Gene–disease association extraction by text mining and network analysis

Changqin Quan
AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine,
School of Computer and Information,
HeFei University of Technology
quanchqin@gmail.com

Fuji Ren
Faculty of Engineering,
University of Tokushima,
ren@is.tokushima-u.ac.jp

Abstract

Biomedical relations play an important role in biological processes. In this work, we combine information filtering, grammar parsing and network analysis for gene-disease association extraction. The proposed method first extracts sentences potentially containing information about gene-diseases interactions based on maximum entropy classifier with topic features. And then Probabilistic Context–Free Grammars is applied for gene-disease association extraction. The network of genes and the disease is constituted by the extracted interactions, network centrality metrics are used for calculating the importance of each gene. We used breast cancer as testing disease for system evaluation. The 31 top ranked genes and diseases by the weighted degree, betweenness, and closeness centralities have been checked relevance with breast cancer through NCBI database. The evaluation showed 83.9% accuracy for the testing genes and diseases, 74.2% accuracy for the testing genes.

1 Introduction

Since the start of Human Genome Project in 1990, over 40 kinds of organism genome have been sequenced. Biological databases expand rapidly with the exponential growth of biological data. For instance, until now, over 260,000 named organisms have their nucleotide sequences in the GenBank (Benson et al. 2008) which integrates data from the major DNA and protein sequence. However, data is not information. Compared with situations before 2003, the key problem today has turned to methods of knowledge extraction. Understanding the role of genetics in diseases is one of the major goals of the post-genome era. The expanding rate of knowledge in gene–disease associations can hardly match up with the growth of biological data. It takes time before new discoveries are included in the databases such as Online Mendelian Inheritance in Man (OMIM), and most of the information represented in these databases is manually collected from literature.

To address this challenge, we proposed an automatic gene-disease association extraction approach based on text mining and network analysis. We combine information filtering, grammar parsing and network analysis. We started by calculating main topics of each sentences in the corpus based on supervised Latent Dirichlet Allocation (sLDA) model (Blei and McAuliffe 2007). The most probable topics derived from sLDA model for each sentence are used as features for training maximum entropy (MaxEnt) (Manning and Schutze, 1999) classifier, which extracts sentences potentially containing information about gene-diseases interactions. After that, Probabilistic Context–Free Grammars (PCFGs) (Klein and Christopher 2003) is applied for sentence grammar parsing. Based on the syntactic tree of each sentence, we extract paths between specific entities such as diseases or genes. The network of all candidate genes and the disease is constituted by the interactions extracted from the sentences in the corpus. Our main hypothesis in network analysis is that the most important and the most central genes in an interaction network are most likely to be related to the disease. Last, network centrality metrics are used for calculating the importance of each gene.

The rest of this paper is organized as follows. Section 2 surveys related work. In Section 3, we introduce the proposed approach of extracting interactions from literature. Section 4 presents gene-disease interaction network analysis. And

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 54–63,
Gothenburg, Sweden, April 26-30 2014.

then Section 5 presents and discusses the experimental results. Lastly we conclude this paper and discuss future work in Section 6.

2 Related Work

Much effort is currently spent on extracting gene–disease associations (Özgür et al. 2008; Chun et al. 2006). Biomedical relation extraction techniques basically include two branches: interaction database based methods and text mining methods. Interaction database based methods rely on the availability of interaction databases, such as OMIM, MINT (Zanzoni et al. 2002), IntAct (Kerrien et al. 2012), BIND (Bader et al. 2003), which predict interactions between entities using sequence, structural, or evolutionary information (Krallinger, Leitner, and Valencia 2010). Although these databases host a large collection of manually extracted interactions from the literature, manually curated databases require considerable effort and time with the rapid increasing of biomedical literature.

Since most biological facts are available in the free text of biomedical articles, the wealth of interaction information provided in biomedical articles motivated the implementation of text mining approaches to automatically extract biomedical relations. Text mining approaches to gene–disease association extraction have shown an evolution from simple systems that rely solely on co-occurrence statistics (Adamic et al. 2002; Al-Mubaid and Singh 2005) to complex systems utilizing natural language processing techniques and machine learning algorithms (Freudenberg and Propping 2002; Glenisson et al. 2004; Özgür et al. 2008). Well-known tools for discovering gene–disease associations include DAVID (Huang et al. 2009), GSEA (Subramanian et al. 2005), GOToolBox (Martin et al. 2004), rcNet (Huang et al. 2011) and many others. However, in many cases, since the existing annotations of disease-causative genes is far from complete (McKusick 2007), and a gene set might only contain a short list of poorly annotated genes, existing approaches often fail to reveal the associations between gene sets and disease phenotypes (Huang et al. 2011).

