Maintenance of Transcription Factor DataBase TFDB

Masako Kaizawa [1] mkaizawa@info.ncc.go.jp
Satoru Watanabe [2] stwatana@gan2.ncc.go.jp
Takahiro Nobukuni [3] tnobukun@gan2.ncc.go.jp
Masami Horikoshi [4] horikosh@gene.selector.trc-net.co.jp
Hiroshi Handa [5] 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 104-0045, Japan
[4] Horikoshi Gene Selector Project, Exploratory Research for Advanced Technology
5-9-6 Tokodai, Tsukuba, Ibaraki 300-2635, Japan
[5] Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology
4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan

1 Introduction

TFD [1]–[4] was a very useful and required database for molecular biologists analyzing transcription mechanisms and gene expressions, which was originally 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) [7, 8], and we established *TFDB Maintenance System* [5, 6] which gathers transcription factor data from articles, to update TFDB systematically.

We have been maintaining TFDB with this system using many journals which we can obtain from MEDLINE database. In this paper, we describe about the TFDB maintenance with *data register group of TFDB*.

2 System

TFDB Maintenance System [5, 6] contains the following subsystems: (1) *Information Retrieval Subsystem* (IR) based on retrieval engine [9]. (2) *Information Extraction Subsystem* (IE) [5, 9] which extract candidates of 'transcription factors' and 'factor binding sequences' from the result of (1), and (3) *Data Registration Subsystem* (DR) [5, 6] which enables to register new data easily and interactively on WWW.

3 Method

We have been maintaining TFDB using our *TFDB Maintenance System* (*IR , IE , DR subsystems*) [6]. We can collect and extract references/data related to transcription factors efficiently with this *TFDB Maintenance System*. But an authorization by the specialists is indispensable in the stage of data registration to the database (TFDB), we formed the *TFDB data registration group* whose speciality is the transcriptional regulation and its mechanisms. We also established a *consensus of data registration* to standardize the quality of the registered data.

3.1 The data registration consensus

Established consensus of data registration is as follows;

1. *The factor* should be described as "transcription factor" in the abstract.
2. *The factor* should be described as "regulating the transcription of a gene" in the abstract.
3. Binding sequences of *the factor* should be describe in the abstract.

If *the factor* satisfies all of the three requirements, we register *the factor*, its binding sequence, bibliographic data and Medline ID to TFDB by WWW interface of *Data Registration Subsystem*.

Table 1: Rate of approval

File No.range	Approved	Pending	Rejected	Approval(%)	Range of KL
1–200	21	44	135	10.5%	0.463–0.753
201–400	20	159	21	10.0%	0.414–0.463
401–600	13	44	143	6.5%	0.381–0.414
601–800	10	74	116	5.0%	0.355–0.381
801–1,000	16	49	135	8.0%	0.333–0.354
1,001–1,200	5	41	154	2.5%	0.314–0.333
1,201–1,400	10	61	129	5.0%	0.297–0.314
1,401–1,600	5	30	165	2.5%	0.284–0.297
1,601–1,800	7	21	172	3.5%	0.272–0.284
1,801–2,000	3	86	111	1.5%	0.262–0.271
total	110	609	1,281	5.5%	0.262–0.753

4 Results and Discussion

We used 265,249 abstracts in MEDLINE 1990, and we chose the top 2,200 abstracts scored by KL from the result of *IR Subsystem*. All of the top 2,200 abstracts related to transcriptional regulatory mechanisms, and analysis of group selection resulted that 5.5% of the top 2,000 abstracts contains new TFDB data with its binding sequences. We can collect references related to transcription factors and its mechanisms efficiently with *TFDB Maintenance system*.

By the tendency of registration pattern in 1990, we decided to use top 1,400 abstracts (KL > 0.3) as they highly includes new TFDB data (Table 1). So, we are continually authorizing the extracted top 1,400 abstracts each year (1991–) by *IR Subsystem*, and *TFDB group* authorizing those abstracts preferentially.

Acknowledgements

We thank Dr. Sarai (The Institute of Physical and Chemical Research, GENE BANK) for giving us advices about this system, and TFDB group. 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 databaase TFDB maintenance system, *Genome Informatics 1997*, 8:292–293, Universal Academy Press, 1997.

[6] Kaizawa, M., Watanabe, S., Nobukuni, T., Horikoshi, M., Kuchino, Y., Sekiya, T., and Mizushima, H., Maintenance of transcription factor database TFDB by *TFDB Maintenance System*, *Genome Informatics 1998*, 9:316–318, Universal Academy Press, 1998.

[7] Mizushima, H., Establishment of a program for searching transcription factor binding region, and analysis of tRNA/Gln gene, *Proceedings of 15th Japanese Molecular Biology Meeting*, 1992.

[8] Mizushima, H., Hayashi, K., Establishment of transcription factor database and human mutation database, *Genome Informatics Workshop 1994*, 142–143, Universal Academy Press, 1994.

[9] 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, AAAI Press, 1997.

[10] Okazaki, T., Kaizawa, M., and Mizushima, H., Establishment and management of transcription factor database TFDB, *Genome Informatics 1996*, 7:218–219, Universal Academy Press, 1996.