Bioinformatics, 32(23), 2016, 3619–3626
doi: 10.1093/bioinformatics/btw503
Advance Access Publication Date: 9 August 2016
Original Paper

Data and text mining

DTMiner: identification of potential disease targets through biomedical literature mining

Dong Xu[1,†], Meizhuo Zhang[2,†] Yanping Xie[1], Fan Wang[1], Ming Chen[2], Kenny Q. Zhu[1,*] and Jia Wei[2,*]

[1]Department of CSE, Shanghai Jiao Tong University, Shanghai 200240, China and [2]R&D Information, Innovation Center China, AstraZeneca, Pudong, Shanghai 201203, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: Jonathan Wren

Received on April 21, 2016; revised on June 7, 2016; accepted on July 19, 2016

Abstract

Motivation: Biomedical researchers often search through massive catalogues of literature to look for potential relationships between genes and diseases. Given the rapid growth of biomedical literature, automatic relation extraction, a crucial technology in biomedical literature mining, has shown great potential to support research of gene-related diseases. Existing work in this field has produced datasets that are limited both in scale and accuracy.

Results: In this study, we propose a reliable and efficient framework that takes large biomedical literature repositories as inputs, identifies credible relationships between diseases and genes, and presents possible genes related to a given disease and possible diseases related to a given gene. The framework incorporates name entity recognition (NER), which identifies occurrences of genes and diseases in texts, association detection whereby we extract and evaluate features from gene–disease pairs, and ranking algorithms that estimate how closely the pairs are related. The F1-score of the NER phase is 0.87, which is higher than existing studies. The association detection phase takes drastically less time than previous work while maintaining a comparable F1-score of 0.86. The end-to-end result achieves a 0.259 F1-score for the top 50 genes associated with a disease, which performs better than previous work. In addition, we released a web service for public use of the dataset.

Availability and Implementation: The implementation of the proposed algorithms is publicly available at http://gdr-web.rwebox.com/public_html/index.php?page=download.php. The web service is available at http://gdr-web.rwebox.com/public_html/index.php.

Contact: jenny.wei@astrazeneca.com or kzhu@cs.sjtu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent scientific discoveries have revealed the molecular, cellular and genetic components of diseases. Researchers have gained many new insights into cellular signaling pathways, genetic alterations and their consequences. Combined with diagnostic breakthroughs, there have been increasing efforts and successes in identifying patient segments defined by biomarkers that are more susceptible to certain diseases or will maximally benefit from certain treatments. Traditionally, oncology was the disease area where the majority of such research was focused. However, in the last decades, there have been considerable advancements in other disease areas, such as respiratory diseases, infectious diseases and inflammatory diseases. All of these new findings are contained in the vast amount of biomedical literature. The effective extraction of gene–disease associations from biomedical literature will potentially enable the discovery and development of new therapeutic targets and patient segment biomarkers.

However, it has remained a daunting task because of the massive volume and the textual complexity of biomedical literature.

Given the strong demand for gene–disease relation extraction from the biomedical literature, over the years ample studies have attempted to tackle it or part of it. The task of gene–disease relation extraction can be broken down into three phases. The first is gene and disease recognition in free text, which can be categorized as a Named Entity Recognition (NER) problem. Machine learning-based methods are one of the approaches used to identify biomedical entities in free text. Settles (2004) proposed a conditional random field model ABNER to recognize biomedical named entities based on orthographic and semantic features. Ju *et al.* (2011) used SVM to recognize biomedical named entities such as proteins and genes in biomedical literature. There are rule-based and pattern-based methods as well. Hanisch *et al.* (2005) developed a rule-based method to recognize genes and proteins in biomedical text. Collins and Singer (1999) discussed the pattern-based method for name entity recognition in their work. In this phase, we used search engines to enrich our disease/gene libraries. Previous efforts have been made to accomplish a similar task, including using the World Wide Web to enrich ontologies (Agirre *et al.*, 2000), using search engine query data to detect epidemics (Ginsberg *et al.*, 2009) and using Google data as a complement to co-occurrence frequency in the literature to identify disease-related genes (Kim *et al.*, 2015).

The second phase is to identify the relationships between genes and diseases. A relationship between a gene–disease pair can include but is not limited to therapeutic targets, prognostic factors etc. For example, in the sentence 'Our prospective findings suggest that individuals carrying the HFE C282Y mutation may be at increased risk of CHD,' a relationship between the gene 'HFE' and the disease 'CHD' is made clear. Kernel methods are widely used in relation extraction (Bunescu and Mooney, 2005; Zelenko *et al.*, 2003). Recently, several systems have been proposed to identify drug-gene relationships (Xu and Wang, 2012), drug–drug interactions (Percha *et al.*, 2012; Segura-Bedmar *et al.*, 2011) and gene–disease relationships (Bravo *et al.*, 2015; Pletscher-Frankild *et al.*, 2015; Ozgur *et al.*, 2008). Kim *et al.* (2010) has successfully implemented dependency kernels to extract protein-protein interaction information. The Shallow Linguistic Kernel is also used to extract drug–drug interaction information (Segura-Bedmar *et al.*, 2011).

