Developing an NLP and IR-based Algorithm for Analyzing Gene-disease Relationships

Y. T. Yen[1], B. Chen[2], H. W. Chiu[1], Y. C. Lee[1], Y. C. Li[1], C. Y. Hsu[1]
[1]Graduate Institute of Medical Informatics, Taipei Medical University, Taipei, Taiwan
[2]Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan

Summary

Objectives: High-throughput techniques such as cDNA microarray, oligonucleotide arrays, and serial analysis of gene expression (SAGE) have been developed and used to automatically screen huge amounts of gene expression data. However, researchers usually spend lots of time and money on discovering gene-disease relationships by utilizing these techniques. We prototypically implemented an algorithm that can provide some kind of predicted results for biological researchers before they proceed with experiments, and it is very helpful for them to discover gene-disease relationships more efficiently.
Methods: Due to the fast development of computer technology, many information retrieval techniques have been applied to analyze huge digital biomedical databases available worldwide. Therefore we highly expect that we can apply information retrieval (IR) technique to extract useful information for the relationship of specific diseases and genes from MEDLINE articles. Furthermore, we also applied natural language processing (NLP) methods to do the semantic analysis for the relevant articles to discover the relationships between genes and diseases.
Results: We have extracted gene symbols from our literature collection according to disease MeSH classifications. We have also built an IR-based retrieval system, "Biomedical Literature Retrieval System (BLRS)" and applied the N-gram model to extract the relationship features which can reveal the relationship between genes and diseases. Finally, a relationship network of a specific disease has been built to represent the gene-disease relationships.
Conclusions: A relationship feature is a functional word that can reveal the relationship between one single gene and a disease. By incorporating many modern IR techniques, we found that BLRS is a very powerful information discovery tool for literature searching. A relationship network which contains the information on gene symbol, relationship feature, and disease MeSH term can provide an integrated view to discover gene-disease relationships.

Keywords

Natural language processing, information retrieval, gene, disease, relationship, MeSH

Methods Inf Med 2006; 45: 321–9

1. Introduction

It is a very important and difficult biological research topic to discover the relationships between genes and diseases. Especially when there are many digital biomedical databases available worldwide, we have the opportunity to extract lots of information to discover gene-disease relationships without bench-work experiments. However, a large amount of the knowledge is only presented in free-text format, and is not readily available for automatic computerized analysis. Therefore, intelligent and efficient information retrieval techniques allowing easy access to huge amounts and various types of biomedical information become highly desired. Systematically analyzing the text-format information will help us to understand the association between genes and diseases.

Some bioinformatics researchers have addressed and used literature mining techniques to automatically extract information from biomedical literature for functional genomics [1-4]. Jenssen et al. created a gene-to-gene co-citation network from publicly available gene and text databases via automated analysis for biological literature [5]. In addition, the use of natural language processing (NLP) techniques for automatically extracting knowledge from biomedical literature has also received much of the attention [6].

Much of the work done so far has focused on discovering gene relations or protein-protein interactions [3, 7] from biomedical documents. Chiang and Yu have developed an ontology-based text mining system named MeKE to extract functions of gene products from biomedical literature knowledge [4]. Although they adopted gene ontology (GO) as the lexicon for constructing the function name index, they also used information retrieval (IR) methods such as sentence alignment and classification for extracting molecular functions, biological processes, and cellular components. The same concept can be applied to the mining of gene-disease relationships.

In this research we selected lymphoma and breast neoplasms-related MEDLINE literatures as our initial disease targets, and applied NLP and IR methods to extract the information for gene-disease relationships.

The research procedures are summarized as follows:

1) Literature collection and data preparation.
2) Gene symbol extraction.
3) Term weighting of literature collection.
4) Semantic analysis.
5) Relationship feature identification.
6) Relationship network construction.

At present study, we have built an IR-based retrieval system to gather relevant information and implemented an N-gram-based algorithm which can extract relationship features from gene-related MEDLINE literatures. Finally, a relationship network for a specific disease showing the association between multiple genes and a disease was generated for analyzing gene-disease relationships.

2. Objectives

The purpose of this research is to extract relevant information from literature and build a knowledge-based inference engine for analyzing gene-disease relationships. Several programs have been developed to

1) extract gene symbols from biomedical literature collection;
2) retrieve relevant information related to a specific gene and disease;
3) extract relationship features which can reveal the relationship between gene and disease.

The biological experts helped us on providing existing knowledge and evaluated the training corpus and models. We expect that our results can be very useful to those researchers on finding gene-disease relationships and prioritizing the valuable information from large databases. The resulting system will help biologists discover gene-disease relationships more efficiently, especially for specific narrow domain knowledge.

3. Methods

Biomedical information exists in both the research literature and various structured databases. Accurate and computationally efficient approaches in discovering relationships between biological objects from text documents are important for biologists to develop biological models. In our research we applied text analysis methods to process the biomedical literature and extract information of gene-disease relationship. Several procedures based on IR and NLP have been implemented for prediction and retrieval of gene-disease relationships. The procedures for reaching these goals are represented as Figure 1 and described in detail below.

