

# A Method of Extracting Sentences Related to Protein Interaction from Literature using a Structure Database

Yoshikazu Kaneta  
Graduate School of  
Information Science and  
Technology, Osaka University  
2-1 Yamadaoka, Suita, Osaka,  
565-0871, Japan  
kaneta@ist.osaka u.ac.jp

Md. Ahaduzzaman  
Munna  
Graduate School of  
Information Science and  
Technology, Osaka University  
2-1 Yamadaoka, Suita, Osaka,  
565-0871, Japan  
munna.md@ist.osaka  
u.ac.jp

Takenao Ohkawa  
Graduate School of  
Information Science and  
Technology, Osaka University  
2-1 Yamadaoka, Suita, Osaka,  
565-0871, Japan  
ohkawa@ist.osaka  
u.ac.jp

## ABSTRACT

As a protein expresses its function through interaction with other chemical compounds, it is important for protein functional analysis to accumulate the interaction information. Because protein interaction information is described in tens of thousands of literature sources, it is impractical to extract all the information manually. Automatic information extraction systems based on the template matching method have already been developed. However, it is not possible to match all the sentences related to interaction information due to the range of the sentence complexity.

We propose a method of extracting sentences related to interaction information. In a protein-compound complex structure, residues appearing in sentences related to interaction are close to their interaction partners. The physical distance between them can be calculated by using the structure data in the PDB database, and the adjacency indicates that the sentence including them describes the interaction information. In a free protein structure, the distance cannot be calculated because the coordinates of its partners is not registered in the structure data. Thus, we use the homologous protein structure data, which is complexed with their partners.

The proposed method was applied to seven papers written about protein-compound complex proteins and four papers written about free proteins, obtaining 71% and 72% sentence extraction F-measures, respectively.

## 1. INTRODUCTION

A protein's function is related to its local interaction site[1], and accumulating and utilizing comprehensive information about protein interaction sites is expected to become a driving force that accelerates protein functional analysis. Such

information can be obtained by carefully examining papers that discuss protein structural analysis. However, constructing such a database by manual extraction from a large number of sources is impractical, leading to the need for automatic information extraction (IE) from literature. Many IE systems have been developed, some of which focus on protein interaction information. In some of these systems, a template matching method is used to extract interaction information[3, 4, 5]. Other systems have been reported, including systems in which a subject word and its object word are extracted[6], and systems in which the frequently appeared verb is considered as a clue for information extraction[7]. However, it is difficult to extract interaction information accurately by using only the text data from literature due to the inherent complexity and ambiguity of the sentences.

This paper proposes an information extraction method, in which we complementarily use protein structure data in the PDB database as the information source to cope with incompleteness of the sentences. First, we aim to extract not the formalized interaction information but the sentences related to the interaction information, which are defined as sentences that mention the interaction site or its interaction partners (such as other proteins, DNA, substrates, ions). As the interaction atoms of a protein are very near to the protein's interaction partner, the distance between them in their structure data provide a hint to specifying which sentences are related to interaction information. However, pairs of adjacent residues or their partners in a sentence do not always mean interaction between them. Hence, we introduce rules for determining the partner interaction in order to recognize the correct interaction pairs. Some structure data contain no coordinates for the interaction partners; therefore, in such structure data for a free protein, it is not possible to calculate the distance. In such cases, we use homologous complex proteins to estimate the interaction site on a free protein.

## 2. PROTEIN STRUCTURE AND LITERATURE

### 2.1 Protein structure

A protein consists of a number of amino acids, which bind each other by peptide bonds. A dehydrated amino

acid is called an amino acid residue, and is identified by the residue number. In literature on protein structural analysis, a residue is described with a three-letter abbreviation and a residue number; for instance, the 100th alanine residue from the terminal of a peptide chain is expressed as “Ala100,” “Ala<sup>100</sup>,” and so on.

## 2.2 Protein interaction

A protein binds to its interaction partners by the interaction between their atoms. Interaction names such as “hydrogen bond” or “van der Waals interaction” commonly appear in literature. Interaction partners are residues of other proteins, bases of DNA strands, substrates, or ions. For example, “Adenine,” “PTR” or “Zn<sup>2+</sup>” appear in literature. Interaction names and the interaction-partner names are not always specified in sentences related to interaction information.

## 2.3 Protein structure data on the PDB database

The protein structure data is registered in the PDB (Protein Data Bank; <http://www.rcsb.org/pdb/>) database. An example of protein structure data is shown in Figure 1. The record name at the line head denotes the type of information in its line; for example, the ‘AUTHOR’ line includes the names of authors who analyze the protein structure, the ‘HET’ line includes the information about compounds that bind to the protein, and the ‘ATOM’ line includes the coordinates of protein atoms, residue names, residue numbers, and so on.

