International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 6, No. 1 (2017): 894-905
Research Article
Open Access

ISSN: 2320-0634

Association Extraction from Biomedical Text using Network Analysis

Kanimozhi U* and Manjula D

Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, India.

*Corresponding author: Kanimozhi U; email: kanimozhiu.03@gmail.com

Received: 30 November 2016 Accepted: 22 December 2016 Online: 02 January 2017

ABSTRACT

Discovery of genes that are responsible for various diseases, becomes an important task. Since the genes are related with many diseases, the gene-disease association should be discovered. To obtain this gene-disease association from available biomedical literature, the relation type between the gene and disease is extracted from the biomedical literature. So, this becomes more and more important to deal with the extraction problem from the biomedical texts in an automatic way. Then the gene-disease association is visualized by network construction and association score matrix is constructed to calculate the gene-disease association score. The gene-disease relation type is identified and then the association score is calculated by integrating disease similarity network and protein-protein interaction network. The candidate genes for the particular disease and the novel genes for various diseases can also be found by calculating the association score and visualizing the dataset network.

Keywords: Association Extraction, Biomedical Literature, Protein-protein Interaction.

1. INTRODUCTION

Determining gene-disease associations will enhance the development of new techniques for prevention, diagnosis and treatment of the diseases. Given that the amount of biomedical literature regarding the identification of disease genes is increasing rapidly. One of the challenges that scientists in this domain face is that most of the relevant information remains hidden in the unstructured text of the published papers. Hence there is an urgent need for a text mining system that extracts both known and novel GDA (Gene Disease Association) and visualization. A new text mining system with network association capability for the visualization of gene disease association is proposed. Recent studies have proposed several approaches to investigating the relationships between genes and diseases. Some previous studies use protein-protein interactions to predict gene-disease relationships. Some researchers compute the similarity values between genes and diseases based on Gene Ontology (GO) or Disease Ontology (DO) terms. Other controlled vocabularies such as MeSH (Medical Subject Headings) have already been utilized for linking proteins to disease terminologies. Some other information like gene expressions, protein/genome sequences and positional information are also served as the important evidences to relating genes and diseases. For gene-disease association discovery from the biomedical literature using biomedical text mining, the first step is Named Entity Recognition (NER) that is the annotation of gene and disease terms by highlighting the gene and disease terms in the biomedical literature. Then the sentences with gene and disease terms have to be extracted by excluding other sentences in that biomedical literature. After extracting the sentences the relationship between the gene and disease has to be identified. The relationship is identified between the gene and disease based on the semantic analysis using relation type dictionary are further included in the network construction phase for identifying the association between them.

Although many approaches have been developed for prioritizing candidate disease-causing genes based on exploiting the protein-protein interaction network and phenotype similarities, most of which deal with the disease-gene association score based on the association between the diseases similar to the query disease and their involved genes independently. In this work, the modular nature of the genetic diseases and the consistency between the disease phenotypic overlap and genetic overlap are fully exploited. For this purpose, the disease similarity network and the protein-protein interaction network are incorporated systematically and comprehensively in a simple and compact manner to formulate the computation of the prioritization scores. As for a single disease gene association score function, both the similar diseases in the disease similarity network and neighboring genes in the protein-protein interaction network are considered because of the modular nature of the genetic diseases. All the association scores between the similar diseases and neighboring genes would be integrated into the iterative computation of this single disease gene association score. And also this system deals with the relation type identification task from the biomedical literature. For constructing disease similarity network, the large-scale disease similarity information is exploited. The disease similarity network and the protein-protein interaction network are coupled in a comprehensive and systematic way for the definition of the disease-gene association score function, and this is well in accord with the consistency between disease phenotypic overlap and genetic overlap. The definition of disease-gene association score makes full use of the information implicated in both disease similarities and neighboring genes comprehensively. On the other hand, not only the noise in the disease similarity information but also the self-loop in the protein-protein interaction network are considered in the computation of the disease-gene association scores. An iterative algorithm is designed for the computation of the disease-gene association score matrix for all the diseases and all the candidate genes in the protein-protein interaction network.

As most of the gene-disease association discovery part depends upon the datasets such as disease similarity network, protein-protein interaction network and the already discovered association dataset from OMIM. Thus the accuracy of the result depends upon the completeness of the datasets. While the calculation of association scores, the correlation between the two networks is calculated. When a candidate gene is prioritized for a disease, the correlation of the two subnetworks are considered separately induced by the neighbors of the gene in the protein-protein interaction network and the neighbors of the disease in the disease similarity network. That is, a single association between a gene and a disease is formulated iteratively by the correlation of the two subnetworks. This constraint can also be described as the fact that a gene is likely to be involved in a disease if the gene's neighbors are associated with the similar diseases. In our method, the association score between disease 'd' and gene 'g' is formulated iteratively as the weighted sum of all the existing association scores between the neighbors of gene 'g' and the diseases similar to disease 'd'. While this calculation, if the dataset does not contain sufficient information then the association score will result in less accurate one. The work is defining about the procedure of finding the association score between gene and disease and also predicting the novel gene for various diseases. And the network construction and visualization using three types of datasets namely disease similarity network, protein-protein interaction network and gene-disease association from OMIM database.

1.1 RELATED WORK

In PRINCE [11], a novel network-based approach for predicting causal genes and protein complexes that are involved in a disease of interest. The method, which is called PRINCE (PRIoritizatioN and Complex Elucidation), generalizes the network-based approaches by both considering the network signal in a global manner and going beyond single genes to the modules that are affected in a given disease. It receives as input a disease-disease similarity measure and a network of protein-protein interactions. It uses a propagation based algorithm, to infer a strength-of-association scoring function that is smooth over the network (i.e., adjacent nodes are assigned similar values) and exploits the prior information on causal genes for the same disease or similar ones. This scoring is then used in combination with a PPI network to infer protein complexes that are involved in the given disease.

