Mining Physical Protein-protein Interactions from Literature

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

[1]Computer Science and Technology Department, Tsinghua University, Beijing,China, 100084

Email: Minlie Huang - aihuang@tsinghua.edu.cn; dsl05@mails.tsinghua.edu.cn; whn03@mails.tsinghua.edu.cn; Xiaoyan Zhu*- zxy-dcs@tsinghua.edu.cn;

*Corresponding author

Abstract

Background: Physical protein-protein interactions are fundamental to understand both the functions of proteins and the entire biological processes. Due to the development of high throughput experimental technologies such as the yeast two-hybrid screening, the interaction data are growing in an increasing speed. Manual curation which spends much time and cost could not keep up with the rapid growing amount of literature and the increasing number of newly discovered proteins. The need for text-mining tools to facilitate the extracting of such information is urgent.

Results: During the benchmark evaluation of BioCreative 2006, all of our results rank at the top three places. In the task of filtering articles irrelevant to physical protein interactions, our method contributes a precision of 79.95%, a recall of 89.33%, and an AUC of 87.46%, which is able to meet the demand of the practical interaction curation. In the task of identifying protein mentions and normalizing the mentions to molecule identifiers, our result (precision=34.83%, recall=24.10%, F-score=28.49%) is one of the best among all submitted runs. In the task of extracting protein interaction pairs, our profile-based method contributes the best result (precision=36.95%, recall=32.68%, F-score=30.42%).

Conclusions: We present a text-mining framework to extract physical protein-protein interactions. Three key issues, i.e., filtering irrelevant articles, identifying protein names and normalizing them to molecule identifiers, and extracting protein-protein interactions are studied in this paper. Our high-performance article filtering algorithms, organism-based protein names normalization and profile-based interaction extraction method

contribute the top three results in the benchmark evaluation of BioCreative 2006. The tool will be practically helpful for the manual interaction curation and greatly facilitate the process of extracting protein-protein interactions.

Background

An important step in functional systems biology is the understanding of the relationships between biomolecules. Interactions between proteins are crucial to biological pathways. The knowledge of the processes in which the proteins are involved is essential for a fundamental understanding of the cellular machinery. The study of protein interactions is one of the most pressing biological 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 [1] [2].

More and more interaction data are published in literature due to the development of high throughput experimental technologies, such as the yeast two-hybrid screening and affinity purification coupled with mass spectroscopy. These experimental techniques make it possible to study protein interactions on a much larger scale, although they suffer from the low accuracy. To provide reliable protein interaction data for biologists, interaction databases such as MINT [3] and IntAct [4], manually detect and curate protein interactions from various information sources; however, it is becoming difficult for database curators to keep up with the rapid growing amount of literature and the increasing number of newly discovered proteins. There are many challenges for the manual interaction curation. Firstly, the rate at which interaction data is being produced is increasing steadily due to the increased use of high throughput techniques for the detection of protein interactions. However, experimental methods are not equally reliable; the curators have to put strong emphasis on a thorough description of the experimental evidence for an interaction. Secondly, many authors continue to use ambiguous gene or protein names in publications or fail to provide the organism or tissue from which the genes or proteins originate. The failure to map gene or protein names to SwissProt/UniProt [5] [6] identifiers results in a heavy workload of an annotator to gather more information from references, supplemental materials and so on. Thirdly, many types of interactions scatter in literature, however, many of which are irrelevant to physical protein-protein interactions. Genetic

interactions, describing functional relationship among genes, are considered distinct from physical interactions between proteins, and are not currently curated. Other interactions such as drug-drug interactions, interactions between protein complexes and proteins, are not relevant to physical interaction curation.

Due to the accumulation of interaction data in biomedical literature and challenges for manual curation of interaction data, the need for text-mining tools to facilitate the extracting of such information is urgent. Particularly, the extraction of physical protein-protein interactions which are defined as the co-localization or direct interaction between protein molecules is becoming extremely important because physical interactions are the most reliable among many data produced by high-throughput experiments. The development of effective text-mining tools could aid the mapping of protein interactors to SwissProt/UniProt identifiers, the discovering of experimental evidence for interactions, as well as the discrimination of physical interactions from other types of interactions. In comparison to the prior work on bio-text mining, the competition BioCreative 2006 [7] [8] addresses such difficulties in normalizing names to molecules, discriminating physical interactions from other interactions, and gathering as much reliable experimental evidence as possible.

The task consists of several sub-tasks: 1) interaction article subtask (IAS) whose goal is to determine whether an article (only abstract) is relevant to some physical interactions; 2) interaction pair subtask (IPS) where interacting protein pairs should be extracted from full-text articles; 3) interaction sentence subtask (ISS) where participants are required to submit a sentence summary for each interaction; 4) interaction method subtask (IMS) where for each interaction, the experimental detection methods should be given. In this paper, we present methods and results during the participation in the protein-protein interaction task of BioCreative 2006 [9]. The contributions of our work are as follows:

1) By using *Kullback Leibler* divergence, we present the quantitative divergence between the training data and final test data in the interaction article subtask and point out the reason originates from not being able to provide an adequate set of irrelevant articles.

