analysis. Multi-level association rule mining requires viewing the GO annotation transactions at multiple levels of abstraction. We have chosen to use a generalization strategy for ontology traversal where the level of abstraction of the annotations is increased one level at a time with the Apriori algorithm applied at each iteration [1]. The protein-protein interaction network was downloaded from BioGrid database (http://thebiogrid.org/) [9]. Meanwhile, we
obtained miRNA-gene interactions and TF-gene interactions from TargetScan (http://www.targetscan.org/) and ENCODE Project (http://encodenets.gersteinlab.org/) [10] we also obtained information from disease related details (https://www.omim.org/), diagnosis (www.rightdiagnosis.com) and other related information (www.dovemed.com/diseases-conditions/).

Figure 1. The flow chart of the retrieval of diseases, symptoms and cure, cross ontology at subontologies, miRNA/gene/TF expression profile analysis and generation of report and graph sorted in the cloud storage.

III. PROBLEM FORMULATION

3.1. Gene Ontology-based Weighted Association Rules

The GO-MOAL algorithm to mine cross-ontology association rules (three sub-ontologies of GO) [1]. Transaction is a gene and the items are GO terms from the three sub-ontologies assigned to that gene. An association rule is defined as an implication of the form \( x \rightarrow y \) where \( x \) (the antecedent) and \( y \) (the consequent) are disjoint subsets of items. Generating GO annotation candidates. Deriving new relationships between terms in the three sub-ontologies of the Gene Ontology(GO). Multi-ontology support is the probability of the two terms in the rule occurring together in the dataset background of the rule. Multi-ontology confidence of a rule is the probability of observing the consequent term given that the antecedent term is present in the dataset background of the rule.

3.1.1. Formulating Multi-ontology support and confidence

Multi-ontology support

The Multi-Ontology Support (MO Support) of a MO_ML rule, \( x \rightarrow y \) is defined as

\[
\text{MO Support } x \rightarrow y = \frac{|XYx \rightarrow y|}{|MO \text{ Category } x \rightarrow y|}
\]
Multi-ontology confidence
The Multi-Ontology Confidence (MO Confidence) of a MO_ML rule, $x \rightarrow y$ is defined as
\[
\text{MO Confidence } x \rightarrow y = \frac{XYx\rightarrow y}{Xx\rightarrow y}.
\]
MO Confidence of 0.70 [1].

Multi-ontology support is the probability of the two terms in the rule occurring together in the transaction background of the rule. Multi-ontology confidence of a rule is the probability of observing the consequent term given that the antecedent term is present in the transaction background of the rule. Initially we calculate the information content for each GO term [1] then, we extract weighted association rules. Semantic mining for logical analysis.

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. In our project we are proposing gene ontology, User login and register their details and get the gene id from Ontology base. Full details of overall project are maintained our database and ontology base. We are proposing cross ontology to manipulate the Protein values from three sub ontologies for identifying the gene attacked 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.

3.2 Collaborative filtering algorithm
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, get the BP, CC&MF value for gene to identify the gene have Intrinsic or extrinsic.

3.2.1 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.

3.2.2 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.

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, i.e. rules that involve GO terms present in the three sub-ontologies of GO [1]. By using collaborative filtering, user get the details about the gene id for cross ontology technique, 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. Depth first search algorithm allows to formulate the possible solutions for determining the diseases and other factor details. Space needed for searching is less than that of breadth first search. When DF search is performed to a limited depth the time is still linear in terms to number of expanded vertices. The space complexity of this variant of DFS is only proportional to the depth limit.

3.3 Instance based algorithm
In order to mine all possible disease factor details for the particular geneid from the database, predictive modeling is applied. This non-parametric algorithm allows to predict new instance $x$ by searching through the entire data set for the $k$ most instances and summarizing the output variable for the $k$ most instances. Distance metric for continuous variables is by Euclidean distance.
Classification allows to classify the data and searching through the entire dataset based on the instance value. Regression provides the most common value form the large dataset. Previously unseen data can be gathered to obtain more accurate details and also allows to find the similarities between instances. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Y is the class label of X given some norm on and a point, it allows to reordering of dataset and provides instance-based disease details to the user. It also removes erroneous details within the database for efficient data retrieval.

