Using Maximum Entropy Model to Extract Protein-Protein Interaction Information from Biomedical Literature

Conference Paper · August 2007

DOI: 10.1007/978-3-540-74171-8_72 · Source: DBLP

Chengjie Sun
Harbin Institute of Technology
105 PUBLICATIONS **1,527** CITATIONS
SEE PROFILE

Lei Lin
Peking University Third Hospital
86 PUBLICATIONS **968** CITATIONS
SEE PROFILE

Xiaolong Wang
Tsinghua University
307 PUBLICATIONS **6,006** CITATIONS
SEE PROFILE

Yi Guan
Harbin Institute of Technology
110 PUBLICATIONS **806** CITATIONS
SEE PROFILE

Using Maximum Entropy Model to Extract Protein-Protein Interaction Information from Biomedical Literature

Chengjie Sun, Lei Lin, Xiaolong Wang, and Yi Guan

School of Computer Science, Harbin Institute of Technology,
150001 Heilongjiang, China
{cjsun,linl,wangxl,guanyi}@insun.hit.edu.cn
http://www.insun.hit.edu.cn

Abstract. Protein-Protein interaction (PPI) information play a vital role in biological research. This work proposes a two-step machine learning based method to extract PPI information from biomedical literature. Both steps use Maximum Entropy (ME) model. The first step is designed to estimate whether a sentence in a literature contains PPI information. The second step is to judge whether each protein pair in a sentence has interaction. Two steps are combined through adding the outputs of the first step to the model of the second step as features. Experiments show the method achieves a total accuracy of 81.9% in BC–PPI corpus and the outputs of the first step can effectively prompt the performance of the PPI information extraction.

Keywords: protein-protein interaction, maximum entropy, text mining, machine learning.

1 Introduction

Protein-Protein interaction(PPI) information are very useful in biological research. Characterizing protein interaction partners is crucial to understanding not only the functional role of individual proteins but also the organization of entire biological processes. Vast number of PPI information are buried in biomedical literature and the number of paper published is growing at a high speed. For example, MEDLINE database has collected eleven million biomedical related records since 1965 and is increasing at the rate of 1500 abstracts a day [1]. So biomedical literature has been exploited as a resource to derive protein interaction records for interaction databases. There are several PPI databases which are literature-derived, such as Reactome [2], BIND [3] and HPRD [4].

Due to the rapid growth of the biomedical literature and the increasing number of newly discovered proteins, it is becoming difficult for the interaction database curators to keep up with the literature by manually detecting and curating protein interaction information. Automatically extracting PPI information from biomedical literature is an urgent task and has become an active research

D.-S. Huang, L. Heutte, and M. Loog (Eds.): ICIC 2007, LNCS 4681, pp. 730–737, 2007.

field recently. Existing PPI information extraction methods can be roughly divided into four categories: co-occurrence based approaches [5, 6], parser-based approaches [7, 8], pattern-based approaches [9, 10] and machine learning based approaches [11, 12].

Co-occurrence based approaches simply use co-occurrence statistics of two proteins to predict their relation. In this way, they can only extract well-known PPIs but may not be able to find new emerging PPIs. Because of the complex of sentences in biomedical literature, parser-based approaches are inherently complicated. Besides, parser-based approaches require many resources and their performances are not satisfied. Pattern-based approaches utilize pre-defined phrase pattern rules. As a result, they are unable to discover new phrase patterns without the known keywords. Besides, once the rule set reaches a certain size, it is very difficult to insert additional rules for further performance improvement. Machine learning based approaches can overcome the limitation of above approaches, but they need high quality annotated training corpora. Fortunately, such corpora have been available recently [13].

