Text Mining for Finding Functional Community of Related Genes Using TCM Knowledge

Zhaohui Wu[1], Xuezhong Zhou[1], Baoyan Liu[2], and Junli Chen[1]

[1] College of Computer Science, Zhejiang University, Hangzhou, 310027, P.R.China
{wzh,zxz,chenjl}@cs.zju.edu.cn
[2] China Academy of Traditional Chinese Medicine, Beijing 100700, P.R.China
liuby@mail.cintcm.ac.cn

Abstract. We present a novel text mining approach to uncover the functional gene relationships, maybe, temporal and spatial functional modular interaction networks, from MEDLINE in large scale. Other than the regular approaches, which only consider the reductionistic molecular biological knowledge in MEDLINE, we use TCM knowledge(e.g. Symptom Complex) and the 50,000 TCM bibliographic records to automatically congregate the related genes. A simple but efficient bootstrapping technique is used to extract the clinical disease names from TCM literature, and term co-occurrence is used to identify the disease-gene relationships in MEDLINE abstracts and titles. The underlying hypothesis is that the relevant genes of the same Symptom Complex will have some biological interactions. It is also a probing research to study the connection of TCM with modern biomedical and post-genomics studies by text mining. The preliminary results show that Symptom Complex gives a novel top-down view of functional genomics research, and it is a promising research field while connecting TCM with modern life science using text mining.

Keywords: Text Ming, Traditional Chinese Medicine, Symptom Complex, Gene Functional Relationships

1 Introduction

The last decade has been marked by unprecedented growth in both the production of biomedical data and the amount of published literature discussing it. In fact, it is an opportunity, but also a pressing need as the volume of scientific literature and data are increasing immensely. Functional genomics and proteomics have been the foci in the post-genomic life science research. However, the reductionism and bottom-up approach are still the infrastructure of life science research since no holistic knowledge is reached. Traditional Chinese Medicine (TCM) is an efficient traditional medical therapy (e.g. acupuncture and Chinese Medical Formula), which embodies holistic knowledge with thousands of years' clinical practice. Symptom Complex (SC) is one of the core issues studied in TCM, which is a holistic clinical disease concept reflecting the dynamic, functional, temporal and spatial morbid status of human body. Moreover, several bibliographic databases have been curated in TCM institutes and colleges since 90s. One main database is the TCM bibliographic database[1] built by Information

[1] http://www.cintcm.com/index.htm

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 459–470, 2004.

Institute of China Academy of TCM, which contains about one half million records from 900 biomedical journals published in China since 1984, and 50% of the records have abstracts. These large amounts of literature storages with high quality will be a good text data sources for text mining.

How to connect TCM with modern life science and using the holistic knowledge of TCM to big biology research is an excited open question, which worthy of large research efforts. The experimental approaches to this objective are even extremely difficult since most TCM concepts are qualitative and systematic complicate. However, automated literature mining offers a yet untapped opportunity to induce and integrate many fragments of information gathered by researchers from multiple fields of expertise, into a complete picture exposing the interrelated relationships of various genes, proteins and chemical reactions in cells, and pathological, mental and intellective states in organisms. Many researches [1,2,3,4,5,6,7,8,9] have focused on the gene or protein name extraction, protein-protein interaction and gene-disease relationship extraction from biomedical literature (e.g. MEDLINE). This paper aims to provide a probing text mining approach to identify the gene functional relationships from MEDLINE using TCM knowledge, which is discovered in TCM literature. A simple but efficient bootstrapping method is used to facilitate the extraction of SC-disease relationships from TCM literature. We obtain the gene nomenclature information from the HUGO Nomenclature Committee[2], which has 17,888 approved gene symbols. The term co-occurrence is used to identify the relationships between disease and gene from MEDLINE. Then we get the SC-gene relationships by one-step inference, that is, to compute the genes and SCs with the same relevant disease. After that, the related gene networks of SCs are computed according to some graph algorithms such as[3]. As the ambiguities and polysemy of terminology using in literature, noise and false positive examples will surely be existed in the simple term match method. Currently, we only use co-occurrence frequency threshold to filter the low frequent relationships, and no other methods be considered for these problems.

