Target Identification and Target-centered Network Construction from Biomedical Literature

Lejun Gong[1,2] Yunyang Yan[1] Xiao Sun[2*]
[1]Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian 223003, P.R. China
Email: glj98226@163.com
[2]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, P.R. China
Email: xsun@seu.edu.cn

***Abstract*—With the explosion of biological data and information in the omics era, text mining is becoming increasingly crucial for biomedical researchers to find relevant biomedical knowledge. This paper presents an approach for target identification and target-centered network construction from biomedical literature. The approach can identify several types of biomedical targets using finite state machine and ontology-based approach, and offer the implementation of target-centered network which could directly illustrate the relationships among targets. To validate the approach, we developed a system which achieved a recall 79.5%, a precision 83.1%, and an F-score 81.0% on average for the test datasets. Experimental results show that our approach is promising to develop text mining tool for biomedical researchers.**

***Index Terms*—target identification; text mining; biomedical ontology; finite state machine; target-centered network**

I. Introduction

Target discovery is the crucial step in the biomarker and drug discovery pipeline to diagnose and fight human diseases[1]. In biomedical science, the "target" is defined as a broad concept ranging from molecular entities, such as genes, and proteins, to biological phenomena, such as molecular functions, and phenotypes[2]. However, it is generally too expensive and time-consuming to perform experiments for discovering targets. The information in biomedical area is increasing at a considerable rate. This wealth of biological data and information presents immense new opportunities for target discovery in biomedical literature.

Currently, natural language processing techniques that are typically linguistically inspired, have demonstrated the potential to biomedical target discovery**.** Most studies focuses on extracting protein-protein interactions [3-4] and interactions that involve molecular mechanisms between proteins[5]. Very few works involve the targets along with diseases, genes and Gene Ontology[6] (GO) terms. This paper described the extraction of genes, diseases, and GO terms, which embody gene products and its functions, from free text. Focusing on the three types of targets in this work will help to understand pathogenic mechanism and discovery key targets. We proposed an approach for target identification and target-centered network construction. First, we used a finite state machine (FSM) to determine noun phrase boundaries, and then filtered these candidate targets to reduce the number of false positives as much as possible via ontology-based method. Finally, we constructed the target-centered network by statistic theory. A big advantage of this work is the implementation of the visualizations both multiclass identified targets and target-centered network with weights. Another advantage is that the mined information is combined to construct new information. In other words, the output of target identification is used as input for construction of the target-centered network.

*Corresponding author

II. Materials And Methods

This paper focuses on the identification of targets, including genes, diseases, and GO terms, as well as the construction of target-centered network. An overview of our proposed method for automatically extracting targets and constructing target-centered network is shown in Fig. 1.

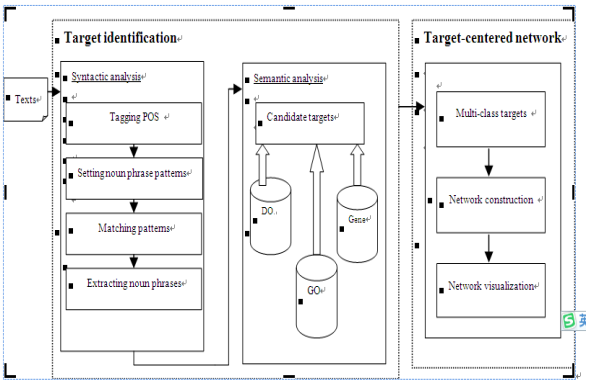

Fig. 1 Overview of our proposed approach.

Target identification is implemented by natural language processing containing syntactic and semantic analysis. The identified targets are used as the input for the construction of target-centered network, which is defined by the model of an undirected weighted graph. Further details of each module are presented in the following sections.

doi:10.4304/jsw.8.2.316-319

A. Target Identification

The goal of target identification is to identify interesting biomedical concepts in text. As already mentioned, target identification consists of the two processes: syntactic analysis and semantic analysis. The syntactic analysis is further divided into several main steps, including tagging Part-Of-Speech (POS), setting noun phrase patterns, matching patterns and extracting noun phrases.

