Building an Annotated Corpus in the Molecular-Biology Domain

Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, Jun-ichi Tsujii
Department of Information Science
Graduate School of Science
University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Abstract

Corpus annotation is now a key topic for all areas of natural language processing (NLP) and information extraction (IE) which employ supervised learning. With the explosion of results in molecular-biology there is an increased need for IE to extract knowledge to support database building and to search intelligently for information in online journal collections. To support this we are building a corpus of annotated abstracts taken from National Library of Medicine's MEDLINE database. In this paper we report on this new corpus, its ontological basis, and our experience in designing the annotation scheme. Experimental results are shown for inter-annotator agreement and comments are made on methodological considerations.

1 Introduction

In the field of molecular biology there have recently been rapid advances that have motivated researchers to construct very large databases in order to share knowledge about biological substances and their reactions. A large part of this knowledge is only available in unformalized research papers and information extraction (IE) from such sources is becoming crucial to help support timely database updating and to help researchers avoid problems associated with information overload.

For this purpose, various NLP techniques have been applied to extract substance names and other terms (Ohta et al., 1997; Fukuda et al., 1998; Proux et al., 1998; Nobata et al., 1999) as well as information concerning the nature and interaction of proteins and genes (Sekimizu et al., 1998; Blaschke et al., 1999; Hamphrays et al., 2000; Thomas et al., 2000; Rindflesch et al., 2000). The nomenclatures of genes and associated proteins for model organisms such as *S. Cerevisiae* (yeast) and *D. Melanogaster* (fruit fly) are established so that good dictionaries for those names have been constructed. However nomenclatures for humans are not yet available as the whole picture of the human genome has yet to be revealed, this results in arbitrary names being used by researchers who identified the structure of proteins and genes, so dictionary-based approaches might not be as effective as in the case of model organisms. Thus many of the previous researchers either limit their scope to extracting information on substances like enzymes which have established naming conventions (Hamphrays et al., 2000) or extracting information on 'substance' giving up the distinction between the class of substance like protein and DNA (Fukuda et al., 1998; Proux et al., 1998; Sekimizu et al., 1998; Thomas et al., 2000).

Term identification and classification methods based on statistical learning seem to be more generalizable to new knowledge types and representations than the methods based on dictionaries and hand-constructed heuristic rules. We think that a corpus-based, machine-learning approach is quite promising, and to support this we are building a corpus of annotated abstracts taken from National Library of Medicine (NLM)'s MEDLINE database.

Corpus annotation is now a key topic for all areas of natural language processing and linguistically annotated corpus such as treebanks are now established. In information extraction task, annotated corpora have been made mainly for the judgment set of information extraction competitions such as MUC (Chinchor, 1998). We think that technical terms of a scientific domain share common characteristics with the "Named Entities" and the tasks we attempt in-

volve recognition and classification of the names of substances and their locations, just as named entity recognition task in MUC conferences. We therefore try to model our annotation task after the definition of "EnameX" (Chincor, 1998a) of MUC conferences. Unlike in MUC conferences, we don't make a precise definition of how the recognized names are used in further information extraction task such as event identification, because we want the recognition technology to be independent of the further task. Our work is also compared to word-sense annotation (e.g.,(Bruce and Wiebe, 1998)) where instances of words that have multiple senses are labelled for the sense it denotes according to a certain dictionary or thesaurus.

We first built a conceptual model (ontology) of substances and sources (substance location), and designed a tag set based on the ontology which conforms to SGML/XML format. Using the tag set, we annotated the entities such names that appears in the abstracts of research papers taken from the MEDLINE database. In this paper we report on this new corpus, its ontological basis, and our experience in designing the annotation scheme. Experimental results are shown for inter-annotator agreement and comments are made on methodological considerations.

2 Design of The Tag Set

2.1 Underlying Ontology

The task of annotation can be regarded as identifying and classifying the names that appears in the texts according to a pre-defined 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.