In other network related work they used a plethora of network-based approaches to investigate the underlying molecular mechanisms of various human diseases. They perform a bipartite, topological and clustering graph analysis in order to gain a better understanding of the relationships between human genetic diseases and the relationships between the genes that are implicated in them. For this purpose, disease-disease and gene-gene networks were constructed from combined gene-disease association networks. Collecting and integrating data from three diverse resources, each one with different content covering from rare monogenic disorders to common complex diseases. Identifying important topological properties of the biological networks and uncover noticeable disease-disease and gene-gene associations. In particular, based on the topological analysis of networks, it is provided that numeric evidence on the assumption that many genes can be causative for a human disease.

A multi-level network model that integrates drugs, diseases and genes together, called a drug–disease–gene network (DrDiGeN)[6]. The network consists of three subnetworks, a drug–drug network (DrDrN), a disease–disease network (DiDiN) and a gene–gene network (GeGeN). The statistic characteristics show

that node degree in most of the subnetworks approximately follows a power-law distribution. The results also indicate that if genes in the GeGeN occupy important topological positions, then their associated drugs and diseases always hold critical roles in the DrDrN and the DiDiN respectively. In addition, most drug target genes and disease-causing genes are always different and nonessential, and the both show a lower likelihood to encode hub proteins in human protein–protein interaction (PPI) network, while a little higher tendency is observed in the GeGeN. Gene modules extracted from the GeGeN are highly enriched in Gene Ontology (GO) terms, but poor co expressed in human tissues compared with that of the PPI network. Furthermore, diseases (or drugs) associated with similar genes highly interact with each other such that tightly related drugs, diseases and genes can easily form co-modules, in which they share a similar pattern. The conserved structures are helpful for the understanding of the interaction mechanisms of drug–disease–gene as well as drug applications and disease treatments in a network-based level.

A work on 'Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD)' [2]. The Comparative Toxicogenomics Database (CTD) is a publicly available resource that promotes understanding about the etiology of environmental diseases. It provides manually curated chemical-gene/protein interactions and chemical- and gene-disease relationships from the peer-reviewed, published literature. Journal articles are prioritized for curation by chemicals of interest. They are identified by querying titles and abstracts from MEDLINE using PubMed and controlled chemical terms and synonyms from the National Library of Medicine's Medical Subject Headings (MeSH). Documents are ranked in date order (the default order from PubMed). Bio curators read abstracts and full-text articles from which they capture chemical-gene/protein interactions and disease relationships. All curated interactions and relationships are captured using controlled vocabularies and ontologies to maximize consistency among biocurators, ensure reproducible data retrieval by users, and enable integration of CTD data with other databases. The CTD chemical vocabulary derives from a modified subset of the chemicals and supplementary concepts in the "Drugs and Chemicals" category of MeSH. For genes and proteins, CTD uses official gene symbols and names from the National Center for Biotechnology Information's (NCBI) Entrez-Gene database .Where possible Entrez Gene entries representing orthologs are merged into a single, cross-species gene entity in CTD. Curators use these cross-species genes in CTD to capture chemical interactions and disease relationships. The CTD disease vocabulary uses terms from MeSH and OMIM. CTD interaction types are described using terms from a hierarchical vocabulary of 50 diverse relational terms (e.g., "binding," "phosphorylation") developed by CTD curators. Organisms in which chemical-gene interactions are curated are specified using terms from the Eumetazoa portion (vertebrates and invertebrates) of the NCBI Taxonomy database.

A hybrid named entity tagger for tagging human proteins/genes [7] that do analysis of scientific literature is the extraction of gene/protein names in biomedical texts. Which includes three processing steps: Conditional random fields (CRF) for initial learning and labelling, rule based tagging to improve the performance of initial tagging process by checking the specific patterns related to human proteins and genes, and a two stage abbreviation identification algorithm which resolves both short form and long form abbreviations. An integrated approach for human proteins and protein kinases normalization [8]. The task of recognizing and normalizing protein name mentions in biomedical literature is a challenging task and important for text mining applications such as protein–protein interactions, pathway reconstruction and many more. In Homo sapiens, a greater number of biological processes are regulated by a large human gene family called protein kinases by post translational phosphorylation. Recognition and normalization of human protein kinases (HPKs) is considered to be important for the extraction of the underlying information on its regulatory mechanism from biomedical literature. ProNormz distinguishes HPKs from other HPs besides tagging and normalization. ProNormz is the first normalization system available to distinguish HPKs from other HPs in addition to gene normalization task. ProNormz incorporates a specialized synonyms dictionary for human proteins and protein kinases, a set of 15 string matching rules and a disambiguation module to achieve the normalization.

2. MATERIALS AND METHODS

The system consists of the main modules (i) gene name identification (ii) relation type identification and the other one is (iii) gene-disease association extraction. Fig. 1 depicts the detailed architecture of the system. The initial input of the system is PubMed biomedical abstracts. Because the system is mainly deals with extracting the relation type and association between gene and disease from the biomedical abstracts. Most of the biomedical information's can be extracted for various research purposes to uncover the hidden information's about the gene and disease. Such biomedical abstracts contain various gene names, human proteins and human protein kinases. The first task of this system is identifying the human genes and human protein kinases in the biomedical document .once the gene names are identified further tasks such as relation identification becomes easier. For the gene name and protein name identification task a tool called Pronormz is used. This tool has an in build tool called NAGGNER for performing the named entity recognition task that is highlighting the gene and protein names in the biomedical abstract. The tool Pronormz consists of four major functional components for the recognition of human gene/protein mentions and mapping them to their corresponding EntrezGene unique identifier, A set

of string matching rules to compare the tagged entity with all entries in the dictionary, A disambiguation method if tagged entity maps more than one gene in the dictionary.

The next major task of this system is identification of the relation type between the identified gene and the disease mentioned in the biomedical document. The relation type recognition is further helpful to find the association between the particular gene and disease identified in the document abstract. For the relation identification, the dictionary [10] is classified as four categories namely altered expression, genetic variation, and regulatory modification and unrelated. The relation identification in the biomedical abstract is done using the Stanford parser. The abstract is read and parsed. The parsed sentences are compared with the dictionary that is stored in the hash map. If match occurs with the dictionary while matching, then the relation type is tagged in the corresponding sentence.

