Liver Cancer Related Gene Analysis Based on Literature Mining

Xuan Liu, Wei Zou, and Jiajun Wang

***Abstract*—To identify liver cancer related genes and understand their interactions, we introduce a method based on literature mining. The genes are extracted with the disease-gene association classifier based on Bayesian network, and then the interaction network between the corresponding proteins is built. This paper has extracted 464 genes which are related to the liver cancer, and found genes such as p53, VEGF, TNF, AKT are hub proteins, which play important roles in the network. The KEGG pathway analysis shows that 19 enriched pathways may participate in liver carcinogenesis. Network analysis and pathway analysis implies the complexity of the occurrence, progression and prognosis of liver cancer.**

***Index Terms*—Bayesian classifier, genes and pathways analysis, liver cancer, literature mining.**

I. Introduction

In the past ten years, with the development of the biomedical science technology, the number of the biomedical literature has been growing exponentially. How to extract the needed knowledge and to obtain new knowledge quickly and efficiently from massive articles has become a very important research area. The interaction between diseases and related proteins is one of the main research directions which mean a lot to disease prevention, diagnosis, therapy and designing biomedical experiment and so on [1].

Liver cancer is one of the most common malignant tumors in China. It is reported that the number of people who die of liver cancer is about 110,000 every year and is 55% of the global patients. What's worse, the early symptoms of liver cancer are not apparent and the condition of patients deteriorates quickly. On average, patients can only live about 6 months after the cancer was diagnosed. Therefore, with all these mentioned above, research on liver cancer is necessary.

So far, the existing study has been focusing on the interaction between individual protein and liver cancer. The present problem is how to integrate the experiment information gained and researched to help analyze molecular interactions, pathways and their influence on the occurrence, progression as well as prognosis of liver cancer.

A method of identifying disease-gene associations based on literature mining which was applied in prostate cancer has been proposed by A.Ozgur *et al.* [1]. This method needs seed genes related to diseases as prior knowledge, and the extracted genes must appear with at least one of seed genes in the same sentence, which leads to the limit of that method.

Manuscript received January 24, 2013; revised March 24, 2013.

The authors are with the School of Electronic and Information Engineering, Soochow University, P. R. China (e-mail: liuxuan1220@yahoo.cn, zouwei_198107@163.com, jjwang@suda.edu.cn).

In order to solve this problem, we try to extract genes related to liver cancer without priors from the massive literatures and automatically build the interaction network between the corresponding proteins.

II. Methods

A. Search for Literatures on Liver Cancer

By searching the key words and free words of liver cancer in MeSH database of NCBI and collecting aliases of liver cancer from the reference [2] to conclude 14 aliases of liver cancer, such as liver neoplasm, hepatic cancer and so on after removing redundancy.

Based on the different names of liver cancer, 195730 articles related to liver cancer are found by searching them in PubMed.

Download these articles related to liver cancer (MEDLINE pattern). Articles in MDELINE pattern contain the titles, authors, abstracts, published journals, PMID and so on. Since the needed information is only titles and abstracts, it is necessary to extract them and cut abstracts into sentences by programming.

B. Filter Candidate Sentences

In [1], the relationship between genes and diseases is predicted by mining genes interaction networks. An assumption was proposed before extracting gene-gene interaction: the genes relationship must contain at least two genes and one interactive word. In this paper, another assumption is proposed: the sentences which describe the relationship between liver cancer and genes must contain a gene name, a cancer name and a diseases-gene interactive word.

There are 19,366 genes in the gene dictionary organized by HUGO Gene Nomenclature Committee [3] and all the genes belong to the type of protein-coding. Interactive words dictionary contains 118 words including verb, noun and different forms of them.

We try to filter out the sentences which contain cancer's name and interactive words. Firstly, the sentences that contain cancer names are filtered out from titles and abstracts. Secondly, the sentences with interactive words are filtered out from the results of first step. Lastly, the genes in these sentences are labeled. The method of labeling genes is based on dictionary. Although the most common method is entity recognition, the results of labeling with that method is not perfect. Meanwhile, the method of entity recognition cannot provide the identifying information of the entity such as GenBank ID and SwissProt ID, which is necessary in information fusion [4].

