BIOINFORMATICS

Vol. 17 no. 2 2001
Pages 155–161

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

Toshihide Ono[1], Haretsugu Hishigaki[1,2], Akira Tanigami[1] and Toshihisa Takagi[2,*]

[1]Otsuka GEN Research Institute, Otsuka Pharmaceutical Co. Ltd, 463-10 Kagasuno, Kawauchi-cho, Tokushima, 771-0192, Japan and [2]Human 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 *Saccaromyces 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.

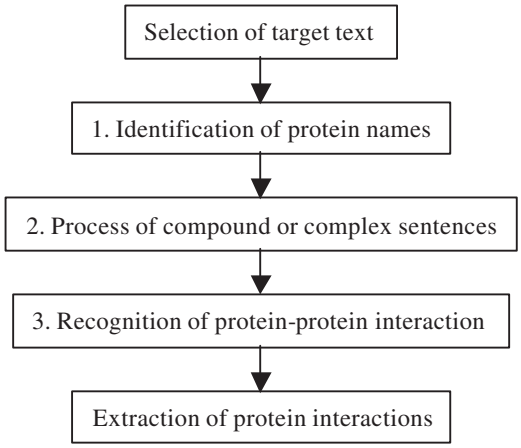

Fig. 1. Flowchart of the method for extracting protein–protein interaction data from text. Information is extracted in three steps.

Table 1. Definition of symbols

Symbol	Definition
,	comma
:	colon
;	semi-colon
CC	coordinating conjunction
DT	determiner
IN	preposition or subordinating conjugation
JJ	adjective
NN	Noun, singular or mass
NNP	proper noun, singular
NNS	noun, plural
$P(1/2)$	phrase
$P(3/4/5)$	phrase without verb
VB(1/2)	verb
VBN	verb, past participle
VBZ	verb, 3rd person singular present

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).

- $P1$ [(, CC DT) | (, IN) | : | ;] $P2$

  (i) $P1$
  (ii) $P2$

Symbols in the patterns are referred to in Table 1.

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).

- $P3$ VB1 $P4$ VB2 CC $P5$

  (i) $P3$ VB1 $P4$
  (ii) $P3$ VB2 $P5$

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.

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

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 Rad54p'.

- *PROTEIN1*.* *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:

$$\text{recall} = TP/(TP + TN) \quad (1)$$

$$\text{precision} = TP/(TP + FP) \quad (2)$$

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 keywords 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

Keyword	Pattern	Example of sentence
Interact	*A* interact with *B*	*Spc97p interacts with spc98* and *Tub4* in the two-hybrid system.
	interaction of *A* (with|and) *B*	The *interaction of Cet1 with Ceg1* elicits...
	interaction (between|among) *A* and *B*	Functional and physical *interaction* between *Rad24* and *Rfc5*...
	A–*B* interaction	These data suggest that the *Cert1-Ceg1 interaction* is...
	A and *B* interact	*Stn1 and Cdc13* proteins displayed a physical *interaction* by...
Associate	*A* associate with *B*	*Atx1* also *associated* directly *with* the cytosolic domains of *Ccc2*.
	association between *A* and *B*	Physical *association between GCN5 and ADA2*.
	association of *A* (with|and) *B*	*Association of Vma12p* with *Vph1p*.
	A and *B* association with each other	The *SET4 and STE18* gene products *associated with each other*.
Bind	*A* bind to *B*	*GCN binds to ADA2*...
	bind of *A* to *B*	The *binding of Met28* to *DNA*.
	A and *B* bind	*Cdc24p and Bem1p bind* to each other
	bind between *A* and *B*	*Binding between TIF34* and *TIF35* in_vitro.
	A bind *B*	the N-terminal of *SINI* is suffisient to *bind SAP1*.
Complex	*A*(-|/)*B* complex	*Pc11, 2-Pho85* kinase *complexes* become essential...
	A and *B* complex	*Cdc46p* and *Cdc47p*... *complex* with each other.
	complex *A* and *B*	*Poll* and *Pob3* may form a *complex*...
	A complex with *B*	*GCG20* was... *complex* formation *with GCN1*.
	A complex... contain *B*	*Boilp* is part of a larger *complex* that *contains Cdc42p*.
	A complex *B*	*Ste11 complexed* to *Ste7*...