A protein is classified into one of two types (a complex protein or a free protein) according to its structure data. A complex protein consists of multiple polypeptide chains except for its own chain, or includes the coordinates of compounds. A free protein consists of its own polypeptide chains.

In the PDB database, a protein is identified by four character codes (PDB-ID). For example, the protein shown in Figure 1 has PDB-ID ‘1EFT’. PDB-ID is used in this paper to identify a protein.

## 2.4 Protein structure literature

The papers that discuss protein structural analysis are referred to in the ‘JRNL’ record of the PDB structure data. Each paper describes the experimental details of the relevant protein structural analysis, such as a method of protein structure determination, location of the interaction site, and types of interaction between the protein and its interaction partner. Though data on experimental details have already been registered in the database, it also is important to extract other information, such as binding and function information about proteins, and the interaction information in the interaction site. This paper focuses on extracting sentences that describe interaction information.

## 2.5 Sentences related to interaction information

In sentences related to interaction information, words that indicate the location of an interaction site, such as names of atoms or identifiers of residues, are described. In literature written about a particular complex protein, the interaction partners often appear in the sentence related to interaction information. An example of such a sentence is shown as follows:

- The methyl group of the inhibitor is hydrogen-bonded

```

HEADER      ELONGATION FACTOR TU (EF-TU) COMPLEXED WITH          24-AUG-93  1EFT
COMPND      2 GUANOSINE-5'-(BETA,GAMMA-IMIDO) TRIPHOSPHATE (GDPNP)
SOURCE      (THERMUS AQUATICUS)
AUTHOR      M.KJELDGAARD,P.NISSEN,S.THIRUP,J.NYBORG
REVDAT      1 31-AUG-94 1EFT 0
JRNL        AUTH M.KJELDGAARD,P.NISSEN,S.THIRUP,J.NYBORG
JRNL        TITL THE CRYSTAL STRUCTURE OF ELONGATION FACTOR EF-TU
JRNL        TITL 2 FROM THERMUS AQUATICUS IN THE GTP CONFORMATION
.
.
.
HET         GNP      406      32      SEE REMARK 7.
HET         MG      407      1       MAGNESIUM ++
.
.
.
ATOM        1  N      ALA      1       75.082  -7.178  43.255  1.00  27.28
ATOM        2  CA     ALA      1       74.276  -6.678  42.092  1.00  34.44
ATOM        3  C      ALA      1       75.143  -5.790  41.184  1.00  33.12
ATOM        4  O      ALA      1       76.370  -5.912  41.224  1.00  33.94
ATOM        5  CB     ALA      1       73.025  -5.872  42.611  1.00  29.66
ATOM        6  N      LYS      2       74.494  -5.260  40.142  1.00  31.69
ATOM        7  CA     LYS      2       74.851  -4.000  39.455  1.00  29.67
ATOM        8  C      LYS      2       74.040  -3.970  38.149  1.00  45.70
.
.
.

```

Figure 1: Example of protein structure data

to the oxygen atom of Ile 60.

This sentence mentions the interaction between “methyl group” of “inhibitor” and “Ile 60.” On the other hand, the interaction partners are seldom specified in literature written about a free protein. An example of a sentence about a free protein is as follows:

- The active site triad consisting of Asp 64, Asp 121 and Glu 157 plays an important role in the catalytic function.

This sentence asserts that three residues (Asp 64, Asp 121 and Glu 157) on the protein belong to its interaction site.

## 3. THE METHOD FOR SENTENCE EXTRACTION

### 3.1 Overview of sentence extraction

It is known that in a complex protein the distance between protein residues and their binding partners is small[8]. The distance between them helps in determining whether a sentence is related to interaction.

In a free protein, the distance cannot be calculated due to the nonexistence of coordinate data on the interaction partners in the structure data. Homologous complex proteins are often registered in the PDB database, and the interaction site in a free protein is similar to the corresponding site in homologous complex proteins. Sentences written about similar interaction sites can be candidates for sentences related to interaction information.