In order to improve the accuracy of labeling genes, Brill's tagger tools are applied to analyze sentences before

DOI: 10.7763/IJBBB.2013.V3.246

recognizing genes based on dictionary. Some parts of speech of the words, including noun, adjective, proper noun, foreign word and numeral are kept as candidate genes.

C. Extract the Gene-Disease Association

The method used to extract gene-disease association is based on Bayesian network classifier. One of the methods for building and training the gene-disease association classifier based on Bayesian network is to use Weka tools [7]. During the building and training progress, 12 feature vectors are used which are shown in Fig. 1, such as disease-gene interactive words, the distance of the first two factors, and order of these three factors and so on. The training set consists of 1713 statements (positive: 910 negative: 803) that describe the gene-disease association. The tenfold cross validation accuracy and recall ratio are both 88.2%.

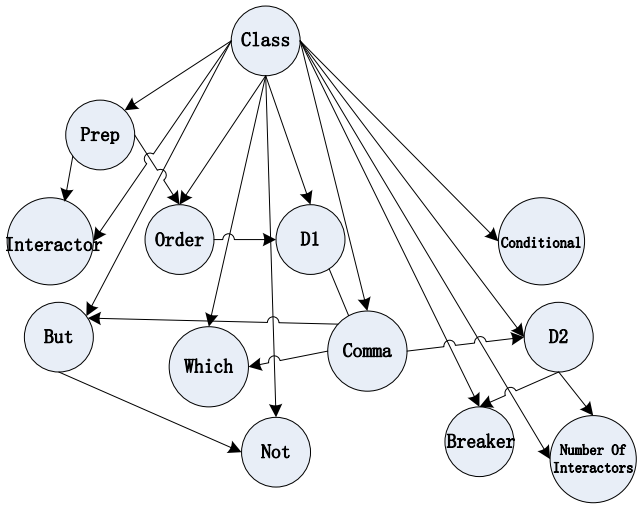

Fig. 1. Feature vectors of the classifier.

D. Build the Interaction Network between the Corresponding Proteins

Protein-protein interactive words library contains 191 words organized by reference [8]. Firstly, filter out the sentences that include protein-protein interactive words and at least two proteins related to liver cancer from the literature of liver cancer. Then, protein-protein association classifier based on Bayesian network is used to extract the interaction between proteins related to liver cancer. Lastly, the software Cytoscape is used to draw the interaction network between the corresponding proteins [9].

E. GO Analysis

Gorilla developed by Eden, is a GO enrichment recognize tool based on web page [10]. This paper intends to find out the hub nodes of the interaction network and then extract the proteins related to these hub nodes. At last, the Gorilla tool is applied to make the GO enrichment analysis for these proteins.

F. KEGG Enrichment Analysis

KEGG database contains metabolism network, cellular process, human disease and many other pathways [11]. 186 pathways and related gene data are downloaded from MSigDB database [12]. The enrichment of the genes related to liver cancer in KEGG pathways is measured by calculation of the P-value by Fisher bilateral accurate testing.

III. Results

A. Genes Related to Liver Cancer

The paper filters out 5249 sentences that include at least a gene name, a cancer name and a disease-gene interactive word. There are 609 non-redundant genes in these sentences. After the classification by the disease-gene association classifier based on Bayesian network, 2675 sentences are judged to be true in describing the relationship between the candidate genes and liver cancer. After removing redundancy of the results, 464 genes are found related to liver cancer.

