Profile-feature Based Protein Interaction Extraction from Full-Text Articles

Shilin Ding[1] Minlie Huang[1] Hongning Wang[1] Xiaoyan Zhu[1,*]

[1]State Key Laboratory of Intelligent Technology and Systems (LITS),
Department of Computer Science and Technology, Tsinghua University,
Beijing, 100084, China

Email: dingsl@gmail.com, {aihuang,zxy-dcs}@tsinghua.edu.cn, whn03@mails.tsinghua.edu.cn

ABSTRACT
Various methods have been proposed to extract genetic protein-protein interactions from abstracts. These methods are unable to specify the interactions in which molecules are physically related and fail to explore the abundant evidence all over the articles. In this paper, we present a method of mining physical protein-protein interactions by exploiting profile feature from full-text articles during our participation in the second task of BioCreAtIvE Challenge 2006. This method synthesizes the features from the whole article as the protein pair's profile to extract the physical interactions, and specifies the SwissProt AC of the molecules involved in the interaction to help biologists make use of the information of the molecules, such as the sequence and cross reference. Compared with the other methods' performance released in BioCreAtIvE 2006, our method has shown very promising results.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Biology and Genetics; I.5.4 [**Pattern Recognition**]: Applications – *Text Processing*

General Terms
Algorithms, Experimentation

Keywords
Protein-Protein Interaction, Text Mining, Information Extraction

1. INTRODUCTION
The study of Protein-Protein Interaction (PPI) is one of the most pressing problems. Characterizing protein interaction partners is crucial to understanding not only the functional role of individual proteins but also the organization of entire biological processes. In the past years, the high throughput technologies have generated large amount of information. However, the information is buried in millions of peer-reviewed literatures. Without efficient management, the biological knowledge in the literatures is of little use to the researchers. A lot of knowledge databases, such as BIND [1], IntAct [11], and MINT [28] have been constructed to this end, but it costs a lot of time and expense to manually review and extract the important information from the literatures. So, automatically mining protein-protein interactions from bioscience literature is crucial and challenging [16].

There are two types of protein interactions: *Genetic Interaction* which is functional relationship among genes revealed by phenotype of cell, and *Physical Interaction* which is interaction among molecules. The task we participated in BioCreAtIvE 2006 is focused on mining physical interactions from the text because the genetic interactions are 1) not direct (the interaction may be through signaling cascades), thus, 2) not always trustworthy for biologists [30]. The abstracts with concentrated and limited information from MEDLINE are not capable to provide enough information to accomplish this task, while the full-text articles are more comprehensive to provide the evidence, such as the biological experiment which verifies the existence of the physical interaction. So the major problem here is how to exploit the physical interactions from the evidence synthesized from the full-text articles.

Various methods have been proposed to extract protein-protein interaction. But most of them are focused on abstract and fail to differentiate the physical interaction from the genetic interaction. In this paper, we describe a profile-feature based method to mine physical protein-protein interactions by exploiting abundant features from full-text articles.

The paper is organized as follows: The related works are discussed in Section 2. Section 3 presents the method to recognize the protein molecule names in text and normalize to them to entries in SwissProt. The profile-feature based method to extract the physical interactions from the evidence of the whole article is discussed in Section 4. In Section 5, we show the experiment and evaluation. And we draw our conclusions and discuss the future work in Section 6.

2. RELATED WORK
The researches of exploiting the information from the full-text articles are limited due to full texts' availability and complexity. SGPE [27] used abstracts and full-text articles to extract gene and protein synonyms, and Yu reported that the system performs

BIOKDD'07, August 12, 2007, San Jose, California, USA.

* Corresponding author: zxy-dcs@tsinghua.edu.cn
Tel: 86-10-62796831 Fax: 86-10-62782266

better on full-text articles because the names are more frequently listed in full-text articles. Schuemie [24] in their study of information content in abstracts versus that in full-text articles argued that the information density is higher in abstracts but the information coverage is much greater in full-text articles which indicates that the IE tools will perform better with the various information resources in the full-text articles. And Natarajan [20] used text mining of full-text articles to help generate novel hypothesis for the guide of gene-relation detection experiment and argued that the full-text articles are more comprehensive than the abstracts. So, the previous studies showed that the full-text articles are more effective for the extraction of physically interacted protein pairs.