The rest of article is structured as follows. To introduce the importance of connecting TCM with modern biomedical research, we give an overview of TCM and biomedical research, and some discussions are also proposed on why and how to do synergistic research of these two fields in Section 2. We review the related research work on biomedical literature mining, which have inspired the work of this paper, in Section 3. Bootstrapping technique and term co-occurrence are two approaches used in text mining process. We introduce them in Section 4. The preliminary text mining results to demonstrate the effective text mining process are proposed in Section 5. Finally, in Section 6, we take a conclusion and give the future work.

2 Modern Biomedical Research and Traditional Chinese Medicine

In the last century, modern biomedical research follows the reductionism and qualitative experimental approach, which is called molecular biology to anatomize the whole

[2] http://www.gene.ucl.ac.uk/nomenclature/

[3] http://www.research.att.com/sw/tools/graphviz/

human body to partial organs, tissues, cells and components. Great achievements have been acquired. Furthermore, the working draft of Human Genome sequence has been the crest of molecular biology. However, no other than this huge sequence data prompts the biologists to grasp the life in a reversed approach, which is called holism. System Biology [10] is such a research activity, which is an academic field that seeks to integrate biological data as an attempt to understand how biological systems function. By studying the relationships and interactions between various parts of a biological system (e.g. organelles, cells, physiological systems, organisms etc.) it is hoped that an understandable model of the whole system can be developed. System biology is a concept that has pervaded all fields of science and penetrated into popular thinking. As every biologist knows, there is still a long way to go before understanding biological systems in systematic approaches.

On the other hand, TCM studies the morbid state of human body by clinical practice and holistic quantitative cognitive process. TCM embodies rich dialectical thought, such as that of the holistic connections and the unity of Yin and Yang (two special concepts of TCM, which reflect the two essential states of human body and general material). Other than the disease concept in orthodox medicine, SC reflects the comprehensive, dynamic and functional disease status of live human body, which is the TCM diagnosis result of symptoms (TCM has developed a systematic approach to acquire the symptoms). In clinical practice, there will be several SCs on one specific disease while in different morbid stage. Also, one SC will occur in several different diseases. For example, *kidney YangXu* SC is a basic SC in TCM studies, which will refer to tens to hundreds diseases. Meanwhile, it has been proved that diabetes has several SCs such as *Kidney YangXu*, *YingYangLiangXu* and *QiYingLiangXu* SCs etc.. The therapies (e.g. Acupuncture and Chinese Medical Formula) based on SC are popular and effective in Chinese medical practice. Moreover, the characteristics of SC have much in common with that of genome and proteome such as polymorphism, individuality and dynamics etc.. We believe that SC will provide much more functional and holistic knowledge about the human body, which will largely support the functional genomics and proteomics research since TCM has the thousands of years experience to study the SC status of human body.

There are some studies such as that of Prof. Shen [11] on connecting TCM with molecule biology by experimental approaches. Prof. Shen has spent his fifty years to study the *kidney YangXu* SC at the molecular level. He found that *kidney YangXu* SC is associated with the expression of CRF (C1q-related factor). This paper will have some comparative study with the results of Prof Shen, and produce the novel functional gene relationships plus several novel relevant genes for *kidney YangXu* SC. It is very difficult and a long way to synergize the researches of these two fields in experimental approaches, because still no good approaches to model the dynamic and temporal qualitative organism concepts at molecular level. However, this paper suggests and proposes a synergistic approach of TCM and modern biomedical research using text mining techniques. We take the assumption that since biomedicine and TCM are both focusing on the study of disease phenomenon, we can regard the disease concept as the connecting point of biomedicine and TCM. There is huge literature and clinical bibliographic records on the research of SC-disease relationships in TCM, and the

genetic pathology of disease is also intensively studied by modern biomedical research. Therefore, when analyze the relationship between SC and gene or protein through disease concepts, it maybe generate some novel hypothesis knowledge, which is not conceived in TCM or modern biomedical field. To have some probing research with system biology, we focus on connecting SC with gene to get novel gene temporal and spatial interaction information, which cannot be easily acquired by large-scale genomics or proteomics techniques.

3 Related Work on Biomedical Literature Mining

The post-genomic topics have been the main issue in biomedical and life science research. The human genomic sequence and MEDLINE propose the two most important shared knowledge sources, which will greatly contribute to the development and progress of big modern biology research. Knowledge discovery from all kinds of the huge biological data such as genomic sequence, proteomic sequence and biomedical literature (e.g. MEDLINE, the annotations of Swiss-Prot, GenBank etc.) has been the foci of bioinformatics research. Text mining from biomedical literature is one of the most important methods assisting for hypothesis driven biomedical research.

