Establishment of Transcription Factor DataBase TFDB Maintenance System

Masako Kaizawa [12] **Tomoko Okazaki** [1] **Hiroshi Mizushima** [1]
mkaizawa@info.ncc.go.jp tokazaki@ncc.go.jp hmizushi@ncc.go.jp

[1] Bioinformatics Section,Cancer Information and Epidemiology Division
National Cancer Center Research Institute
5-1-1 Tsukiji, Chuo-ku, Tokyo 104, Japan

[2] System Science Department, Mitsubishi Research Institute, Inc.
2-3-6 Otemachi, Chiyoda-ku, Tokyo 100, Japan

Abstract

Transcription Factor Database (TFD) was originally maintained by D.Ghosh of National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institute of Health, but the updating was terminated in 1993. Therefore, we started a new database TFDB using SYBASE System 11.

This time, we established TFDB Maintenance System which gathers transcription factor data from written articles, to update TFDB systematically.In this paper, we describe about TFDB Maintenance System, which contains the following subsystems: (1) Information Retrieval Subsystem, (2) Information Extraction Subsystem, and (3) the Java applet based Data Registration Subsystem by which the operators can register new data easily and interactively. This system is suitable to read and authorize the data easily by molecular biologists at different places through internet.

1 Introduction

TFD [1]–[4] was a very useful and required database for molecular biologists analyzing transcription mechanisms and gene expressions. TFD had been maintained by David Ghosh at NCBI until 1993. We took over his work to establish a new database system named TFDB which is based upon the site table of the original database [5, 6]. It was difficult to collect information about TFD systematically because of so many reports about transcription factor. Therefore, we established TFDB Maintenance System which gathers transcription factor data from written articles, to update TFDB.

2 System

2.1 Overview

This TFDB Management system handles database developed using SYBASE System 11. The system contains the following subsystems: (1) Information Retrieval Subsystem based on retrieval engine described in [7] that collects references related to transcription factor from MEDLINE correctly and efficiently, (2) Information Extraction Subsystem based on Information Extraction System described in [7] to extract candidates of 'transcription factors' from the collected references, and (3) the Java applet based Data Registration Subsystem by which the operators can register new data easily and interactively.

2.2 Database

The original data of TFD is converted into TFDB created by SYBASE System 11. The TFDB consists of 'Factor ID', 'Factor Name', 'Factor binding sequence', 'Reference', and 'MEDLINE ID'.

2.3 TFDB Maintenance System

This system consists of following three subsystems.

2.3.1 Information Retrieval Subsystem

This subsystem is based on retrieval engine described in [7], focuses on text retrieval which is based on text similarity. It also employs the basic vector space model.

2.3.2 Information Extraction Subsystem

This subsystem is based on Information Extraction System described in [7]. In this process, candidates of 'transcription factor' are extracted automatically from the collected references.

On the other hand, candidates of 'transcription factor binding sequence' are extracted by perl script [8], using pattern matching. Extracted reference text data by above process, is converted into HTML format by using perl script. Those candidates of 'transcription factor' and 'binding sequence' are marked with HTML tag which is included "On click" event handler of java script, to make clickable on WWW interface.

2.3.3 Data Registration Subsystem

We asked molecular biologists working at the bench, who have expert knowledge of transcription mechanisms, to read and authorize of extracted data to be able to input data in different places. We developed a data registration system using WWW on the Internet.

Submitted data is registered to TFDB, by SYBASE Web SQL.

3 Conclusions

With this system, the operators can register new data easily and interactively. This system is suitable for asked molecular biologists to read and authorize the data, and for management of the database.

Although, we have not analyzed performance of this system, we can collect good references and extract candidate of 'factors' and 'binding sequences' very easily. We will enhance the system performance as the next step.

In future, we are going to develop useful systems on a WWW interface, (1) transcription factor searching system, and (2) transcription factor binding site prediction system.

The former system is able to search TFDB data by WWW.

The latter one is a system, that return results of predictional transcription factor binding site, transcription factor name, and relationship of diseases when users input DNA sequence with unknown functional sites and regulatory sites.

These system will be opened for use near future on National Cancer Center WWW server.

Acknowledgments

We thank Takagi-laboratory (Human Genome Center, Institute of Medical Science, The University of Tokyo) for helping to establish this system. 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] Ghosh, D.,*Nucleic Acid Research*, Vol.18, pp.1749–1456, 1990.
[2] Ghosh, D.,*Trends in Biochemical Sciences*, Vol.16, pp.455–457, 1991.
[3] Ghosh, D.,*Nucleic Acid Research*, Vol.20S, pp.2091–2093, 1992.
[4] Ghosh, D.,*Nucleic Acid Research*, Vol.21S, pp.3117–3118, 1993.
[5] Mizushima, H., *Proceeding of 15th Japanese Molecular Biology Meeting*, 1992.
[6] Mizushima, H., *Genome Informatics Workshop 1994*, 1994.
[7] Ohta, Y. Yamamoto, Y. Okazaki, T. Uchiyama, I Takagi, T., Proc. 5th International Conference on Intelligent System for Molecular Biology *ISMB '97*, pp218–225, 1997.
[8] Okazaki, T. Kaizawa, M., Mizushima, H., *Genome Informatics Workshop 1996*, 1996.