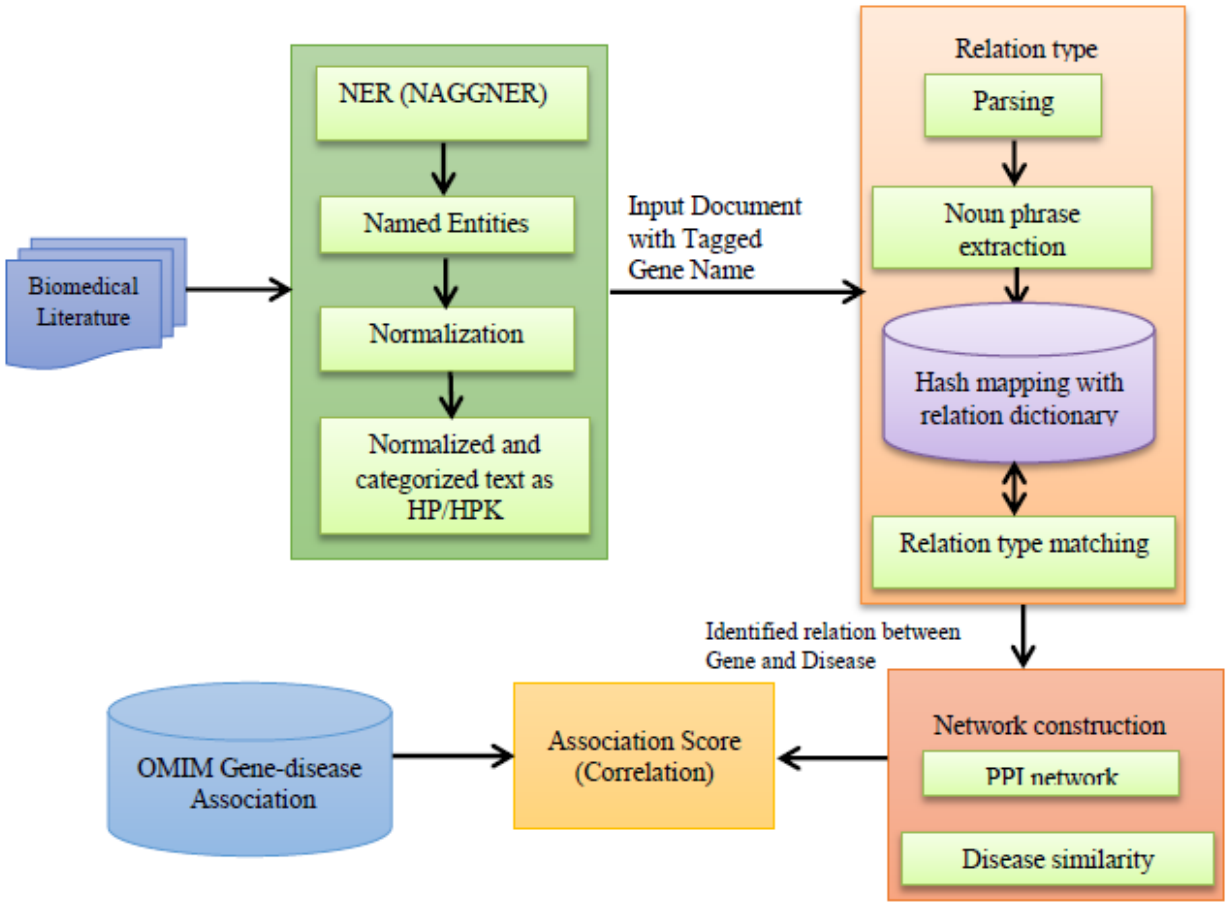

Figure 1: Work flow of the proposed model.

The final major task of the system is association score calculation. After the relation type identification, the association score is calculated between the gene and the disease that is mentioned in the abstract. The major input for the association score module is disease similarity dataset and the protein-protein interaction network. The gene disease association dataset that contains already found association between gene and disease from OMIM [4]. The association score is calculated in form of matrix using an iterative algorithm. For association score calculation the neighboring genes and the diseases related or similar diseases are also considered. These related genes and diseases are visualized using the software called Cytoscape 3.4.0 that is used for biomedical network construction tasks.

2.1 Gene Name Identification

The database of biomedical abstracts contains scientific knowledge about thousands of interacting genes and proteins. Automated text processing can aid in the comprehension and synthesis of this valuable information. The fundamental task of identifying gene and protein names is a necessary first step towards making full use of the information encoded in biomedical text. This remains a challenging task due to the irregularities and ambiguities in gene and protein nomenclature. The task of recognizing and normalizing protein name mentions in biomedical literature is a challenging task and important for text mining applications such as protein–protein interactions, pathway reconstruction and many more.so, a tool called ProNormz is used, an integrated approach for human proteins (HPs) tagging and normalization. In

Homo sapiens, a greater number of biological processes are regulated by a large human gene family called protein kinases by post translational phosphorylation. Recognition and normalization of human protein kinases (HPKs) is considered to be important for the extraction of the underlying information on its regulatory mechanism from biomedical literature. ProNormz distinguishes HPKs from other HPs besides tagging and normalization. ProNormz incorporates a specialized synonyms dictionary for human proteins and protein kinases, a set of 15 string matching rules and a disambiguation module to achieve the normalization.

For the gene name tagging module the input is a bio medical literature documents. ProNormz is different from other GN systems in two different ways. First, it is specific system to normalize human proteins and also distinguishes human protein kinases. Second, ProNormz has two in built named entity taggers, NAGGNER and BANNER for GM task. So, it can process and normalize text with pre-tagged protein/gene mentions (GM text) as well as raw text.

Input: Biomedical abstracts (PubMed)
Steps:

1) Named Entity Recognition (NER) using in build tool Naggner
2) Applying dictionary rules and entity rules on NER text
3) Disambiguation task on NER text
4) NER text is normalized using the specialized synonyms dictionary for human protein (HP) and human protein kinases (HPK)

Output: The normalized biomedical abstract with highlighted gene and proteins

2.2 Relation Type Recognition

Current biomedical research needs to leverage and exploit the large amount of information reported in scientific publications. Automated text mining approaches, in particular those aimed at finding relationships between entities, are key for identification of actionable knowledge from free text repositories. The relationship identification is mainly based on identifying relationships between biomedical entities with a special focus on genes and their associated diseases. The relation type is identified using the relation type dictionary, in which the relation type is categorized as four types namely altered expression, genetic variation, regulatory modification and unrelated which is given in Table 1. Biomedical text relation type keywords occur as noun phrases but in special cases they occur as verb phrases. The relation type recognition is performed using Stanford parser.

