An integrated text mining system based on network analysis for knowledge discovery of human gene-disease associations (GenDisFinder)

Suresh Subramani[#1], Jeyakumar Natarajan[*2]

Department of Bioinformatics, Bharathiar University
Coimbatore, Tamil Nadu, India

[#] Current address: Division of Bioinformatics and Biostatistics,
U. S. Food & Drug Administration,
National Center for Toxicological Research,
Jefferson, AR, USA.

[1]sureshsubramani@hotmail.com;
[*2]n.jeyakumar@yahoo.co.in

Abstract. We introduce an automated text mining tool named 'GenDisFinder' that aids in the extraction of human gene-disease associations from biomedical literature and further categorize them as three classes known, inferred or novel using network analysis. The main modules of GenDisFinder are named entity tagging of gene/protein and disease names, gene-disease relation extraction, gene-disease network construction and analysis to predict various association types. It also provides an interface to view the interaction network. URL: http://biominingbu.org/GenDisFinder

Keywords. Text mining; Gene-disease associations; Relation extraction; Network analysis

1 Introduction

In this post genomic era, a vast amount of biomedical information from several experimental findings, clinical case report and therapeutics has been stored in text databases such as MEDLINE, PubMed Central, BioMed Central, etc. Proper use of existing domain knowledge from the literature is a prerequisite for any novel research [1]. One of the major tasks in biomedical text mining is the extraction of underlying relation between several genes and disease phenotypes. Gene-disease association (GDA) data from various high-throughput experi-

ments are hidden in the literature and lack a structured form needed for easy information extraction and visualization [2]. Hence there is an urgent need for a text mining system that extracts both known and novel GDA and visualization. In this paper, we introduce a new text mining system with network association capability, named GenDisFinder for the visualization of known/ inferred /novel gene disease association. This tool aims to discover novel associations between genes and diseases based on direct/neighborhood associations in the network.

2 Methods

GenDisFinder automatically extracts and visualizes gene-disease associations and association networks from biomedical literature and it utilizes the extracted associations to discover a novel gene-disease relation which is not reported in databases. GenDisFinder's web interface is shown in Fig. 1.

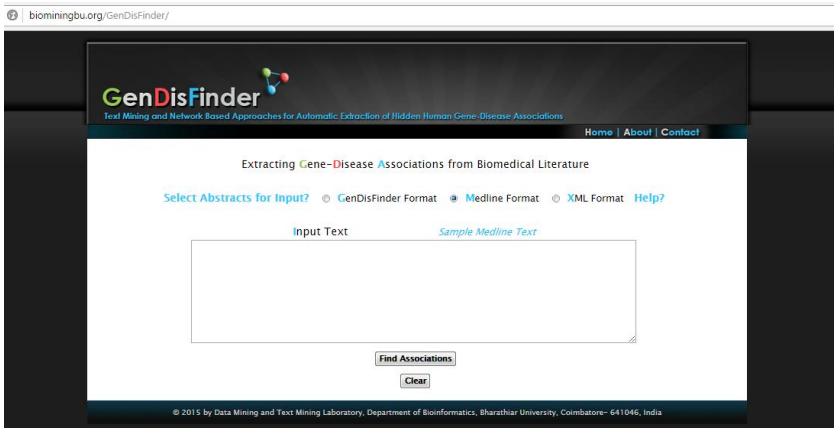

Fig. 1. GenDisFinder web interface.

2.1 System overview and architecture

The main five modules of GenDisFinder detailed below and the overall system architecture of GenDisFinder are shown in Fig. 2. The major modules were implemented in Perl and Java. The database of gene-disease is organized using MySQL. The web interface is implemented using Perl/CGI and JavaScript. Network visualization is accomplished by Cytoscape web [3].

