社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－BIO－3（13）
2005／12／22

蛋白質機能情報抽出支援システムPROFESSにおけるSVMを利用した機能情報文特定方式

Md. Ahaduzzaman Munna[1], **大川　剛直**[2]

[1] 大阪大学大学院情報科学研究科
[2] 神戸大学大学院自然科学研究科
munna.md@ist.osaka-u.ac.jp, ohkawa@cs.kobe-u.ac.jp

概要

我々は、蛋白質構造解析に関する文献から、蛋白質機能情報の抽出を支援するシステムPROFESS (PROtein Functional site information Extraction Support System) を開発している。PROFESSは、機能情報記述文の特定機能、機能情報記述文からの情報抽出機能、抽出結果の手動編集機能を備え、機能情報のデータベース化を総合的に支援する。本稿では、SVMを利用した機能情報記述文の特定手法について述べる。PROFESSが対象とする文献には、必ず対応する立体構造データが存在する。このことを利用し、提案手法では、機能情報関連キーワードや記述パターンといった「文に関する特徴」に加え、文中に記述されている残基と相互作用対象の3次元空間上での距離といった「構造に関する特徴」をもとに、各文を特徴ベクトルで表現し、SVMによる学習を実現する。

提案手法を蛋白質の構造解析関連の7編の文献に適用した結果、機能情報記述文の特定に関する平均再現率は0.72、平均F値は0.70となった。

SVM-based Sentence Distinguishing in PROFESS, a System to Support Extraction of PROtein Functional Site Information from Literature

Md. Ahaduzzaman Munna[1] **and Takenao Ohkawa**[2]

[1]Graduate School of Information Science and Technology, Osaka University
[2]Graduate School of Science and Technology, Kobe University
munna.md@ist.osaka-u.ac.jp, ohkawa@cs.kobe-u.ac.jp

Abstract

We are developing PROFESS, a system to assit with the extraction of protein functional site information from the literature related to protein structural analysis. PROFESS supports distinguishing the sentences related to functional site information, extracting functional site information from the distinguished sentences, and modification of the extracted information to make a correct database of functional site information. In this paper, we describe the method which uses SVM to distinguish the sentences related to functional site information. In this method, each sentence in the literature is expressed by a vector for SVM considering the following features: (1) feature related to structure that is the distance on the structure between the two interacting objects which are written in the sentence, (2) keywords related to functional site information, (3) expression patterns of the functional site information related sentences.

The proposed method was applied to seven documents related to structural analysis of protein for distinguishing sentences related to functional site information, where the average recall value and F value were 0.72 and 0.70, respectively.
respectively.

Keywords: protein functional site, literature, structure data, SVM, 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. So, an automatic and effective support system to extract this information is essential.

In the literature, sentences are not always simple rather very long and complex in structure. Besides, the functional information also is not always explicitly written. The method of keyword matching or template matching to extract this information has

shown limitation in these cases[3]. This functional site information related sentence bears a unique feature in it. The name of the residues and their interaction partners (another residue, compound, dna, etc) are written in it. Using the protein structural data, the distance between the interacting objects can be calculated and comparing that to the threshold value, whether a sentence relates to functional site information or not can be determined[4]. But, this method has limitation when the two interacting objects are having large distance among them even though in the sentence the interaction information among them is clearly written. Besides, sentences not related to function information, for example experimental information, protein structural information are also extracted by this method. So, to solve the effective extraction problem, an unique method is needed which can solve both of the above method's limitations.

The above methods bear two strong features for the sentences related to functional site information and if both of these features are taken into account a higher performance can be expected for extraction. Beside the distance feature between the interacting objects and the functional site information related keyword feature, it is also found that there are specific patterns in the functional site information related sentences. If some correct examples of this functional site information related sentence can be learnt by a learner, then the learner can be used as a classifier to distinguish the correct sentences. Support vector machine (SVM) [5] which is both a learner and a classifier, has been proved successful in the field of pattern recognization and information retrieval. To use the SVM, first the sentence in the literatures can be expressed by vector by the features mentioned above and then the correct examples can be learnt which is defined as the "Learning Phase" and finally by the "Testing Phase" the correct sentences can be distinguished from the literature. In this paper, we mainly focus on this Learning Phase and Testing Phase of the SVM to distinguish the sentences related to functional site information from literature.

We have a further intention about the extension of our proposed system. The functional site information related sentences which are distinguished by the proposed method, would be deleted if they are found unnecessary or the necessary sentences that are lacking would be complemented by the system operator. We want to use this modification information as feedback from the operator and learning those by SVM again, and repeating this feedback and learning for few different literatures, finally a system which is very accurate in distinguishing protein functional site information related sentence can be expected.