3.1 Literature Collection and Data Preparation

Intelligent and efficient information retrieval techniques allow easy access to a huge amount and various types of biomedical information. For the purpose of specializing the knowledge scope, we used specific MeSH term classifications for lymphoma and breast neoplasms in this research. The same algorithms are expected to be applicable for other diseases. We collected text information related to lymphoma and breast neoplasms from the MEDLINE citation database(http://www.ncbi.nlm.nih.gov/PubMed) provided by the National Library of Medicine (NLM). Those articles are constrained to be the subset of MeSH tree number (C04.557.386) and (C04.588.180). The following are the tree structures of MeSH terms for lymphoma and breast neoplasms:

Lymphoma [C04.557.386]
- Histiocytosis, Malignant [C04.557.386.345]
- Hodgkin Disease [C04.557.386.355]
- Immunoproliferative Small Intestinal Disease [C04.557.386.390]
- Letterer-Siwe Disease [C04.557.386.435]
- Lymphoma, Non-Hodgkin [C04.557.386.480] +
- Plasmacytoma [C04.557.386.720] +
- Reticuloendotheliosis [C04.557.386.802] +

Breast Neoplasms [C04.588.180]
- Breast Neoplasms, Male [C04.588.180.260]
- Mammary Neoplasms [C04.588.180.520]
- Mammary Neoplasms, Experimental [C04.588.180.525]
- Phyllodes Tumor [C04.588.180.762]

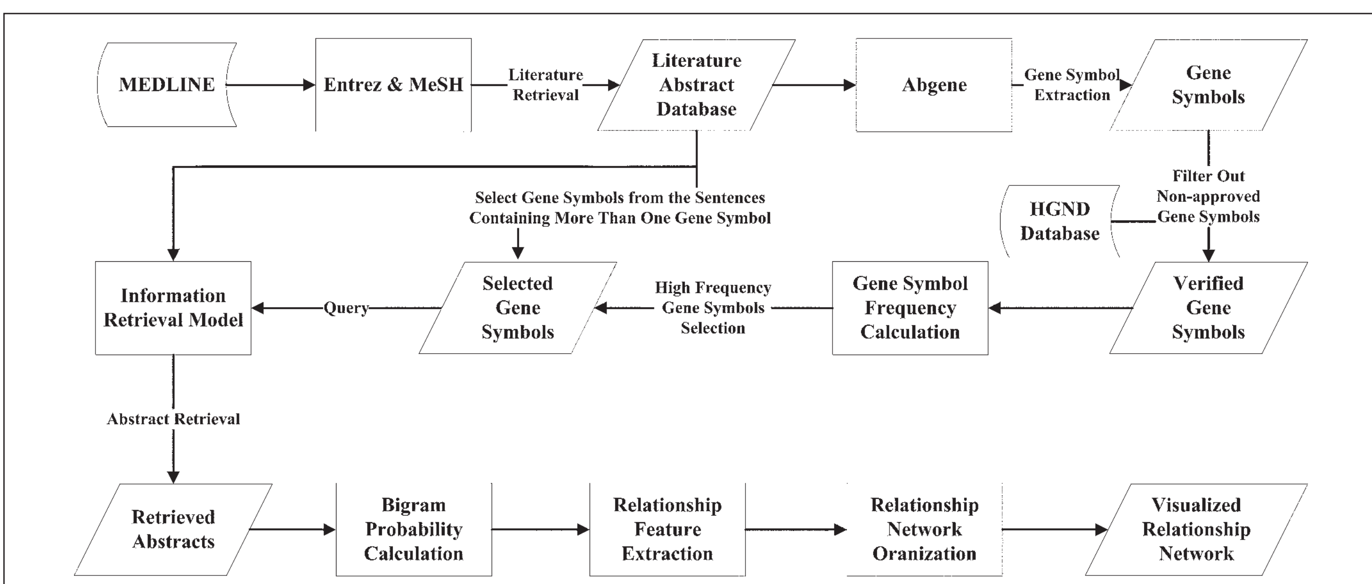

Fig. 1 The overall procedure for prediction and retrieval of gene-disease relationship. 1) Retrieve MEDLINE abstracts related to a specific disease by Entrez and MeSH to build a literature abstract database. 2) Using Abgene to extract gene symbols from abstract database. 3) Calculate frequency of gene symbol and select the gene symbol with high frequency as the query of Information Retrieval model. 4) Select the gene symbols from the sentences that contain more than one gene symbol. 5) Retrieve abstracts related to specific gene symbols from database by using Information Retrieval model. 6) Calculate bigram probability and extract relationship features from retrieved abstracts. 7) Organize relationship network by using selected gene symbols and their relationship features.

