Finding the evidence for protein-protein interactions from PubMed abstracts

Hyunchul Jang¹,*, Jaesoo Lim¹, Joon-Ho Lim¹, Soo-Jun Park¹, Kyu-Chul Lee² and Seon-Hee Park¹

¹Bioinformatics Team, Electronics and Telecommunications Research Institute (ETRI), Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea and ²Department of Computer Engineering, Chungnam National University, Gung-Dong, Yuseong-Gu, Daejeon, 305-764, Korea

ABSTRACT

Motivation: Protein-protein interactions play critical roles in biological processes, and many biologists try to find or to predict crucial information concerning these interactions. Before verifying interactions in biological laboratory work, validating them from previous research is necessary. Although many efforts have been made to create databases that store verified information in a structured form, much interaction information still remains as unstructured text. As the amount of new publications has increased rapidly, a large amount of research has sought to extract interactions from the text automatically. However, there remain various difficulties associated with the process of applying automatically generated results into manually annotated databases. For interactions that are not found in manually stored databases, researchers attempt to search for abstracts or full papers.

Results: As a result of a search for two proteins, PubMed frequently returns hundreds of abstracts. In this paper, a method is introduced that validates protein-protein interactions from PubMed abstracts. A query is generated from two given proteins automatically and abstracts are then collected from PubMed. Following this, target proteins and their synonyms are recognized and their interaction information is extracted from the collection. It was found that 67.37% of the interactions from DIP-PPI corpus were found from the PubMed abstracts and 87.37% of interactions were found from the given full texts.

Availability: Contact authors.

Contact: janghc@etri.re.kr

1 INTRODUCTION

An uncountable number of protein-protein interactions are buried in research papers published thus far, and the number of papers published is growing continuously. Although there are stored data in verified databases such as BIND (Bader et al., 2001), KEGG (Kanehisa et al., 2002), SwissProt (Bairoch et al., 2000), and the Database of Interacting Proteins (Xenarios et al., 2001), these sources occasionally do not satisfy researchers. Even if the data is very useful, easily searchable and well structured, these databases nonetheless do not store the whole data, and most of protein interactions remain as unstructured text from scientific abstracts and full papers (Blaschke et al., 2001, 2002; Temkin et al., 2003). Moreover, most of the data exist only in the scientific literature. They are scattered in throughout the scientific literature and written in natural language. Accordingly, automated extraction information from the PubMed abstracts is preferable, and research that consolidates the set of known protein interactions using biomedical literature is necessary (Jenssen et al., 2001; Hirschman et al., 2002; Rzhetsky et al., 2004; Ramani et al., 2005).

In recent years, many researches have proposed to extract this information regarding protein interactions with automatic tools. However key issues such as the detection of protein names are not completely resolved with the use of such tools, thus they remain far from perfect (Blaschke et al., 2001, 2002).

Various techniques for recognizing protein names have been proposed. The use of standardized dictionaries containing the names and synonyms of proteins has been shown to be effective for recognizing these entities in text (Blaschke et al., 1999; Rindflesch et al., 1999, 2000). This technique remains limited as protein names not present in the dictionaries produce large amounts of false negatives. Others have proposed approaches using templates capable of recognizing common naming patterns for proteins (Fukuda et al., 1998; Ng et al., 1999; Yu et al., 2002). These techniques have also been shown to generate a large number of false positives by recognizing words that match the templates but are in fact not proteins. Alternative approaches have proposed machine learning methods (Proux et al., 1998; Hatzivassiloglou et al., 2001), and statistical methods (Krauthammer et al., 2000; Tanabe et al., 2002). Although these techniques have reported incremental gains in overall recall and precision over the template and dictionary based approaches, it has been shown that these techniques are also limited by the quality and extent of the training sets used to train the algorithms (Tanabe et al., 2002).

Similar to the limits inherent in the recognition of protein names, there have been various approaches published for extracting relationships from scientific literature. Several researches have shown that template and simple rule based algorithms can be used to extract interactions (Sekimizu et al., 1998; Blaschke et al., 1999; Ng and Wong 1999; Thomas et al., 2000; Friedman et al., 2001; Ono et al., 2001; Wong 2001; Pustejovsky et al., 2002). These approaches are, however, limited to a set of interactions by the predefined extraction rules or templates. Complicated cases are often solved by applying complex tools such as the use of domain-specific expert knowledge, which is often based on human annotation.

