Gene Mutation Data using Multiplicative Adaptive Algorithm and Gene Ontology

Jeevitha K[1], Sirushtika A[2], Vijayalakshmi K[3], Anitha moses[4]

[1,2,3]Student, Department of Computer Science and Engineering, Panimalar Engineering College, Tamil Nadu, India
[4]Associate Professor, Dept. of CSE, Panimalar Engineering College, Tamil Nadu, India

***

Abstract - *Gene Ontology (GO) concepts are associated to one or more gene products through a process referred to as annotation. There are different approaches of analysis to get bio information. 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) for extracting Weighted Association Rules from ontology-based annotated datasets is used. 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. Cross ontology is proposed to manipulate the Protein values from three sub ontologies for identifying the gene attacked disease. The proposed system, focus on intrinsic and extrinsic. Based on cellular component, molecular function and biological process values intrinsic and extrinsic calculation would be manipulated.*

The Co-Regulatory modules between miRNA(microRNA), TF (Transcription Factor) and gene on function level with multiple genomic data. Compare the regulations between miRNA-TF interaction, TF-gene interactions and gene-miRNA interaction with the help of integration technique. This interaction could be taken the genetic disease like breast cancer, etc. Iterative Multiplicative Updating Algorithm is used to solve the optimization module function for the above interactions. After that interactions, compare the regulatory modules and protein value for gene and generate Bayesian rose tree for efficiency of result.

Key Words: **Gene Ontology, Sub Ontologies, Depth First Search, Regulatory modules, Integration Technique, Multiplicative Update Algorithm, Tree Representation**

1. INTRODUCTION

Usage of computer and technology, huge amount of medical data creates a huge scope for data mining techniques. Data mining techniques are widely used and are popular among the medical research groups. To obtain the solution from large amount of knowledge base, relationship / association among the variables, predict a specific disease based on historical datasets, assigning weightage to the variable etc data mining techniques are used. To identify the genetic disorders the objective is to develop a common knowledge base for genomic and protein patterns. Also, integration of graph-based clustering and Bayesian rose tree would provide an efficient and easy solution for representing the gene terms.

There is large scope for gene analysis due to increasing huge amount of bimolecular valuable data and information in life sciences. The DNA compromises of gene and proteins. Identifying the association between the biomolecular entities is focused by gene analysis. Thus, in existing system the multiple genomic data are scattered in many distributed systems. In our proposed architecture, a common knowledge base for genomic and proteomic analysis is to be developed, which can be accessed by doctors, scientists, researchers and others to provide solution for more genetic disorders. To produces biologically meaningful data and solutions analysis of gene and protein data provides vital opportunity for bioinformatics domain which. To systematize the knowledge base ontology methods and techniques are used to handle the gene terms.

The proposed architecture helps in understanding complex biological patterns and associations. For grouping of same gene information for a particular gene disorder graph clustering approach is been used. A group of similar data elements or data elements somewhere interconnected is called clustering. Graphs are nothing but structures, combination of set of vertices and set of edges. Graph clustering is defined as identifying and grouping the vertices from the input graph into clusters. Graph clustering technique is quite popular and widely used for data clustering.

Gene ontology is a collaborative process of work to identify the need and descriptions about the gene products along its databases. The process of collaboration started with three model organisms. The GO process has grown on a high rate by incorporating many databases, which includes world's major repositories for plant, animal, and microbial genomes. There are separate aspects for maintenance of all the gene products. Every structure of the gene products is assigned with separate properties. The scope of GO is that protein- protein interactions. And also, anatomical features above the level of cellular components includes all cell types. There are different types of Ontology to elucidate different fields. GO is the largely used frameworks around the world.