B. Build and Analyze the Interaction Network between Corresponding Proteins

As mentioned above, 464 genes related to liver cancer are found and protein-protein interaction network is drawn by using Cytoscape software shown in Fig. 2. The results show that among these 464 genes, 368 genes have corresponding proteins with 5100 pairing interaction network and 14 genes are regarded as hub proteins because the number of their interactive proteins is beyond 30. The hub nodes are p53 (TP53，connection:68), TG (68), TNF (44), gamma (41), AKT (40), VEGF (40), AFP (34), BCL2 (34), p38 (33), p21 (33), beta-catenin (33), STAT3 (32), ERK (30), TRAIL (30) and so on. These hub proteins are in the center of the network and are very important in the network. They can be assumed as the key proteins in the occurrence, progression and prognosis of liver cancer.

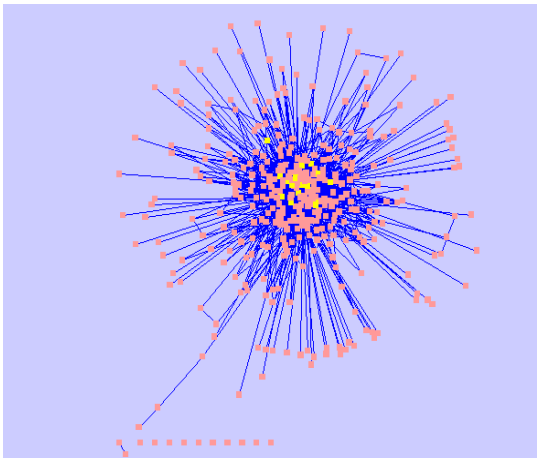

Fig. 2. Protein-protein interaction network in liver cancer. (yellow nodes: hub proteins, red nodes: non-hub proteins, blue line: the interaction of proteins, undirected graph).

C. GO Enrichment Analysis of Proteins Interactive with Hub Proteins

The result of GO enrichment analysis for hub nodes shows that feature of proteins which are interactive with VEGF concentrates on enzyme binding (*P*-value: 3.71E-13) and so on. VEGF has been proved to be very important to liver cancer, which can provide more information for tumor angiogenesis, tumor growth and metastasis [13].

D. Enrichment Analysis of Genes Related to Liver Cancer in KEGG Pathway

From Table I, it can be seen that the results of the KEGG enrichment analysis show that genes which are related to liver cancers are active in 19 pathways. Many of these

pathways have been reported to have closely interaction with the occurrence process of cancer. For example, CYTOKINE CYTOKINE RECEPTOR INTERACTION has been repeated to be related to the occurrence of lung cancer [14], CHEMOKINE SIGNALING PATHWAY is related to cardiovascular, and FOCAL ADHESION is related to liver cancer and gastric cancer and so on.

TABLE I: ENRICHED KEGG PATHWAYS OF LIVER CANCER RELATED GENES

Pathway	Number of liver cancer related genes in the pathway	Number of liver cancer related genes	Number of genes in pathway	Number of genes in human	*P*-value
PATHWAYS_IN_CANCER	47	464	328	19366	7.42E-20
CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	39	464	267	19366	4.30E-17
PANCREATIC_CANCER	20	464	70	19366	3.17E-14
CHEMOKINE_SIGNALING_PATHWAY	29	464	190	19366	1.26E-13
PROSTATE_CANCER	21	464	89	19366	1.81E-13
ADIPOCYTOKINE_SIGNALING_PATHWAY	17	464	67	19366	1.24E-11
TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	20	464	102	19366	1.25E-11
NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	16	464	62	19366	3.93E-11
SMALL_CELL_LUNG_CANCER	17	464	84	19366	2.64E-10
FOCAL_ADHESION	25	464	201	19366	2.72E-10
COLORECTAL_CANCER	15	464	62	19366	3.39E-10
BLADDER_CANCER	13	464	42	19366	3.88E-10
MAPK_SIGNALING_PATHWAY	28	464	267	19366	8.66E-10
T_CELL_RECEPTOR_SIGNALING_PATHWAY	18	464	108	19366	1.27E-09
MELANOMA	15	464	71	19366	1.71E-09
ENDOMETRIAL_CANCER	13	464	52	19366	3.51E-09
NON_SMALL_CELL_LUNG_CANCER	13	464	54	19366	5.19E-09
FATTY_ACID_METABOLISM	11	464	42	19366	3.65E-08
METABOLISM_OF_XENOBIOTICS	13	464	70	19366	7.59E-08

