The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain

Tomoko Ohta
University of Tokyo
Hongo 7-3-1, Bunkyo-ku,
Tokyo
113-0033, Japan
okap@is.s.u-tokyo.ac.jp

Yuka Tateisi
CREST, JST
Hongo 7-3-1, Bunkyo-ku,
Tokyo
113-0033, Japan
yucca@is.s.u-tkyo.ac.jp

Jin-Dong Kim
CREST, JST
Hongo 7-3-1, Bunkyo-ku,
Tokyo
113-0033, Japan
jdkim@is.s.u-tokyo.ac.jp

ABSTRACT

With the information overload in genome-related field, there is an infreest need for natural language processing technology to extract information from literature and various attempts of information extraction using NLP has been being made. We are developing the necessary resources including domain ontology and annotated corpus from research abstracts in MEDLINE database (GENIA corpus). We are building the ontology and the corpus simultaneously, using each other. In this paper we report on our new corpus, its ontological basis, annotation scheme, and statistics of annotated objects. We also describe the tools used for corpus annotation and management.

1. INTRODUCTION

Recently, rapid advances in genome-related field have caused a problem of information overload. Although the databases of proteins, genes, etc. have been constructed, the work of updating them with the latest results published in journals is still largely a manual process, and the use of natural language processing technology to help this process is expected and various attempts of information extraction using NLP has been being made.

We are currently working on the task of extracting event information by using general-purpose natural language processing method[2]. At the same time we are developing the necessary resources or knowledge including domain ontology and annotated corpus from research abstracts in MEDLINE database (GENIA corpus) that can be used as a gold standard for evaluation, training data for machine learning programs, and for other various purposes. We are building the ontology and the corpus simultaneously, using each other; we enhance the ontology with the terms in the annotated corpus, and the tags set for annotation with enhanced ontology.

Text corpus annotated with various levels of linguistic information is required as the process of event information extraction requires syntactic, semantic, and other broad range of levels of natural language processing. As the first step of information extraction, we focused on extracting and classifying the "actors" of events, such as substances. The current version of GENIA corpus is aimed for such application, in which we annotated the named entities in the research abstracts taken from MEDLINE database. We first built a conceptual model (ontology) of substances and sources (substance location), and based on that ontology, we have annotated such named entities involved in biological reactions concerning to transcription factors in human blood cells. The preliminary version of the corpus[15] was used as a learning and test set of a system that extracts the names of substances involved in and their sources from MEDLINE abstracts[3]. In this paper we report on the corpus, its ontological basis, annotation scheme, and statistics of annotated objects. We also describe the tools used for corpus annotation and management.

2. THE TAG SET AND UNDERLYING ONTOLOGY

Named entity annotation task can be regarded as identifying and classifying the names that appears in the texts according to a predefined classification. For a reliable classification, the classification must be well-defined and easy to understand by the domain experts who annotate the texts. To fulfill this requirement, we create a concrete data model (ontology) of the biological domain on which the tag sets are based.

In defining the ontology, we required that the classification must be mutually exclusive. For example, the substances can be classified by their biological role or by their chemical structure, which are both important. However, in our initial annotation work, we rather concentrated on one criteria because using two might prevent the mutually exclusive classification or require more complicate tag structures and context dependent semantic tags.

Ontologies of biological terminology has been created to provide a model of biological concept that can be used for integrating the heterogeneous information sources such as multiple databases[1, 13, 14], as a controlled vocabulary for annotating database entries[6, 7, 16, 10], or as templates in information extraction task[11]. However, existing ontologies were not suitable for our task. For example, there are classes of "Organic Chemicals [D02]", "Enzymes, Coenzymes, and Enzyme Inhibitors [D08]", and "Amino Acids, Peptides, and Proteins [D12]" at the same level right under the "Chemicals and Drugs [D]" category in the MeSH term hierarchy, a contorolled vocabulary developed at National Library of Medicine for indexing MEDLINE abstracts. Because an enzyme is a protein in chemical structure and a protein is an organic compound, we must allow a multiple tag or a structured tag to annotate the name of an enzyme if we use this level of classification in annotation. Otherwise, there would be an ambiguity as to which tag

.

(class) should be used to annotate the name. Similar problems were found in several other ontologies[1, 13].