The class label of X, so that X|y = r ~ P_r for r = 1, 2 (and probability distributions P_r) given some normalization on R_d let (X (1), Y (1), ..., X (n), Y (n)) be a reordering of the training data such that ||X (1) - x|| <= ... <= ||X (n) - x||.

Complexity of Instance based algorithm is O(n).

3.4 Iterative Multiplicative Updating method

To identify co-regulatory modules, on-negative matrix factorization (NMF) are a powerful method for data reduction and clustering that has widely been used to analyse high-throughput genomic data [11]. Much of a cell's activity is organized as a network of interacting modules: sets of genes co-regulated to respond to different conditions. We present a probabilistic method for identifying regulatory modules from gene expression data. Our procedure identifies modules of co-regulated genes, their regulators and the conditions under which regulation occurs, generating testable hypotheses in the form ‘regulator X regulates module Y under conditions W’. We propose an integrative framework that infers gene regulatory modules from the cell cycle of cancer cells by incorporating multiple sources of biological data, including gene expression profiles, gene ontology, and molecular interaction. Among 846 human genes with putative roles in cell cycle regulation, we identified 46 transcription factors and 39 gene ontology groups.

Sparsity penalization are given to manipulate microRNA iterative Multiplicative update algorithm allows to update and optimise the values.

Here we provide the final optimization function.

\[
\begin{align*}
\min & \ W, H_1, H_2, H_3 = \sum_{i=1}^{3} \|X_i - W H_i\|_F^2 - \lambda_1 T_r (H_2 A H_2^T) \\
& \quad - \lambda_2 T_r (H_1 B H_2^T) - \lambda_3 T_r (H_3 C H_2^T) \\
& \quad + \gamma_1 \|W\|_F^2 + \gamma_2 (\sum_j \|h_j\|_1^2 + \sum_{j'} \|h_{j'}\|_1^2) \\
& \quad + \sum_{j''} \|h_{j''}\|_1^2, \\
\text{W} & \geq 0, \text{H}_i \geq 0, i = 1, ..., 3 \text{ where expression matrix X1,2,3 is decomposed by basic matrix W with size of N \times K and coefficient matrix H1,2,3 with size of K \times M. The following subsections will provide the details as well as the solution of objective function.}
\end{align*}
\]

Necessary to analyze the modules distributions and its topological characteristics. Accordingly, we analyze the modules densities to observe the tightness of modules. The density function is defined as 2l/n(n−1), where l is the number of edges in the module, n represents the number of nodes including miRNAs, TFs and genes. We further employed miRNA-gene expression
correlation (MiMEC) for each module to compute the sum value of (anti)-correlations between miRNAs and genes. Based on these values the associated genes and other factors associated with the modules are determined. It provides more details about the cancer related diseases factor details.

3.5 Bayesian rose tree

A Bayesian network is a directed acyclic graph which is drawn to represent the overall characteristics of gene details manipulated. Context-Specific Independent representation of data allows to formally characterize this structured representation and catalogue a information processed. From the gene-id from the user, the manipulation of intrinsic/extrinsic characteristics od gene-id possible diseases and the symptoms and cure details the diseases are drawn in the form graph shown in figure 2. Generated report is stored in the cloud in order provide the access to the document in the efficient a manner. It allows the consultant to analysis the information of the patient details in the secure manner and to provide the details to the user through the web service, since the information automatically stored in the cloud which is dynamic updatable and also provides protected access of data.

Client as to register the details in the webservice with name and password, so that details can be accessed by only authorised users.it forms the two-factor authentication of details. Only authorised consultant and users can access the information.

![Figure 2](image)

**Figure 2. Graphical representation of gene-id and characteristics of gene, disease, symptoms.**

V. ANALYSIS

We reconstructed regulatory modules to infer the underlying regulatory relationships. Four regulatory network motifs were identified from the interaction network. The relationship between each transcription factor and predicted target gene groups was examined by training a recurrent neural network whose topology mimics the network motif(s) to which the transcription factor was assigned [13]. Inferred network motifs related to eight well-known cell cycle genes were confirmed by gene set enrichment analysis, binding site enrichment analysis, and comparison with previously published experimental results. It allows to provide details regarding the micro RNA diseases, symptom and cure details applying methods and algorithms in fig 1.
Based on the geneid which is the unique id given the specific genes based on the sub-ontologies biological function, cellular components and molecular process manipulation by cross ontology technique allows to determine the characteristics the genes whether it is extrinsic or intrinsic. Micro-RNA and transcription factor analysis allows to manipulate micro disease associations based on the details list of possible diseases and cure symptom and cure are provided through the web application.