Various methods and systems [3, 5, 7, 9, 13, 14, 19, 21] have been proposed for protein interaction extraction, but few of them are focused on physical interactions by exploring the evidence synthesized from the full-text articles. One class of these approaches is based on machine learning models. For example, Craven [4] employed a Naïve Bayes Classifier to predict relations from sentences.

Another class of methods for relation extraction is rule-based or pattern-based. The simplest method of this category is to extract relations from co-occurrence of entities in sentences [6, 15]. This method generates high sensitivity but low specificity.

Pattern based methods adopt hand-coded or automated patterns and then use pattern matching techniques to capture relations. Ono [21] manually constructed lexical patterns to match linguistic structures of sentences for extracting protein interactions. Similar hand-coded pattern based systems were also proposed by Rindflesch [23] and Pustejovsky [22]. Such methods contribute high accuracy but low coverage, and moreover, the construction of patterns is time-consuming and requires much domain expertise. Methods which can learn patterns automatically for general relation extraction include SPIES [14], ONBIRES [13, 7], Chiang [3], and Daraselia [5]. Most of them take annotated texts as input, and then learn patterns semi-automatically (starting from some pattern seeds) or automatically. Most of these methods focus on extracting one specific type of relations and can only explore the information confined in one sentence.

The third class of methods analyzes the syntax structures and semantics of the sentences to extract the relations [9]. This method strongly rely on the Natural Language Processing techniques, such as dependence parse trees [18], to get the structure of a particular sentence. This method has promising performance and is able to extract deeper semantic relations from the text. But it is also focused on single sentence and fails to explore the evidence from the whole articles.

In this paper, we describe a method to mine physical protein-protein interactions by exploiting abundant features. A profile-feature based method is adopted to extract the physical interactions from the full-text articles. Every sentence where the candidate molecule pairs co-occur is considered as a piece of evidence. And the profile, which is defined as the representation of the pair's features all over the article, is constructed based on all of the evidence. Thus, the method is able to exploits the document-level information instead of focusing on the features on sentence level. Here, we use SVM for training and classifying.

Although the information from the whole article is exploited, another difficulty facing physical interaction extraction is how to recognize the molecules in the articles. Since the physical interaction is the interaction between molecules, the identified names should be normalized to entries in a standard database, such as SwissProt. Thus, the biologists can easily get the whole information of the molecules, such as the sequence and taxonomy information, or other abundant cross-reference information.

Previous Named Entity Recognition methods [8, 25, 26, 29] can find out the protein names, but fail to specify what exact molecules these names refer to. The statistical based method is the most prevalent method to recognize named entities in the text. It exploits abundant word form features and context features to train a model [29, 25]. It has promising performance and flexibility but needs a large scale of annotated corpus. The rule based method is fast and highly accurate in a specific domain, but costs a lot of efforts to construct the rules [8]. These two methods are unable to normalize the names to database entries because the lack of reference to protein database. And the dictionary based method has the potential to map the names to the database entries, but the previous ones are only focused on find out the names.

The difficulty is due to extensive ambiguity in names and overlap of names with common English terms [12]. The use of phenotypic description, the conventional abbreviations lead to various synonyms that are difficult to differentiate. Our Named Entity Recognition and Normalization (NER/N) method is a dictionary matching method based on the organism information from the full-text article. We curated the SwissProt database to boost coverage and accuracy of the terms in the database. Then various rules are applied to solve naming convention related problem. The organism information is used to improve the NER/N process in terms of both time and accuracy.

Our contributions in this paper include 1) the novel NER/N method based on the organism information from the full-text article to recognize the protein name and specify the corresponding entry in SwissPort; and 2) the profile-feature based method which exploits the evidence all over the article to extract the physical interaction. In comparison to the average performance of all the submitted runs in BioCreAtIvE 2006, our method shows promising results and is ranked top in the official evaluation.

3. NAMED ENTITY RECOGNITION AND NORMALIZATION (NER/N)

Different from traditional NER, this task requires the protein names be normalized to primary Access Numbers (AC) of SwissProt entries, not just find the original names in the text. The motivation of this task is to help biologists identify the exact molecule of the mentioned protein, so they can use other information of the molecule, such as the sequence and taxonomy, and cross-reference information like protein structure. The major problem here is how to associate the name in the article with the entry in SwissProt.

- First, the inconsistent naming conventions and various usages in text cause a lot of ambiguous terms. For example, *TCF*, *PAL*, and *PKB* may refer to different entities.