Table 1: Relation Type Dictionary Feature

Relation Type	Feature
Regulatory Modification	Dephosphorylation, Demethylation, Epigenetic, Methylation, Hyper methylation, Hypo methylation, Phosphorylation, Hyper phosphorylation.
Genetic Variation	Allele, Alteration, Autosomal, Biallelic, Amino Acid, Carriers Of, Exchanges, Frameshift, Genotypes, Genotype-Phenotype, Duplication etc.,
Altered Expression	Activation Of, Alternative Splice, Co-Expression, Cross-Regulation, Differentially Expressed, Down-Regulate, Co-Regulate, Expressions, Immunoexpression, Inactivate, MRNA etc.,
Unrelated	But not, No, Neither, Nor, Independently, Not likely, Unlikely that, Unrelated, Does not etc.,

Input: Biomedical document (PubMed abstracts)
Steps:

1) Loading the relation dictionary using hash map
2) Hash Map <String, String> relation_dict = load Relation( )
3) Read the input document
4) String content = new Scanner(new File(filename))
5) Load the Stanford parser model
6) Lexicalized Parser lp = Lexicalized Parser .load Model(parser Model)
7) Load Treebank Language Pack tlp = lp. Treebank Language Pack( )
8) Parsing tree to insert relation id

Output: sentences of abstracts with relation id tagged in it

2.3 Gene-Disease Association Extraction

A disease-gene association matrix in favor of computing and storing the association scores. The disease similarity network and the protein-protein interaction network are also constructed and incorporated into the formulation of the disease gene association matrix in a simple and compact manner of matrix multiplication. As a result, an iterative algorithm is designed for the computation of the disease-gene association matrix. The disease similarity network and the protein-protein interaction network are coupled in a comprehensive and systematic way for the definition of the disease-gene association score function. Third, an iterative algorithm was designed for the computation of the disease-gene association score matrix for all the diseases and all the candidate genes in the protein-protein interaction network.

Input: Identified gene-disease relationship
Steps:

1. Construct matrix $psim[i,j]$, the disease similarity network
2. Construct matrix $G[i,j]$, the protein-protein interaction network
3. Calculate adg,where $Adg{=}psim{*}adg0{*}G$

for $(k-th$ iteration)
$\text{adg}^k = (1-\alpha) \times psim \times adg^{k-1} \times G + \alpha \times adg^0$.
where α €(0,1)

for $(k+1-th$ iteration)

$$adg[i,j]^{k+1} = \sum_{l=1}^{G} \left(\sum_{k=1}^{|psim|} psim[i,k] \times adg[k,l^k] \right) \times \text{G[l, j]}$$

Output: The association score between the gene and disease

2.4 Knowledge Discovery

Knowledge discovery is the task of determining whether the discovered association between the gene and disease is new, inferred or already known association. This is determined by the visualization of the network which is constructed by integrating the disease similarity network and protein-protein interaction network. The association type can be visualized using the network construction software Cytoscape 3.4.0. The network contains direct link between the gene (4049) and the disease (607507), this network is constructed from the already available dataset. So the association already exist, thus this belongs to known association. If the associations are not exists in databases but inferred by network analysis of first neighborhood association between genes/diseases and newly retrieved from the literature, then that type of association is called inferred association. If there is no direct or inferred association from the network, then that type of associations are called novel association that are newly retrieved from the literature.

3. RESULTS AND DISCUSSION

Two disease related vocabulary resources: Online Mendelian Inheritance in Man (OMIM) and Comparative Toxicogenomics Database (CTD). An OMIM based human disease synonyms dictionary is incorporated .The relation type dictionary developed by Bundschus et al is used for relation type identification. (i)Human Disease-Disease similarities (D2D) from MimMiner ii) Protein-Protein interactions (PPI) from Human Protein Reference Database (HPRD) and iii) Gene-Disease associations (G2D) from OMIM morbid map.

Table 2: Datasets Description

Dataset	Description	Number Of Items
Relation type dictionary	Four types of relations namely altered expression, genetic variation, regulatory modification and unrelated.	197
Disease similarity	Similarity score between two disease	27092
Protein-protein interaction	Two genes that interacts with each other	39142
Gene-disease association	Already found association between gene and disease	4456

Table 3: Disease Similarity Dataset Feature

Disease Similarity	Similarity Score Range
Most similarity	0.9 - 1.0
Moderate similarity	0.6 - 0.8
Least similarity	0 - 0.5

3.1 Disease Similarity Network

The disease similarity network is introduced, where the node in the network represents a disease, the edge connecting two nodes indicates that the two diseases are similar, and the weight of the edge indicates to what extent the two diseases are similar. A disease similarity matrix Psim to model this network, in which Psim[i,j] is the similarity score between disease i and disease j. Table 3 defines about the similarity score ranges.

Fig. 2. Represents the matrix for disease similarity dataset, in which both column and row represents same set of diseases and each cell value represents the similarity score of the two set of diseases in the respective row and column. In fig. 7. Column A and row 1 both represents the same disease.

3.2 Protein-Protein Interaction Network

The protein-protein interaction network is modelled as matrix G fig. 3. In which the value of $G[i,j]$ indicates whether the interaction between proteins i and j exists. The value ‘‘1’’ denotes that the interaction exists, and ‘‘0’’ denotes that the interaction does not exist. In this method, with regard to the association between disease ‘d’ and gene ‘g’, the associations between the diseases similar to ‘d’ and the neighbors of ‘g’, the associations between the diseases similar to ‘d’ and the neighbors of ‘g’, and the associations between ‘d’ and the neighbors of ‘g’ all need to be considered. So, the protein-protein interaction network is extended by adding the self-interactions of all the proteins into the interaction network. In fig. 8 each row defines the two types of genes that interact with each other in the protein-protein interaction network. Column A and C defines the gene name or symbol, column B and D defines the genes id.

