APPLICATION NOTE Genomics & Informatics Vol. 6(2) 95-97, June 2008

# A Bio-database Management System for the Monitoring and Automatic FTP of Public Databases

**Hongseok Tae[1], Jeong-Min Han[2], Bu-Young Ahn[3] and Kiejung Park[1]***

[1]Information Technology Institute, SmallSoft Co. Ltd., Daejeon 305-343, Korea, [2]Korean Medicine Information Division, Korea Institute of Oriental Medicine, Daejeon 305-811, Korea, [3]Contents Convergence Team, Korea Institute of Science and Technology Information, Daejeon 305-806, Korea

## Abstract

Many bioinformatics sites have managed local bio-databases, including major databases such as GenBank and PIR with update load. We have developed several programs to monitor the update status of these databases and to FTP them automatically. These programs can be used for maintaining local bio-databases as recent versions and providing up-to-date databases through FTP sites. Currently, the program serves major bio-databases and will extend to accommodate many more bio-databases.

***Availability:*** The trial version of this system is available from http://gate.smallsoft.co.kr:8088/bioftp.

***Keywords:*** bio-database, FTP, GenBank, monitoring, scheduling

## Introduction

Bio-databases have been produced very rapidly in number and quantity. Each original database that has produced a site for each database has developed its own updating system for the database internally and provides it publically. Additionally, many bioinformatics sites have tried to maintain many databases locally to provide information services based on them. As each database has its own update schedule or an irregular update scheme, the need to monitor and update many bio-databases has been a burden for such sites.

CCBB of the Korea Institute of Science and Technology Information has provided major bio-databases and information services as a Korean bioinformatics site. We have developed and constructed a management system for the automatic update of public bio-databases by implementing several programs to monitor and automatically FTP them at CCBB (www.ccbb.re.kr). Currently, the system supports major databases, such as GenBank (Benson, 1994), PIR (McGarvey, 2000), and PDB (Berman, 2000), and will accommodate many more bio-databases.

*Corresponding author: E-mail kjpark@smallsoft.co.kr
Tel +82-42-864-2524, Fax +82-42-385-9240
Accepted 16 June 2008

## Features and Results

The system is composed of three major parts (Fig. 1). The first part monitors each database. A web robot program checks the version of each database by a scheduled plan that is controlled by a scheduling server program. The web robot program accesses the version information of the each database and compares the most recently updated information with a parser. The comparison and scheduled access are managed by each public DB entry, which is stored in MySQL DB (Table 1).

If an update is detected by the web robot program, the second part, the automatic FTP, is triggered, and raw data from each database are transferred from its FTP site. The progress is monitored by a web status re-

