Development Information Extraction Tool: Toward Construction of Protein-Protein Interaction Database

Tomoko Okazaki-Ohta[1 2] **Toshihisa Takagi**[1]

okap@ims.u-tokyo.ac.jp takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

[2] JSPS Research Associate of JSPS Research Project for the Future

1 Introduction

In connection with the development of the molecular biology in recent years, the amount of information in genome area has been growing rapidly. Although many data about biomolecular structures have been comprehensively compiled into public databases, most of data as biological functions such as molecular interactions are still only in the biological papers. Reading every articles in the world requires too much time and labor. An interesting information extracting system is needed [1, 2]. Since there are many descriptions of various research backgrounds and hypotheses in papers, in order to extract a certain knowledge and to build the database, the logical structure of a paper needs to be understood and the sentence suitable for information extraction needs to be specified. Since there is structure decided as the paper to some extent, while understanding the structure of the whole paper, understanding the structure of the sentence expressing knowledge is connected with information extraction. We propose here a technique of specifying the sentence which has the description of a protein-protein interaction using the figure legend having shown the experiment result (Fig. 1).

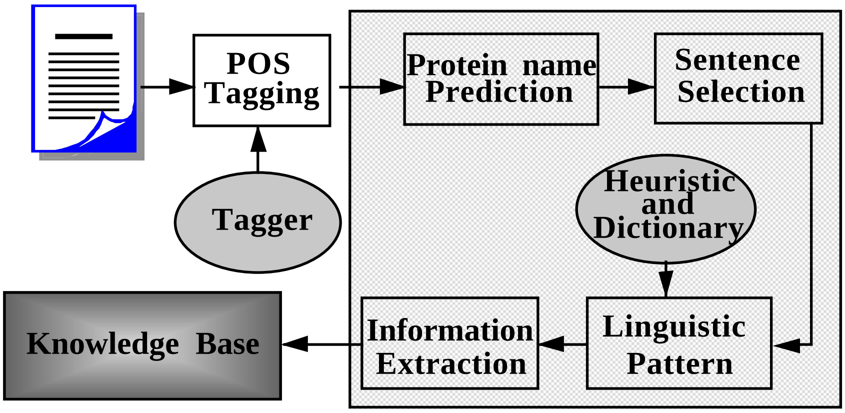

Figure 1: Overall Architecture.

2 Overview

The titles and abstracts of biological papers are compiled in MEDLINE database. However, these are inadequate to extract the knowledge like protein-protein interaction. So we digitized the papers relevant to protein-protein interaction by OCR, and created the full text database. At first, figure legends were searched using keywords such as an experimental method, to specify the sentence which is describing the fact checked by the experiment, except for the research background and hypothesis. And, the sentence containing the number of the figure having the searched figure legend was taken from the result section. Then, the sentence containing the information on a target was narrowed down. We have developed the tool which helps a man in reading such selected sentence and extracting knowledge.

Our final purpose is to extract the knowledge written in papers and to create the entry of a database automatically, and it is a target to automate the shaded portion which was shown in Fig. 1.

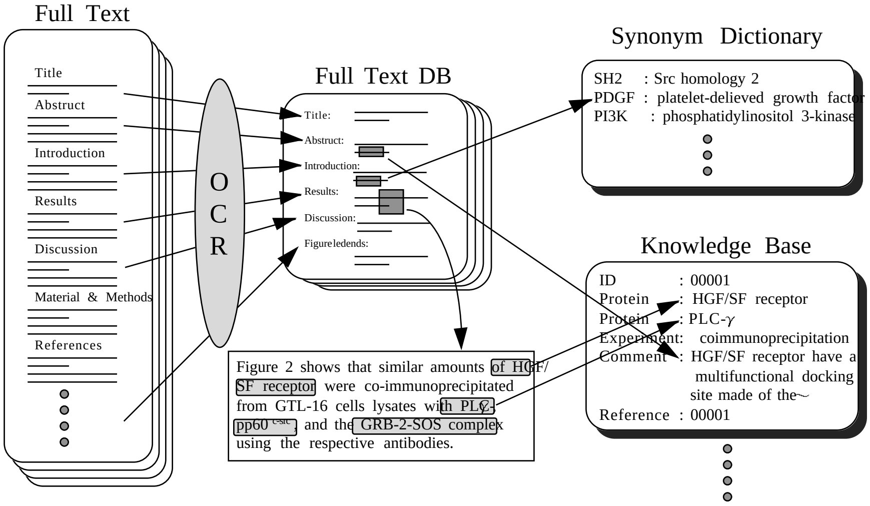
Full Text
Title
Abstruct
Introduction
Results
Discussion
Material & Methods
References
OCR
Full Text DB
Title:
Abstruct:
Introduction:
Results:
Discussion:
Figureledends:
Synonym Dictionary
SH2 : Src homology 2
PDGF : platelet-delieved growth factor
PI3K : phosphatidylinositol 3-kinase
Knowledge Base
ID : 00001
Protein : HGF/SF receptor
Protein : PLC-γ
Experiment: coimmunoprecipitation
Comment : HGF/SF receptor have a multifunctional docking site made of the~
Reference : 00001
Figure 2 shows that similar amounts of HGF/SF receptor were co-immunoprecipitated from GTL-16 cells lysates with PLC-γ, pp60 c-src, and the GRB-2-SOS complex using the respective antibodies.

Figure 2: Information Extraction.

Acknowledgments

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

References

[1] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T., Toward Information Extraction: Identifying protein names from biological papers, *Proc. of the Pacific Symposium on Biocomputing '98* (PSB'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 the Fifth International Conference on Intelligent Systems for Molecular Biology* (ISMB'97), 218-225, 1997