Network-based approaches (Wuchty, Oltvai, and Barabási, 2003; Schwikowski et al. 2000; Chen et al. 2006) is performed by assessing how much genes interact together and are close to known disease genes in protein networks. Relation extraction among genes is the fundamental step for gene-interaction network creation. Recently, syntactic analysis has been considered for relation extraction, and different parsing grammars have been applied. Temkin and Gilder (2003) used a full parser with a lexical analyzer and a context free grammar (CFG) to extract protein-protein interactions. In Yakushiji et al. (2005)'s work, they proposed a protein-protein interaction extraction system based on head-driven phrase structure grammar (HPSG). Although the pattern generation is complicated, the performance is not satisfactory. In addition, dependency grammar is used frequently in this domain. Erkan et al. (2007) proposed a semi-supervised classification for extracting protein interaction sentences using dependency parsing. Katrin et al. (2007) defined some rules based on dependency parse tree for relation extraction. The problem of those systems using dependency parse is that they cannot treat non-local dependencies, and thus rules acquired from the constructions are partial (Yakushiji et al. 2005). Differently, in this work, we apply sentence filtering based on topics and phrase structure parsing for relation extraction. The extracted sentences potentially contain information about gene-diseases interactions. Phrase structure grammars are based on the constituency relation, as opposed to the dependency relation associated with dependency grammars. Phrase structure parsing is full parsing, which takes into account the full sentence structure.

In addition, many researches (Aerts et al. 2005; Chen et al. 2009; Ma et al. 2007; Hutz et al. 2008; Morrison et al. 2005; Özgür et al. 2008) used an initial list of seed genes to build a disease-specific gene-interaction network, and thus they are biased in favor of the seed genes, consequently the results also depend on the pickup seed genes.

3 Extracting interactions from literature

3.1 The Corpus

We used 44,064 articles from PubMed Central (PMC) Open Access which is a free full-text archive of biomedical and life sciences journal literature. All articles were extracted by querying the keyword of "breast cancer". We applied a segmentation tool Splitta for segmenting articles into sentences which includes proper tokenization and models for high accuracy

sentence boundary detection with reported error rates near 0.25% coded by Gillick (2009).

A gene name dictionary was built from OMIM database. The disease name dictionary was built based on Genetic Association Database (GAD) which is an archive of human genetic association studies of complex diseases and disorders.

3.2 Key sentences extraction

We applied MaxEnt classifier with topic features for key sentences extraction. The extracted sentences potentially contain information about genes and breast cancer interactions.

A Latent Dirichlet Allocation (LDA) model was used to infer topics of sentences. Three most probable topics of each sentence were put into trained MaxEnt classifier as features for extracting sentences that potentially contain interaction relationship between genes and diseases.

3.2.1 Key words annotation

We assume that each sentence indicating interactions should contain at least one gene and target disease name. Key words are the words increasing possibility of sentence containing interaction relationships, such as genes and diseases. As mentioned above, we built the gene name dictionary with data from OMIM database and disease name dictionary from Genetic Association Database (GAD). All gene names and disease names were considered as key words.

3.2.2 Topic model based on Gibbs Sampling

Latent Dirichlet Allocation (LDA) was applied based on Gibbs Sampling method in our system. Compared with algorithm obtaining approximate maximum-likelihood estimates for topics-words distribution and the hyperparameters of the prior on documents-topics distribution given by Blei, Ng and Jordan (2002), Gibbs Sampling method doesn't need to explicitly represent the model parameters which effect on the final results (Griffiths, 2002).

For a word w in a specific article, the possibility it belongs to topic j can be given by :

$$P(z_i = j \mid z_{-i}, w) \propto P(w_i \mid z_i = j, z_{-i}, w_{-i}) P(z_i = j \mid z_{-i}) \quad (1)$$

where z_i represents current topic, z_{-i} represents all topics except for i, w represents all words in the article, w_i represents current word and w_{-i} represents all words except for w_i.

Formula (1) could be represented as follow after derivation:

$$P(z_i = j \mid z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,\bullet}^{(d_i)} + T\alpha} \quad (2)$$

where $n_{-i,j}^{(\bullet)}$ represents count of words belong to topic j except for current word. $n_{-i,j}^{(w_i)}$ represents count of word w_i belong to topic j in the article except for current one. $n^{(d_i)}$ represents total of words in article d_i, while $n_{-i,\bullet}^{(d_i)}$ represents count of words in document d_i not including the current one. α and β are hyperparameters that determine extent of smooth of this empirical distribution, and how heavily this distribution can be chosen to give the desired resolution in the resulting distribution. W stands for count of words while T stands for count of topics.

