Genetic Disease Identification and Medical Diagnosis using MF, CC, BP, MicroRNA and Transcription Factors

J. Jensy Celestina[1] R. Karthiga[2] Dr. Adlin Sheeba[3]
[1,2]UG Student [3]Professor
[1,2,3]Department of Computer Science & Engineering
[1,2,3]St.Joseph's Institute of Technology, Chennai – 600 119, India

Abstract— There are a lot of upcoming and new evolving diseases that the humans cannot predict. These diseases are found and analysed in the last stages and feel helpless in saving the humans. If we could determine the disease and predict in the early stages, human lives can be saved much more than we can imagine. So Genomic and Proteomic Dataset (GPD) is created, which integrates the most relevant sources of bio information. This dataset helps the user to find their genetic diseases that they obtained from their families in advance. The hospitals can make use of these data to keep track on their patient's health. Cross ontology is used to evaluate the gene test values from three ontologies for determining the genetic disease. We compare the cellular component, molecular function and biological process values between the input values and the average values. If any two of the input values are greater than the normal genetic values then the person is prone to extrinsic diseases or else intrinsic diseases. The diseases obtained as the result of cross ontology is integrated with the miRNA – transcription factor interactive diseases using the fusion technique. Disease updation is used to optimize the diseases from the above interactions and update the diseases whenever new diseases are encountered. Tree Representation gives us the complete structure of genetic diseases identified by various process, their symptoms and cure for the particular diseases to the associated user.

Key words: MF, CC, BP, MicroRNA, Transcription Factors, Genetic Disease, Medical Diagnosis

I. Introduction

Many genomic and proteomic data are scattered in many distributed and heterogeneous data sources. This is due to the developing technologies in the field of medical research. There are different approaches of analysis to get bio information[1]. Data mining techniques are used to gather information from various data sources. This is the most popular technique used in the medical field[5]. The genomic and proteomic dataset that we created contains all the bio information necessary for the biologist to diagnose the genetic diseases. These information are gathered from a lot of databases and websites all over the world[2]. Cross ontology provides the list of intrinsic and extrinsic diseases according to the molecular function, biological process and cellular components[4]. The diseases from the cross ontology is ordered in the database according to the severity of their impact on the human body [3]. These diseases are integrated with the diseases of miRNA values and Transcription Factor (TF) values by the fusion technique. This fusion technique provides us more accurate diseases. Then the diseases are ordered by their priority and finally the gene attacked diseases for the person is found and represented as a tree[6]. This helps every individual human being to determine their genetic diseases without any difficult analysis.

II. Proposed Methodology

A. System Architecture

In Fig.1, the geneID is obtained from the user and the associated geneID is retrieved from the dataset along with the diseaseID and disease names. Filtering based on diseaseID is performed on all the diseaseID related to the associated geneID and stored in the database. Then the Molecular Function (MF), Biological Process (BP), Cellular Components (CC) values are obtained from the user and cross ontology is performed to differentiate between extrinsic and intrinsic diseases. Fusion technique is used to integrate the diseases obtained out of the values from miRNA, transcription factor and cross ontology. Disease Updation is used to update the recent diseases according to the values. Finally the genetic diseases, their preventive measures, symptoms are displayed in the form of tree.

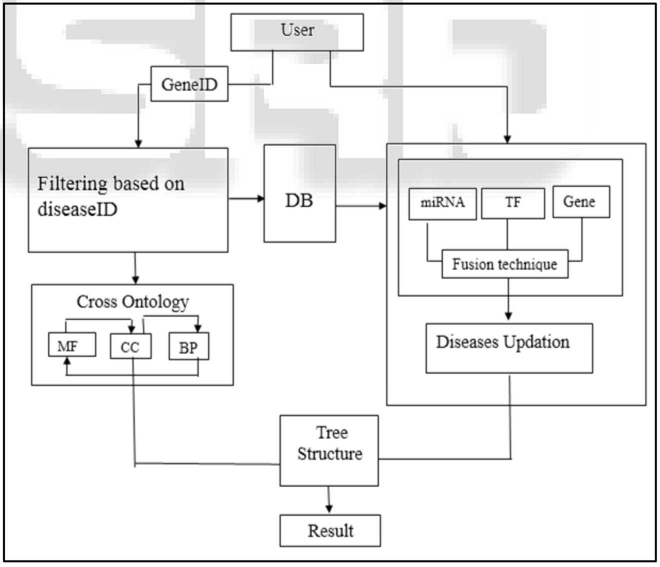


Fig. 1: Genetic Disease Identification and Medical Diagnosis

B. Methodology Used

1) Gene Ontology

User enters their geneID and their molecular function values, biological process values and cellular component values in order to find whether they have intrinsic or extrinsic diseases as in Fig.2. In cross ontology MF, BP, CC values of the users geneID is compared with Related geneID's(RgID) in the database and average values for MF, BP, CC is found.