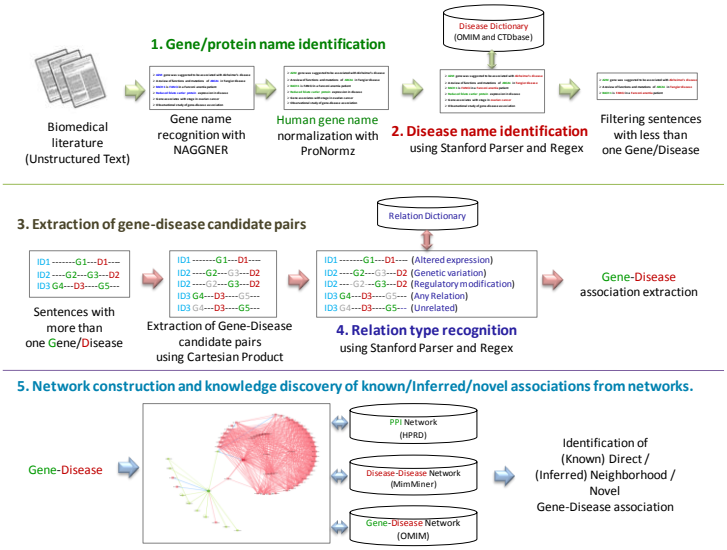

Fig. 2. Architecture of GenDisFinder.

First module involves named entity recognition of gene/protein in the text which is accomplished by in-house developed tools NAGGNER [4] for gene mention (GM) task and ProNormz [5] for gene normalization (GN) task.

Second module is disease name recognition and normalization. This task was accomplished by two disease related vocabulary resources: Online Mendelian Inheritance in Man (OMIM) [6] and Comparative Toxicogenomics Database (CTD) [7]. We incorporated an OMIM based human disease synonyms dictionary (Suppl. 1). It was performed using two Perl modules, Stanford parser (NLP::StanfordParser) and Regex (Regex::PreSuf). The identified disease names were normalized to unique OMIM identifiers.

Third module involves the filtration of sentences that does not contain at least one gene and one disease. However, some sentences may contain more than one gene or disease or both. In order to get the gene disease candidate pairs, we employed two dimensional coordinate system of Cartesian product.

Fourth module is necessary to identify the relation keyword between the gene and disease. We have used the relation type dictionary developed by Bundschus et al., [8]. The relation dictionary synonyms are grouped into four relation types namely altered expression, genetic variation, regulatory modification and unrelated (Suppl. 2). Generally, in

biomedical text relation type keywords occur as noun phrases but in special cases they occur as verb phrases [9]. The relation type recognition was also performed similar to disease name recognition by using Stanford parser and Regex.

Final module is the construction of three heterogeneous networks namely (Suppl. 3), i) human Disease-Disease similarities (D2D) from MimMiner [10], ii) Protein-Protein interactions (PPI) from Human Protein Reference Database (HPRD) [11], and iii) Gene-Disease associations (G2D) from OMIM morbid map [6]. Then we performed the integrated network analysis of the above three networks using state-of-the-art network association method described by [12] and classified extracted gene-disease associations as:

(i) **Known**: Already a direct association exists between the gene-disease pairs based on the information from databases.
(ii) **Inferred:** 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.
(iii) **Novel:** There is no direct or inferred association from the network. These associations are newly retrieved from the literature.

Our network construction and analysis methods for classifying newly extracted gene/diseases associations from literature as 'Inferred' or 'Novel' is the iterative computation is performed until the occurrence of first neighborhood association in order to efficiently determine the most informative associations and further iterations after the second stage is terminated. For the neighborhood association, a threshold of 0.5 is assigned for each neighborhood gene/disease that remains as intermediate connections for the formation of a link between indicative genes/disease in the network. These associations were referred as 'Inferred' associations. If there is no indicative association found in the first neighbor and the associations are retrieved only from the literature, then those associations are referred to as 'Novel' associations.

3 Results and Discussion

3. 1 Datasets and evaluation

To our our knowledge there is no gold standard OMIM based phenotype disease-gene association corpus available for evaluation task. Hence, we created our own gene-disease association corpus named as Human Gene-Disease Association (HGDA) corpus. For HGDA corpus

construction, first we took the pre-annotated GeneRIF corpus [13]. We randomly selected 500 sentences which were manually annotated by three domain experts in our lab. The annotations were jointly carried out all the three curators to avoid any conflicts. The final HGDA corpus contains 157 unique genes, 96 unique diseases and 206 relations between them from 182 sentences (Suppl. 4).Precision, recall and F-score were used as measure of performance assessment metrics. The evaluation results of all the four text mining modules of our system were shown in Table 2. The result shown indicates that they were equivalent and comparable to the state-of-the art systems in each task.