*To whom correspondence should be addressed.
missed by these approaches. Others have proposed the use of parts of speech analysis (Humphreys et al., 2000), and natural language based approaches (Rindflesch et al., 2000; Friedman et al., 2001). Huang et al., proposed a method for automatically generating patterns and extracting protein interactions (Huang et al., 2004; Hao et al., 2005). Bunescu et al., showed that various rule induction methods are able to identify protein interactions with higher precision than manually-developed rules (Bunescu et al., 2004). Ramani et al., used a set of 230 Medline abstracts manually tagged for both proteins and interactions to train an interaction extractor (Ramani et al., 2005). However, machine learning techniques are also limited by the quality and extent of the training sets used to train the algorithms.

A lack of standard common corpus, techniques and equations for reporting recall and precision has made comparative analysis of different approaches a difficult job (Hirschman et al., 2002).

Most of the current biological knowledge can be retrieved from the MEDLINE database, which now has records from more than 4,800 journals accounting for nearly 15 million articles. These citations contain thousands of experimentally recorded protein interactions. However, because of the large number of articles and the lack of formal structure, it is difficult to retrieve the data. A method to validate interactions from PubMed abstracts with the limits listed above is proposed.

2 METHODS

The present protein-protein interaction validation system consists of the following components, as shown in Fig. 1:

(i) A PubMed collector
(ii) A PPI extractor
(iii) A PPI validator

The abstracts collection component generates a PubMed query from the given two protein names and then collects abstracts from PubMed. The interaction extraction phase divides abstracts into sentences and recognizes protein names in sentences. Following this, sentences that have both proteins are selected, morphologically tagged and syntactically parsed after sentence simplification. As the last step of the extraction component, interactions between two proteins are extracted from the syntactically parsed sentences. The conflict resolution component detects false-positive interactions that were extracted, removes these false interactions, and decides whether the wanted interaction exists.

Brill’s transformation-based part-of-speech tagger\(^1\) (Brill 2002) was utilized, and was trained with the GENIA corpus\(^2\) (Kim et al., 2003). Its precision was 98.35% after training with the GENIA corpus and 83.73% with the WSJ corpus. The Stanford Parser\(^3\) version 1.4 with probabilistic context free grammar (PCFG) was also used.

2.1 PubMed abstracts collection

Simple queries for two proteins were generated in the forms of ‘‘A and B’’. In addition, two proteins A and B are expanded automatically with their synonyms. In the query step, users can add additional missed synonyms or abbreviations. The final query strings are in a form that resembles ‘‘(A or A1 or A2 or . . . or Aa) and (B or B1 or B2 or . . . or Bb)’’ under ‘‘A1’, ‘‘A2’, . . .

\(^{2}\)GENIA corpus: http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
\(^{3}\)The Stanford Natural Language Processing Group: http://www-nlp.stanford.edu/software/lex-parser.shtml

2.2 PPI extraction

The sentences are parsed syntactically and interactions are extracted from them. The result of a parser in the form of the Penn Treebank syntactic tags (Marcus et al., 1994) is then applied. Fig. 3 (a) is an example sentence, and Fig. 3 (b) shows the parsing result for it. This shows the syntactic tree structure and how the interaction is extracted between two proteins through the traversing of the tree. This is similar to finding a path between two leaf nodes.

Many existing full parsers that are not tuned to the biomedical domain frequently fail to parse, or the parsed results are often incorrect. This result occurs as most sentences in the biomedical literature are syntactically complex, or because words in sentences are tagged incorrectly. The sentence in Fig. 4 (a) is an example of this. This sentence has 43 tokens when the parentheses are tokenized and the minus symbols are not tokenized. To avoid this problem, sentences are made simple by the proposed method by substituting one word for complex words, i.e., protein names and nouns.

Protein names recognition The protein name extractor tags proteins using the words that were used for the PubMed query. Capitalized characters...
are ignored in the string matching step. Sentences those have both proteins are selected to extract interactions.

The use of dictionaries containing the names and synonyms of proteins has been shown to be effective for recognizing entities in free form text (Blaschke et al., 1999; Kindflesch et al., 1999). However, applications of this technique remain limited for the reason that protein names not present in the dictionaries produce large amounts of false negatives. This technique has reported high rates of recall and precision, and the proposed method relates to only two proteins at the validation step.