POS tagging refers to mapping individual words to their syntactic classes (for example nouns, verbs, and adjectives). Each sentence is parsed to generate POS tags using the Stanford POS Tagger[7]. Generally, a biomedical target term consists of a noun phrase. In accordance with English grammar, a noun phrase may be the combination of nouns and noun modifiers (such as IL-2 gene expression, reactive oxygen production), including: adjectives, nouns, quantifiers, participles, verbal nouns, and prepositions. GENIA[8] is the largest available corpus, comprising 2000 abstracts with approximately 19000 sentences containing 36 class targets annotated by manual analysis. In the corpus, the 34,585 tagged biomedical targets consist of noun phrases. Statistics show several types of POS labels including adjectives, nouns, quantifiers, participles, verbal nouns, and prepositions, occupy more than 96% tagged entities share. Moreover, analyzing these biomedical targets, we found generally nouns, quantifiers, and verbal nouns could be as the end words of noun phrases, and prepositions occur behind the following words: nouns, quantifiers, and verbal nouns. Through syntactic analysis, noun phrase patterns are set by the FSM in terms of the combination of POS labels, as shown in Fig. 2.

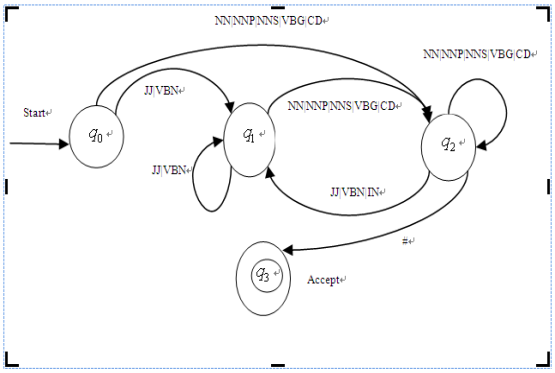

Fig.2 Noun phrase patterns defined by FSM.

Let us suppose the following: $M = (\Sigma, Q, \delta, q_0, F)$ is a 5-tuples;

$\Sigma = \{\text{JJ}, \text{NN}, \text{NNS}, \text{NNP}, \text{VBN}, \text{VBG}, \text{IN}, \text{CD}, \#\}$ is a finite set of labels; the tag "JJ" is defined as adjectives; "NN" as normal nouns, "NNS" as plural nouns, "NNP" as proper nouns, "VBN" as past participles, "VBG" as verbal nouns, "IN" as prepositions; "CD" as cardinal numbers; the notation "#" is defined as the end label; $T = \{\text{JJ}, \text{NN}, \text{NNS}, \text{NNP}, \text{VBN}, \text{VBG}, \text{IN}, \text{CD}\}$ is a set of POS tags; $Q = \{q_0, q_1, q_2, q_3\}$ is a state set; q_0 is the start state; $F = \{q_3\}$ is the set of the accepted state where $F \subseteq Q$, and δ is the state transition function with respect to $Q \times \Sigma \to Q$.

To exemplify the flow of setting patterns, the instances of noun phrase patterns are shown in Fig. 3 for the sentence "IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production".

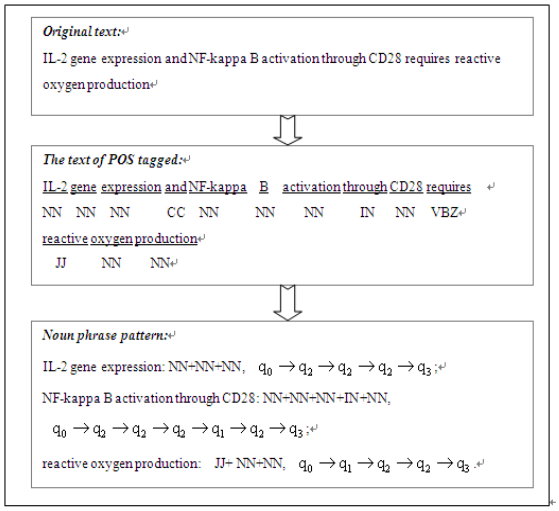

Fig.3 The instances of noun phrase patterns.