(a) Input: Co-immunoprecipitation experiments using in vitro translated proteins showed that **Pc19** 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.

pattern of 'association of *A* with *B*' in Table 2. Then, our method extracts wrong information of interaction between 'Ste4p' and 'Ste18p'. But, by using the rule, the sentence is divided into two parts which contain only one protein respectively, and they do not match the pattern in Table 2. As a result, we can avoid the wrong information extraction. Similarly, Figure 4 shows that when part-of-speech rule is not used, the method extracts interactions between 'STD1' and 'TBP', and 'STD1' and 'SUC2'. The latter relationship is incorrect, because information of direct interaction is not described in the sentence. These results indicate that dividing the sentence with the rules allows us to retain correct information and eliminate inaccurate extraction. If these rules are not applied, our method fails to recognize protein–protein interaction and extracts inaccurate information.

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

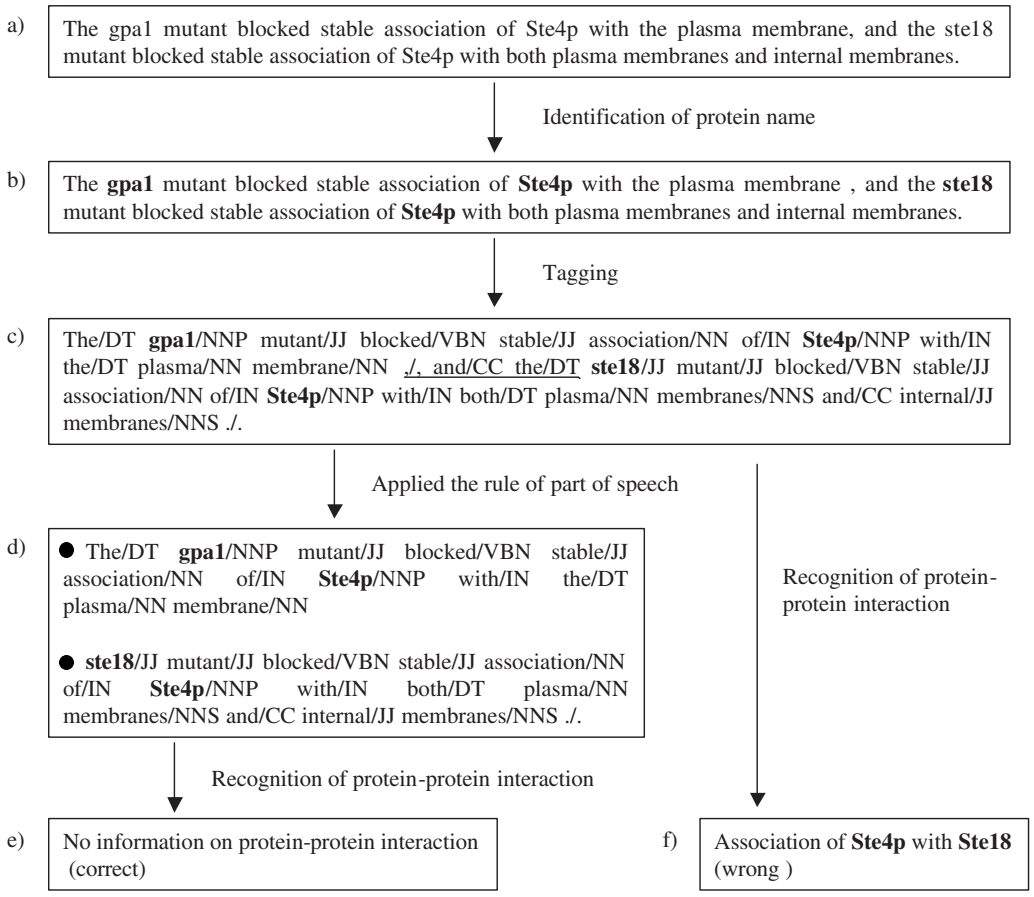

Fig. 3. An example of a procedure for information extraction using the part-of-speech rule 1. (a) Target sentence. (b) The result of protein name identification. Protein names are indicated by bold type. (c) The result of tagging. Underlined words match the pattern of rule 1. The tagged text takes the 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

Key word	TP	$TP+TN$	$TP+FP$	Recall (%)	Precision (%)
			(a)		
Interact	198	222	206	89.1	96.1
Associate	55	68	61	80.9	90.2
Bind	103	119	108	86.6	95.3
Complex	152	176	164	86.4	92.7
Total	508	585	539	86.8	94.5
			(b)		
Interact	173	208	180	83.2	96.1
Associate	34	44	38	77.3	89.4
Bind	133	166	139	80.1	95.7
Complex	155	182	172	85.2	90.1
Total	495	600	529	82.5	93.5