Making sentences simple Biomedical sentences are generally complex. One reason for this is that named entities in biomedical texts are usually not simple and consist of many words that have various morphological tags. This makes it difficult to parse biomedical texts.

One named entity can be divided into different phrases. This causes the structure to collapse. Accordingly the sentences are made simple by the following steps. First, recognized protein names are substituted with one predefined word. Second, noun phrases are substituted with one predefined word. Third, parenthesis phrases that are not a part of a named entity are removed. Following these steps, a more simplified sentence is created. The sentence in Fig. 4 (a) is changed to that in Fig. 4 (b). The new sentence now has 27 tokens. The parser can then process this sentence correctly. Lexicons were modified to tag substituted named entity words as NNPs and to tag substituted noun phrase words as NNS.

Extracting protein-protein interactions It is not straightforward to decide whether extracted paths between leaf nodes of a syntactic tree structure denote meaningful relationships. However, if two leaf nodes are proteins, it is easier to decide whether a relationship between them exists or not. If verbs or nouns are one of the predefined keywords, the extractor considers them as interaction events. The keywords are manually listed and based on the research by Temkin et al., and Hakenberg et al., (Temkin et al., 2003; Hakenberg et al., 2005). The keyword list was acquired via the Internet, from the homepage of Jörn Hakenberg5.

First, the extractor finds NP tags, and then checks whether NP belongs to any of these three cases: NP+VP, NP+PP or NP+CC+NP. Most of the interactions belong to one of these types; others usually belong to the following two cases. The first is similar to ‘is-a’ semantically, as in: ‘JAB phosphorylates JAK2’. JAB has recently been identified as a regulator of JAK2 phosphorylation and activity by binding phosphorylated JAK2 and inducing its degradation. This sentence contains ‘JAB phosphorylates JAK2’ information. The second is ‘is-a’, in: ‘CD38-associated Lck’. These two types are processed by a template-based method.

In Fig. 3, the sentence has NP and VP tags. ‘GAS41’ is the first noun of the first NP in the top NP and finds the first noun, NNP, in it. It finds the verb, VBP in the VP and ‘binds’ is extracted. Finally, it looks for PP after VBP and finds NP. ‘NuMA’ is extracted from the NP. This sentence presents in the form of NP+VP. ‘GAS41’ represents the NP phrase, ‘binds’ and ‘NuMA’ represents the VP and PP phrases in VP phrase. This is not a passive form and there is no negative expression. Therefore, ‘GAS41’ is the subject of the ‘binds’ event and ‘NuMA’ is the object.

In Fig. 5, a NP+VP structure is detected, and the PPI extractor finds ‘NED’ as a subject and ‘activated’ as an event. From the VP that has VBN ‘activated’, ‘INTERACTION’ is found as an object. Finally, the subject and object are exchanged due to the IN, ‘by’ and VBN tags.

Negative expressions can be extracted from the RB or DT tags in any phrase. Each type of required PP phrases was manually defined. Thus, the extractor continues to search for one more PP phrase after extracting ‘interac- tion’ and ‘FKBP12’, shown in Fig. 6. ‘RyR1’ and ‘IP3R1’ are found from the next PP phrase.
In all cases, a NNP-tagged protein is extracted as a subject or an object only when it is the first noun in the NP, and this NP is the first NP in its parent NP or PP. If a protein or a NP follows CC, they and their parent are available as a subject or an object.

2.3 Conflict resolution

Protein names recognition and their interactions extraction are very complex, and can be ambiguous. As a result of these processes, false positives in the extracted information may occur. In the proposed method, these are usually caused by parsing errors or rules for high recall.

Protein names conflict

One protein name may indicate more than one protein that is different in terms of species. The same string can signify another category. Therefore, it is necessary to confirm that subjects and objects of extracted interactions are the truly wanted proteins.

Conflicts in protein names are caused by false positively recognized names, such as different species or categories, abbreviations, inaccurate boundaries, or homonyms. For now, it is considered that identical strings indicate identical proteins and that there are no distinctions among species and categories.

Relation events conflict

For interactions, several types of interactions can be extracted as a false positive, or correctly extracted as two identical proteins, as shown in Table 1. The most critical conflict comes when they are opposites. Incorrectly extracted reverse interactions have to be removed.