On the other hand, the ontologies for annotating entries in biological databases[16, 7, 6] have focused on describing specific class of entities in depth, because they must be as fine-grained as to distinguish between the entities differently categorized in existing databases, but do not have to treat the entities that are not in the databases. Thus, these ontologies are very specific in one area but not enough for other: some information necessary for information extraction task can be completely out of the scope of these ontologies. Thus, for our current annotation purpose, we could not choose an ontology of this type. We need an ontology that covers broader range of concept at shallower (more general and coarse-grained) level than these ontologies. The ontologies developed at Columbia University[11] or Institute of Medical Science in University of Tokyo[14] are more suitable for out task, but they were not widely available at the time when we started the project.

Thus, we first defined the base ontology as a taxonomy shown in Figure 1. The top level of the ontology is divided into *substance* and *source* subhierarchy, and *other* that is not either a substance or a source.

At the outset of our annotation work we chose to simplify the classification of substances by concentrating on the chemical structure. This is because the chemical structure is more independent of the biological context that the substance appears, and is therefore more stably defined. For example, the chemical characteristics of a protein can be easily defined, but its biological role may vary depending on the biological context, e.g., a substance may work as an enzyme for one species but not for others. In our model we do not classify substance as enzymes, transcription factors, genes, etc. but as proteins, DNAs, RNAs, etc. They are further classified into families, complexes, individual molecules, subunits, domains, and regions.

Based on the ontology, names of substances that appear in MEDLINE abstracts are classified into the names of atoms (e.g., *calcium*), inorganic compounds (e.g., *H2O2*), proteins (e.g. *NF-kappa B*), peptides (e.g., *pml/RAR-alpha peptides*), amino acids(e.g., *tyrosine*), DNAs (e.g., *PML gene*), RNAs (e.g., *Egr-1 mRNA*), nucleotide polymers other than DNAs and RNAs (e.g., *CpG dinucleotide*), nucleotide monomers (e.g., *thymidine*), lipids (*lipopolysaccharide*), carbohydrates (*3-O-methyl-D-glucose*), and other organic compounds (*2,3,7,8-tetrachlorodibenzo-p-dioxin*). Similarly, the names of sources are classified into the names of multi-cell organisms, mono-cell organisms other than the virus, virus, body parts, tissues, cell types, cell subcomponents, cell lines, and other artificial sources.

Sources are biological locations where substances are found and their reactions take place. They are also hierarchically sub-classified into organisms (e.g., *human*), body parts (e.g., *head*), tissues (e.g. *brain*), cells or cell types (e.g., *leukocyte*), cell sub-cations (e.g., *me mbrane*), or artificially cultured cell lines (e.g., *HeLa*). Organisms are further classified into multi-cell organisms, mono-cell organisms other than viruses, and viruses. There is obviously relations other than taxonomy between the classes in the source hierarchy. For example, (multi-cell) organisms, tissues, cells, and sublocations are interrelated with *part-of* relation, a cell line is made from a certain cell type, and so on. However, these relationships are not in Figure 1.

We mark up the terms that appear in abstracts and classify them with this classification. Also, we mark up other terms that are not the names of substances or sources, but yet considered important in the task of information extraction. They are used for further enhancing the ontology.

We adopt the DARPA Agent Mark-Up Language (DAML)[4] as the ontology-defining language for GENIA project. The DAML has been designed as an XML-based se mantic language and the latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontologies and to markup information. Many ontologies have been written or are being written in DAML. We have therefore decided to manage our ontology in DAML so as for us to utilize the existing ontologies in DAML ontology library and for others to access our products easily.

3. GPML

The classification described in previous section is encoded in GENIA Project Markup Language (GPML)[8], a document type definition (DTD) of XML language that consists of definitions of structural elements for defining logical division of documents, linguistic elements for annotating linguistic (both syntactic and semantic) information, and resource elements for defining language resources that are not the document itself such as ontologies or thesauri. These elements provide the way of annotating corpora with linguistic information and document structure information simultaneously. The structure of GPML elements also supports our strategy to build language resources simultaneously by providing a mecha nism of extending existing ontology and lexicon while annotating corpora. The extended parts of ontology can be defined using the resource elements while annotating the corpus, and will be collected later for complementing the main ontology. In designing GPML, simplicity and extensibility were regarded highly. For the simplicity, we didn't define such entities that are just useful for potential extension but not used for the current GENIA resources. For example, we didn't define structural entities for chapters or sections because the current GENIA corpus consists of MEDLINE abstracts and there are no such structures. For the extensibility, we tried to secure reliable guidelines instead of relying on imperfect predictions about potential extensions. For example, we decided to maintain the structural part of GPML as compatible as possible with DocBook[5], a markup language for technical documentation maintained by OASIS. It is one of th e most popular XML applications supported by a large base of users and developers.