2) We propose solutions to overcome the above issue. In addition to the term features, other levels of features such as the string, entity, and template features are studied to reduce the distribution divergence. Information fusion from both the feature perspective and classifier perspective is studied. Our results rank at the first place in terms of accuracy and the second place in terms of AUC (area under receiving operator characteristic curve) in the benchmark evaluation. The performance is highly improved and this tool will be useful in the practical interaction curation.

3) We propose a named entity recognition framework which utilizes the organism information in articles. Further we present the quantitative analysis of how the extraction of physical interactions is influenced by the errors caused by the named entity recognition module. We point out the framework is extremely important because in the interaction curation task protein names need be normalized to molecule identifiers and therefore molecular properties such as sequence can be easily identified.
4) A profile-based method is proposed in the interaction pair subtask. It is not always possible or easy to identify a single sentence that clearly describes an interaction in a paper. In many cases the evidence for an interaction is dispersed throughout multiple sentences in the full-text articles. Inspired by the curation experience, we construct for each candidate interaction a profile vector from the whole article. By integrating evidence from the whole article, a more reliable prediction is achieved for robust interaction curation. Our results rank at the first place in terms of F-score in the official test.

Results and Discussion

Article filtering for efficient interaction curation

One of the most useful text-mining tools for interaction curation is to rule out articles irrelevant to physical interactions. According to the reports from database curation projects, it takes 2-3 hours to finish a paper for even highly qualified curators [1]. Obviously, high-performance article filtering tools will reduce a large amount of time and cost. The IAS of BioCreative 2006 addressed this issue precisely. The task is difficult because 1) some articles can not be determined relevant or irrelevant given only abstracts, and curators usually figure out evidence from full texts; 2) articles describing genetic interactions are hard to separate from those with physical interactions; 3) irrelevant articles are distributing much more broadly and randomly than relevant ones, leading to the difficulty in computational modeling.
There are 3536 articles relevant to physical interactions and 1959 irrelevant ones in the training dataset. The official test dataset consists of 375 relevant and 375 irrelevant articles. The serious problem in this task is that the performance on the training data is much better than that on the official test data (0.95 vs. 0.80 in terms of F_1-score), which is also reported by [10]. To analyze the problem, 750 articles (375 positive) are randomly taken out of the training corpus as the leave-out dataset. The top 50 features whose significance is measured by the chi-square statistics, is selected from the remaining training dataset. Based on the 50 features, three probability distributions are estimated on the leave-out dataset, remaining training dataset and official test dataset respectively, by using Eq. 3. Then we compute the average Kullback Leibler divergence (defined by Eq. 4) between two distributions to measure how different two

distributions are. The results are shown in Table 1 where the term features are unigrams/bigrams and the string features are strings with 7 characters.

For $Pr(x|c_+)$, the probability of a feature x occurring in the relevant articles, there is no significant difference between term distributions estimated on the leave-out dataset, remaining training dataset, or official test dataset. In other words, the three different datasets have almost the same term distribution. However, the case is significantly different for $Pr(x|c_-)$ (the probability of a feature x occurring in the irrelevant articles) whose distributions are illustrated by Fig. 1. There is a much larger divergence between the distribution estimated on the official test set and that on the training data set (0.992 vs. 0.188). We conjecture that there is a different term distribution on the official test set, and this may be the reason why the model degraded markedly on the official test dataset. This was also verified by the experiments from [10], where much better performance was obtained if the training corpus and final test dataset were reversed. When the string is selected as features, the divergence is much less (0.992 vs. 0.188). This might explain why the string feature even excelled the term feature in these runs as shown in Table 2.

We may conclude here the problem originates from the irrelevant articles which can not represent the entire sample space sufficiently. In the interaction curation task, irrelevant articles distribute more randomly, where some articles are describing genetic interactions which are very similar to physical interactions, some are discussing other types of interactions (e.g. drug-drug interactions), and some are totally dissimilar and can be easily filtered out. It is difficult to provide a good set of representative irrelevant articles. In other words, these irrelevant articles bring more uncertainty and bias to the learning machines.

We proposed several solutions to overcome the problem. The first attempt is to select strings as features based on the previous analysis. Further, we propose a new feature computing schema (defined by Eq. 5) to reduce the distribution divergence between the training data and test data. The new schema takes into account the probability of a feature observed both in relevant articles and in irrelevant ones, instead of simply using *TF*IDF*. The second attempt is to incorporate more high-level semantic features such as the named entity features, and template features. Entities including the *protein*, *DNA*, *RNA*, *cell line*, and *cell type* are recognized by using ABNER [11]. Term frequency (TF) is calculated for these entity features. Template features are exploited, to represent the specific syntactic dependency among entities. The third attempt is to integrate more information together from different classifiers. By fusing the various description powers of different classifiers, the performance can be highly boosted.

We firstly studied how the features influence the classification performance. The classification model is the SVM with a linear kernel. The classification results are shown in Table 2. The string features easily defeat

the term features due to the distribution divergence between the training data and test data, as mentioned in the above analysis. Note that the entity features are attractive since a very high recall (0.96) is obtained, indicating that almost all the original relevant articles have been mined out. This is very useful if the precision of the classification can be improved in the further process. By integrating all the features together, the AUC is further improved to 86.08%.

