PROFESS, a System to Support Extracting Protein Function Information from Literature

Yoshikazu Kaneta[1], Masayuki Numa[1], Md. Ahaduzzaman Munna[1], Yohe Sakurai[1], Takenao Ohkawa[2]

[1]Graduate School of Information Science and Technology, Osaka University, Japan

[2]Graduate School of Science and Technology, Kobe University, Japan

{munna.md,sakurai.yohei}@ist.osaka-u.ac.jp, ohkawa@cs.kobe-u.ac.jp

Abstract

Protein functional site information described in thousands of literatures plays an important role in protein functional analysis. In this paper, we present PROFESS, a system to assist with extraction of functional information from literature. In PROFESS, first the functional information is extracted using the protein's structure data, then the extracted information is corrected or complemented by the operator.

The proposed system was applied to 11 documents related to structural analysis of protein for extracting functional site information, where the average recall value and F value were 0.927 and 0.782, respectively.

Keywords: protein functional site, literature, structure data, information extraction

1 Introduction

As a protein expresses its function through the binding of various compounds to its functional site [1], a database of functional site information plays an important role in protein functional analysis [2]. However, such functional site information is described in thousands of literatures and it is thus impractical to extract all the information manually. In this paper, we present PROFESS, a system to assist with extraction of protein functional site information from literature concerned with protein structural analysis. The process of the proposed system is divided into two parts. First, the system automatically extracts the functional site information, after which the improper information is corrected or the information that is lacking is complemented by the system operator. In such a process, we expect improvement in the efficiency of the operation and high accuracy in extraction of the functional site information.

Through only checking the extracted information, it might be difficult for the operator to determine whether the extracted information is indeed functional site information, since the operator needs to refer to the position of the site in the structure or in a figure. To help with this, we have built a graphical user interface on which the operator can check the extracted information and related figures simultaneously. In addition, to achieve high accuracy in the extraction process, we introduce a novel method that considers the distance between a protein and a compound on a given structure to extract relevant functional site information automatically from the literature.

2 Structure of protein and functional site information

A protein is a long chain of amino acid residues linked together. In the chain, each residue is given a number counted from one end of the chain. For example, Alanine, the 100th residue would be written as "Ala-100" or "Ala^{100}." PDB (Protein Data Bank) is a repository of the data on structurally analyzed proteins. In the structure data, information about relevant literature, and the three-dimensional coordinates of the atoms in the protein and the compound, are registered.

A protein is classified as a complex protein or a free protein according to its structure data. A complex protein consists of multiple polypeptide chains in addition to its own chain, or includes the coordinates of the compounds. On the contrary, a free protein consists of only polypeptide chains of its own type.

Literature that discusses protein structural analysis is referred to in PDB structure data. Each paper describes the experimental analysis, such as the method of protein structure determination, location of the functional site, and the types of interaction between the protein and its interaction partner on that site, etc. Our objective is to support the operator in extracting information related to the functional site and interactions that occur on it.

Functional site information can be defined into three categories: positional information on the protein, positional information about the compound, and the relation among them. The positional information on the protein (the positional information on the compound) includes the name of the protein (compound), residues, and atoms. Information about the relations among them includes the name of the interaction and the function that occurs on their binding. For example, in a sentence, "The methyl group of PTR is hydrogen-bonded to the oxygen atom of Ile 60," the positional information on the protein is the name of the residue"Ile 60," the positional information about the compound comprises the name of the compound "PTR" and

the name of the functional group "methyl group," and finally the information on the relation among them is the name of the interaction, "hydrogen-bond."

3 PROFESS

3.1 Outline of PROFESS

Figure 1 shows the outline of PROFESS. In this system, the functional site information is automatically extracted from the literature related to the protein structural analysis by complementary use of the protein's structure data. The extracted information is displayed to the operator, who corrects the information by modifying, adding, or deleting manually. Following that, the functional site information can be completed.

Because PROFESS aims to reduce the time required to extract the information, effective support of the extraction operation and high accuracy of information extraction are expected. In Section 3.2, we describe the development of the user interface for effective support, and in Section 3.3, we explain the module of automatic functional site information extraction, the purpose of which is to achieve accurate extraction.

3.2 Display components in PROFESS

The display components in PROFESS are shown in Figure 2. The main screen of PROFESS comprises four sections. (1) for relevant literature (area A in Figure 2); (2) for extracted protein functional site information (area B); (3) for figures related to the functional site information (areas C and D); and (4) for the structure data (area E).

(1) Section for relevant literature

This section displays the text of the literature. To determine whether the extracted information is indeed functional site information, the operator needs to check adjacent sentences. For this, not only the sentences on functional site information, but also any sentence from the literature, can be referred to by the operator. Besides, the words are marked in different colors on the basis of type (for example, the words for the name of the protein, residue, function, and interaction are marked in different colors). This makes scanning functional site information easier.

