Method of Retrieving Articles on Protein Structure Analysis from User Intention

T. Aso and T. Ohkawa

***Abstract*—In recent years, the number of articles that describe protein structure and function is increasing. Since the articles are written using polysemic and complex terminologies, however, such naive retrieval methods as full text search often fail to find appropriate articles. We propose a new method in which the structural and functional concepts of proteins are considered using Gene Ontology (GO) and other databases. In our proposed method, articles of interest are submitted as queries to solve the problem caused by the ambiguity of the terminologies, and then articles similar to the query article are retrieved. In addition, giving another article as an additional query article clarifies the user intention and improves retrieval accuracy. The effectiveness of our proposed method was confirmed by evaluating its accuracy through retrieval experiments, especially for retrieving new articles.**

***Index Terms*—Article retrieval, ontology, similarity between articles, user intention.**

I. Introduction

The number of articles on protein structure analysis has reached about 30,000 and is increasing rapidly. Researchers spend much time finding articles from the huge amount of published articles. Hence, a system that helps them retrieve articles easily might solve this problem. Many methods retrieve articles by keywords; one is a full text search from abstracts or whole documents. However, it often fails to find appropriate articles because biological articles contain polysemic and complex terminologies.

In this study, we propose a method in which the structural and functional concepts of proteins are utilized to retrieve articles on protein structure analysis by identifying user intentions. As mentioned above, keyword-based retrieval cannot get high accuracy. To obtain better retrieval results, we focus not on keyword-based but article-based retrieval, in which an article of interested is considered a query to retrieve articles that resemble the query article. Since articles on protein structure analysis are annotated with many concepts (e.g., target protein, disease, species, etc.), the retrieval results reflect the aspects from which users intend to refer to query article.

In our method, multiple articles are treated as a query to identify the user intention that is evaluated by calculating the similarity between query articles.

In a previous study, we proposed a similar method [1], where only limited concepts that resemble the concepts observed in the input articles were used to evaluate user intentions. All of the generalized concepts about input articles are used in our new method, where concepts similar to input articles are considered more important sources of user intentions than in our previous method. Our proposed method introduces a filtering mechanism that can only extract candidate articles from target articles for faster and more accurate retrieval, which has not been mentioned in [1].

Manuscript received November 14, 2012; revised February 15, 2013. This work was partly supported by Grant-in-Aid for Scientific Research (B) (24300056) from MEXT.

The authors are with the Graduate School of System Informatics, Kobe University, Kobe, Japan (e-mail: t-aso@cs25.scitec.kobe-u.ac.jp, ohkawa@kobe-u.ac.jp).

II. Methods

A. Articles on Protein Structure Analysis and Related Databases

Retrieval our targets are articles that describe protein structure analysis from the protein structure database (PDB[1]) entries. An article contains much information from such aspects as biological activities, protein structures, protein functions, and experimental methods. Such a wealth of information can be obtained not only from the articles themselves but also from the databases related to them.

In this study, we use the following biological databases to evaluate the relationships among articles from various aspects:

- Biomedical articles (MEDLINE, PubMed)[2] [2]
- Consistent descriptions to annotate gene products (Gene Ontology/GO) [3]
- Protein structure information (Protein Data Bank/PDB) [4]
- Protein domains, families, and functional sites (PROSITE) [5]
- Protein sequence and function information (UniProt) [6]

Since no naming schema exists for biological terminology, researchers often use idiomatic terminology in their own fields. For this reason, unified concepts or terms for representing articles are required for article-based retrieval. GO is one significant database for providing unified biomedical terminology. GO terms are introduced to uniquely represent biomedical concepts, and the relationship among GO terms is also defined.

This relationship is provided as a parent-child in which the parent term implies a broader concept than the child term, and the child term is more specific than the parent term. In such a manner, GO is represented by directed acyclic graphs, called GO DAGs, where nodes and edges correspond to GO terms and their relationships. A GO DAG consists of three graphs, each of which has one of the following terms as a root: biological process, cellular component, and molecular

[1] http://www.pdb.org/pdb/home/home.do
[2] http://www.ncbi.nlm.nih.gov/pubmed/

function.

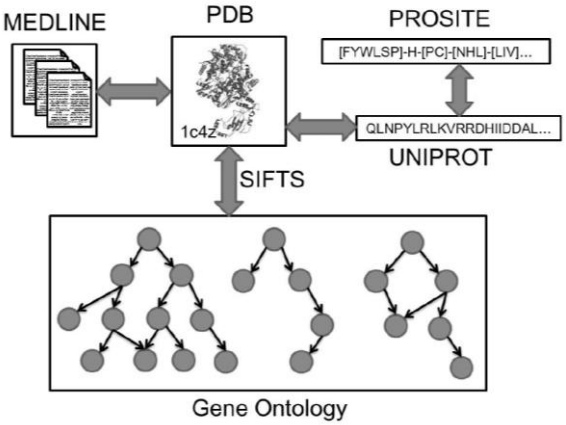

Fig. 1. Cooperative use of various databases.

PROSITE, which is a database of protein domains, families, and functional sites and consists of many biologically significant patterns or motifs, is a useful tool to identify to which known protein family a new sequence belongs. The sequence information of proteins is collected in a database called UniProt, which correlates the protein sequence with function information to utilize sequence patterns in PROSITE.

Fig. 1 shows an overview of the cooperative use of various databases.

B. Outline of Proposed Method

Fig. 2 outlines our proposed method for similar article retrieval. In this framework, a query is submitted as articles on protein structure analysis, and the articles that are similar to the query articles are retrieved.

Since the similarity between articles depends on the viewpoint from which the user refers to the query articles, however, we must identify the user intention before similarity calculation. For such identification, a query is submitted as two articles: primary and additional. A primary article functions as a key, which means that articles similar to it are retrieved. An additional article is another query article that is considered similar to the primary article by the users, whose intention can be identified by extracting the concepts that are included both in the primary and additional articles.

Since each PDB entry is annotated with one or more GO terms, the article referred from the PDB entry is also related to these GO terms. Therefore, the similarity between articles can be evaluated with the similarity between sets of GO terms, each of which is related to the article. If two articles, namely, a primary article and an additional article, are submitted as a query, the concept that is included in both of the submitted query articles, which is defined as a pair of similar GO terms related to the articles, is extracted to identify the user intention. The extracted concept affects the calculation of the similarity between the query and each of target articles. Topological path in GO DAG

We use the topological path [7] of GO DAG to calculate the semantic similarity between two nodes. The topological path between two GO terms, t_1 and t_2, consists of two paths between t_c, which is a common ancestor of t_1 and t_2, and each of the two nodes, where the sum of the length of the paths is minimum. Based on this definition, for example, the topological path between GO:0006629 and GO:0044237 is expressed with deep black arrows (Fig. 3).

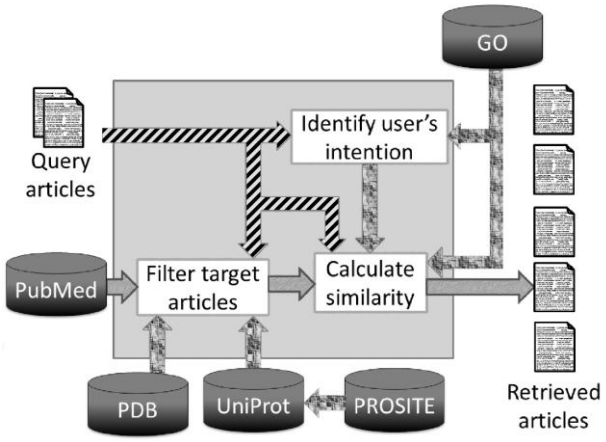

Fig. 2. Outline of similar article retrieval.

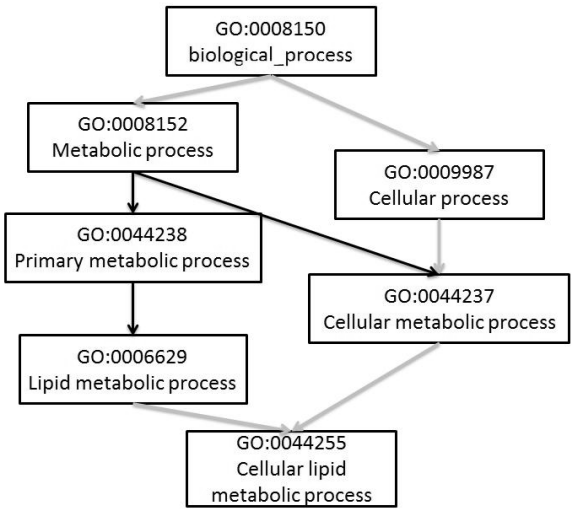

Fig. 3. Topological path between two GO terms, GO:0006629 and GO:0044237.

C. Semantic Similarity Using Ontology

1) Identification of user intention

Users have intentions when they submit a query to retrieve articles. Even if the same query article is submitted, desirable retrieval results depend on individual user intentions, because the query article can be interpreted from various aspects. Therefore, the retrieval results should be arranged based on the user intention that is identified by extracting similar concepts from more than one query article. If a user submits two articles as a query, the concepts in them have to share an intention. In other words, the GO terms related to these articles should be close to each other, which can be formalized by giving a length that is less than 1.0 to the edges of the topological paths among the terms related to the query articles. Shortened edges affect the calculation of the semantic similarity between articles (the query article and the target articles). As a result, appropriate retrieval results can be provided that reflect user intentions.

Let T_p be a set of terms used to annotate a primary article and let T_a be a set of terms used to annotate an additional query article. Let t_p be a term in T_p and t_a be a term in T_a. The

ancestor edges of t_p and t_a are the candidates to be shortened, where the edges close to t_p or t_a are shortened, but remote edges are not. The length of an edge is calculated as the average of the edge length shortened using each GO term that is used to annotate the proteins of query articles. $len^*(p,t,T_a,T_p)$, which is the length of edge p that is shortened by term t in $T_p \cup T_a$, is defined as follows:

$$len^*(p,t,T_a,T_p) = \begin{cases} 1 - \frac{\omega}{2}\left(\frac{1}{curv * l(t,t_1)^2 + 1} + \frac{1}{curv * l(t,t_2)^2 + 1}\right) \\ \qquad if\ t_1, t_2 \in isUp(T_a) \cup isUp(T_p) \\ 1 \qquad otherwise \end{cases} \quad (1)$$

where t_1 and t_2 are both ends of edge p and $isUp(T)$ is a set of all the ancestor terms in T. $l(t, t_i)$ $(i=1, 2)$ is the number of edges on the topological path of t and t_i. ω and $curv$ are pre-defined parameters. ω represents the maximum value of the length from 0.0 to 1.0. When ω is 0.0, len^* is always 1.0; in this case, the retrieval does not consider the user intention. When $curv$ is a large number, a narrow range of edges around t is shortened. To strongly shorten the edges around t_1 or t_2, l of denominator is squared.

$len(p,T_p,T_a)$, which is the length of edge p shortened by the set of GO terms T_p and T_a, is defined as follows:

$$len(p,T_p,T_a) = \frac{\sum_{t \in T_p, T_a} 1 - len^*(p,t,T_p,T_a)}{|T_p| + |T_a|} \quad (2)$$

where $|T|$ is the size of set T.

2) Length of topological path between two concepts

Before calculating the similarity between articles, note how we calculate the semantic similarity between two GO terms. Let t_p be one of the GO terms that is used to annotate a primary article, let t_t be one of the GO terms that is used to annotate a target article, $dis(t_p,t_t,T_p,T_a)$, which is the semantic distance between t_p and t_t, provided that some edges are shortened by T_p and T_a, as follows:

$$dis(t_p,t_t,T_p,T_a) = \min_P \sum_{p \in P} len(p,T_p,T_a) \quad (3)$$

where P is a set of edges comprising the topological path between t_p and t_t. The length of edge p is shortened based on the user intention introduced in the previously mentioned method.

$SimGO(t_p, t_t, T_p, T_a)$, which is the semantic similarity between t_p and t_t, provided that some edges are shortened by T_p and T_a, is defined as follows:

$$Sim^{GO}(t_p,t_t,T_p,T_a) = \begin{cases} \frac{1}{dis(t_p,t_t,T_p,T_a)} & if\ t_t \in c(t_p) \\ 0 & otherwise \end{cases} \quad (4)$$

where $c(t)$ is a set of all the terms in the connected DAG including term t.

3) Semantic similarity between sets of terms

Most articles are annotated with more than one term. The similarity between articles can be evaluated based on the similarity between sets of terms, each of which is related to the article. Let T_t be a set of terms used to annotate a target article. $SimGOs(T_p,T_t,T_a)$, which is the semantic similarity between sets of terms T_p and T_t when the sets of terms T_p and T_a are given as query articles, is defined as follows [8]:

$$Sim^{GOs}(T_p,T_t,T_a) = \frac{\sum_{t_p \in T_p} \max_{t_t \in T_t} Sim^{GO}(t_p,t_t,T_p,T_a)}{|T_p|} \quad (5)$$

where $|T_p|$ is the size of set T_p.

D. Filtering Articles

The retrieval targets are all the articles that are cited from each entry in PDB. The number of articles exceeds 30,000 and continues to increase. It is meaningless to calculate the semantic similarity between the query articles and articles that don't seem to be related to them.

The target articles are extracted by filtering for speed and accuracy. The proteins with similar sequence patterns are classified into the same groups in the PROSITE database. In the first filter, the articles that describe the structure analysis of proteins in the same family as the proteins treated in the query article are extracted as target articles.

On the other hand, some protein families only contain a few proteins. In this case, users cannot get enough results. Accordingly, the PDB Descriptor, which is a keyword that represents the features of the biological function and the structure used by PDB, is used to increase the target articles. Table I shows examples of the PDB Descriptor. The articles related to proteins with the same words as the proteins in the query articles in the PDB Descriptor are added to the retrieval targets.

III. Results and Discussion

The target articles are all those that are registered in PubMed and cited from PDB. The number of articles as the retrieval target is 28,133. We conducted retrieval experiments on a single computer with an Intel Core i7 950 (quad core), 3.06 GHz, and 12 GB RAM. Our proposed method was implemented using Java programming language.

TABLE I: Example of PDB Descriptor

PDB ID	PDB Descriptor
1c4z	UBIQUITIN - PROTEIN LIGASE E3A / UBIQUITIN CONJUGATING ENZYME E2
1yh2	HSPC150 protein similar to *ubiquitin-conjugating* enzyme
1e0d	UDP-N-ACETYLMURAMOYLALANINE--D-GLUTAMATE *LIGASE*
1dgs	DNA *LIGASE* FROM T. FILIFORMIS
1dhp	DIHYDRODIPICOLINATE SYNTHASE
2oni	E3 *ubiquitin-protein* ligase NEDD4-like protein (E.C.6.3.2.-)
1uby	FARNESYL DIPHOSPHATE SYNTHASE, DIMETHYLALLYL DIPHOSPHATE

A. Retrieval Example

Table II shows an example of a retrieval result when the query consists of only one article: "Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the

E2-E3 enzyme cascade (PDB: 1c4z)." This Table I shows the top five article retrieval results.

TABLE II: TOP FIVE OUTPUT ARTICLES WHEN INPUT ONLY ONE QUERY; PDB: 1C4Z

PDB ID	Title
1z5s	Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex.
1fbv	Structure of a c-Cbl-UbcH7 complex: RING domain function in ubiquitin-protein ligases.
2nvu	Basis for a ubiquitin-like protein thioester switch toggling E1-E2 affinity.
2c2v	Chaperoned ubiquitylation-crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex.
2grn	Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway.

In this table, the titles of PDB: 1z5s and PDB: 2grn contain the word, SUMO, which is the abbreviation of Small Ubiquitin-related (like) Modifier, and is strongly related to Ubiquitin, which is found in the title of PDB: 1c4z. Such words as SUMO are difficult to be detected from the keyword "Ubiquitin" by a full text search. The articles that are related to the query article were retrieved successfully.

B. Retrieval Result from Primary and Additional Query Articles

To evaluate the accuracy of our retrieval results, a set of correct articles must be prepared for each query article. We constructed sets of correct articles based on citation information obtained from CiteSeer [9] in the following manner. Articles that were cited by an article (not limited to articles on protein structure analysis) that also refers to all of the query articles were selected as the set of correct articles for the query articles.

TABLE III: QUERY ARTICLES SETS

Primary	Add	Primary	Add	Primary	Add
1c4z	1ayz	1fbv	1d5f	1ldk	1d5f
1c4z	1fbv	1fbv	1fqv	1ldk	1fbv
1c4z	1fxt	1fbv	1fxt	1ldk	1fqv
1c4z	1kps	1fbv	1ldk	1ldk	1nex
1c4z	1nd7	1fbv	1nd7	1ldk	1p22
1c4z	1u9a	1fbv	2e2c	1ldk	1u6g
1c4z	1y8q	1fbv	2esk	1ldk	1vcb

Table III shows the query article set in our experiment. The query set is prepared as pairs of one of three articles (primary articles) selected randomly and an article (additional article) that is related to the primary article in the PubMed database.

The retrieved articles are ranked by similarity scores. The retrieval accuracy is evaluated using Mean Average Precision (MAP). Parameters ω and *curv* are changed from 0.0 to 1.0. Fig. 4 shows the MAP values of the retrieval results. When ω is 1.0 and *curv* is 0.2, a maximum value of MAP (0.5375) is obtained. When ω is 0.0, which means that no user intention was considered, the MAP value is 0.5204.

The characteristic of our proposed method are clarified by comparing it with a simple citation-based method, where the score of each article is defined as the number of articles referred from or that refer to both this article and query articles.

Each article is ranked by its score. In this method, the MAP value is 0.7296, which is higher than the value in our method. On the other hand, the citation-based method assumes the existence of articles published after the query articles were published, but this assumption is not always valid if the query articles are recent. If newer articles than the query articles are unavailable, the MAP value decreases to 0.5379, even in the citation-based method. This MAP value remains slightly higher than the value in our method. Since the sets of correct articles have been constructed based on the citation information, however, the citation-based method tends to be overestimated in comparison with our method. Although our method uses no citation information, the results both of our method and of the citation-based method are almost the same, which suggests that our method has enough accuracy to retrieve new articles.

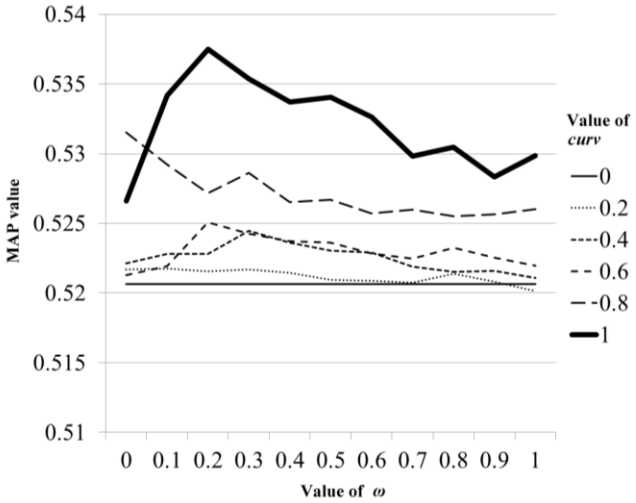

Fig. 4. Mean average precision of retrieval results.

The previous method [1] only shortens a set of edges that comprise the topological path between a term in the primary article and a term in the additional article. Moreover, the edges around the terms in the input articles are shortened less than our method. As a result, the MAP value of the previous method for the same input articles as the above experiment is 0.3982, which shows less accuracy than our new method presented in this paper.

C. Filtering Results

The target articles were filtered out in the preprocessing stage. Table IV shows the number of remaining articles after filtering. Fig. 5 compares the results with/without filtering for the query articles (Table III). Fig. 5 (a) and (b) show the macro averages of the recall at top *k* and of precision at top *k*, respectively. Filtering improved the retrieval accuracy in terms of both recall and precision. While it takes about 27.3 minutes to calculate the retrieval result without filtering, filtering reduced the calculation time to about 14.2 seconds on average for all queries in Table III.

TABLE IV: NUMBER OF EXTRACTED ARTICLES

Primary articles	Extracted by		
	PROSITE	PDB Descriptor	Total (percentage of extracted articles)
1c4z	76	73	127 (0.45%)
1fbv	98	83	166 (0.59%)
1ldk	42	232	273 (0.97%)