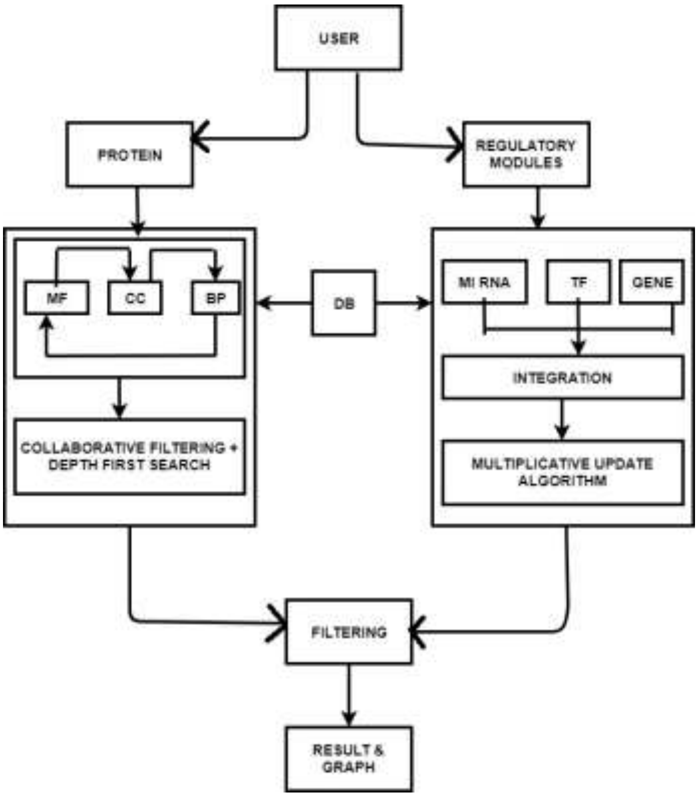

Fig-1: Architecture Diagram

There are three main Go sub ontologies and they are Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Each ontology has separate GO Terms which are used for describing the functions. There is separate code for every term of Gene and also a textual description for the same. The biological process of the Gene is performed through a process known as annotation which are stored in databases which is known as the Gene Ontology Annotation database (GOA). GO evaluates the annotation consistency to avoid the inconsistent of the process. Weighted-association rule mining was introduced to identify the gene associations in proposed system. Thus, 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, in the proposed system microRNA (miRNA), gene and transcription factor (TF) are considered for gene regulation. the relationship between miRNA and TF are identified for particular genetic disorder. The proposed work is web application based.

1.1 Related Work

Hongbao Cao[3], has proposed a sparse representation based clustering (SRC) method for integrative data analyses, and applied the SRC method to the integrative analysis of 376821 SNPs in 200 subjects (100 cases and 100 controls) and expression data for 22283 genes in 80 subjects (40 cases and 40 controls) to identify significant genes for osteoporosis (OP).

Mingon Kang(2015)[16], introduced a multi-block bipartite graph and its inference methods, MB2I and sMB2I, for the integrative genomic study.

Mingon Kang(2016)[10], the proposed methods not only integrate multiple genomic data but also incorporate intra/inter-block interactions by using a multi-block bipartite graph.

Jiawei Luo (2017)[4], propose a method called SNCoNMF (Sparse Network regularized non-negative matrix factorization for Co-regulatory modules identification) which adopts multiple non-negative matrix factorization framework to identify co-regulatory modules including miRNAs, TFs and genes.

Arif Canakoglu(2013)[1], developed a software architecture to create and maintain updated a Genomic and Proteomic Data Warehouse (GPDW), which integrates several of the main of such dispersed data. It uses a modular and multi-level global data schema based on abstraction and generalization of integrated data features.

Xiaoyu Jiang[5], developed Bayesian framework combining relational, hierarchical and structural information with improvement in data usage efficiency.