- Second, abbreviated terms, such as *p53*, may cause difficulty for normalization, although domain experts can infer from the context what molecules the author is discussing.
- Third, the same protein name is used to identify different molecules that are from the same or related gene but different organisms. For example, *PI3K* may refer to different molecules in mouse (**P42337**), human (**P42336**), bovine (**P32871**), produced by the same gene *PIK3CA*.
- Fourth, the same protein name is used to identify different molecules of different isoforms. For example, PI3K is referred to **Q8BTI9** which is the beta isoform of the protein in mouse, and **O35904** which is the delta isoform.

As shown in Figure 1, there are mainly four processes in this module: 1) database curation; 2) organism detection; 3) dictionary-matching based name recognition; and 4) normalized names disambiguation. The process is as follows:

1) The SwissProt is curated to incorporate gene names/synonyms and unify the written form;
2) Find all the organisms that are mentioned in the article, mark their positions as an index;
3) The organism list is used to filter out irrelevant SwissProt entries for the matching of current article;
4) The article is processed by the same unification rules and matched by the filtered entries;
5) Disambiguate the multi-mapped names by the organisms in the context

3.1 Database Curation

During database curation, two main procedures below are done to improve the quality and coverage of the terms in SwissProt database:

- Curate entry terms in the SwissProt entries. The gene names/synonyms, gene product names/synonyms of the same entry are included. Addition of gene names may cause ambiguity since a gene may encode several proteins.
- Unify the written form of the entry terms based on rules. The same rules are applied to articles to maintain consistency.
    1) Prefixes and suffixes which are not critical for entity identification are removed. For example, prefix *c*, *n* and *a* of PKC, known as Protein Kinase C, which mean *conventional*, *novel* and *atypical* respectively, are removed.
    2) Terms with digits or Roman/Greek numbers are transformed into a unified format: Alphabet + white space + digits. This rule implies such normalization: IL-2, IL2, IL 2→IL 2; CNTFR alpha, CNTFR A, CNTFR I→CNTFR 1.
    3) Terms not in abbreviated forms are converted to lowercases.

The curation helps to improve the coverage because the official SwissProt names are descriptive and too long to use in articles. And it also helps to solve the nonstandard writing habits due to the rule-based unification.

3.2 Dictionary Matching

After curation, there are totally 230,000 entries, and more than 1 million terms. Obviously, it is not feasible for all the terms to be used during dictionary matching with the articles. To improve computation efficiency, we first detect the organisms in an article, and then use the information to rule out irrelevant entries. Our assumption here is that physical interactions described in one article would belong to a limited number of organisms. The organism database used as the controlled vocabulary is NCBI taxonomy [31]. A dictionary matching method is used to detect organisms, and five most frequent organisms are left, marked with their positions in the article. When matching the articles with SwissProt to find the ACs of the protein names mentioned, only the entries belonging to these organisms are used.