Each abstract (article element) in GENIA corpus comprises four elements: *title*, *articleinfo*, *abstract* and *localresource* (See Figure 2) for an example). A *title* element contains the title of the abstract and an *abstract* element contains the main body of the abstract. An *articleinfo* element contains bibliographic information about the article; currently we use only the *bibliomisc* element (as defined in DocBook) in it, which contains the unique identifier of the abstract in MEDLINE assigned by the National Library of Medicine. A *localresource* element contains ontologic and lexical information for the article.

An abstract element comprises one or more *sentence* elements. A *sentence* element contains a sentence in the abstract, and contain *term* elements and *cons* elements. T he *term* elements are semantically atomic terms which are basic units of linguistic analysis. The *cons* elements are phrases which consist of other *term* or *cons* elements. The *term* and *cons* elements can have their syntactic and semantic information as their attributes.

In the current GENIA corpus, technical terms in the title and the body of the abstracts are tagged as *term* or *cons* elements with conceptual information from the ontology encoded in the *sem* attribute, which is a pointer to an element within the *localresouce* elements as described later. A technical term is considered to be a semantic unit, so that they can be naturally regarded as a *term* element. However, in coordinating clauses, a term is often separated into two or more parts. In such case, we just isolate each part as it is and label

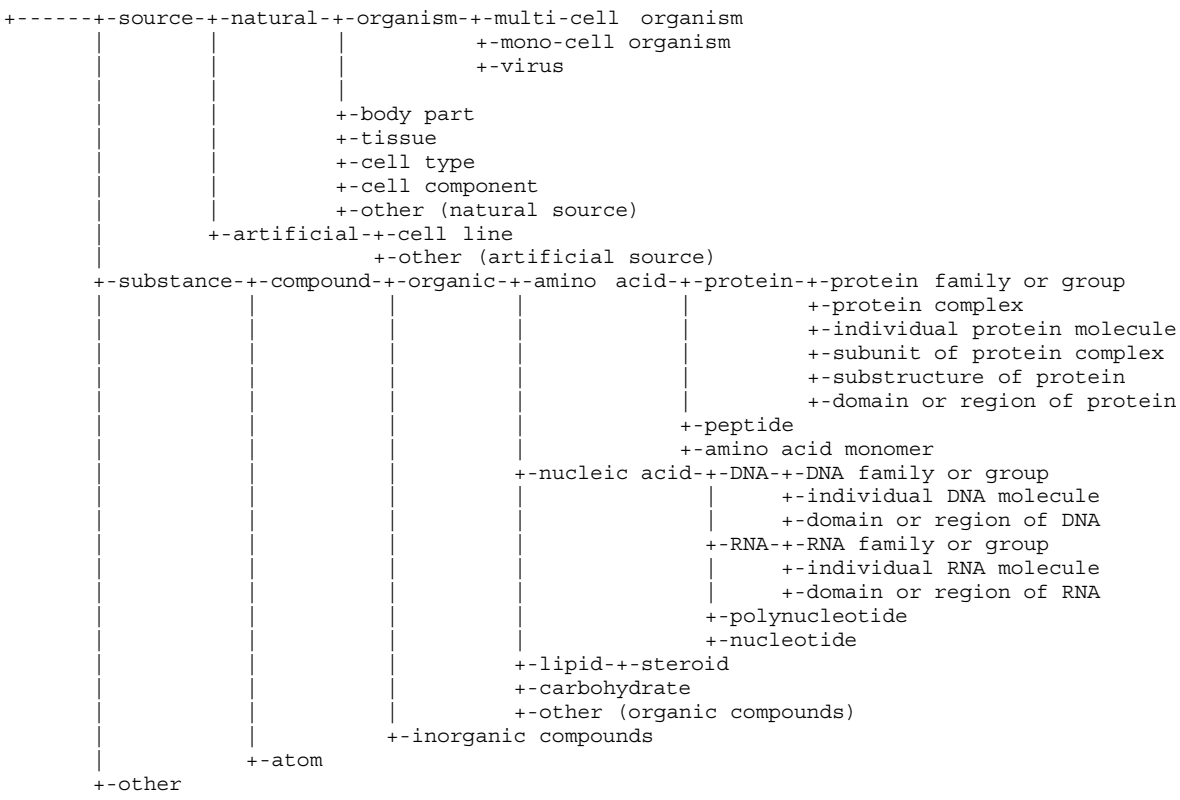

Figure 1: The underlying ontology

each of them as indicating they are parts of certain terms. After doing so, we assemble or iginal terms with their parts. When we annotate assembled terms, we use *cons* elements instead of *term* elements.

For example, in Figure 2, the term *mitogen- and stress-activated protein kinases* in the title actually means *mitogen-activated protein kinases and stress-activated protein kinases*. In the title, the term *mitogen- and stress-activated protein kinases* is annotated as a *cons* element that is an *and*-coordination of two terms (denoted with pointers #3 and #4 in the *sem* attribute). The actual entities are declared in the *localresource* element (the *class* elements with id=3 and id=4).

The *localresource* element is for createing new concepts reqquired but are not in the base ontology, which is the case of almost every time in current version of the corpus. Lexical variation information has been also stored up in the localresource elements. A newly created subclass or an instance of a class in the base ontology is declared in *class* elements which has a *label* element, the name of the newly created entity (class or instance), and *SubClassOf* element, the superclass of the entity. For example, in the case of the coordinated terms mentioned above, we need to create the concepts of *mitogen-activated protein kinase* and *stress-activated protein kinase*. The full forms of the terms are used a s the label, and in the *SubClassOf* element they are declared as instances of the *protein* class of the GENIA ontology. The base ontology to be used is declared at the top of the *localresource* element as an *imports* element.

