Extraction of protein-protein interactions using natural language processing based pattern matching

Kaixian Yu
Department of Statistics
Florida State University
Tallahassee, US
kaixianyu@stat.fsu.edu

Tingting Zhao
Department of Geography
Florida State University
Tallahassee, US
tzhao@fsu.edu

Peixiang Zhao
Department of Computer Science
Florida State University
Tallahassee, US
zhao@cs.fsu.edu

Jinfeng Zhang
Deptment of Statistics
Florida State University
Tallahassee, US
jinfeng@stat.fsu.edu

***Abstract*— a significant part of our knowledge is relationships between two terms. However, most of these information is documented as unstructured text in various forms, like books, online articles and webpages. Extract those information and store them in a structured database could help people utilize these information more conveniently. In this study, we proposed a novel approach to extract the relationships information based on Nature Language Processing (NLP) and graph theoretic algorithm. Our method, Grammatical Relationship Graph for Triplets (GRGT), extracts three layers of information: the pairs of terms that have certain relationship, exactly what type of the relationship is, and what direct this relationship is. GRGT works on a grammatical graph obtained by parsed the sentence using Natural Language Processing. Patterns were extracted from the graph by shortest path among the words of interests. We have designed a decision tree to make the pattern matching. GRGT was applied to extract the protein-protein-interactions (PPIs) from biomedical literature, and obtained better precision than the best performing method in literature. Beyond extracting PPIs, our method could be easily extended to extracting relationship information between other bio-entities.**

Keywords*— *Information extraction, relationship extraction, protein-protein-interactions, Nature Language Processing, graph theorem algorithm

I. INTRODUCTION

A significant part of biological knowledge lies on relationships among different biological terms including proteins, genes, small molecules, pathways, diseases, and gene ontology (These terms are called bio-entities in this paper). Information on bio-entity relationships such as protein-protein interactions (PPIs) is indispensable for our understanding of biological processes, disease complexes, and development of drugs [2]. Manual annotation has been used to extract this information from scientific literature and deposit into various database [3-7].

However, it is quite time and resource consuming to manually annotate these PPIs. And it has become much more difficult to keep pace with the ever increasing amount of publications in biomedical sciences. In recent years, computational methods have been developed to automatically extract molecular interaction information and other bio-entity relationships from the literature, and have been used to assist the human annotators in building databases [8-13]. Many computational studies have recently attempted to extract PPIs from published literature, mostly PubMed abstracts due to the easy accessibility of deposited articles [14]. All of the computational methods detecting PPIs based on some rules (or patterns, templates etc.). To specify the rules, there are two major approaches: either specified manually [15-18], or inferred/learned computationally from sentences that are manually annotated [19-21].

Initial efforts of computationally detecting PPIs were based on simple rules, such as co-occurrence, which assumes that two proteins likely interact with each other if they co-occurred in the same sentence/abstract [22]. These approaches tend to produce a large number of false positives, and still require a significant amount of manually annotated examples.

Later studies, aiming to reduce the high false-positive rate of earlier methods, used manually-specified rules. Such methods sometimes achieved a higher accuracy than co-occurrence methods by extracting cases satisfying the rules. However, they have disadvantages, such that the rules are not general enough due to missing cases not covered by the limited number of manually-specified rules [15-18, 23-27].

Recently, machine-learning-based methods were introduced to extract PPI information automatically. By learning the language rules using annotated texts, the machine learning involved methods have performed better

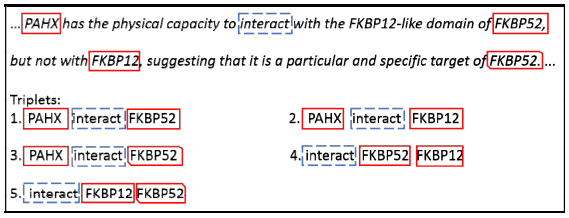

Figure 1: Example of PPIs, where the context [1] row shows the relative order of appearance in the context, and the circle represents the second occurrence of FKBP52.

than other methods in terms of both decreasing false-positive rate and increasing the coverage [19-21]. Huang *et al [19]* uses a dynamic programming algorithm, similar to that used in sequence alignment, to extract patterns in

This study was partially supported by a grant from NIGMS of National Institute of Health under award number R01GM126558.

sentences tagged by part-of-speech tagger. Kim *et al* [21] uses a kernel approach for learning genetic and protein-protein interaction patterns.