Ontologies have been developed in the biomedical sciences for several applications. Such ontologies include conceptual hierarchies for databases covering diseases and drug names. Construction of a more general ontology e.g. (Baker et al., 1999) is being attempted by several groups interested in interconnecting databases under a uniform view.

We start from a taxonomy illustrated in Fig-

ure 1[1]. In this taxonomy, we classify substances according to their *chemical* characteristics rather than their biological role. This is unlike other existing ontologies in the biology field (Baker et al., 1999; Schulze-Kremer, 1998), which mix the classification by biological role and by chemical structure. The reason that we have adopted this approach is that we consider mixing two criteria prevents the mutually exclusive classification and thus makes the annotated task more complicated by introducing nested tag structures and context dependent semantic tags. In our initial annotation work we therefore chose to simplify the classification by concentrating on the chemical structure.

Chemical classification of substances is quite independent of the biological context in which it 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., it may work as an enzyme for one species but a poison for others. Therefore, 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, because these super- and sub- structures often have separate names. This classification is non-controversial among biologists and can be easily expanded into other ontologies.

Sources are biological locations where substances are found and their reactions take place, such as *human* (an organism), *liver* (a tissue), *leukocyte* (a cell), *membrane* (a sub-location of a cell) or *HeLa* (a cultured cell line). Organisms are further classified into multi-cell organisms, mono-cell organisms other than viruses, and viruses. Organism, tissue, cell, sub-locations are interrelated with *part-of* relation but that relation is not shown in Figure 1. Based on this domain model, we annotate the names of proteins, DNAs, RNAs, and sources using the tags shown in Table 1.

An example of an annotated text is shown in Figure 2: the UI number is a unique identifier of the abstract in MEDLINE assigned by

---

[1]In Figure 1 the concepts represented in **bold** are reflected in the tag set and the concepts represented in *italic* are reflected in the attributes.
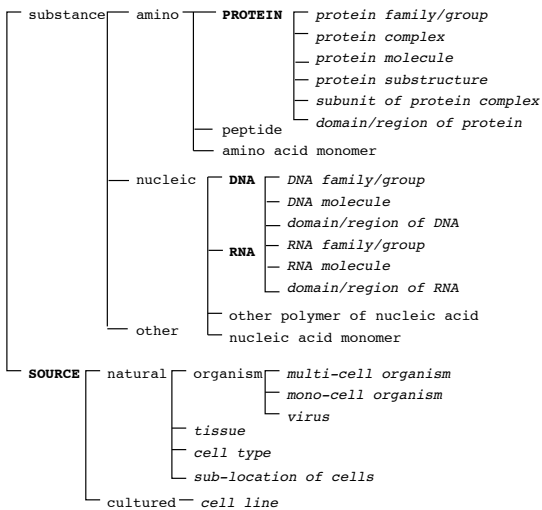
Figure 1: The taxonomy used as a domain model of our tagging scheme