Currently detected are only conflicts in which some interactions are positive and others are negative. In addition, interactions by type or polarity are not distinguished.
279 abstracts were collected from PubMed with a query ‘MEK1 and ERK2’, limited to only items with abstracts. The proposed system extracted 20 interactions between ‘MEK1’ and ‘ERK2’ in the abstracts. In the Yapex testing corpus, five interactions were extracted.

Due to the 20 extracted events, the proposed system can validate that interaction between ‘MEK1’ and ‘ERK2’ exist. However, understanding whether two phosphorylation interactions, ‘MEK1 phosphorylate ERK2’ and ‘ERK2 phosphorylate MEK1’, are in conflict is not easy to determine. In this case, interactions were correctly extracted when the experimental conditions were ignored. It is nearly impossible to decide that some interactions are not facts.

3 RESULTS AND DISCUSSION

3.1 Full parsing sentences
The Yapex corpus was selected to evaluate the effect of sentence simplification. The Yapex corpus is used for the purpose of evaluating named entity recognition methods. It consists of 99 abstracts for training and 101 abstracts for testing. 101 testing abstracts were utilized for the evaluation. The Yapex testing corpus has 962 sentences, including abstract titles. The number of sentences that have more than two protein names is 532. The parser processed 439 sentences, and did not process 93 sentences. The percentage of parsed sentences was 82.5% and the average number of tokens per sentence was 24.97. The percentage of failed sentences was 17.5% and the average number of tokens per sentence was 49.07.

After sentence simplification, the parser could parse additional 62 sentences, and only 31 of 93 sentences were left out. The average number of tokens was 26.15 in 501 sentences, and 53.38 words in 31 sentences. The parser success rate is higher when the morphological tags are given by the tagger. The precision of parsed results was not evaluated. However, 62 sentences (11.7%) could be parsed after simplification. This indicates that the sentences could be parsed more correctly.

3.2 Extracting protein-protein interactions
The BC-PPI corpus was selected in order to evaluate the proposed protein-protein interaction validation method. This corpus consists of 1,000 sentences, with annotated genes/proteins and interactions. It contains 255 interactions and 173 sentences contain at least one interaction. If a sentence includes more than one interaction, all interactions were counted as answers. Additionally, the present system tried to extract all.

The value of a recall was calculated to be TP/(TP+FN)×100, and the value of a precision was calculated to be TP/(TP+FP)×100. TP indicates the total number of interactions extracted correctly and tagged in the corpus, TP+FN indicates the total number of interactions tagged in the corpus, and TP+FP indicates the total number of interactions extracted correctly or incorrectly by the proposed method. The rate of recall and precision of extraction with the sentence simplification were 42.74% and 81.34%, respectively. The BC-PPI corpus has no negatively tagged interaction; hence any extracted negative interactions were excluded from TP+FP. The TP was 109, the TP+FN was 255, and the TP+FP was 134, as shown in Table 3. The proposed method was not evaluated without the sentence simplification. Extracted protein names can be scattered over the syntactic tree and the proposed interaction extraction method does not address this problem.

Some false positively extracted interactions were caused by parsing fail or error. A parsing failure indicates that the parser can not parse, and parsing error signifies that it does not parse correctly. The false positively extracted interactions are caused by a parsing error, as in: ‘We concluded that the two NF-IL6 sites mediate induction of IL-1 beta in response to the stimuli LAN, LPS, and TNF-alpha.’ The parser returned ‘two NF-IL6 sites mediate TNF-alpha’.

Most missed interactions are caused by semantic problems. The proposed extractor does not account for semantic relations; as well, and syntactic tags don’t indicate them. The following sentences are examples:

(1) ‘Receptor activation by the haematopoietic growth factor proteins interleukin 5 (IL-5) and granulocyte-macrophage colony-stimulating factor (GM-CSF) leads to phosphorylation of JAK2 as a key trigger of signal transduction.’

(2) ‘We analyzed the abilities of fibrillins and LTBPs to bind latent TGF-beta by their 8-Cys repeats.’

(3) ‘In vitro GAS41 bound to the C-terminal part of the rod region of NuMA.’

These sentences need to be handled semantically, or errors occur. For examples, The proposed system was not able to determine that ‘leads to phosphorylation of’ is equivalent to ‘phosphorylate’ in sentence (1), or that ‘the abilities of fibrillins to bind’ corresponds to ‘fibrillins binds’ in sentence (2). In addition, it did not determine that ‘to the C-terminal part of the rod region of NuMA’ meant that ‘to NuMA’ in sentence (3).