2. METHODOLOGY

A) find nearest neighbours of geneid (KNN algorithm)

KNN algorithm is used for classification and regressive predictive problem. There are three most important aspect,

1.Ease to interpret output

2. Calculation time

3. Predictive Power

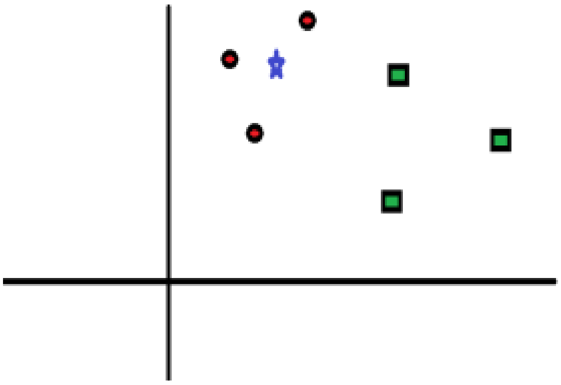

Fig-2: KNN Example

“K” in KNN algorithm is the nearest neighbours.

Gene id is given and find the nearest neighbours using this algorithm from the database.

Following is the curve for the training error rate with varying value of K:

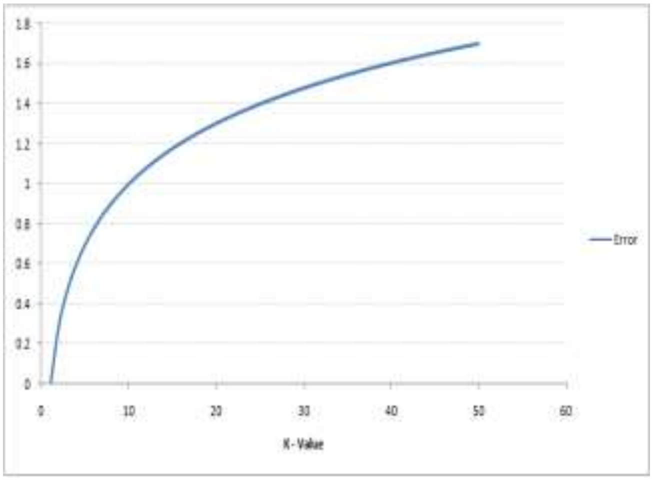

Chart-1: Error Rate

Following is the validation error curve with varying value of K;

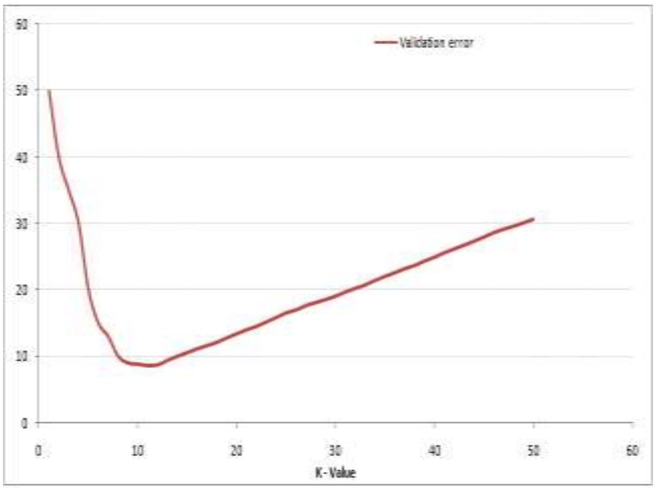

Chart-2: Validation Error

Algorithm:

```
k-Nearest Neighbour

Classify (X, Y, x)//X: training data, Y: class labels of X, x: unknown sample

for i=1 to m do

        Compute distance d(Xi, x)

End for

Compute set I containing indices for the k smallest distances d (Xi, x)

Return majority label for {Yi where I £ I
```

Intrinsic diseases

Intrinsic disease is a disease not caused directly by outside (extrinsic) factors, but by internal factors.eg. Intrinsic asthma, hemophilia.

Extrinsic diseases

Extrinsic disease is a disease caused directly by outside (extrinsic)factors. e.g. diseases caused by bacteria, fungi, parasite etc.,

B) Iterative multiplicative updating algorithm

Initialization:

Fix an ≤1/2. For each expert, associate the weight ≔1. For = , ,…,

Algorithm:

1. The weighted majority of the experts' predictions based on the weights is to be predicted. Depending on which prediction has a higher total weight of experts advising it, makes a binary prediction (breaking ties arbitrarily).
2. For every i predicts wrongly, decrease his weight for the next round by is obtained by multiplying it by a factor of (1-η): = (update rule).