3.2.3 Training of topic model

We randomly selected sentences from 8000 documents in our corpus as training set and set number of topics as 10. Topic that contains most words in gene name dictionary and disease name dictionary was treated as a key topic. Then we manually assigned each word in gene name dictionary or disease name dictionary to key topic, and each word doesn't belong to the two dictionaries was assigned to the most probable topic of itself.

3.2.4 Prediction of key sentences

The sentences containing interactions among genes or diseases were marked as 'Key' and others were marked as 'None'. A MaxEnt classifier [1] was trained based on the topic distribution.

3.3 Extracting interactions from key sentences

In order to extract interactions from sentences, we used phrase structure parsing which generates parse tree of a sentence that can be analyzed for relationships among words. Stanford parser tool[2] (de Marneffe et al. 2006) is employed for sentence parsing. Figure 1 shows an example of phrase structure parse tree.

We extracted interactions by depth-first search in the parse tree. Each path between keyword nodes (e.g. gene or disease) and the root node were collected. A list of interaction verbs

[1] http://morphix-nlp.berlios.de/manual/node36.html
[2] http://nlp.stanford.edu/software/stanford-dependencies.shtml

were compiled from VerbNet[3], which consists of 1048 verbs. We captured interactions from the paths which contain an interaction verb.

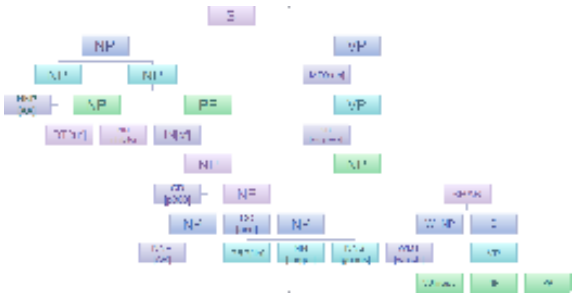

Figure **1**. Part of the phrase structure parse tree of the sentence "AA, an inhibitor of p300, can suppress AR and its target genes, which can induce cells cycle arrest and apoptosis of Lncap cells through AR signaling."

For instance, two genes 'AA' and 'AR' could be extracted from sentence "AA, an inhibitor of p300, can suppress AR and its target genes, which can induce cells cycle arrest and apoptosis of Lncap cells through AR signaling". The path from 'AA' to 'AR' in the syntactic tree is "NP(AA) ->NP ->NP ->S ->VP(can) ->VP(suppress) ->NP ->NP ->NP(AR)", where 'suppress (VP)' is an interaction verb. Therefore, we consider there is a 'suppression' interaction between 'AA' and 'AR'.

4 Interaction network analysis

The extracted interactions can be represented by an adjacency matrix, where $A_{i,j}=1$ if there is an edge between node i and j, and $A_{i,j}=0$ if there is no edge between node i and j. We establish disease-specific interaction network through searching for nodes within 3 distance unit from the target disease node. To gain the most related gene of the target disease, Centrality approach is used for calculating correlation of each gene based on its weight in this specific disease network.

4.1 Degree centrality

Degree centrality represents central tendency of each node in the network, the more direct connects it has, the more power it has in the network and so the more important it is. The degree centrality $C_D(v)$ of node v is calculated as follows.

$$C_D(v)=\sum_{j=1}^{n} A_{ij} \tag{3}$$

4.2 Betweenness centrality

Betweenness centrality reflects the ability of a node taking control of other nodes' communication and the capability of controlling resources in the network. The more nodes that shortest paths pass through, the more communications of other nodes depend on it, and the more betweenness centrality the node has. The betweenness centrality $C_B(v)$ of node v is calculated as follows:

$$C_B(v)=\sum_{s\neq v\neq t\in V}\frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4}$$

where σ_{st} is the total number of shortest paths from node s to t and $\sigma_{st}(v)$ is the number of paths that pass through v.

4.3 Closeness centrality

Closeness centrality reflects the ability a node has of not being controlled by other nodes. The closeness centrality of a node measures how close it is to other nodes in the whole network. The smaller the total distance from a node to other nodes in the network, the less dependency the node has on nodes in the network, and thus the higher its centrality is. The closeness centrality $C_c(v)$ of node v is calculated as follows.

$$C_C(v)=\sum_{t\in V\setminus v} 2^{-d_G(v,t)} \tag{5}$$

where $d_G(v,t)$ represents distance from node v to node t.

4.4 Weighted centrality

Formula (6) is applied to assigne weights for each measure of centrality equally:

$$C_A(v)=\frac{C_D(v)}{3C_D}+\frac{C_B(v)}{3C_B}+\frac{C_C(v)}{3C_C} \tag{6}$$

where C_D represents the largest degree centrality of all nodes in the network, C_B represents the largest betweenness centrality of the whole network and C_C represents the largest closeness centrality among all nodes.

5 Results and Discussion

As a common disease with high incidence, breast cancer gains much attention among researchers and has a rather large literature accumulation.