The third phase is to rank the gene–disease relations obtained from the previous phase in an intelligent manner. A frequency-based ranking system is commonly adopted in earlier works. BeFree (Bravo *et al.*, 2015) ranks the relationships according to the frequency of occurrence. Clematide *et al.* (2012) uses a logistic regression-based method to optimize ranking of relations from curated abstracts.

There has been some work dedicated to the gene–disease relation extraction problem, which is the end-to-end problem that we try to solve in this paper. CoPub (Frijters *et al.*, 2008) provides a text mining tool that detects co-occuring biomedical concepts in abstracts from the MEDLINE literature database. BeFree (Bravo *et al.*, 2015) can be used in a text mining workflow aimed at extracting information on biological associations from scientific publications. But all current work is limited both in coverage and ranking precision.

Herein, we present a framework for extracting gene–disease relationships from biomedical literature that addresses all three phases. It takes large biomedical literature repositories and the name of a gene or a disease of interest as inputs and produces a meaningful ranked list of diseases or genes that are related to the input entity with supporting evidence. In the NER phase, we implemented an algorithm that combines dictionary-based fuzzy matching and conditional random fields (CRF) to recognize genes and diseases in free text. Next, we trained a SVM model combining lexical features and syntactic features to identify the relationships between genes and diseases. Finally, we proposed a ranking algorithm to rank the disease-related genes based on co-occurrence frequency, paper citations and author information.

2 Materials and methods

In this section, the data sources used in this work and the detailed algorithms and methods designed for this framework will be introduced.

2.1 Data sources

The data used in this work mainly include the gene/disease term libraries, the biomedical literature database, the annotated data and the ground truth.

2.1.1 Gene library

The gene library used in this study is a combination of three publicly available gene/protein databases, namely the NCBI-gene database (Brown *et al.*, 2015), the HGNC gene dataset (Gray *et al.*, 2015) and the UniProt knowledge base (Uniprot Consortium, 2015), with cross-references. NCBI-gene database contains almost all of the publicly available nucleotide sequences and their protein translations. By August, 2014, HGNC provided the names and symbols of 39 135 genes. UniProt is a comprehensive resource of protein sequence and functional information. Since genes are sometimes named after their protein names, UniProt is also used as a data source for gene recognition.

The combined gene library has 60 197 genes. Each gene and its synonyms were cross-referenced among all three data sources mentioned above.

2.1.2 Disease term library

Disease Ontology (Mitraka and Schriml, 2015), MedDRA(The Medical Dictionary for Regulatory Activities) (Brown *et al.*, 1999), UMLS (The Unified Medical Language System) (Bodenreider, 2004) and IDDB (Infectious Disease Database) were cross-referenced to yield the disease database used in this study. The disease term library is hierarchical. If disease A belongs to a parent class B, an attribute 'is a' with the value being B will be attached to the disease A. In total, the disease library includes 22 831 diseases. For each disease, the library includes its unique identifier, disease name, disease synonyms, ID in each source disease database and its parent classes.

2.1.3 Biomedical literature database

The biomedical literature database used in this study is MEDLINE (https://www.nlm.nih.gov/bsd/mms/medlineelements.html), the U.S. National Library of Medicine (NLM) journal citation database. It includes citations from more than 5600 scholarly journals, over 25 million references to biomedical and life science journal articles from as early as 1946. The downloadable database contains 779 files in the XML format in chronological order. Every piece of data contains the PMID, the publication date, the author information, the citations, etc. PubMed (http://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch) provides retrieval APIs to MEDLINE.

2.1.4 Annotated data

In order to train and evaluate the methods and tools used in this work, we gathered annotated texts containing a total of 2340 positive disease-gene relationship labels and 1437 negative relationship labels. 2113 of the positive labels and 1010 of the negative labels are from Genetic Association Databases (GAD) (Becker *et al.*, 2004), a database of genetic association data from complex diseases and disorders. Since GAD did not label all gene/disease entities contained within a sentence, which will bias our NER training, many of the data from GAD are re-annotated manually by domain experts for gene entities, disease entities, positions of genes and positions of diseases at the sentence level. The remaining training data are sentences randomly selected from the biomedical abstracts in MEDLINE. Gene entities, disease entities, positions of genes, positions of diseases and the relationships between genes and diseases are manually labeled in each sentence.

2.1.5 Ground truth

Human annotated gene–disease associations for 10 randomly selected diseases obtained from DisGeNET (http://www.disgenet.org/) serve as our ground truth dataset. The diseases include Retinitis Pigmentosa, Adrenal Gland Chromaffinoma, Bipolar I Disorder, Hyperlipidemia, Papilloma, Thrombocytopenia, Glioblastoma, Hernia Diaphragmatic, Brain Ischemia and Cerebrovascular Accident. The ground truth is used to evaluate the end-to-end results of a system.

2.2 Extraction workflow

To extract gene–disease associations from biomedical literature, we aim to extract from MEDLINE triples of the format (disease; gene; score). ‘Disease’ and ‘gene’ denote the unique identifiers of diseases and genes respectively in our disease library and gene library. ‘Score’ refers to the plausibility of gene–disease associations. When the user queries the relevant genes of a particular disease, the framework will return the results based on ‘score’.