Prof. Swanson is one of the first researchers who use MEDLINE to find novel scientific hypothesis [12]. He proposed the concept of complementary literature probably with innovative knowledge[13] and provided a system named ARROWSMITH to help the knowledge discovery in medical literature of MEDLINE[14]. The work of Swanson gave a set up of knowledge discovery research in medical literature. Gordon and Lindsay got into the literature-based discovery research by applying Information Retrieval (IR) techniques to Swanson's early discoveries [15][16]. Weeber et al proposed the architecture of DAD-system, a concept-based Natural Language Processing system for PubMed citations, to discover the knowledge of drug and food. They claimed that the system could generate and test Swanson's novel hypothesis [17].

Fukuda et al [6] are one of the first researchers using information extraction techniques to extract protein names from biological papers. Since then, the extraction of terminological entities and relationships from biomedical literature is the main research efforts in biomedical literature mining. The research instances include words/phrases disambiguation[1], gene-gene relationships [1][8], protein-protein [3,4,5,7,9], and gene-protein interactions or specific relationships between molecular entities such as cellular localization of proteins, molecular binding relationships[18], and interactions between genes or proteins and drugs[19]. Bunescu et al [20] have a comparative study on the methods such as relational learning, Naïve Bayes, SVM etc. in biomedical information extraction. Hirschman et al [21] have a survey of biomedical literature data mining from natural language processing. They argued that there need a challenge evaluation framework to boost the promising researches. Yandell[22] takes a felicitous discussion of biomedical text mining as a new emerging field-biological natural language processing that will do great help to biology research. Sehgal [23] uses Mesh headings to create concept profiles to compute the similar genes and drugs. Perez-Iratxeta [24] uncovers the relation of inherited disease and gene by Mesh terms of MEDLINE and RefSeq. While, Freudenberg et al [25] got disease

clusters according to the fuzzy similarity between phenotype information of disease, which is extracted from OMIM (Online Mendelian Inheritance in Man) then predict the possible disease relevant genes based on the disease clusters.

Most of the recent related work is focus on extraction of terminology concept or concept relationship knowledge, which has been existed in the literature. Moreover, natural language processing techniques (NLP) are preferred since the knowledge is conceived in the sentences. Such analysis is not only computationally prohibitive but also error prone when using NLP, and it is not applicable for large-scale literature mining. We address the large-scale literature analysis by very simple method, which only considers the co-occurrence of terms. Currently, the dictionary-based term extraction method is used. The most similar work is that of Jenssen et al [8] and Wilkinson [26]. They provide the simple approach to build large scale literature network of genes while only considering the gene co-occurrence as the view of gene interaction and dictionary based term extraction method is used. Furthermore, Wilkinson follows the method of [27] to resolve some of the gene terminological ambiguities. However, in this paper we take advantage of TCM knowledge (e.g. SC and the SC-disease relationships) to consider the related genes by a temporal and spatial holistic perspective. The method is based on the assumption that the genes relevant to the same SC will have some temporal or spatial connections since SC reflects the holistic functional state of morbid human body. Currently, there are two aims of the work of this paper. One is to find the relevant genes of SC, and the other is to find the communities of related genes. Bootstrapping technique is used to extract the disease names from TCM literature since no Chinese disease dictionary is available and the irregular using of terminology in clinical literature. The experimental results show that bootstrapping is much suitable to TCM terminology name extraction. The communities of related genes can be modeled as sub-graphs as in [8] and [26]. We believe that SC as a core TCM clinical concept will give novel approach to discovery of gene functional relationships from literature. This paper gives the framework of this approach and some preliminary results.

