Maintenance of Transcription Factor DataBase TFDB by *TFDB Maintenance System*

Masako Kaizawa [14]
mkaizawa@info.ncc.go.jp
Satoru Watanabe [2]
stwatana@gan2.ncc.go.jp
Takahiro Nobukuni [3]
tnobukun@gan2.ncc.go.jp
Masami Horikoshi [5]
horikosh@gene.selector.trc-net.co.jp
Hiroshi Handa [6]
hhanda@bio.titech.ac.jp
Yoshiyuki Kuchino [2]
ykuchino@ncc.go.jp
Takao Sekiya [3]
tsekiya@ncc.go.jp
Hiroshi Mizushima [1]
hmizushi@ncc.go.jp

[1] Cancer Information and Epidemiology Division, National Cancer Center Research Institute
[2] Biophysics Division, National Cancer Center Research Institute
[3] Oncogene Division, National Cancer Center Research Institute
5-1-1 Tsukiji, Chuo-ku, Tokyo 104, Japan
[4] System Science Department, Mitsubishi Research Institute, Inc.
2-3-6 Otemachi Chiyoda-ku, Tokyo 100, Japan
[5] Horikoshi Gene Selector Project, Exploratory Research for Advanced Technology
5-9-6 Tokodai, Tsukuba, Ibaraki, Japan
[6] Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology
4259 Nagatsuta-cho, Midori-ku, Yokohama, Japan

1 Introduction

TFD [1]–[4] was a very useful and required database for molecular biologists analyzing transcription mechanisms and gene expressions, which was maintained by David Ghosh at NCBI until 1993. We took over his work as TFDB (which is based upon the *'sites'* table of the TFD [6, 7]), and we established *TFDB Maintenance System* described in [5] which gathers transcription factor data from articles, to update TFDB systematically.

We started the maintenance of TFDB with this *TFDB Maintenance System* using many journals which we can obtain from MEDLINE database. In this paper, we describe how we maintain the TFDB using this system.

2 System

TFDB Maintenance System contains the following subsystems: (1) *Information Retrieval Subsystem* based on retrieval engine described in [8] which collects references related to transcription factor from MEDLINE correctly and efficiently, (2) *Information Extraction Subsystem* described in [5, 8] which extracts candidates of 'transcription factors' and 'target(transcription factor binding) sequences' from the result of *Information Retrieval Subsystem*, and (3) *Data Registration Subsystem* [5] which enables to register new data easily and interactively on WWW (the interface of *Data Registration Subsystem*, see Fig. 1).

3 Method

We are maintaining TFDB by the following method. (1) Collecting abstracts of the various fields, published in 1990 – 1998 from MEDLINE database as a text data format. (2) Scoring abstracts related to the transcriptional regulation using *Information Retrieval Subsystem*. (3) The words related to