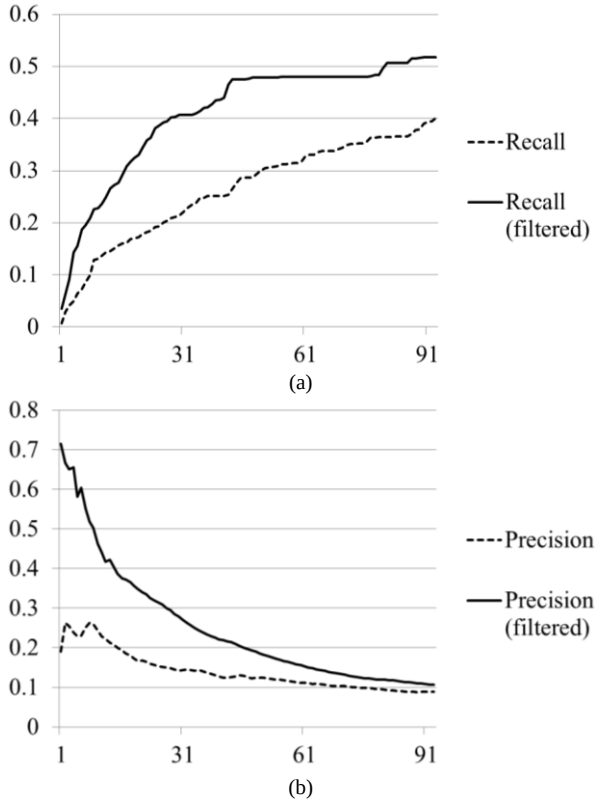

Fig. 5. The value of recall and precision with/without filtering.

IV. Conclusions

In this paper, we present a new method of similar-article retrieval that considered user intention using functional information from Gene Ontology (GO). Our proposed method has the following notable features:

1) Retrieval based on function information over GO is more convenient than a full text search.
2) An additional query article helps identify user intention to obtain articles that the user really wants.
3) Filtering target articles improves retrieval accuracy and efficiency.

On the other hand, if a protein is annotated with just a few GO terms, our proposed definition of similarity between query and each of target articles is too sensitive to update similarity values based on user intention. To cope with this situation, the introduction of a modified definition of similarity is one important subject.

Moreover, GO consists of three separated graphs, but the similarity among terms is evaluated using only one of the three in our proposed framework. Integrating the three directed acyclic graphs will provide more fruitful similarity evaluation schema.

In our experiments, the database in which the retrieval targets are stored is relatively large, but the query data sets are very small. To clarify the effectiveness of our proposed method, larger scale experiments are required, in which is compared with other similar methods using larger query data sets. This is one important future works. In addition, the method of constructing sets of correct articles should be improved because the current temporary method overestimates the links of citations.

References

[1] R. Kyogoku, R. Fujimoto, T. Ozaki, and T. Okawa, "A method for supporting retrieval of articles on protein structure analysis considering user's intention," *BMC Bioinformatics*, vol. 12, Suppl 1, S42, Feb. 2011.
[2] J. McEntyre and D. Lipman, "PubMed: bridging the information gap," *CMAJ*, vol. 164, no. 9, pp. 1317-1319, May 2001.
[3] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25-29, May 2000.
[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235-242, Jan. 2000.
[5] C. J. A. Sigrist, L. Cerutti, E. Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo, "PROSITE, a protein domain database for functional characterization and annotation," *Nucleic Acids Research*, vol. 38(Database issue), pp. D161-166, Jan. 2010.
[6] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40(Database issue), pp. D71-5, Jan. 2012.
[7] H. Yu, R. Jansen, G. Stolovitzky, and M. Gerstein, "Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications," *Bioinformatics*, vol. 23, no. 16, pp. 2163-2173, May 2007.
[8] A. Zhang, *Protein Interaction Networks*, New York: Cambridge University Press, 2009, pp. 216-242.
[9] C. L. Giles, K. D. Bollacker, and S. Lawrence. "Citeseer: an automatic citation indexing system," in *Proc. 1998 the third ACM conference*, pp. 89-98.

Tomoki Aso obtained his Bachelor of Engineering from Kobe University in 2011. He is currently a student of the master course at Graduate School of System Informatics, Kobe University (Kobe, Japan). His research interests include intelligent data processing and bioinformatics.

Takenao Ohkawa received his B.E, M.E., and Ph.D. degrees from Osaka University in 1986, 1988, and 1992, respectively. He is currently a professor in the Department of Information Science, Graduate School of System Informatics, Kobe University. His research interests include intelligent data processing and bioinformatics. He is a member of the IEEE, the Information Processing Society of Japan, the Japanese Society for Bioinformatics, the Institute of Electronics, Information, and Communication Engineers, the Institute of Electrical Engineers in Japan, and the Japanese Society for Artificial Intelligence.