Specifically, to accomplish the goal, we first pre-process the data from MEDLINE and obtain three kinds of information: (i) Title and abstract of articles, (ii) author information and (iii) article reference information. Then, the gene/disease recognition module processes the titles and abstracts using both a Stanford NER tool (Finkel *et al.*, 2005) trained on 2000 annotated sentences (1000 from GAD and 1000 from our manually labeled data library) and a dictionary-based longest match strategy using the gene and disease libraries. Recognition results of the two methods are combined. In case of discrepancies, heuristic rules are applied to resolve them. Next, in the association detection phase, all recognized gene–disease pairs that co-occur within the same sentence are considered as candidate evidence. A binary SVM classifier, which extracts two types of features, namely local lexical features and global syntactic features, is used to determine the plausibility of the candidate pairs. Finally, positive pairs are ranked by three methods. The basic method is counting the co-occurrence frequency. The second method is to weigh each co-occurrence by the PageRank of the paper from which the evidence was extracted, in a paper citation network constructed from PubMed. The last and most advanced method considers the duplicated evidence published by the same author, and thus suppresses the contribution of such evidence. Details of each part of the workflow will be presented in the sections below.

2.3 Gene/disease recognition

The gene/disease recognition task is an NER task where the entities being recognized are genes and diseases. We brought up a hybrid recognition method in this study. A CRF-based NER tool, namely the Stanford NER tool (Finkel *et al.*, 2005), is combined with a dictionary-based longest match strategy.

The Stanford NER tool provides a general implementation of (arbitrary order) linear chain CRF sequence models. It was trained on 2000 annotated sentences from MEDLINE before being used to recognize genes and diseases names in the titles and abstracts of articles from MEDLINE. Recognized genes and diseases are then mapped to the unique identifiers in the gene and disease libraries. The Stanford NER is further enhanced in the following fashion. If a gene or a disease recognized by the Stanford NER tool cannot be found in the libraries, a web crawler will automatically query Bing (https://www.bing.com/) to search for information that can relate the gene/disease name to one of the unique identifiers. A gene/disease is added to the results only when it can be related to a unique identifier. If a gene/disease that can be related to a unique identifier is currently not in its synonym set, it will be added as a synonym.

Next, a dictionary-based longest match algorithm is implemented using the gene and disease libraries as dictionaries. The longest match strategy uses a sliding window. A fixed-length window slides within the sentence and the words within the window are fuzzy matched to the items included in the gene and disease libraries. The ‘fuzzy match’ strategy picks up both terms that are exact matches and terms with small discrepancies in some punctuation or singular/plural form. (For example, ‘lung type-i cell membrane-associated glycoprotein’ is considered a match to ‘lung type i cell membrane-associated glycoprotein’, and ‘benign gastric tumours’ is considered a match to ‘benign gastric tumour’.) If the words covered by the window are matched to a gene or a disease, they will be marked to make sure that other matched words will not overlap with them. When the window reaches the end of the sentence, the window length is reduced by one, and the window resumes sliding from the start of the sentence, until the length is zero.

The results of the two methods are combined and a number of heuristic rules are applied to resolve possible discrepancies: (i) If a term’s length is less than four characters, it is likely to be either an acronym or an incorrect recognition. It will be treated as an acronym if its longer synonym (complete form) occurs in the former context. (ii) If a term with less than four characters does not have a longer synonym occurring in the text before it, but it is recognized by the enhanced Stanford NER, then it will be added to the final results. (iii) If a term is recognized by both the dictionary-based method and the enhanced Stanford NER, the result of the enhanced Stanford NER is used. The rationale is that the enhanced Stanford NER takes advantage of web search engines, which encompasses more knowledge and therefore is potentially more accurate in recognizing named entities.

Genes and diseases may sometimes share the same synonyms. Thus, a term may be mapped to several unique identifiers. However, if a synonym of the term occurs in the previous context and the synonym can be mapped to a specific unique identifier, the term will be considered as belonging to the same unique identifier. Otherwise, the term is mapped to each of the unique identifiers once. Considering the hierarchies among genes and diseases, a term will be mapped not only to its unique identifier, but also its parent class’s unique identifier.

Gene–disease pairs recognized in this phase are considered candidate evidences for association between the gene and the disease.

2.4 Association detection

In this step, a binary SVM classifier is implemented to determine the plausibility of association between a gene and a disease based on the gene–disease pairs extracted from the previous step. Given a sentence $S = w_1,\ldots,g,\ldots,w_i,\ldots,d,\ldots,w_n$, the classifier decides whether there is a gene–disease relation between entities g and d.

The classifier utilizes two types of features, namely local lexical and global syntactic features. The local lexical feature contains words surrounding the gene or the disease terms in the original text. The global syntactic feature, which contains unigrams, bigrams and trigrams drawn from (i) the shortest path between the gene and the disease terms in the dependency tree, and (ii) the path between the least common ancestor (LCA) of the two terms and the root of the dependency tree. The features are detailed in Table 1. Feature 1 is the local lexical feature and contains the context information of the gene and the disease terms, while features 2 and 3 are the global syntactic features and contain rich syntactic information from the dependency tree. The lemmas and the dependency tree are generated by the Stanford CoreNLP tool (Manning *et al.*, 2014).