3.2 Gene Symbol Extraction

We used "Abgene" [8], which is an extraction tool developed by the National Center for Biotechnology Information (NCBI) to extract gene symbols from collected articles. It is designed for extracting gene symbols and protein names from MEDLINE articles. Furthermore it is based on the transformation-based part of speech (POS) tagger (Brill tagger) and is trained by 7000 sentences from biomedical articles. After adding biomedical lexicons to Brill package, gene symbols and protein names in the biomedical articles can be identified. The following is an example output of Abgene. Obviously "Aw24", "Aw33", and "Bw44" are identified by the "GENE" label.

> In/IN somewhat/RB weaker/JJR association/NN ,/, Aw24/**GENE** and/CC Aw33/**GENE** are/VBP elevated/VBNin/IN follicular-center-cell/ JJ lymphomas/NNS ,/, while/CC Bw44/**GENE** is/VBZ depressed/VBN ./.

Although Abgene can extract most gene symbols, gene symbols associated with multiple names is still a critical issue to be tackled. More names are added as new functional or structural information is discovered. For example, the gene symbol, TNFSF5, has four gene names, tumor necrosis factor (ligand) superfamily, member 5 (hyper-IgM syndrome), T-B cell-activating molecule, TNF-related activation protein and CD40 antigen ligand and ten aliases, CD40L, TRAP, gp39, hCD40L, IGM, IMD3, HICM1, T-BAM, CD40LG and CD154. To confirm the gene symbols from Abgene's output, the Human Gene Nomenclature Database (HGND) was referred to filter out non-approved gene symbols.

3.3 Term Weighting of Literature Collection

In order to decide which article is most likely related to specific gene symbols extracted by Abgene, we applied IR methods such as the Vector Space Model [9] to calculate a term weighting for each gene symbol and build a ranking algorithm for sorting the articles. The result is a procedure that can easily select articles related to specific gene symbols from large literature collections.

We have calculated a lexicon for the collected abstracts by eliminating the selected stop words. The lexicon contains 4,696,163 entries for lymphoma and 353,251 entries for breast neoplasms. The lexicon was used as the index terms for establishing the Vector Space Model. The model is described as follows:

N: the number of total document.

$W_{i,j}$: the weight of index term k_i in document d_j.

$W_{i,q}$: the weight of index term k_i in query q.

$\vec{d_j}$: the document vector ($W_{1,j}$, $W_{2,j}$, ..., $W_{t,j}$).

The vector $W_{i,j}$ is influenced by two factors:

1) *tf* (term frequency): the frequency of index term k_i in document d_j (intra-document).
2) *idf* (inverse document frequency): the frequency of the document which contains index term k_i (inter-documents).

Thus, $W_{i,j} = tf_{i,j} \times idf_i$ (1)

If n_i is the number of documents in which the index term k_i appears and $freq_{i,j}$ is the raw frequency of k_i in document d_j, the normalized frequency of k_i document d_j is

$$f_{i,j} = \frac{freq_{i,j}}{\max freq_{l,j}} \quad (2)$$

If idf_i is the inverse document frequency of k_i, we have

$$idf_i = \log \frac{N}{n_i} \quad (3)$$

The term-weighting of k_i is

$$W_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \quad (4)$$

In our research, each component $w_{i,j}$ of the feature vector $\vec{d_j}$ for a document D is associated with the weighted statistics of a specific index term:

$$w_{i,j} = (1 + \ln(f_{i,j}) \cdot \ln((N+1)/n_i) \quad (5)$$

A query Q is also represented by a set of feature vectors $\vec{q_j}$ constructed in the same way. The cosine measure is used to estimate the query-document relevance for each type of index terms:

$$R_j(\vec{q_j}, \vec{d_j}) = (\vec{q_j}, \vec{d_j}) / (\|\vec{q_j}\| \cdot \|\vec{d_j}\|) \quad (6)$$

The overall relevance is then the weighted sum of the relevance scores of all types of index terms:

$$R(Q, D) = \sum_j W_j \cdot R_j(\vec{q_j}, \vec{d_j}) \quad (7)$$

We used this model to calculate the weighted ranking scores of genes and selected the most likely articles related to specific genes for lymphoma and breast neoplasms.

3.4 Semantic Analysis

The N-gram model was used for contextual analysis to reveal the association between genes and diseases. This model is a word prediction model that given a word sequence (W_1, W_2, …, W_n) and guess the next word by calculating the following probability:

$$P(W_1, W_2, ..., W_{n-1}, W_n) \quad (8)$$

Using chain rule of probability to decompose the probability:

$$P(W_1^n) = P(W_1)P(W_2|W_1)P(W_3|W_1^2)\ldots P(W_n|W_1^{n-1}) = \prod_{k=1}^{n} P(W_k|W_1^{k-1}) \quad (9)$$

It is difficult to compute the probability of a word given a long sequence of preceding words. To solve this problem, we used the Unigram and Bigram models to estimate the N-gram model.