Inspired by [12], we cast PPI information extraction task as a binary classification problem and use an machine learning method —– Maximum Entropy model to solve it in this work. Maximum Entropy can conveniently involve rich features to do classification and has better classification performance compared to other classifiers in most applications. In this work, we focused on the extraction of PPI information contained in one sentence. The whole process is divided into two steps. The first step was designed to judge whether current sentence contains PPI information. The detail analysis on a corpus shows that it is also an important and challenge task. The sentence features generated in the first step will be helpful for the second step. The second step determines whether there is an interaction between every two proteins in the current sentence. Both steps use Maximum Entropy models and are combined by adding the outputs of the first step to the model of the second step as features. Experiments show the method achieves a total accuracy of 81.9% in BC–PPI corpus[1] and the outputs of the first step can effectively prompt the performance of the PPI information extraction.

The rest of this paper is organized as follows. We introduce Maximum Entropy model in section 2. Section 3 describes the data set used in this work. The proposed methods are detailed in section 4, followed by experiment results in section 5. Section 6 gives a brief conclusion.

2 Maximum Entropy Model

The motivating idea behind Maximum Entropy (ME) is that one should prefer the most uniform models that also satisfy any given constrains. ME model can easily combine diverse features, so it is widely adopted in many natural language processing and text mining tasks, such as Part of Speech(POS) tagging [14], named entity recognition [15] and relation extraction [16].

[1] BioCreAtIve-PPI corpus: http://www.informatik.hu-berlin.de/˜hakenber/corpora/

The maximum entropy is defined as:

$$p(c|x) = \frac{1}{Z} \prod_{j=1}^{K} \alpha_j^{f_j(c,x)} \tag{1}$$

$$Z = \sum_{i=1}^{N} p(c|x) = \sum_{i=1}^{N} \prod_{j=1}^{K} \alpha_j^{f_j(c,x)} \ . \tag{2}$$

where c is the outcome label, x is the given observation. Z is a normalization factor. N is the number of outcome labels. $f_1, f_2, ..., f_K$ are feature functions and $\alpha_1, \alpha_2, ..., \alpha_k$ are the model parameters. Each parameter corresponds to exactly one feature and can be viewed as a "weight" for that feature. All features used in the maximum entropy model are binary, which is defined as:

$$f_j(c,x) = \{ \begin{matrix} 1, & if\ c = true\ \&\ currentword = \text{“}Gal4\text{”} \\ 0, & otherwise\ . \end{matrix} \tag{3}$$

The parameters estimation process or training process in ME model is to maximize the likelihood of the training data $\prod_{i=1}^{M} p_\alpha(c|x)$, where M is the number of training instances and α stands for the model parameters. In the test process, the label c which has the highest probability calculated by Eq.1 will be choose as the label of the test observation.

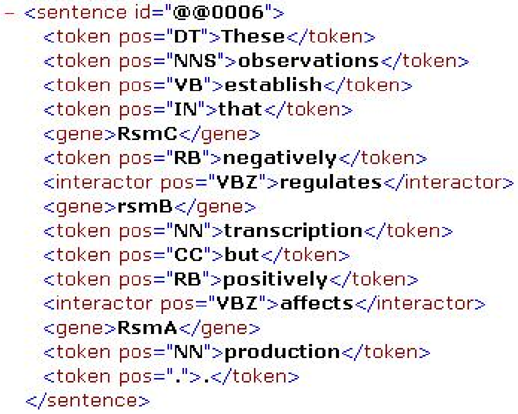
In PPI information extraction task, c is either "+"$(true)$ or "−"$(false)$ according to whether the current protein pair has interaction relationship. x is the given sentence, f_j is a feature function. We used Zhang's maximum entropy package[2] in our approach.

3 Data Set

The BC–PPI corpus was selected to evaluate the proposed protein-protein interaction extraction method. This corpus consists of 1,000 sentences randomly selected from biocreative 2004 task 1a corpus [17]. All the gene/protein (hereafter, simply use protein) names and interaction words in each sentence are annotated by experts, other tokens in the sentence are also annotated with POS tags. An example of annotated sentence is shown in Fig.1. An additional file describes the PPI information contained in the BC–PPI corpus and can be used to identify those sentences containing PPI information. BC–PPI corpus contains 255 interactions in 173 sentences. If a sentence includes more than one interaction, all interactions will be counted. The distribution of protein names and interaction words contained in the BC–PPI corpus was analyzed (Tab.1).