To help with the modification process, we have built a graphical user interface on which the operator can check the distinguished sentences and carry out the addition or the deletion process with a click of the mouse.

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", "Ala^{100}", etc. 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

Figure 1 shows the outline of PROFESS. In the extraction module, the functional site information related sentence is automatically distinguished from the literature related to the protein structural analysis by complementary use of the protein's structure data, keywords related to functional site information and patterns which are frequently expressed in the sentences related to functional site information, where the SVM is used as a learner and a classifier. Then from the distinguished sentences, functional site information is extracted by the process of distinguishing the named entities. This 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. In this paper, we explain the method of distinguishing the sentences related to functional site information only. The feedback from the operator is not described in this paper which is our future work for the proposed system.

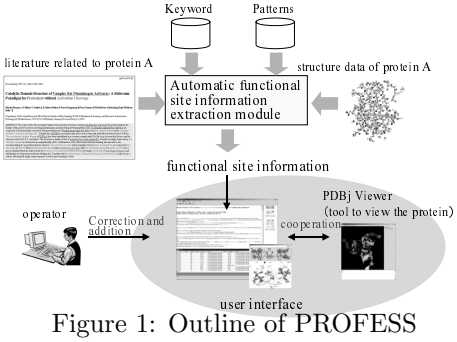

Figure 1: Outline of PROFESS

The display components in PROFESS are shown in Figure 2. The main screen of PROFESS comprises four sections. (1) for relevant literature (region 1 in Figure 2); (2) for extracted protein functional site information (region 2); (3) for figures related to the functional site information (region 3 and region 4); and (4) for the structure data (region 5).

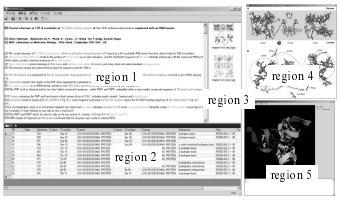

Figure 2: Display components of PROFESS

4 Module for automatic extraction of functional site information

In the module of automatic extraction of functional site information, the process of distinguishing the sentences related to functional site information is devided into two phases. The first phase is the "Learning Phase" where the learning corpus is learnt by the SVM tool. The second phase is the "Testing Phase" where the sentences related to functional site information are distinguished from the testing corpus using the SVM tool. Figure 3 provides an outline of the Learning Phase. In this phase, the Learning corpus made from literature related to the structural analysis of protein, the structure data of protein, the functional site information related keywords and the patterns which are often expressed in the functional site information sentences are given as input to the Vector Expression Module.

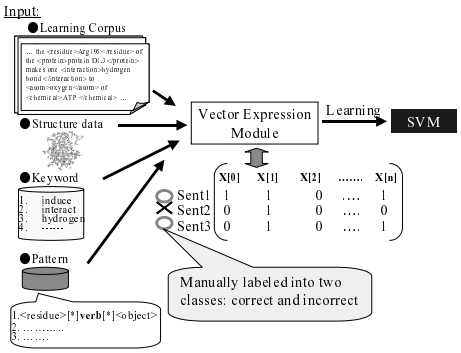

Figure 3: Learning phase using SVM

The input corpus is an NNP-tagged document, in which the named entities are tagged corresponding to their meaning (for example <protein>, <atom>, <residue>, etc) [6]. In the Vector Expression Module the learning corpus is expressed in vector which is later learnt by the SVM tool. In the same way, in the testing phase, the testing corpus is expressed by vector by the Vector Expression Module which is then given as input to the SVM tool. The SVM tool which has been learnt already by the Learning phase, distinguish sentences related to functional site information automatically and extracts it. Finally, in the extracted sentences, 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. We describe each of these procedures elaborately in the following subsections.

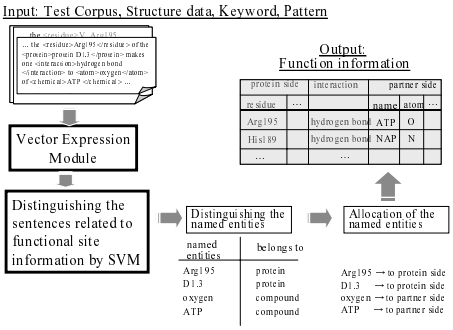

Figure 4: Testing Phase using SVM

4.1 Support Vector Machine

SVM is a classification method which aims to estimate a classification function $f : \chi \rightarrow \{\pm 1\}$ using labeled training/learning data from $\chi \times \{\pm 1\}$. Learning the labeled training data, SVM creates a hyperplane in the feature space so that the examples of the two different class have the maximum distance which is defined as "Margin" from it. The examples which touch the margin line are called the support vectors. In Figure 5, the $\circ$ and the $\diamond$ on the dotted line are called the support vectors. SVM has been very popular recently as a learning method in the research field of pattern recognition.

