Establishment and Management of Transcription Factor Database TFDB

T. Okazaki [1] tokazaki@ncc.go.jp
M. Kaizawa [2] kaizawa@mri.co.jp
H. Mizushima [1] 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] Systems Science Department, Mitsubishi Research Institute, Inc.,
2-3-6 Otemachi Chiyoda-ku, Tokyo 100, Japan

Abstract

D. Gohsh of National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institute of Health originally maintained 'TFD (Transcription Factor Database)' from 1990. As NCBI stopped its maintenance since 1993, we started a new database, TFDB (Transcription Factor DataBase), to take over some parts of the database focusing to the DNA binding sequence data. To update the database with recent data, we developed system which search literature database exhaustively and extract related information from the abstracts of collected articles. We also developed mail server to search target sequence of transcription factor using this database.

1 Introduction

TFD [1]–[4] maintained by D. Gohsh of NCBI was a database with nine tables (clones, domains, factors, polypeptides, sites, methods, n_proteins, references and x_proteins). We previously made a program using 'sites' table to search potential DNA binding sites in a promoter region of a DNA sequence [5]. There are some other programs for easier TFD search. For molecular biologists working at the bench, especially those who are analyzing mechanisms of transcriptional regulation, these systems are important and useful.

In 1993, NCBI stopped the maintenance of TFD, but there were many requests for its update from researchers. So, we started to maintain a new database TFDB using data in TFD as a starting point [6].

Our final objective for this project is to elucidate the mechanisms of transcriptional regulation using this database.

2 Methods and Results

The original data in TFD is converted into TFDB on SYBASE System11. TFDB consists of factor_ID, factor_name, DNA-binding_sequence and reference_data.

Recent data is updated by searching the literature database 'MEDLINE'. To investigate the algorithm for searching the MEDLINE, we surveyed all articles in 'Molecular and Cellular Biology' published in 1995, and we modified the query for searching the database by analyzing the terms in the selected articles. Furthermore, we developed a system which extracts transcription factor name and target sequence from the abstracts of the collected articles by using several perl scripts.

We also developed a mail server which receives the query DNA sequence from users, and searches for the existence of target sequence of transcription factor using TFDB.

3 Discussion

As it has been a long time since TFD was not updated, we have to collect all missing data in this period. To update the database with recent data, it is important to get related information from articles without omissions. We tried to improve the precision and exhaustiveness of the algorithm to search the literature database.

We are going to develop a world wide web (WWW) interface for accessing this database.

References

[1] Gohsh, D., *Nucleic Acid Research*, Vol. 18, pp. 1749-1756, 1990.

[2] Gohsh, D., *Trends in Biochemical Sciences*, Vol. 16, pp. 455-457, 1991.

[3] Gohsh, D., *Nucleic Acid Research*, Vol. 20S, pp. 2091-2093, 1992.

[4] Gohsh, 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.