Secondly, we studied how the issue is influenced by different classification models. Each model has a different point of view on the data. SVM learns discriminatively to separate data by a decision hyperplane, while the Naïve Bayes classifier and multinomial classifier estimate probability distributions and try to interpret data from the probability perspective. The linear kernel SVM requires the data being represented as feature vectors, while the p-spectrum kernel SVM simply views an example as a string. The different description powers can be combined by AdaBoost in this manner [12]. The best performance, a precision of 80% and a recall of 90%, are approaching the needs of practical usage.

Normalizing protein names to SwissProt identifiers

It is extremely useful to normalize protein names to molecule identifiers, which will ease the process of interaction curation to a large degree. However, the task is challenging because inconsistent naming terminologies are used. To name a few, non-standard abbreviation terms, protein mentions without specifying species or organisms, and mentions without specifying definite isoforms, are very common. The following are some examples:

(1) Common terms, such as *p53*, are not easy to be normalized without any contextual information.

(2) The same term is used to name different molecules that are from the same or related genes but different organisms. For example, *PI3K* may refer to different molecules in mouse (*P42337*), human (*P42336*), bovine (*P32871*), encoded by the same gene *PIK3CA*.

(3) The same term is used to name 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.

Two important steps are indispensable to the normalization of names to SwissProt identifiers. First, curate the terms of database entries to canonical forms and use the new terms to detect protein mentions. Second, disambiguate multiple mapping of protein mentions to molecule identifiers using the organism and contextual information. The following rules are used to curate database entries:

(1) The gene names/synonyms, gene product names/synonyms of the same entry are included.

(2) Prefixes and suffixes which are not crucial for entity identification are removed. For example, prefix c, n

and *a* of *PKC*, which mean *conventional*, *novel* and *atypical* respectively, are removed.

(3) Terms with digits or Roman/Greek numbers are transformed into a unified format: Alphabet + white space + digits. This rule implies such examples: *IL-2*, *IL2* → *IL 2*; *CNTFR alpha*, *CNTFR A*, *CNTFR I* → *CNTFR 1*.

(4) Terms that are not in abbreviated forms are converted to lowercases.

About 23,000 normalized entries are produced from SwissProt database. As mentioned before, even with normalized terms, there are many ambiguous mappings to database identifiers. To solve the ambiguities, the principle of nearest neighbor is used, based on the organism context. The presumption here is that every protein name belongs to a particular organism context. Organisms in each sentence are identified. The organism context of a protein name is made up of organisms occurring at adjacent sentences. The organism of candidate proteins is determined by the nearest neighbor principle.

In the interaction pair subtask of BioCreative 2006, 740 full-text articles are provided for training and 358 for testing. These articles are manually annotated for the experimentally confirmed physical interactions. There is no separate evaluation for the protein name recognition and normalization. The results presented in Table 3 differ from the traditional evaluation of named entity recognition algorithms, because the gold standard set is only the protein molecules which have annotated interactions. Other correctly identified proteins but without interaction annotation are not included in the gold standard set. The average results are based on 45 runs from 16 teams. Obviously, our results are much better than the average results (ours>Mean+Dev).

Physical Interaction Extraction

The module of interaction extraction will greatly facilitate the process of interaction curation for database curators. It is not always easy to identify a single sentence that clearly describes an interaction in a paper. In many cases the evidence for an interaction is dispersed throughout multiple sentences in the full-text articles. However, most previous methods extract interactions at the sentence level [13–19], where each sentence is handled independently. Inspired by the fact that curators have to gather sufficient supporting evidence to decide whether a physical interaction is claimed in an article, we propose a profile-based method to extract physical interactions at the document level, by integrating evidence across the whole document. The results are shown in Table 4. Our method is obviously much better than others (Ours>Mean+2*Dev). In the benchmark evaluation, our results rank at the first place in terms of F-score, the second place in terms of precision, and the third place in terms of recall. Also, our method outperforms traditional

template-based methods (we believe ONBIRES represents the state-of-the-art performance [19]).

There are three reasons leading to the success of our profile-based method. First, the method is less sensitive to the errors caused by the protein normalization module whose performance is far from satisfactory. Second, by integrating evidence from the whole article, the method is more robust to extract physical interactions. For example, for sentence "*A interacts with B*", template-based methods will definitely take it as a positive example; however, it may describe a genetic interaction. In the profile-based method, other evidence is required to make such a claim. Third, abundant features such as the term features, entity features, template features, and position features are all integrated in the method. Here, we analyze the errors in detail to identify the problems hampering the overall performance. There are 798 manually annotated interaction pairs in the 358 test articles. 339 protein pairs are extracted, 100 of which are true positive pairs. There 8172 co-occurred pairs, many of which include incorrectly recognized names. The statistics is show in Fig. 2.

For the 239 false positive errors (area III), we manually checked the first 50 errors, which fall into 3 categories:

(1) 22 incorrectly normalized names. For example, in sentence "BAF60c interacts directly with PPAR gamma in vitro", the annotated interaction is *Q6STE5*(BAF60c)-*P37231*(PPAR gamma). We correctly extracted the interaction but unfortunately normalized the names to *Q6P9Z1-O19052*.