4 Bootstrapping and Term Co-occurrence

As many modifiers, which reflect the genre, state or characteristic of a specific clinical disease, are used in clinical literature representation, and irregular clinical disease names are also popular, the clinical disease names are far more various than standard disease dictionary such as that of TCM headings terminology. Dictionary based automatic name extraction cannot meet for this situation. We use bootstrapping to extract the disease names from TCM literature, and then extract the relationship of SC and disease from TCM bibliographic records based on metadata of TCM headings and SC standard terminological database, which includes about 790 standard SC names. The term bootstrapping here refers to a problem setting in which one is given a small set of seed words and a large set of unlabeled data to iterated extract the objective patterns and new seeds from free texts. Current work has been spurred by two papers namely that of Yarowsky [28] and Blum [29]. Bootstrapping methods have been used to automatic text classification [30], database curation[31][32] and knowledge base con-

struction[33] from World Wide Web. Next we introduce the bootstrapping method of this article in detail.

Bootstrapping is an iterative process to produce new seeds and patterns when provided with small set of initial seeds. The initial seed information gives the objective semantic type, which bootstrapping technique should extract from the text. In this paper, we aim to extract the terminological name of disease. So some initial disease name seeds are given before the iterative procedure is set up (the initial seeds has 19 disease names). How to define, evaluate and use pattern is one of the core issue of bootstrapping technique. Since TCM literature is written in Chinese, the process of pattern will surely be different from that of bootstrapping used in English like literature. This paper takes a simple but efficient pattern definition and evaluation method, and uses a search-match method without any shallow parsing processing. We define the TCM terminological pattern as a 5-tuple as

$$P =< \text{lPstr}, \text{TermType}, \text{rPstr}, \text{RefCount}, \text{FreqCount} > \quad (1)$$

In which *lPstr* represents the left Chinese string of the seed tuples, *rPstr* represents the right Chinese string of the seed tuples (we only extract two and three Chinese character to form the left and right string respectively called *Two-Left and Three Right pattern* (*TLTR*)), *TermType* represents the semantic type of seed name such as disease and Chinese Medical Formula etc., *RefCount* and *FreqCount* represent the number of seeds produced by pattern and the occurrence of pattern respectively (we also define the corresponding *RefCount* and *FreqCount* of seed).

To keep the bootstrapping procedure be robust in extraction of high quality new seed tuples, we use a dynamic bubble up evaluation method to assure the high quality pattern contribute to the iterative process. Currently, *RefCount* is computed before the next iteration and considered as the only quality criterion to keep the pattern (when *RefCount* is above a pre-assigned threshold) to attend the next step bootstrapping procedure. It is very simple while produce patterns from seeds. Next, we give some descriptions of the procedure of patterns to new seeds. First, the *lPstr* string is used as the search string after the record texts are split by regular punctuations and get the sentences, which have *lPstr* string as their sub-string. Second, the *rPstr* string is used to match the sentences of first step. Finally, we extract the strings between the *lPstr* and *rPstr* string as the objective new seeds. Meanwhile, we refresh the value of *RefCount* and *FreqCount* of pattern. It is experimentally concluded that *TLTR* is very robust and efficient pattern in TCM terminological term extraction as Table 1 shows. The recall of bootstrapping in TCM literature may be due to the iterative threshold, data source quantity and quality. The bootstrapping recall of disease on TCM literature of 2002 (*WX__2002*) is obviously better than that of other years. One reason is that *WX__2002* has high quality data and most of the records have abstracts, but the number of abstracts of the others is small.

Based on the latest version approved gene symbol vocabulary from HUGO in Feb. 2004, and the bootstrapped disease term database, we develop a Perl program to search the title and abstract fields of online MEDLINE to acquire the disease and gene relevant PubMed citations. Before search MEDLINE, Chinese disease name is translated partial automatically to formal English disease name according to the TCM

headings database and manually check by TCM terminological expert is needed when no TCM headings of disease exist. Followed we induce the disease and gene term co-occurrences by computing the same PubMed identifiers between each disease and gene term. Currently, no disambiguation method is used and alias gene symbols are not yet considered to search MEDLINE. Fig.1 shows the related data collections used and the related modules for generating novel hypothesis from modern biomedical and TCM literature.