[3] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

We used breast cancer as testing disease for system evaluation.

The corpus contains 3,209,385 sentences from 44,064 articles. All articles were extracted from PMC with keyword of "breast cancer" (search date: March 1 2013). The gene name dictionary consists of 19,195 gene names searched from OMIM database while the disease dictionary consists of 5644 disease names from Genetic Association database (GAD).

5.1 Evaluation on key sentence extraction

MaxEnt classifier is applied with topic features for key sentences extraction. We randomly selected sentences from 8000 documents in our corpus as training set. We set number of topics K as 10. The results of topics-words distribution predicted by Gibbs Sampling based topic model and topic correction are shown in Table 1.

Topic0	**Topic1**	**Topic2**	**Topic3**	**Topic4**
molecul	use	increase	cancer	cluster
receptor	analysis	rate	organis	compari
body	table	exhibit	gene	melanog
clone	differen	consider	MLL	identical
organis	significa	evolutio	HBB	place
mutator	set	degree	DLC1	share
band	map	due	GRXCR	rDNA
expressi	group	position	XRCC1	parental
replicate	score	distance	GST01	pattern
Topic5	**Topic6**	**Topic7**	**Topic8**	**Topic9**
indicate	observe	control	chromos	growth
test	Demons	express	carry	medium
line	dominan	suppress	male	assay
determi	fact	elegans	female	conditio
experim	reductio	germlin	cross	colony
represen	weak	deficien	homozy	culture
measure	strong	distinct	segregat	syntheti
derive	enhance	close	recover	survival
conversi	still	segment	hybrid	cell

Table 1: The results of topics-words distribution predicted by Gibbs Sampling based topic model and topic correction.

There are totally 1037,637 key sentences were extracted, and the extraction precision is 66.4%.

5.2 Interaction network analysis

5.2.1 Degree centrality

The breast cancer related gene-interaction network consists of 4636 distinct gene nodes and 19,972 interactions extracted among them. Figure 2 illustrates degree centrality of the interaction network of breast cancer. Different color and size indicate different degree centrality of each node. The node in red with the largest degree centrality 1069 in the figure represents breast cancer. This indicates that 1069 genes have direct interactions with breast cancer referred in all sentences.

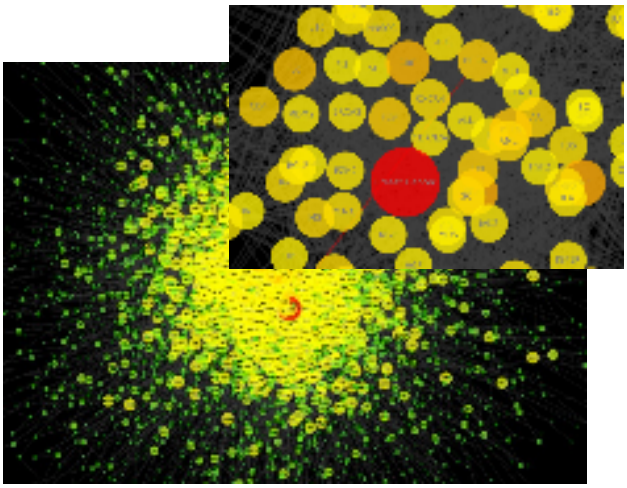

Figure 2. Degree centrality of the gene-breast cancer interaction network.

Figure 3 shows the relationship between each degree centrality and its count of nodes.

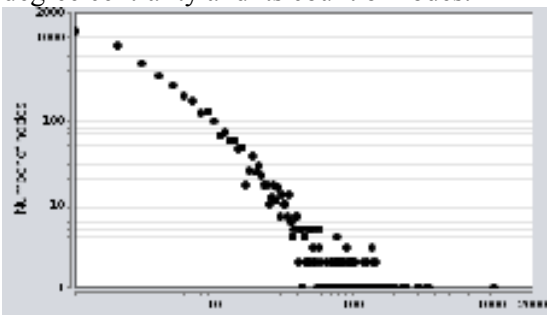

Figure 3. The relationship between each degree centrality and its count of nodes.

As shown in Figure 3, the node with maximum degree centrality 1069 is target disease while most of other nodes distribute from degree centrality of 1 to 10 which are considered as least related genes. Table 2 lists part of ranks of all 1069 genes in the order of degree centrality.

Gene	**Degree Centrality**
TNF	359
EGFR	342
CRC	301
IL-6	245
EGF	200
BRCA1	195
HR	193
GAPDH	190
AR	188
ATM	148
TP53	138
BRCA2	94

Table 2: Part of ranks of all 1069 genes in the order of degree centrality.