(2) 12 errors due to false positive names, where cases are in sentences where protein A and B physically interacted, a false positive recognized protein C coupled with A or B.

(3) 16 errors due to the classifier, which includes classifying non-physical interaction pairs as physically interacted, and other problems. This problem is partly due to the classification model and partly because of the incompleteness of the training set which doesn't provide the evidence of truly interacted samples.

Among the 698 false negative errors (area II+V), the majority, 532 errors, are caused by protein mention identification and normalization while 166 are due to the interaction extraction model. In these 166 co-occurred pairs, we found 37 are classified as negative because the co-occurred sentences do not contain sufficient evidence. Examples of these sentences are like A activates B or Camptothecin-induced nuclear export of A does not require B. Also, this problem is due to the fact that the evidence of physical interactions is not confined in a single sentence. The rest 129 errors are believed to be caused by our classifier.

From the above analysis, we conclude that the difficulty of protein name normalization leads to the majority of errors, although our module has relatively higher performance in all submitted runs. It caused

about 64% (34/50) false positive errors and 76.2% (532/698) false negative errors. The second problem is the incompleteness of annotation. Since the annotation only specifies the interacted protein ID in the article without the evidence passages or the location of these molecules, it makes not only the training process untraceable but also the process of error analysis extremely difficult. Currently, one major limitation of our method is the requirement of the protein co-occurrence within a sentence. This is not always true in the practical interaction curation where curators often find evidence from contextual sentences, each of which may contain only one protein of the interacting pair. For example, sentence "the two proteins are co-purifying together" definitely describes a physical interaction, however, both protein names appear in the preceding sentences instead of in this sentence. The problem can be solved by extending to neighbor sentences in our method.

Conclusions

In this paper, we have discussed three key issues in the practical interaction curation, i.e., filtering articles irrelevant to physical protein-protein interactions, identifying protein mentions and normalizing them to molecule identifiers, and extracting experimentally verified interactions. Different levels of features, including the string, term, named entity, and template features are exploited to study the problem of distribution divergence between the training data and test data. AdaBoost based information fusion technique is studied to integrate various description powers of different classifiers. Through these improvements, high-performance article filtering will greatly facilitate the process of interaction curation. Although the current state of protein name identification and normalization techniques still have huge room to enhance, our proposed method which utilizes the organism information to reduce ambiguities, outperforms the state of the art methods and will be helpful for biologists. The profile-based interaction extraction method combines evidence from multiple sentences across the whole document, and achieves much better prediction of physical interactions than other systems. In comparison to traditional methods that extract interactions at the sentence level, our method utilizes information from the whole article and it is less sensitive to the errors caused by the named recognition module.

There are still many difficulties and challenges in the extraction of biologically meaningful knowledge, for example, recognizing biological molecules with wide-accepted identifiers, and mining physical interactions with experimentally verified evidence. The paper contributes efforts to solve these issues from both the feature perspective and the classifier perspective.

Methods

In this part, we will firstly present the system architecture of our method (shown in Fig. 3), and then describe the models and algorithms used in our method. There are three major modules in our framework: the first for filtering irrelevant articles, the second for identifying and normalizing protein mentions to SwissProt identifiers, and the third for extracting protein-protein interactions.

Article filtering module

We studied three models in the article filtering module, i.e., the naïve bayes classifier, multinomial classifier, and SVM classifier [20]. All these classifiers require the prior selection of features to represent the data to be classified. As mentioned before, the amount of the term, string, and template features is very large, thus we use the chi-square statistics to select the most significant ones. The naïve bayes model for article filtering is defined as follows:

$$\begin{aligned} R_{NB}(d) &= \log \frac{\Pr_{NB}(c_+|d)}{\Pr_{NB}(c_-|d)} \\ &= \log \frac{\Pr(c_+)}{\Pr(c_-)} + \sum_{i=1}^{n} \log \Pr(w_i|c_+) - \sum_{i=1}^{n} \log \Pr(w_i|c_-) \end{aligned} \quad (1)$$

where d denotes an article, c_+/c_- denote relevant/irrelevant articles, w_i indicates a feature, and $R_{NB}(d)$ is the output score indicating the degree of relevance. The multinomial model implies a different distribution:

$$\begin{aligned} R_{MN}(d) &= \log \frac{\Pr_{MN}(c_+|d)}{\Pr_{MN}(c_-|d)} \\ &= \log \frac{\Pr(c_+)}{\Pr(c_-)} + \sum\nolimits_{i=1}^{n} x_i \log \Pr(w_i|c_+) - \sum\nolimits_{i=1}^{n} x_i \log \Pr(w_i|c_-) \end{aligned} \quad (2)$$

where x_i denotes the number of times that feature w_i appears in document d. In these two models, we have to estimate the probability of each feature in both relevant and irrelevant articles. This can be easily implemented by the below equation:

$$\begin{aligned} \Pr(w_i|c_+) &= \frac{1 + N(w_i, c_+)}{V + \sum\limits_{w_i} N(w_i, c_+)} \\ &= \frac{1 + \sum\limits_{d_j \in POS} TF(w_i, d_j)}{V + \sum\limits_{w_i} \sum\limits_{d_j \in POS} TF(w_i, d_j)} \end{aligned} \quad (3)$$

where V is the total number of features, POS is the set of relevant documents, and $TF(w_i, d_j)$ is the frequency of feature w_i observed in document d_j.

