# Extraction of Biomedical Informtion Related to Breast Cancer Using Text Mining

Lejun Gong*
Jiangsu High Technology Research Key Lab for Wireless Sensor Networks，College of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, china

Ronggen Yan
College of Intelligent Science and Control Engineering, Jinling Institute of Technology, Nanjing, 211169, China

Quan Liu
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, china

Haoyu Yang
College of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, china

Gene Yang
Jiangsu High Technology Research Key Lab for Wireless Sensor Networks，College of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, china

Kaiyu Jiang
College of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, china

*Abstract*—**In this paper, we provide an approach to extract biomedical information related to breast cancer using text mining technology. We first extract entities related to breast cancer, then find these relationships, and visualize these biomedical information. Finally, these extracted biomedical information are annotated in the original experimental dataset which offer researchers as breast cancer biomedical information corpus. the approach provide an new way to obtain biomedical information of breast cancer. Moreover, it is promising for development of biomedical text mining.**

***Keywords- breast cancer, entity recognition, relationship extraction, visualization***

## I. INTRODUCTION

Breast cancer is very common and highly fatal in women and may lead to death. Almost 7% of breast cancers are diagnosed among women age 40 years and younger in Western populations[1].It has received great attention from the research community. Researchers from various domains are still working hard to discover treatment which leads to guaranteed cure from breast cancer. With the vast amount of scientific publications on breast cancer, the text mining technology make the daunting task be possible. Text mining concentrate on solving a specific problem in a specific domain using natural language processing (NLP) techniques which attempts to understand the meaning of text as a whole[2]. It is typically comprised of the following stages: (1) Information Retrieval (IR), (2) Named Entity Recognition (NER), (3) Information Extraction (IE)[3]. Information retrieval is to gather the papers which are relevant to the topic of interest by Boolean model and vector model. The best known one is PubMed, which searches the Medline database. There are other some IR systems, for example, PubMed Central(UKPMC)[4], which retrieves full-text documents from PubMed , PolySearch[5], which retrieves information including documents and database entries according to particular patterns of queries . NER is to find the biomedical entities that are mentioned within the gathered texts, for instance, the names of genes and proteins. Typical NER system includes ABNER[6] which could identify six categories entities based on conditional random fields (CRF)[7]. GARSCORE[8] could identify gene and protein names in text based on statistical model with an F-score of 82.5% for partial matches in test dataset. Information extraction aims to extract relationships among the named entities including two approaches: co-occurrence processing and natural language processing (NLP). iHOP [9] could process MEDLINE abstracts and generate a hyperlinked set of data for protein interactions with interactive functionality. Recent developments in molecular biomedicine and knowledge discovery have introduced new avenues for exploring the breast cancer. In this work[10], the researchers detected CyclinD1 expression using immunohistochemistry and CCND1 gene copy number in 355 invasive breast cancers. This study indicated loss expression of CyclinD1 might be an important event in the tumorigenesis in basal-like breast cancer. Moreover, they confirmed there are intense relationship between CyclinD1 protein and CCND1 gene in breast cancer. Fauteux[11] applied computational methods to select candidtate target overexpressed in three major breast cancer subtypes. Their results indicate that mining human gene

* Corresponding author.

expression data has the power to select and prioritize breast cancer antiboby-drug conjugate targets. The work[12] indicate the clinical relevance of CXCL13 to young breast cancer as a potential therapeutic target for young breast cancer. New genetic biomarks could be find in the existing literature.

This paper presents a text mining approach to extract biomedical information for potential biomarkers related to breast cancer. The biomedical information include entities and the relationships among them. The following sections describe the details.

## II. METHODS

Biomedical information related to breast cancer contain two types of information : entity and relationships among them. The pipeline of extraction of biomedical information related to breast cancer is the Figure 1.

