International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

PREDICTION SYSTEM FOR GENETIC DISEASES USING GENOME MAPPING AND ITERATION OF REGULATORY MODULES

[1]Niveditha.S, [2]Sushmit Chakraborty, [3]Amit Jaiswal, [4]Vrindha Ramesh
[1]Assistant Professor, Department of Computer Science and Engineering,
SRM Institute of Science and Technology
[2,3,4]Student, Department of Computer Science and Engineering,
SRM Institute of Science and Technology
Corresponding Author: vrindha96@gmail.com

Abstract

Gene Ontology (GO) has a depository of existing concept that is structured which is associated to genes and it's genomic products through a process which is called Annotation. There are many approaches that are unique to the analysis for getting bio-information. One of these is using Association Rules (AR). This discovers biologically matching associations among GO terms. Here, we adapt MOAL algorithm for mining cross-ontology association rules, which is to find rules that are Gene Ontology terms that are present in three sub-ontologies. We are proposing cross ontology to manipulate the Protein values from three sub ontologies for identifying the gene attacked disease. Also our proposed system, focuses on intrinsic and extrinsic. Based on cellular component, molecular function and biological process (10) values intrinsic and extrinsic calculation are made. In this article, we have used 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: Gene Ontology (GO)**,** MOAL algorithm, Ontologies, microRNA, gene-miRNA

1. Introduction

Ontologies are specifications of a relational vocabulary. Gene-Ontologies (GO) is now a majority of bioinformatic initiatives with the purpose of unifying the representations of genes and genes products attributes across all of the species. The GO projects have now developed a three structure controlled vocabulary (ontology) that describes the genes product in the term of their associate biological and biochemical processes, cellulars component and molecular functions in species-codependent manner. There are three of these domains in ontology:

- Cellular Components (CC): All parts in a cell or the extra cellular environment;

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

• Molecular Functions (MF): All activities at an elemental level of a gene and it's products at a molecular level, example binding and catalysts.
• Biological Processes (BP): Operations and the sets of events which are molecular with a definite and regulated beginning and finish, related to the function of living units which are integrated such as cells, organs, tissues and organisms.

An introduction of throughput technologies which are high in molecular and cellular biology that produces an accumulation of many large sets of data which is experimental in nature. Such a large amount of data that is experimental has to be integrated with information which is additional and able to verify data of similar sorts. For example, genes and many types of proteins are accompanied by storing the additional information which has been used in the illumination in the role of the molecules that are being investigated. In the event of systemising such important knowledge, instruments which are formal for example vocabularies that are being controlled and such ontologies that have been used for managing the previously used terms. Different such ontology have been proposed in order to illuminate different such fields. In an instance, the Gene Ontology (GO) is one such frame works that is very largely used now. Gene Ontology also includes three important sub-ontology which are mainly called: Biological Processes (BP), Cellulars Components (CC), and Molecular Functions (MF).

Each and every ontology which stores and then organises the biological and molecular concepts, known as Gene Ontology Terms, which are used for describing such functions, localization and processes of biomolecular and cellular molecule. Every single Gene Ontology term has been uniquely identified by using an unique code, which ensures that it belongs to a single ontology. For every Gene Ontology Term there is also an important description which is textual that is also simultaneously available. As an example we can use GO : 0006915 which represents an apoptosis process.

There is a continuously increasing large amounts of heterogeneous, sparse and very largely valuable bio molecular information and data which are very essential to characterise sciences of life. In particular, bio molecular entities annotations which are controlled by bio molecular entities and semantics, that is the associations among bio molecular entities (which is mainly the genes and the gene products) along with controlling the terms that are used for describing the bio molecular entity functions and features, that are of extremely great and important value; this is used to support many scientists who are equipped in many important terminologies along with ontologies which are used to describe functional, structural, phenotypic and biological features of these said entities which are valuable (e.g. these sequences of polymorphisms and expression of different and unique tissues and it's involvement in many biological and biochemical processes and biochemical and biomolecular pathways and disorders which are mainly genetic).

These annotations which are semantic can surely and effectively secure and support interpretations of such genomic and proteomic tests and their results and these extractions of bio molecular informations, which are also used to formulated and validated biological and biochemical hypotheses and possible discovery of new and essential biomedical information.