Table 1: Tags and their target objects

tag	object
`<PROTEIN>`	the names of proteins, including protein groups, families, molecules, complexes, and substructures
`<DNA>`	the names of DNAs, including DNA molecules, DNA groups, DNA regions, and genes
`<RNA>`	the names of RNAs, including DNA molecules, RNA groups, RNA regions, and genes
`<SOURCE>`	the sources of substances, i.e., the names of organisms, tissues, cells, sub-locations of cells, and cell lines

```
UI - 91012785
TI - <PROTEIN unsure=ok>Lymphotoxin</PROTEIN>
activation by <SOURCE subtype=cl unsure=ok>human
T-cell leukemia virus type I-infected cell
lines</SOURCE>:  role for <PROTEIN unsure=ok>NF-kappa
B</PROTEIN>.  AB - <SOURCE subtype=cl
unsure=ok>Human T-cell leukemia virus type
I (HTLV-I)-infected T-cell lines</SOURCE>
constitutively produce high levels of biologically
active <PROTEIN unsure=ok>lymphotoxin</PROTEIN>
(<PROTEIN unsure=ok>LT</PROTEIN>; <PROTEIN
unsure=ok>tumor necrosis factor-beta</PROTEIN>)
protein and <RNA unsure=ok>LT mRNA</RNA>.
To understand the regulation of <PROTEIN
unsure=ok>LT</PROTEIN> transcription by <SOURCE
subtype=vi unsure=ok>HTLV-I</SOURCE>, we analyzed
the ability of a series of deletions of the
<DNA unsure=ok>LT promoter</DNA> to drive the
<DNA unsure=ok>chloramphenicol acetyltransferase
(CAT) reporter gene</DNA> in <SOURCE subtype=cl
unsure=ok>HTLV-I-positive MT-2 cells</SOURCE>.  The
smallest <DNA unsure=ok>LT promoter fragment</DNA>
(-140 to +77) that was able to drive CAT activity
contained a site that was similar to the <DNA
unsure=ok>immunoglobulin kappa-chain NF-kappa
B-binding site</DNA>.
```

Figure 2: Example of Annotated Text

stracts. The abstracts were 116 words long on average. One of the authors, who has a doctorate in molecular biology, manually tagged the abstracts. The process took about 40 hours. 2125 proteins, 358 DNAs, 30 RNAs, and 801 SOURCEs are tagged.

Ten abstracts out of the 100 were randomly chosen and three other volunteers, two medical science researchers and one biology researcher, were asked to annotate them with our tagging scheme. We gave a brief explanation on the tagging task and scheme to each annotator. The annotators were asked to annotate the text independently in one weeks' time.

After the annotation was done, we sent a questionnaire to annotators to ask for their comments on the tagging task and the guide. From the feedback of the questionnaire, we learned that the annotators felt the task to be relatively easy, but there are several cases where the they were unsure about which tags to be assigned where. The cases include:

- where two or more names are conjoined with *and* or *or*, e.g., `IRF-1 mRNA and protein`

- the ambiguity in some papers concerning

the National Library of Medicine, `TI` is the title, and `AB` is the abstract text. The `unsure` attribute shown in the text is optional. This is used when annotators are unsure about whether a name should be tagged or whether the boundary of the tagged name is correct, and when the annotator was sure about the instance of the markup, `unsure` attribute can be omitted (or can be assigned the value `ok`).

3 Tagging Task

Before beginning the tagging process we made a preliminary experiment by tagging 100 ab-

Table 2: The percentage of inter-annotator agreement on 10 abstracts

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Mean
A0-A1	100.00	69.05	38.18	82.76	69.81	83.87	74.07	83.33	88.31	91.67	77.29
A0-A2	100.00	60.98	66.13	67.65	80.49	72.31	72.73	90.11	84.21	71.43	76.78
A0-A3	95.24	59.09	57.63	96.55	86.05	83.82	69.64	79.55	85.71	84.91	78.44
A1-A2	100.00	83.78	41.18	60.61	62.00	67.21	83.02	77.65	80.82	78.72	72.55
A1-A3	95.24	52.50	66.67	78.57	61.54	76.56	77.78	70.73	79.73	76.47	72.76
A2-A3	95.24	51.28	47.27	63.64	85.00	76.12	85.45	82.02	83.56	65.38	73.85
Mean	97.62	62.78	52.84	74.96	74.15	76.65	77.16	80.57	83.72	78.10	75.85

whether names denote DNAs, RNAs or proteins,

The annotators also said that the concrete example of tagged texts are more useful than descriptions and more examples should be included in the manual.