For the Unigram model, we calculated unigram probability of each word (w_i):

$$P(W_i) = \frac{N(W_i)}{\sum_{j=1}^{|V|} N(W_j)} \quad (10)$$

$\forall W_i \in V, i \in [1, ..., |V]$

where V is the set of words we collected from the sample articles.

For the Bigram model, we calculated the bigram probability of each word:

$$P(W_n|W_{n-1}) \quad (11)$$

In order to estimate the N-gram model, we combined the bigram sequence:

$$P(W_1^n) = \prod_{k=1}^{n} P(W_k|W_{k-1}) \quad (12)$$

Table 1 Gene symbol counts for MeSH terms, "Hodgkin Disease", "Lymphoma, B-Cell", and "Lymphoma, T-Cell" under disease classification of lymphoma (C04.5557.386). We classified lymphoma-related MEDLINE articles into three MeSH term classifications and used Abgene to extract gene symbols from these MEDLINE articles. The number of articles classified by each MeSH term was listed in the third column.

MeSH term under lymphoma	Gene symbol counts	Article counts
Hodgkin Disease	1604	8618
Lymphoma, Non-Hodgkin		
Lymphoma, B-Cell	2494	4377
Lymphoma, T-Cell	1607	3065

Table 2 Selected gene symbols for locating the most relevant articles. In three MeSH term classifications of lymphoma disease, the first ten gene symbols with the highest occurrence frequency and the gene symbols from the sentences containing more than one gene symbol were selected as query strings to retrieve relevant articles by the retrieval system.

MeSH term under lymphoma	Selected gene symbols
Hodgkin Disease	P53, IL-2, CD4, IL-6, HD, TNF, GM-CSF, ALK, CD43, IL-10
Lymphoma, Non-Hodgkin	
Lymphoma, B-Cell	CD5, CD10, P53, CD19, IL-2, BCL2, CD43, BCL6, IL-6, C-MYC
Lymphoma, T-Cell	CD4, CD56, IL-2, CD5, CD2, P53, CD7, CD43, IL-4, IL-6

Then we calculated bigram probabilities from the relevant articles to extract candidates of relationship features that were related to a specific gene and disease.

A bigram probability [10] is composed of forward and backward probability which is defined as follows:

Bigram probability: $P(W_i, W_j) =$

$$\sqrt{P_f(W_j\,|\,W_i) \times P_b(W_i\,|\,W_j)} \quad (13)$$

Forward probability: $P_j(W_j|W_i) =$

$$\frac{P(W_{t+1} = W_j, W_t = W_i)}{P(W_t = W_i)} \quad (14)$$

Backward probability: $P_b(W_i|W_j) =$

$$\frac{P(W_{t+1} = W_j, W_t = W_i)}{P(W_{t+1} = W_j)} \quad (15)$$

As the result of calculation, the term with high bigram probability related to a gene symbol was selected as a candidate of relationship feature.

3.5 Relationship Feature Definition and Identification

A relationship feature is a functional word or a word with semantic property that is adjacent to a gene symbol with high bigram probability and is confirmed by biology experts. It can be used to reveal the relationship between gene and disease.

According to the calculated bigram probabilities which were related to a specific gene symbol, the average bigram probability was used as a cut-off threshold to filter out the lower bigram probability terms adjacent to a gene symbol in documents. We selected the terms whose bigram probabilities were higher than the average bigram probability as the candidates of relationship feature.

In order to identify correct terms which can properly represent relationships between genes and diseases, the candidates were reviewed by biology experts to extract the most relevant terms as the final relationship feature collection. The final relationship feature collection was used to build a so called "relationship network" which represented the relationship between gene and disease.

3.6 Relationship Network Construction

The components of a relationship network for a disease includes: sub-networks determined by MeSH term classifications, gene symbols, relationship features, and connections between gene symbols.

The co-occurrence gene symbol between the disease MeSH term classifications is used to combine the sub-networks. And the connection between two gene symbols is established while these gene symbols have the same relationship feature. In the present study, we built this relationship network manually and an automatic procedure will be developed in the future.

4. Results

4.1 MEDLINE Article Collection

We have worked on several topics related to the exploration of the relationship between genes and diseases. Currently we focused on lymphoma and breast neoplasms; the MeSH tree number leads off with C04.557.386 and C04.588.180. We used Entrez System to extract articles from the PubMed database and the following is the PubMed query language for lymphoma:

((((((((((“Histiocytosis, Malignant” [MeSH Terms] OR “Hodgkin Disease”[MeSH]) OR “Immunoproliferative Small Intestinal Disease”[MeSH]) OR “Letterer-Siwe Disease”[MeSH]) OR “Lymphoma, Non-Hodgkin”[MeSH]) OR “Multiple Myeloma”[MeSH]) OR “Mast-Cell Sarcoma”[MeSH]) AND (“1960/1/1”[PDat] : “2003/7/1”[PDat])) AND hasabstract[text]) AND English[Lang]) AND (“hominidae”[MeSH Terms] OR “Human”[MeSH Terms]))

According to the query, those articles retrieved are constrained by publication date (1960/1/1 to 2003/7/1), are companied with abstracts, and the language must be English. From this query, we collected 37,810 articles for lymphoma and 2782 articles for breast neoplasms.

