Automatic Construction of Biological Abbreviation Dictionary from Abstracts of Biomedical Papers

Mikio Yoshida **Kenichiro Fukuda** **Toshihisa Takagi**
mikio@ims.u-tokyo.ac.jp ichiro@ims.u-tokyo.ac.jp takagi@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

Integration of knowledge described in a huge number of biomedical papers is a very important subject in the genome informatics area [1, 2]. Therefore, we need an intelligent information extracting system to save time. In order to extract information or knowledge from texts and utilize them, it is necessary to use various kinds of dictionaries, which are helpful in eliminating obscurities contained in natural languages (e.g. ambiguity and polysemy). We sought to develop a protein name abbreviation dictionary (PNAD) in an automatic way.

2 System

2.1 Approach

The target of our proposed system is extraction of abbreviation defining expressions. Such expressions are often observed in a typical pattern, in which an original term is followed by a parenthesis including its abbreviation (e.g. "`Thyrotrophin-releasing hormone (TRH)`"). We will call this type of expression *"Parenthetical-Paraphrase"*, the phrase enclosed in brackets *"Inner-Phrase"*, and the phrase followed by brackets *"Outer-Phrase"*. When a Parenthetical-Paraphrase is used to define an abbreviation, we will call it an "Abbreviation-Parenthetical-Paraphrase".

Abbreviation-Parenthetical-Paraphrase	:	Original-Term	(Abbreviation)
Parenthetical-Paraphrase	:	*Outer-Phrase*	(*Inner-Phrase*)

2.2 System Architecture

Fig. 1 shows the overall architecture of the PNAD construction system. In this system, we assume that the *Inner-Phrase* represents an abbreviation and the *Outer-Phrase* involves its Original-Term.

- Step(1) Extract protein names from machine readable biomedical papers.
- Step(2) Extract the *Parenthetical-Paraphrase* from the protein names extracted in Step (1).
- Step(3) Excise *Outer-Phrase* and *Inner-Phrase* from the *Parenthetical-Paraphrase* extracted in Step (2).
- Step(4) Extract a candidate of the Original-Term from the *Outer-Phrase*, and validate if the *Inner-Phrase* is an abbreviation of the candidate.
- Step(5) Register the validated abbreviation and the candidate of its Original-Term into the PNAD.

To extract protein names from machine readable papers, we used the PROPER system [1] that can extract parenthetical expressions with 94.70% precision and 98.84% recall.

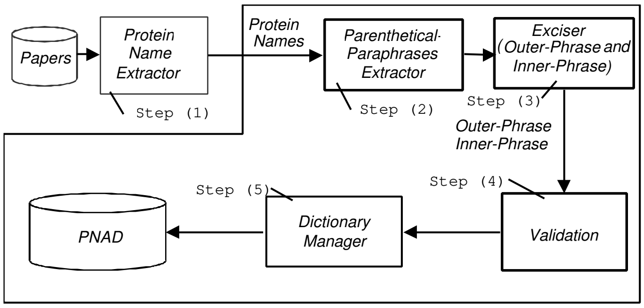

Figure 1: System Architecture.

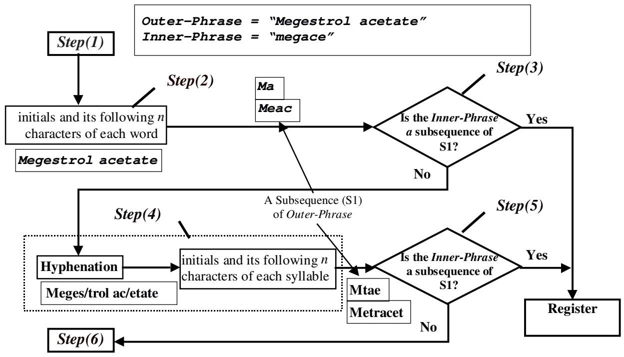

Figure 2: A block diagram of the validation method.

2.3 Validation Method

In Fig. 1, the validation block outputs an abbreviation and its original term. Arguments given to the validation block are two groups of words which are excised from the *Parenthetical-Paraphrase*; *Inner-Phrase* and *Outer-Phrase*. Therefore, it is possible to confirm if the *Inner-Phrase* is an abbreviation of a candidate of the Original-Term extracted from the *Outer-Phrase* by checking whether the *Inner-Phrase* is a subsequence of a candidate of the Original-Term or not.

Fig. 2 shows our proposed validation method. The validation process and subsequence extraction process are done simultaneously. To hyphenate words, we adopted a hyphenation method which had been used in TEX82 [3].

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Fukuda, K., Tamura, A., Tsunoda, T. and and Takagi, T., Toward Information Extraction Identifying protein names from biological papers, *Proc. of Pacific Symposium on Biocomputing '98*, 707–718, 1998.

[2] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I. and Takagi, T., Automatic construction of knowledge base from biological papers, *Proc. of ISMB–97*, 5:218–225, 1997.

[3] Knuth, D.E., *The TEXbook*, Addison-Wesley, Massachusetts, 1984.