Using the *localresource* element, we can add the concepts to the base ontology while annotating the instances which appear in the real-world texts. Later, the information in the *localresource* elements can be gathered to enhance the base ontology itself. In this way, we can create the corpus and the ontology simultaneously.

4. THE CURRENT STATUS OF THE GENIA CORPUS

So far, we have annotated 2,500 abstracts related to the transcription factors in human blood cells. We have marked up 146,505 names with 31 different semantic classes. 77,480 proteins, 23,628 DNAs, 2,432 RNAs, 27,542 sources, 36,683 others are marked up. We have made 670 abstracts out of 2,500 (GENIA Corpus ver. 1.1) available to public[17].

We are currently trying to enhance the corpus in both quantity and quality. As well as increasing the number of abstracts, we have started to annotate syntactic and co-reference structures to the same set of abstracts that the terms were annotated. Then, the corpus can be further annotated with event structures.

Part-of-speach annotation to the same set of abstracts annotated for named entities is under way using Penn Treebank set[12]. The preliminary version of annotation to the base text of GENIA Corpus ver 1.1 was done and the annotated texts are used internally for the experiment of named entity extraction. We just started to define the scheme for annotating the deeper level of syntactic feature and co-reference structures.

The underlying ontology is also under enhancement. We have collected the terms in *other* class in the current corpus and have been re-classifying the terms to broden the coverage of the ontology. Deailed quantitive analysis is yet to be done, but we have found certain classes, like diseases, interaction of substances and cells, and experimental methods, should be added to the base ontology.

```
<article>
<title>
Effects of <term sem="#000">Ara-C</term> on <term sem="#002">neutral sphingomyelinase</term>
and <cons sem="(AND #3 #4)"><term sem="#005">mitogen-</term> and <term sem="#006">stress-</term>
term sem="#007">activated protein kinases</term></cons> in <term sem="#008">T-lymphocyte cell lines</term>.


</title>
<articleinfo>
<bibliomisc>MEDLIN:97107281</bibliomisc>
</articleinfo>
<abstract>
<sentence>
<term sem="#002">Neutral sphingomyelinase</term> (<term sem="#009">SMase</term>) can be
activated by <term sem="#010">extracellular signals</term> to produce <term sem="#011">ceramide</term>,
which may affect <term sem="#012">mitogen-activated protein kinase (MAPK) activities</term>.
</sentence>
        :
</abstract>
<localresource>
<imports resource="http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/GENIA.daml"  prefix="GENIA"/>
<class id="000"><label>Ara-C</label><subClassOf  resource="GENIA#other_organic_compounds"/></class>
<class id="001"><label>molecule</label><subClassOf  resource="GENIA#protein"/></class>
<class id="002"><label>neutral_sphingomyelinase</label><subClassOf  resource="#001"/></class>
<class id="003"><label>mitogen-_activated_protein_kinase</label>
   <subClassOf resource="GENIA#protein"/></class>
<class id="004"><label>stress-_activated
_protein_kinase</label>
   <subClassOf resource="GENIA#protein"/></class>
:
</localresource>
</article>
</set>
```