From Table 2, we can find that BRCA1 and BRCA2 are known familial breast cancer genes which have gained authority validation. Although their mutations are not common in sporadic breast cancer patients, they accounts for approximately 80% to 90% among all hereditary breast cancer.

TP53 is a kind of mutant gene with high penetrance which has also been verified association with breast cancer in genetics. Moreover, ATM and AR are low frequency genes belong to specific loci, about 5% to 10% of breast cancer relate to at least one or more changes in the susceptibility genes mentioned above.

The result of CRC in contrast is more like some kind of institution's name: Cooperative Research Centre for Discovery of Genes for Common Human Diseases or the abbreviation of another disease: Colorectal Cancer (CRC). There haven't been any evidence reveals direct correlation between CRC gene and breast cancer, we can only consider this as a misrecognition.

In addition to genes described above, other genes in the list have also been verified in authoritative sites or papers. These results preliminarily verified the accuracy of our system.

5.2.2 Betweenness centrality

Figure 4 illustrates betweenness centrality of the interaction network of breast cancer. Color and size of each point reflect betweenness of the node, which indicate the ability to control other nodes in the network. Nodes in green have the minimum betweenness centrality while the color of jade-green shows larger betweenness centrality. Yellow nodes indicate betweenness centrality larger than jade-green and orange represents the largest.

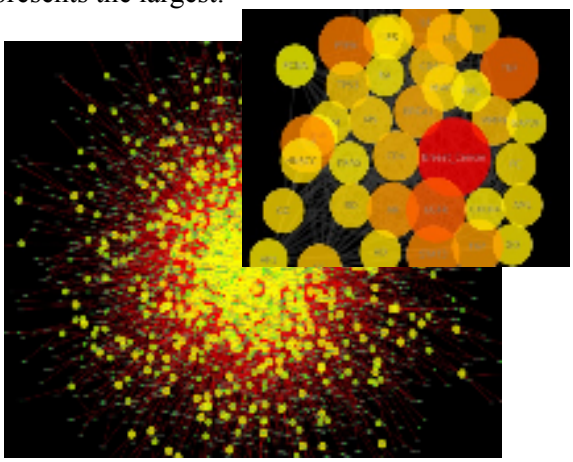

Figure 4. Betweenness centrality of the gene-breast cancer interaction network

Figure 5 shows relationship between each betweenness centrality and its count of neighbors.

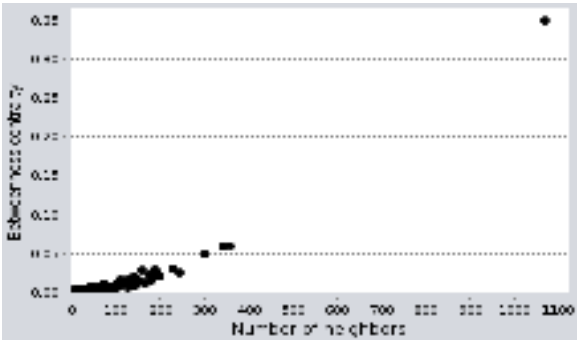

Figure 5. Relationship between each betweenness centrality and its count of neighbors.

As shown in Figure 5, the more adjacent nodes, the larger betweenness centrality. The node with most neighbors of 1068 has maximum betweenness centrality of 0.35 while most nodes in the network have the count of neighbors from 0 to 200 with their betweenness centrality between 0 and 0.04. Table 3 lists part of ranks of all 1069 genes in the order of betweenness centrality.

Gene	Betweenness Centrality
TNF	0.05981684
EGFR	0.05912439
CRC	0.04896846
AR	0.02892632
GAPDH	0.02877095
AD	0.02863766
IL-6	0.02545676
HR	0.02381936
BRCA1	0.02202402
TP53	0.01603455
ATM	0.01566084
BRCA2	0.00507333

Table 3: Part of ranks of all 1069 genes in the order of betweenness centrality.

As can be seen from Table 3, the rank of betweenness centrality is approximately matched with the rank of degree centrality. TNF, EGFR and CRC are still the highest ranked genes while IL-6, AR, HR , GAPDH and ATM simply exchanged their order. AR, androgen receptor, has a quick raise in the rank list. It plays a vital role in the development and maintenance of male reproductive function and the cause of prostate cancer, but the effect and function on breast cancer of AR have not been clear until 2010 (most of the literature published before 2010). This result shows that the genes excavated by our system not only include genes in the known interaction network, but also reflect research

tendency at present or in a certain period of time. This also indicates the effectiveness of understanding scientific research tendency of our system.

As the definition of betweenness centrality, it reflects the ability to affect other nodes in the network. If a gene interacts with another gene through an intermediate gene such as suppression or promotion, then the role played by this intermediate gene is decisive in this association. The more intermediate roles played in associations, the greater the influence of the gene in the network. Similarly, among all genes in the neighborhood of a specific gene, the greater the betweenness centrality of a gene, the more influence it has on that specific gene.