The effect of words with the part-of-speech (POS) tag 'neg' or 'advmod' that modify verbs is taken into account in feature extraction. (For example, the word 'not' in 'does not relate with' or the word 'rarely' in 'rarely indicates the association' strongly implies the negative association of a candidate pair.). Words with POS tag 'neg' or 'advmod' that modify verbs are included in the paths used by global features. Besides, the gene and the disease terms in the paths are replaced with the general names 'GENE' and 'DISEASE', because specific gene/disease names do not contain information concerning association detection (Fig. 1).

Take the following sentence as an example:

All three complementary approaches employed (family-based, case-control and quantitative trait design) suggest a role for the MAO A promoter-region polymorphism in conferring risk for ADHD in our patient population.

In this sentence, 'MAO A' is a gene term and 'ADHD' is a disease name. A part of the dependency tree of the sentence is drawn in Figure 2. 'for', 'the', 'promoter-region', 'polymorphism', 'risk', 'for', 'in' and 'our' are extracted for feature 1. From the shortest dependency path between the gene and the disease terms, n-grams like 'risk for DISEASE' and 'in risk for' are extracted for feature 2. The LCA of the gene and the disease terms in the dependency tree is 'suggest'. N-grams including 'suggest role for' and 'role for polymorphism' are extracted for feature 3.

Table 1. Features extracted for association detection

Feature	Type	Description
1	Local lexical feature	Lemmas of the two words in front of the gene term and the two words behind the gene term, and lemmas of the two words in front of the disease term and the two words behind the disease term
2	Global syntactic feature	Unigram, bigram and trigram of lemmas on the shortest path between the gene and the disease terms in the dependency tree
3	Global syntactic feature	Unigram, bigram and trigram of lemmas on the path between the LCA of the gene and the disease terms and the root of the dependency tree

Data Preprocessing
Gene/Disease Recognition
Association Detection
Ranking

Fig. 1. Work flow of extracting gene–disease associations from MEDLINE

After the features are extracted, libsvm (Chang and Lin, 2011) is used to train the binary SVM classifier. The kernel function of SVM grades the local lexical feature (i.e. feature 1) and the global syntactic features (i.e. feature 2 and 3) and combines the scores linearly. The features are treated as a bag of words. Every possible word or n-gram is considered a dimension in a vector. If the feature contains a particular word/n-gram, the value of the corresponding dimension is set to one; otherwise, it is zero. The similarity between two instances of features is quantified by the cosine value between vectors.

A set of positive gene–disease pairs along with the unique identifiers of journal articles in which they co-occur are generated and passed on to the next step.

2.5 Ranking

For a given disease, hundreds of positive gene–disease pairs are generated from the previous steps, and vice versa. The ranking of genes for a particular disease and the ranking of diseases for a particular gene is done using the following three ranking methods.

The basic ranking method uses the co-occurrence frequency. Different pairs with the same disease are ranked by the number of distinct journal article in which they co-occur.

The second method is to grant a weight to each co-occurrence according to the PageRank of the journal article, in a paper citation network constructed from PubMed. PageRank was proposed by Larry Page and Sergey Brin (Page et al., 1999). It is used by Google for website ranking. PageRank operates on the idea that the more important a website is, the more websites will link to it. So it utilizes the number and importance of websites that link to a given website to assess the importance of this website, or the PageRank of the website. In this study, we apply the principle of PageRank to the task of publication citation. If a journal article is cited by more articles and the articles that cite it are more influential, the article itself is more influential and its PageRank is higher.

According to the reference information obtained from MEDLINE, we first built a paper citation network, as shown in Figure 3. Nodes in the network denote the journal articles, and the

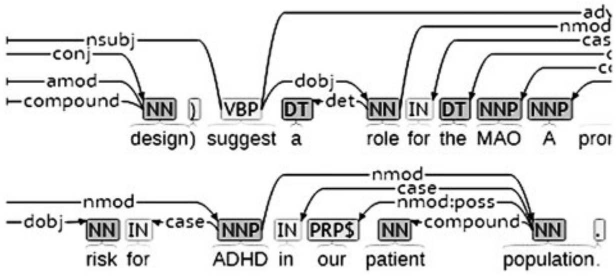


Fig. 2. Part of the Dependency Tree of sentence 'All three complementary approaches employed (family-based, case-control and quantitative trait design) suggests a role for the MAO A promoter-region polymorphism in conferring risk for ADHD in our patient population'

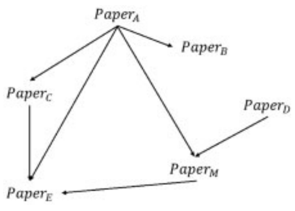


Fig. 3. Simplified paper citation network

edge from node A to node B means that Paper A cites Paper B. Based on the paper citation network, the PageRank of each node (i.e. each article) is calculated.

