Automated extraction of information on protein–protein interactions from the biological literature

Toshihide Ono¹, Haretsugu Hishigaki¹,2, Akira Tanigami¹ and Toshihisa Takagi²,*

¹Otsuka GEN Research Institute, Otsuka Pharmaceutical Co. Ltd, 463-10 Kagasuno, Kawauchi-cho, Tokushima, 771-0192, Japan and ²Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

Received on September 11, 2000; revised on September 13, 2000; accepted on September 28, 2000

ABSTRACT

Motivation: To understand biological process, we must clarify how proteins interact with each other. However, since information about protein–protein interactions still exists primarily in the scientific literature, it is not accessible in a computer-readable format. Efficient processing of large amounts of interactions therefore needs an intelligent information extraction method. Our aim is to develop an efficient method for extracting information on protein–protein interaction from scientific literature.

Results: We present a method for extracting information on protein–protein interactions from the scientific literature. This method, which employs only a protein name dictionary, surface clues on word patterns and simple part-of-speech rules, achieved high recall and precision rates for yeast (recall = 86.8% and precision = 94.3%) and Escherichia coli (recall = 82.5% and precision = 93.5%). The result of extraction suggests that our method should be applicable to any species for which a protein name dictionary is constructed.

Availability: The program is available on request from the authors.

Contact: ono@otsuka.gr.jp

INTRODUCTION

Recently, vast amounts of sequences have accumulated in public databases through the efforts of various genome sequencing projects. The next step in genome analysis requires not only defining the function of each gene but also determining its role in biological pathways. In particular, the study of protein–protein interactions is important to the understanding of biological process. These interactions form the basis of phenomena such as DNA replication and transcription, metabolic pathway, signaling pathway, and cell cycle control.

Protein–protein interaction data have been collected through both biochemical and genetic approaches, including the widely used yeast two-hybrid test. Several databases that accumulate these data are currently under development, including the FlyNets for Drosophila melanogaster (Sanchez et al., 1999), the MIPS interaction table for Saccharomyces cerevisiae (Mewes et al., 1999), and metabolic databases such as EcoCyc and KEGG (Karp et al., 1999; Ogata et al., 1999). The data stored in these databases are almost assembled manually. Because most of the interaction data still exists only in the scientific literature, which is written in a natural language that computers cannot easily manipulate, the collection of these data takes too much time and labor. Efficient processing of large amounts of scientific text therefore requires an intelligent information extraction method.

In this report, we describe a method for automated extraction of information on protein–protein interaction from text sources. Our method circumvents the complexities of natural language processing (NLP) techniques by focusing on a particular area of interest (protein–protein interactions) and using only simple rules for information extraction.

In the following section, we illustrate our method for information extraction, and show the results of applying it to the abstracts described on yeast and E.coli protein interaction.

METHODS

The overall architecture of our method is shown in Figure 1. First, our method identifies protein names in a sentence. Next, the sentence is processed by part-of-speech rules. Finally, information about protein–protein interaction is extracted by pattern matching. We describe the detail for each step in the following subsections.

*To whom correspondence should be addressed.
Selection of target text

1. Identification of protein names
2. Process of compound or complex sentences
3. Recognition of protein–protein interaction

Extraction of protein interactions

Fig. 1. Flowchart of the method for extracting protein–protein interaction data from text. Information is extracted in three steps.

**Step 1. Identification of protein names**

To extract information on protein–protein interactions from literature, it is necessary to identify protein names first. The issue of name and synonym identification remains as one of the big problems, because the standard nomenclature is often only loosely followed by authors naming new proteins (Fukuda et al., 1998; Proux et al., 1998). In this study, we identify protein names in the literature using a dictionary of protein names which is constructed manually. The process of name identification is based on pattern matching between the dictionary entries and words in sentences. Our method references a genetic nomenclature guide for pattern matching (Cherry, 1995; Chater et al., 1995). The examples of processing the sentence are shown in Figures 3a,b and 4a,b.