5.2.3 Closeness centrality

Figure 6 illustrates closeness centrality of the interaction network of breast cancer.

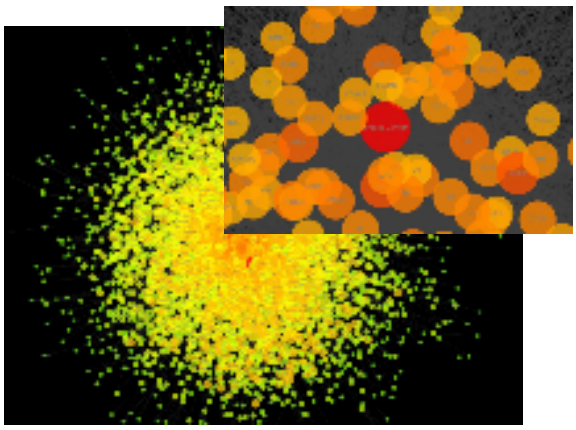

Figure 6. Closeness centrality of the gene-breast cancer interaction network.

As can be seen from Figure 6, red node at the center of the network represents breast cancer and neighboring orange nodes stand for direct related genes while peripheral nodes in green represents least related genes. Figure 7 shows relationship between each closeness centrality and its count of neighbors.

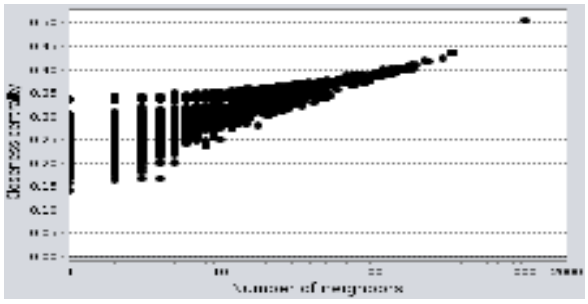

Figure 7. Relationship between each closeness centrality and its count of neighbors.

Figure 7 shows the tendency of closeness centrality in the network while number of neighbors increases. There is an approximate positive correlation between the count of neighbors and the closeness centrality of nodes but not so obvious compared with betweenness centrality or degree centrality. For instance, the closeness centrality ranges from 0.14 to 0.34 for nodes with only one neighbor. This tendency represents that closeness centrality reflect geographical centricity of each node more efficiently compared with degree centrality and betweenness centrality with less dependence on count of neighbors. For example, if a node has only one edge to the center of the network, this node is bound to own large closeness centrality even though this edge is the only edge it has. Meanwhile, another node has much more than one edge but far away from the center of the network, the closeness centrality of it can never be larger than the former one. Table 4 lists part of ranks of all 1069 genes in the order of closeness centrality.

Gene	Closeness Centrality
TNF	0.43612418
EGFR	0.43550963
CRC	0.4247366
PTEN	0.41920608
IL-6	0.41814738
AR	0.41092005
EGF	0.40954064
BRCA1	0.40914306
STAT3	0.4088544
MMP-9	0.40386793
HR	0.40330579
MMP-2	0.40031085

Table 4: Part of ranks of all 1069 genes in the order of closeness centrality.

Table 4 shows that list ordered by closeness centrality is generally similar to list ordered by degree centrality and betweenness centrality. TNF, EGFR and CRC are still highest ranking genes. However, genes like STAT3, MMP-9 and MMP-2 appear firstly in the list where STAT3 ranks 18 in degree centrality and 14 in betweenness centrality. The details of STAT3 has been clearly described in Hsieh FC et al. STAT3 full-called signal transducer and activator of transcription 3, which is often detected in breast cancer tissues and its cell lines. STAT3 has already been defined as an oncogene since its activated form in nude mice can produce malignant transformation of cultured cells and ultimately form tumors. MMP-9 and MMP-2 are gelatinase, proteolytic enzymes involved in

process of tumor invasion which is considered as a potential tumor marker in breast cancer.

All these three genes can be identified as direct related genes with breast cancer. These associations which are not obvious in degree centrality and betweenness centrality indicating the effectiveness of closeness centrality in finding related gene to a specific disease.

5.3 Result Evaluation

We enumerate 31 top genes ranked with weighted centrality considered as related to breast cancer due to our system. Table 5 lists the gene or disease symbol, ID, and full name from OMIM database.