Consequently, the sentence extraction system we propose consists of two processes according to protein type. Figure 2 provides an overview of the method. The literature and the structure data of the target protein are the input data. In this case, the input literature is an NNP-tagged document, in which the proper nouns are tagged corresponding to their

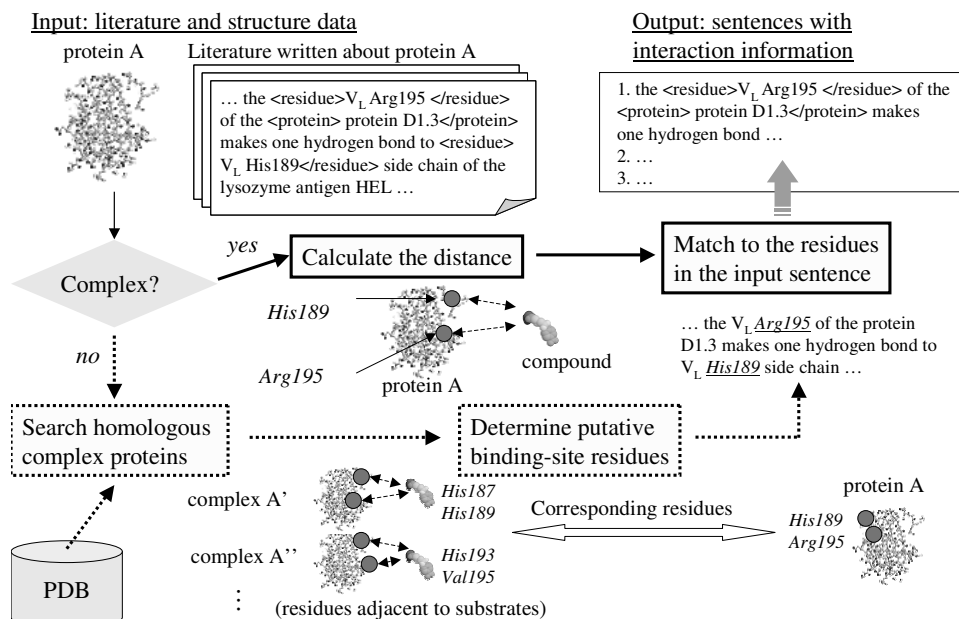


Figure 2: Sentence extraction system

Table 1: List of proper noun tag

proper noun tag	contents	examples
<protein>	name of a protein	antibody D1.3
<chemical>	name of a substrate	esterase
<peptide>	name of a peptide	Glu-Gly-Arg-chloromethyl ketone
<atom>	name of an atom	oxygen
<ion>	name of an ion	zinc ion
<molecule>	name of a molecule	water molecule
<group>	name of a group	hydroxyl group
<residue>	name of a residue	Tyr100
<chain>	chain information	main chain
<sec_structure>	name of a secondary structure	$\alpha$ -helix, $\beta$ -sheet
<ter_structure>	name of a tertiary structure	oxyanion hole, pocket
<domain>	name of a domain	CDR L1

meanings. The list of NNP tags is shown in Figure 1, but for further details of the method of attaching the NNP tags to the original text data, see reference[9].

The problems with the proposed method are that a sentence including interaction residues is not always an important one. For instance, residue identifiers are often used in a sentence that describes intra-interactions of a protein. Thus, it is important to consider the condition for determining whether the sentence does indeed relate to interaction information.

## 3.2 A method of sentence extraction for complex proteins

### 3.2.1 Outline of the method

Figure 3 shows an outline of the sentence extraction method for a complex protein. First, the residue and its interaction partner are selected from a sentence. The dis-

### Sentence S

The interaction shields <residue>Arg21</residue> completely from the solvent, by which <chemical>PTR</chemical> binding site turns away from <residue>His23</residue>.

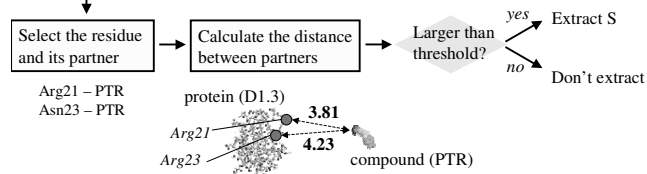


Figure 3: Sentence extraction for complex

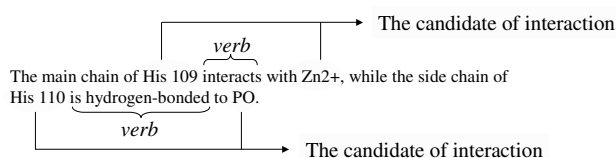
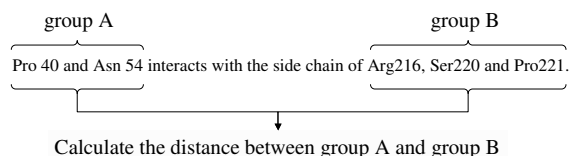
tance between them is then calculated, and if the distance is smaller than  $T_{Dc}$  (a distance threshold value for complex proteins), the sentence is extracted as a sentence related to interaction information.

If the distance between all combinations of pairs of residues and their partners is used for the threshold judgment, many unnecessary sentences may be extracted. For example, the sentence in Figure 3 contains two residues (Arg21 and His23) and one substrate (PTR). Since the distance between “Arg21” and “PTR” is comparatively small, the sentence may be extracted; however, this sentence does not represent the actual interaction information between them.