In experiment, the matching process saves about three quarters

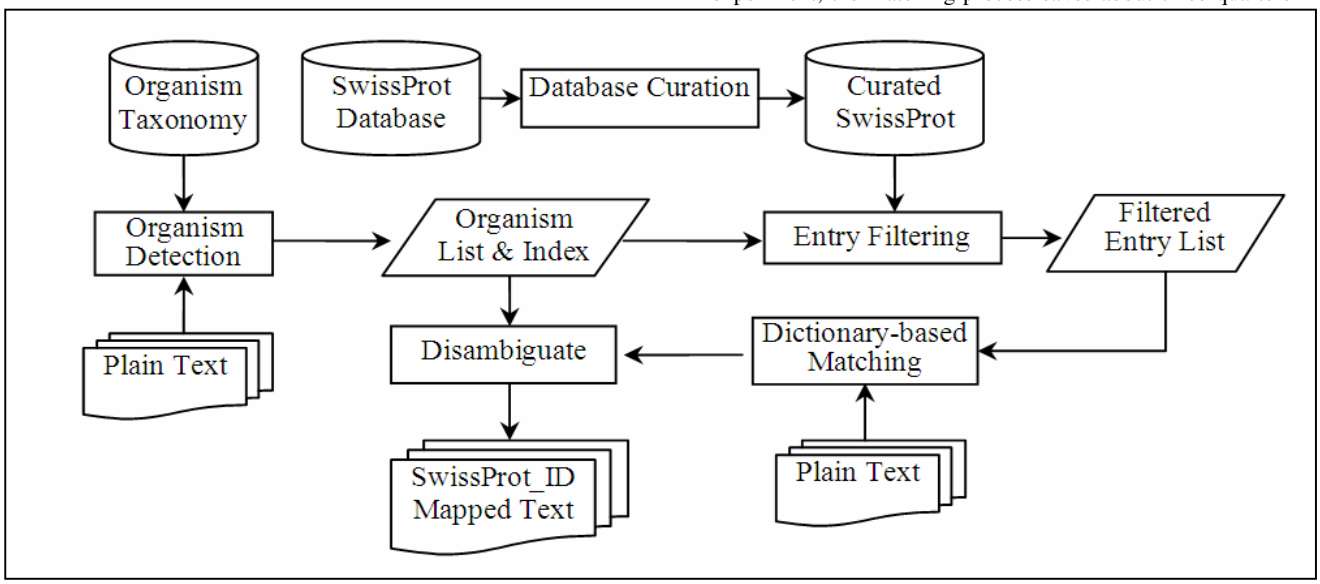

Figure 1: The Flowchart of NER. First, curate the terms in the SwissProt database; second, find the names and map them to SwissProt entries; and third, disambiguate the multi-mapped names by zone of control information from the organism contexts.

of the time due to the filtering. The time consumed by matching 740 articles with all entries is 460 minutes on a normal Pentium 4 2.0G processor. Through the filtering process before dictionary matching, the time is reduced to 125 minutes in the same condition.

3.3 Disambiguation

One protein name, particularly in abbreviated form, may correspond to multiple SwissProt entries. This is common in cases when the gene products in different organisms are similar (refer to the 3rd and 4th NER problems in Section 2). To solve the disambiguation, the principle of nearest neighbor is used, based on the organism's zone of control. The presumption here is that every protein name belongs to a particular organism's context. This context can be determined by the organism's zone of control (ZOC): beginning from the sentence that mentions the organism till the sentence that mentions another organism. When a multi-mapped name is met, we calculate which organism's zone the name belongs to based on the nearest neighbor rule, and filter out other maps to SwissProt entries with different organisms.

The disambiguation can't solve the isoform problems because the name is mapped to different isoforms that belong to the same organism. However the method is efficient because the isoform problems are not prevalent. We will see later in the experiment that this disambiguation method improves the precision greatly with only a little loss in recall.

From the discussion above, it can be inferred that our NER/N method outperforms other methods because: 1) carefully designed curation greatly improves the database's coverage and eliminates lots of naming inconsistency due to writing habit; 2) the dictionary matching method efficiently maps the name to the SwissProt entries based on the organism information from the full-text article.

4. PROFILE-FEATURE BASED EXTRACTION

Previous methods to extract protein interactions are based on sentence level, thus fail to synthesize the information from the whole articles. However, the topic-level interactions will be discussed at several places across the article, and these places will provide different sources of evidence, such as the experiment support and cross-reference evidence. The basic idea here is to extract interactions by using profile features derived from the whole document. The classifier is trained to make the decision based on the features all over the article. The profile-feature based extraction is more robust than pattern based extraction and other methods focused on the evidence from single sentence.

First, the goal is to extract physical interactions, so the single description as "*PTN1* binds to *PTN2*" does not necessarily indicate the existence of a physical interaction between *PTN1* and *PTN2*. However, if there is other evidence in the document, such as "The bind of *PTN1* to *PTN2* is determined by two hybrid screen", then the interaction is more probably to be true. So, different evidence will strengthen the validation of the physical interaction.

Second, the profile-feature based extraction is more robust when NER performance is far from satisfactory. The false positive protein names will falsely pair with other recognized names. But the pairs of the false positive proteins will be less statistically significant all over the document. Their profile features will be more random and less significant. For example, "The Y2H experiment proved the interaction between *PTN1* and *PTN2*, *CGA* ... …". The underlined term "*CGA*" that is the sequence of *PTN2* will be recognized as a protein, because *CGA* is the synonym of *Chromogranin A precursor*, which is P05059 in SwissProt. This false positive protein will be falsely paired with *PTN1* and *PTN2*. The previous method is hard to filter out the pair even though the pair only appears once in the article. However, the profile-feature based method is able to solve the problem by incorporate the evidence from the whole article.

