BioPPIExtractor: A protein–protein interaction extraction system for biomedical literature

Zhihao Yang *, Hongfei Lin, Baodong Wu

Department of Computer Science and Technology, DaLian University of Technology, No. 2 LingGong Road, ShaHeKou district, DaLian 116023, China

Abstract

Automatic extracting protein–protein interaction information from biomedical literature can help to build protein relation network, predict protein function and design new drugs. This paper presents a protein–protein interaction extraction system BioPPIExtractor for biomedical literature. This system applies Conditional Random Fields model to tag protein names in biomedical text, then uses a link grammar parser to identify the syntactic roles in sentences and at last extracts complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations. Experimental evaluations with two other state of the art extraction systems indicate that BioPPIExtractor system achieves better performance.

Keywords: Conditional Random Fields; Link grammar parsing; Interaction extraction; DIP

1. Introduction

Along with the rapid expansion of biomedical literatures, the demand for efficiently extracting biomedical information from the huge amount of resources offers an excellent opportunity for biomedical text mining, i.e., the automatic discovery of biomedical knowledge. Among others, automatic extracting protein–protein interaction (PPI) from biomedical literatures helps to build protein relation network, predict protein function and design new drugs and, therefore, has become a research focus.

Existing PPI works can be roughly divided into three categories: Manual pattern engineering approaches, Grammar engineering approaches and Machine learning approaches.

Manual pattern engineering approaches employ shallow parsing with patterns to extract the interactions. The SUISEKI system of Blaschke, Andrade, Ouzounis, and Valencia (1999) uses regular expressions, with probabilities that reflect the experimental accuracy of each pattern to extract interactions into predefined frame structures. The GeneScene system (Leroy et al., 2003) extracts interactions using frequent preposition-based templates. The BioRAT system (Corney, Buxton, Langdon, & Jones, 2004) uses manually engineered templates that combine lexical and semantic information to identify protein interactions. Such manual pattern engineering approaches for information extraction are very hard to scale up to large document collections since they require labor-intensive and skill-dependent pattern engineering.

Grammar engineering approaches use manually generated specialized grammar rules that perform a deep parse of the sentences. Temkin and Gilder (2003) used a Context Free Grammar that is designed specifically for parsing biological text. Recently, extraction systems have also used link grammar (Grinberg, Lafferty, & Sleator, 1995) to identify interactions between proteins (Ding, Berleant, Xu, & Fulmer, 2003). Their approach relies on various linkage paths between named entities such as gene and protein names. The IntEx system (Ahmed, Chidambaram, Davulcu, & Baral, 2005) extracts interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations via link grammar.
Machine learning approaches have also been used to learn extraction rules. Marcotte’s (Marcotte, Xenarios, & Eisenberg, 2001) supervised learning text classification can decide PPI information which is mentioned in the text. Xiao, Su, Zhou, and Tan (2005) used Maximum-Entropy models to combine diverse lexical, syntactic and semantic features for PPI extraction.

One of existing problems in current PPI research is the lack of defined criteria for evaluating the PPI systems: researchers develop and test on their own corpus and, therefore, their results are not comparable. Blaschke and researchers develop and test on their own corpus and, therefore, their results are not comparable. Blaschke and Valencia (2001) recommend using DIP (Xenarios et al., 2000) as a way of evaluating biological IE systems, because it represents a realistic problem of practical interest to biological researchers. IE researchers can use their systems to extract protein–protein interactions, and then compare these with the records in DIP.

Corney et al. (2004) presented a PPI system BioRAT which used templates and achieved a recall of 20.31% and a precision of 55.07% on 389 interactions from the DIP database correspond to 229 Medline abstracts. On the same dataset, the IntEx system which used link grammar to identify interactions between proteins achieved a recall of 26.94% and a precision of 65.66%.

Most errors encountered in BioRAT and IntEx occur in protein name recognition stage. In BioRAT the precision errors caused by protein name recognition account for around two-thirds of all the precision errors (Corney et al., 2004) while in IntEx the errors caused by protein name recognition account for 45% of all the errors (Davis et al., 2004). The reason is that their methods are dictionary-based whose performance depends badly on the size and quality of the dictionary. Currently, the popular protein name recognition methods are machine learning techniques including HMM (Zhou & Su, 2004), SVM (Lee, Hwang, & Rim, 2003), MEMM (Finkel et al., 2004), CRFs (Settles, 2004), etc. For example, in BioCreative 2004 task 1A (Hirschman, Colosimo, Morgan, & Yeh, 2005) the best system (Finkel et al., 2005) obtained an F-score of 83.2% using a maximum-entropy based system incorporating a diverse set of features.