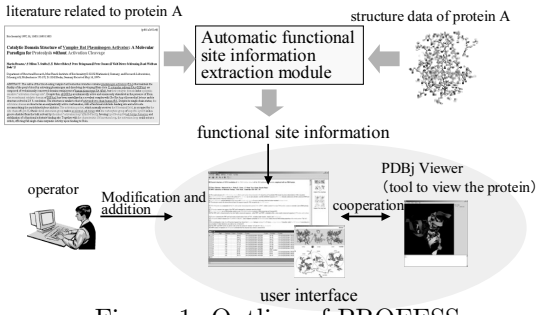

Figure 1: Outline of PROFESS

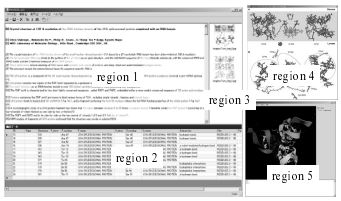

Figure 2: Display components of PROFESS

(2) Section for extracted protein functional site information

This area displays the extracted functional site information. The data extracted by the automatic extraction module is initially loaded into cells in this frame, thus reducing the extraction burden on the operator. If the operator points to a sentence missing functional site information in the present literature display, the system automatically displays a list of probable extraction candidates from the highlighted sentence and its neighbors.

(3) Section for figures related to the functional site information

This section displays figures related to relevant functional site information. Literature concerning the structural analysis of protein includes figures in which functional site information is delineated in detail. By referring to these figures, supplementary functional site information might be noticed and thus added to previously extracted information. To browse the figures efficiently, the figures are displayed as thumbnails (area C) and in an individual window (area D).

(4) Section for structure data

This window displays structure data on the protein. Sentences related to functional site information are often complex in structure and also complicated in content, which might make it difficult for the operator to grasp the content of the sentence. Considering the positions of the residues and atoms in the structure, the operator can easily make a decision about the sentence. In this window the relevant parts linking the sentence and the figure would be highlighted. To display the structure, we used jV version 2, an advanced version of PDBjViewer [3] provided by Protein Data Bank Japan.

3.3 Module for automatic extraction of functional site information

Figure 3 provides an outline of the module responsible for extracting the functional site information automatically. In this module, literature related to the structural analysis of protein and the protein's structure data is given as input data. The input literature is an NNP-tagged document, in which the named entities are tagged corresponding to their meaning (for example <protein>, <atom>,

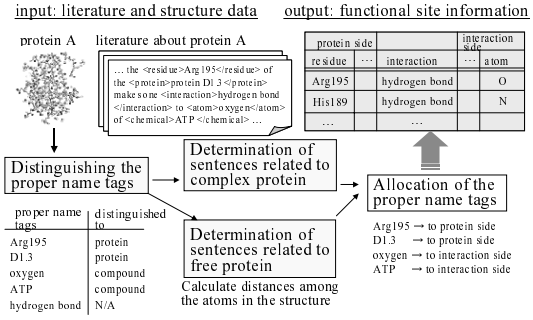

Figure 3: Outline of the automatic information extraction module

<residue>, etc) [4]. First, we will distinguish between the protein-related tags and compound-related tags. Then, calculating the distances among the atoms in the structure, whether or not a sentence relates to functional site information can be determined, where different methods would be applied in the case of complex proteins and free proteins, since their structure data are different. Finally, if the named entity relates to a protein, it will be extracted to the field for protein, and in the same way it will be extracted to the field for compounds if it relates to a compound.

3.3.1 Distinguishing the named entities

In general, the name of the residue is written along with the chain information. For example, "ArgH20" indicates that residue "Arg20" exists in the chain "H." Furthermore, the tag of the residue can be examined to ascertain whether chain H belongs to the protein. If the atoms and the functional groups do belong to the protein, they are written close to the residues (for example, "oxygen of Tyr35," etc). Therefore, the sentence is divided into segments according to the verbs and prepositions (except "of") and then in each segment, the tags for the named entities are ascertained with respect to the tags of the other named entities in the same segment.

3.3.2 Functional site information for complex proteins

As the coordinates of the compound are written in the complex protein structure data, choosing a pair comprising a residue and a compound in a sentence, then calculating their distance of separation in the structure, and finally comparing this distance with the threshold value, it can be determined whether the sentence describes the functional site information. Figure 4 shows an overview of the method. However, there are special patterns where the residues' names are enumerated in a sentence and the compound names are omitted. Consequently, mistakes might occur mistakes in selecting the interaction candidates. To solve this problem, we introduce the two rules given below:

(1) Rule regarding grouping:

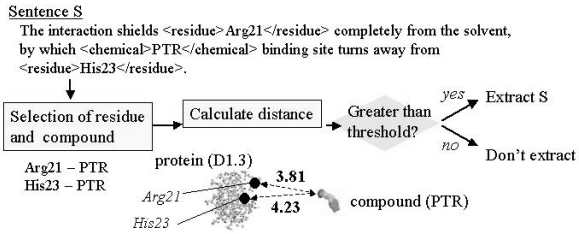

Figure 4: Determination of sentences related to complex protein

The residues enumerated in a sentence simultaneously take part in interaction with the compound. Therefore, if the residues are written together and they belong to the same protein, they are defined as a group. The compounds written together also compose a group.