When we get the PageRank of the articles, the following formula is used to calculate the score of the gene–disease pair, where g denotes the gene, d denotes the disease, $C_{(g,d)}$ denotes the set of all the articles that contain the (g,d) pair and pr(a) denotes the PageRank of paper a.

$$Score(g, d) = \sum_{a \in C_{(g,d)}} pr(a) \tag{1}$$

The third ranking method takes the authors into consideration. Biomedical researchers, who focus on specific diseases or genes, may write about the same gene–disease pair in multiple publications. For instance, Mark J Sarnak mentioned the association between Kidney Failure and Cystatin C in up to twelve distinct papers. They should not be considered as twelve independent evidences. In the third method, if a gene–disease pair is repeatedly mentioned by the same author, the contribution of the duplicated evidence will be suppressed.

We assume that the sum of weights an author can grant to different evidences of the same pair is one. If an author mentions the same pair in multiple publications, the weights of the evidence from each of the publications is equal and sums to one. The weight of each paper where the evidence of (g,d) is extracted is

$$w_a(g, d) = \frac{\sum x \in l \frac{1}{|c_x|}}{|l|} \tag{2}$$

where *l* denotes the author list of paper a, $|c_x|$ denotes the number of papers author x wrote about (g,d). Then we can modify the score function of gene–disease pairs proposed in the formula (1).

$$Score(g, d) = \sum_{a \in C_{(g,d)}} w_a(g, d) \times pr(a) \tag{3}$$

The fourth method uses a PageRank function that is adjusted by a time factor. This is based on the observation that recently published papers may have 'less exposure' for citation than those published before them. The time-weighted PageRank is defined as below.

$$pr(u) = d \sum_{v \in B(u)} \frac{pr(v)}{N_v} + (1 - d) \times T_u \tag{4}$$

where T_u denotes a smoothed time factor related to each paper's year of publication. Disease gene pairs mentioned by the same author is also be suppressed in this method.

2.6 Significance testing

The micro sign test (Yang *et al.*, 1999) is used to examine whether the improvements in F1-scores are statistically significant. A one-sided *P*-value that is less than 0.05 is considered as statistically significant.

3 Results

The number of associations extracted by our method is shown in Table 2. The number of associations extracted by our method is significantly larger than Befree (Bravo *et al.*, 2015). At the meantime, the precision of our extraction is better than Befree, which is detailed in Section 3.3.

Table 2. Results of extracted associations compared to BeFree

	Associations	Genes	Diseases
DTMiner	**1 728 535**	**16 893**	**9950**
BeFree	131 012	2803	2751

Bold text signifies the best performer in the column.

Table 3. Results of gene/disease recognition

	Precision	Recall	F-score
ABNER	0.593	0.549	0.57
Only dictionary	0.839	0.659	0.738
Stanford NER tool	**0.954**	0.524	0.673
Before enriched by Bing	0.851	0.875	0.863
After enriched by Bing	0.87	**0.885**	**0.877**

Bold text signifies the best performer in the column.

Table 4. Results of gene/disease relation extraction

Feature	Precision	Recall	F-score
Local Lexical Feathers	0.761	0.748	0.755
Global Syntactic Features	0.827	0.853	0.839
Local+Global features	**0.846**	**0.88**	**0.863**

Bold text signifies the best performer in the column.

Table 5. Comparison of relation extraction performance

Framework	F-score	Training Time (s)	Testing Time (s)
DTMiner	0.863	**24**	**241**
BeFree	**0.898**	384	4393

Bold text signifies the best performer in the column.

3.1 Gene/disease recognition

First, we evaluated the Gene/Disease recognition module on 800 annotated sentences. In these 800 sentences, 592 diseases and 525 genes were labeled. The recognition results are shown in Table 3. We found that Stanford NER tool has high precision but poor recall. By combining it with dictionaries, our hybrid recognition method has about a 0.125 enhancement on the F-score than the dictionary-only method, which is similar to the method used by BeFree (Bravo *et al.*, 2015) and CoPub (Frijters *et al.*, 2008). The F-score improved about 0.014 ($P < 0.01$) after we enriched our gene and disease dictionary using Bing. Our method also performs much better than ABNER (Settles, 2004) on these sentences.

3.2 Relation extraction

We evaluated the performance of the SVM classifier through 10-fold cross validation on a training set with 2080 positive samples and 1277 negative samples. The result is shown in Table 4. We found that the F-score can only reach 0.755 if we use the local lexical features alone. However, if we use both the local features and the global features, the SVM classifier achieves the best performance.

Due to the large volume of biomedical literature, computation time becomes an important issue in the relationship extraction task. A comparison between the state-of-art system BeFree and our work is shown in Table 5. Although the F-score of our work is slightly lower than that of BeFree, the training speed of the classifier and the testing speed on 420 randomly selected instances of our work are significantly (more than 10x) faster.

3.3 Ranking

The ground truth is obtained from DisGeNET. We selected 10 diseases and computed the mean reciprocal rank (MRR)of ranks obtained by the three methods mentioned above. The result is shown in Table 6. For most diseases, the suppressed PageRank achieves the best performance. Overall, compared with the frequency-based method, the MRR increased 10.6% if we weighted the paper with PageRank score. And, after suppressing the contribution of the same author, the MRR increased by 8.4% .