This paper aims to introduce CRFs-based protein name recognition method and evaluate its contribution to the overall PPI performance. Our experiment results show that introduction of this method indeed helps to improve the PPI performance.

The remaining part of this paper is organized as follows: Section 2 describes the BioPPIExtractor system and its key components. Section 3 presents the experiment and discusses the results. Section 4 concludes the paper.

2. System work flow

The BioPPIExtractor system work flow is shown in Fig. 1. The system consists of six main steps to extract interaction information from the input sentences: pronoun resolution, protein name recognition, interaction word recognition, link grammar parsing, complex sentence processing, and interaction extraction. The details are described in the following sections.

2.1. Pronoun resolution

Extracting PPI from text should take into account the resolution of pronominal references to entities since interactions are often specified through these references. Our anaphora resolution module currently focuses on third person pronouns and reflexives since the first and second person pronouns are frequently used to refer to the authors of the papers. In our pronoun resolution module, noun and noun phrase in text are identified using GENIA Tagger (Tsuruoka et al., 2005) which is specifically tuned for biomedical text such as MEDLINE abstracts and achieves an F-score of 98.20% on GENIA corpus (Kim, Ohta, Tateisi, & Tsujii, 2003). Then the nearest noun phrase that matches the number of the pronoun is considered as the referred phrase.

2.2. Protein name recognition

In biomedical domain protein name recognition remains a challenging task due to the irregularities and ambiguities in protein nomenclature. In BioCreative 2004 task 1A the best system obtained an F-score of 83.2% using matching and this score reduced to 74.3% using exact match (Tsai et al., 2006). In BioPPIExtractor, a CRFs-based protein name recognition method is used.

2.2.1. Conditional Random Fields

Bio-entity recognition can be thought of as a sequence segmentation problem: each word is a taken in a sequence to be assigned a label (e.g. protein, RNA, DNA or other). Conditional Random Fields are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine. Such models are well suited to sequence analysis. They have recently been applied to the more limited task of finding gene and protein mentions (McDonald & Pereira, 2005) with promising early results.

Let $\sigma = \langle o_1, o_2, \ldots, o_n \rangle$ be a sequence of observed words of length $n$. Let $S$ be a set of states in a finite state machine, each corresponding to a label $\in L$. Let $s = \langle s_1, s_2, \ldots, s_n \rangle$ be the sequence of states in $S$ that correspond to the labels assigned to words in the input sequence $\sigma$. Linear chain CRFs define the conditional probability of a state sequence given an input sequence to be:

$$P(s|\sigma) = \frac{1}{Z} \exp \left( \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{jk} f_k(s_{i-1}, s_i, o_i) \right)$$ (1)

where $Z$ is a normalization factor of all state sequences, $f_k$ is one of $m$ functions that describes a feature, and $\lambda_{jk}$ is a learned weight for each such feature function.
The training process is to find the weights that maximize the log likelihood of all instances in training data:

\[
LL(D) = \sum_{j} \log(P(s_j | o_j)) - \frac{\sum k \lambda_k^2}{2\sigma^2}
\]

The second term in formula (2) is a spherical Gaussian prior over feature weights. Once these settings are found, the labeling for a new, unlabeled sequence can be done using a modified Viterbi algorithm. CRFs are presented in more complete detail by Lafferty, McCallum, and Pereira (2001).

2.2.2. Feature set

Feature based statistical models like CRFs reduce the problem to finding an appropriate feature set. The following features are used in our CRFs model:

1. **Surface word features**: We use words themselves as features. All the words are lower-cased so that the dimension of features can be decreased and the loss of information can be compensated through its combination with other features.

2. **Orthographic features**: The purpose of this feature is to capture capitalization, digitalization and other word formation information.

3. **Prefix/suffix features**: The three and four characters’ prefix and suffix of each word is used as feature.

4. **Word shape features**: Word shapes refer to mappings of each word to a simplified representation that encodes attributes such as its length and whether it contains capitalization, numerals, greek letters, and so on. For example, capital letters are replaced with ‘A’, lower-case letters with ‘a’, digits with ‘0’, and all other characters with ‘x’. Thus “Varicella-zoster” would become Xx-xxx, and “CPA1” would become XXX0.

5. **Compound features**: To model local context simply, neighboring words in the window [-1,1] are also added as features.

6. **Part-of-speech features**: POS may provide useful evidence about the boundaries of biomedical entity names. Here GENIA Tagger is applied again.

7. **Keyword features**: Some words occur more frequently in the entity names. These words (we called keywords) such as “factor”, “receptor”, “site”, etc. can help to identify entity names. We automatically extract unigram and bigram keywords which occur more than 20 times from the training data.

