GPKB- for Gene Disease Identification and Medical Diagnosis using MF, CC, BF, Micro RNA and Transcription Factors

M. Ramamoorthy

Received: 10 Jan 2018 • Revised: 15 March 2018 • Accepted: 08 April 2018

Abstract: Multiple genomic and proteomic semantic annotations scattered in many distributed and heterogeneous data sources; such heterogeneity and dispersion hamper the biologists’ ability of asking global queries and performing global evaluations. To overwhelm this problem, we developed a software planning to create and maintain a Genomic and Proteomic Knowledge Base (GPKB), which integrates several of the most relevant sources. 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. There are different methods of analysis to get bio information. One of the methods is the use of Association Rules (AR) which discovers biologically applicable associations between terms of GO. In existing work we used GO-WAR (Gene Ontology-based Weighted Association Rules) for extracting Weighted Association Rules from ontology-based annotated datasets. We here adapt 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. Also our proposed system, focus on intrinsic and extrinsic. Based on cellular component, molecular function and biological process values intrinsic and extrinsic values would be calculated. For each proteomic analysis for every gene disease, we analyze 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. In this paper, 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. Iterative Multiplicative Updating Algorithm is used in our paper 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: GO-WAR, MIRNA, Transcription Factor.

INTRODUCTION

Ontologies are specifications of a relational vocabulary. Gene ontology (GO) is a major bioinformatics capability to unify the description 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 processes vital for a living organism to live that are persistence an transformation of life forms.

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

Increasingly large amounts of valuable, but heterogeneous and sparse, biomolecular data and information are characterizing life sciences [1]. 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 (e.g. their sequence polymorphisms, expression in different tissues, or involvement in biological processes, biochemical pathways and genetic disorders).

![Figure 1: Biology and Genetic Endowment](image)

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.

A comprehensive approach to such data integration, querying and analysis can help understanding complex biological processes and their pathological alterations, by answering related complex biomedical questions. Yet, the scattering of genomic and proteomic annotation data in many complementary but also overlapping sources is an important and not yet completely solved challenge. Specifically, data source heterogeneity in data representation and format, their fast evolution in number, data content and structure, the high variety of available data types, and also the great amount of data produced over time, are the facets of a very hard data integration problem.

Taking advantage of our previous experience with the GFINDer system, we developed a software architecture to create and maintain an updated and publicly available integrative data warehouse of genomic and proteomic semantic annotations. It adopts a modular and multilevel global schema that we propose for integrated data management. This data schema supports integration of data sources, possibly overlapping, which are fast evolving in data content, structure and number, and assures provenance tracking of all the **Davis MJ, Sehgal MS, Ragan MA (2010) Automatic, context-specific generation of Gene Ontology slims. BMC Bioinformatics 11: 498.**

This research article 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.
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. Giuseppe Agapito, Mario Cannataro, Petro Hiram Guzzi, Marianna Milano, Using GO-WAR for mining crossontology weighted association rules, Elsevier, 2015 Developed GO-WAR, i.e. Gene Ontology-based Weighted Association Rules Mining, a novel data-mining approach able to extract weighted association rules starting from an annotated dataset of genes or gene products. The proposed approach is based on the following steps: (i) initially we calculate the information content for each GO term; (ii) then, we extract weighted association rules by using a modified FP-Tree like algorithm able to deal with the dimension of classical biological datasets. We use publicly available GO annotation data to demonstrate our method.

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. DRAWBACKS: The disadvantage of this method is that Classical association rules mining algorithms are not able to deal with different sources of production of GO annotations. Consequently, when used on annotated data they produce candidate rules with low Information Content. Consequently, a large amount of information is usually missed.

PROPOSED SYSTEM

In this paper, we proposed co-regulatory modules between Transcription Factor, gene and MiRNA on functional level with genomic data. The integration technique is implemented between miRNA, Transcription Factor (TF) and gene. After integration, Iterative Multiplying update algorithm is used to check the optimization function between the regulatory modules. We get the expression or some value from this algorithm then compare to protein values. The protein value get from Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) with the help of cross ontology technique. At last we generate a bayesian rose tree structure for the relation between regulatory modules and protein values of our gene. By this structure we

![System Architecture Diagram]

SYSTEM MODEL

Know our disease which was affected in our chromosome and also know how to cure. Also we can identify the symptoms applicable for our gene by our proposed system.
**Advantages**

- It is important to detect interaction effect of multiple genes during certain biological process.
- Given a time series gene expression data a hidden markov model Bayesian model was developed to calculate the observed data.
- Co regulatory calculate gene similarities based on common neighbor of any two gene.

**MODULE DESCRIPTION AND IMPLEMENTATION**

**Gene Ontology**

Gene Ontology is the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects: molecular function molecular activities of gene products, cellular component where gene products are active, biological process pathways and larger processes made up of the activities of multiple gene products. The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. In our paper 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 paper 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.

**Collaborative Filtering**

In this paper 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 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.

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.

**Depth first search**

Depth first search in relation to specific domains such as searching for solutions in artificial intelligence. The graph to be traversed is often either too large to visit in its entirety or infinite. 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 but the space complexity of this variant of DFS is only proportional to the depth limit. 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.

DFS is an iterative approach, the DFS implementation are as follows: DFS(G, u)

\[ u \text{.visited} = \text{true} \]

for each \( v \in G \text{.Adj}[u] \)
if v.visited == false
    DFS(G,v)

init () {
    For each u ∈ G
        u.visited = false
    For each u ∈ G
        DFS(G, u)
}

**Regulatory Modules**

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 applied the method to a Saccharomyces cerevisiae expression data set, showing its ability to identify functionally coherent modules and their correct regulators. We present microarray experiments supporting three novel predictions, suggesting regulatory roles for previously uncharacterized proteins.

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. 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. 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.

**I) 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.

**II) 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.

**Tree Representation**

In this section, we briefly introduce the BRT algorithm. Bayesian hierarchical clustering algorithm that can produce trees with arbitrary branching structure at each node known as rose trees. We interpret these trees as mixtures over partitions of a data set and use a computationally efficient greedy agglomerative algorithm to find the rose trees which have high marginal likelihood given the data.

**CONCLUSIONS**

In this paper relevant progresses in biotechnology and system biology are creating remarkable amount of bimolecular dated semantic annotations; they increase in number and quality, but are dispersed and only partially connected. Integration and mining of these 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. Our work has tackled them by developing a novel and generalized way to define and easily maintain updated and extend an integration of many evolving and heterogeneous data sources.
REFERENCES