Then, we randomly chose five different diseases (cerebrovascular accident, brain ischemia, hernia diaphragmatic, thrombocytopenia and retinitis pigmentosa) and evaluated the top K ranking of our results. The result is shown in Table 7. If we select the top-50 genes, the Suppressed PageRank method achieves the better result compared with other methods. While K increases, the performance of these four methods' is similar. This is because when K increases, most of the known disease-related genes are ranked on top. The numbers of true positives, false positives and false negatives for the Suppressed PageRank method are listed in the supplementary material, as well as some examples.

We also compared our results with BeFree (Bravo *et al.*, 2015) and Copub (Frijters *et al.*, 2008). The results are shown in Figure 4.

Table 6. MRR results of different ranking methods

Disease	Frequency-based	PageRank-based	Suppressed PageRank	Weighted PageRank
Retinitis pigmentosa	0.111	0.141	0.146	0.156
Adrenal gland chromaffinoma	0.161	0.194	0.212	0.207
Bipolar I disorder	0.26	0.268	0.273	0.26
Hyperlipidemia	0.267	0.279	0.325	0.325
Papilloma	0.183	0.24	0.321	0.226
Thrombocytopenia	0.089	0.072	0.063	0.063
Glioblastoma	0.278	0.3	0.346	0.274
Hernia diaphragmatic	0.026	0.028	0.029	0.027
Brain ischemia	0.052	0.076	0.067	0.06
Cerebrovascular accident	0.178	0.178	0.147	0.175
Overall	0.161	0.178	0.193	0.177

Table 7. F-score of top-K of the rankings

Top K	Freq-Based	PR-Based	Weighted PR	Sup-PR	BeFree	CoPub
K = 50	0.221	0.228	0.237	**0.241**	0.213	0.212
K = 100	0.236	0.235	0.230	**0.238**	0.214	0.211
K = 150	0.22	**0.221**	0.216	0.216	0.186	0.192
K = 200	0.196	0.196	**0.201**	0.197	0.167	0.175

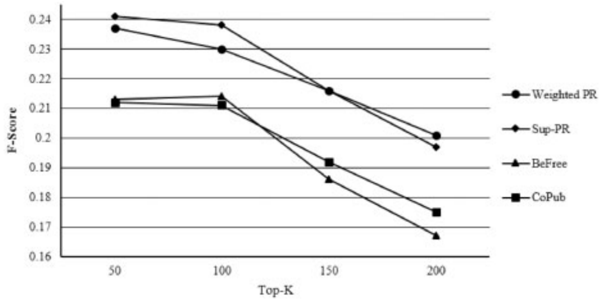


Fig. 4. F-score of Top-K Rankings

In most cases, Suppressed PageRank results are about 0.03 higher than CoPub and BeFree. In addition, our system extracts many more genes associated with a specific disease. (cerebrovascular accident: 1238 associated genes; brain ischemia: 1588; hernia diaphragmatic: 266; thrombocytopenia: 1049; retinitis pigmentosa: 818) For example, we extracted 818 genes associated with *retinitis pigmentosa*, while BeFree only extracts 193 genes (of which 142 are in our results) and Copub extracts 179 genes (of which 124 are in our results). Manual inspection of 40 of the 818 genes reveals that 34 of these associations are correct. This shows that our approach not only achieves much better coverage, but also competitive accuracy.

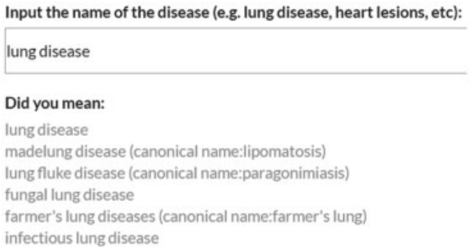
3.4 Web-based service

The DTMiner web server is a gene–disease association discovery platform using the U.S. National Library of Medicine (NLM) journal citation database, MEDLINE, as a data source. It is accessible from the website: http://gdr-web.rwebox.com/public_html/index.php. DTMiner allows user-friendly access to a gene–disease relationship database. The associations between genes and diseases are represented in a bipartite graph and it permits both queries of genes and diseases. For a disease (or gene) query, DTMiner will provide multiple disease (or gene) choices ranked by the string similarity of the users' input. Next, the website displays all the genes (or diseases) related to the users' selection according to their relative weighted PageRank. Further, a user can find the evidence of each disease-gene pair (the papers where they co-occur) for further details. See Figures 5–7 for example.

We also provide a URL for users to submit a query directly from a program and return Json-encoded results to the program, which is http://gdr-web.rwebox.com/public_html/get-disease.php for disease and http://gdr-web.rwebox.com/public_html/get-gene.php for gene. Please see http://gdr-web.rwebox.com/public_html/index.php?page=help.php for more details.