	A	B	C	D	E	F	G	H	I
1	1	0.535411	0.999943	0.999986	0.503653	0.767047	0.999943	1	0.500497
2	0.535411	1	0.535668	0.535547	0.500146	0.546096	0.535668	0.535411	0
3	0.999943	0.535668	1	0.999986	0.50359	0.767177	1	0.999943	0.50092
4	0.999986	0.535547	0.999986	1	0.503628	0.767123	0.999986	0.999986	0.500716
5	0.503653	0.500146	0.50359	0.503628	1	0	0.50359	0.503653	0
6	0.767047	0.546096	0.767177	0.767123	0	1	0.767177	0.767047	0
7	0.999943	0.535668	1	0.999986	0.50359	0.767177	1	0.999943	0.50092
8	1	0.535411	0.999943	0.999986	0.503653	0.767047	0.999943	1	0.500497
9	0.500497	0	0.50092	0.500716	0	0	0.50092	0.500497	1
10									

Figure 2: Matrix for Disease Similarity Dataset

	A	B	C	D	E	F	G	H	I
1	1	0	0	0	1	0	0	0	0
2	0	1	1	0	1	0	0	0	0
3	0	1	1	0	1	1	0	0	0
4	0	0	0	1	1	0	1	0	0
5	1	1	1	1	1	1	1	1	1
6	0	0	1	0	1	1	0	0	0
7	0	0	0	1	1	0	1	0	0
8	0	0	0	0	1	0	0	1	0
9	0	0	0	0	1	0	0	0	1

Figure 3: Protein-protein Interaction Network as a Matrix Model

3.3 Disease-Gene Association Network

The gene disease association dataset consists of already discovered gene-disease association data from OMIM. The disease-gene association network is constructed such as the one where the node in the network can be either a disease or a gene and the weighted edge connecting a disease and a gene indicates to what extent the gene is involved in the disease. This network can also be regarded as a bipartite graph. In this method the disease-gene association network is expressed by a disease-gene association matrix Adg, in which the element Adg[i,j] stores the association score of gene 'j' and disease 'i' indicating the association strength between the gene and the disease. Fig. 10 represents the gene-disease association dataset in which each row defines the one set of already associated gene and disease. Column A defines the gene name or symbol, column B defines the gene id and column C defines the disease OMIM ID that is associated with the respective gene.

3.4 Gene Name Tagging

The gene name tagging module involves named entity recognition of gene/protein in the text which is accomplished by tools NAGGNER for gene mention (GM) task and Pronormz for gene normalization (GN) task. For Pronormz, PubMed biomedical abstract is given as input and the correct gene name tagging is compared with other tools. Precision, recall and F-score are used as evaluation metrics. TP denotes numbers of true positives, FP denotes the number of false positives and FN denotes the numbers of false negatives. The F-score is the harmonic mean of recall and precision.

Table 4: Gene Normalization Task Evaluation of PRONORMZ

Gene Normalization Task	Precision%	Recall%	F-Score%
PRONORMZ	86.66	80.25	83.33
BANNER	65.14	74.52	69.52
NAGGNER	66.19	71.33	68.66

The evaluation of pronormz and its in build tools namely BANNER and NAGGNER for gene normalization task are shown in Table 4. The evaluation of the pronormz and its in build tools are done using the test set of BioCreAtIvE-II Gene Normalization. Pronormz shows comparatively higher precision and recall value than other tools namely BANNER and NAGGNER. Table 5 shows the comparison of Pronormz tool with other tagging tool, in which pronormz shows comparatively better precision and recall value.

Table 5: Comparison of PRONORMZ with other tools

Tool Name	Precision	Recall	F-Score
Hybrid Named Entity Tagger	80.47	71.60	75.77
Abner	86.93	51.49	64.68
Ling Pipe	76.61	83.64	79.97
Memm(Mallet)	84.12	81.75	82.91
Pronormz	86.66	80.25	83.33

Table 6: Relation Type Identification Evaluation

Evaluation Metrics	Relation Type Identification
Precision%	94.32
Recall%	66.83
F-score%	78.23

3.5 Relation Type Identification

The relation type is tagged in the biomedical abstract using the relation type dictionary using the Stanford parser packages. To identify the relation keyword between the gene and disease. We have used the relation type dictionary developed by Bundschus [10]. The relation dictionary synonyms are grouped into four relation types namely altered expression, genetic variation, and regulatory modification and unrelated. Table 6 shows the evaluation for relation type identification module.

3.6 Association Score

The association score is calculated as a matrix by giving the disease similarity matrix and protein-protein interaction matrix as input. The association score is calculated using the iterative algorithm which defines the correlation between the disease similarity network and protein-protein interaction network. In figure. 4, the disease similarity network for psoriasis disease is constructed in the form of matrix. In disease similarity network each cell represents the similarity score between two diseases. The disease that are similar to the disease psoriasis are discovered from the network visualization in Cytoscape 3.4.0 which is depicted in fig. 5 and the disease similarity matrix is constructed. The matrix row and column both contains the disease names that are similar to the disease psoriasis. The matrix cells contain values of only "0" and "1"."0" represents that there is no interaction between two genes and "1" represents that interaction exist between two gene. The row and column both contains genes that are responsible for disease psoriasis and the genes that interact with the gene LTA. The association score is calculated using an iterative algorithm by including the disease similarity network for psoriasis and the protein-protein interaction network for the gene LTA. Each matrix cell represents the association score between the gene and disease in which the row contains diseases and column contains genes. The association score value defines the association strength between the corresponding gene and disease.