We can found that 56.5% sentences contain less than 2 proteins names among 1000 sentences in BC–PPI corpus from Tab.1. Also, among 430 sentences which do not contain interaction words, 351 sentences contain less than 2 proteins names. Considering that at least two protein names and one interaction word

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent.html



```
- <sentence id="@@0006">
    <token pos="DT">These</token>
    <token pos="NNS">observations</token>
    <token pos="VB">establish</token>
    <token pos="IN">that</token>
    <gene>RsmC</gene>
    <token pos="RB">negatively</token>
    <interactor pos="VBZ">regulates</interactor>
    <gene>rsmB</gene>
    <token pos="NN">transcription</token>
    <token pos="CC">but</token>
    <token pos="RB">positively</token>
    <interactor pos="VBZ">affects</interactor>
    <gene>RsmA</gene>
    <token pos="NN">production</token>
    <token pos=".">.</token>
  </sentence>
```

Fig. 1. An example of annotated sentence in BC–PPI corpus

Table 1. Distribution of proteins names and interaction words in BC–PPI corpus ("interactor" means "interaction word")

# of proteins	0	1	2	3	4	5	6	>6
# of sentences	411	154	203	110	57	32	16	17
# of interactor	0	1	2	3	4	5	6	>6
# of sentences	430	290	162	72	31	11	3	1

are the minimum requirements for a sentence to contain PPI information, it can be concluded that at least 64.4% of the sentences contained in the BC–PPI corpus do not contain PPI information. Out of the 356 remaining sentences, only 173 sentences (less than 50%) contain PPI information. Thereby, it is a challenging task to determine whether a sentence contains PPI information in BC–PPI corpus.

4 Methods

In this work, PPI information extraction task is addressed as a binary classification problem. ME model is adopted to solve this problem. Our approach includes two steps. The first step was designed to judge whether a sentence contains PPI information and in the second step the PPI information is extracted from the given sentence. The whole sentence information can be helpful for PPI information extraction and the first step affords such information. So the outputs of the first step are used as features in the second step.

4.1 Whether a Sentence Contains PPI Information

Given a sentence, we want to know whether the sentence contains PPI information in this step. This is also a binary classification problem. If a sentence contains PPI information, the output label will be "+"; otherwise, "−" will be the output. The number of protein names and the number of interaction words in a sentence are used as features in the ME model for this task. When counting the number of protein names, simple anaphora resolution strategy is taken. The strategy is that if there is only a left parenthesis between two adjacent protein names, the two protein names will be counted once instead of twice. For the sentence in Fig.1, its values of features are: "# of protein names = 3" and "# of interaction words = 2".

4.2 PPI Information Extraction

At this stage, the target is to extract all PPI information in a sentence. Each pair of protein names in a sentence will be extracted, then ME model is used to classify them. The true PPI pairs will be classified into "+" class, and false PPI pairs will be classified into "−" class.

According to [12], the use of shallow lexical features contributes a large portion of performance improvements in contrast to the use of parsing or partial parsing information. So we only use words features (WF) and sentence features (SF) which are the outputs of the first step in our ME model for sentence level PPI information extraction.

Table 2. Feature value examples for protein pairs in the sentence of Fig.1

+ pr=RsmC pr=rsmB sf=1 hw=establish hw=that bw=negatively interactor=regulates pw=transcription pw=but
+ pr=RsmC pr=RsmA sf=1 hw=establish hw=that bw=negatively interactor=regulates bw=transcription bw=but bw=positively interactor=affects pw=production pw=.
− pr=rsmB pr=RsmA sf=1 hw=negatively hw=regulates bw=transcription bw=but bw=positively interactor=affects pw=production pw=.