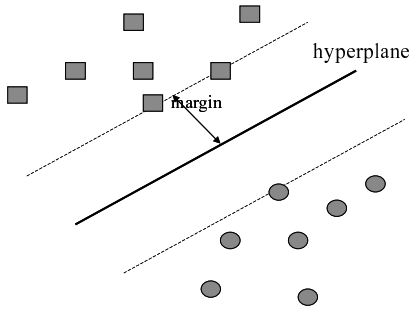

Figure 5: Example of SVM

4.2 Vector Expression using the Structure data of Protein

In a sentence related to functional site information, it is described the interaction between two objects, for example: interaction between "residue and residue", interaction between "residue and compound". When two objects interact to each other, they are very close to each other on the structure. So, we can calculate the distance between the two interaction candidates using the structure data and comparing the distance with the threshold value we can consider the sentence as a probable functional site information sentence. Therefore, if any sentence i shows this special feature, in the sentence's vector expression(x[i][0] x[i][1] ....x[i][n]) the first vector element x[i][0] is set as "1" or "0" for the distance smaller than the threshold and greater than the threshold respectively. However, there are special patterns where the residues' names are enumerated in a sentence and the compound names are omitted. Consequently, mistakes might occur in selecting the interaction candidates. To solve this problem, we introduce the two rules given below:

(1) Rule regarding grouping:

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.

4.3 Vector Expression using the functional site information related keywords

In the sentences related to protein functional site information, it is found that words like "interact", "bind", "induce", "hydrogen bond" etc are written to express the interaction between two objects. Although, there is not any specific list of these words, a frequent expression of a good number of words are observed in the functional site information related sentences. So, these words might be considered as important words or keywords for the functional site information related sentences and thus expressing a sentence in the literature into vector considering the presence of these keywords a functional site information related sentence might be more distinguishable than any other non-functional site information related sentence. In our research, we extracted automatically the high frequency words that are written in the sentences related to the functional site information and from those words, finally we selected 29 keywords manually as most probable candidates to appear in the sentences related to functional site information. When any sentence i is expressed by vector, its vector elements from x[i][1] to x[i][29] are allotted for these keywords and are expressed by the binary value "1" or "0" depending on the frequency of the corresponding keyword. Using the following equation we get the value of x[i][j] for $1 \leq j \leq 29$, where $f_{i,j}$ is the frequency of the keyword j in sen-

tence i .

$$x[i][j] = \begin{cases} 1 & (\mathbf{f}_{i,j} > 0) \\ 0 & (\mathbf{otherwise}) \end{cases}$$

4.4 Vector Expression using patterns

In the sentences related to protein functional site information the following three patterns are frequently observed. So, when a sentence is expressed by vector, three of its elements can be expressed by"1" or "0" considering whether these three patterns are written in that sentence or not.

1. <residue>[*] [VERB] [*] <object> or <object> [*] [VERB] [*] <residue>
   This pattern is frequently written in the sentence to express the interaction between any residue and its interaction partner. For identifying the verb we did morphological analysis by the Brill's tagger[7]. If this pattern is present in a sentence i , its vector element x[i][30] is expressed by the binary value "1", otherwise by "0".

2. between [*] <residue> and <object>
   Usually when an information about the interaction between two different objects are written in the sentence, the pattern above is written to express it. Considering the presence of this pattern in sentence i , the x[i][31] element of its vector expression is expressed by the binary value "1" or "0".

3. "residue"
   Protein functional site information are mostly the interaction information between residues and other objects on the structure. Thus, in a sentence related to functional site information, very frequently the word "residue" is written in it. So, the word "residue" can be a special feature for a sentence related to functional site information. When a sentence i is expressed by vector, its x[i][32] element is expressed by the binary value "1" or "0" considering the word "residue" is written in that sentence or not.

4.5 Distinguishing the named entities

This is the procedure of extracting functional site information from the sentences distinguished as sentences related to functional site information. 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.