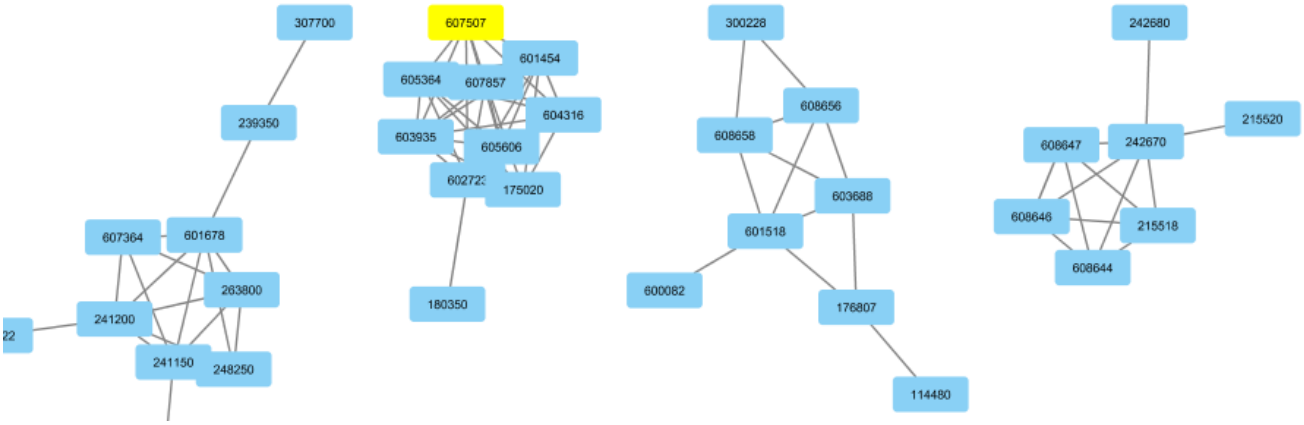

Figure 4: Disease similarity network visualization

3.7 Network Construction

The network construction is visualized using Cystoscape 3.4.0 software which is a software used for biomedical network construction tasks. The network is constructed by giving the disease similarity network and protein-protein interaction network as input.it will merge the two networks and give the visualization of combined network for the given dataset. From that we can select the subset of network. Fig.5 depicts the Cytoscape network visualization of the disease similarity network which is constructed from the disease similarity dataset. The information provided in the dataset of disease similarity network is constructed as the network by converting the various diseases as nodes and the similarity score information as edges. From this network the subset of the network can be selected for any particular disease.

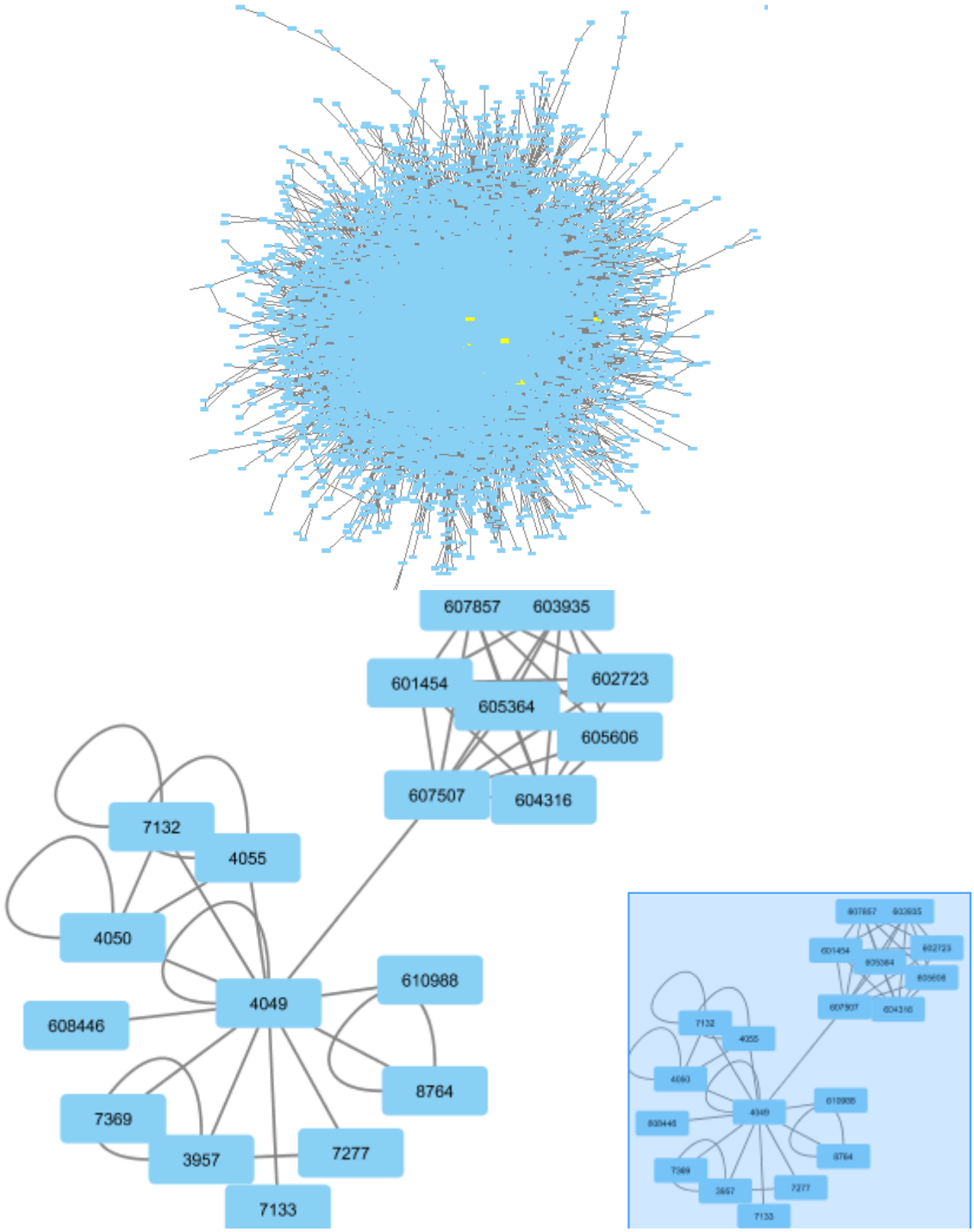

Figure 5: Protein-protein interaction network visualization

Fig. 6 defines the network construction task of the protein-protein interaction dataset. The network nodes represents the various genes in the dataset and the edges represents the interaction between the two genes. The interaction of a particular gene with others can be viewed separately from the whole network by selecting subset of the network.