These two models are called *probabilistic* models, since they interpret data by estimating a probability distribution. As mentioned before, to analyze the quantitative difference of two distributions, we define the average *Kullback Leibler* divergence as follows:

$$AKL(q,p) = \frac{1}{2}\sum_{x}\left(q(x)\log\frac{q(x)}{p(x)} + p(x)\log\frac{p(x)}{q(x)}\right) \tag{4}$$

where p and q are two distributions. If two distributions are identical, the AKL is 0, otherwise positive. The SVM is a *discriminative* model, which constructs a hyper-plane in the feature space to separate data into categories. The classification decision is made by calculating the distance of a sample to the hyper-plane. In this module, we investigate two types of SVM models. The first one is a traditional SVM with the linear kernel, where each sample is represented as a feature vector. Instead of using *TF*IDF*, we proposed a new computing schema to overcome the issue of the distribution divergence between the training set and test set, as shown in the following:

$$TF * MLP(w_i, d_j) = TF(w_i, d_j) * \log\frac{\Pr(w_i|c_+)}{\Pr(w_i|c_-)} \tag{5}$$

where $TF(w_i, d_j)$ is the frequency of feature w_i observed in document d_j. Comparative results show that the computing schema is much better than *TF*IDF* in the benchmark evaluation. The decision variable in SVM model is

$$R_{SVM}(x) = b + \sum\nolimits_{i\in SV} y_i\alpha_i K(x_i, x), \tag{6}$$

where SV means the support vectors. The kernel function provides an alternative mechanism to represent data in a composite manner in addition to the feature-vector representation. As an instance, the p-spectrum kernel computes the number of common sub-strings shared by two input samples [21]:

$$K_p(x,y) \;=\; <\phi^p(x), \phi^p(y)> \;=\; \sum\nolimits_{u\in\Theta^p} \phi^p_u(x) * \phi^p_u(y) \tag{7}$$

$$\phi^p_u(x) = |\{(v_1, v_2)|x = v_1 u v_2, u \in \Theta^p\}| \tag{8}$$

where x and y are two strings (or documents) defined on the alphabet Θ, and Θ^p indicates all possible sub-strings of length p. In our method, we take p=7, which is about 1.5 times of the average length of unigrams. An article here is treated as a string, and no other semantics is considered. This low-level representation is able to reduce the distribution divergence between the training and test data.

Molecule recognition module

Identifying protein mentions and normalizing them to molecule identifiers is the necessary step to the extraction of protein interactions. Different from traditional named entity recognition tasks, this task

requires the submitted protein pairs be mapped into unique SwissProt identifiers, instead of presenting the original names in the text. We not only need to identify named entities but also map them to unique molecule identifiers. As shown in Fig. 4, there are four main processes in the molecule recognition module: database curation, organism detection, dictionary-based matching, and mapped name disambiguation. After curation, there are totally 230,000 protein identifiers, and more than 1 million terms. Obviously, it is not feasible if all the terms are used during the dictionary-based matching process. Moreover, the same terms, particularly abbreviations, may correspond to many protein identifiers. This is common in cases that the same gene products only differ in organisms. Thus the organism context is crucial to remove such ambiguities. We first detect the organism information in an article, and then use the information to rule out irrelevant database entries and further to remove ambiguities when terms are mapped to multiple protein identifiers. Our assumption here is that physical interactions described in one paper should be within only a few organisms. The organism database used here is the NCBI taxonomy [22]. A dictionary-based matching is used to detect organisms, and five most frequent organisms are left. Each sentence is linked with several detected organisms. To disambiguate mapping from identified names to molecule identifiers, the principle of nearest neighbor is used, implying that the organism of a recognized name is the organism of the nearest sentence where the name is observed.

Profile-based PPI extraction module

In the practical interaction curation, to extract experimentally verified physical interactions, curators usually collect evidence from multiple sentences. Previous methods to extract protein interactions are all at the sentence level, where each sentence is processed independently, and thus fail to synthesize the information from the whole article. Our profile-based method is able to exploit profile features from multiple sources of evidence across the whole document. For each candidate interacting protein pair, a profile vector is constructed from multiple sentences. In comparison to the traditional methods, the profile-based method is more robust and less sensitive to the local errors caused by the molecule recognition module.

Every protein pair co-occurred in a sentence is viewed as an interaction candidate. For each pair, profile features are calculated from all the sentences in which the pair co-occurs. The corresponding bit is set to 1 if the feature is found in these sentences (Fig. 5). Through such a representation, information from the whole document has been integrated together. A SVM with the linear kernel is trained on the profile feature vectors. There are three types of profile features:

(1) 168 unigram/bigram features. 100 of these features are selected by the chi-square statistics, and 68 are manually taken from the branches of Physical Interaction and Detection Method in the Molecular Interaction (MI) ontology [23].
(2) 91 template features. These features are generated in a semi-supervised manner [24]. They have a form as "$Protein_1$ * bind to * $Protein_2$", where * means that any word can be omitted.
(3) 2 position features. One is whether the two proteins co-occur within the title; the other is whether they co-occur within the abstract.
Our method is more robust than the traditional methods because: 1) the single description as "$Protein_1$ binds to $Protein_2$" does not necessarily indicate the existence of a physical interaction. However, if there is other evidence, such as "The bind of $Protein_1$ to $Protein_2$ is determined by Y2H", the interaction is more trustworthy. Obviously, more evidence will strengthen the confidence of the interaction; 2) our algorithm is more robust when the performance of the molecule recognition module is far from satisfactory. For example, in sentence "The Y2H experiment proved the interaction between $Protein_1$ and $Protein_2$, *CGA* ...", *CGA* that is the sequence of $Protein_2$ will be recognized as *Chromogranin A precursor*. Then it will co-occur with $Protein_1$ and $Protein_2$. The previous methods will fail though these false pairs are less statistically significant all over the document. Our method is able to solve the problem by incorporating the evidence from multiple sentences.

Authors contributions

The second author (Mr. Shilin Ding) implemented the modules of named entity recognition and interaction extraction. The third author (Mr. Hongning Wang) finished the algorithms and methods for the article filtering. Prof. Xiaoyan Zhu proposed many valuable suggestions and provided many supports in all the work.

Acknowledgements

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 and No. 053220002.

References

1. Chatr-aryamontri A, Ceol A, Licata L, Cesareni G: **Annotating molecular interactions in the MINT database**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:55–59.
2. Khadake J, Aranda B, Derow C, Huntley R, Kerrien S, Leroy C, Orchard S, Apweiler R, Hermjakob H: **IntAct - serving the text-mining community with high quality molecular interaction data**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:55–59.
3. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database**. *FEBS Letters* 2002, **513(1)**:135–140.
4. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff A P Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database**. *Nucleic Acids Research* 2004, **32**:D452–D455.
5. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Research* 2003, **31(1)**:365–370.
6. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, S YL: **UniProt: the Universal Protein Knowledgebase**. *Nucleic Acids Research* 2004, **32**:D115–D119.
7. Krallinger M, Valencia A: **Evaluating the detection and ranking of protein interaction relevant articles: the BioCreative challenge interaction article sub-task (IAS)**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:29–38.
8. Krallinger M, Valencia A: **Assessment of the second BioCreative PPI task: automatic extraction of protein-protein interactions**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:41–54.
9. Huang ML, Ding SL, Wang HN, Zhu XY: **Mining physical protein-protein interactions by exploiting abundant features**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:237–245.
10. Cohen AM: **Automatically expanded dictionaries with exclusion rules and support vector machine text classifiers: approaches to the biocreative 2 GN and PPI-IAS tasks**. In *Proceedings of the BioCreative Workshop: 22-25 April 2007; Madrid*. Edited by Krallinger M, Spanish National Cancer Research Centre 2007:169–174.
11. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text**. *Bioinformatics* 2005, **21(14)**:3191–3192.
12. Ratsch G, Onoda T, R MK: **Soft Margins for AdaBoost**. *Machine Learning* 2001, **42**:287–320.
13. Ono T, Hishigaki H, Tanigami A, Takagi T: **Automated extraction of information on protein-protein interactions from the biological literature**. *Bioinformatics* 2001, **17(2)**:155–161.
14. Marcotte EM, Xenarios I, Eisenberg D: **Mining literature for protein-protein interactions**. *Bioinformatics* 2001, **17**:259–263.
15. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CW: **PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine**. *BMC Bioinformatics* 2003, :4–11.
16. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I: **Extracting human protein interactions from MEDLINE using a full-sentence parser**. *Bioinformatics* 2004, **20**:604–611.
17. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C: **GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data**. *Journal of Biomedical Informatics* 2004, **37**:43–53.
18. Huang ML, Zhu XY, Payan DG, Qu KB, Li M: **Discovering patterns to extract protein-protein interactions from full texts**. *Bioinformatics* 2004, **20**:3604–3612.

19. Huang ML, Zhu XY, Ding SL, Yu H, Li M: **ONBIRES: ONtology-based BIological Relation Extraction System**. In *Proceedings of the 4th Asia Pacific Bioinformatics Conference: 13-16 February 2006; Taipei, Taiwan*. Edited by Jiang T, Yang UC, Chen YP, Wong LM 2006:327–336.

20. Joachims T: **Text categorization with support vector machines: learning with many relevant features**. In *Proceedings of 10th European Conference on Machine Learning: Chemnitz, Germany,1998* 1998:137–142.

21. Leslie C NW Eskin E: **A string kernel for SVM protein classification**. In *Proceedings of the Pacific Symposium on Biocomputing* 2002:566–575.

22. **The NCBI Entrez Taxonomy** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy].

23. **The Molecular Interaction Ontology** [http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI].

24. Ding SL, Huang ML, Zhu XY: **Semi-supervised Pattern Learning for Extracting Relations from Bioscience Texts**. In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference* 2007:307–316.

Figures

Figure 1 - The probability of a feature x occurring in irrelevant articles ($Pr(x|c\_)$) on different datasets (only 40 features are listed here)

The figure shows three distributions on the leave-out dataset, remaining training dataset, and official test dataset.