2. Literacture Survey

• 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. This research article have used association rule mining to identify relationships

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

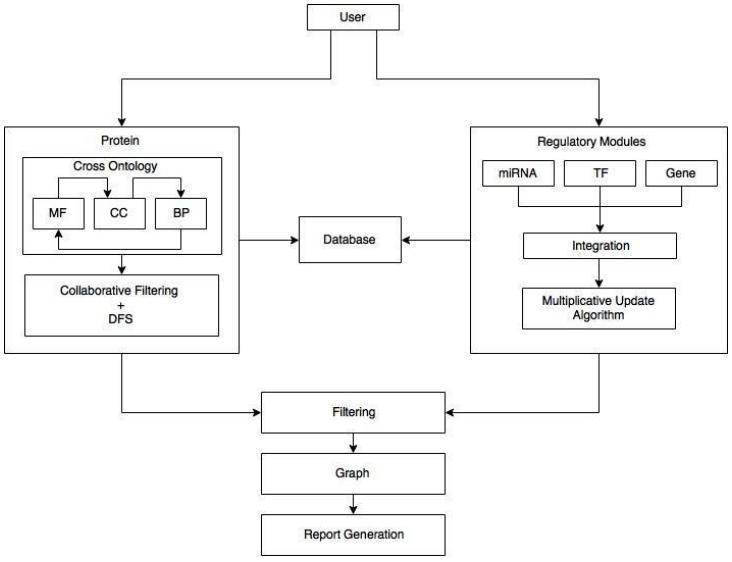
among up and down regulated genes in gene expression studies. These studies do not make use of the GO and its hierarchical structure. Previous research applying association rule mining to the GO includes studies mining single level, multi-level and cross-ontology association rules.

- 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.This paper 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.
- M. Hahsler, B. Grün, K. Hornik, SIGKDD Explorations, pp. 0-4.This article elaborates 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.

3.Existing System

Our existing system that has been proposed has made use of associations rule in order for supporting curators in GO.In this evaluation of the consistency of the annotation is done so that we can reduce all possible inconsistencies as well as annotations that are redundant.To achieve this methods such as association rules (Classical) for algorithm mining

We are using SN Co NMF which is an algorithm that requires us to set many different parameters which are described inside a pseudo code.It can be important for deciding the value in a reduced dimensions of a matrix factorisations K.We then created a cluster of miRNA analysis that needed the clustered data of miRNA from a website ("www.mirbase.org").The disadvantages of the SN Co NMF (1) is that associations rule (classical) has algorithms for mining for dealing with various sources where



productions of annotations of GO occur.Consequently data that is annotated when used can produce many rules that are candidated due to very low content in information More over very large amounts of information were lost. The merging of theses different existing algorithms work.

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

4. Proposed System

4.1 System Architecture:

4.2 Gene Ontology

Gene Ontology is the frameworks for any models of biology. The GO (2) defined concepts and it's classes are used for describing the gene function, and it's relationships between these concepts. It has been classified functioning along three aspects: molecular function molecular activities of gene products, cellular component where gene products are active, biological process pathways and many large process has made up of the activities of multiple gene products. The Gene Ontology (GO) project has a collaborative effort to address the needed for consistent descriptions of the gene products in many different databases. 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. We are proposing cross ontology to manipulate the Protein values (9) 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.

4.3 MOAL (Multi Ontologies Datas Mining At All Level) Algorithm for mining all cross ontology relationship between the ontologies. MOAL algorithm (3) is used for mining cross-ontologies associations rule, i.e. the rules that are involved in the GO terms are present in the three sub-ontology of GO.By using collaborative filtering (7), 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.

4.4 Depth first search:

Depth first search (4) is in relational to any specific domain such as searching for solutions in any artificiial intelligence. The graphs are to be traversed is often either too large to visit in its entirety or infinite. In such cases a search is done to a certain particular depth because of it's limited and important resources. For example in memory and disk-based space an individual usually does not exactly use many data structures for keeping track of the set of all previously visited vertice. When DF search has been performed on to a limited and particular depth the time is generally linear in accordance to the number of the expanded vertice. Although the number of edges is not necessarily the same as the specific capacity of an entire graphs. This is due to the fact that some vertices might be searched once or more and others vertices might not be searched at all. The space complexity of the variants of DFS is said to be proportional to it's depth limit. As a result of this,it is significantly smaller than the prescribed space needed for search to the said depth using BFS. DF searchs has also optimal foor heuristic search(6) options in order to find a similar looking branch.

DFS is an iterative approach. the DFS implementation is as follows:

```
DFS(G, u)
   u.visit = t
   for each v ∈ G.Adj[u]
      if v.visit == f
         DFS(G,v)
```

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

```
init () {
   For each u ∈ G
      u.visit = f
  For each u ∈ G
     DFS(G, u)
}
```

4.5 Multiplicative algorithm

The idea of the algorithm is to maintain a distributions over any certain sets in various fields and uses the updating multiplicating rule (5) for changing the weights in an iterative fashion.The algorithm increases in an exponential function .It works on a very simple idea of divide and conquer but its importance is undefined as it is simple used in most of the surveys to achieve the desired result.
Initialisation: Let $\eta \leq 1$. With every decisions i, associated with the wt wi (1) := 1.2

For t = 1,2,…,T:

1. Choose decision i with probability proportional to its weight i(t). I. e., use this distributions over decisions p(t) = {w1(t)/Φ(t),...,wn(t)/Φ(t)} where $\Phi(t) = \sum i\ wi(t)$.
2. Observe the costs of the decisions m(t).
3. Penalize the costly decisions by updating their weights as follows: for all decision i, set wi(t+1) = wi(t)(1−ηmi(t))

Proof:
$\Phi(t) = \sum i\ wi(t)$:
$\Phi(t+1) = \sum wi(t+1)\ i$
$= \sum wi(t)(1-\eta mi(t))\ i$
$= \Phi(t) - \eta\Phi(t)\sum mi(t)pi(t)\ i$
$= \Phi(t)(1-\eta m(t) \bullet p(t))$
$\leq \Phi(t)\exp(-\eta m(t) \bullet p(t))$.
It is known that pi(t) = wi(t)/Φ(t). Thus, by induction, after T rounds, we have □T□ □T□
$\Phi(T+1) \leq \Phi(1)\ \exp -\eta \sum m(t) \bullet p(t) = n \bullet \exp -\eta \sum m(t) \bullet p(t)$. (2.2) t=1 t=1
Next we use the following facts, which follow immediately from the convexity of the exponential function:
$(1-\eta)x \leq (1-\eta x)$ if $x \in [0,1]$, $(1+\eta)-x \leq (1-\eta x)$ if $x \in [-1,0]$.
Since $mi(t) \in [-1,1]$, we have for every decision i,
$\Phi(T+1) \geq wi(T+1) = \prod(1-\eta mi(t)) \geq (1-\eta)\sum_{\geq 0} mi(t) \bullet (1+\eta)-\sum_{<0} mi(t)$, $t \leq T$
where the subscripts "≥ 0" and "< 0" in the summations refer to the rounds t where mi(t) is ≥ 0 and < 0 respectively. Taking logarithms in equations (2.2) and (2.4) we get:
T lnn−η ∑m(t) •p(t) ≥ ∑mi(t) ln(1−η)−∑mi(t) ln(1+η).
t=1 ≥0 <0 Negating, rearranging, and scaling by 1/η: T
∑m(t)•p(t) ≤
lnn1 11
η +η∑mi(t)ln1−η+η∑mi(t)ln(1+η) ≥0 <0
lnn + 1 ∑mi(t)(η +η2)+ 1 ∑mi(t)(η −η2)
t=1 ≤ η +∑mi(t)+η∑mi(t)−η∑mi(t) t=1 ≥0 <0
lnnT T η +∑mi(t)+η∑|mi(t)|.
t=1 t=1 In the second inequality we used the facts that

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

1□2 2 ln 1−η ≤η+η and ln(1+η)≥η−η (2.5)
for η ≤ 1/2

5. Experiments

5.1 Datasets
In the experiments, we validate our prediction system using the two kinds of real world datasets.

Table 1
Gene and Gene id

DISEASE NAME	ASSOCIATED GENES	INRINSIC/EXTRINSIC	SYMPTOMS	TREATMENTS
BECKWITH-WIEDEMANN SYNDROME	CDKN1C (1028), H19 (283120), KCNQ1OT1 (10984), H19-ICR (105259599)	Intrinsic	and/or genetic alterations that dysregulate imprinted genes on chromosome 11p15.5. Molecular subgroups are associated with different	a clinically unaffected monozygotic twin of a patient, but should not be guided by genotype/phenotype correlations at this time. Screening for hypoglycemia should be undertaken in the neonatal period if there are suggestive or diagnostic prenatal findings, and even for clinically unaffected
BILE ACID SYNTHESIS DEFECT, CONGENITAL, 3	CYP7B1 (9420)	extrinsic	Diarrhea. Loss of liver functStunted or abnormal	We are in the process of getting a bile acid, called cholic acid
OSTEOPETROSIS, AUTOSOMAL RECESSIVE 8	SNX10 (29887)	Intrinsic	disease (CGD). It is also used to slow down the progression of severe, malignant osteopetrosis (SMO). Interferon gamma ...	At present, there is no effective medical treatment for osteopetrosis. Management is supportive and aims at providing multidisciplinary surveillance and symptomatic treatment.
ATPASE DEFICIENCY, NUCLEAR-ENCODED	ATPAF2 (91647)	extrinsic	cryptorchidism,hypospadias,microcephaly,low-set ears	Treatment of simple aldosterone deficiency should be first
PYRUVATE CARBOXYLASE DEFICIENCY	PC (5091)	extrinsic	abdominal pain, vomiting, tiredness and muscle weakness. Children with this type of PC deficiency usually die in infancy or early childhood, but some survive to adulthood.	The aim of treatment is to provide alternative energy sources and to correct acute metabolic acidosis. Other management and treatment options depend on the type of PC deficiency. Current symptomatic and supportive treatments are generally ineffective.
LYSINURIC PROTEIN INTOLERANCE	SLC7A7 (9056)	intrinsic	psychotic episodes,osteoporosis,cutis laxa	Treatment revolves around protein-restricted diet and supplement of lysine, ornithine, and citrulline. The complication of pulmonary alveolar proteinosis has been reported to be successfully treated by whole lung lavage.