4.1 Profile Feature

Profile features are selected to represent the evidence of a physical interaction. There are 3 types of profile features:

- 168 Unigram/Bi-gram Features
  100 of these features are selected by chi-square statistics of distinctiveness [18], and the rest 68 features are selected from Molecular Interaction (MI) ontology's [30] definition of Physical Interaction and Detection Method.
- 91 Pattern Features
  These features are generated in a semi-supervised manner [7]. These features have a form as "PTN * bind to * PTN", where PTN indicates a protein entity, and * means any word that can be skipped. The pattern feature is matched against the sentences as a regular expression.
- 2 Position Features
  One is whether the two proteins co-occur within the title; the other is whether they co-occur within the abstract.

These features eventually comprise a 261-dimensional feature vector, where each dimension is 1 or 0 indicating the presence or absence of a feature. Examples of these features are shown in Table 1.

Table1: Feature examples

Unigram/Bigram	Pattern
aggregation	activation of **PTN1* *by **PTN2*
crystallography	*PTN1* bind **PTN2*
elongation	*PTN1* *interact with **PTN2*
circular dichroism	*PTN1* *form complex with **PTN2*

4.2 Feature Construction

Every protein pair occurred within a sentence is viewed as a candidate. These sentences are considered as evidence. For each pair, *profile features* are extracted from all the sentences in which the pair appears. The corresponding bit is set as 1 if the feature is found in these sentences, see Figure 2. Through such a representation with abundant features, information from the whole document has been incorporated.

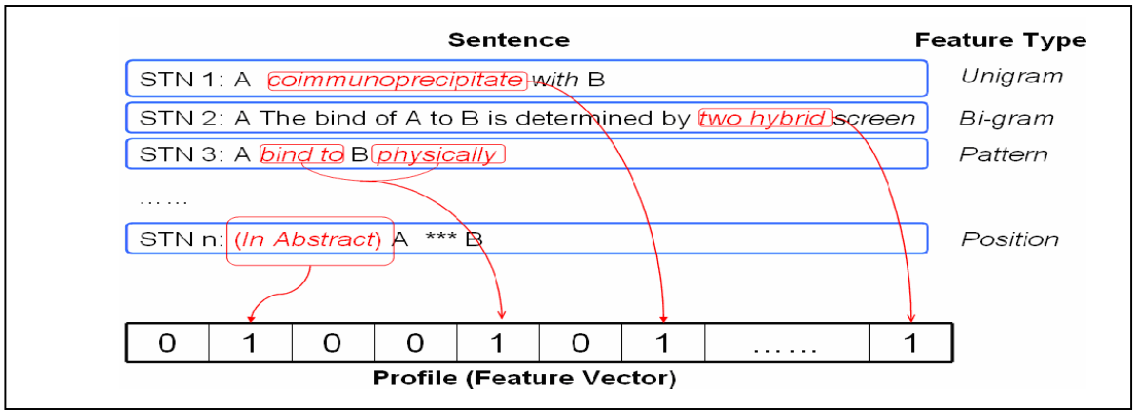

Figure 2: Feature Construction

4.3 Training

We use SVM-Light as our classifier [17]. In this part, we will discuss the construction of the training set.

The problem of the training corpus is that the supervised information is not given at the sentence level but only at the document level. The annotations from MINT and IntAct only specify the database ID (mainly SwissProt AC) of the interactors in the article, which means they do not provide the evidence texts that support the existence of the physical interaction, neither do we know where the interactors appear in the texts. So the annotation of the training corpus can not be used directly.

To establish the training set that the classifier can make use of, the protein names are first extracted and mapped to primary Access Number of SwissProt entries by our NER module. The protein pairs[1] which are annotated by domain experts are considered as positive samples. The other protein pairs in the text are treated as negative samples. Since lots of proteins are not part of a physical interaction, the number of negative samples overwhelms that of positive samples, which will lead to a biased distribution of training set. So from 740 training articles we randomly choose the negative samples twice as many as the positive samples and finally get 701 positive samples and 1402 negative samples as the training set for SVM.

5. EXPERIMENT AND EVALUATION

Data used in the experiments are introduced in Section 4.1. Evaluation methods are presented in detail in Section 4.2. The experiments of NER/N and Physical Interaction Extraction are discussed in Section 4.3 and 4.4. The evaluation results are officially published by BioCreAtIvE 2006.