According to the patterns, noun phrases can be extracted by the process described as follows

1. Preprocess the text and split it into sentences.
2. Process each sentence with POS tagging.
3. Separate the POS from the tagged sentence.
4. Set the POS patterns of noun phrases.
5. Feed the separated POS to the FSM.
6. Extract the combinations of POS labels of the noun phrases.
7. Identify the noun phrases via the corresponding combinations of POS.
8. Store these noun phrases as candidate biomedical targets.

Semantic analysis involves looking for biomedical concept of each noun phrase, which is accomplished by applying a specific domain ontology controlled vocabulary. Ontology-based approach attempts to map a noun phrase occurring in a text to a concept in biomedical ontology. Linking domain-specific terms to their descriptions in the ontologies provides a platform for the semantic interpretation of textual information. In this step, the following resources are used: Disease Ontology[9] (DO) downloaded from the website (http://sourceforge.net/projects/diseaseontology/files/), is a shared resource related to human disease. Entrez gene information[10] and HGNC[11], compile as gene information associated with human. Finally, GO (http://www.geneontology.org) provides a controlled vocabulary for the description of the gene products and function corresponding to cellular component, molecular function, and biological process. These mentioned biomedical ontologies are compiled as biomedical concept dictionaries, allowing the candidate biomedical targets (extracted noun phrases) to be refined.

B. Target-centered Network

Multiclass biomedical targets are generated by the process of target identification, and act as nodes for constructing the target-centered network. An undirected weighted graph is used as the topological model for the target-centered network, which is defined as follows.

Let us suppose the following: the undirected weighted graph $G=\{N,E,W\}$ is a 3-tuple; $N=\{n_i \mid i=1,2,...\ \mathrm{k}\}$ is a set of target nodes; $E=\{(n_i,n_j) \mid n_i,n_j \in N\}$ is a set of edges; f_{ij} represents scoring based co-occurrence frequency with the fine granularity of sentence level in the text between nodes n_i and n_j; W is a upper triangular matrix, $W_{ij}=f_{ij}$ where $i=1,2,...\ \mathrm{k}$ and $j=i+1,...\ \mathrm{k}$. The formula $f_{ij}=0$ shows there does not exist an edge (relationship) between nodes n_i and n_j, $f_{ij}\neq 0$ shows there is an edge (relationship) between nodes n_i and n_j. Moreover, the bigger value of f_{ij}, the stronger correlation is between nodes n_i and n_j.

III. Results And Discussion

In this paper, we use JSP and MySQL technology to implement target identification and construct target-centered network. Screenshots of the implementation are shown in Fig. 4.

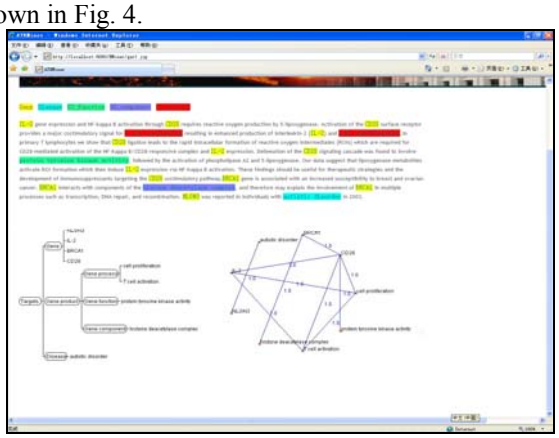

Fig. 4 Screenshots of the implementation of target identification and target-centered network.