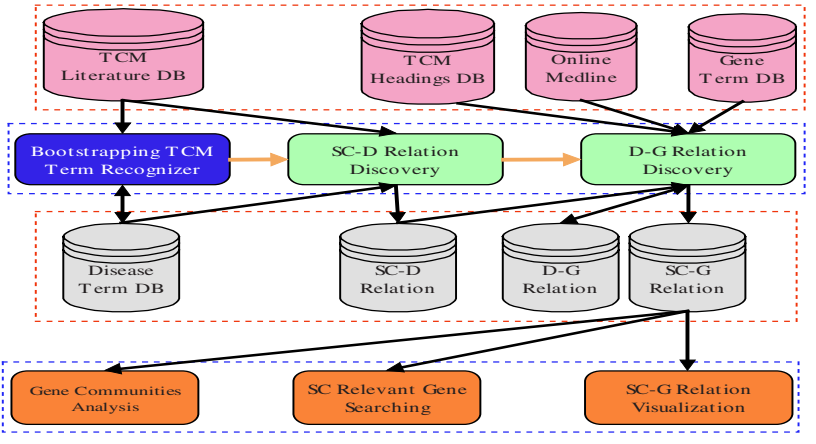

Fig. 1. The relevant data collections and related modules supporting knowledge discovery in biomedical literature. Blue box indicates a bootstrapping term recognizer to extract clinical disease terms. Green boxes represent SC-Disease and Disease-Gene relationships extraction modules. Orange boxes represent the modules that will be used to support biomedical research.

Table 1. Bootstrapping-based TCM terminological name recognition results, the TCM bibliographic database of year 1999, 2001, 2002 and 2003 has (38,937), (43,266), (44,315) and (16,151) records, respectively. P/S is the abbreviation of Pattern/Seed.

Years	Term Type	Iterated P/S Threshold	Iterated P/S Number	Precision	Recall	Pattern Number	Seeds Number
2002	Disease	8/8	8/20	92.6%	48.6%	3153	1018
2002	Disease	7/7	10/20	98.2%	55.5%	3153	1097
2002	Disease	6/6	73/109	90.8%	80.7%	5807	1753
2001	Disease	7/7	9/19	98.9%	30.1%	2430	915
1999	Disease	7/7	4/19	99.4%	35.3%	1459	853
2003	Disease	4/4	27/39	97.6%	21.1%	1684	416

5 Results

We have compiled about two and half million disease relevant PubMed citations, and 1,479,630 human gene relevant PubMed citations in local database. All of the citations are drawn from online MEDLINE by a Perl program. Meanwhile, we get about 1,100

SC-disease relationships from TCM literature (*WX__2002*). Fig.2 gives a picture of the concept associations supported by existed biomedical information. The SC relevant genes are all novel scientific discovery since there are no experiments to study the connections, and obviously, it is vital to boost SC from qualitative sense to quantitative experimental research. Furthermore, the related genes of the same SC will have some functional interactions, which is very different from the literature network built on gene term co-occurrence. And in a way, the appropriate subsets of the specific related genes have some selection of the huge literature gene network conceived in MEDLINE (e.g. the PubGene). To have a demonstration of the work of this article, we take the *kidney YangXu* SC as an instance since it is an important SC involving caducity, neural disease and immunity etc. in TCM, and has been studied by experimental approaches. Table 2 lists the 71 related genes of *kidney YangXu* SC (we exclude the "T" gene symbol because it seems no much information can be given by it). Inspired by the previous studies of Prof. Shen [11], we have an analysis whether the text mining method could find some novel knowledge from MEDLINE compared with experimental approaches. That is, if we can find the relevant genes such as CRF of *kidney YangXu SC*. Because the capacity of our gene symbols vocabulary is limit and alias gene symbols have not been considered (e.g. CRF is an alias name of CRH and is not in our gene symbol vocabulary), we have a confirmation process based on the work of PubGene [8] to have a relative complete view. Before the demonstration, we propose our basic assumption that polygenic etiology or gene interaction network will contribute to the phenotype of SC. We follow the next several steps to verify the efficiency of the text mining research. First, we select some important genes namely CRP (C-reactive protein, pentraxin-related), CRH (corticotropin releasing hormone), IL10 (interleukin 10), ACE (angiotensin I converting enzyme), PTH (parathyroid hormone), MPO (myeloperoxidase) in *kidney YangXu* SC. Second, we search the PubGene for subset network using each of the above genes. We get the subset networks as Fig 3 shows. The third step is analyzing the extracted knowledge. Now suppose that we don't know CRF (literature alias name of CRH) is a relevant gene of *kidney YangXu* SC. By analyzing the six subset networks in the left part and the CRF subset network of Fig 3, we may in a way get the hypothesis that CRF is some relevant to *kidney YangXu* SC, because that the subset network, which reassembled with the gene nodes such as IL10, CRAT, CRF/CRH/ ACE/MPO/PTH etc., constitute possible functional gene communities that contribute to *kidney YangXu* SC. No existed literature reporting the relationship between CRF and *kidney YangXu* SC is used to generate the novel knowledge since CRF is not in our gene vocabulary. It is excited that this simple demonstration has shown the primary text mining results will largely decrease the labor in molecular level SC research. The presented work of this paper gives a tool for the TCM researchers to rapidly narrow their search for new and interesting genes of a specific SC. Meanwhile, we give the specific functional information to the literature networks and divide the large literature networks to functional communities (e.g. the community containing IL10, CRAT, CRH/CRF etc. genes for *kidney YangXu* SC), which cannot be identified in the current PubGene. Moreover, through SC perspective, we can easily have a subset selection or explanation of giant literature based gene