Two-way agreement rate is scored according to the scheme used in MUC conferences(Chincor, 1998b). This scoring scheme uses the F-measure derived from recall and precision. Recall R and precision P are given by:

$$R = |X \cap Y|/|X| \qquad (1)$$

and

$$P = |X \cap Y|/|Y| \qquad (2)$$

where X is the set of 'correct' objects and Y is the set of 'retrieved' objects. The F-measure is the harmonic mean of R and P given by

$$F = 1/(1/P + 1/R) = 2 \times |X \cap Y|/(|X| + |Y|) \qquad (3)$$

and this F can be used to measure the agreement of two sets of objects neither of which are considered 'correct' (note that F is symmetric with regards to X and Y).

The F-measures multiplied by 100 to show the percentage of the agreement between annotators for the 10 abstracts are shown in Table 2. In Table 2, T1, ..., T10 denotes the abstracts and A0, ..., A3 denotes annotators. The table shows that the agreement rate, comparable to man-machine agreement of systems participated in MUC, is not good for inter-annotator agreement rate. The disagreement indicate that there are several problems in the definition of the target and the description in the manual,

some of which seem to be specific to this domain[2].

We investigate into the case of disagreement by aligning the tagged text and examining the disagreed parts by hand. We found that the disagreement could be classified into several patterns enlisted below. The numbers in the parentheses in the items are the number of the occurrence of the disagreement in total 10 texts. See Table 3 for examples[3].

Division (27): The cases where a same part of a text is tagged as one by some annotators but divided into two (or more) parts by others. They were further classified into the following cases.

D-1 (13) parenthesized abbreviations, full forms, and synonyms

D-2 (3) appositive phrases

D-3 (6) names of a substance which includes SOURCE names

D-4 (2) names of a complex

D-5 (3) conjoined names

Part (60): The cases where a part of phrases is included between <TAG> and </TAG> by some annotators but not by others. They were further classified into the following cases.

P-1 (30) the cases where the substances designated by the tagged part are changed by whether the words following a name are tagged together or not: in 10 cases, different tags are used by the annotators; in

---

[2] Though it may not be directly compared, inter-annotator agreement for the judgment set of IREX conference on Japanese information extraction(Sekine, 1999) is reported to be around 97% in F-measure.

[3] Attributes are omitted in the examples.

Table 3: Examples of disagreement

Cases	Examples
D-1	`<SOURCE>Mycobacterium avium complex (MAC)</SOURCE>` `<SOURCE>Mycobacterium avium complex</SOURCE> (<SOURCE>MAC</SOURCE>)`
D-2	`<SOURCE>U937, a human monocytoid cell line</SOURCE>` `<SOURCE>U937</SOURCE>, <SOURCE>a human monocytoid cell line</SOURCE>`

D-3	`<PROTEIN>Human erythroid 5-aminolevulinate synthase</PROTEIN>`
	`<SOURCE>Human erythroid</SOURCE> <PROTEIN>5-aminolevulinate`
	`synthase</PROTEIN>`

D-4	`<PROTEIN>p50-p65</PROTEIN>` `<PROTEIN>p50</PROTEIN>-<PROTEIN>p65</PROTEIN>`
D-5	`<RNA>ferritin or transferritin receptor mRNAs</RNA>` `<PROTEIN>ferritin</PROTEIN> or <RNA>transferritin receptor mRNAs</RNA>`
P-1 (different tags)	`<DNA>AP-2 consensus binding sequences</DNA>` `<PROTEIN>AP-2</PROTEIN> consensus binding sequences`
P-1 (same tags)	`<PROTEIN>IRF-2 repressor</PROTEIN>` `<PROTEIN>IRF-2</PROTEIN> repressor`
P-2	`<PROTEIN>Stat91 protein </PROTEIN>` `<PROTEIN>Stat91</PROTEIN> protein`
P-3	`<RNA>housekeeping ALAS mRNA</RNA>` `housekeeping <RNA>ALAS mRNA</RNA>`
P-4	`<PROTEIN>transcription factor AP-2</PROTEIN>` `transcription factor <PROTEIN>AP-2</PROTEIN>`
P-5	`<DNA>the terminal protein 1 gene promoter</DNA>` `the <DNA>terminal protein 1 gene promoter</DNA>`
Class	`<RNA>TAR</RNA>` `<DNA>TAR</DNA>`
Missing	`<DNA>21 bp repeats</DNA>` `21 bp repeats`

the other 20 cases, the same tags are used by the annotators.