We have predicted the microRNA diseases based on the genomic data microRNA diseases, symptoms and cure factor details that are applicable for gene can be determined shown fig2. Medical description for particular gene disease can be easily accessible by uploading to the cloud. Cloud allows the fast retrieval of data and fast access to data. This method allows to predict the diseases in effective manner decreases the complexity. Preplanned medical description saves upcoming generations. It is easy to maintain extend and update the application.

Figure. 3 Geneid and characteristics of gene, disease, symptoms and cure factors

V. ACKNOWLEDGEMENT

We wish to acknowledge and our gratitude to head of department, Associate professor, Mrs. P. Veeralakhmi, M.E., Department of Information Technology, for her support and Dr. P. Balakumar, Ph.D., Associate Professor, Department of Computer Science and Engineering, for her guidance throughout the course of our project. To see the list of contributors, please refer to the top of file in the distribution.

IV. CONCLUSION AND FUTURE WORKS

Relevant progress in biotechnology and system biology are creating a remarkable amount of biomolecular data and semantic annotation [14]. They increase in number and quality, but are dispersed and only partially connected. Integration and mining of distributed and evolving data and information have the high potential of discovering hidden biomedical knowledge useful in understanding complex biological phenomena, normal or pathological and ultimately of enhancing
diagnosis, prognosis and treatment, but such integration poses huge challenges. By applying data mining algorithms on genomic data, micro RNA diseases, symptoms and cure factor details that are applicable for our gene are provided. Our work has tackled them by developing a novel and generalized way to define and easily maintain update and extend evolving heterogeneous data sources our approach proved useful to extract biomedical knowledge about complex biological processes and diseases. Enhancements with application can by providing the additional information by analysis of other factors of diseases and report with visual interactions and query interfaces.

REFERENCES
I. P. Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi *, Marianna Milano, Using GO-WAR for mining cross-ontology weighted association rules, Elsevier, 2015.
II. 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
III. Y. Zhou, J. Ferguson, J. T. Chang, and Y. Kluger, “Inter-and intra-combinatorial regulation by transcription factors and microRNAs,” BMC genomics, vol. 8, no. 1, p. 1, 2007.
IV. C.-Y. Chen, S.-T. Chen, C.-S. Fuh, H.-F. Juan, and H.-C. Huang, “Coregulation of transcription factors and microRNAs in human transcriptional regulatory network,” BMC bioinformatics, vol. 12, no. 1, p. 1, 2011.
V. D. H. Tran, K. Satou, T. B. Ho, and T. H. Pham, “Computational discovery of miRNA regulatory modules in human genome,” Bioinformation, vol. 4, no. 8, pp. 371–377, 2010
VI. O. Hobert, “Gene regulation by transcription factors and microRNAs,” Science, vol. 319, no. 5871, pp. 1785–1786, 2008
VII. Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, “Hmdd v2.0: a database for experimentally supported human microRNA and disease associations,” Nucleic acids research, p. gkt1023, 2013.
VIII. L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, “Semfunsim: A new method for measuring disease similarity by integrating semantic and gene functional association,” PLoS ONE, vol. 9, no. 6, p. e99415, June.
IX. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “Biogrid: a general repository for interaction datasets,” Nucleic acids research, vol. 34, no. suppl 1, pp. D535–D539, 2006.
X. E. P. Consortium, “The encode (encyclopedia of dna elements) project,” Science, vol. 306, no. 5696, pp. 636–640, 2004.
XI. S. Yoon and G. D. Micheli, “Prediction of regulatory modules comprising microRNAs and target genes,” Bioinformatics, vol. 21, no. 2, pp. i93–i100, 2005.
XII. J. LU, G. Getz, and E. A. Miska, “MicroRNA expression profiles classify human cancers,” nature, vol. 435, no. 9, pp. 834–838, Jun 2005.
XIII. Jiawei Luo, Gen Xiang and Chu Pan “Discovery of microRNAs and transcription factors co-regulatory modules by integrating multiple types of genomic data” IEEE TRANSACTION ON NANOBIOSCIENCE, VOL.16, NO.1, JAN.2017.
XIV. 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.