The words features include the two protein names (pr), the interaction words (interactor) between two protein names and the surrounding tokens which consist of the tokens between two protein names (bw), the two tokens right before the first protein name (hw) and the two tokens right after the second protein names (pw). These surrounding tokens make a remarkable promotion in the performance of PPI information extraction in [12]. The SF is quite simple: if current sentence contains PPI information, a "sf=1" feature value will be added, otherwise a "sf=0" feature value will be added. For the sentence in Fig.1, it contains three protein names, so there will be three pairs of proteins needed to be judged. Tab.2 shows their formats in training data or test data.

5 Experiments and Results

All the experiments were conducted on BC–PPI corpus. The precision, recall, F-measure and accuracy are used to evaluate the performance. For one class, if we use ST to stand for the number of instances correctly labeled by the system, GT to stand for the number of ground truth instances and SA to stand for the number of instances labeled by the system, we have $precision = ST/SA$, $recall = ST/GT$, F-measure $= 2 * precision * recall/(precision + recall)$. The accuracy is the ratio of the number of correctly labeled instances for all classes and the total number of all instance.

5.1 Experiments for Step 1

We first conducted the experiment to evaluate the performance of step 1. Using the two simple kinds of features mentioned in section 4.1, the accuracy of 10-fold cross validation in BC–PPI corpus is 88.5%, which is much higher than 64.4% — the accuracy of the baseline mentioned in section 3. The precision, recall and F-measure of each class are shown in Tab.3.

Table 3. The results of first step

class	precision	recall	F-measure
+	0.701	0.584	0.637
−	0.916	0.948	0.932

5.2 Experiments for PPI Information Extraction

1877 protein pairs from BC–PPI corpus are extracted, which contains 255 positive examples (interaction pair) and 1622 negative examples (non-interaction pair). When training ME model, 127 positive examples and 182 negative examples are used. Generalized Iterative Scaling algorithm is adopted to do parameter estimation and the number of iteration is set to 100. The test set contains another 128 positive examples and 180 negative examples. We measured the performance using precison, recall, F-measure and total accuracy. The results are shown in Tab.4.

Table 4. The results of PPI information Extraction on BC–PPI corpus

Feature Set	class	precision	recall	F-measure	total accuracy
WF only	+	0.813	0.711	0.758	0.812
	−	0.812	0.884	0.847	
WF+SF	+	0.816	0.727	0.769	0.819
	−	0.868	0.801	0.833	
WF+true SF	+	0.819	0.742	0.779	0.825
	−	0.829	0.884	0.856	

In Tab.4, “WF only” means the feature set only contains words features mentioned in section 4.2. Using this feature set, we got the baseline results. The F-measure of positive class is 75.8%, which is much better than 56.0% in [8](Recall and precision are 42.7% and 81.3% respectively). This is also shows that machine learning based approaches for PPI information is better than parser-based approaches. “WF+SF” means the outputs of first step was added besides words features. The F-measure of positive class has a slight promotion with 1 percent under this feature set, which means the outputs of the first step are effect in improving the performance of the model. “WF+true SF” means the oracle outputs of first step was added besides words features, i.e. we use the true information got from the corpus. This feature set makes the F-measure of positive class increase to 77.9% from 75.8% with a promotion of 2.8%. Meanwhile the F-measure of false class also increases. So, whether a sentence contains PPI information which stands for whole sentence information is helpful for PPI information extraction.

5.3 Discussion

Currently, the F-measure of “+” class in the first step is 63.7%, which could be further prompted by involving more features. Also, our method could be augmented to judge whether a paragraph or an article contains PPI information, which is the Protein Interaction Article subtask (IAS) in BioCreative 2006[3]. Also from Tab.4, we found that step 2 can be benefit from the performance improvement of step 1. A problem of our method is that usually the amount of sentences not containing PPI information is much larger than that of sentences containing PPI information, which could make the results of our method bias to the class with large amount. This need to be solved in the future work.

6 Conclusion

In this paper, we proposed a two-step approach to extract PPI information from biomedical literature. Both of the two steps use a supervised learning approach —– Maximum Entropy model. The task of the first step is to determine whether a sentence contains PPI information. The second step is to extract PPI information from a sentence. The outputs of the first step will be used as features in the second step. Experiment results show that the outputs of the first step, which embody whole sentence information, can effectively prompt the performance of PPI information extraction. Our approach achieves an F-measure of 76.9% in the positive class, which is much better than 56.0% achieved by a parser-based approach in the same corpus.