**Step 2. Processing compound or complex sentences**

A sentence which contains at least two proteins identified by Step 1 (Figures 3b and 4b) is parsed with simple part-of-speech rules to avoid the difficulty of extracting information on protein–protein interactions from compound or complex sentences using only word pattern-matching rules. We apply the Brill POS tagger package (Brill, 1994) to analyze parts of speech. The sentences are parsed using the following two rules:

**Rule 1.** If the sentence matches the following part-of-speech pattern as indicated by regular expression of Perl language, it is divided into two parts of (i) and (ii).

- \[ P^1 (/, CC DT) | (, IN) | : | :) P^2 \]
  - (i) \[ P^1 \]
  - (ii) \[ P^2 \]

Symbols in the patterns are referred to in Table 1.

**Step 3. Recognition of the protein–protein interaction**

The sentences processed by Step 2 (Figures 3d and 4d) are parsed using a simple pattern-matching rule to recognize the protein–protein interaction described in a sentence. This rule is based on the arrangement of protein names, prepositions, and keywords that indicate the type of relationship between proteins. Examples of keywords include ‘interact’, ‘associate’ and ‘bind’. To solve the problem of inflection of keywords during pattern matching, suffixes

<table>
<thead>
<tr>
<th>Table 1. Definition of symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
</tr>
<tr>
<td>: .</td>
</tr>
<tr>
<td>: :</td>
</tr>
<tr>
<td>: CC</td>
</tr>
<tr>
<td>CC</td>
</tr>
<tr>
<td>DT</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>JJ</td>
</tr>
<tr>
<td>NN</td>
</tr>
<tr>
<td>NNP</td>
</tr>
<tr>
<td>NNS</td>
</tr>
<tr>
<td>P(1/2)</td>
</tr>
<tr>
<td>P(3/4/5)</td>
</tr>
<tr>
<td>VB(1/2)</td>
</tr>
<tr>
<td>VBN</td>
</tr>
<tr>
<td>VBZ</td>
</tr>
</tbody>
</table>

The example of processing a sentence is shown in Figure 3c,d.

The sentence of Figure 3c matches the above pattern. The words which conform to the pattern are underlined in Figure 3c. By applying this rule, this sentence divides into the two parts shown in Figure 3d.

**Rule 2.** If the sentence matches the following part-of-speech pattern, it is divided and built again into two parts of (i) and (ii).

- \[ P^3 VB1 P^4 VB2 CC P^5 \]
  - (i) \[ P^3 VB1 P^4 \]
  - (ii) \[ P^3 VB2 P^5 \]

The example of processing the sentence is shown in Figure 4c,d.

The sentence of Figure 4c matches the above pattern. The word ‘interact’ and ‘modulates’ are assigned as VB1 and VB2, and ‘STD’ is assigned P3. In the same way, the staves, that are ‘directly with the TBP’ and ‘transcription of the SUC2 gene of S.cerevisiae’, are allotted to P4 and P5, respectively. By applying this rule, this sentence is transformed into the two parts shown in Figure 4d.
are removed using the Porter stemming algorithm (Porter, 1980). This method can remove the more common morphological and inflectional endings from words.

Moreover, to increase precision, we incorporate processing of negative sentences into this step. Negative sentences, which describe a lack of interaction, or ‘non-interaction’, constitute a well-known problem in language understanding. For this reason, processing of negative sentences has not been integrated into many related studies. As a result, the previously proposed methods often extract inaccurate information.

To address this problem, we have constructed patterns of regular expression:

- **PROTEIN1.** not (interact|associate|bind|complex).
- **PROTEIN2**

The example is shown as follows:

**Dmc1** does not interact in the two-hybrid assay with **Rad52p** or **Rad54p**.

"*\** indicates that the character immediately to its left may be repeated any number of times, including zero and ‘ ’. Indicates an arbitrary string. Protein names are indicated in bold type, and underlined words indicate the pattern of regular expression. Through pattern matching, we obtain the following information: ‘Dmc1 does not interact with Rad52’ and ‘Dmc1 does not interact with Rad54’.