Gene Symbol	Gene ID	Gene Full Name
TNF	*191160	TUMOR NECROSIS FACTOR
EGFR	*131550	EPIDERMAL GROWTH FACTOR RECEPTOR
CRC		COLORECTAL CANCER
PTEN	+601728	PHOSPHATASE AND TENSIN HOMOLOG
IL-6	*147620	INTERLEUKIN 6
AR	*313700	ANDROGEN RECEPTOR
BRCA1	*113705	BREAST CANCER 1 GENE
EGF	*131530	EPIDERMAL GROWTH FACTOR
GAPDH	*138400	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
HR	*602302	HAIRLESS, MOUSE, HOMOLOG OF
AML	#601626	LEUKEMIA, ACUTE MYELOID
CD4	*186940	CD4 ANTIGEN
STAT3	*102582	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 3;
AD	#104300	ALZHEIMER DISEASE
MMP-9	*120361	MATRIX METALLOPROTEINASE 9
MS	#126200	MULTIPLE SCLEROSIS, SUSCEPTIBILITY TO
RD	#111620	RADIN BLOOD GROUP ANTIGEN
MYC	*190080	V-MYC AVIAN MYELOCYTOMATOSIS VIRAL ONCOGENE HOMOLOG
S6	*185520	SURFACE ANTIGEN 6
TP53	*191170	TUMOR PROTEIN p53
ATM	*607585	ATAXIA-TELANGIECTASIA MUTATED GENE
IL-8	*146930	INTERLEUKIN 8
AP1		activator protein-1
MMP-2	*120360	MATRIX METALLOPROTEINASE 2
GC	+139200	GROUP-SPECIFIC COMPONENT
FBS	#227810	FANCONI-BICKEL SYNDROME
ES	#612219	EWING SARCOMA
RA	#180300	RHEUMATOID ARTHRITIS
CXCR4	*162643	CHEMOKINE, CXC MOTIF, RECEPTOR 4
IL-10	*124092	INTERLEUKIN 10
BRCA2	*600185	BRCA2 GENE

Table 5: The gene or disease symbol, ID, and full name from OMIM database.

The Genes and diseases in Table 5 inferred by degree, betweenness, closeness centralities and the relevance are listed in Table 6.

Gene	Degree	Betweenness	Closeness	Relevance
TNF	359	0.05985761	0.43401678	Yes
EGFR	342	0.05904224	0.4332496	Yes
CRC	301	0.04875035	0.4225186	No
PTEN	229	0.03029572	0.41695765	Yes
IL-6	245	0.02541463	0.41613797	Yes
AR	188	0.02883127	0.40890333	Yes
BRCA1	195	0.02190664	0.40704484	Yes
EGF	200	0.01992148	0.40747222	Yes
GAPDH	190	0.02868382	0.39946818	Yes
HR	193	0.02371613	0.40136172	Yes
AML	177	0.02417702	0.39779619	Disease
CD4	179	0.01865428	0.40467501	Yes
STAT3	182	0.01563346	0.40683148	Yes
AD	159	0.02853342	0.39769428	Yes
MMP-9	160	0.01347212	0.40188126	Yes
MS	148	0.01806096	0.39967388	Disease
RD	166	0.0113587	0.3970162	No
MYC	141	0.02132884	0.39052411	Yes
S6	136	0.01504618	0.39912581	Yes
TP53	138	0.01607533	0.39607076	Yes
ATM	148	0.01556309	0.39170662	Yes
IL-8	146	0.00944026	0.40108518	Yes
AP1	141	0.01531257	0.39286317	Yes
MMP-2	138	0.01241541	0.39837468	Yes
GC	131	0.01515181	0.39055686	No
FBS	126	0.0117904	0.39749061	No
ES	128	0.01325333	0.39283003	No
RA	133	0.01256221	0.3894464	Disease
CXCR4	138	0.01019905	0.39039316	Yes
IL-10	128	0.00680617	0.39045862	Yes
BRCA2	94	0.00504479	0.38194046	Yes

Table 6: Genes inferred by degree, betweenness, and closeness centralities and the relevance.

As results listed in Table 6, all 31 top ranked genes and diseases have been checked relevance with breast cancer through NCBI database. Terms marked as 'No' are none-relevant to breast cancer and words marked as 'disease' are related diseases to breast cancer. The accuracy rate is 83.9% for these top 31 genes and diseases and 74.2% for these top 31 genes.

6 Conclusion

Understanding the role of genetics in diseases is one of the major goals of the post-genome era. We have proposed an automatic gene-disease association extraction approach based on text mining and network analysis.

Gene-breast cancer interaction network analysis demonstrated that degree, betweenness, and closeness centralities can estimate disease related genes effectively. And closeness centrality is able to find disease related genes which are not obvious ranked by degree centrality and betweenness centrality. In addition, this result showed that the genes excavated by our system not only include genes in the known interaction network, but also reflect research tendency at present or in a certain period of time. This also indicates the effectiveness of understanding scientific research tendency of our system.