Acknowledgments. We extend our special thanks to Jörg Hakenberg for his information about BC–PPI corpus. We also thank three anonymous reviewers for many helpful comments on an earlier version of this work. This work is supported by National Natural Science Foundation of China (60435020 and 60673019).

[3] http://biocreative.sourceforge.net/bc2_ppi_ias.html

References

1. Tsai, T.H., Chou, W.C., Wu, S.H.: Integrating Linguistic Knowledge into a Conditional Random Field Framework to Identify Biomedical Named Entities. Expert Systems with Applications. 30(**1**) (2006) 117–128
2. Joshi-Tope, G., Gillespie, M., Vastrik, I.: Reactome: a Knowledgebase of Biological Pathways. Nucleic Acids Research. **33** Database Issue (2005) D428–432
3. Bader, G.D., Betel, D., Hogue, C.W.: Bind: the Biomolecular Interaction Network Database. Nucleic Acids Research. 31(**1**) (2003) 248–250
4. Peri, S., Navarro, J.D., Kristiansen, T.Z.: Human Protein Reference Database as a Discovery Resource for Proteomics. Nucleic Acids Research. **32** Database Issue (2004) D497–501
5. Bunescu, R., Mooney, R., Ramani, A.: Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline. In Proc of BioNLP–2006. (2006) 49–56
6. Ramani, A., Bunescu, R., Mooney, R.: Consolidating the Set of Know Human Proteinprotein Interactions in Preparation for Large-scale Mapping of the Human Interactome. Genome Biology. 6(**5**) (2005) r40
7. Temkin, J.M., Gilder, M.R.: Extraction of Protein Interaction Information from Unstructured Text Using a Context-free Grammar. Bioinformatics. 19(**16**) (2003) 2046–2053
8. Jang, H., Lim, J., Lim, J.H.: Finding the Evidence for Protein-protein Interactions from PubMed Abstracts. bioinformatics. 22(**14**) (2006) e220–e226
9. Huang, M., Zhu, X., Hao, Y.: Discovering Patterns to Extract Protein-Protein Interactions from Full Biomedical Texts. Bioinformatics. 20(**18**) (2004) 3604–3612
10. Hao, Y., Zhu, X., Huang, M.: Discovering Patterns to Extract Protein-protein Interactions from the Literature: Part II. Bioinformatics. 21(**15**) (2005) 3294–3300
11. Mitsumori, T., Murata, M., Fukuda, Y.: Extracting Protein-Protein Interaction Information from Biomedical Text with SVM. IEICE Trans Inf & Syst. E89–D (2006) 2464–2466
12. Xiao, J., Su, J., Zhou, G.D.: Protein-Protein Interaction Extraction: A Supervised Learning Approach. In Proc Symp. on Semantic Mining in Biomedicine. (2005) 51–59
13. Berleant, D., Ding, J., Fulmer, A.W.: Corpus Properties of Protein Interaction Descriptions in MEDLINE. http://class.ee.iastate.edu/berleant/home/me/cv/papers/corpuspropertiesstart.htm. (2003)
14. Ratnaparkhi, A.: A Maximum Entropy Model for Part-of-Speech Tagging In Proc of the Conference on Empirical Methods in Natural Language Processing. (1996) 133–142
15. Chieu, H.L., Ng, H.T.: Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In Proc of the Seventh Conference on Natural Language Learning. (2003) 190–203.
16. Nanda, K.: Combining Lexical, Syntactic and Semantic Features with Maximum Entropy Models for Extracting Relations. In Proc of the ACL-2004 Interactive Posters/Demonstrations Session. (2004) 178–181
17. Yeh, A.S., Morgan, A., Colosimo, M.: BioCreAtIvE task 1A:Gene Mention Finding Evaluation. BMC Bioinformatics 6(Suppl 1) (2005)