Table 2
Gene and diseases:

GENE NAME	GENE ID
NSD1	64324
CDKN1C	1028
H19	283120
KCNQ1OT1	10984
H19-ICR	105259599
CYP7B1	9420
SNX10	29887
ATPAF2	91647
PC	5091
SLC7A7	9056
CPOX	1371
OCLN	100506658
GBA	2629
SLC39A4	55630
LYST	1130
FUCA1	2517
ABCD3	5825
PIGM	93183
SLC22A5	6584
FAH	2184

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

6. Result

By taking the probability and calculating the gene and it's gene ids from the three protein values such as Molecular Function, Biological Process and Cellular Component. Using these values along with the Co-Regulatory modules such as Transcription Factor, MicroRNA and gene on gene interaction we can get a result of all probable diseases with high efficiency. With future advancements in biological and bio-molecular fields the efficiency of genome mapping could be greatly maximised.

7. Future Work

In the future with advancements in bio-molecular and biochemical technologies and sciences we can expect to integrate the concept of Machine Learning. This can be used to enable the system to map and learn the frequently occurring patterns in the gene ontology sets to improve the efficiency and accuracy in finding the particular genetic diseases. Gene ontology sets include genes, gene interactions, gene id and the respective diseases. Machine Learning is specifically useful and effective in detecting and recognising patterns and associations. This can be useful to reduce the time and space complexity which forms an integral part of any software development process.

8. Conclusion

Vast multitude of progress has now been made in the fields of biotechnology and in it's system biology. These are now creating a very significant amounts of bio molecular and biochemical informations and datas and it's annotations which are semantic. Increases in the quality and the number , but these are dispersed and partially only connecting. Integrations and the minings of some of these evolving and distributed datas and it's informations have the highest possible potential of being discovered in hidden bio medical information and knowledge that is useful in the understanding of complex and integrated biologic-al phenomenon,pathological or normal,which ultimately is used for enhancing the prognosis,diagnosis, and treatment. Such type of integration also poses very huge and complex challenges. The work we have done has been tackled by now developing a generalised and novel approach to define easily and to maintain the updated and extended integration of many heterogeneous and the ever evolving data sources. This approach has proved to be useful for the extraction of such bio medical information and knowled-ge related to the complex biological process and the genetic diseases.

REFERENCES

[1] Melissa J Davis, Muhammad Shoiab B Sehgal MS, Mark A Ragan, "Automatic, Context-Specific Generation of Gene Ontology Slims,"BMC Bioinformatics 2010
[2] Nailing Su and Yufu Wang "Combinational regulation of transcription factors and microRNAs" BMC Systems Biology 2010, 4:150.[2]
[3] Peter M, "MOLE: A Varono1 diagram-based explorer of molecular channels, pore and tunnel," National Centre for Biomolecular Research.
[4] Jaime Luo, Di Dai, Buwen Cao, "Inferring human miRNA functional similarity based on gene ontology annotations," IEEE International Conference 2016.
[5] Hasna Njah, Salma Jamoussi, Mohamed Elati, Walid Mahdi, "A Bayesian Approach to Construct Context-Specific Gene Ontology: Application to Protein Function Prediction," IEEE Conference October 2016.
[6] Acón Man-Sai, Mora Rodreiguez, "A biocomputational platform for the automated construction of large scale mathematical models of mi-RNA-transcription factor networks for studies on gene dosage compensation" IEEE Conference 2016.

International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 1, March - May, 2018
Pages 77-84 | Cosmos Impact Factor (Germany): 5.195
Received: 23.03.2018 Published : 01.04.2018

[7] Jiawei Luo, Gen Xiang and Chu Pan, "Discover of microRNAs and Transcription Factors Co-Regulatory Modules by Integrating Multiple Types of Genomic Data," IEEE Transaction Jan 2017.
[8] Qi, H. Ge, "Modularity and dynamics of cellular networks", PLoS Comput. Biol., vol. 2, no. 12, pp. e174, 2006.
[9] X. Yang et al., "miR-449a and miR-449b are direct transcriptional targets of E2F1 and negatively regulate pRb–E2F1 activity through a feedback loop by targeting CDK6 and CDC25A", Genes Develop., vol. 23, no. 20, pp. 2388-2393, Oct. 2009
[10] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data", Nucleic Acids Res., vol. 40, no. 19, pp. 9379-9391, Oct. 2012.