Figure 2: An annotated abstract in GPML format

5. TOOLS FOR CORPUS ANNOTATION

Although a XML-tagged text can be created by using text editors, semantically annotated corpora must be created by domain experts who are not always familiar with XML tag scheme. A tool for management of the tagged texts is also indispensable for controlling the quality of the corpus and taking the statistics of tag data.

To help annotators, we use a GUI-based tag definition tool TagEdit (Figure 3) and tagging tool JTag (Figure 4), developed by Hitachi Co Ltd.

In the tag definition tool TagEdit, definition of new tag-set, refinement of definition, enhancement of the tag-set by adding or removing tags, and enrichment of tags by adding or removing attributes are available. In TagEdit, a tag set is assumed have hierarchical structure. That is, a tag is defined as a "child tag" of another. The defined tag set is directly corresponding to a taxonomy in an ontology, and this feature is suitable for semantic annotation of named entities.

The tagging tool JTag allows the annotators to choose tags from a tagset defined with TagEdit and annotate texts with them. JTag supports the exporting the annotated data in two forms: An annotated text can be saved as tag-embedded XML document, or the data including the class of tag, the position of tag in the text, and values of attributes can be saved separately from the text to be stored in external database.

The separation of tag information from the text itself allows users to perform various operations on tags, e.g. to compare the annotation by different annotators and to take statistics. For this kind of operations, we developed Tag Information Management System (TIMS) Workbench[9] to perform/view tagging on a particular document. Since tag informations stored separately from original doc-

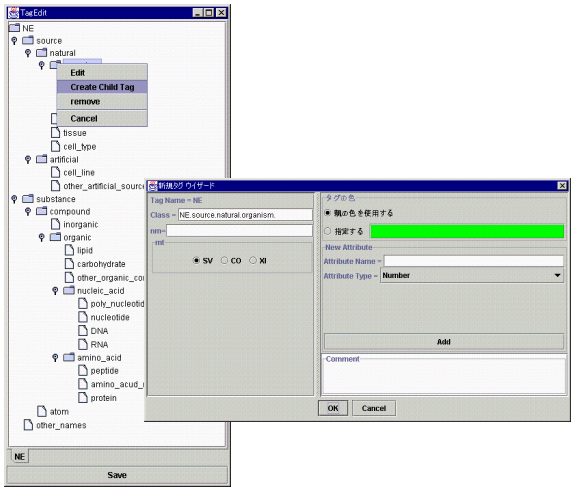

Figure 3: Definition of Tags with TagEdit: select a tag in hierarchy and "create child tag" to add a tag for its child.

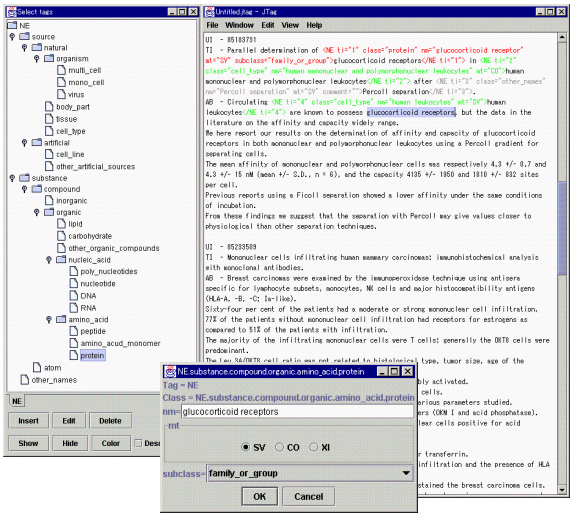

Figure 4: Annotation with JTag: select a range of text in right window and insert one of the tags shown in the left window

uments and managed using an external database software, various different types of tags for the same document can be added. The system also keeps track of the Audit Trail or History, i.e., the date and time, the user or system that performed the tagging etc. TIMS also has facility to search logically for tags, which we call Interval Operations. They are XML specific text/data mining operations which can be performed over documents. The operations can be done over open documents or in batch mode, selecting the files from a list. The TIMS workbench is also available to public[17].