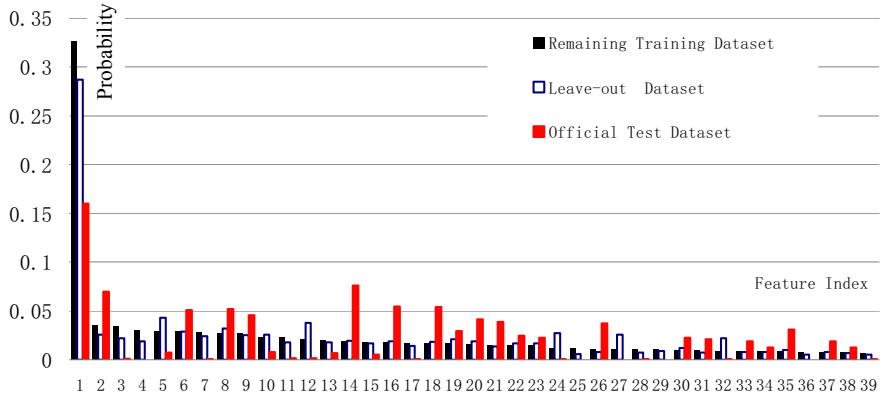

Figure 1: The probability of a feature x occurring in irrelevant articles ($Pr(x|c\_)$) on different datasets

Figure 2 - Errors of interaction pair extraction. Blue ellipse: 798 annotated Pairs; Yellow ellipse: 8172 co-occurred pairs; Green circle: 339 extracted pairs. I: 100 True Positive samples; II: 166 co-occurred but false negative samples; III: 239 False Positive samples; IV: 7135 True Negative samples; V: 532 false negative samples but never co-occurred.

The figure shows the distribution of errors in the interaction pair extraction.

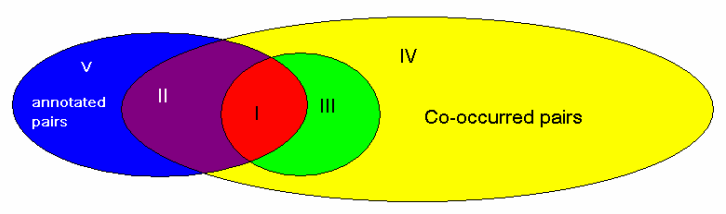

Figure 2: Errors of interaction pair extraction. Blue ellipse: 798 annotated Pairs; Yellow ellipse: 8172 co-occurred pairs; Green circle: 339 extracted pairs. I: 100 True Positive samples; II: 166 co-occurred but false negative samples; III: 239 False Positive samples; IV: 7135 True Negative samples; V: 532 false negative samples but never co-occurred.

Figure 3 - The system architecture of our method. PPI=Protein-Protein Interaction; MR=Molecule Recognition. Blue Rectangles are the three main modules in our system.

The figure shows the architecture of our system. There three main modules in the system, which have been colored as blue.

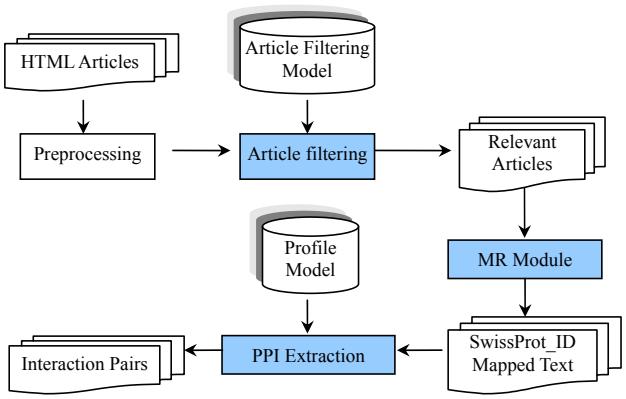

Figure 3: The system architecture of our method. PPI=Protein-Protein Interaction; MR=Molecule Recognition. Blue Rectangles are the three main modules in our system.

Figure 4 - The flowchart of the molecule recognition module. Gray boxes are the input of our molecule recognition module.

The figure illustrates the flowchart of the molecule recognition module.

Figure 5 - The profile vector in the extraction of interaction protein pairs.

The construction of the profile vector for each candidate protein pair is shown in this figure. The term feature (unigram/bigram), template feature, and position feature are used in this process.

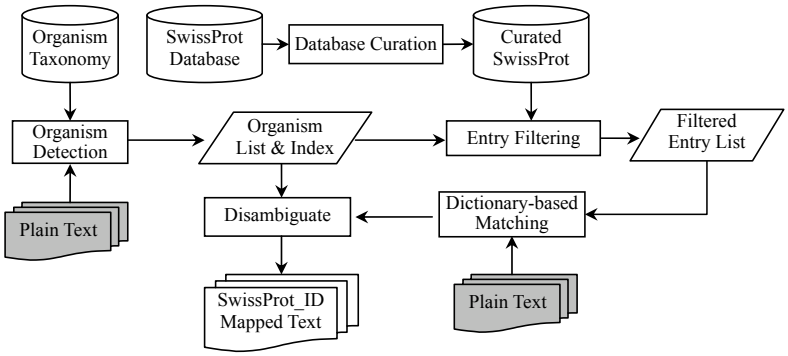

Figure 4: The flowchart of molecule recognition module. Gray boxes are the input of our molecule recognition module.