- **PROTEIN1.** PATTERN for PATTERN. but not PROTEIN2

**PATTERN** is one of the patterns in Table 2.

The example is shown as follows:

**Bnr1p** interacts with another Rho family member.

**Rho4p** but not with **Rho1p**.

Through pattern matching, we obtain the following information: ‘Bnr1p interacts with Rho4p’ and ‘Bnr1p does not interact with Rho1p’.

**Evaluation of information extraction**

To evaluate our extraction method, we calculate recall and precision based on the following formula:

\[
\text{recall} = \frac{TP}{TP + TN} \quad (1) \\
\text{precision} = \frac{TP}{TP + FP} \quad (2)
\]

where **TP**, **TP + TN** and **TP + FP** indicate as follows:

- **TP** = the number of sentences extracted correctly by our method;
- **TP + TN** = the total number of sentences containing information on protein–protein interactions;
- **TP + FP** = the total number of sentences retrieved by our method.

In this study, we measured the value of **TP**, **TN** and **FP** by hand.

**IMPLEMENTATION**

In this study, we performed information extraction for yeast and **E.coli** proteins, because protein names for these two species are managed well in public databases. The yeast protein name dictionary was derived from entries in the **Saccharomyces** Genome Database (SGD) (Cherry et al., 1998). The gene symbols also have variations, called synonyms, which are also managed by SGD. The dictionary we constructed contained 6084 molecules and 16,722 synonyms. The **E.coli** protein name dictionary was constructed using K-12 data (Blattner et al., 1997) and contains 4405 entries. The protein names were gathered from WWW sites (http://genome-www.stanford.edu/Saccharomyces, http://www.genome.wisc.edu/html/k12.html). Next, we manually defined common word patterns for recognition of protein–protein interactions. We selected four key-words indicating the relation between proteins, those were ‘interact’, ‘associate’, ‘bind’, ‘complex’, and inflections of these words. Pattern matching rules were defined by the order of protein names, these keywords and prepositions. Table 2 shows the word patterns used to extract information.

Analyzed sentences were obtained by a MEDLINE search using the following key words, ‘protein binding’ as a MeSH term, and ‘yeast’ (in case of yeast), ‘**E.coli**’ (in case of **E.coli**), ‘protein’, and ‘interaction’. We filtered the corpus and retained only those sentences containing at least two protein names and one of the keywords described above. Such sentences are believed to have a higher probability of describing interactions among proteins. We obtained 834 and 752 sentences for yeast and **E.coli**, respectively.

**RESULTS**

We tested our extraction method for selected sentences using yeast and **E.coli** protein name dictionaries, the set of pattern matching rules and part-of-speech rules.

Figure 2 shows the examples of information extraction from some sentences.

In the case of Figure 2a, the protein names ‘Pc19’ and ‘**Pho85**’ are recognized initially. Next, the part-of-speech rule is applied, but the sentence remains largely unchanged. Following comparison with patterns outlined in Table 2, the sentence matches the pattern of ‘A and B complex’. As a result, information about the interaction between ‘Pc19’ and ‘**Pho85**’ is extracted. If there are multiple relationships between proteins in a sentence, our method extracts each relationship (Figure 2c).

Figures 3 and 4 show how information is extracted from a compound and complex sentence using the part-of-speech rules. As shown in Figure 3, when part-of-speech rule is not applied, this sentence matched the
Table 2. A set of word patterns for recognition of protein–protein interaction. A and B indicate the protein name