(2) Rule regarding omission of the compound name:

In a sentence where the residue name is written but the compound name is omitted, it is difficult to determine the compound interacting with the residue. In this case, all pairs comprising residues and compounds described in the structure data should be considered.

3.3.3 Functional site information for free protein

In the structure data on a free protein, since the coordinates for the compound are not specified, we cannot calculate the distance as we did above. However, if the free protein and its homologous complex protein are registered in PDB, the functional site can be considered as a homologous site between the free protein and the complex protein. Hence, as described in Figure 5, by searching the complex proteins homologous to the target free protein and then presuming the functional site of the free protein from the sites close to the compounds in each complex protein, we can determine an appropriate sentence.

The "BLAST" program [5] is used for retrieving homologous proteins from the database. The search result illustrates the homology between the free protein and the complex proteins. On the basis of this homology, the residues of the putative binding site in the free protein are predicted using the residues related to the interaction in each homologous protein.

If the residues of the putative interaction site are described in a sentence, the hit rate for residues in that sentence can be calculated. The hit rate is the ratio of the number of hit residues in the sentence to the number of residues at the putative interaction site. Finally, if the hit rate is higher than the threshold, we may conclude that the sentence de-

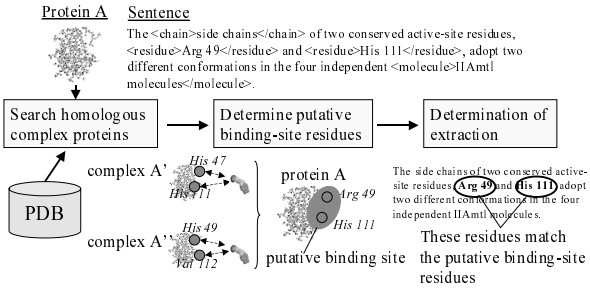

Figure 5: Determination of sentences related to free protein

scribes the functional site information.

4 Evaluation

We evaluate the accuracy of the module for functional site information extraction experimentally using literature referred by PDB.

The PDB-ID of the protein and the number of words, sentences, pages, and sentences containing correct information are summarized in Table 1, while the evaluation result is shown in Table 2. The experimental result reveals that the recall value is high, which means that most of the functional site information can be covered. Since extraction leakage might be a high risk for a system that supports information extraction, it is necessary to increase the recall value, even if the precision value decreases slightly . Our result shows that the support system performs well.

An example of a sentence mistakenly extracted from a paper related to protein, "1a3a" is, "... Arg 49 is turned away and cannot form hydrogen bonds with the phosphoryl group." This sentence means that the residue "Arg 49" does not form a hydrogen bond. However, as this residue exists in the presumed functional site of the homologous complex protein "1j6t," an error had occurred. To cope with this problem, negative sentences require special treatment.

5 Conclusion

In this paper, we proposed PROFESS, a system that supports extraction of protein functional site information. PROFESS provides an effective user interface that, for example, displays tagged named entities as probable candidates for extraction, and features a function that enables reference to structure data. In addition, PROFESS has a facility for extracting functional site information automatically. The automatic extraction was evaluated experimentally using 11 documents related to structural analysis of protein. The average recall value and F value were 0.927 and 0.782 respectively, which confirms the effectiveness of the method.

Our future work will include improving the system's accuracy. We will consider the learning of mistakenly extracted patterns and the automatic analysis of information that failed to be extracted.

Table 1: Literature data used in the experiment

PDB-ID	♯of words	♯of sente -nces	♯of pages	♯of correct sente -nces
1a0h	9534	359	13	11
1a0q	7569	389	10	18
1a26	5308	359	9	12
1a3l	3025	340	7	16
1a4k	5498	190	5	10
1a5i	8903	324	11	22
1a5y	7502	302	9	10
1a03	7335	498	9	2
1a3a	8535	545	12	14
1a4l	9550	365	11	25
1a58	4338	199	6	6
Average	7008.81	351.81	9.27	13.27

Table 2: Results

PDB-ID	Precision	Recall	F-measure
1a0h	0.625	0.909	0.741
1a0q	0.750	1.000	0.857
1a26	0.800	1.000	0.889
1a3l	0.867	0.813	0.839
1a4k	0.769	1.000	0.870
1a5i	0.870	0.909	0.889
1a5y	0.471	0.800	0.593
1a03	0.667	1.000	0.800
1a3a	0.400	0.857	0.546
1a4l	0.512	0.913	0.656
1a58	0.857	1.000	0.923
Average	0.690	0.927	0.782

Acknowledgement

The authors wish to thank Prof. Norihisa Komoda who offered useful advice related to this research. A part of this research was supported by BIRD of the Japan Science and Technology Corporation and Grant-in-Aid for Scientific Research.

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 Infomatics*, Vol. 12, pp. 508-509 (2001).

[3] K. Kinoshita, H. Nakamura: "eF-site and PDBjViewer: database and viewer for protein functional sites," *Bioinformatics* Vol. 20, pp. 1329-1330 (2004).

[4] 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).

[5] 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).