8. **Boundary word features**: Quite a few annotation errors are partial match errors. So we extract the common boundary words which occur more than five times from the training data to alleviate the problem.

Trained on BioCreative 2004 task 1A training set, our CRFs model achieved an F-score of 81.32% (using relax matching) on BioCreative test set. The performance is improved to 83.7% via the exploitation of the contextual cues including bracket pair, heuristic syntax structure and interaction words cue.

2.3. Interaction word recognition

In BioPPIExtractor system, a sentence is considered to include a PPI only if the sentence has at least two protein names and an interaction word (e.g. “bind”, “down-regulate”, “interact” and so on). So interaction words in sentences should also be recognized. Our gazetteer for interaction words recognition contains a total of approximately 150 entries including interaction verbs and their variants (for example, interaction verb “bind” has variants such as “binding” and “bound”).

2.4. Link grammar parsing and complex sentence processing

Link grammar was first introduced by Sleator and Temperley (1991) to simplify English grammar with a context free grammar. The basic idea of link grammar is to connect pairs of words in a sentence with various links. Each word is viewed as a block with connectors coming out. There are various types of connectors, and connectors may point to the right or to the left. A link consists of a left-pointing connector connected with a right-pointing connector of the same type on another word. A valid sentence is one in which all the words are connected in some way. The parsing result of a sentence “Bovine PRION protein as a modulator of protein KINASE CK2 is described.” is shown in Fig. 2.

Link grammar parser has the ability to detect multiple verbs and their constituent linkage in complex sentences and can be used to resolve of complex sentences into their multiple simple sentence clauses which contain a subject and a predicate. The complex sentence processing model follows a verb-based approach to extract the simple clauses. A sentence is identified to be complex it contains...
more than one verb. A simple sentence is identified to be one with a subject, a verb, objects and their modifying phrases.

The link grammar parser used in BioPPIExtractor was developed by Grinberg et al. (1995). The parsers’ dictionary can also be easily enhanced to produce better parses for biomedical text. Owing to the dictionary, parser can recognize most words in biomedical domain.

2.5. Interaction extraction

Our interaction extractor module works like the one in IntEx system: it extracts interactions from simple sentence clauses produced by the complex sentence processing module. Link grammar considers a thorough case based analysis of contents of various syntactic roles of the sentences like their subjects (S), verbs (V), objects (O) and modifying phrases (M) as well as their linguistically significant and meaningful combinations like S–V–O, S–V–M, illustrated in Fig. 3, for finding and extracting protein–protein interactions. Only if a syntactic role (or meaningful combination) has at least two protein names and an interaction word can a protein–protein interaction be extracted. However, unlike IntEx, BioPPIExtractor do not consider extracting interaction from combinations of S–O and S–M since we found they would introduce many extraction errors.

3. Experiment and discussion

3.1. Corpus

We conducted experiments using the same dataset as the BioRAT and IntEx evaluation so that the results are comparable. For BioRAT and IntEx evaluation, authors identified 389 interactions from the DIP database such that both proteins participating in the interaction had SwissProt entries. These interactions correspond to 229 abstracts from the PubMed. Since we did not achieve the responding full papers, BioPPIExtractor was tested only on the abstracts.

3.2. Results

The interactions extracted by the BioPPIExtractor system were manually examined for precision and recall. The sensitivity of the system is given by the recall measure, calculated as the ratio of true positives to the sum of true positives and false negatives. Precision is a measure of correctness of the system, and is calculated as the ratio of true positives to the sum of true positives and false positives. If an interaction extracted by BioPPIExtractor is not found in DIP, it could be that (a) it is a false-positive example, reducing the precision of BioPPIExtractor; or (b) the interaction is missing from DIP. The latter case consists of interactions that are mentioned in papers, but have not been added to DIP. Like BioRAT and IntEx, We manually reanalysed these records with no reference to DIP but instead we counted how many of BioPPIExtractor’s predictions were correctly extracted from the text. Tables 1 and 2 present the evaluation results as compared with the BioRAT and IntEx systems.

Table 1 shows the recall from these abstracts by BioPPIExtractor, namely 41.62%, which is much higher than BioRAT (20.31%) and IntEx (26.94%). Table 2 shows the precision from these abstracts by BioPPIExtractor, 55.41%, which is a bit higher than BioRAT (55.07%) and but lower than IntEx (65.6%). However, in the term of F-score, the harmonic mean of precision and recall (F-score = (2PR)/(P + R), where P denotes Precision and R Recall), the performance of BioPPIExtractor (47.53%) is better than BioRAT (29.68%) and IntEx (38.2%).

3.3. Discussion

Confined to the complicity of natural language, extracting PPI interactions from biomedical literatures is a challenging task and it is difficult to achieve a good performance. A detailed analysis of the sources of all types of recall errors of BioPPIExtractor is shown in Table 3.