5.1 Data Setup

BioCreAtIvE 2006 provided 740 full-text articles for training and 358 articles for testing from MINT and IntAct (The annotations of the testing articles are not released until the end of BioCreAtIvE 2006). These articles are manually annotated by database curators. The interaction pairs are only annotated from the full text articles in case there was an experimental confirmation for this interaction mentioned in the article.

[1] Protein Pair is defined as two proteins which co-occur in at least one sentence in the name-mapped text.

5.2 Evaluation

Due to the annotation methods applied by MINT and IntAct, the evaluation in BioCreAtIvE 2006 is different from previous evaluation of PPI extraction tools. Traditionally, the annotation will focus on one sentence and provide the *position* of the interactors and their *relations* (such as "induce" or "bind"). Thus the evaluation requires the exact match of these criteria to mark the result as true positive [13]. However, the current annotation in MINT and IntAct is focused on document level and provide the normalized database ID of the physically interacted proteins. So, the evaluation requires the detection of normalized interaction pairs of the document.

The evaluation for NER/N provided by BioCreAtIvE 2006 is also different from that of traditional NER task, because it only considers the physically interacted protein ACs as reference. So a lot of correctly recognized and normalized proteins are evaluated as false positive because they are not annotated as part of a physical interaction. Thus, the data of the evaluation can't represent the absolute performance of a NER/N module, but the comparison can reveal the difference of these NER/N methods.

5.3 Named Entity Recognition And Normalization (NER/N)

The performance of our NER/N module is shown in Table 2. The average results are calculated on 45 runs from 16 teams. Our performance is much better than the mean/median performance. From the comparison, it's obvious that our contributions to NER/N are database curation and organism-based disambiguation.

The curation will improve the database entries' accuracy and coverage, because the official names of the SwissProt entries are very long, descriptive and formal. The addition of synonyms and gene names will significantly increase the coverage. The unification of the various writing habits helps a lot to improve the matching accuracy. The F-score after database curation is improved by 77.3% compared to the naïve match.

The disambiguation based on organism information collected from the whole article greatly improves the NER/N's precision with slight loss in recall. The F-score is improved by 14.6% after disambiguation. Thus, the disambiguation by organism is efficient.

Although our method outperforms other methods (Our > Mean + Dev), the result is far from satisfaction. One problem is the wide spread synonyms which are hard to differentiate, such as PKB, Akt, and CGA. Another problem lies in the disambiguation. One protein name may refer to multiple entries in SwissProt, such as protein isoforms, which make the disambiguation method hard to handle.

Table 2: Overall performance vs. our overall results of NER/N

Score	Proteins normalized to SwissProt entries			
	Precision	Recall	F-score	
Mean	0.1495	0.2828	0.1707	
Std. Dev	0.0963	0.1294	0.0764	
Median	0.1337	0.2723	0.1683	Improv.
Naïve Match	0.2223	0.1024	0.1402	N/A
Prev. +Curation	0.2345	0.2648	0.2487	**+77.3%**
Prev. +Disambiguation	**0.3483**	**0.2410**	**0.2849**	**+14.6%**

5.4 Physical Interaction Extraction

To illustrate the effectiveness of profile-feature based method, we compare our methods with other methods submitted by other 45 runs from 15 teams in BioCreAtIvE 2006. Moreover, we adopt the results of pattern based method derived from ONBIRES [13, 7] as the baseline. The pattern based method learns lexicon-syntactic patterns describing interactions in a semi-supervised way: it first learns the patterns from large amount of unlabeled texts and then uses relatively small amount of labeled texts to select the candidate patterns. After that, the patterns are aligned against the sentences to extract interactions, where the matching score must exceed a pre-specified threshold. In this model, interactions are extracted at the sentence level. Thus, the approach is sensitive to the performance of NER which is far from satisfactory.

Table 3 shows the overall performance for both average results of all runs and our submitted results (two results by pattern based method, ONBIRE, and one result by profile-feature based method). It is worth noting that our results are much better than mean performance across all runs from all teams. And our system based on profile-feature excels others significantly (Our > Mean + 2*Dev) and is ranked top in the evaluation.

One reason for the whole system achieving higher performance is our effective NER/N module. To illustrate the contribution of profile-feature based method alone, we compare it with our pattern based method.