### 3.2.2 Rules for interaction partner determination

To solve the problem mentioned above, the structural patterns in the sentence that should be extracted are introduced. Patterns for consideration are as follows:

1. A verb often appears between residues and their partners. Hence, if a verb does appear between residues and their partners, the distance between them is sufficiently meaningful to be calculated. Figure 4 repre-


**Figure 4: Verb between proper nouns**

**Figure 5: Grouping words with same NNP tag**

sents an example of a verb between residues and their partners.

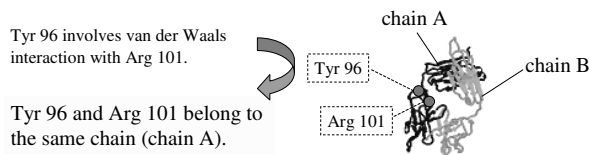
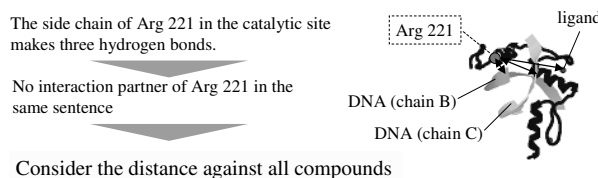
- The interaction between coordinated multiple residues and their partners is often described. In such a sentence, all of the coordinated residues are related to the interaction with their partner. Therefore, the coordinated residues or their partners should be grouped, and the distance between the groups is significant. However, since the pattern “between *A* and *B*” represents the interaction between *A* and *B*, the *A* and *B* must not be grouped. Figure 5 shows an example of grouping.
- Some sentences describe an interaction between residues in the same protein chain, though such sentences are not a targets for extraction. Hence, it is checked whether the residues belong to the same chain. If they do, they are excluded from the distance calculation target. An example of a sentence written about residues in the same chain is shown in Figure 6.
- Occasionally the interaction partners are omitted, and only the residue names are described. In such a sentence, the interaction partner cannot be determined. Hence, combination between all of the residues and all of the chemical compounds in the structure data should be considered. An example is shown in Figure 7.

Based on these patterns, four rules are defined as follows, where [A] means the strings that are attached by the NNP tag <A>, and [any-pn] means one of [residue], [chemical], [peptide] or [ion]. VERB represents a verb, \* means any number of any words, and ... means words except for “...”. “Brill’s Tagger” [10] is used to analyze the category of each word in a sentence.

#### Rule for the order of verbs and proper nouns

If a sentence matches to the sequences “[residue] \* VERB \* [any-pn]” or “[any-pn] \* VERB \* [residue],” then [any-pn] is regarded as an interaction partner of [residue].

#### Rule for grouping words


**Figure 6: Interaction in one chain**

**Figure 7: Lack of the interaction partner**

If a sentence matches to the sequence “between [any-pn1] and [any-pn2],” then [any-pn1] and [any-pn2] are regarded as an interaction pair. If a sentence matches “[any-pn1], [any-pn2]” or “between [any-pn1] and [any-pn2]” and [any-pn1] has the same NNP tag as [any-pn2], [any-pn1] and [any-pn2] are grouped.

#### Rule for residues in the same chain

If a sentence matches “[residue1] \* [residue2],” it is determined whether [residue1] and [residue2] belong to the same chain. If both [residue1] and [residue2] are in the same chain, they are not regarded as an interaction pair.

#### Rule for lack of partner

If a sentence matches “\* [residue] \*” and [residue] has not been matched to other rules, [residue] and all possible partners in the structure data are regarded as interaction pairs.

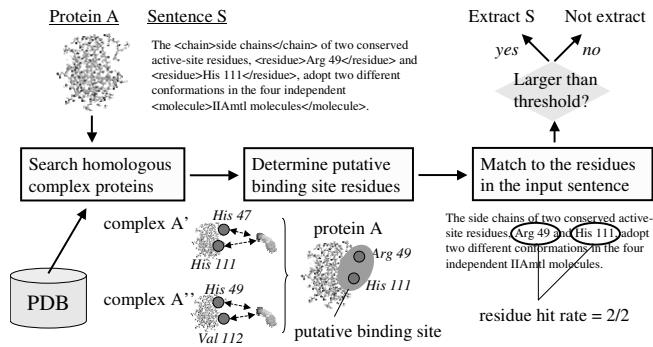
### 3.2.3 Distance calculation between components of interaction pair

After the interaction pairs are determined, the sentence is evaluated by calculating the distance between them. Because the residues and their partners contain many atoms, the distance between a residue and its partner is defined as the minimum distance between atoms of a residue and its partner, since two adjacent atoms may indicate the interaction between them. In the case of groups, all residues or their partners in a group are related to the interactions. Thus, the distance between two groups is defined as the average of all combinations of atom pairs in the residues and their partners.