4 Discussion

In the named entity recognition phase, the dictionary-only method, which is similar to the method BeFree (Bravo *et al.*, 2015) and



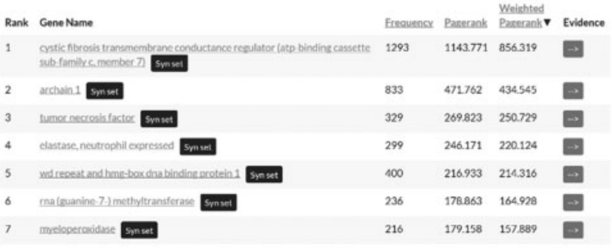
Input the name of the disease (e.g. lung disease, heart lesions, etc):

lung disease

Did you mean:

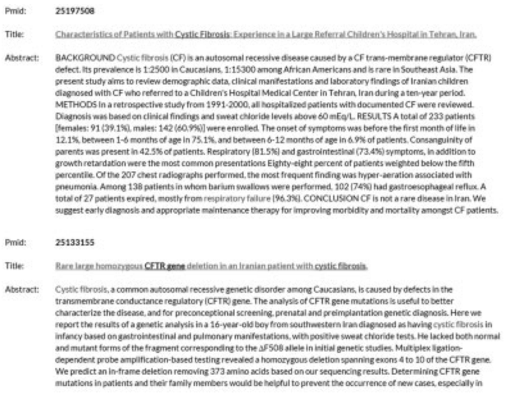
lung disease
madelung disease (canonical name:lipomatosis)
lung fluke disease (canonical name:paragonimiasis)
fungal lung disease
farmer's lung diseases (canonical name:farmer's lung)
infectious lung disease

Fig. 5. Search for lung disease



Rank	Gene Name	Frequency	Pagerank	Weighted Pagerank▼	Evidence
1	cystic fibrosis transmembrane conductance regulator (atp-binding cassette sub-family c, member 7) Syn set	1293	1143.771	856.319	-->
2	archain 1 Syn set	833	471.762	434.545	-->
3	tumor necrosis factor Syn set	329	269.823	250.729	-->
4	elastase, neutrophil expressed Syn set	299	246.171	220.124	-->
5	wd repeat and hmg-box dna binding protein 1 Syn set	400	216.933	214.316	-->
6	rna (guanine-7-) methyltransferase Syn set	236	178.863	164.928	-->
7	myeloperoxidase Syn set	216	179.158	157.889	-->

Fig. 6. Select the first choice



Pmid: **25197508**

Title: Characteristics of Patients with **Cystic Fibrosis**: Experience in a Large Referral Children's Hospital in Tehran, Iran.

Abstract: BACKGROUND Cystic fibrosis (CF) is an autosomal recessive disease caused by a CF trans-membrane regulator (CFTR) defect. Its prevalence is 1:2500 in Caucasians, 1:15300 among African Americans and is rare in Southeast Asia. The present study aims to review demographic data, clinical manifestations and laboratory findings of Iranian children diagnosed with CF who referred to a Children's Hospital Medical Center in Tehran, Iran during a ten-year period. METHODS In a retrospective study from 1991-2000, all hospitalized patients with documented CF were reviewed. Diagnosis was based on clinical findings and sweat chloride levels above 60 mEq/L. RESULTS A total of 233 patients [females: 91 (39.1%), males: 142 (60.9%)] were enrolled. The onset of symptoms was before the first month of life in 12.1%, between 1-6 months of age in 75.1%, and between 6-12 months of age in 6.9% of patients. Consanguinity of parents was present in 42.5% of patients. Respiratory (81.5%) and gastrointestinal (73.4%) symptoms, in addition to growth retardation were the most common presentations Eighty-eight percent of patients weighted below the fifth percentile. Of the 207 chest radiographs performed, the most frequent finding was hyper-aeration associated with pneumonia. Among 138 patients in whom barium swallows were performed, 102 (74%) had gastroesophageal reflux. A total of 27 patients expired, mostly from respiratory failure (96.3%). CONCLUSION CF is not a rare disease in Iran. We suggest early diagnosis and appropriate maintenance therapy for improving morbidity and mortality amongst CF patients.

Pmid: **25133155**

Title: Rare large homozygous **CFTR gene** deletion in an Iranian patient with **cystic fibrosis**.

Abstract: Cystic fibrosis, a common autosomal recessive genetic disorder among Caucasians, is caused by defects in the transmembrane conductance regulatory (CFTR) gene. The analysis of CFTR gene mutations is useful to better characterize the disease, and for preconceptional screening, prenatal and preimplantation genetic diagnosis. Here we report the results of a genetic analysis in a 16-year-old boy from southwestern Iran diagnosed as having cystic fibrosis in infancy based on gastrointestinal and pulmonary manifestations, with positive sweat chloride tests. He lacked both normal and mutant forms of the fragment corresponding to the ΔF508 allele in initial genetic studies. Multiplex ligation-dependent probe amplification-based testing revealed a homozygous deletion spanning exons 4 to 10 of the CFTR gene. We predict an in-frame deletion removing 373 amino acids based on our sequencing results. Determining CFTR gene mutations in patients and their family members would be helpful to prevent the occurrence of new cases, especially in

Fig. 7. Click evidence for more details

CoPub (Frijters *et al.*, 2008) used, shows an F-score of 0.738. We achieved a 0.125 improvement on the F-score by combining the Stanford NER tool with the dictionary-based method. The CRF model behind the Stanford NER tool takes advantage of the distribution information of words and thus compensates the dictionary-only method. Incorporating the Internet search engine further improved the F-score by 0.014, for that information from the Internet enriched our gene/disease libraries.