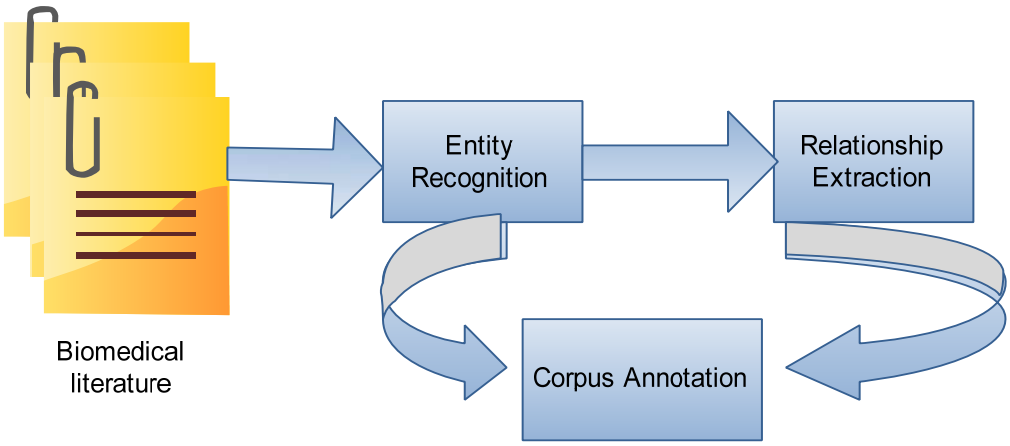


Figure 1. Pipeline of Extraction of biomedical information

### A. Dataset

The first step in the pipeline is to gather dataset which are relevant to the topic of interest. A query for"breast cancer"to PubMed could retrieve over 300,000 biomedical articles up to the end of December 2015 which could be acted as the experimental dataset.

### B. Entity Recognition

Entity recognition is the key step in the work, we used Conditional Random Fields (CRFs) to identify entity recognition.We developed a system called BerMiner based on CRFs model involving with six types of entities with a precision of 76.48%, a recall of 72.64 %, and a F-measure of 74.50% in protein entity recognition . In the work, we also integrate the prior entity recogntion tool ABNER for accurately identify entity.

### C. Relationship Extraction

Relationship extraction is to find the relationships between the biological entities mentioned in the text. There are two approaches to do this task: co-occurrence processing and natural language processing (NLP). In this study, we use the co-occurrence statistics to extract relationship among entities. The co-occurrence approach considere if two entities often occur in an article,a paragraph, a sentence, or a phrase, then the two entities have different level's relationship with different level's weights.

### D. Corpus Annotation

Identified entities and relationships related to breast cancer could not only offer references to breast cancer's researchers, but also develop biomedical text mining technology. We annotated the experimental dataset using identified entities and relationships with some tags. The offered annotated corpus could promote biomedical text mining technology.

## III. RESULTS

Aiming at the experimental dataset, we develop a system to validate our above mentioned approach. Our extracted entities contain 115214 entities by ABNER, 419539 ones by BerMiner, 61298 mutual entities between ABNER and BerMiner.

In relationship extraction, we used co-occurrence statistics to extract entities' relationships by weights to measure the relevance considering two level's relationship: article and sentence. The one effect visualization of extracted relationships is shown in Figure 2.