Corpus	No. of sentences	No. of Triplets	No. of true PPI
HPRD50	145	954	126
IEPA	374	1,341	164
LLL	79	977	106

Table1: Dataset information.

Up to present, there are few methods that extract both the protein names and the interaction word at the same time. However, being given only the protein names may not be sufficient to understand the PPI. Therefore, to extract the PPI triplet (two different protein names and one interact word) is needed to describe how the proteins are interacted [28]. Temkin and Gilder, 2003

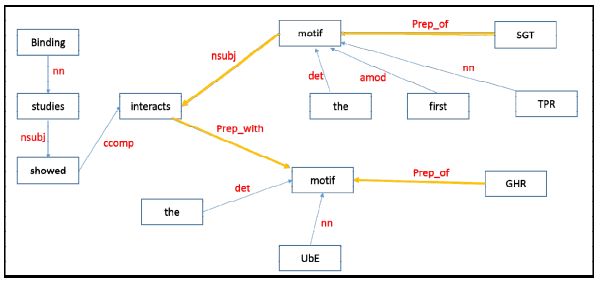

Figure 2: Grammatical dependencies graph.

There is a practical issue in extracting PPI triplets if we do not care about the structure of a sentence. For example, the sentence in figure 1 contains four protein names (FKBP12-like is not considered as a protein name) PAHX, FKBP52, FKBP12, and FKBP52 (the second occurrence of FKBP52 in the sentence) and one interaction word *interacts*. There are five PPI triplets (figure 1), only one of the triplets correctly describes this specific PPI (triplet 1 in figure 1).

Recently Natural Language Processing (NLP) techniques have been utilized in many machine learning approaches [15-18] to parse the sentence into dependency trees or constituent trees, which later could be used in pattern matching or rule-based searching. However, to our knowledge all the methods use some given rules/patterns. The given rules are rather general; therefore, they do not represent all the patterns in training sentences.

In this paper, we propose a method based on NLP and automatically learned rules/patterns to extract the PPI triplets from sentences, and to classify them as true or false with probability values based on whether the interaction word correctly describes the interaction relationship between the two protein names.

II. Method

Our approach used natural language processing (NLP) techniques and a graph theorem algorithm (shortest path algorithm). We used dictionaries for the names of proteins and interaction words, with variants, in this study.

At present, there have been some methods that used NLP techniques in protein-protein interaction extraction [8, 15-18, 29]. For our method, we first performed sentence parsing using Stanford Sentence Parser, and obtained the dependencies (grammatical relations) among the words (tags) in the sentences. For example, the sentence, *"Binding studies showed that the first TPR motif of SGT interacts with the UbE motif of the GHR,"* can be parsed to the following graph (figure 2) representing the grammatical relationships between the words (tags) in the sentence, where the words in red are typed dependencies defined in. The typed dependencies have a hierarchical structure themselves. Due to the limitation of space, we only introduce some necessary facts here. The top level of the hierarchical structure is dependent (*dep*), which has the following types: auxiliary (*aux*), argument (*arg*), coordination (*cc*), conjunct (*conj*), expletive (*expl*), modifier (*mod*), parataxis (*parataxis*), punctuation (*punct*), referent (*ref*) and semantic dependent (*sdep*). Each of the above types may have subtypes themselves. For example, *arg* has subtypes: agent (agent), complement (*comp*) and subject (*subj*), where *subj* has nominal subject (*nsubj*) and clausal subject (*csubj*) as its subtypes. For example, "Binding studies" is *nsubj* of "Showed," and "the first TPR motif of SGT interacts with the UbE motif of the GHR," is a clausal complement (*ccomp*) (figure 2).