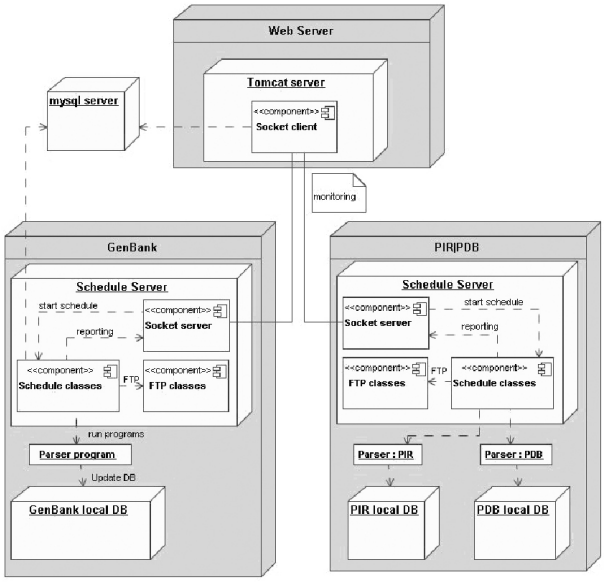


**Fig. 1.** The system structure for scheduled monitoring and automatic updating of bio-databases.

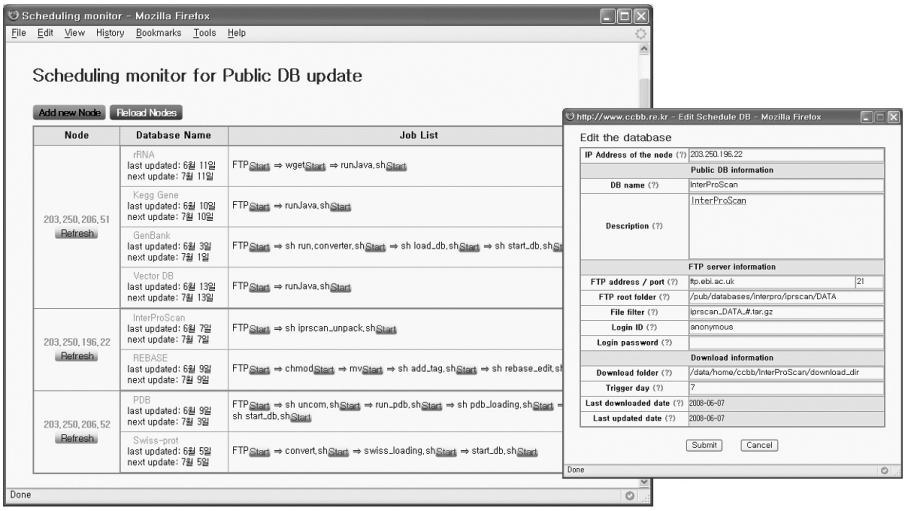


**Fig. 2.** (A) Monitoring the progress of the database update. (B) Editing database entry information.

**Table 1.** Summary of control data for scheduled monitoring and automatic update of bio-databases

| PublicDB | Type | Public DB Information |
|---|---|---|
| db_id | SMALLINT | public DB ID |
| db_name | VARCHAR (20) | public DB name ex) GenBank |
| description | VARCHAR (255) | public DB description |
| ftp_addr | VARCHAR (20) | public DB ftp address ex) ftp.ncbi.nih.gov |
| ftp_port | SMALLINT | public DB ftp port number |
| ftp_root_dir | VARCHAR (255) | Corresponding root dir of the public DB |
| file_filter | VARCHAR (255) | file type of public DB data ex) *.gz |
| login_id | VARCHAR (20) | Login id of public DB if necessary |
| login_pwd | VARCHAR (20) | Password of public DB if necessary |
| download_ip | VARCHAR (20) | ip address of a downloading server (the node with parsers and schedule programs) |
| download_dir | VARCHAR (255) | Downloading directory |
| trigger_day | TINYINT | download start day of each month |
| downloaded_date | DATE | The last downloaded date of public DB |
| updated_date | DATE | The last locally updated date of public DB |

port (Fig. 2A).

After the raw database is downloaded, the third part, the local database update, is triggered, and parsing and loading programs for public databases are executed to construct the final local database.

Many databases can be added to the system for automatic update using a database entry management program (Fig. 2B), which supports the editing of database information (Table 1), including addition, deletion, and modification.

## Discussion

Due to the exponential growth of public bio-databases, it has been requested that the integrated management and service system for those data function automatically. To address this need, we have developed a bio-database management system for the monitoring and automatic FTP of public databases. It includes the scheduled monitoring of raw data updates, file transfer of raw databases, and execution of updating programs to reconstruct local databases. The developed system can be generally used for most public bio-databases.

Local databases are managed by several DBMS, such as MySQL, Oracle, and KRISTAL. In fact, some databases are managed by MySQL, and some databases are managed by KRISTAL at KISTI. Because the local database update programs are independent of the developed system, the system can be useful for the update of most bio-databases without a change in the programs. A more convenient program can be developed to improve the system.

### Acknowledgements

This work was supported by the Korea of Science and Technology Information grant from MOST.

## References

Benson, D.A., Boguski, M., Lipman, D.J., and Ostell, J. (1994). GenBank. *Nucleic Acid Res.* 22, 3441-3444.

McGarvey, P.B., Huang, H., Barker, W.C., Orcutt, B.C.,

Garavelli, J.S, Srinivasarao, G.Y., Yeh, L.L., Xiao, C., and Wu, C.H. (2000). PIR: a new resource for bioinformatics. *Bioinformatics* 16, 290-291.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235-242.