network. And it is promising that even the low gene term co-occurrence network is functionally identified by our approach.

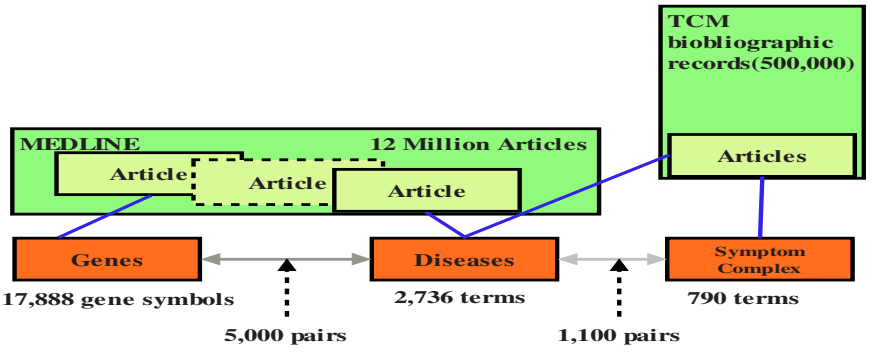

Fig. 2. Components used for identifying associations between genes and Symptom Complex. Green boxes represent databases. Orange boxes indicate the associated concepts and gray arrows represent the association used. Disease is the connecting point concept between gene and Symptom Complex.

Table 2. The 72 relevant genes of *kidney YangXu* SC filtered by total co-occurrence is above 10 or the number of relevant diseases is above 2. NP—526—2 represents the gene NP has co-occurred 526 times with the two diseases in *kidney YangXu* SC. Now there are only 7 relevant diseases of *kidney YangXu* SC, which have related genes in MEDLINE.

ID	Relevant Genes	ID	Relevant Genes	ID	Relevant Genes
1	NP--526--2	25	FCP--18--1	49	GC--11--3
2	IV--137--5	26	TG--18—4	50	IKBKAP--11--1
3	CCR3--61--2	27	MTHFR--18--2	51	EGF--11--4
4	ACE--56--3	28	FAT--18--1	52	AS--9--4
5	CD68--53--4	29	C3--17—6	53	CRP--9--5
6	C5--42--2	30	PON1—17--1	54	IL4--9--3
7	TNF--42--6	31	AHR--16--1	55	MCP--8--4
8	SD--34--5	32	PI3--16—2	56	PCNA--8--4
9	ALK--31--1	33	TAP1--16--1	57	HP--7--4
10	NOS1--31--1	34	HR--15—2	58	TF--7--4
11	SRS--31--2	35	STAT6--15--2	59	VEGF--6--4
12	CD34--30--3	36	STAR—15--1	60	MIP--5--3
13	AA--29--3	37	MPO--15--4	61	NPY--5--3
14	CD28--29--2	38	CD72--14--1	62	SDS--5--3
15	CD4--25--6	39	PTH--14--3	63	PC--5--3
16	CD14--24--3	40	NPHS2--14--2	64	CD2--4--3
17	MLN--24--1	41	LTB--13--3	65	CP--4--3
18	CXCR3--23--1	42	SEA--13--1	66	MIF--4--3
19	CD80--22--1	43	CCR2—12--2	67	DBP--4--3
20	FH--22--2	44	SC--12—4	68	HD--4--3
21	GSTP1--21--2	45	PAX2—12--2	69	EPO--4--3
22	TAT--21--3	46	IL13--12--2	70	CD63--3--3
23	CD86--19--1	47	EGFR—12--2	71	PGM1--3--3
24	ACTN4--18--2	48	CXCR4--11--1	72	T—161--7