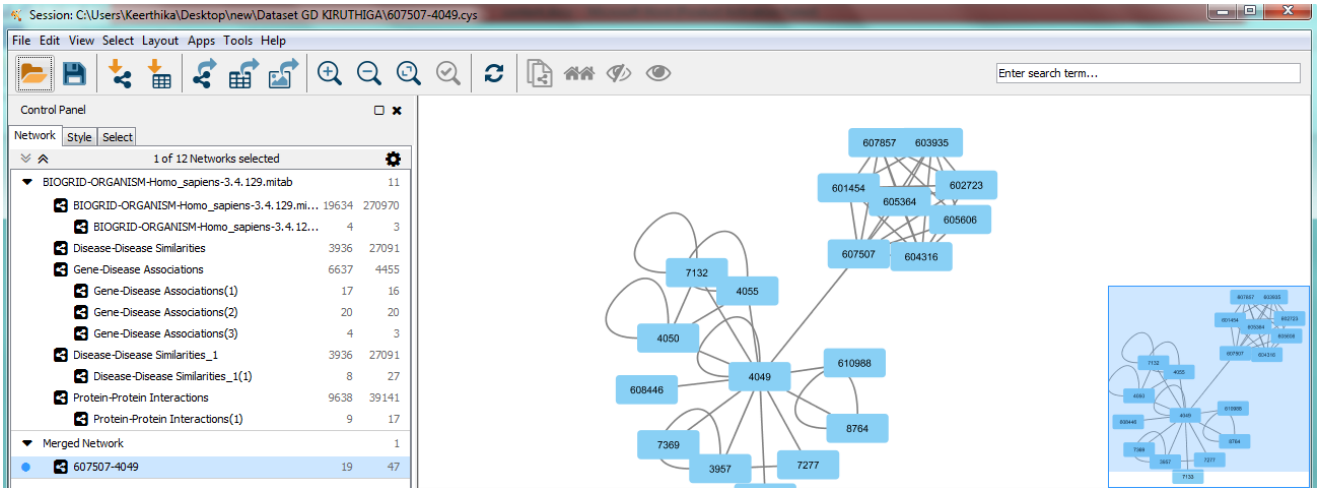

Figure 6: Network Construction for protein-protein interaction

3.8 Precision

Precision is the fraction of retrieved instances that are relevant. In this system the precision value is calculated for namely three tasks gene identification, relation type identification and association extraction which is plotted as bar chart in fig. 7. The precision value for gene identification task is calculated as the number of genes correctly highlighted in the given input abstract by the tool Pronormz. For the relation type identification task the precision value defines the correctly tagged relation types in the paragraph of abstract.

$$Precision = TP/(TP+FP) \quad (1)$$

True Positive (TP) measures the proportion of positives that are correctly identified as such. A False Positive (FP) is an error in some evaluation process in which a condition tested for is mistakenly found to have been detected.

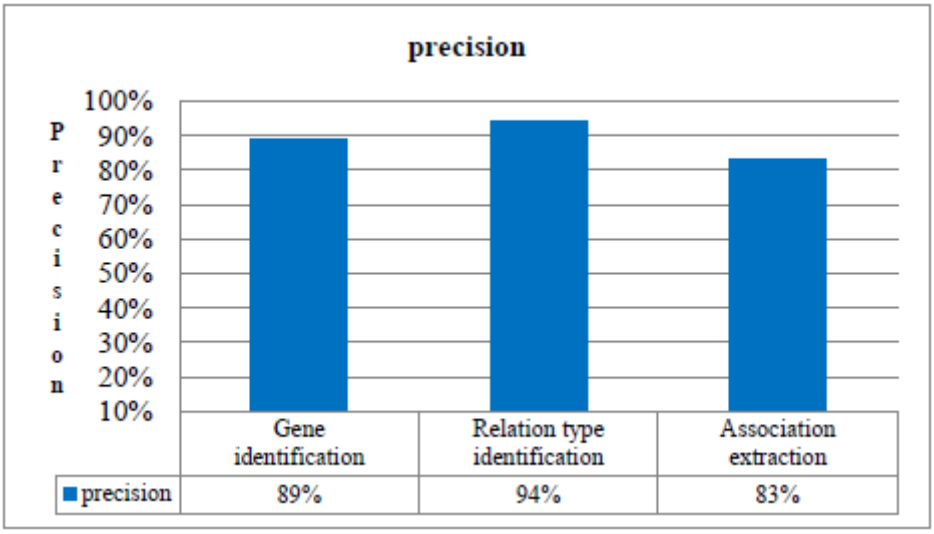

Figure 7: Precision

3.9 Recall

Recall is the fraction of relevant instances that are retrieved. In fig. 8 the recall value is plotted for gene tagging, relation type identification and association extraction module. Both precision and recall value depends upon the datasets used for this system namely relation dictionary, disease similarity network and protein interaction network. It is the ratio of the number of relevant records retrieved to the total number of relevant records in the database.

$$Recall = TP/(TP+FN) \quad (2)$$

True Positive (TP) measures the proportion of positives that are correctly identified as such. False Negative (FN) is result that appears negative when it should not.

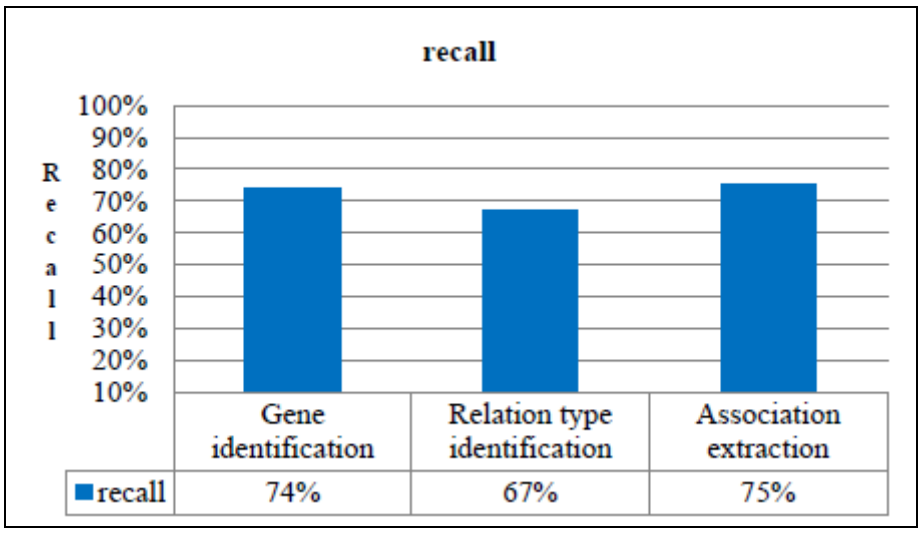

Figure 8: Recall

3.10 F-score

F-score is a measure of a test's accuracy, interpreted as a weighted average of the precision and recall