4.2 Gene Symbol Extraction

Because there are differences among subtypes of a disease, we used disease MeSH term classifications to discover gene-disease relationship in a histological level. For example, there are two major subtypes of lymphoma disease, "Hodgkin" and "Non-Hodgkin". Both are malignant growth of cells in a lymph system; however Non-Hodgkin's lymphoma is focus on B or T cells. We classified the MEDLINE article collections according to the MeSH term classifications, "Hodgkin Disease", "Lymphoma, B-Cell", and "Lymphoma, T-Cell" and extracted gene symbols from MEDLINE articles for each classification (Table 1). Then we selected the first ten gene symbols with the highest occurrence frequency from each MeSH term classification. These gene symbols were used to locate the most relevant articles through our retrieval system (Table 2).

4.3 Retrieval System

For the purpose of retrieving relevant articles, we implemented a Vector Space Model-based retrieval system, Biomedical Literature Retrieval System (BLRS), and applied the Porter stemming algorithm [9] to deal with morphological problems.

The Porter stemming algorithm is a process for removing the commoner morphological endings from words in English. It is based on the idea that suffixes in English are mostly made up of a combination of smaller and simpler suffixes. For example, the term, "connect" has the following morphological forms: "connected", "connecting", "connection", and "connections". This algorithm has five steps applying rules within each step. Within each step, if a word is matched a suffix rule, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule.

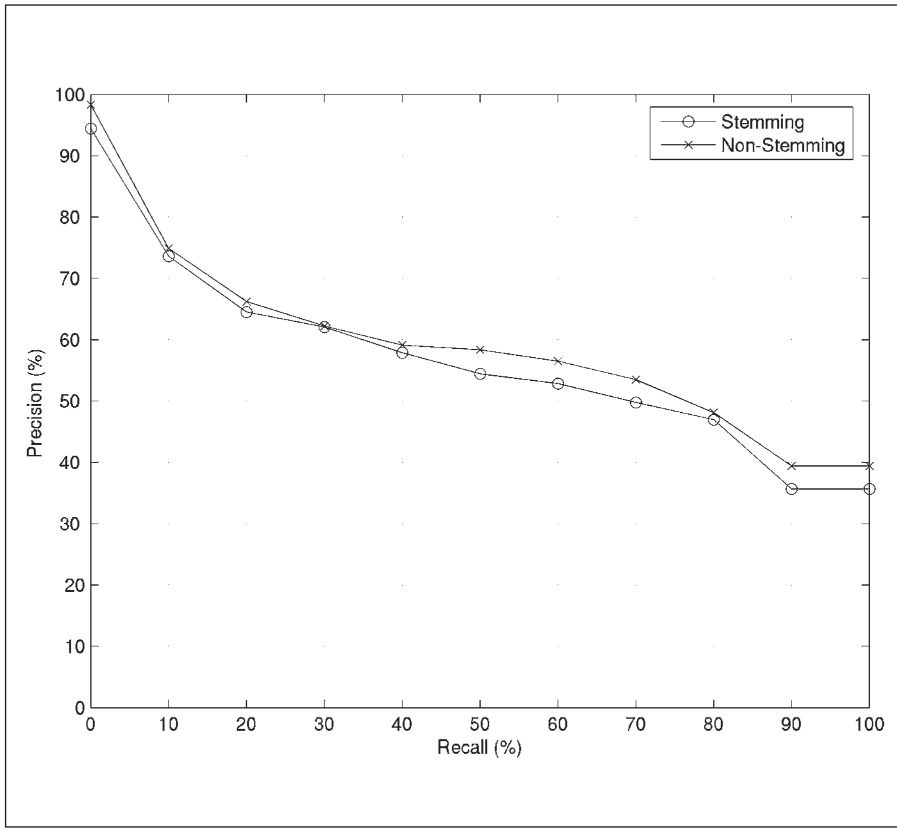

Fig. 2 The BLRS retrieval system has been evaluated by recall-precision curve. The result shows the system has high precision at low recall rate. The comparison between stemming and non-stemming algorithm is plotted. The precision of the system using stemming is a little lower than that of not applying.

Table 3 Comparison of the retrieval system with and without applying the stemming algorithm. The mean average precision decreased about 4.2% for using stemming but vocabulary was reduced about 20% when applying the stemming algorithm.