6. CONCLUSION AND FUTURE WORK

We have made a corpus from research abstracts in molecular biology domain annotated with named entities classified based on a ontology of substances and sources, and the tools for tag definition, annotation and corpus management. We have annotated 2500 abstracts from MEDLINE database, and made 670 of them available to public. Since its release in July 2001, the corpus was downloaded 67 times, and TIMS workbench was downloaded 42 times[1].

We have started to annotate syntactic and co-reference information to the same set of abstracts. Using GPML DTD, the syntactic and co-reference information will be integrated to the current corpus as an attribute to *term* and *cons* elements. We plan to anntate the corpus with event structures. The future work also include the enhancement of the underlying ontology and enrichment of the tag set.

7. ACKNOWLEDGMENTS

The research is partially supported with Genome Information Science Project (MEXT, Japan) and Information Mobility Project (CREST, JST, Japan).

8. ADDITIONAL AUTHORS

Additional authors: Hideki Mima (CREST, JST, email: `mima@is.s.u-tokyo.ac.jp`) and Jun-ichi Tsujii (University of Tokyo and CREST, JST, email: `tsujii@is.s.u-tokyo.ac.jp`).

[1] as of the end of February 2002

9. REFERENCES

[1] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass. An ontology for bioinformatics applications. 15:510–520, 1999.

[2] N. Collier, H. Mima, S.-Z. Lee, T. Ohta, Y. Tateisi, A. Yakushiji, and J. ichi Tsujii. The genia project: Knowledge acquisition from biology texts. pages 448–449, 2000.

[3] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedins of COLING '2000*, pages 201–207, 2000.

[4] DARPA. The DARPA agent markup language. http://www.daml.org/.

[5] DocBook Technical Committee. http://www.oasis-open.org/docbook/.

[6] V. Giudicelli and M.-P. Lefranc. Ontology for immunogenetics: the imgt-ontology, bioinformatics. 15(12):1047–1054, 1999. http://imgt.cines.fr.

[7] P. D. Karp. An ontology for biological function based on molecular interactions. 16(3):269–285, 2000. http://www.ai.sri.com/pkarp/interactions.html.

[8] J.-D. Kim, T. Ohta, Y. Tateisi, H. Mima, and J. Tsujii. Xml-based linguistic annotation of corpus. In *Proceedins of the First NLP and XML Workshop*, pages 47–53, 2001.

[9] H. Mima, S. Ananiadou, G. Nenadic, and J. Tsujii. Xml tag information management system: A workbench for ontology-based knowledge acquisition and integration. In *Proceedins of Human Language Technology Conference (HLT 2002)*, 2002.

[10] National Library of Medicine. Medical subject headings. http://www.nlm.nih.gov/mesh/meshhome.html.

[11] A. Rzhetsky, T. Koike, S. Kalachikov, S. M. Gomez, M. Krauthammer, S. H. Kaplan, P. Kra, J. J. Russo, and C. Friedman. A knowledge model for analysis and simulation of regulatory networks. 16(12):1120–1128, 2000. http://genome6.cpmc.columbia.edu/ tkoike/ontology.

[12] B. Santrini. Part-of-speech tagging guidelines for the Penn Treebank project, 1991. included in Penn Treebank II CD-ROM.

[13] S. Schulze-Kremer. Ontologies for molecular biology. In *Proceedins of 3rd Pacific Symposium on Biocomputing*, pages 695–706, 1998.

[14] T. Takai-Igarashi and T. Takagi. Signal-ontology: Ontology for cell signalling. 11:440–441.

[15] Y. Tateisi, T. Ohta, N. Collier, C. Nobata, and J. ichi Tsujii. Building an annotated corpus from biology research papers. In *Proceedins of Workshop on Semantic Annotation and Intelligent Content*, pages 28–34, 2000.

[16] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. 25:25–29, 2000. http://www.geneontology.org/.

[17] The GENIA project. http://www-tsujii.is.s.u-tokyo.ac.jp/genia/.