P-2 (18) the cases where the substances designated by the tagged part are not affected by whether the words following a name are tagged together or not

P-3 (6) the preceding attributive phrase that narrows the meaning of the phrase

P-4 (5) the preceding appositive phrase

P-5 (1) determiners

Class (19): The same part of text is tagged with different tags

Missing (25): A part of text is tagged by some annotators but not by others

The result shows that most of disagreement involves recognizing the names, i.e., identifying the range of words in sentence that are part of the names. On the other hand, there are relatively few cases where classification of the names alone is the problem.

The disagreement involving abbreviation and synonym (case D-1) will be simply solved by explicitly giving an instruction as to whether a full form and its abbreviation (or a name and its synonym) should be separated or not. The case of appositives (cases D-2 and P-5) and determiners are also easy to solve by giving explicit instruction, though the distinction between appositives or determiners and other attributive phrases (case P-4) must be carefully stated in the instruction. The cases involving words that follow a name that do not affect the substance the name designates (P-2) should be handled similarly with a careful description of such cases in the instruction.

The cases that involve the source names (case D-3) and the following words that modify the

meaning of the phrase (P-1) are more difficult, because the names with or without the modifying phrases are recognized by the annotators. One solution would be to allow nesting tags, but this might complicate the tagging scheme and be the cause of another type of error. Simple heuristics of 'taking the longest phrase' might work here, but in the case of preceding modifiers (P-3) the heuristic is not desirable, because most of the preceding modifiers are just description of a characteristics of a substance.

The names tagged by some annotators but not by others (case M) were mostly the terms that describes the parts of a gene as in the example above, or the terms that denotes a family or a class of substances. Such parts or families are considered to be the 'substance' by some annotators but not by others. Incorporating the distinction between families, individual substances, and parts of the substance would help to make the classification of names clearer and result in more consistent annotation.

One of the difficulties of this task compared to MUC named entity extraction is that our targets are inherently unique names of classes, whereas the targets of MUC named entity extraction are names of unique entities. When we refer to a specific protein or DNA, we don't refer to a specific molecule, but rather a class of molecules that have the same characteristics. As the name of a class, when a researcher finds a new substance, the substance is often named after the combination of its function, location, etc. For example, "B-cell specific transcription factor" is a name of a protein (there is an entry in the SwissProt database). This results in the difficulty of distinguishing the names of substances from general description of the substance. In cases such as "Human erythroid 5-aminolevulinate synthase", some researchers recognize it as a name but some only recognize "5-aminolevulinate synthase" as a name and "Human erythroid" as just a description and separate the part as different entity. Also the prenominal modifiers are recognized or not recognized as a part of the name depending on whether the names with or without the modifying phrases are recognized by the annotators.

The classification error, though relatively few, also might be from the nature of this domain. Most of the inconsistency are suspected to be from conventional use of the protein names to denote the genes that transcribe the protein. For example, `NF-kappa B gene` is a name of a gene that transcribe the protein NF-kappa B, and the authors often omit the word `gene` where they think it is clear from the context that the particular occurrence of `NF-kappa B` denotes the DNA. This require the annotators good background knowledge and careful reading, and sometimes the cause of annotation errors. Even the participated annotators, who are qualified specialist of the domain, are sometimes unsure about the target, according to the questionnaire. This might be resolved if the full paper could be referenced in the process of annotation.