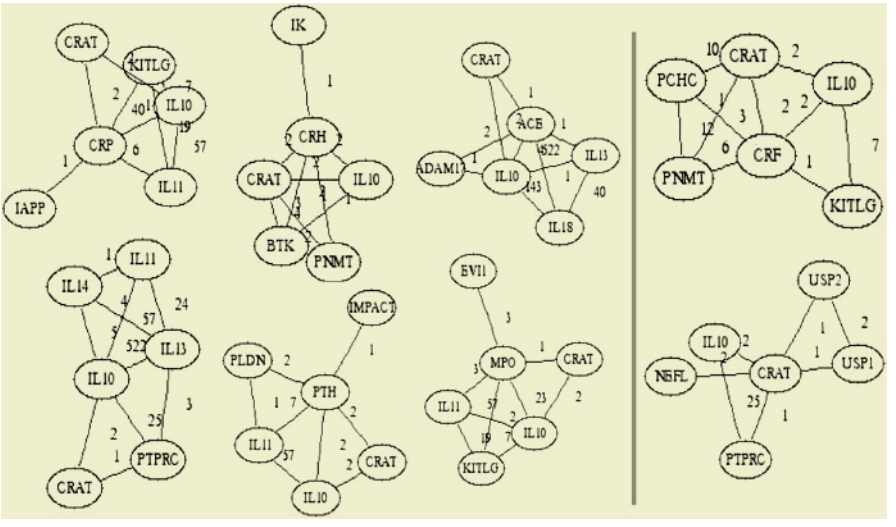

Fig. 3. The six subset networks of each selected genes (left part of the vertical yellow line). Gene CRAT, which in all of the six subset networks, may be a novel relevant gene of *kidney YangXu* SC. The right part gives the subset network of the already known relevant gene CRF, and the subset network of CRAT.

6 Concluding Remarks and Future Work

The exponential growth of MEDLINE and GenBank is rapidly transforming bio-NLP from a research endeavor into a practical necessity. Most of the recent biomedical literature mining studies are focused on information management, and the development of such tools is a necessary and laudable goal. Nevertheless, if bio-NLP (natural language processing) is to achieve its full potential, it will have to move beyond information management to generate specific predictions that pertain to gene function that can be verified at the bench. The synergistic use of sequence and text to extract latent information from the biomedical literature holds much promise in this regard. In this paper, we give a more novel text mining approach to generate large-scale biomedical knowledge by using TCM holistic knowledge. Bootstrapping techniques is used to extract the Chinese disease names from literature, and simple term co-occurrence is used to extract the relationship information. We believe that much more valuable hypothesis will be generated by enhancing the synergy of TCM and modern life science. As protein-protein interactions are central to most biological processes, the systematic identification of all protein interactions is considered a key strategy for uncovering the inner workings of a cell. The future work will focus on studying the protein-protein interactions using SC. Also we will try to use multistrategy learning method to extract more reliable gene-disease or protein-disease relationships from MEDLINE. Because many new terminological names are used in literature, but absent in vocabulary, vocabulary-based name extraction is not applicable to extract the latest relationships. Therefore, we will study on the novel using of bootstrapping techniques to gene or protein name extraction.

Acknowledgements

The authors thank all the researchers of Information Institute of China Academy of Traditional Chinese Medicine for the TCM bibliographic databases and discussion of TCM topics. This work is partially supported by National Basic Research Priorities Programme of China Ministry of Science and Technology under grant number 2002DEA30042.