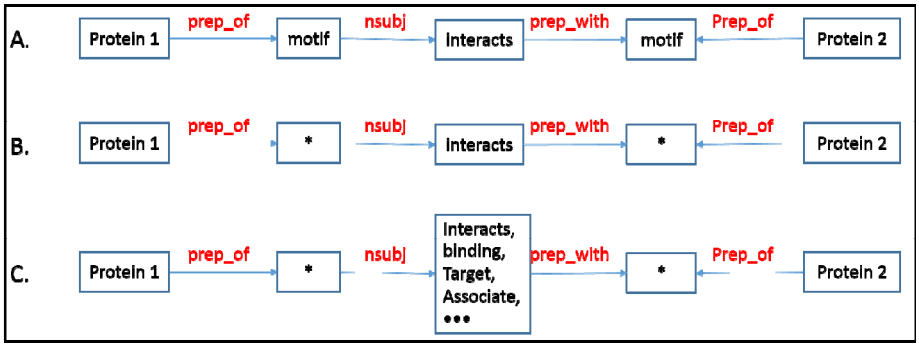

Figure 3: Grammatical dependencies sub-graph.

From the relationship graph (figure 2), we extracted a sub-graph containing the triplet (two protein names and the interaction word). To do so, we found the three pairwise shortest paths between pairs of the triplet elements. The

obtained sub-graph is the grammatical relationship graph for triplets (GRGT) (figure 3).

The GRGT of figure 3 describes the meaning *"motif of P1 (SGT) interacts with motif of P2 (GHR)."* The information in this graph is all the information we need to know to infer the interaction between SGT and GHR. In fact, this is true for majority of the triplets and their corresponding GRGT. A new triplet that matches the pattern (figure 3) is classified as true. Given a set of manually annotated true triplets, we could use a pattern matching approach to classify new triplets. The directions of the sub-graph can also be inferred at the same time, since the information of the direction of the true patterns can also be annotated.

So far, what we had was an exact matching. To account for similar but not exact matches, we could relax the sub-graph (figure 3A) by allowing the *motif* to differ from the annotated samples (figure 3B). Beyond that, we could make the matching even more general by further relaxing the exact interaction word to a set of similar interaction words (figure 3C). In this more general version, we manually grouped the interaction words into 20 groups by the similarity of their grammatical properties.

To implement the general version, we designed a simple decision tree, which has one decision node at each level representing the patterns at different levels of details. To demonstrate how the decision tree works, here we use the above interaction sentence *"motif of P1 (SGT) interacts with motif of P2 (GHR)"* as an annotated sample. The procedure is shown below:

- The first level of the decision tree will be the exact pattern in figure 3A . If the sentence does not match the pattern exactly, send the sentence to the second level. Therefore, *"motif of P1 interacts with motif of P2"* is a match, and the probability of triplet *"P1-interacts-P2"* being true is the proportion of this pattern being true. However, *"X-terminal domain of P1 interacts with residue 30-50 of P2"* does not match the pattern, thus should be passed to the next level.
- The second level is the relaxed graph as shown in figure 3B. At this level, the previous example, *"X-terminal domain of P1 interacts with residue 30-50 of P2"* is a match; therefore, the probability of this triplet *A-interacts-B* classified as true triplet is the proportion of the pattern being true. However, the sentence *"X-terminal domain of P1 associates with residue 30-50 of P2"* does not match the pattern from this level since the interaction word is different. Therefore, it is passed to the next level.
- The third level, as described in figure 3C, is the more relaxed version. In this level we allow the interaction words to differ from the annotated example for example, the above sentence *"X-terminal domain of P1 associates with residue 30-50 of P2"* is a match in level 3, although it is not a match in level 1 or 2. Therefore the triplet P1-*associates*-P2 is given the probability being true as the proportion of the pattern being true. If a sentence fails to match the pattern in this level (in practice there may be much more levels), we mark the triplet contained in this sentence as a false triplet.