Table 2. Evaluation of GenDisFinder on HGDA corpus

HGDA Corpus - (206 Relation sentence)			
Dataset	**Precision %**	**Recall %**	**F-score %**
Gene identification	89.30	73.57	80.68
Disease identification	96.85	75.12	84.61
Relation type identification	94.32	66.83	78.23
Association extraction	**82.84**	**74.07**	**78.20**

3. 2 Evaluation by Biocurators of BioCreative V IAT Track

Prior to the BioCreative 5 challenge workshop 2015, curators and user advisory group of interactive curation track (IAT) assigned five different tasks for 1. Evaluation process such as find documentation, 2. Finding genes associated to a disease, 3. Goal: review the association networks, 4. Editing information and 5. Exporting the annotations. In task 2, 3 and 4, a set of five recently published abstracts (PMID: 26091350, 26087562, 26085869, 26082485, 25909225) on prostate cancer from PubMed was chosen for the evaluation. These abstracts were submitted to GenDisFinder and the evaluation was performed by the curators.

In this survey genes associated to prostate cancer and their association type were determined along with other queries. Reports are generated from certain criteria such as compatibility with various browsers, classification of extracted relations as known, novel, and unknown and its association network and finally rating the system under each category. Eight different curators evaluated the system and their evaluation survey reports are given in supplementary file (Suppl. 5). From these

results, we inferred that some users were not clear with the classification terms of extracted gene/disease associations. Hence, the associations are renamed as known, inferred, and novel to provide more clarity as described in the materials and methods section. Overall experience of the 8 curators with the system is rated as 1 very positive, 2 positive, 2 neutral, 2 negative and 1 unanswered. As a result, we find that most users give positive feedback with possible recommendation of the system to their colleagues.

In addition, few different data curator groups evaluated our system with their own corpus of different gene/disease categories. However, the evaluation task is still not completed and we are awaiting for the results. Once the results will be available, they will be uploaded in the GendisFinder web site http://www.biominingbu.org/GenDisFinder/

3. 3 Outputs

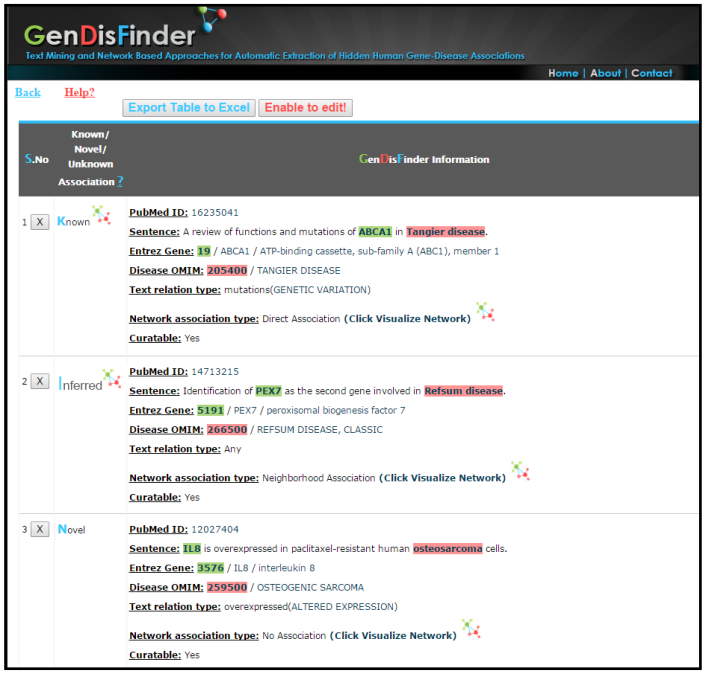

Fig. 4. Screenshot of GenDisFinder output: Gene-Disease association extraction from the biomedical text.

GenDisFinder's first output window was the mined human GDA information as shown in Fig. 4. In the output window, the gene and dis-

ease names were being normalized to official ID in order to offer links to unique references. In addition to the extracted information, the system will display highlighted association sentence with PubMed ID for cross references, Relation type of association (altered expression, genetic variation, regulatory modification and unrelated) and the predicted the association type [Known/Inferred/Novel]. On selecting particular association information will lead to the next level output window of network view of extracted associations as shown in Fig 5.