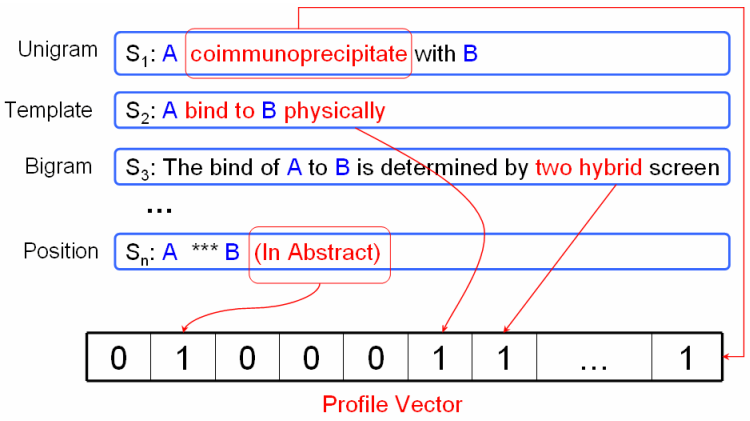

Figure 5: The profile vector in the extraction of interaction protein pairs.

Tables

Table 1 - The Average Kullback Leibler divergence between distributions on different data sets. Dis. = distribution

The table shows the AKL divergence among three distributions estimated on the leave-out dataset, remaining training dataset, and the official test data.

Table 1: The Average Kullback Leibler divergence between distributions on different data sets

Compared Distributions	Term Feature		String Feature	
	$Pr(x\|c_+)$	$Pr(x\|c_-)$	$Pr(x\|c_+)$	$Pr(x\|c_+)$
Dis. on the remaining training dataset vs. Dis. on the leave-out dataset	0.0216	0.0703	0.0029	0.0163
Dis. on the remaining training dataset vs. Dis. on the official test dataset	0.0369	0.9926	0.0357	0.1887

Table 2 - The performances of Article Filtering with different features and classifiers. AUC is the area under the receiving operator characteristic curve.

This table shows experimental results for the task of article filtering.

Table 2: The performances of Article Filtering with different features and classifiers.

Model	Precision	Recall	F_1-score	AUC
Mean	0.6642	0.7636	0.6868	0.7351
Standard Deviation	0.0810	0.1926	0.1035	0.0741
Term (baseline)	0.7016	0.8213	0.7568	0.8037
String	0.7044	0.8960	0.7887	0.8416
Named Entity (NE)	0.5815	**0.9600**	0.7243	0.7570
Template	0.7841	0.7653	0.7746	0.8239
String + NE	0.7360	0.8773	0.8005	0.8479
String + Template	0.7416	0.8880	0.8082	0.8372
String + NE + Template	0.7585	0.8373	0.7959	0.8507
String+Term+NE+Template	**0.7432**	**0.8720**	**0.8025**	**0.8608**
Naïve Bayes Classifier	0.6321	0.8613	0.7291	0.7884
Multinomial Classifier	0.6264	0.8720	0.7290	0.7770
Linear Kernel SVM	0.7016	0.8213	0.7568	0.8037
p-spectrum Kernel SVM (p=7)	.7352	0.8293	0.7794	0.8376
Integration of the above four classifiers (AdaBoost)	**0.7995**	**0.8933**	**0.8438**	**0.8746**

Table 3 - Comparative results for protein name normalization.

The table shows the comparative results for identifying and normalizing protein names.

Table 3: Comparative results for protein name normalization.

		Precision	Recall	F_1-score
Average	Mean	0.1495	0.2828	0.1707
	Std. Dev	0.0963	0.1294	0.0764
	Median	0.1337	0.2723	0.1683
Our method	Baseline	0.2223	0.1024	0.1402
	+Entry Curation	0.2345	0.2648	0.2487
	+Organism Context	**0.3483**	**0.2410**	**0.2849**

Table 4 - Comparative results for interaction pair extraction. "Whole collection" means all the articles have been considered. "SwissProt only article collection" include articles containing interaction pairs which can be normalized to SwissProt entries.

The table shows the comparative results for the extraction of interaction pairs.

Table 4: Comparative results for interaction pair extraction.

Compared Models	Whole Collection			SwissProt only article collection		
	Precision	Recall	F_1-score	Precision	Recall	F_1-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
Template-based method (th=0.0)	0.1373	0.2905	0.1578	0.1566	0.3189	0.1784
Template-based method (th=80.0)	0.2177	0.2651	0.2038	0.2434	0.2828	0.2247
Profile-based method	**0.3096**	**0.2935**	**0.2623**	**0.3695**	**0.3268**	**0.3042**

List of Abbreviations

TF	Term Frequency
IDF	Inverse Document Frequency
NE	Named Entity
NER	Named Entity Recognition
SVM	Support Vector Machine
AUC	Area Under receiving operator characteristic Curve
BioCreative	Critical Assessment for Information Extraction in Biology
IAS	Interaction Article Subtask
IPS	Interaction Pair Subtask
ISS	Interaction Sentence Subtask
IMS	Interaction Method Subtask