Profile-feature based model achieves the best results compared to the other two runs submitted by pattern based system, ONBIRES. These three results are achieved by the same NER/N module, so the NER/N does not impact the comparison of different extraction methods. It is obvious that the profile-feature based model contributes a much better precision than others. This is mainly because the model is more rational by synthesizing the evidence from the whole article, thus causes less false positive results.

So, the conclusion can be made from the evaluation that the profile-feature based method outperforms the traditional extraction methods, such as the pattern based method. The main advantage is that profile-feature is able to encode various features from the whole article. Because the task is focused on physical interactions, extraction methods which only exploit single evidence is prone to generating false positive results, while profile based method can incorporate lots of evidence and extract the semantic relations more rational.

6. DISCUSSION

To extract physically interacted protein pairs from the full-text articles has two major challenges: 1) recognizing protein named entities and mapping each entity to a unique entry in the SwissProt database; 2) identifying protein pairs which have been experimentally confirmed to have physical interactions. These challenges can lead to *Biologically Meaningful Knowledge*, which requires deeper understanding of semantic relations in the text.

First, NER/N is a most challenging task, and is obvious the bottleneck of the system. The difficulty to recognize and normalize the names to SwissProt entries is due to various synonyms and ambiguity in names. Database curation and organism based disambiguation are exploited as solutions. However, since the conventional naming of biomedical entities is far from standardized, the curation procedure lacks unified guides and fails to help the database to cover all the terms. Moreover, the normalization of the protein names to the unique entries in SwissProt database requires deeper understanding of the semantics buried in natural language. Future work will be focused on exploiting semantic information of the article for NER/N. The third problem is that the processing speed is not suitable for real-time application. We will try to speed up the NER/N process in the future by 1) indexing the protein terms in SwissProt and 2) dictionary matching by suffix tree.

Second, the profile based method is superior to previous ones because it incorporates evidence all over the article. However, one problem is that the model considers the article as a linear structure and misses a lot of useful information such as the positioning feature. The future work will focus on using more information from different regions of the full texts, such as the table/figure captions and cross-reference information to extract the interactions. Another problem is the lack of understanding of the syntactic structure and semantics of the sentence. This is a common problem because of the immature of Natural Language Understanding. We will try to develop novel method to capture the deeper semantics of the document by NLP techniques, such as the semantic lexicon/role defined in FramNet [2].

We believe that the text mining in biomedical area is to extract and manage the biological meaningful knowledge from the literatures. This knowledge can be used to integrate with the high-throughput experimental data for validation, hypothesis generation and biological discovery, and finally make the text mining really helpful to biologists.

Table 3: Physical interaction extraction performance averaged on 45 runs from 16 teams vs. our overall results. "Whole collection" means all the articles have been considered. "SwissProt only article collection" means articles containing exclusively interaction pairs which can be normalized to SwissProt entries have been scored.

Score	Whole collection			SwissProt only article collection		
	Precision	Recall	F-score	Precision	Recall	F-score
Mean	0.1062	0.1858	0.1035	0.1160	0.2000	0.1127
Std. Dev	0.0945	0.1001	0.0761	0.1035	0.1062	0.0836
Median	0.0755	0.1961	0.0788	0.0808	0.2156	0.0842
ONBIRES (th=0.0)	0.1373	0.2905	0.1579	0.1566	0.3189	0.1784
ONBIRES (th=80.0)	0.2177	0.2651	0.2039	0.2434	0.2828	0.2247
Profile-feature	**0.3096**	**0.2935**	**0.2623**	**0.3695**	**0.3268**	**0.3042**
Rank (in 45 runs)	**2**	**4**	**2**	**2**	**3**	**1**

7. ACKNOWLEDGEMENT

The work is supported by Chinese Natural Science Foundation under grant No. 60572084 and 60621062, National High Technology Research and Development Program of China (863 Program) under No. 2006AA02Z321, as well as Tsinghua Basic Research Foundation under grant No. 052220205.

8. REFERENCE

[1] Bader, G.D., Donaldson, I., Wolting, C., Quellette, B.F., Pawson, T. and Hogue, C.W. BIND –The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 29(1), 2001, pp. 242–245.

[2] Baker, C., Fillmore, C., and Lowe, J. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pp. 86-90, 1998

[3] Chiang, J.H., and Yu, C.Y. Literature extraction of protein functions using sentence pattern mining. *IEEE. Trans. On Knowledge and Data Enginerring*, 17 (8), 2005, pp. 1088-1098.