The identified multiclass biomedical targets are shown in Fig.4. Different colors represent different types of biomedical targets. For example, the phrase "T cell activation" has been recognized as a GO term relating to biological process with red color. The extracted multiclass targets are also presented in the form of a tree diagram. Furthermore, our system also generates a visualization of target-centered network. In the network visualization, nodes represent biomedical targets, and edges represent the relationships among the nodes. Moreover, the weight over the edge represents the correlation between two nodes in the network. The bigger the weight, the stronger the correlation is between nodes. For example in Fig. 4, node "IL-2" co-occurs with "CD28" in the same sentence in the biomedical text, which shows that they have a kind of relationship. Hence, there is an edge between the two nodes in the target-centered network. The weight 3.0, presents the correlation between nodes "IL-2" and "CD28". This two nodes are represented the most significant correlation as measured by weights in the target-centered network visualization.

The output of target identification is used as the input for constructing the target-related network in our approach. Therefore, the task of target identification is the key to measuring our proposed approach. To evaluate the system performance, we use three popular measurements of the quality of the results, Precision, Recall and F-measure. Precision is the number of correctly extracted biomedical targets (TP) divided by the total number of biomedical targets extracted (sum of TP and FP). Recall is the number of correctly extracted biomedical targets divided by the total number of biomedical targets in the test datasets (sum of TP and FN). F-measure is a combination of the two values, defined as the harmonic mean of precision and recall. Experimental results have been evaluated in terms of Recall, Precision and F-measure as defined below:

$$\mathrm{Precision(P)}=\frac{\mathrm{TP}}{\mathrm{TP+FP}}\times 100\% \quad (1)$$

$$\mathrm{Recall(R)}=\frac{\mathrm{TP}}{\mathrm{TP+FN}}\times 100\% \quad (2)$$

$$\mathrm{F-measure(F)}=\frac{2\mathrm{PR}}{\mathrm{P+R}}\times 100\% \quad (3)$$

We used 500 abstracts related to biomedical literature as test datasets from PubMed site. Experimental results are shown in Table I and Table II for the test datasets.

TABLE I.
COUNTS OF MULTI-CLASS TARGETS

Target Class	TP	FP	FN
Gene	436	60	113
GO terms	22	6	8
Disease	383	80	60

TABLE II.
PERFORMANCE OF OUR SYSTEM

Target Class	P	R	F
Gene	88.0%	79.0%	83.0%
GO terms	78.6%	73.0%	75.6%
Disease	82.7%	86.4%	84.5%
Average	83.1%	79.5%	81.0%

In the test datasets, the counts of gene targets are more than that of other two types of targets, and the counts of gene products are few. According to the above mentioned measurements, our system achieved a recall 79.5%, a precision 83.1%, and an F-score 81.0% on average for the test datasets.

Simultaneously, we also used the BioCreAtIvE II [12] Gene Normalization (GN) Task as a gold standard. The

gold standard set contains 785 human gene targets in 262 abstracts. Our system obtain 78.60% recall, 74.97% precision and 76.73% F-measure.

Some biomedical targets in text aren't identified by our system yet. The causes are as follows.

Incomplete vocabularies. In many cases, hyponyms are used in the abstract. For example, "Humly9", which is not contained in gene ontology dictionary, is used in one abstract to represent gene "Ly9", "a second type (Mr approximately 60,000) of IL-1 receptor" to describe "IL1R2 interleukin 1 receptor, type II".

Tagging errors. Noun phrase recognition largely depends on the combination of the POS. the Stanford POS Tagger which is statistics model based on maximum entropy can also generate a few errors. For example, most of words of "A" occurring in the biomedical literature often act as articles in text. Occasionally word "A" could be used as a proper noun in biomedical targets. A sample "FA proteins A" in one abstract is not recalled owing to the cause that the word "A" is tagged as the indefinite article by the tagger. The target "HIV promoter construct" is not recalled in that the word "construct" is tagged as the verb.