$$F = 2\times(precision.recall)/(precision+recall) \quad (3)$$

Fig. 9 defines the F-score value, which is the calculated mean value of both precision and recall. F-score is a measure of a test's accuracy. It considers both the precision P and the recall R of the test to compute the score. P is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F-score can be interpreted as a weighted average of the precision and recall, where an F- score reaches its best value at 1 and worst at 0.

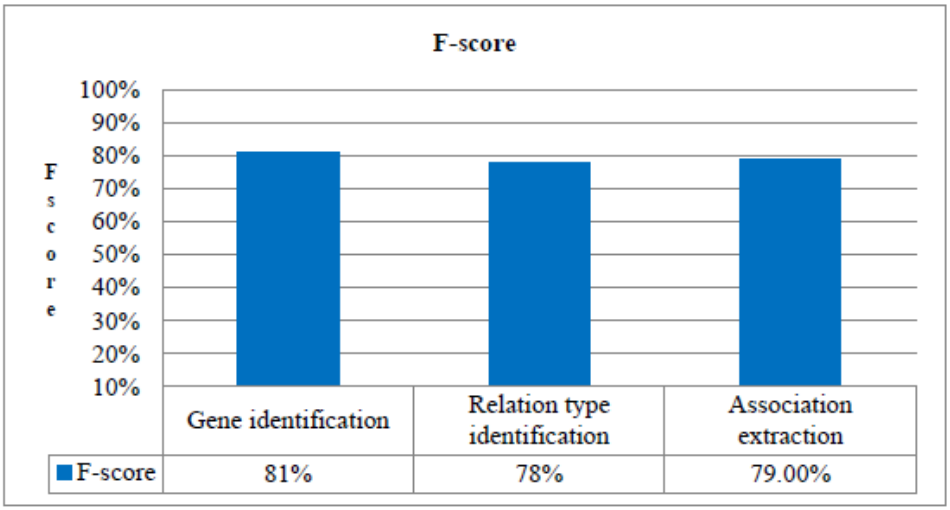

Figure 9: F – Score

4. CONCLUSION

The proposed system deals with the gene-disease association extraction along with gene and disease name tagging in the biomedical literature and also the relationship identification between the disease and gene also discovered from the biomedical literature. The gene name tagging and normalization is done using a tool called Pronormz which has an in build tool Naggner for gene name tagging. For association discovery, the disease similarity network and the protein-protein interaction network are coupled together. The association score is calculated using a neighborhood iterative algorithm by constructing three matrices namely disease similarity matrix, protein-protein interaction matrix and the gene-disease association matrix. The definition of disease-gene association score makes full use of the information implicated in both disease similarities and neighboring genes comprehensively. The self-loop in the protein-protein interaction network are considered in the computation of the disease-gene association scores. Advantages of this method are, the prioritization score for candidate genes can give some suggestions for further investigation. Second, the prioritization score can be exploited to identify disease-causing protein subnetworks, which are valuable for the study of the

multifactorial diseases. In future, the system is going to constructed in such a way that predicting more number of novel genes and also the identification of gene and disease along with the drugs and identifying gene to drug relation and the link prediction concept can also be included in the future work.

5. REFERENCES

1. Allen PD., Wiegers TC., Johnson RJ., Lay JM., Lennon-Hopkins K., Saraceni-Richards C., Sciaky D., Murphy CG and Mattingly CJ. (2013), 'Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database', PLoS One., pp 8, e58201.
2. Hamosh A., Scott AF. Amberger JS. Bocchini CA and McKusick VA. (2005), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', Nucleic acids research, pp 33, D514-517.
3. Kalpana Raja, Suresh Subramani and Jeyakumar Natarajan (2014) 'A hybrid named entity recognition for tagging human proteins/genes', International journal of data mining and bioinformatics 10.3, pp 315-328.
4. Markus Bundschus ,Mathaeus Dejori, Martin Stetter,Volker Tresp and Hans Peter Kriegel (2008), 'Extraction of semantic biomedical relations from text usingconditional random fields', BMC Bioinformatics.,pp 9, 207.
5. Panagiota I. Kontou , Athanasia Pavlopoulou , Niki L.Dimou , Georgios A. Pavlopoulos and PantelisG.Bagos (2016), 'Network analysis of genes and their association with diseases', Gene 590, pp 68–78.
6. Peng Gang Sun (2015), 'The human Drug–Disease-Gene Network', Information Sciences 306, pp 70–80
7. Suresh Subramani, Raja Kalpana and Jeyakumar Natarajan (2014), 'ProNormz – An integrated approach for human proteins and protein kinases normalization', Journal of Biomedical Informatics 47, pp 131–138.
8. Suresh Subramani and Jeyakumar Natarajan (2015) 'An integrated text mining system based on network analysis for knowledge discovery of human gene-disease associations (GenDisFinder)', Proceedings of the fifth BioCreative challenge workshop 2015.
9. Vanunu O., Magger O., Ruppin E., Sholomi T., and Sharan R. (2010) ,'Associating genes and Protein Complexes with Disease via Network Propagation', PloS Comput Biol 6(1): e1000641.
10. Xingli Guo,Lin Gao ,Chunshui Wei,Xiaofei Yang,Yi Zhao,Anguo Dong (2011) 'A computational method based on the integration of heterogeneous networks for predicting disease-gene associations'. PLoS One., 6, e24171.

*****