4   Conclusion and Future Work

We are in the process of developing a high-quality tagging scheme for semantic annotation of substances and their sources which play an important role in molecular-biology events. We have shown the results of initial inter-annotator agreement tests using the current scheme. After the initial experiment, we revised the tagging manual to give more precise definitions and more examples, and also added attributes to denote the distinction of whether the protein (DNA, RNA) is a molecule, complex, substructure, region, etc. We tagged 500 abstracts according to the revised manual and tagging-scheme, which are in the process of cross-checking and cleaning up the errors. When they are done we plan to make the corpus available to the public along with the tagging manual.

Establishing the training process of annotators, including communication between annotators to get agreement on tagging strategies, which is reported to improve the agreement rate (Dan Melamed, 1998; Wiebe et al., 1999) should also be necessary to help them make consist annotation.

One of the concerns that we have is that our target task is more difficult than the traditional named entity recognition task, because of the naming convention (or the lack of it) of the molecular-biology domain and because the task requires very precise knowledge of the specialist. To solve this problem, tagging tools that incorporates the reference function to the external sources such as substance databases, on-line

glossaries, and full-text of the paper should also be of great help.

The preliminary corpus, though it may be 'noisy', can be useful as a training set for recognition program of biological names and terms. The preliminary corpus can also be used to gain the knowledge of how the tagged names are related to each other and other names, in order to give feedback to the annotators and enhance the domain model and enables us to annotate more rich information such as biological roles.

References

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

C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proc. 7th International conference on Intelligent Systems for Molecular Biology*, pages 60–67.

R. Bruce and J. Wiebe. 1998. Word sense distinguishability and inter-coder agreement. In *Proc. 3rd Conference on Empirical Methods in Natural Language Processing*, pages 53–60.

N. Chinchor. 1998. Overview of MUC-7. In *Proceedings of 7th Message Understanding Conference*. available at http://www.muc.saic.com/proceedings.

N. Chinchor. 1998a. MUC-7 named entity task definition version 3.5. In *Proceedings of 7th Message Understanding Conference*. available at http://www.muc.saic.com/proceedings.

N. Chincor. 1998b. MUC-7 test scores introduction. In *Proceedings of 7th Message Understanding Conference*. available at http://www.muc.saic.com/proceedings.

I Dan Melamed. 1998. Manual annotation of translation equivalence:the blinker project. Technical Report IRCS-98-07, IRCS, University of Pennsylvania. available at ftp://ftp.cis.upenn.edu/pub/ircs/tr/98-07/.

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Towards information extraction: Identifying protein names from biological papers. In *Proc. 3rd Pacific Symoisium of Biocomputing*, pages 707–718.

K. Hamphrays, G. Demetriou, and R. Gaizauskas. 2000. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Proc. 5th Pacific Symoisium of Biocomputing*, pages 72–80.

C. Nobata, N. Collier, and J. Tsujii. 1999. Automatic term identification and classification in biology texts. In *Proc 5th Natural Language Processing Pacific Rim Symoisium*, pages 369–374.

Y. Ohta, Y. Yamamoto, T. Okazaki, and T. Takagi. 1997. Automatic construction of knowledge base from biological papers. In *Proc. 5th International Conference on Intelligent Systems for Molocular Biology*, pages 218–225.

D. Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq. 1998. Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. In *Genome Informatics*, pages 72–80. Universal Academy Press.

T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. 2000. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proc. 5th Pacific Symposium on Biocomputing*, pages 514–525.

S. Schulze-Kremer. 1998. Ontologies for molecular biology. In *Proc. 3rd Pacific Symposium on Biocomputing*, pages 695–706.

T. Sekimizu, H. S. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts. In *Genome Informatics*, pages 62–71. Universal Academy Press.

S. Sekine. 1999. Analysis of the answer of named entity extraction. In *Proceedings of the IREX workshop*, pages 129–132. in Japanese.

J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. 2000. Automatic extraction of protein interactions from scientific abstracts. In *Proc. 5th Pacific Symposium on Biocomputing*, pages 538–549.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Meeting of ACL*, pages 246–253.