Structure analysis errors. The errors are generated due to the tokenization by punctuation mark as delimiters. For example, the phrase "SMADs 1, 5, 8" as a biomedical target in gold standard is broken into three constituent tokens by delimiters and the target "UDP-galactose: Ga1beta-4G1cbeta1-cer alpha1, 4-galactosyltransferase" in text is broken into three constituent tokens.

Partial matching. In this study, the extraction of targets is via noun phrases matching the target name and symbol in the ontology dictionaries. Some biomedical targets are described rather than referred to by target name or symbol as in "insulin- and ECG-receptor". Only "ECG-receptor" is identified as a target in our system. The phrase "light chain-3 of microtu-bule-associated proteins 1A" in text is described as the target name "microtubule-associated protein 1 light chain 3 alpha".

IV. Conclusions

In this paper, an approach is proposed for target identification and target-centered network construction. The approach identifies biomedical targets using natural language processing technology including: POS tagging, syntactic analysis, and semantic analysis, and constructs target-centered network defined by the topological model of the undirected weighted graph. Our proposed approach can be used to identify multiclass targets, and generate a graphic visualization for target-centered network, which can directly illustrate the relationships among targets. The approach may be offer a new path to discovery targets and to understand disease etiology aiming at a large volume of biomedical literature. Our experimental results show that the proposed approach is promising for the development of biomedical text mining technology, which will help to understand molecular mechanism for disease researchers.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Project No. 61073141, 60971099), and Natural Science Foundation of the Higher Education Institutions of Jiangsu Province in China (Project No. 12KJB520002).

References

[1] Lindsay M.A. Target discovery. Nat Rev Drug Discov. 2003, 2:831-838.

[2] Yang Y., Adelstein S.J., Kassis A.I. Target discovery from data mining approaches. Drug Discov Today. 2004,14:147-154.

[3] Chowdhary R., Zhang J., Liu J.S. Bayesian inference of protein-protein interactions from biological literature. Bioinformatics. 2009, 25(12):1536-42.

[4] Donaldson I., Martin J., de Bruijn B., Wolting C., Lay V., Tuekam B.,et al. PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics. 2003,4:11-23.

[5] Jiao D., Wild D.J. Extraction of CYP chemical interactions from biomedical literature using natural language processing methods.J Chem Inf Model.2009,49:263-269.

[6] Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet.2000; 25:25-29.

[7] Toutanova K., Manning C.D. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000,63-70.

[8] Kim J.D., Ohta T., Tateisi Y., Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics. 2003,19 Suppl 1:i180-182.

[9] Marquet G., Mosser J., Burgun A. A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: the case of OBO disease ontologies. Int J Med Inform.2007,76: 353-361.

[10] Maglott D., Ostell J., Pruitt K.D., Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2007,35:D26-31.

[11] Bruford E.A., Lush M.J., Wright M.W., Sneddon T.P., Povey S.,Birney E., et al. The HGNC Database in 2008: a resource for the human genome.Nucleic Acids Res.2008, 36:D445-448.

[12] Morgan A.A., Lu Z., Wang X., Cohen A.M., Fluck J., Ruch R., et al. Overview of BioCreative II gene normalization. Genome Biol. 2008,9 Suppl 2:S3.

Lejun Gong The lecture of Huaiyin Institute of Technology, her research area is bioinformatics. She is PHD student of State Key Laboratory of Bioelectronics, Southeast University. Currently, she is focused primarily on biomedical text mining and pattern recognition.

Yunyang Yan The professor of Huaiyin Institute of Technology, his research area is pattern recognition. Currently, he is focused primarily on imaging processing.

Xiao Sun The professor of State Key Laboratory of Bioelectronics, Southeast University, his research area is bioinformatics. He has published many papers in these research areas. Currently, he is focused primarily on epigenetic regulation and high-throughput sequencing data processing.