References

1. Stephens, M. et al., Detecting Gene Relations from MEDLINE Abstracts. PSB 2001, pp: 483-95.
2. Hatzivassiloglou V., Duboue P.A., Rzhetsky A., Disambiguating proteins, genes and RNA in Text: a machine learning approach. Bioinformatics, vol.17 Suppl. 1 2001, pp:S97-S106.
3. James Thomas, et al., Automatic Extraction of Protein Interactions from Scientific Abstracts. psb2000.
4. Bunescu R. et al, Learning to Extract Proteins and their Interactions from MEDLINE Abstracts. Proceedings of ICML-2003 Workshop on Machine Learning in Bioinformatics, Washington DC, August 2003,pp: 46-53.
5. Marcotte E.M., Xenarios L. and Eisenberg D., Mining literature for protein-protein interactions. Bioinformatics,Vol. 17 no.4 2001,pages 359-363.
6. Fukuda K., et al, Toward information extraction: Identifying protein names from biological papers. In Proc. PSB 1998, Maui, Hawaii, January 1998, pp: 707-718.
7. Blaschke M. et al, Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions, Proc. of ISMB'99, pp. 60-67.
8. Jenssen T.-K., et al, A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics 28. 21-28(2001).
9. Humphreys, K., Demetriou, G., & Gaizayskas, R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. Pac. Symp. Biocomput.5. 505-516(2000).
10. Ideker, T., Galitski, T., Hood, L., A new approach to decoding life: Systems Biology. Annu. Rev. Genomics Hum.Genet. 2001,2.
11. Shen Ziyin, The continuation of kidney study. Shanghai, Shanghai scientific & Technical Publishers.1990.3-31.
12. Swanson, D.R., Two medical literature that are logically but not bibliographically connected, Journal of the American Society for Information Retrieval, 1987,38 (4), 228-233.
13. Swanson, D.R, Complementary structures in disjoint science literature. SIGIR-91, pp: 280-289.
14. Swanson, D.R. and Smalheiser, N.R., An interactive system for finding complementary Literature: a stimulus to scientific discovery, Artificial Intelligence, 1997, 91, 183-203.
15. Gordon M.D., Lindsay R.K., Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. J Am Soc Inf Sci 1996; 47 (2):116–128.
16. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 47 (2): 116–128.
17. Weeber, M. et al. Text-based discovery in biomedicine: the architecture of the DAD-system. In: Proceedings of AMIA, November 4-8, 2000, 903-907.

18. Rindflesch, T.C., Rayan, J.V. & Hunter, L. Extracting molecular binding relationships from biomedical text. Association for Computational Linguistics, Seattle, 2000,pp. 188-195.
19. Rindflesch, T.C., et al, EDGAR: Extraction of drugs, genes and relations from the biomedical literature. PSB 2000, 5:514-25.
20. Bunescu R., et al, Comparative Experiments on Learning Information Extractors for Proteins and their Interactions, Special Issue in JAIM on Summarization and Information Extraction from Medical Documents.25, August 2003.
21. Hirschman L., et al., Accomplishments and challenges in literature data mining for biology. Bioinformatics Review.Vol.18 no.12 2002, Pages 1553-1561.
22. Yandell M.D. and Majoros W.H., Genomics and Natural Language Processing. Nature Reviews Genetics, 2002, 3: 601-610.
23. Sehgal A., Qiu X.Y., Srinivasan P., Mining MEDLINE Metadata to Explore Genes and their Connections, SIGIR-03 Workshop on Bioinformatics.
24. Perez-Iratxeta C., Bork P. & Andrade M. A., Association of genes to genetically inherited diseases using data mining, letter to nature genetics, volume 31, july 2002.
25. Freudenberg J. and Propping P., A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics, Vol. 18 Suppl.2 2002, Pages S110-S115.
26. Wilkinson D. and Huberman B. A., A Method for Finding Communities of Related Genes. Proc. Natl. Acad. Sci. USA, 10.1073/pnas.0307740100.
27. Adamic L.A., et al., A Literature Based Method for Identifying Gene-Disease Connections.Proceedings of the IEEE Computer Society Conference on Bioinformatics, August 14-16, 2002, pp: 109.
28. Yarowsky D., Unsupervised word sense disambiguation rivaling supervised methods. ACL-95, pp. 189–196.
29. Blum A. and Mitchell T., Combining labeled and unlabeled data with co-training. In COLT: Proceedings of the Workshop on Computational Learning Theory. 1998.
30. Jones, R., et al. Bootstrapping for Text Learning Tasks. In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications.
31. Riloff E., Jones R., Learning Dictionaries for Information Extraction by Multi-level Bootstrapping, AAAI-99, pp. 474-479.
32. Brin, S. Extracting Patterns and Relations from the World Wide Web. WebDB Workshop at EDBT-98.
33. Craven, M. et al. Learning to Extract Symbolic Knowledge from World Wide Web. AAAI-98, pp: 509-516.