Table 1 Recall comparison of IntEx and BioRAT from 229 abstracts when compared with DIP database

<table>
<thead>
<tr>
<th>Recall results</th>
<th>BioPPIExtractor</th>
<th>IntEx</th>
<th>BioRAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Percent (%)</td>
<td>Cases</td>
</tr>
<tr>
<td>Match</td>
<td>164</td>
<td>41.62</td>
<td>142</td>
</tr>
<tr>
<td>No match</td>
<td>230</td>
<td>58.38</td>
<td>385</td>
</tr>
<tr>
<td>Totals</td>
<td>394</td>
<td>100.00</td>
<td>527</td>
</tr>
</tbody>
</table>

The total interaction number we obtained from Dr. David Corney (the author of BioRAT) is 394, a bit different from 389 (the number used in BioRAT evaluation). However, we do not know the reason why the number used in IntEx evaluation is 527 since we did not get into touch with the authors.
DIP contains protein interactions from both abstracts and full text. Since BioPPIExtractor system was tested only on the abstracts, the system missed out on some interactions that were only present in the full text of the abstract. This accounts for more than half of the all recall errors (53.91%). If those interactions are excluded, BioPPIExtractor can have a recall of 60.74%.

In addition, recall errors occur in all the PPI processing stages: pronoun resolution, protein name recognition, interaction word recognition, link grammar parsing, complex sentence processing, and interaction extraction. Among others, the number of errors generated in interaction extraction stage is the biggest (21.74%). The reason is that due to the complicity of the protein interaction expression it is rather difficult to compile the appropriate extraction rules and, therefore, many interactions are missed out. The errors generated in protein name recognition stage account for 7.83% while the number is about 60% and 45% in BioRAT and IntEx separately. This shows that with the introduction of CRFs-based protein name recognition method, the performance of protein name recognition module is significantly improved, which will correspondingly improve the overall PPI performance.

The errors generated in link grammar parsing and complex sentence processing account for 10.43%. Among them link grammar parser itself may make some mistakes. For example, when dealing with too long sentences link grammar parser can parse even very long sentences quickly, but with considerably reduced accuracy. In addition, the errors generated in pronoun resolution and interaction word recognition account for 2.61% and 3.48% separately.

The leading cause of precision errors is our not perfect extraction rules. As discussed above, due to the complicity of the protein interaction expression it is difficult to compile the proper extraction rules to extract protein–protein interactions. In addition, link grammar parsing and complex sentence processing caused some precision errors.

4. Conclusion

This paper presents a protein–protein interaction extraction system specially designed to process biomedical literature BioPPIExtractor. The distinguishing feature of BioPPIExtractor is that it introduces CRFs-based protein name recognition method, and extracts interactions via link grammar parser. Experimental evaluations of the BioPPIExtractor system with the state of the art systems – the BioRAT and IntEx indicate that BioPPIExtractor's performance is better.

We have shown that with the introduction of CRFs-based protein name recognition method BioPPIExtractor achieved better PPI performance than BioRAT and IntEx. However, its current limitations are also evident, as highlighted by the moderate performance gain in our experiment. In the next step, we plan to introduce machine learning techniques to replace extraction rules, which may hopefully further improve the performance of protein–protein interaction extraction.

Acknowledgements

This work is supported by grant from the Natural Science Foundation of China (No. 60373095 and 60673039) and the National High Tech Research and Development Plan of China (2006AA01Z151). We would like to thank to Dr. David Corney for sharing the evaluation datasets and results.

References


<table>
<thead>
<tr>
<th>Table 2</th>
<th>Precision comparison of IntEx and BioRAT from 229 abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision results</td>
<td>BioPPIExtractor</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
</tr>
<tr>
<td>Correct</td>
<td>543</td>
</tr>
<tr>
<td>Incorrect</td>
<td>437</td>
</tr>
<tr>
<td>Totals</td>
<td>980</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Analysis of types of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error cause</td>
<td>Error number</td>
</tr>
<tr>
<td>Pronoun resolution</td>
<td>6</td>
</tr>
<tr>
<td>Protein name recognition</td>
<td>18</td>
</tr>
<tr>
<td>Interaction word recognition</td>
<td>8</td>
</tr>
<tr>
<td>Link grammar parsing and complex sentence processing</td>
<td>24</td>
</tr>
<tr>
<td>Interaction extraction</td>
<td>50</td>
</tr>
<tr>
<td>No included in abstracts</td>
<td>124</td>
</tr>
<tr>
<td>Totals</td>
<td>230</td>
</tr>
</tbody>
</table>