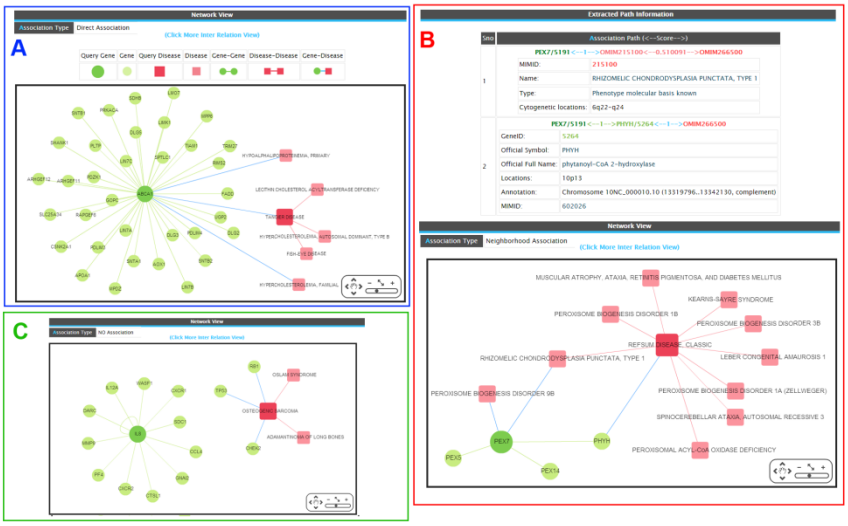

Fig. 5. A) Network visualization of direct association (Known). B) Extraction of neighborhood association and network visualization (Inferred). C) Extraction of unknown association and network visualization (Novel).

4 Conclusion

In this paper we present 'GenDisFinder' a text mining system for identifying known, inferred and novel associations between genes and diseases mined through text mining and network analysis. The novelty of this tool is threefold; it is a first system that integrates both text mining and network analysis modules. Second, it discovers highly informative associations at first iteration level (first neighborhood association) based on protein-protein interaction, gene-disease phenotype association and phenotype similarity. Third, it finds the inferred associations reported only in literature. In future, we intend to discover more relations between biological entities like genes-to-drugs relations for the development of novel treatments and cure for many diseases.

5 Acknowledgment

This work has been carried out at Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, India. The research has received funding from the Department of Biotechnology (DBT), Government of India, Grant No. BT/PR15378/BID/07/361/2011.

SUPPLEMENTARY MATERIALS

Available at http://www.biominingbu.org/GenDisFinder/suppl/
Supplementary Material S1: Disease synonyms dictionary
Supplementary Material S2: Relation type dictionary
Supplementary Material S3: D2D, PPI and G2D Network pairs
Supplementary Material S4: HGDA corpus
Supplementary Material S5: Biocreative survey result

REFERENCES

1. Shaidah, J. et al. (2012) Techniques, Applications and Challenging Issue in Text Mining. IJCSI International Journal of Computer Science., 9, 431-436
2. Li, C. et al. (2014) Biological network extraction from scientific literature: state of the art and challenges. Brief Bioinform. 15(5), 856-77
3. Lopes, C.T. et al. (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics., 26, 2347-2348.
4. Raja, K. et al. (2014) A hybrid named entity recognition for tagging human proteins/genes. Int J Data Min Bioinform., 10, 315-32.
5. Subramani, S. et al. (2014) ProNormz - An integrated approach for human proteins and protein kinases normalization. J Biomed Inform., 47, 131-8.
6. Hamosh, A. et al. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res., 33, D514-517.
7. Davis, A.P. et al. (2013) Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database. PLoS One., 8, e58201.
8. Bundschus, M. et al. (2008) Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics., 9, 207.
9. Grobe, S.J. (1990) Nursing intervention lexicon and taxonomy study: language and classification methods. ANS Adv Nurs Sci. ;13, 22-33.
10. Van Driel, M.A. et al. (2006) A text-mining analysis of the human phenome. Eur J Hum Genet., 14, 535-542.
11. Keshava Prasad, T.S. et al. (2009) Human Protein Reference Database--2009 up-date. Nucleic Acids Res., 37, D767-772.
12. Guo, X. et al. (2011) A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. PLoS One., 6, e24171.
13. Mitchell, J.A. et al. (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc., 460-464.