Although only a small number of interactions are expressed with anaphora terms, they were not analyzed, though unquestionably this should be addressed. The following sentence is an example of this.

(4) ‘Deletion of the binding site from MEK1 reduced its phosphorylation by ERK2, but had no effect on its phosphorylation by p21-activated protein kinase-1 (PAK1).’

---

Table 2. Parsing before and after sentence simplification

<table>
<thead>
<tr>
<th>Sentence simplification</th>
<th>Full parsing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>439(82.51%)</td>
<td>93(17.48%)</td>
</tr>
<tr>
<td>After NES</td>
<td>455(85.52%)</td>
<td>77(14.47%)</td>
</tr>
<tr>
<td>After NES+NPS</td>
<td>474(89.09%)</td>
<td>58(10.90%)</td>
</tr>
<tr>
<td>After NES+NPS+PPR</td>
<td>501(94.17%)</td>
<td>31(5.82%)</td>
</tr>
</tbody>
</table>

Table 3. Recall and precision for BC-PPI corpus

<table>
<thead>
<tr>
<th>TP+FN</th>
<th>TP</th>
<th>TP+FP</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>255</td>
<td>109</td>
<td>134</td>
<td>42.7%</td>
<td>81.3%</td>
</tr>
</tbody>
</table>

NES: named entity substitution, NPS: noun phrase substitution, PPR: parenthesis phrase removal.

---

Finding the evidence for protein-protein interactions from PubMed abstracts

3.3 Finding the evidences for PPIs

The DIP-PPI\(^8\) corpus was selected to evaluate the proposed validation method. The DIP-PPI corpus is based on protein-protein interactions from the DIP\(^9\), and is restricted to proteins from yeast. The full texts are included in the corpus, rather than the abstract only. DIP uses IDs from the SGD\(^10\) for nodes. The DIP-PPI corpus contains 297 interactions. For protein synonyms, the DIP synonyms\(^11\) from SGD of the DIP-PPI corpus were used, and a number of missed synonyms and aliases were added from the SGD Gene Names\(^12\).

20 interactions from the DIP-PPI corpus are composed of one protein. These are interactions in which the first partner and the second partner have the same SGD ID, and they were excluded from the validation.

An abstract vs. a full text vs. abstracts  In addition, 87 interactions are valid but the corpus contains no text for these. 107 interactions were totally excluded while 190 interactions were included to compare the effects of an abstract, a full text and abstracts for the interactions.

As shown in Table 4, from among 190 interactions, 166 interactions were extracted from the full text given in the corpus, with the rate of 87% as shown in Table 4 (B). When using only each abstract instead of the given full text for an interaction, only 83 interactions were extracted, as shown in Table 4 (A). When using all collected abstracts for an interaction, 128 interactions were extracted, as shown in Table 4 (C). These results show that using a number of collected abstracts for an interaction is more effective naturally compared to using an abstract, and less compared to the use of full text versions.

When abstracts collected from PubMed were used, no abstract was collected for 11 interactions, and no target interaction was extracted from the collected abstracts for 51 interactions. 13 from 51 had no sentence that had both proteins, and 38 from 51 had more than one sentence that had both proteins; however, no wanted interaction was extracted.

PubMed returned at least one abstract for 179 interactions, and no wanted interaction was extracted from sentences found in the PubMed abstracts. From seven invalidated interactions, more than forty abstracts were collected, but the wanted interactions were not extracted.

In real cases, a user can edit the proposed query for the PubMed collection. However, the query is generated from the given protein names automatically.

In case no relationship is extracted from sentences in which two proteins are present, the co-occurrence information may be useful in a statistical method. However, this was not calculated at this point.

Although more than thirty sentences in which both proteins were present were collected, the interaction between the two proteins could not be validated. Only 11 of 164 interactions were validated from more than thirty sentences. 153 of 164 interactions were validated in less than thirty sentences. This indicates that the validation possibility is not very dependent on the number of collected sentences in which both proteins were present.

From seven invalidated interactions, more than forty abstracts were collected, but the wanted interactions were not extracted. From 131 validated interactions, less than thirty abstracts were collected for each interaction. This signifies that the validation possibility is not overly dependent on the number of abstracts collected.