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

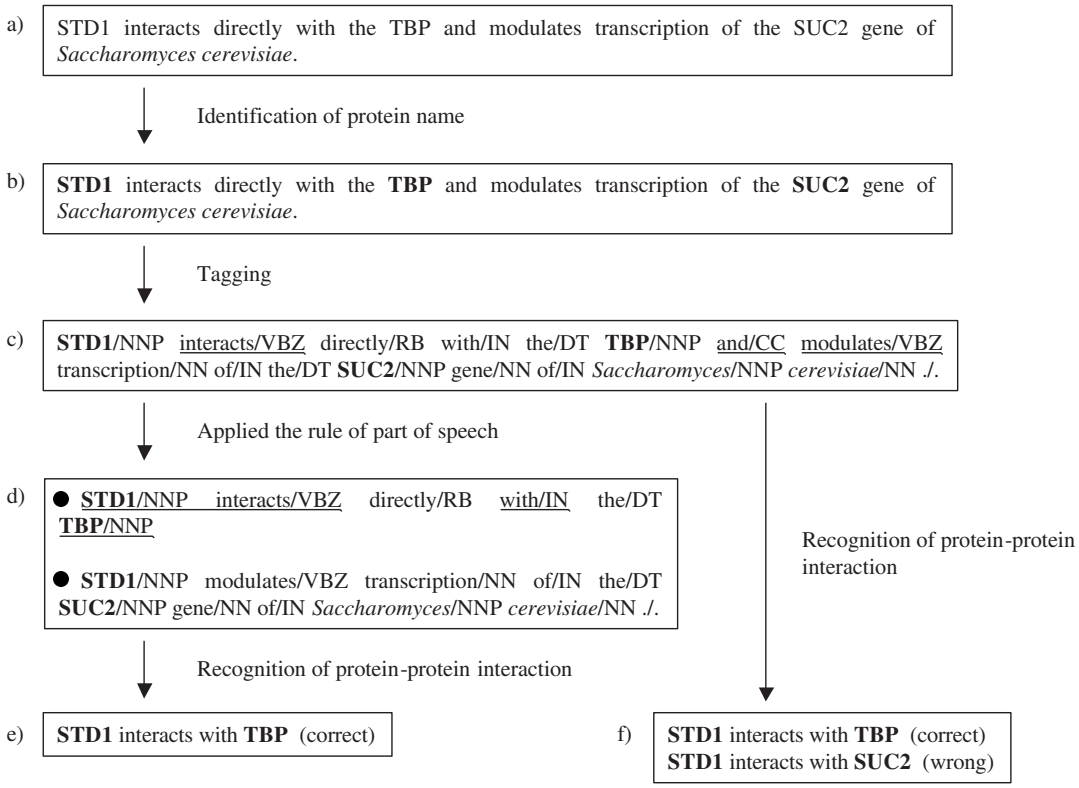

Fig. 4. An example of a procedure for information extraction using the part-of-speech rule 2. (a) Target sentence. (b) The result of identification of protein names. Protein names are indicated by bold type. (c) The result of tagging. Underlined words match the pattern of rule 2. The tagged text takes 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

Blaschke,A., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB99)*. AAAI Press, pp. 60–67.

Blattner,F.R., Plunkett,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Brill,E. (1994) Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press.

Chater,K., Berlyn,M. and Bachmann,B. (1995) Genetic nomenclature guide, bacteria. *Trends Genet*. Mar, pp. 5–8.

Cherry,J.M. (1995) Genetic nomenclature guide. *Saccaromyces cerevisiae*. *Trends Genet*. Mar, pp. 11–12.

Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) SGD: *Saccharomyces* genome database. *Nucleic Acids Res.*, **26**, 73–79.

Fukuda,K., Tsunode,T., Tamura,A. and Takagi,T. (1998) Toward information extraction: identifying protein names from biological papers. In *Proceeding of the Pacific Symposium on Biocomputing (PSB98)*, pp. 707–718.

Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1999) Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.

Lappin,S. and Leass,H.J. (1994) An algorithm for pronominal anaphora resolution. *Comput. Linguistics*, **20**, 535–561.

Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**, 44–48.

Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.

Porter,M.F. (1980) An algorithm for suffix stripping. *Program*, **14**, 127–130.

Proux,D. *et al.* (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome Informatics*, 72–80.

Sanchez,C. *et al.* (1999) Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.*, **27**, 89–94.

Sekimizu,T., Park,H.S. and Tsujii,J. (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Informatics*, 62–71.

Tomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interaction from scientific abstracts. *Proc. Pacific Symp. Biocomput. (PSB2000)*, **5**, 538–549.