With reduced representation, some patterns will have both true and false cases. When a triplet is matched to a pattern, the probability of the triplet being true can be assigned as the proportion of true cases with that pattern, as in a standard classification tree. If a triplet cannot be matched with any existing pattern, then it is classified as false.

Corpus	HPRD50			IEPA			LLL		
	F	P	R	F	P	R	F	P	R
Bui *et al.*	71.7	62.2	84.7	73.4	62.9	88.1	83.6	81.9	85.4
GRGT	64.0	**86.5**	50.8	64.0	**91.0**	63.6	63.8	**91.2**	77.1

Table 2: Performance comparison of our method (GRGT) with Bui. et al on four benchmark datasets. F: F1 measure, P: precision, R: recall. The measurement is out of 100

III. Results

We compared the performance of our method with [8], the best performing method in literature, on several benchmark datasets (Table 1): HPRD50, IEPA, LLL.

Our method had better precision for all the benchmark databases (Table 2). Most misclassified cases by our method were true triplets that cannot be matched to any known patterns. Higher precision is very important since discovered (classified) results are quite often used as prior knowledge to guide experiment design. If precision of a certain method is low, then it is likely the researchers received incorrect information. As a consequence, the experiment may be incorrectly designed. However, one could more or less tolerate recall rate being lower since interactions (PPI triplets) often occur more than once in literature. As long as one of them is classified as true, the interaction is extracted.

IV. Discussion

We could further simplify the patterns so that more true triplets can be matched if they are similar to true patterns, but not exactly the same. One option is to use the

hierarchical structure of the typed dependencies. For example, *nsubj* (nominal subject) can be reduced to *subj* (subject) or even further to *arg* (argument). Of course, by simplification, we would expect to improve the recall rate; however, the precision rate will be sacrificed, which means we have to balance these two. We will conduct some more experiments with various ways of reducing the exact patterns, and do experiments on how to combine the new relaxed patterns with our existing patterns by designing different decision trees to achieve better performance.

This method can be used to extract other relationships as well, as long as the triplet is well defined and the library for terms and interaction words are given.

REFERENCES

[1] B. Chambraud, C. Radanyi, J. H. Camonis, K. Rajkowski, M. Schumacher, and E. E. Baulieu, "Immunophilins, Refsum disease, and lupus nephritis: the peroxisomal enzyme phytanoyl-COA alpha-hydroxylase is a new FKBP-associated protein," (in eng), *Proceedings of the National Academy of Sciences of the United States of America,* vol. 96, pp. 2104-2109, Mar 2, 1999 1999.

[2] M. G. Kann, "Protein interactions and disease: computational approaches to uncover the etiology of diseases," (in eng), *Briefings in bioinformatics,* vol. 8, pp. 333-346, Sep 2007 2007.

[3] B. Aranda *et al.*, "The IntAct molecular interaction database in 2010," (in eng), *Nucleic acids research,* vol. 38, pp. D525-531, Jan 2010 2010.

[4] A. Chatr-Aryamontri *et al.*, "The BioGRID interaction database: 2013 update," (in eng), *Nucleic acids research,* vol. 41, pp. D816-823, Jan 2013 2013.

[5] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database--2009 update," (in eng), *Nucleic acids research,* vol. 37, pp. D767-772, Jan 2009 2009.

[6] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," (in eng), *Nucleic acids research,* vol. 36, pp. D684-688, Jan 2008 2008.

[7] C. Stark *et al.*, "The BioGRID Interaction Database: 2011 update," (in eng), *Nucleic acids research,* vol. 39, pp. D698-704, Jan 2011 2011.

[8] Q.-C. Bui, S. Katrenko, and P. M. A. Sloot, "A hybrid approach to extract protein-protein interactions," (in eng), *Bioinformatics (Oxford, England),* vol. 27, pp. 259-265, Jan 15, 2011 2011.