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Pattern</th>
<th>Example of sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interact</td>
<td>A interact with B</td>
<td>Spe97p interacts with spe98 and Tab4 in the two-hybrid system.</td>
</tr>
<tr>
<td></td>
<td>interaction of A (with) B</td>
<td>The interaction of Cet1 with Ceg1 elicits…</td>
</tr>
<tr>
<td></td>
<td>interaction between A and B</td>
<td>Functional and physical interaction between Rad24 and Rfc5…</td>
</tr>
<tr>
<td></td>
<td>A–B interaction</td>
<td>These data suggest that the Cet1-Ceg1 interaction is…</td>
</tr>
<tr>
<td></td>
<td>A and B interact</td>
<td>Sm1 and Cdc13 proteins displayed a physical interaction by…</td>
</tr>
<tr>
<td>Associate</td>
<td>A associate with B</td>
<td>Axl1 also associated directly with the cytosolic domains of Cce2.</td>
</tr>
<tr>
<td></td>
<td>association between A and B</td>
<td>Physical association between GCN5 and ADA2.</td>
</tr>
<tr>
<td></td>
<td>association of A (with) B</td>
<td>Association of Vma12p with Vph1p.</td>
</tr>
<tr>
<td></td>
<td>A and B association with each other</td>
<td>The SET4 and STE18 gene products associated with each other.</td>
</tr>
<tr>
<td>Bind</td>
<td>A bind to B</td>
<td>GCN binds to ADA2…</td>
</tr>
<tr>
<td></td>
<td>bind of A to B</td>
<td>The binding of Met28 to DNA.</td>
</tr>
<tr>
<td></td>
<td>A and B bind</td>
<td>Cdc24p and Bem1p bind to each other</td>
</tr>
<tr>
<td></td>
<td>bind between A and B</td>
<td>Binding between TIF34 and TIF35 in vitro.</td>
</tr>
<tr>
<td></td>
<td>A bind B</td>
<td>the N-terminal of SIN1 is sufficient to bind SAP1.</td>
</tr>
<tr>
<td>Complex</td>
<td>A(-)B complex</td>
<td>Pk11, 2-Pho85 kinase complexes become essential…</td>
</tr>
<tr>
<td></td>
<td>A and B complex</td>
<td>Cdc46p and Cdc47p… complex with each other.</td>
</tr>
<tr>
<td></td>
<td>complex A and B</td>
<td>Pob1 and Pob3 may form a complex…</td>
</tr>
<tr>
<td></td>
<td>A complex with B</td>
<td>GCG20 was… complex formation with GCN1.</td>
</tr>
<tr>
<td></td>
<td>A complex… contain B</td>
<td>Bollp is part of a larger complex that contains Cdc42p.</td>
</tr>
<tr>
<td></td>
<td>A complex B</td>
<td>Ste11 complexed to Ste7…</td>
</tr>
</tbody>
</table>

(a) Input: Co-immunoprecipitation experiments using in vitro translated proteins showed that C19 and Pho85 form a complex.
Output: (complex: Pc19, Pho85)

(b) Input: We define a Nab2p sequence that binds to Kap104p.
Output: (bind: Nab2p, Kap104p)

(c) Input: Association of UBE2I with RAD52, UBL1, p53, and RAD51 proteins in a yeast two-hybrid system.
Output: (associate: UBE2I, RAD52), (associate: UBE2I, UBL1), (associate: UBE2I, p53), (associate: UBE2I, RAD51)

Fig. 2. Example of the information extraction from some sentences. Protein names are indicated by bold type. Underlined regions match the pattern for recognition of protein–protein interaction.

Table 3 shows the recall and precision of extraction for each keyword. Both recall and precision share similar values between yeast and E.coli and usually exceed 80%. The word ‘interaction’ gives a particularly high extraction result (96.1% precision for both yeast and E.coli). On the other hand, the keyword ‘associate’ gives a lower precision, because sentences containing this word sometimes refer to relationships other than protein–protein interactions. For example, the sentence ‘Mso1p is functionally associated with Sec1p’ (Mso1p and Sec1p are protein names) matches the word patterns shown in Table 2, but this sentence does not describe a direct interaction.