In the relation extraction phase, we achieved an F-score of 0.863 using both the local lexical features and global syntactic features. Using the local lexical features or the global syntactic features alone gives a lower F-score, which indicates that, both the local and the global features are effective and necessary. While the F-score of our method is nearly as good as that of BeFree (Bravo *et al.*, 2015), our speed is more than 10 times faster. It is because we utilized fewer but more representative features than BeFree so that we can achieve this advantage in speed while maintaining the F-score.

In terms of the MRR results, the suppressed PageRank method achieved the best overall performance. The PageRank-based method increased the overall MRR by 10.6%. This is because our PageRank-based method weighted the trustworthiness of each paper according to the citation information. The suppressed PageRank method further improved the overall MRR by 8.4%. This supports our assumption that some authors wrote multiple papers about the same gene–disease association and such behavior will falsely boost the ranking of the aforementioned association in the PageRank-based method. The same judgement on the performance of the four ranking methods is given by the F-score results. Adding a time factor in the PageRank function does not seem to improve the MRR. This may suggest that cutting edge research get cited frequently despite the short amount of time since publication.

Furthermore, our end-to-end results showed that our system extracted much more genes related to a given disease than BeFree (Bravo *et al.*, 2015) and CoPub (Frijters *et al.*, 2008) did. This is due to the large scale of our input and the good performance of our NER method and relationship extraction method.

5 Conclusion

In this paper, we proposed a framework for the automatic extraction of gene–disease relation from biomedical literature. For the gene and disease recognition, we built large gene and disease libraries by combining and cross referencing existing biomedical knowledge bases. We enriched our dictionary by identifying new synonyms using search engines. For association detection, we designated effective features and built an efficient SVM classifier. For the ranking, we considered the weight of papers and the other contributions by the same authors and proposed three different algorithms for ranking. In addition, we launched a web service for public access to our results. Overall, our system created a disease-gene association dataset that is several times larger than any previously reported dataset of similar nature, achieved a good balance between accuracy and computation time, and outperformed existing state-of-the-art systems on similar tasks. The system has been made available online for free public access, which will potentially enable the discovery and development of new therapeutics and breakthrough in diagnostics.

Funding

This work has been supported by AstraZeneca.

Conflict of Interest: none declared.

References

Agirre,E. *et al.* (2000) Enriching very large ontologies using the WWW. arXiv preprint cs/0010026.

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Bravo,A. *et al.* (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinf.*, **16**, 55.

Brown,E.G. *et al.* (1999) The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, **20**, 109–117.

Brown,G.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.

Bunescu,R.C. and Mooney,R.J. (2005) Proceedings of the conference on human language technology and empirical methods in natural language processing. *Association for Computational Linguistics*, pp. 724–731.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)*, **2**, 27.

Clematide,S. *et al.* (2012) Ranking relations between diseases, drugs and genes for a curation task. *J. Biomed. Seman.*, S5.

Collins,M. and Singer,Y. (1999) Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, Chicago, IL, 100–110.

Finkel,J.R. *et al.* (2005) *Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling*. ACL' 05, Stroudsburg, PA, pp. 363–370.

Frijters,R. *et al.* (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.*, **36**, W406–W410.

Ginsberg,J. *et al.* (2009) Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012–1014.

Gray,K.A. *et al.* (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, D1079–D1085.

Hanisch,D. *et al.* (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinf.*, **6**, S14.

Ju,Z. *et al.* (2011) *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on IEEE*, pp. 1–4.

Kim,J. *et al.* (2015) LGscore: A method to identify disease-related genes using biological literature and Google data. *J. Biomed. Inf.*, **54**, 270–282.

Kim,S. *et al.* (2010) Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinf.*, **11**, 107.

Manning,C.D. *et al.* (2014) *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Mitraka,E. & Schriml,L.M. (2015) 2015 Disease Ontology update: DO's expanded curation activities to connect disease-related data.

Ozgur,A. *et al*. (2008) Identifying gene–disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**, i277–i285.

Page,L., Brin,S., Motwani,R. & Winograd,T. (1999) The PageRank citation ranking: bringing order to the web.

Percha,B. *et al*. (2012) Discovery and explanation of drug–drug interactions via text mining. Pacific Symposium on Biocomputing. *Pac. Symp. Biocomput*., 410–421.

Pletscher-Frankild,S. *et al*. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.

Segura-Bedmar,I. *et al*. (2011) Using a shallow linguistic kernel for drug–drug interaction extraction. *J. Biomed. Inf*., **44**, 789–804.

Settles,B. (2004) Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. *Association for Computational Linguistics*, pp. 104–107.

Uniprot Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res*., **43**, D204–D212.

Xu,R. and Wang,Q. (2012) A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text. *J. Biomed. Inf*., **45**, 827–834.

Yang,Y. *et al*. (1999) A re-examination of text categorization methods, *SIGIR'99*, ACM, New York, NY, 42–49.

Zelenko,D. *et al*. (2003) Kernel methods for relation extraction. *J. Mach. Learn. Res*., **3**, 1083–1106.