IV. CONCLUSION

In post-genomics era, the number of the biomedical literature is growing exponentially into a huge knowledge base. It is difficult to derive information that people are interested in only by manual reading from so massive articles. In this paper, literature mining has quickly and efficiently extracted 464 genes related to the liver cancer and built the interaction network between the corresponding proteins. In addition, with the help of the recognizing hub proteins, GO and KEGG enrichment analysis, this paper has studied molecular interactions, pathways and their influence on liver cancer with bioinformatics strategy and provided more information for doctors to diagnose and remedy.

REFERENCES

[1] A. Ozgur *et al.*, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277-85, 2008.

[2] D. Cheng *et al.*, "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic Acids Res*, vol. 36 (Web Server issue), pp. 399-405, 2008.

[3] E. A. Bruford *et al.*, "The HGNC Database in 2008: a resource for the human genome," *Nucleic Acids Res*, vol. 36 (Database issue), pp. D445-8, 2007.

[4] Y. Tsuruoka and J. Tsujii, "Improving the performance of dictionary-based approaches in protein name recognition," *J Biomed Inform*, vol. 37, no. 6, pp. 461-70, 2004.

[5] E. Brill, "Some advance in transformation based part of speech tagging," in *Proc. the 12th National Conference on Artificial Intelligence* (AAAI-94), Seattle, 1994, pp. 722-7.

[6] V. Jensen, *An introduction to Bayesian networks and decision graphs*, Springer-Verlag, 2001.

[7] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second edition, 2005, pp. 271-283.

[8] R. Chowdhary, J. Zhang, and J. S. Liu, "Bayesian inference of protein-protein interactions from biological literature," *Bioinformatics*, vol. 25, no. 12, pp. 1536-42, 2009.

[9] M. Kohl *et al.*, "Cytoscape: software for visualization and analysis of biological networks," *Methods Mol Biol*, vol. 696, pp. 291-303, 2011.

[10] E. Eden *et al.*, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, pp. 48, 2009.

[11] M. Kanehisa *et al.*, "KEGG for linking genomes to life and the environment," *Nucleic Acids Res*, vol. 36, pp. D480-4, 2008.

[12] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," in *Proc Natl Acad Sci USA*, vol. 102, no. 43, 2005, pp. 15545-50.

[13] A. Rapisarda and G. Melillo, "Role of the VEGF axis in cancer biology and therapy," *Adv Cancer Res.*, vol. 114, pp. 237-67, 2012.

[14] M. Y. Chang *et al.*, "Combined oligonucleotide microarray-bioinformatics and constructed membrane arrays to analyze the biological pathways in the carcinogenesis of human lung adenocarcinoma," *Oncol Rep*, vol. 18, no. 3, pp. 569-79, 2007.

Xuan Liu was born in Jiangsu Province, China, in December 1988. She received her B.Sc in 2011 from Soochow University, China.

She is currently a master student with the School of Electronic and Information Engineering, Soochow University, China. Her research is mainly focused on bioinformation and medical image processing.

Wei Zou was born in Jiangsu Province, China, in July 1981. He completed the Ph.D studies as a joint Ph. D student by Soochow University and the University of Sydney. He received the Ph.D degree in School of Electronic and Information Engineering from Soochow University in 2010.

He is currently a lecturer in school of electronic and information engineering of Soochow University. His research interest includes the image reconstruction, image processing and bioinformation.

Jiajun Wang was born in Jiangsu Province, China, in 1969. He received his BSc and MSc both in physics in 1992 and 1995 from Soochow University, China and his PhD in Biomedical Engineering from Zhejiang University in 1999.

He is currently a professor with the school of Electronic and Information Engineering, Soochow University, China. His research is mainly focused on medical imaging, image processing, pattern recognition and bioinformation. He has published more than 40 scientific journal or conference papers.