[9] A. Ceol *et al.*, "MINT, the molecular interaction database: 2009 update," (in eng), *Nucleic acids research,* vol. 38, pp. D532-539, Jan 2010 2010.

[10] X. Hu, X. Zhang, I. Yoo, X. Wang, and J. Feng, "Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule," *Int. J. Intell. Syst.,* vol. 25, pp. 207–223, February 2010 2010.

[11] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of BioCreative II," (in eng), *Genome biology,* vol. 9 Suppl 2, p. S4, 2008 2008.

[12] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser, "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature," (in eng), *PLoS computational biology,* vol. 6, p. e1000837, 2010 2010.

[13] L. Wong and G. Liu, "Protein Interactome Analysis for Countering Pathogen Drug Resistance," (in en), *Journal of Computer Science and Technology,* vol. 25, pp. 124-130, 2010/01/01 2010.

[14] A. Skusa, A. Rüegg, and J. Köhler, "Extraction of biological interaction networks from scientific literature," (in eng), *Briefings in bioinformatics,* vol. 6, pp. 263-276, Sep 2005 2005.

[15] Y. Miyao, K. Sagae, R. Saetre, T. Matsuzaki, and J. Tsujii, "Evaluating contributions of natural language parsers to protein-protein interaction extraction," (in English), *Bioinformatics,* vol. 25, no. 3, pp. 394-400, Feb 1 2009.

[16] H. T. Zhang, M. L. Huang, and X. Y. Zhu, "A Unified Active Learning Framework for Biomedical Relation Extraction," (in English), *Journal of Computer Science and Technology,* vol. 27, no. 6, pp. 1302-1313, Nov 2012.

[17] J. Lee, S. Kim, S. Lee, K. Lee, and J. Kang, "On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach," *BMC Med Inform Decis Mak,* vol. 13 Suppl 1, p. S7, 2013.

[18] K. Raja, S. Subramani, and J. Natarajan, "PPInterFinder--a mining tool for extracting causal relations on human proteins from literature," *Database (Oxford),* vol. 2013, p. bas052, 2013.

[19] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," (in eng), *Bioinformatics (Oxford, England),* vol. 20, pp. 3604-3612, Dec 12, 2004 2004.

[20] R. Malik, L. Franke, and A. Siebes, "Combination of text-mining algorithms increases the performance," (in eng), *Bioinformatics (Oxford, England),* vol. 22, pp. 2151-2157, Sep 1, 2006 2006.

[21] S. Kim, J. Yoon, and J. Yang, "Kernel approaches for genic interaction extraction," (in eng), *Bioinformatics (Oxford, England),* vol. 24, pp. 118-126, Jan 1, 2008 2008.

[22] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," (in eng), *Nature genetics,* vol. 28, pp. 21-28, May 2001 2001.

[23] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature," (in eng), *Bioinformatics (Oxford, England),* vol. 17, pp. 155-161, Feb 2001 2001.

[24] J. C. Park, H. S. Kim, and J. J. Kim, "Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar," (in eng), *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 396-407, 2001 2001.

[25] G. Leroy and H. Chen, "Filling preposition-based templates to capture information from medical abstracts," (in eng), *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 350-361, 2002 2002.

[26] J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran, "Robust relational parsing over biomedical literature: extracting inhibit relations," (in eng), *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 362-373, 2002 2002.

[27] J. M. Temkin and M. R. Gilder, "Extraction of protein interaction information from unstructured text using a context-free grammar," (in eng), *Bioinformatics (Oxford, England),* vol. 19, pp. 2046-2053, Nov 1, 2003 2003.

[28] V. Hatzivassiloglou and W. Weng, "Learning anchor verbs for biological interaction patterns from published text articles," (in eng), *International journal of medical informatics,* vol. 67, pp. 19-32, Dec 4, 2002 2002.

[29] S. Kim, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, and B.-T. Zhang, "PIE: an online prediction system for protein-protein interactions from text," (in eng), *Nucleic acids research,* vol. 36, pp. W411-415, Jul 1, 2008 2008.