C) SNCoNMF rule

SNCoNMF (Sparse Network regularized non-negative matrix factorization for Co-regulatory modules identification) used in

- Multiple non-negative matrix factorization framework
- To identify co-regulatory modules including miRNAs, TFs and genes.

3. CONCLUSIONS

Relevant progress in biotechnology and system biology is creating a remarkable amount of bimolecular data and semantic annotations. They increase in number and quality but are dispersed and only partially connected. Integration and mining of these distributed and devolving data and information have the 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. Our work has tackled them by developing a novel and generalized way to define and easily maintain updated and extended integration of many evolving and heterogeneous data sources. Our approach proved useful to extract biomedical knowledge about complex biological processes and diseases.

REFERENCES

[1] Arif Canakoglu, Marco Masseroli, Stefano Ceri, Luca Tettamanti, Giorgio Ghisalberti, and Alessandro Campi, "Integrative Warehousing of Biomolecular Information to Support Complex Multi-Topic Queries for Biomedical Knowledge Discovery",2013

[2] C. H. Ooi, H. K. Oh, H. Z. Wang, A. L. K. Tan, J. Wu, M. Lee, S. Y. Rha, H. C. Chung, D. M. Virshup, and P. Tan, "A densely interconnected genome-wide network of micro ran's and on cogenic path ways revealed using gene expression signatures," PLoS genetics, vol. 7, no. 12, p. e1002415, December 2011.

[3] Hongbao Cao, Shufeng Lei, Hong-Wen Deng, and Yu-Ping Wang, Identification of Genes for Complex Diseases by Integrating Multiple Types of Genomic Data, 2012

[4] Jiawei Luo, Gen Xiang and Chu Pan, "Discovery of microRNAs and transcription factors co-regulatory modules by integrating multiple types of genomic data", 2017

[5] Jia-Juan Tu, Le Ou-Yang, Xiaohua Hu and Xiao-Fei Zhang, Inferring gene network rewiring by combining gene expression and gene mutation data,2016

[6] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," PLoS Comput Biol, vol. 4, no. 7, p. e1000029, 2008

[7] K. Mohan, P. London, M. Fazel, D. Witten, and S. s. Lee, "Node-based learning of multiple gaussian graphical models," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 445–488, 2014.

[8] L. Ou-Yang, D. Q. Dai, X. L. Li, M. Wu, X. F. Zhang, and P. Yang, "Detecting temporal protein complexes from dynamic protein-protein interaction networks," BMC Bioinformatics, vol. 15, no. 1, p. 335, 2014.

[9] M. J. Ha, V. Baladandayuthapani, and K. A. Do, "Dingo: differential network analysis in genomics," Bioinformatics, vol. 31, no. 21, pp.3413–3420, 2015.

[10] Mingon Kang, Juyoung Park and Dong-Chul, "Multi-Block Bipartite Graph for Integrative Genomic Analysis", 2016

[11] M. Ye, X. Zhang, N. Li, Y. Zhang, P. Jing, N. Chang, J. Wu, X. Ren, and J. Zhang, "Alk and ros1 as targeted therapy paradigms and clinical implications to overcome crizotinib resistance," Oncotarget, vol. 7, no. 11, p. 12289, 2016.

[12] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 76, no. 2, pp. 373–397, 2014.

[13] Siva Ratna, Kumari Narisetti and Shuai Zeng, "Development of KB commons"-universal informatics framework for multi-omics translational research

[14] 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.

[15] Y. Zhang, Z. Ouyang, and H. Zhao, "A statistical framework for data integration through graphical models with application to cancer genomics," The Annals of Applied Statistics, vol. 11, no. 1, pp. 161– 184, 2017

[16] Mingon Kang, J uyoung Park, Dong-Chul Kim, Ashis K. Biswas, "An Integrative Genomic Study for Multimodal Genomic Data Using Multi-Block Bipartite Graph",2015.