DISCUSSION
We have described a method for automatically extracting information on protein–protein interactions from text...
Extraction of information on protein–protein interaction

Fig. 3. An example of a procedure for information extraction using the part-of-speech rule 1. (a) Target sentence. (b) The result of protein name identification. Protein names are indicated by bold type. (c) The result of tagging. Underlined words match the pattern of rule 1. The tagged text takes the form of ‘word/part-of-speech’. Tags are shown in Table 1. (d) The result of applying the part-of-speech rule. Underlined words match the pattern for recognition of protein–protein interaction. (e) The result of information extraction. (f) The result of information extraction without implementing the part-of-speech rule.

Table 3. Results of information extraction. (a) The value of recall and precision for yeast proteins. (b) The value of recall and precision for of E.coli proteins

<table>
<thead>
<tr>
<th>Key word</th>
<th>TP</th>
<th>TP + TN</th>
<th>TP + FP</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interact</td>
<td>198</td>
<td>222</td>
<td>206</td>
<td>89.1</td>
<td>96.1</td>
</tr>
<tr>
<td>Associate</td>
<td>55</td>
<td>68</td>
<td>61</td>
<td>80.9</td>
<td>90.2</td>
</tr>
<tr>
<td>Bind</td>
<td>103</td>
<td>119</td>
<td>108</td>
<td>86.6</td>
<td>95.3</td>
</tr>
<tr>
<td>Complex</td>
<td>152</td>
<td>176</td>
<td>164</td>
<td>86.4</td>
<td>92.7</td>
</tr>
<tr>
<td>Total</td>
<td>508</td>
<td>585</td>
<td>539</td>
<td>86.8</td>
<td>94.5</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Key word</th>
<th>TP</th>
<th>TP + TN</th>
<th>TP + FP</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interact</td>
<td>173</td>
<td>208</td>
<td>180</td>
<td>83.2</td>
<td>96.1</td>
</tr>
<tr>
<td>Associate</td>
<td>34</td>
<td>44</td>
<td>38</td>
<td>77.3</td>
<td>89.4</td>
</tr>
<tr>
<td>Bind</td>
<td>133</td>
<td>166</td>
<td>139</td>
<td>80.1</td>
<td>95.7</td>
</tr>
<tr>
<td>Complex</td>
<td>155</td>
<td>182</td>
<td>172</td>
<td>85.2</td>
<td>90.1</td>
</tr>
<tr>
<td>Total</td>
<td>495</td>
<td>600</td>
<td>529</td>
<td>82.5</td>
<td>93.5</td>
</tr>
</tbody>
</table>

(b)

Sources. The basic idea of our approach is that sentences will contain a significant number of protein names and word patterns that indicate the type of relationship between them. Focusing on a particular area of interest (such as protein–protein interactions) and pre-specifying a limited number of keywords circumvent the complexities of NLP technique like semantic and discourse analyses.

As interest in extraction of information on protein–protein interaction has grown recently, several other research groups have proposed systems for information extraction from the scientific literature. Sekimizu et al. (1998) describes a method to parse, determine noun phrases, spot the commonly-occurring verbs and choose the most likely subject and object from the candidate noun phrases in the surrounding text. They report precision results ranging from 67.8 to 83.3% across a range of verbs. Blaschke et al. (1999) try to do without NLP
STD1 interacts directly with the TBP and modulates transcription of the SUC2 gene of Saccharomyces cerevisiae.

Identification of protein name

STD1 interacts directly with the TBP and modulates transcription of the SUC2 gene of Saccharomyces cerevisiae.

Tagging

STD1/NNP interacts/VBZ directly/RB with/IN the/DT TBP/NNP and/CC modulates/VBZ transcription/NN of/IN the/DT SUC2/NNP gene/NN of/IN Saccharomyces/NNP cerevisiae/NN ./.

Applied the rule of part of speech

Recognition of protein-protein interaction

● STD1/NNP interacts/VBZ directly/RB with/IN the/DT TBP/NNP
● STD1/NNP modulates/VBZ transcription/NN of/IN the/DT SUC2/NNP gene/NN of/IN Saccharomyces/NNP cerevisiae/NN ./.