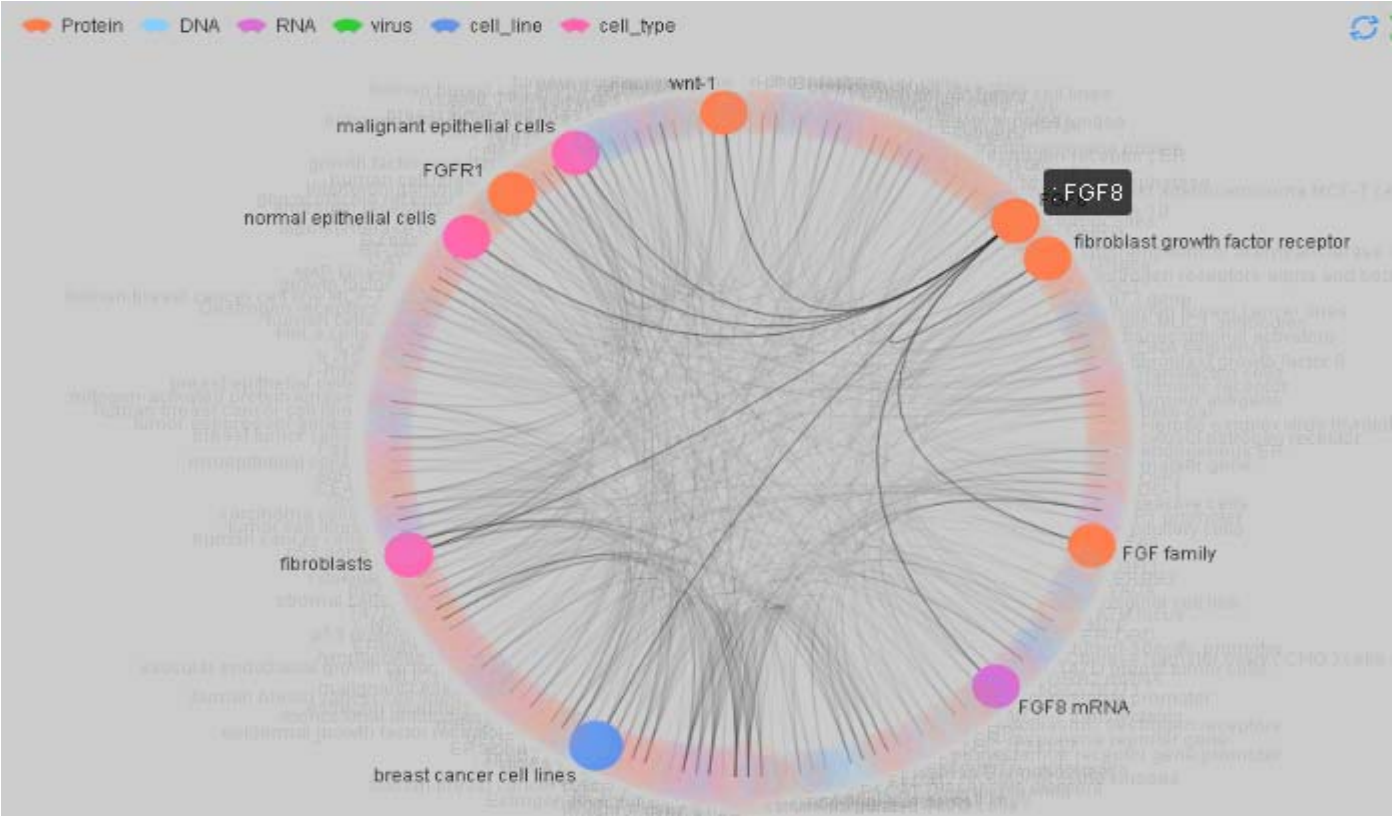


Figure 2. One effect visualization of extracted relationships

In Figure 2 show the extracted relationship visualization between FGF8 protein and wnt1, FGF8 mRNA, breast cancer cell lines,

The annotated results are shown in Figure 3 and Figure 4 using XML tags format. Figure 3 shows the annotated entities, and Figure 4 show the annotated relationship among entities. For example, Tag "<b class=\"protein"> …</b> " annotated protein entities in annotated entity corpus. Tag "EA" represents the entity could have the relationship of article level. Tag "ES" indicates the entity could have the relationship of sentence level. The number present the frequences of the level relationship.

In corpus annotation, we used extracted entity and relationship to annotate original experimental unstructured texts. using the Sentence class and the the annotated algorithm in the anotated corpus is shown in Figure 5.:


Figure 3. annotated entities' corpus


Figure 4. annotated entities and their relationships' corpus

```
Class Sentence    {
    int id;                    //   biomedical text's sentence identification
    String sentence;                                          // sentence's contents.
    List<String> sentenceEntity = new ArrayList<String>(); // stored the sentence's entities.
    void distinguishOneself() ;    // identify the sentence's entity, and add them to the data structure "sentenceEntity".
   String insertArticleTag(List<String> articleEntity) ;  //add tags to the sentence, and return the embedde tags's sentence.
    String SA; // stored the annotated text by tags.
}
```

The annotated algorithm

Input : biomedical texts;

Output: SA; // annotated texts by tags;

1. Split the *Text* with *id* to sentence set;
2. distinguishOneself();
3. For each sentence $S_i$
4. Identify each entity in the sentence and append to the *sentenceEntity* set;
5. insertArticleTag =Annotate the entity in $S_i$;
6. $Sa_i$=Annotate the relationship between the entities in $S_i$
7. Append $Sa_i$ to SA;
8. Return SA;