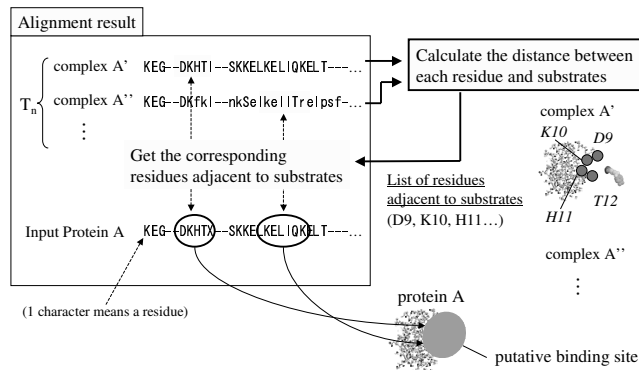
## 3.3 Method of sentence extraction for free protein

### 3.3.1 Outline of the method

Figure 8 illustrates the outline of the sentence extraction method for a free protein. First, homologous complex proteins are retrieved from the database, after which the residues of the putative binding site in a free protein are determined, using the residues related to the interaction in each homologous protein. If the residues of the putative



**Figure 8: Outline of sentence extraction for free protein**



**Figure 9: putative interaction site**

interaction site are described in a sentence, the hit rate of residues can be calculated. Finally, if the hit rate is higher than  $T_S$  (a threshold value for the hit rate), the sentence is extracted as the sentence related to interaction.

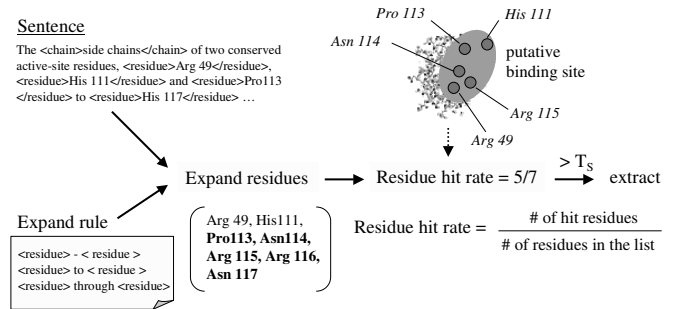
The “BLAST” [11] program is used for retrieving homologous proteins from the database. It calculates the alignment between two protein sequences.

### 3.3.2 Determination of the putative interaction site

Figure 9 shows the outline of the method for determining the putative interaction site on a free protein. Many homologous proteins are retrieved as the result of a homologous protein search. Considering the interaction site predicted from low-homologous proteins does not show high reliability, we use only the highest  $T_N$  (a threshold value for the number of homologous proteins) numbers of proteins that have at least 30% homology. The residues, whose distance from their partners is smaller than  $T_{Df}$  (a distance threshold value for free proteins), are selected from each homologous protein in order to specify the protein’s interaction site. Then, the residues in the free protein corresponding to the interaction site of the homologous proteins are specified by using the alignment result. These residues are regarded as being responsible for constructing the putative interaction site in the free protein.

### 3.3.3 Judgment of sentence extraction

Figure 10 presents a method of determining which sen-



**Figure 10: Judgment of sentence extraction**

tence is related to the interaction information. Most of the residues in the sentence related to interaction information are involved in the interaction, hence the residue hit rate is introduced to determine whether or not the sentence should be extracted, as follows:

$$R_{hit} = \frac{n_i}{n_s},$$

where  $n_s$  denotes the number of residues in the sentence,  $n_i$  means the number of residues that exist both in the sentence and in the putative interaction site. Since  $R_{hit}$  is larger than threshold  $T_S$ , the sentence is extracted as the sentence related to interaction information.

The location of the interaction sites are sometimes described in a sentence by the range of residue numbers, and in such a sentence, it is unsatisfactory to match only the start residue and end residue of the range. As a result, we introduce the residue-expanding rule as follows to satisfy the all of the residues in the range.

### Rule for expanding residues

If a sentence matches “[residue1] - [residue2],” “[residue1] to [residue2]” or “[residue1] through [residue2],” and the residue number of [residue1] ( $R1$ ) is smaller than that of [residue2] ( $R2$ ), the residues having a residue number between  $R1$  to  $R2$  are regarded as appearing in the sentence.