[4] Craven, M. Learning to extract relations from Medline. *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[5] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. Extracting Human Protein Interactions from MEDLINE Using a Full-Sentence Parser. *Bioinformatics*, vol. 20(5), 2004, pp. 604-611.

[6] Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. Mining medline: abstracts, sentences, or phrases? In *Proceedings of the 7th Pacific Symoisium of Bio-computing*, pp. 326–337, 2002

[7] Ding, S.L., Huang, M.L., and Zhu, X.Y. Semi-supervised Pattern Learning for Extracting Relations from Bioscience Texts. In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference*, pp. 307-316, 2007.

[8] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. Toward information extraction: identifying protein names from biological papers. In *Proceedings of the 3rd Pacific Symposium on Biocomputing*, pp. 707-718, 1998.

[9] Fundel K., Küffner R., and Zimmer R. RelEx - Relation extraction using dependency parse trees. *Bioinformatics*, vol. 23, 2007, pp. 365-371

[10] Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R., and Fluck, J. ProMiner: Rule-based Protein and Gene Entity Recognition. *BMC Bioinformatics* 2005, 6 (Suppl 1): S14.

[11] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. IntAct: an open source molecular interaction database. *Nucleic Acids Research*, vol. 32, 2004, pp. D452-D455.

[12] Hirschman, L., Yeh1, A., Blaschke, Y., and Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 2005, 6 (Suppl): S1

[13] Huang, M.L., Zhu, X.Y., Ding, S.L., Yu, H., and Li, M. ONBIRES: ONtology-based BIological Relation Extraction System. In *Proceedings of the Fourth Asia Pacific Bioinformatics Conference*, pp. 327-336, 2006.

[14] Huang, M.L., Zhu, X.Y., Hao, Y., Payan, D.G., Qu, K., and Li, M. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, vol. 20, 2004, pp. 3604-3612.

[15] Jelier,R., Jenster, G., Dorssers, L.C., van der Eijk, C.C., van Mulligen, E.M., Mons, B., Kors, J.A. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, vol. 21, 2005, pp. 2049–2058.

[16] Jensen, L.J., Saric, J., and Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, Vol. 7(2), 2006, pp. 119-129.

[17] Joachims, T. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods* - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[18] Manning, C., and Schütze, H. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press. 1999.

[19] Marcotte, E.M., Xenarios, I., and Eisenberg, D. Mining literature for protein-protein interactions. *Bioinformatics*, vol. 17, pp. 259-363, 2001

[20] Natarajan, J., Berrar, D., Hack, C., and Dubitzky, W. Knowledge Discovery in Biology Texts: Applications, Evaluation Strategies, and Perspectives. *Critical Reviews in Biotechnology*, vol. 25(1-2), 2005, pp. 31-52.

[21] Ono,T., Hishigaki,H., Tanigami,A., and Takagi,T., Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2), 2001, pp. 155-161.

[22] Pustejovsky, J., Castano, J., and Zhang, J. Robust relational parsing over biomedical literature: extracting inhibit relations. In *Proceedings of the seventh Pacific Symposium on Bio-computing*, pp 362-373, 2002.

[23] Rindflesch, T., Hunter, L., and Aronson, L. Mining molecular binding terminology from biomedical text. In *Proceedings of the AMIA Symposium*, Washington, D.C., pp. 127–131, 1999.

[24] Schuemie, M.J., Weeber, M., Schijvenaars, B.J., van Mulligen, E.M., van der Eijk C.C., Jelier, R., Mons, B., and Kors, J.A. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, vol. 20(16), 2004, pp. 2597-2604.

[25] Settles, B. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (*NLPBA*), Geneva, Switzerland, pp. 104-107, 2004.

[26] Yoshimasa, T., and Tsujii, J. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine,* pp. 41-48, 2003.

[27] Yu, H., Hatzivisaailoulou, V., Friedman, C., Rzhetsky, A., and Wilbur, W.J. Automatic Extraction of Gene and Protein Synonyms from Medline and Journal Articles. In *Proceedings of AMIA Symposium*, pp. 919-923, 2003

[28] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Letters*, vol. 513(1), 2002, pp. 135-140.

[29] Zhou, G.D., Zhang, J., Su, J., Shen, S., and Tan, C.L. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, Vol. 20 (7), 2004, pp. 1178-1190.

[30] http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI

[31] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy