International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2004, Orlando

Automated Template Discovery for Information Extraction from Biomedical Literature

Satoshi Kamegai
Advanced Research Group, INTEC Web and Genome Informatics Corporation,
3-23, Shimoshin-machi, Toyama 930-0804, Japan
Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation,
5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Kenji Satou
School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation,
5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Akihiko Konagaya
Genome Science Center, RIKEN Yokohama Institute,
1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

ABSTRACT

We propose a method to automatically extract templates from biomedical literature without background knowledge. The proposed method automatically extracts verbs and templates indicating interactions between biomolecules with a large dictionary called an extensional ontology. We applied our method to two datasets: one comprised 299 full texts from *Cell* (1998–2002) and 13,818 entries from OMIM (Online Mendelian Inheritance in Man); the other included 33,622 abstracts from Medline (2002). Experimental results showed that our method could extract verbs and templates that had been manually collected in related works. For extracting templates, our method only needs to prepare ontology (or dictionary) and a large body of texts. Consequently, it can be applied to those of other fields as well as the biomedical literature.

Keywords: Ontology, Information Extraction, Text Mining

1. INTRODUCTION

Extracting information from biomedical literature, especially that addressing interactions between biomolecules (e.g., protein–protein interaction) is important for advancing genome analysis research. Many methods for extracting such information from the literature have been investigated so far.

The most powerful method of extracting information on the interactions between biomolecules from the biomedical literature is to extract named entities (NEs) representing biomolecules and the verb representing their relationship. To recognize NEs, natural language processing (NLP)- and dictionary (or ontology)-based approaches are being tried. Fukuda et al. [1] proposed a method of extracting NEs (e.g., protein names) through an NLP approach, using surface clues of character strings. With this method, they succeeded in obtaining NEs without using any background knowledge. The NLP approach can recognize unknown words and coinages. In comparison, the method of using NEs in a dictionary- or ontology-based approach as proposed by Rindflesch [2] could be performed at low computation cost because NEs were recognized by simple matching. However, if the dictionary or ontology is not sufficiently large and up-to-date, the analysis will fail to recognize important NEs.

Much of the past research on the extraction of biomolecular interaction adopted template-matching a pproaches. That is, first NEs were extracted using methods introduced above, then NE–verb–NE sequences were extracted that contained verbs included in a list previously prepared by domain experts. Using a template-matching approach, Sekimizu et al. [3] retrieved a corpus of around one million words from Medline abstracts. They then adopted a shallow parsing technology using a system called EngCG from Lingsoft to find subject and object terms for frequently seen verbs (e.g., *activate*, *bind*, *interact*, *regulate*, *encode*, *signal*, and *function*) and saved the resulting information as sentence-like assertions in a database. Consequently, some frequently seen verbs were extracted as indicative of interactions between genes and gene products. Thomas et al. [4] analyzed around 200 abstracts without the aid of computer programs to find common ways of describing interactions. Approximately 30 different verbs including *activate*, *inhibit*, *modulate*, *suppress*, *isolate*, *promote*, and *characterize* were examined, and three templates—*interact (with)*, *associate (with)*, and *bind (to)*—were considered to indicate protein–protein interactions. These templates then were used to extract information on protein–protein interactions. However, because of the wide variety of expressions that represent interactions between biomolecules, it is practically impossible to manually prepare all the necessary verb lists or templates for extracting these interactions by only domain experts.

Table 1: Our selected NEs from categories by Yagyuu.

Category	Database: Field
organism	*[1]GenBank: organism
	GenBank: variety
	GenBank: lab_host
	GenBank: specific_host
	GenBank: sub_species
	*[2]RefSeq: organism
	RefSeq: variety
	RefSeq:lab_host
	RefSeq:specific_host
	RefSeq: sub_species
	*[3]GENOME: NAME
	GENOME: DEFINITION
	*[4]PMD: SOURCE
	PMD: EXPRESSION-SYSTEM
	*[5]Swiss-Prot: OS

Category	Database: Field
organism	*[6]BRITE: ORGANISM
	*[7]EPD: OS
	*[8]TRANSFAC: OS
organism class	GENOME: LINEAGE
	Swiss-Prot: OC
	TRANSFAC: OC
protein	GenBank: product
	RefSeq: product
	PMD: PROTEIN
	TRANSFAC: DE
	*[9]ENZYME: NAME
	*[10]PRF: NAME
compound	*[11]COMPOUND: NAME
gene	GenBank: gene
	RefSeq: gene

*[1] GenBank, http://www.ncbi.nih.gov/Genbank/index.html
*[2]RefSeq (Reference Sequences), http://www.ncbi.nlm.nih.gov/RefSeq/
*[3]GENOME, http://www.genome.ad.jp/dbget-bin/www_bfind?genome
*[4]PMD (Protein Mutant Database), http://pmd.ddbj.nig.ac.jp/
*[5]Swiss-Prot http://kr.expasy.org/sprot/
*[6]BRITE (Biomolecular Relations in Information Transmission and Expression), http://www.genome.ad.jp/brite/
*[7]EPD (The Eukaryotic Promoter Database), http://www.epd.isb-sib.ch/
*[8]TRANSFAC http://www.gene-regulation.com/
*[9]ENZYME http://kr.expasy.org/enzyme/
*[10]PRF (Protein Research Foundation), http://www.prf.or.jp/en/
*[11]COMPOUND http://www.genome.ad.jp/dbget-bin/www_bfind?compound

To extract information on the relationships among biomolecules, we adopted an ontology-based approach to NE recognition. Biological ontology is one of the most important and interesting subjects in today's bioinformatics. Efforts have already been focused on the need to construct biological ontology (TaO [5], Gene Ontology [6], EcoCyc [7], etc.) and to develop tools (GKB-Editor [8]). These efforts are bearing fruit; however, the basic philosophy of a biological ontology is oriented toward the construction of a reliable and carefully screened hierarchy of biological concepts by domain experts. For this reason, building a large amount of biological ontology is difficult. In contrast, by collecting NEs from biological databases in GenomeNet, Yagyuu et al. [9] have constructed an extensional ontology database that, although not well organized yet, covers nearly 2,000,000 NEs. In this paper, we propose a method for automatically extracting templates from the biomedical literature by using this large body of NEs from the extensional ontology.

2. MATERIALS AND METHODS

Our approach to extract verbs and templates indicative of biomolecular interaction proceeds as follows.

a) Filtering extensional ontology NEs
The extensional ontology can provide a massive number of NEs, but it contains many terms whose categories are not clear. For example, a term taken from the keyword field (KW) of the SWISS-PROT database can be a protein, a gene, a function, a concept, and so on. To concentrate on the extraction of relationships among substantial objects in biology and medical science, we filtered extensional ontology NEs (Table 1) based on the categorization performed by Yagyuu et al. We selected five categories (*organism*, *organism class*, *protein*, *compound*, and *gene*) that we expected to consist mainly of NEs for biomedical substances. Consequently, we extracted 1,082,830 NEs from the extensional ontology.

b) Extracting the interval between two NEs
We surmised that between two NEs in a sentence, a word (typically a verb) characterizing their interaction often occurs. From this viewpoint, by simple matching of NEs and given texts (e.g., abstracts), sequences of words (so-called intervals) between two NEs were extracted. In the example in Figure 1, four NEs are bolded, three intervals are underlined, and three important words characterizing (three) relationships are italicized.

finally it was discovered recently that **apc** *binds* to **asef** an exchange factor that apparently *activates* the **small g protein rac** which in turn *controls* the **actin** cytoskeleton kawasaki et al. 2000

Figure 1: Extraction of interval

The texts used for interval extraction were:

Dataset 1—299 complete articles (full texts) from *Cell* (1998–2002) and 13,818 entries from OMIM [10]
Dataset 2—33,622 abstracts from Medline (2002)

To avoid problems associated with case and special characters in NEs and texts, we converted all letters to lowercase and removed (converted to white spaces) all special characters. In addition, the following grammatical words were converted into general terms (Table 2).

Table 2: General terms and Grammatical words

General terms	Grammatical words
ARTICLE	a, an, the
RELATIVE	who, whose, whom, …
PRONOUN	it, they, itself , …
AUXILLARY VERB	can, might, shall, …
CONJUNCTION	and, or, but , …
PREPOSITION	after, from, in, …
BE	is, are, was, …
HAVE	have, has, had, having

c) Extracting specific and frequent words

The next problem is finding the important word in an interval. We expected that such a word would occur frequently and specifically in the interval. We then evaluated each of the words in intervals (except for one- or two-letter words and non-alphabetical terms) by using Equation 1, where *a* denotes how many times a word occurs in intervals, and *b* denotes how many times the word occurs in texts.

$$\frac{a \times \log(a)}{b} \times 100 \qquad (1)$$

Words obtaining high scores tend to be frequent and specific in intervals. We adopted the 100 highest-scoring words (excluding general terms) for the next step.

d) Template extraction

Template extraction was performed as follows:

1. Input a word as an initial template, whose length is one.
2. Retrieve templates containing the word from all intervals.
   (e.g., word = *bind*, interval = *would bind to*
   ⟶ template = *bind to*, *would bind*, *would bind to*)
3. Evaluate each template by using Equation 2:
   $$\frac{a \times \log(a)}{b} \times 100 + \log(c) \times 100 \qquad (2)$$
   where *a*, *b*, and *c* denote how many times a template occurs in intervals, how many times the template occurs in texts, and the length of template (e.g., *bind to* = 2), respectively.
4. Extract the highest-scoring template.

3. EXPERIMENTAL RESULTS

Using the previously described method, we ordered all words in intervals in datasets 1 (32,952 words; Table 3) and 2 (108,888 words; Table 4). After the word ranking, we adopted the 100 top-scoring words (excluding general terms) for subsequent template extraction (Tables 5 and 6). Finally, we analyzed the extracted templates with regard to the categories of NEs (*organism*, *organism class*, *protein*, *compound*, and *gene*). Two examples follow:

A is regulated by B
gene gene

A is bound to B
protein protein

where A and B are NEs. If a template specifically occurs in intervals between NEs for genes (proteins), the template might be useful to extract relationships between genes (proteins). Some examples are shown in Figures 2 and 3.

Table 3: Top words from intervals in dataset 1

Evaluation	Word	Evaluation	Word
477	roychoudhury	298	bound
413	CONJUNCTION	298	product
411	hybrids	296	required
403	PREPOSITION	295	via
402	deficient	294	telomeric
359	symbolized	294	site
351	binds	293	codes
349	ARTICLE	293	coded
345	located	293	inhibits
333	RELATIVE	292	catalyzes
333	ceacam	292	induced
324	encoded	291	stimulates
324	chromosome	291	express
320	bind	290	chains
319	homolog	289	increase
318	produced	288	tabulated
316	lacking	288	produce
314	electrophoresis	288	resulting
311	carrying	288	prime
308	mediated	288	deficiency
307	homologous	287	hybridize
305	activates	287	substitution
301	containing	287	encodes
299	activate	286	ternary
299	expressing	286	promotes

Table 4: Top words from intervals in dataset 2

Evaluation	Word	Evaluation	Word
598	encodes	488	express
587	deficient	477	blocked
569	mediated	477	induces
556	binds	474	encoded
552	induced	471	plays
545	expressing	470	via
538	encoding	469	mediates
531	phosphorylation	469	stimulates
530	catalyzes	469	regulate
528	stimulated	467	reporter
523	suggesting	467	signaling
522	regulates	467	PREPOSITION
521	inhibited	465	bound
518	nick	464	phosphorylated
515	interacts	463	homologue
515	expression	459	member
509	inhibits	459	inhibit
506	containing	458	activate
501	CONJUNCTION	458	inhibitors
498	expressed	454	transfected
496	kinases	454	regulated
493	activates	454	production
493	bind	452	antagonist
491	activation	452	indicating
489	dodecyl	452	promoter

Table 5: Template with highest score for each word from intervals in dataset 1

Word	Extracted template
hybrids	hybrids PREPOSITION
deficient	deficient PREPOSITION ARTICLE
symbolized	BE CONJUNCTION symbolized
binds	binds PREPOSITION ARTICLE
located	BE located PREPOSITION
encoded	BE encoded PREPOSITION ARTICLE single
bind	PREPOSITION bind PREPOSITION
produced	produced PREPOSITION
lacking	lacking ARTICLE
electrophoresis	electrophoresis CONJUNCTION
carrying	carrying ARTICLE
mediated	mediated cleavage PREPOSITION
activates	activates ARTICLE
activate	PREPOSITION activate
expressing	expressing ARTICLE
bound	BE bound PREPOSITION
required	BE required CONJUNCTION
codes	codes CONJUNCTION

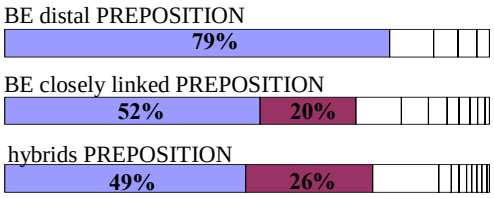

Figure 2: Examples of templates with high specificity to categories of NEs

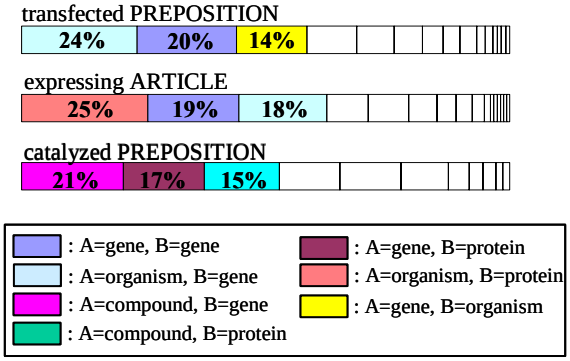

Figure 3: Examples of templates with low specificity to categories of NEs

4. EVALUATION

We compared the words and templates we extracted with those of Thomas et al. [3] and Sekimizu et al. [4]. First, we compared the results obtained by Thomas et al., who manually extracted verbs and templates indicative of protein–protein interactions (e.g., *interact with*, *associate with*, and *bind to*), with those we obtained.

Table 6: Template with highest score for each word from intervals in dataset 2

Word	Extracted template
encodes	encodes ARTICLE
deficient	deficient PREPOSITION
mediated	mediated PREPOSITION
binds	binds PREPOSITION
induced	induced activation PREPOSITION
expressing	expressing ARTICLE
encoding	encoding ARTICLE
phosphorylation	phosphorylation PREPOSITION
catalyzes	catalyzes ARTICLE
stimulated	stimulated PREPOSITION
regulates	regulates ARTICLE
inhibited	BE inhibited PREPOSITION
interacts	interacts PREPOSITION
expression	expression CONJUNCTION
inhibits	inhibits ARTICLE
containing	containing ARTICLE
expressed	expressed PREPOSITION
activates	activates ARTICLE
bind	bind PREPOSITION

Table 7: Evaluation of the verbs extracted by Thomas et al.

Thomas's stem words	Our words from dataset 1	Our words from dataset 2
interact	interacts (56) interaction (329) interact (589) interactions (603)	interacts (15) interact (81) interaction (294) interacted (378) interacting (460) interactions (561)
associate	associated (167) associates (355) associate (500)	associates (137) associated (372) associate (435)
bind	binds (5) bind (10) binding (719)	binds (4) bind (22) binding (132)

Numbers in parenthesis express the score ranking by our method.

Table 8: Comparison of the templates extracted by Thomas et al.

Templates by Thomas	Our templates from dataset 1	Our templates from dataset 2
interact with	CONJUNCTION interacts PREPOSITOIN ARTICLE	interacts PREPOSITION PREPOSITON interact PREPOSITION
bind to	PREPOSITION bind PREPOSITION binds PREPOSITION ARTICLE	bind PREPOSITION binds PREPOSITION

Table 7 shows the words extracted by Thomas et al. and their score ranking by our method. We see in Table 7 that the words related to Thomas's stem words (e.g., *interact*, *associate*, and *bind*) have higher ranks than other word ranks. Next, we investigated templates extracted by our method that contain *bind* and *interact* (Table 8). Figure 4 shows the breakdown of

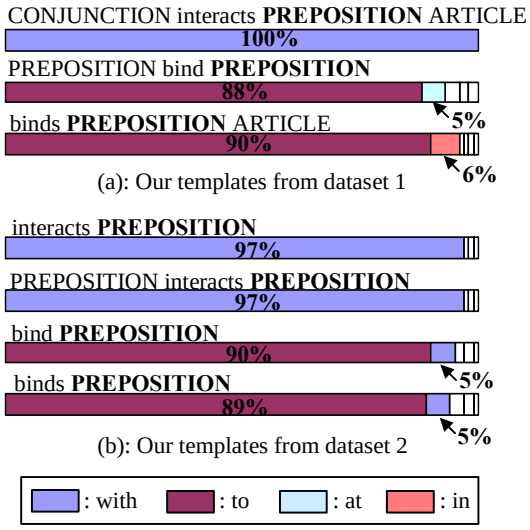

Figure 4: Ratio of PREPOSITON

Table 9: Evaluation of the verbs extracted by Sekimizu et al.

Sekimizu's stem words	Our words from dataset 1	Our words from dataset 2
activate	activates (18) activate (20) activation (61) activators (703) activating (972)	activate (21) activation (23) activates (41) activating (114) activators (483)
bind	see Table 7	
interact	see Table 7	
regulate	regulate (57) regulates (71) regulated (212) regulation (278) regulating (502) regulatory (975)	regulates (12) regulate (33) regulated (44) regulating (54) regulation (59) regulators (222) regulatory (426)
encode	encoded (8) encodes (44) encoding (215) encode (383)	encodes (1) encoding (7) encoded (28) encode (154)
signal	signaling (165) signals (953)	signaling (35) signals (849)
function	function (214) functions (538) functional (694) functionally (834)	function (434) functions (519) functionally (783)

prepositions in the templates of Table 8. In most cases, our words for **PREPOSITION** agree with those in the templates discovered by Thomas et al., and this agreement demonstrates the power of our method.

Second, we checked the order of the words extracted by Sekimizu et al., who extracted the verbs (e.g., activate, regulate, encode, and so on) indicative of relationship between genes and gene products (Table 9). We found that the words related to most of Sekimizu's stem words (*activate*, *bind*, *interact*, *regulate*, *encode*, and *signal*) have higher ranks too. The rank of *function* is not high because this word is often used as a noun as well as a verb. Therefore, because this word appeared in non-interval as well as interval regions, it ranked lower.

Table 10: Newly extracted verbs

Stem word	Extracted word from dataset 1	Extracted word from dataset 2
mediate	mediated (16) mediates (74)	mediated (3) mediates (31)
express	expressing (21) express (34)	expressing (6) express (25)
contain	containing (19)	containing (18)
induce	induced (32) induces (78) induce (87)	induced (5) induces (27) induce (57)
catalyze	catalyzes (31)	catalyzes (9)
inhibit	inhibits (32) inhibit (91)	inhibits (17) inhibit (40)
stimulate	stimulated (50) stimulates (33) stimulate (67)	stimulated (10) stimulates (32) stimulate (94)
lack	lacking (13)	lacking (99)
release	release (58)	release(58)
promote	promotes (46) promoter (95)	promotes (48) promoter (90)
culture	cultured (64)	cultured (95)

Compared with related works, our experimental results showed that our approach could extract verbs and templates that were manually discovered by others. However, we can see that the success of our method depends on the quality and quantity of the input texts. Because dataset 1 is biased and smaller than dataset 2, ranks in the second column of Table 9 tend to be lower than those in the third column.

In addition, our method extracted many verbs overlooked so far (Table 10).

Finally, the power of our method is demonstrated in Figure 5 by using one of the pathway diagrams in KEGG [11]. In the figure, black circles show molecules whose names are recognized by NEs from extensional ontology. A star indicates that our method extracted one or more templates from intervals between two NEs at the start and end points of the flagged arrow.

5. CONCLUSION

For recognition of NEs, we used a subset of a large dictionary known as an extensional ontology. In related works, domain experts manually collected verbs and templates indicative of interaction; we sought to extract them automatically. We first extracted intervals between two NEs and then extracted verbs and templates indicating biomolecular interaction from the intervals. Our experimental results showed that our method could extract the verbs and templates that had been manually prepared in related works. Furthermore, our method extracted a wide variety of previously unidentified interaction-indicative verbs and templates. Hence, our approach to template extraction, which doesn't require any background knowledge, can be used for large-scale extraction of information regarding biomolecular interactions. Extraction pattern-based approaches have been used to extract to various relationships (e.g., company–headquarters relation, management succession) with sufficient performance [12, 13]. However, for success, our method requires a large body of unbiased texts.