Acknowledgment

This research has been partially supported by the National High-Tech Research & Development Program of China 863 Program under Grant No. 2012AA011103, National Natural Science Foundation of China under Grant No. 61203312, National Program on Key Basic Research Project of China (973 Program) under Grant No. 2014CB347600, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Key Science and Technology Program of Anhui Province under Grant No. 1206c0805039.

References

Adamic, L.A., Wilkinson, D., Huberman, B.A., and Adar, E. 2002. A literature based method for identifying gene-disease connections. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, Stanford, CA, pp. 109–117.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. & Moreau, Y. 2006. Gene prioritization through genomic data fusion. *Nature biotechnology* 24(5):537–544.

Al-Mubaid, H., and Singh, R.K. 2005. A new text mining approach for finding protein-to-disease associations. *Am J Biochem Biotechnol*, 1:145–152.

Bader, G., Betel, D., Hogue, C. 2003. Bind – the biomolecular interaction network database. *Nucleic Acids Research*, 31, pp. 248–250.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Ouellette, B.F.F., Rapp, B.A. and Wheeler, D.L. 2000. GenBank. *Nucleic Acids Research*, 28. pp. 15–18.

Blei, D. and McAuliffe, J. 2007. Supervised topic models. Neural Information Processing System 21.

Blei, D.M., Ng, A., Jordan, M.I. 2002. Latent Dirichlet Allocation. NIPS.

Chen, J.Y., Shen, C., Sivachenko, A.Y. 2006. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac. Symp. Biocomput.*, 11, 367–378.

Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. 2009. Toppgene suite for gene list enrichment analysis and candidate gene prioritization, *Nucleic Acids Research*, 37(Web Server issue): gkp427+.

Christopher D. Manning and Hinrich Schjtze. 1999. Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Chun, H., Tsuruoka, Y., Kim, J. Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. 2006. Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 4–15.

Erkan, G., Radev, D., Ozgur, A. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pp. 228–237.

Freudenberg, J., and Propping, P. 2002. A similarity-based method for genomewide prediction of disease-relevant human genes. *Bioinformatics*, 18 (Suppl. 2), pp. S110–S115.

Gillick, D., Sentence Boundary Detection and the Problem with the U.S. NAACL 2009. pp. 241–244,

Glenisson, P., Coessens, B., Vooren, S. V., Mathys, J., Moreau, Y., and De Moor, B. 2004. TXTGate: profiling gene groups with text-based information. *Genome Biol*., 5, R43.

Griffiths, T., 2002. Gibbs sampling in the generative model of Latent Dirichlet Allocation. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3760.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.*, 4: 44–57.

Hutz, J., Kraja, A., McLeod, H. & Province, M. 2008. Candid: a flexible method for prioritizing candidate genes for complex human traits., *Genetic Epidemiology* 32(8): 779–790.

Kerrien, S., Aranda, B., Breuza L., Bridge, A., Broackes-Carter, F., and Chen, C. 2002. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40, pp. 841–846.

Klein, D. and Christopher D. M. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430.

Krallinger, M., Leitner, F., Valencia, A. 2010. Analysis of biological processes and diseases using text mining approaches. *Methods Mol Biol,* 593: 341–82.

Ma, X., Lee, H., Wang, L. & Sun, F. 2007. Cgi: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, *Bioinformatics* 23(2): 215–221.

Martin, D., Brun. C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. 2004. GOToolbox: functional

analysis of gene datasets based on gene ontology. *Genome Biol.*, 5, R101.

McKusick,V. 2007. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, 80, pp. 588–604.

Morrison, J. L., Breitling, R., Higham, D. J., and Gilbert, D. R. 2005. Generank: using search engine technology for the analysis of microarray experiments., BMC Bioinformatics 6: 233. URL: http://www.biomedsearch.com/nih/GeneRank-using-search-engine-technology/16176585.html

OMIM. 2007. Online Mendelian inheritance in man, OMIM (TM).McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Özgür, A., Vu, T., Erkan, G., and Radev D. R. 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24. pp. 277–285.

Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein-protein interactions in yeast. *Nat.Biotechnol.*, 18, pp. 1257–1261.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad.* Sci. USA, 102, pp. 15545–15550.

Hwang, T., Zhang, W., Xie, M., Liu, J., and Kuang, R. 2011. Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics*, 27(19): 2692–2699.

Temkin, J. and Gilder, M. 2003. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19:2046–2053.

Wuchty, S., Oltvai, Z.N., Barabási, A.L. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.*, 35:176–179.

Yakushiji, A., Miyao, Y., Tateisi, Y., and Tsujii J. 2005. Biomedical information extraction with predicate argument structure patterns. In *Proceedings of the Eleventh Annual Meeting of the Association for Natural Language Processing,* pp. 93–96.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. 2002. Mint: A molecular interaction database. FEBS Letters, 513: 135–140.