## 4. EVALUATION

The proposed method is evaluated in terms of the precision  $P$ , recall  $R$ , and F-measure [12]  $F$ , which are defined as follows:

$$P = \frac{COR}{SYS}, \quad R = \frac{COR}{GLD}, \quad F = \frac{2 \times P \times R}{P + R},$$

where  $GLD$  represents the quantity of correct data,  $SYS$  the quantity of data that is extracted by the proposed method, and  $COR$  represents the quantity of correct data extracted by the proposed method. To estimate the effectiveness of the proposed method, the manually and correctly tagged documents are used as the input data.

### 4.1 Evaluation for complex proteins

#### 4.1.1 Parameter adjustment

Ten papers describing structural analysis of complex proteins (PDB-IDs are 1a07, 1a0b, 1a0f, 1a0i, 1a0l, 1a0m, 1a0o,

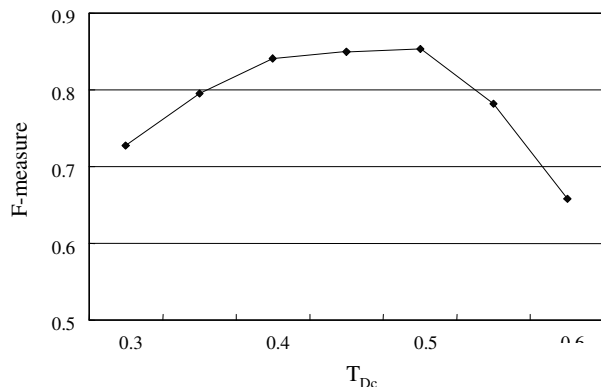


Figure 11: The parameter tuning result for

Table 2: Literature data (complex)

ID	# of Word	# of Sentence	# of Page
1a0h	9534	359	13
1a0q	7569	389	10
1a26	5308	359	9
1a3l	3025	340	7
1a5v	5779	303	6
1a5y	7502	302	9
1a5z	10221	428	13

1a0s, 1a13, 1a15) were prepared to determine the threshold  $T_{Dc}$ . The F-measure of the correctly extracted sentences was calculated as changing the threshold  $T_{Dc}$  by 0.05. Figure 11 shows the result, which determines that  $T_{Dc} = 5.0$ .

#### 4.1.2 Evaluation of sentence extraction

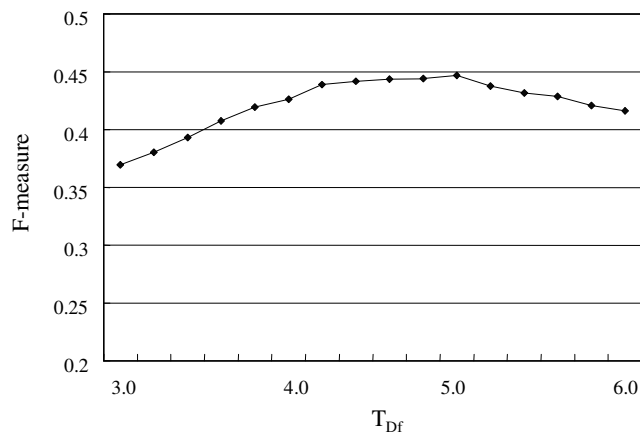
The proposed method was applied to seven papers about the structural analysis of complex proteins. The literature data and PDB-IDs used in this experiment are shown in Table 2. The result is shown in Table 3, in which the value in the parenthesis indicates the result without rules. The precision of the proposed method was better than result without rules, indicating the effectiveness of the proposed method.

Some sentences related to interaction information could not be extracted. An example in the paper describing “1a3l” is as follows: “... asparagine-L91 and aspartic acid-H50 form hydrogen bonds to the carboxylate side chain that substitutes for the carbamate diene substrate.” This sentence means that there are “hydrogen-bond” interactions between “asparagine-L91 and aspartic acid-H50” and “carboxylate side chain.” The reason for this failure is that “form” is considered as a noun. The improvement is anticipated by elevating the precision in POS tagging.

On the contrary, the following sentence in the paper about “1a5y” was extracted incorrectly: “The catalytic Asp residue (Asp-181 of PTP1B) contributes to the basic limb of the pH activity profile, and its substitution to Ala-21 causes a 105-fold reduction in kcat, suggestive of a role as an acid catalyst.” This sentence mentions the catalytic function information of “Asp-181” and “Ala-21.” Because the two catalytic residues are adjacent to each other, the sentence was extracted incorrectly. It is necessary to classify the extracted sentences into the interaction information or the function in-

Table 3: Experimental result (complex protein)

ID	GLD	SYS	COR	P	R	F
1a0h	10	20	10	0.50(0.18)	1.00(1.00)	0.67(0.30)
1a0q	19	18	18	1.00(0.50)	0.95(1.00)	0.97(0.67)
1a26	10	18	10	0.56(0.48)	1.00(1.00)	0.71(0.65)
1a3l	17	19	15	0.79(0.55)	0.89(1.00)	0.83(0.71)
1a5v	10	12	6	0.50(0.25)	0.60(1.00)	0.55(0.40)
1a5y	16	29	10	0.34(0.16)	0.63(1.00)	0.44(0.28)
1a5z	3	3	3	1.00(0.13)	1.00(1.00)	1.00(0.22)
Ave.	12.1	17	10.3	0.61(0.27)	0.85(1.00)	0.71(0.43)


 Figure 12: The parameter tuning result for  $T_{Df}$ 

formation.

