Automatic Extraction of Information on Protein-Protein Interaction from Scientific Literature

Toshihide Ono [1] ono@otsuka.gr.jp
Haretsugu Hishigaki [12] hisigaki@ims.u-tokyo.ac.jp
Akira Tanigami [1] atanigam@otsuka.gr.jp
Toshihisa Takagi [2] takagi@ims.u-tokyo.ac.jp

[1] Otsuka GEN Research Institute, OTSUKA Pharmaceutical Co., Ltd., 463-10 Kagasuno, Kawauchi-cho, Tokushima 771-0192, Japan
[2] Laboratory of Genome Database, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan

1 Introduction

Biological processes are controlled by direct and specific molecular interactions, involving DNA, RNA and proteins. The integration of the structure and function of them with the knowledge of macromolecular interactions and networks is an important step towards the construction of a unified and physiological view of the organism. Most of the data resources on biological functions such as expressed patterns and interactions are still only in the biological literature. The rapid growth of these collections makes it difficult for human beings to access the required information in a convenient and effective manner. The problem is that most of the documents are written in a natural language that computers can not deal with easily, and it is time-consuming for human beings to extract the knowledge from them. Therefore, the biologists have started asking for an intelligent information extraction system be developed method to save time and labor [1, 2, 3]. Here, we propose a system for the information extraction of protein-protein interactions from scientific text. We think that the system presented here could support biological researchers in various situations, e.g. constructing a database on protein interaction.

2 Method

Fig. 1 shows the architecture of the system. Our technique used only surface clues based on the word patterns that were presented by the word positions. The data sources of information used were a Medline abstract subset. First of all, we constructed the protein synonym database based on the yeast proteins from Saccharomyces genome database (SGD) [4]. Then, we collected the several frequently encountered words, called keywords (such as “interact” and “associate”), that were related to protein interaction from biological literatures. In the next step we searched for particular patterns including the keyword. The patterns representations was defined by the position between the keyword, protein names, and other characteristic word, such as prepositions in the sentences. The example of pattern is “proteinA interact with proteinB”. After that, each sentence contained the pattern was filtered with the rules based on the grammatical part of speech information. The last step was to search whether each sentence had any identifiable protein name and the pattern described above.

3 Result

We are developing a system of extracting the relationships between proteins by searching frequently seen keywords, their patterns created by surface clues, and a protein dictionary. In this study, we

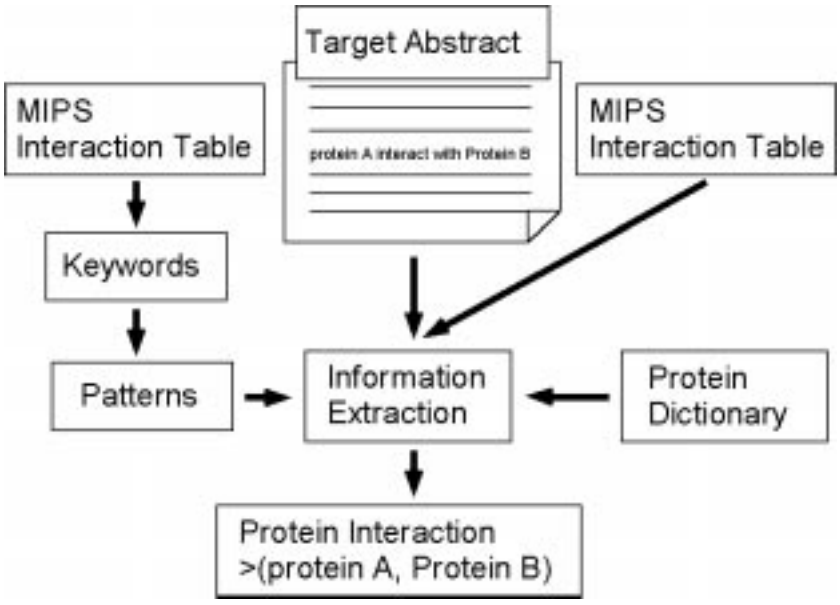

Figure 1: Overall Architecture.

obtained a recall of 86.8% and a precision of 94.4% for all keywords when using the yeast gene symbols as the protein dictionary. Although our system needs a complete dictionary of the protein names, we can obtain a high precision and recall without complicated NLP technique. This suggests that it can be practically used as support to extract protein interaction data when a protein dictionary becomes available. We think that this system may become a powerful tool for creating a database, such as protein interaction, from a huge variety of public databases.

Acknowledgement

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] Andrade, M.A. and Valencia, A., Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, *Bioinformatics*, 14:600–607, 1998.

[2] Craven, M. and Kumlien, J., Constructing biological knowledge bases by extracting information from text sources, *Proc. ISMB '99*, *AAAI Press*, 77–86, 1999.

[3] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T., Toward information extraction: identifying protein names from biological papers, *Proceeding of the Pacific Symposium on Biocomputing (PSB98)*, 697–718, 1998.

[4] Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D., SGD: Saccharomyces genome database, *Nuclecic Acids Res.*, 26:73–79, 1998.