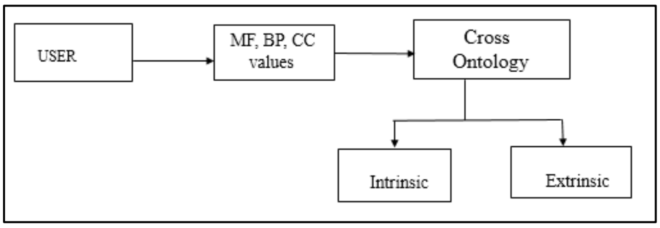


Fig. 2: Flow diagram of gene ontology

If the normal protein value of human is lower than that of calculating cross ontology value is said to be Intrinsic. If the normal protein value of human is higher than that of calculating cross ontology value is said to be extrinsic and the algorithm is given below:

```
get MF, BP, CC values from user
   find average values of MF, BP, CC from RgID
        if user(MF, BP, CC) greater than RgID(MF, BP, CC)
            display extrinsic diseases
        else
             display intrinsic diseases
```

C. Filtering Based on DiseaseId

The geneID obtained from the user is matched against the database and the related geneID, disease name, diseaseID is retrieved as shown in Fig.3. The filtering based on diseaseID is performed on the retrieved diseaseID and the sorted disease list is obtained as the result.

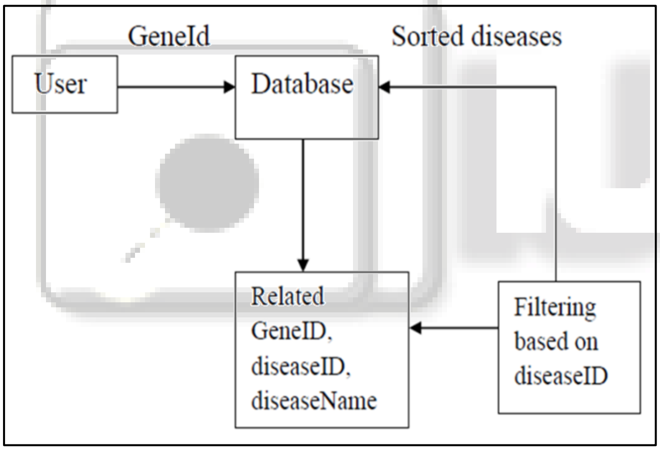

GeneId
Sorted diseases
User
Database
Related GeneID, diseaseID, diseaseName
Filtering based on diseaseID

Fig. 3: Filtering based on diseaseID

The Pseudo code for filtering based on diseaseID is given below:

```
for each geneID check RgID
   if RgID exists
      select diseaseID(RgID)
      for each diseaseID(RgID)
            if diseaseID(RgID) equals diseaseID(RgID)
            increment
display diseaseID(RgID) in descending order
```

D. Fusion Technique

In this module, we use a fusion technique to integrate both cross ontology and miRNA-TF interactive diseases. The resulting diseases obtained from the miRNA, TF values are integrated along with the diseases obtained from the cross ontology. In Fig.4, the disease A represents the disease from miRNA and TF values and disease B represents the diseases from the cross ontology values.

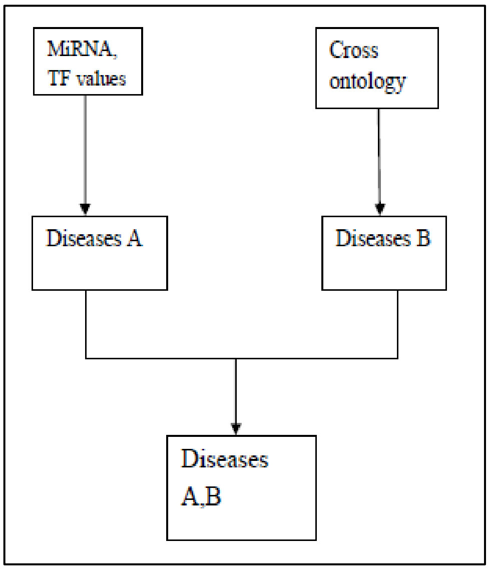


Fig. 4: Fusion Technique

E. Disease Updation

An approach to identify miRNA and transcription factors co-regulatory modules (miRNA-TF-gene) is essential. The diseases common between miRNA, TF and the cross ontology's result is obtained first and then the diseases common between any two of these is obtained second and the remaining diseases are obtained next. As the value changes the result keeps on updating.

F. Tree Structure

Tree gives us the complete structure of the process performed in the system along with the genetic diseases found, their symptoms, images and the preventive measures. And also the detailed text document of the diseases for the future use by the people.

III. EVALUATION

The graph gives us the relationship between the values obtained from the user and the average values that we found from the system. If any two values obtained from the user is greater than the average value then user will be attacked by extrinsic gene diseases (see Fig.5).From the graph (see Fig.6) we can easily find if any two values are less than the average values then it is intrinsic diseases. If the input values are equal to the average values then the person is normal and free from any genetic diseases.