Evaluation factors	Non-stemming	Stemming	Reduction
Mean average precision	0.5667	0.5427	–4.2%
Vocabulary	2,948,668	2,363,979	–20%

Table 4 A statistics of bigram and relationship feature for lymphoma. From the MEDLINE articles of three MeSH term classifications retrieved by BLRS, the bigram probabilities related to the selected gene symbold from Table 4 were calculated. The number of bigram in each MeSH term classification is listed in the second column. The third column is the number of the candidates of relationship feature whose bigram probabilities are higher than the average bigram probability. The candidates were reviewed by biology experts to extract the most relevant terms as the final relationship feature collection listed in the fourth column.

MeSH term	Bigram	Candidate	Relationship feature
Hodgkin Disease	978	301	29
Lymphoma, Non-Hodgkin			
Lymphoma, B-Cell	1702	454	33
Lymphoma, T-Cell	1004	301	10

Table 5 A detail list of the final relationship feature collection for the gene symbols from Table 4. The * symbol is represented if the relationship feature in the first column is belonging to the gene symbol. The gene symbols which have no relationship feature are omitted.

Relationship feature	Hodgkin Disease										Lymphoma, Non-Hodgkin													
											B-Cell									T-Cell				
	P53	IL-2	CD4	IL-6	HD	TNF	GM-CSF	ALK	CD43	IL-10	CD5	P53	CD19	IL-2	BCL2	CD43	BCL6	IL-6	C-MYC	CD4	CD56	IL-2	CD2	P53
Accelerate							*																	
Accumulate	*																							
Activation			*												*									
Adsorb		*																						
Alteration	*				*							*					*							
Amplification																			*					
Bind		*																						
Blockage	*																							
Cocrosslink													*											
Co-expression	*		*								*					*				*	*			
counter-receptor																							*	
differentiation								*																
diminish		*																						
disruption		*																						
down-regulate																		*	*					
dysfunction												*												
enrichment			*																					
induce																			*					
infusion		*																						
inhibition						*				*														
lesion					*																			
leydig																					*			
modulation													*											
mutate	*											*					*		*	*				*
oncogene																			*					
oncosuppressor	*																							
overexpression	*														*	*			*					
overlap		*																						
phosphorylation															*									
promotor				*									*											
proto-oncogene																	*		*					
rearrangement															*		*		*					
receptor																						*		
receptor-positive		*																						
reduction		*													*									
regulate																	*							
reinstitution		*																						
repress																						*		

Table 5 Continued

Relationship feature	Hodgkin Disease										Lymphoma, Non-Hodgkin													
											B-Cell									T-Cell				
	P53	IL-2	CD4	IL-6	HD	TNF	GM-CSF	ALK	CD43	IL-10	CD5	P53	CD19	IL-2	BCL2	CD43	BCL6	IL-6	C-MYC	CD4	CD56	IL-2	CD2	P53
reverse				*																				
trans-activate													*											
transduction						*																		
transmembrane																								
trigger														*									*	
tumoral																				*				
up-regulate								*					*	*	*									

The retrieval system has been evaluated by precision and recall calculations from 967 MEDLINE articles and the answer set was confirmed by medical experts. According to the precision at 11 standard recall levels, the recall-precision curve (Fig. 2) demonstrates the system has high precision at low recall. In other words, it can find more relevant information by retrieving lesser documents.

We also compared two conditions, the systems with and without applying stemming algorithm (Table 3). Although the result showed that the stemming algorithm reduced the mean average precision by 4.2%, 20% of vocabulary has been reduced. Therefore we could compensate a little precision rate to increase the efficiency of the system by reducing the amount of vocabulary.

4.4 Relationship Feature

From the MEDLINE articles retrieved by the selected gene symbols, we calculated bigram probability related to these gene symbols and filter out the terms whose bigram probabilities were lower than the average bigram probability in each MeSH term classification. The remaining terms were selected as candidates of relationship features. Then these candidates were reviewed to extract the most relevant terms as the final relationship feature collection by biological experts. A statistics of bigram and relationship feature for lymphoma is shown in Table 4 and the detail list of final relationship feature collection is shown in Table 5.

4.5 Relationship Network

A relationship network of a disease is composed of several sub-networks determined by disease MeSH term classifications. The relationship network for lymphoma disease shown in Figure 3 is composed of three sub-networks, “Hodgkin Disease”, “Lymphoma, B-Cell” and “Lymphoma, T-Cell” represented as three circles. The gene symbols which have the highest occurrence frequency from each MeSH term classification are represented as different color hexagons. The gene symbols in the intersection of three circles are the co-occurrence gene symbols in these three MeSH term classifications. The connection between two gene symbols means that these gene symbols have the same relationship feature in the literature collection. There are three kinds of relationship features represented as rounded rectangles with red, blue, and without border. The red one only precedes a gene symbol and the blue one only follows a gene symbol in the literature collection. The rounded rectangle without border can either precede or follow a gene symbol in the literature collection.