4 CONCLUSION

A PubMed abstract-based protein-protein interaction validation method is presented. The basic idea of this approach is that sentences in the biomedical literature are simplified after multi-word substitutions. Additionally, a normal full parser can parse these interactions were validated; however, this does not indicate that the only interaction in which the same abstract was given in the corpus could be validated.

Co-occurrence: found vs. not found  No abstract was collected by the query generated in this trial for 27 interactions, and at least one abstract was collected for each of the 250 interactions as shown in Table 5 (D).

In order to validate an interaction between two proteins, the proposed system has to find at least one sentence in which both proteins are present. Among the 250 interactions in Table 5 (E), 221 collections had at least one sentence in which both proteins were present. 164 of 221 interactions that have more than one sentence were validated as shown in Table 5 (F). 57 interactions were not validated from those sentences found in the PubMed abstracts.

In real cases, a user can edit the proposed query for the PubMed collection. However, the query is generated from the given protein names automatically.

In case no relationship is extracted from sentences in which two proteins are present, the co-occurrence information may be useful in a statistical method. However, this was not calculated at this point.

Although more than thirty sentences in which both proteins were present were collected, the interaction between the two proteins could not be validated. Only 11 of 164 interactions were validated from more than thirty sentences. 153 of 164 interactions were validated in less than thirty sentences. This indicates that the validation possibility is not very dependent on the number of collected sentences in which both proteins were present.

From seven invalidated interactions, more than forty abstracts were collected, but the wanted interactions were not extracted. From 131 validated interactions, less than thirty abstracts were collected for each interaction. This signifies that the validation possibility is not overly dependent on the number of abstracts collected.

| Table 4. Number of validated interactions by one abstract, full text, and a number of abstracts |
|---------------------------------|---|---|---|
| 190 interactions                | (A) | (B) | (C) |
| Not Validated                  | 107 | 24  | 62  |
| Validated                      | 83  | 166 | 128 |
|                                | 56.32% | 12.63% | 32.63% |
|                                | 43.68% | 87.37% | 67.37% |

(A) using only one abstract, (B) using full-text and (C) using abstracts collected from PubMed.

| Table 5. Number of validated interactions from the PubMed abstracts |
|---------------------------------|---|---|---|
| 277 interactions                | (D) | (E) | (F) |
| Not Collected                   | 27 | 9.75%  |
| No Sentence                     | 29 | 29  |
| Validated                       | 57 | 57  |
|                                 | 20.58% | 22.80% | 25.79% |
| Not collected                   | 164 | 164 |
|                                 | 59.20% | 65.60% | 74.21% |
| Total                           | 277 | 250 | 221 |
|                                | 100.00% | 100.00% | 100.00% |

(D) total interactions, (E) interactions that abstracts are collected from PubMed and (F) interactions in which both proteins are found in the sentences

---

\(^8\)DIP-PPI corpus: http://www.informatik.hu-berlin.de/~hakenber/corpora/

\(^9\)Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/

\(^10\)Saccharomyces Genome Database: http://www.yeastgenome.org/

\(^11\)gene/protein names from SGD: http://www.informatik.hu-berlin.de/~hakenber/corpora/dipppi/

\(^12\)http://www.yeastgenome.org/gene_list.shtml
simplified sentences even if the parser is not tuned to biomedical sentences. In the next step, the proposed system reads the results from the parser and extracts all existing interactions. For validation, more than one abstract was used and any extracted interactions that were false positives were resolved.

When the recall performance was assessed through the use of the DIP database of protein–protein interactions, the recall for IntEx and BioRAT were approximately 27% and 20%, respectively (Corney 2004). The recall in this study is 44% when only one abstract is used.

The proposed method validated protein-protein interactions at a rate of 63.68% through the use of one given abstract for an interaction, 67.37% through the use of collected PubMed abstracts, and 87.37% through the use of a given full-text paper. This value is different from the normal recall rate. For collected abstracts with proper sentences, the proposed method validated interactions in nearly 75% of the cases. Additionally, for a case in which at least one abstract was collected, the proposed method validated at a rate of 65%.

ACKNOWLEDGEMENTS

The project described in this paper was fully supported by the Korean Institute for Information Technology Advancement (IITA) under the Korean Ministry of Information and Communication.

REFERENCES