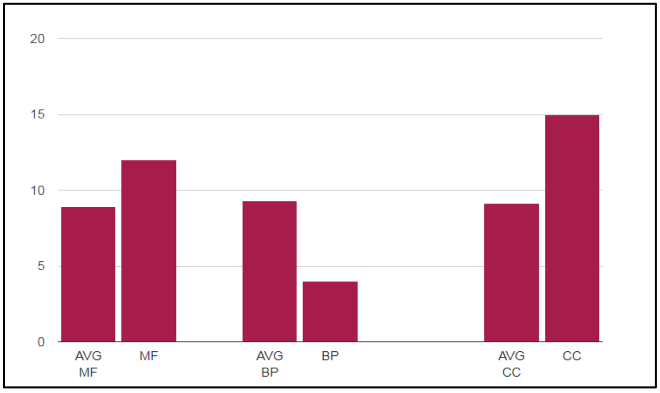


Fig. 5: Extrinsic Disease

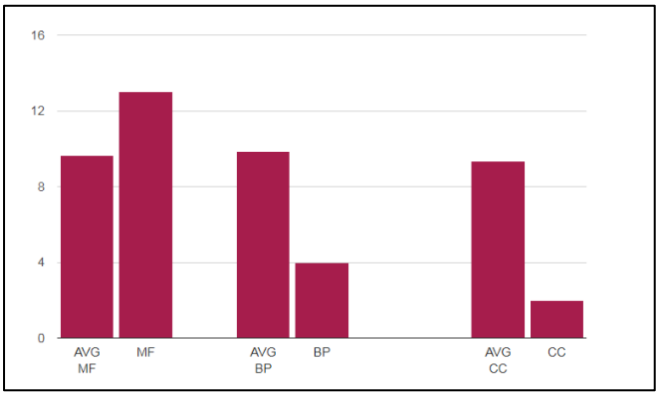
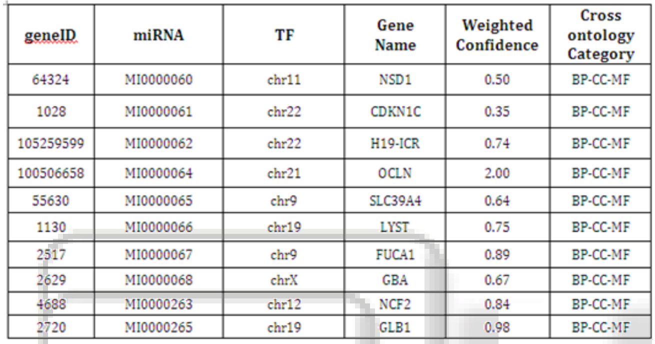


Fig. 6: Intrinsic Disease

In Table 1, the miRNA and tf values for the related gene ID's is difficult to identify. But by our system we have identified the miRNA and tf values for the geneID and also cross ontology category.



geneID	miRNA	TF	Gene Name	Weighted Confidence	Cross ontology Category
64324	MI0000060	chr11	NSD1	0.50	BP-CC-MF
1028	MI0000061	chr22	CDKN1C	0.35	BP-CC-MF
105259599	MI0000062	chr22	H19-ICR	0.74	BP-CC-MF
100506658	MI0000064	chr21	OCLN	2.00	BP-CC-MF
55630	MI0000065	chr9	SLC39A4	0.64	BP-CC-MF
1130	MI0000066	chr19	LYST	0.75	BP-CC-MF
2517	MI0000067	chr9	FUCA1	0.89	BP-CC-MF
2629	MI0000068	chrX	GBA	0.67	BP-CC-MF
4688	MI0000263	chr12	NCF2	0.84	BP-CC-MF
2720	MI0000265	chr19	GLB1	0.98	BP-CC-MF

Table 1: The miRNA, tf, gene, cross ontology for few geneID's.

IV. Conclusion & Future Work

The progresses in biotechnology and genetic engineering has led to the increase in number of homogeneous and heterogeneous data. Our work has tackled them by developing a novel and generalized way to define and easily maintain the updated information 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. Scholars know that common diseases and even rare diseases can run in families. If one generation of a family has heart diseases then the next generation is likely to have the same diseases. This is because of genes of the family so a family history can be used as another diagnostic tool and help guide decisions on genetic testing for the patient and at-risk family members. Family history holds important information about the family members past and future life and we aim to include family history as a part of our system.

References

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

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

[3] Marco Masseroli, Arif Canakoglu, and Stefano Ceri, "Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction", IEEE 2016.

[4] 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, 2015.

[5] Shweta Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", IJCSEIT, Vol.2, No.2, April 2012.

[6] Yangqiu Song, Shixia Liu, Xueqing Liu and Haixun Wang, "Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees", IEEE 2015.