Recognition of protein-protein interaction

c) STD1 interacts with TBP (correct)
f) STD1 interacts with SUC2 (wrong)

Fig. 4. An example of a procedure for information extraction using the part-of-speech rule 2. (a) Target sentence. (b) The result of identification of protein names. Protein names are indicated by bold type. (c) The result of tagging. Underlined words match the pattern of rule 2. The tagged text takes the form of ‘word/part-of-speech’. Tags are shown in Table 1. (d) The result of applying the part-of-speech rule. Underlined words match the pattern for recognition of protein–protein interaction. (e) The result of information extraction. (f) The result of information extraction without implementing the part-of-speech rule.

technology, such as parsing and simple matching approach to extract protein interactions from scientific text. This method is simplified by assuming a pre-existing protein dictionary. It is difficult to compare to any other approach because they present no quantitative results. However, it is obvious that it will not be able to easily cope with a sentence which distances a subject or object from a verb. Tomas et al. (2000) have used Highlight, a general-purpose information extraction engine developed at SRI Cambridge for use in commercial applications, in combination with the NP scoring method, to obtain high precision; their method achieved 77% precision and 58% recall rates. The main causes of low precision and recall are protein identification with NP blanketing and no processing of a negative sentence. The main difference between these approaches and our method lies in the use of part-of-speech rules to process compound and complex sentences. Although the sentences generated by applying part-of-speech rules do not always keep the meaning of the original sentence, information on protein relationships is retained. The accuracy of this process is more than 95%. By using these rules, information can be extracted in the better precision than if they are not used (the precision is 86.2% in the case of yeast proteins). Our results suggest that while these rules are simple, they increase the effectiveness of information extraction.

Moreover, our method can also process negative sentences and extract information about non-interaction between specific proteins. Extraction of negative information is also valuable, because such data can be integrated into global protein interaction maps. The extraction accuracy of this process is 97% precision and 91.1% recall.

Our extraction method improves recall and precision rates compared with other methods, but some errors arise from utilizing only surface clues. The first error arises from semantic differences. For example:

These findings suggest that Msp1p is a component of the secretary vesicle docking complex whose function is closely associated with that of Dec1p.

The current method incorrectly extracts a protein interaction between ‘Msp1p’ and ‘Dec1p’, because the word
pattern in this sentence (underlined) matches the word pattern shown in Table 2. Sentences that conform to our extraction rules do not always describe a protein–protein interaction. We believe that semantic analysis for such sentences is necessary to reduce this type of error.

The second error arises from the processing of anaphoric terms. For example:

They form a complex even in the absence of cross-linker.

Our current method cannot extract the information because the proteins involved in the interaction are defined by the word ‘they’. Anaphoric terms such as pronouns and definite articles are often encountered when processing unrestricted text written in natural language. Improvements in our method will be necessary before it can derive actual protein names from these expressions. Anaphora resolution in NLP is regarded as one of the most difficult problems. To address this problem, Lappin and Leass (1994) described an algorithm that achieved a high rate of correct analysis. Incorporation of this approach will improve our success in this area.

Our method can extract information with high recall and precision for both yeast and *E.coli* proteins (Table 3). It suggests that the accuracy of information extraction based on word patterns and part-of-speech rules is independent of the species examined. We expect that our method can extract information with similar recall and precision rates for other species, including human, mouse and rat, by providing a species-specific protein name dictionary or by automatic identification of protein names (Fukuda et al., 1998). Then, this method should reduce time and labor for construction of protein–protein interaction databases.

CONCLUSION

We describe here an automated method for extracting information about relationships between proteins from scientific text by searching with protein names, word patterns and simple part-of-speech rules. We have eliminated the problem of text understanding by restricting the number of protein names and keywords. This method achieved high recall and precision without incorporating complicated NLP techniques and should be applicable to any species for which a protein name dictionary is constructed.

ACKNOWLEDGEMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, ‘Genome Science’, from the Ministry of Education, Science, Sports and Culture in Japan.

REFERENCES