5. Discussion

Applying knowledge discovered from genomic research to clinical application is crucial and imperiously needed to improve healthcare. Although there are many important pieces of information piling up in the literature databases, such as MEDLINE, how to extract direct and related information linking gene-to-disease relationship is our initiative goal.

MEDLINE, one of the world’s premier biomedical research databases dated from 1966, has provided comprehensive and powerful search ability. However, the search outcome is normally too comprehensive to specify what a user really needs. It is all because the database itself and search engine are not designed for subject-specific purposes. In other words, it is difficult to input a simple word, as simple as a disease name such as “lymphoma” and to get a simple set of related genes, instead of many records of any subject with a title containing “lymphoma”. Therefore it is important to establish a subject-oriented database with simple search interface and object-directed output. Recently, research using gene-expression profiling by DNA microarray to reveal the association between genes and diseases also generates formidable amounts of information [11]. How to integrate the information and databases to establish a system for clinical use is an important issue.

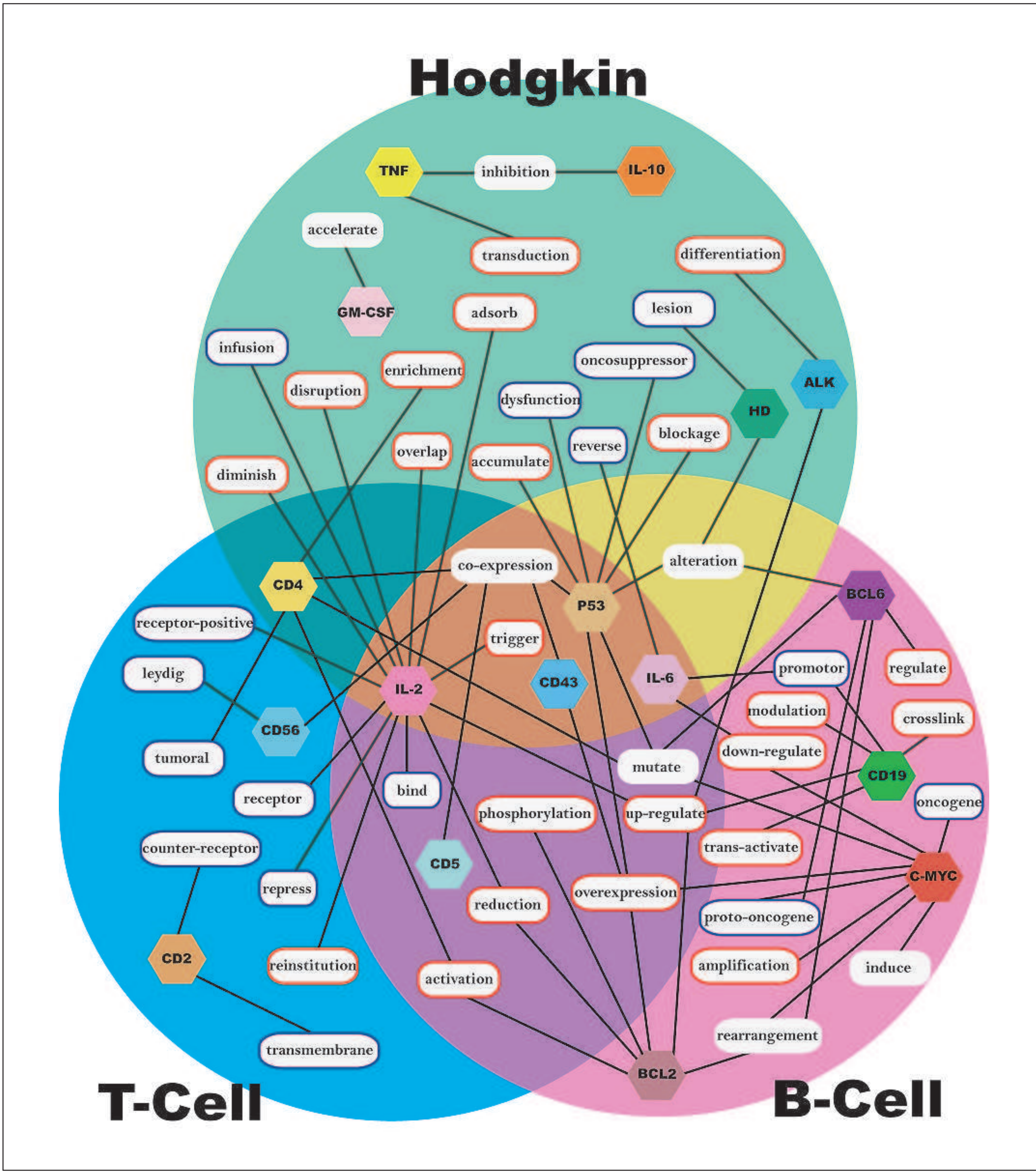

Fig. 3 The relationship network for lymphoma disease is composed of three sub-networks, "Hodgkin Disease", "Lymphoma, B-Cell" and "Lymphoma, T-Cell" represented as three circles. The gene symbols which have the highest occurrence frequency from each MeSH term classification are represented as different color hexagons. The gene symbols in the intersection of three circles are the co-occurrence gene symbols in these three MeSH term classifications. The connection between two gene symbols means that these gene symbols have the same relationship feature in the literature collection. There are three kinds of relationship features represented as rounded rectangles with red, blue, and without border. The red one only precedes a gene symbol and the blue one only follows a gene symbol in the literature collection. The rounded rectangle without border can either precede or follow a gene symbol in the literature collection.