## 4.2 Evaluation for free proteins

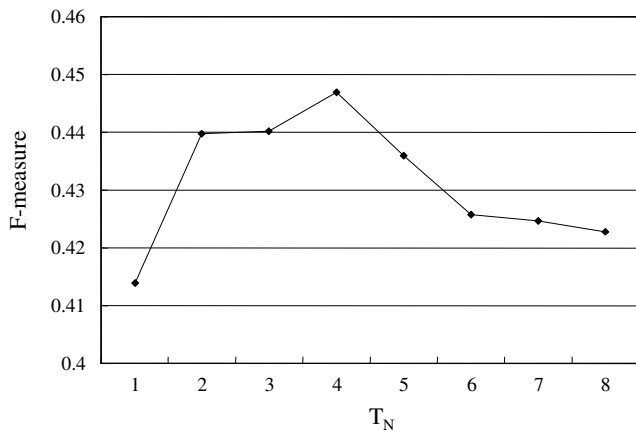
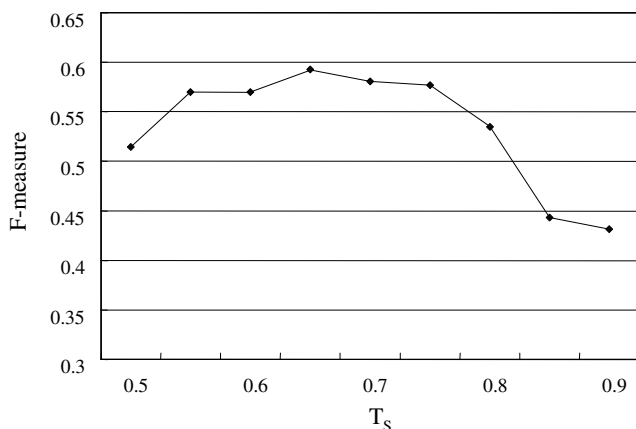
### 4.2.1 Parameter adjustment

Five papers describing structural analysis of free proteins (1a03, 1a0k, 1a3y, 1a4u, 1a5c) were used to determine the threshold  $T_{Df}$ ,  $T_N$  and  $T_S$ . The F-measure for extracting residues in the correct sentence was estimated for adjusting  $T_{Df}$  and  $T_N$ . Here,  $T_N = 4$  and  $T_S = 0.5$  are used in the adjustment of  $T_{Df}$ , while  $T_{Df} = 5.0$  and  $T_S = 0.5$  are used in the adjustment of  $T_N$ . On the other hand, the F-measure for extracting the correct sentences was estimated for the adjustment of  $T_S$ . The results are shown in Figure 12, 13 and 14. In this adjustment,  $T_{Df} = 5.0$  and  $T_N = 4$  are used.

### 4.2.2 Evaluation of sentence extraction

We applied the proposed method to four papers. The literature data and PDB-IDs used in this experiment are shown in Table 4. The result is shown in Table 5, in which the values in parentheses indicate results without a rule. The high recall value indicates that almost all sentences related to interaction information were extracted correctly. In addition, the precision was higher than that of the result without a rule.

An example that could not be extracted from the paper about “1a3a” is as follows: “The IIAmtl-binding site of HPr involves the loop comprising residues 13-21, the helix composed of residues 16-27 and the helix containing residues 48-56 (Figure 7).” This sentence means “IIAmtl-binding site” is included in the domain 13-21, 16-27 and 48-56. The


 Figure 13: The parameter tuning result for  $T_N$ 

 Figure 14: The parameter tuning result for  $T_S$ 

reason for failure is that the domain is larger than the actual interaction site, and the residue hit rate decreased. In order to extract such sentences,  $T_S$  should be changed dynamically depending on the sentences.

In the same paper, the following sentence was extracted by mistake: "... Arg 49 is turned away and cannot form hydrogen bonds with the phosphoryl group." This sentence means the residue "Arg 49" is not related to interactions. The reason is that the residue corresponding "Arg 49" in the homologous protein "1j6t" is related to the interactions. For future works, it is required to recognize the negation of the extracted sentences.