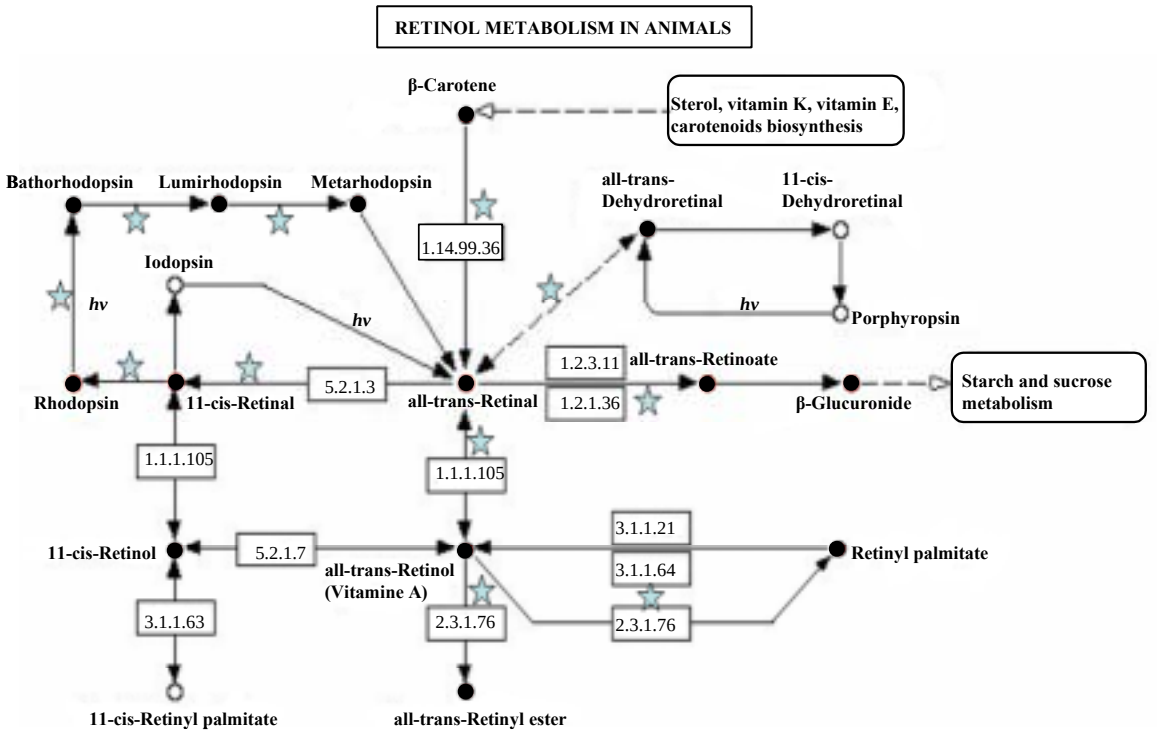

Figure 5: Extracted named entities and templates that occur in a pathway diagram.

Although aspects of our proposed method may need to be improved, we believe it is helpful for extracting large amounts and a wide variety of useful relationships among biomolecules, including protein–protein interactions, protein–gene interactions, and gene–gene regulation. In the next step, we will investigate improving the accuracy and will perform experiments on relationship extraction from biomedical texts.

6. ACKNOWLEDGEMENT

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by BIRD of Japan Science and Technology Agency (JST).

7. REFERENCES

[1] Fukuda, K., Tamura, A., Tsunoda, T., Takagi, T. "Toward Information Extraction: Identifying Protein Names from Biological Papers", Biocomputing, pp.707-718, 1998.

[2] Rindflesch, Thomas, C. "EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature", Biocomputing, pp. 517-528, 2000.

[3] Sekimizu, T., Park, H., Tsujii, J. "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts", Genome Informatics, pp. 62-71, 1998.

[4] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M. "Automatic Extraction of Protein Interactions from Scientific Abstracts", Biocomputing, pp.541-551, 2000.

[5] Baker, P. G. et al. "An Ontology for Bioinformatics Applications", Bioinformatics, Vol. 15, No. 6, pp. 510-520, 1999.

[6] The Gene Ontology Consortium. "Gene Ontology: Tool for the Unification of Biology", Nature Genetics, Vol. 25, pp. 25-29, 2000.

[7] P. D. Krap et al. "The EcoCyc Database", Nucleic Acids Research, 30(1):56 2002.

[8] "Generic Knowledge-Base Editor", http://www.ai.sri.com/~gkb/

[9] Yagyuu, T., Satou, K. "Toward Automatic Construction of Extensional Ontology from Genome Databases", Genome Informatics, pp442-443, 2000

[10] "Online Mendelian Inheritance in Man" http://www.ncbi.nlm.nih.gov/omim/

[11] "Kyoto Encyclopedia of Genes and Genomes" http://www.genome.ad.jp/kegg/

[12] Riloff, E. "Automatically Generating Extraction Patterns from Untagged Text", In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp.1044-1149, 1996.

[13] Yangarber, R., Grishman, R. "Machine Learning of Extraction Patterns from Unannotated Corpora: Position Statement", Proceedings of the 14th European Conference on Artificial Intelligence: ECAI-2000 Workshop on Machine for Information Extraction, Berlin, August 2000.