In addition, gene libraries, such as LocusLink, provide a wealth of information on the gene expression by a particular tissue or in response to a specific disease. Frequently, many of the genes are known in the literature, but until now, no practical method for exploring the possible interconnections between the genes existed.

Recently text mining techniques were applied to extract information from biomedical literature for functional genomics. In most research, MEDLINE abstracts are used as the text corpus for extraction of annotation, protein-protein interaction, and functional gene relationship. We applied MeSH terms to locate specific topics such as diseases and used gene symbols extracted by Abgene to retrieve related MEDLINE abstracts according to the ranked query results generated by an IR model. It is efficient to analyze the abstracts with high similarity instead of all abstract collection.

There is much research focused on how to represent gene-gene or gene-disease relationship by analyzing sentences or context in literature using NLP methods. In this research, we used relationship features to link genes and disease. In the relationship network, every relationship feature represents that the specific gene symbol has possible feature property in the disease. The co-occurrence relationship feature connects to two or more gene symbols representing that the possible gene symbols have the same feature property in the disease. Thus the relationship network can be used to reveal both gene-gene and gene-disease relationships.

Although NLP is considered as a method to raise the potential of text mining from biomedical literature, the lack of extensive annotated corpus for literature causes a major obstacle to apply these techniques. Kim et al. have previously made some efforts in developing a semantically annotated corpus for bio-text mining named GENIA [12]. Currently, GENIA corpus Version 3.0 consisting of 2000 MEDLINE abstracts has released more than 400,000 words and approximately 100,000 annotations for biological terms. The already released information is also helpful in terms of extracting information from biomedical literature when applying NLP or IR techniques.

Another important yet unsolved issue is the complexity of gene names. Genes and proteins are often associated with multiple names and more new names are added continuously as new functional or structural information are being discovered. For this issue, Yu and Agichtein [13] have developed approaches which could be used to improve search, extraction, and analysis for biomedical literature to extract synonymous gene and protein terms from biomedical literature. A clear and systematic nomenclature for gene name will further incorporate into this research in the near future.

6. Conclusion

Presently, information in digital biomedical databases is growing tremendously fast. Most of the data is presented in an unstructured manner such as text-style format. Therefore, it is very helpful if we can develop an efficient information retrieval and analysis tool to facilitate knowledge discovery in biomedical research. Specifically, we have extracted gene symbols from a collection of biomedical literature and statistically mapped it to MeSH classification. We have also applied IR methods to build a ranking algorithm and a retrieval system for this literature. We found it is very helpful to filter out non-relevant information and improve the system performance. A relationship network generated by NLP methods is a big map to represent the relationships between genes and disease. Moreover, the methods we developed in this study can be applied on different diseases. It will be a good experiment that we collect articles for variety of biomedical topics and evaluate the system performance between them in the future. We expect that the system can support biologists to design their experiments and discover gene-disease relationship more efficiently.

Acknowledgment

This work is supported in part by National Science Council (Code 93-2914-I-038–009-A1). The author would like to thank Dr. I. J. Chiang for his useful comments on this research.

References

1. Andrade MA, Valencia A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics 1998; 14: 600-7.
2. Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. Bioinformatics 2001; 17: 359-63.
3. Ono T, Hishigaki H, Tanigami A, Takagi T. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2001; 17: 155-61.
4. Chiang JH, Yu HC. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics 2003; 19: 1417-22.
5. Jenssen TK, Lægreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics 2001; 28: 21-8.
6. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003; 19: 1699-1706.
7. Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. 1999; 60-7.
8. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2000; 18: 1124-32.
9. Baeza RY, Ribeiro BN. Modern information retrieval. Addison Wesley Longman, 1999.
10. Chen B, Kuo JW, Tsai WH. Lightly supervised and data-driven approaches to Mandarin broadcast news transcription. The 29th IEEE Int Conf Acoustics, Speech, Signal processing (ICASSP 2004).
11. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997; 282: 699-705.
12. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 2003; 19: 180-2.
13. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. Bioinformatics 2003; 19: 340-9.

Correspondence to:
C. Y. Hsu
Graduate Institute of Medical Informatics
Taipei Medical University
Taipei
Taiwan
E-mail: cyhsu@tmu.edu.tw