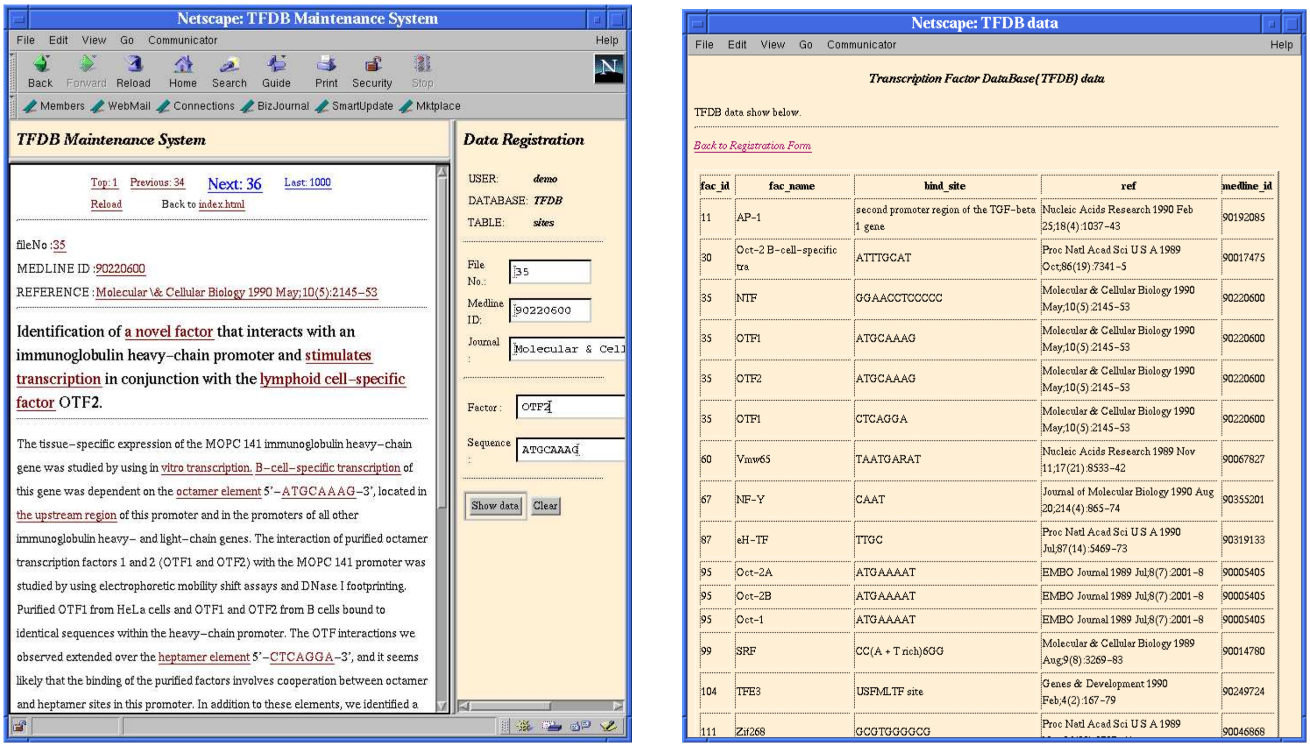

Figure 1: TFDB Data Registration interface and new data of TFDB.

transcription factors and its *binding sequences* are automatically extracted from (2) with *Information Extraction Subsystem*. (4) Specialist authorizes the validity of (3) and register with *Data Registration Subsystem*.

4 Results and Discussion

We used 265,249 abstracts for the purpose of trying the performance of the *TFDB Maintenance System*. We chose the top 150 abstracts from the output result of *Information Retrieval Subsystem*.

All of the 150 abstracts related to transcriptional regulatory mechanisms. About 7 candidates of binding factor and about 1.6 candidates of binding sequence appeared per abstract in average. Analysis of experts selection resulted that 10% of the top 150 abstracts contains new TFDB data with Binding Sequences.

We can collect data related to transcription factors efficiently with *Information Retrieval subsystem* and extract candidate of 'Binding Factors', 'Binding Sequences' very easily with *Information Extraction subsystem*. We are now continually authorizing the extracted data with *Data Registration Subsystem*.

This system is also suitable for maintaining databases in other fields.

Acknowledgements

We thank Dr. Sarai (The Institute of Physical and Chemical Research, GENE BANK) for giving us advices about 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., A relational database of transcription factors, *Nucleic Acid Research*, 18:1749–1456, 1990.

[2] Ghosh, D., New developments of a transcription factors database , *Trends in Biochemical Sciences*, 16:455–457, 1991.

[3] Ghosh, D., TFD:the transcription factors database, *Nucleic Acid Research*, 20S:2091–2093, 1992.

[4] Ghosh, D., Status of the transcription factor database (TFD), *Nucleic Acid Research*, 21S:3117–3118, 1993.

[5] Kaizawa, M., Okazaki, T., and Mizushima,H., Establishment of Transcription Factor DataBase TFDB Maintenance System, *Genome Informatics 1997*, Universal Academy Press, 292–293, 1997.

[6] Mizushima, H., Establishment of a Program for Searching Transcription Factor Binding Region, and Analysis of tRNA/Gln Gene, *Proceeding of 15th Japanese Molecular Biology Meeting*, 1992.

[7] Mizushima, H., Hayashi, K., Establishment of Transcription Factor Database and Human Mutation Database, *Proc. Genome Informatics Workshop 1994*, Universal Academy Press, 142–143, 1994.

[8] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I., and Takagi,T., Automatic Construction of Knowledge Base from Biological Papers, *Proc. 5th International Conference on Intelligent System for Molecular Biology (ISMB '97)*, 218–225, 1997.

[9] Okazaki, T., Kaizawa, M., and Mizushima,H., Establishment and Management of Transcription Factor DataBase TFDB, *Genome Informatics 1996*, Universal Academy Press, 218–219, 1996.