### 4.3 Effectiveness of using structure data

To evaluate the effectiveness of using the PDB structure data, the result of proposed method was compared with the result without using the structure data, which extracts all the sentences that includes one or more <residue> tags. Eleven papers in Table 2 and Table 4 were used for evaluation. As a result, the average precision, recall and F-measure were 0.32, 1.00 and 0.48 respectively. The precision and F-measure of the proposed method in Table 3 and Table 5 is much larger than that of the method without structure data, which indicates the effectiveness of the proposed method.

**Table 4: Literature data (free protein)**

ID	# of Word	# of Sentence	# of Page
1a03	7335	498	9
1a3a	8535	545	12
1a4l	9550	365	11
1a58	4338	199	6

**Table 5: Experimental result (free protein)**

ID	GLD	SYS	COR	P	R	F
1a03	2	3	2	0.67(0.22)	1.00(1.00)	0.80(0.36)
1a3a	14	35	12	0.34(0.32)	0.86(0.86)	0.49(0.47)
1a4l	27	44	24	0.55(0.53)	0.89(0.95)	0.68(0.68)
1a58	6	7	6	0.86(0.67)	1.00(1.00)	0.92(0.80)
Ave.	12.3	22.3	11	0.61(0.44)	0.94(0.94)	0.72(0.58)

## 5. CONCLUSION

This paper proposed a method for extracting sentences with interaction information from literature using protein structure data in the PDB database. For literature on a complex protein, we introduced the interaction partner determination rules to filter the combinations of pairs of residues and their partners, because some unexpected sentences may be extracted as a result of calculating all combinations. For a free protein, we used the homologous protein structure data to determine the putative interaction residues, since structure data of a free protein contains no coordinates of the interaction partner. Future works will be as follows:

1. A sentence that includes the interaction residues may possibly describe the function information. Consequently, it is necessary to distinguish a sentence related to interaction information from a sentence related to function information.
2. Some sentences related to interaction information include negative words, thus we should improve the method for determining whether the information is negative.

## 6. ACKNOWLEDGEMENT

The authors wish to thank Prof. Norihisa Komoda, who offered useful advice related to this research. We would also like to thank Prof. Haruki Nakamura and Prof. Nobutoshi Ito for their professional advice. A part of this research was supported by BIRD of the Japan Science and Technology Corporation and the Japan Society for the Promotion of Science.

## 7. REFERENCES

- [1] S. Goto, T. Nishioka, and M. Kanehisa: "LIGAND: Chemical Database for Enzyme Reactions," *Bioinformatics*, Vol. 14, pp. 591-599 (1998).
- [2] N. Ito, H. Sakamoto, K. Kobayashi and H. Nakamura: "Development of PDBj-ML," *Genome Informatics*, vol. 12, pp. 508-509 (2001).
- [3] C. Blaschke and A. Valencia: "The Frame-Based Module of the SUISEKI Information Extraction System," *IEEE Intelligent Systems*, Vol. 17, No. 2, pp. 14-20 (2002).

- [4] J. Thomas, D. Milward, C. Ouzounis, S. Pulman and M. Carroll: "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Proc. Pacific Symp. Biocomputing*, pp.384-395 (2000).
- [5] Y. Kaneta, M. Numa, and T. Ohkawa; "Automatic Extraction of Protein Active Site Information from Literature Using Template Matching and Anaphora Analysis," in *Proc. of the 2003 Int'l Conf. on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'03)*, pp. 100-106 (2003)
- [6] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raju and Simon Rhodes: "A Multi-level Text Mining Method to Extract Biological Relationships," *IEEE Computer Society Bioinformatics Conference (CSB'02)*, pp.97-108 (2002).
- [7] T. Sekimizu, H. S. Park, J. Tsujii: "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts," *Proc. 9th Workshop Genome Informatics*, pp.62-71 (1998).
- [8] M. Chanda: *Atomic Structure and Chemical Bond including Molecular Spectroscopy*, Tata McGraw-Hill (1972).
- [9] M. Numa, Y. Kaneta, and T. Ohkawa: "Automatic Classification of Proper Names in Protein-related Literatures Using Database Retrieval on WWW," in *Proc. of the 5th Conference on Computational Biology and Genome Informatics (CBGI'03)*, pp. 903-906 (2003).
- [10] E. Brill: "Some Advances in Transformation-Based Part of Speech Tagging," *the 20th National Conference on Artificial Intelligence* (1994).
- [11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman: "Basic Local Alignment Search Tool," *J. Mol. Biol.* Vol. 215, pp. 403-410 (1990).
- [12] M. T. Pazienza: "Information Extraction," Springer-Verlag (1997).