Figure 5 . The Sentence class and the the annotated algorithm in the anotated corpus

## IV. CONCLUSIONS

In this paper, we proposed an approach to extract biomedical information related to breast cancer involving with biomedical entities, biomedical relationship. First, we extract biomedical entities based on CRFs, then in the base of identified entities, the relationships among their entities are obtained by co-occurrence statistics. Finally, these extracted entities and their relationships are annotated in original experimental dataset as annotated breast cancer corpus. The approach offer a new venue for obtaining biomedical information related breast cancer. Experimental results show the approach is promising for development of biomedical text mining.

## ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (Grant Nos: 61272084, 61300240, 61572263, 61502251, 61503195 , 61502247, and 61502243), Natural Science Foundation of the Jiangsu Province (Grant Nos: BK20130417, BK20150863, BK20140895, and BK20140875), China Postdoctoral Science Foundation (Grant No. 2016M590483), Jiangsu Province postdoctoral Science Foundation (Grant No. 1501072B), Nanjing University of Posts and Telecommunications' Science Foundation (Grant Nos: NY214068 and NY213088). It is also supported by the Science Foundation of Zhongxing Telecommunication Equipment Corporation.

## REFERENCES

[1] Brenner DR, Brockton NT, Kotsopoulos J Cotterchio M, Boucher BA Courneya KS, Knight JA, Olivotto IA, Quan ML, Friedenreich CM4. Breast cancer survival among young women: a review of the role of modifiable lifestyle factors. Cancer Causes Control. 2016 Apr;27(4):459-72. doi: 10.1007/s10552-016-0726-5

[2] Rebholz-Schuhmann D, Oellrich A, Hoehndorf R.Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012 Dec;13(12):829-39. doi: 10.1038/nrg3337. Epub 2012 Nov 14. Review.

[3] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006 Feb;7(2):119-29. Review.

[4] McEntyre JR, Ananiadou S, Andrews S, Black WJ, Boulderstone R, Buttery P, Chaplin D, Chevuru S, Cobley N, Coleman LA, Davey P, Gupta B, Haji-Gholam L, Hawkins C, Horne A, Hubbard SJ, Kim JH, Lewin I, Lyte V, MacIntyre R, Mansoor S, Mason L, McNaught J, Newbold E, Nobata C, Ong E, Pillai S, Rebholz-Schuhmann D, Rosie H, Rowbotham R, Rupp CJ, Stoehr P, Vaughan P. UKPMC: a full text article resource for the life sciences.Nucleic Acids Res. 2011 Jan;39(Database issue):D58-65. doi: 10.1093/nar/gkq1063. Epub 2010 Nov 9.

[5] Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS.PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W399-405. doi: 10.1093/nar/gkn296.

[6] Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005 Jul 15;21(14):3191-2.

[7] Sha F , Pereira F. Shallow parsing with conditional random fields, In Proc. of HLT/NAACL 2003,1-8

[8] Chang JT1, Schütze H, Altman RB.GAPSCORE: finding gene and protein names one word at a time.Bioinformatics. 2004 Jan 22;20(2):216-25.

[9] Fernández JM, Hoffmann R, Valencia A.iHOP web services. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W21-6.

[10] Li Z, Cui J, Yu Q, Wu X, Pan A, Li L. Evaluation of CCND1 amplification and CyclinD1 expression: diffuse and strong staining of CyclinD1 could have same predictive roles as CCND1 amplification in ER positive breast cancers. Am J Transl Res. 2016 Jan 15;8(1):142-53.

[11] Fauteux F, Hill JJ, Jaramillo ML, Pan Y, Phan S, Famili F, O'Connor-McCourt M. Computational selection of antibody-drug conjugate targets for breast cancer. Oncotarget. 2016 Jan 19;7(3):2555-71.

[12] Chen L, Huang Z, Yao G, Lyu X, Li J, Hu X, Cai Y, Li W, Li X, Ye C. The expression of CXCL13 and its relation to unfavorable clinical characteristics in young breast cancer. J Transl Med. 2015 May 20;13:168. doi: 10.1186/s12967-015-0521-1.