5 Evaluation

We made the learning corpus using four literatures referred by PDB(1a0h, 1a0q, 1a26 and 1a3l). In the learning corpus total number of sentences were 964, out of which 71 sentences were labeled manually as correct sentence and then the learning corpus was learnt by the SVM tool. For learning we used "LIBSVM tool" [8].

Then, we evaluated the accuracy of the proposed method which uses the SVM tool to distinguish sentences related to functional site information, using the literature referred by PDB.

The PDB-ID of the protein and the number of words, sentences 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 satisfactory. But still there is a scope to obtain a higher value if the correct sentences which the system could not distinguish, are feedback to the module and learnt as correct sentences. Using feedback from the operator both recall and precision can be expected to obtain a higher value.

An example of a sentence mistakenly distinguished from a literature related to protein "1a5i", "As in human tPA , the side chain of Tyr151 residue extends into the S2' subsite". In which what the side chain of the residue "Tyr151" does has been described which has nothing to do with the functional site information. Besides, there were mistakenly distinguished sentences like, "Gly216 does not contribute to binding by hydrogen bond formation...." where straightly describes that the residue "Gly216" does not make a hydrogen bond.

6 Conclusion and Future Work

In this paper, we proposed a method to distinguish protein functional site information related sentences from literature for PROFESS. In the proposed method the SVM tool was used to carry out

Table 1: Literature data used in the experiment

PDB-ID	♯of words	♯of sente -nces	♯of correct sente -nces
1a5i	7638	274	48
1a5h	6859	295	30
1a5y	7331	290	20
1a5z	9521	427	3
1a3r	6988	298	17
1a5v	5769	276	13
2a2g	7701	364	7
Average	7401	317.71	19.71

Table 2: Results

PDB-ID	Precision	Recall	F-measure
1a5i	0.74	0.83	0.78
1a5h	0.51	0.50	0.50
1a5y	0.68	0.85	0.76
1a5z	0.67	0.67	0.67
1a3r	0.73	0.64	0.68
1a5v	0.68	0.84	0.75
2a2g	0.80	0.57	0.67
Average	0.68	0.72	0.70

the Learning Phase and distinguish the correct sentences from the incorrect senteces in the Testing Phase. The method was evaluated experimentally using seven documents related to structural analysis of protein. The average recall value and F value of extraction were 0.72 and 0.70 respectively, which confirms the effectiveness of the method.

Our future work will include improving the system's accuracy. For this, we will consider the feedback from the system operator. At present, any sentence in the literature is expressed by vector using the protein structure data, keywords and patterns where the keywords and patterns are prepared manually. The sentences which are not distinguished by the extraction module contain keywords and patterns which are not included in our prepared list. So, an automatic recognition of the functional site information related keywords and the syntactic patterns from the feedback sentences would be essential. Our perspective is to increase the total number of the vector elements dynamically which would cover any sentence related to functional site information and lead to a higher performance. Besides, the sentences which are not related to functional site information but are distinguished by the proposed method, need also to be considered for the sake of unalloyed information extraction by the system. For this, not only increasing the number of vector elements by adding new keywords and patterns but also to tune each of the vector element's weight so that the vector expression of any literature leads to an accurate and higher performance of distinguishing the correct sentences. The cyclic process of "Adding new vector elements to the learning corpus", "Tuning the weight of each of the vector elements", "Learning the corpus by SVM" and "Distinguishing the correct sentences by the SVM" can be expected to obtain a higher performance and achieve our goal.

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] Y.Kaneta, M.Numa and T.Ohkawa: "Automatic Extraction of Protein Active Site Information from Literature Using Template Matching and Anaphora Analysis," *Proc. The 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, pp.100-106 (2003).

[4] Y.Kaneta, Md.A.Munna and T.Ohkawa: "A Method for Extracting Sentences Related to Protein Interaction from Literature using a Structure Database," *Proc. Second Workshop on Data Mining and Text Mining for Bioinformatics (in conjunction with ECML/PKDD)*, pp.18-25 (2004).

[5] V.N Vapnik: "The Nature of Statistical Learning Theory," *Springer*, (1995).

[6] M.Numa, Y.Kaneta and T.Ohkawa: "Automatic Classification of Proper Names in Protein-related Literatures Using Database Retrieval on WWW," *Proc. 5th International Conference on Computational Biology and Genome Informatics (CBGI)*, pp.903-906 (2003).

[7] E. Brill: "Transformation-based error driven learning and natural language processing: A case study in parts of speech tagging," *Computational Linguistics*, 21, pp.543-565 (1995).

[8] C. Chang and C. Jen Lin: "LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/LIBSVM ,"