APPLICATION NOTE

Genomics & Informatics Vol. 5(4) 196-199, December 2007

Biological Object Downloader (BOD) Service for Easy Download and Management of Biological Databases

Daeui Park[1], Jungwoo Lee[1], Giseok Yoon[2], Sungsam Gong[3] and Jong Bhak[1*]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea, [2]Object Interaction Technologies Inc., Daejeon 305-806, Korea, [3]Biochemistry department, Cambridge University, Cambridge, CB2 1TN, UK

Abstract

BOD is an FTP service management tool on the Internet. It was developed for biological researchers in South Korea. It enables easier and faster access of bioinformation without having to go through foreign FTP sites. BOD includes an automatic downloader with a management and email alert service from which the user can easily select and schedule any biological database. Once listed in BOD, the user can check and modify the download status and data from an additional email alert service.

Availability: http://ftp.kobic.kr, ftp://ftp.kobic.kr, and http://bioftp.org

Keywords: FTP services, biological data, downloader

Introduction

Since the completion of the Human Genome Project, the amount and complexity of biological data have been increasing at exponential rates (Dennis *et al*., 2004, Andreas, 2003). Nowadays, the term "biological data," in fact, means complex information. It often implies a network of databases that need to be downloaded and managed constantly due to rapid update rates. In genomics, data formats often reflect genome sequences derived from automatic experiments and biological chip data such as cDNA microarrays and SNP arrays. In the proteomics field, data are derived from measures applied to a more complex condition: 2D-PAGEs, mass spectrometers, and 3D protein structures. Also, data describing such as processes as promoter and protein-protein interactions are produced by computational predictions. These data collectively form omics resources and are deposited in various public databases for sharing with other researchers.

However, there are too many omics and ome databases to download, and it is impractical for biologists to manage them. Currently, there are several programs for downloading and managing biological data, such as BioMirror (Gilbert *et al*., 2004), which is a public service for high-speed access to up-to-date DNA and protein sequence databanks. However, the user cannot directly update the download lists in this service. In addition, BioDownloader (Shapovalov *et al*., 2007) is a program for easy download and update of files from FTP and http servers. BioDownloader can be installed on a local computer using either the Microsoft Windows or Mac operating system.

Since 2003, we have developed a web service tool, the Biological Object Downloader (BOD), which easily downloads and updates large amounts data from the various FTP servers for Korean bioinformatists, biologists, and medical scientists. BOD can be used for accessing the newest biological data abroad without installing any databases at the local computer. BOD was initially designed for local database administrators, but it also can be a useful data warehouse for end users. BOD mainly aims to automate downloads and management of various biological databases. BOD not only enables any user to automatically download databases, but also notifies the user about the latest updates of these databases via email. BOD is part of an effort to share bioinformation openly and freely. It is a daughter project of BioFTP (http://bioftp.org), wherein global biological information resources can be updated and downloaded with maximum convenience. BOD will serve as a file download and management branch of the biological community cluster (BioCC, Gong *et al*., 2006).

Results

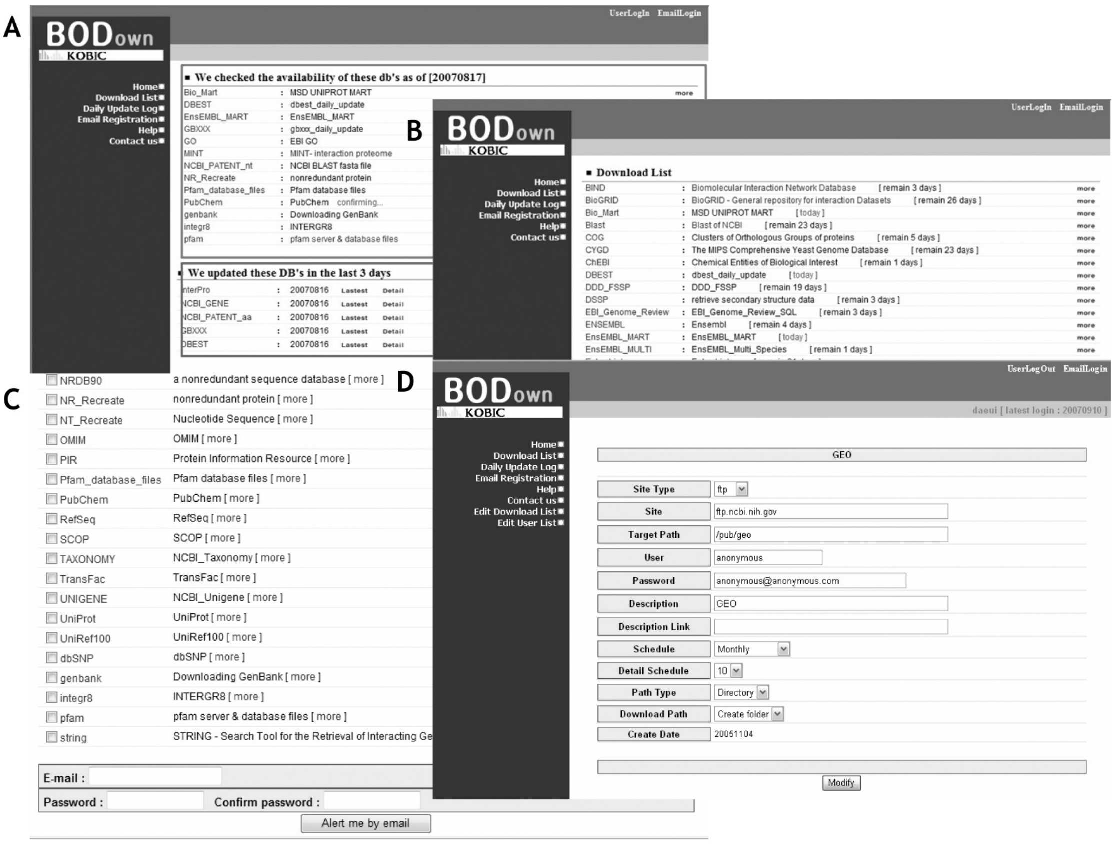
BOD is a useful biological database download and management system that can be accessed through the web. BOD uses the Apache web server with the MySQL database. Its main role is to act as a hub of biological database downloads for biologists. It is automated and effective for large-scale database handling. It serves genomic (NCBI, Pruitt *et al*., 2001; PrimateDB, Woo *et al*., 2005), transcriptomic (dbEST, Boguski *et al*., 1993; GAzer, Kim *et al*., 2007), proteomic (SCOP, Murzin *et al*.,

*Corresponding author: E-mail jong@kribb.re.kr or j@bio.cc
Tel +82-42-879-8500, Fax +82-42-879-8519
Accepted 22 October 2007

1995; PSIbase, Gong *et al*., 2005; InterPare, Gong *et al*., 2005), interactomic (InterAct, Hermjakob *et al*., 2004), variomic (dbSNP, Sherry *et al*., 2001; SNP@Ethnos, Park *et al*., 2007; SNP2NMD, Han *et al*., 2007; D2GSNP, Kang *et al*., 2006; SNP@Domain, Han *et al*., 2006), and textomic (EntrezLink, Maglott *et al*., 2007) bioinformation. BOD's unique characteristics are the following: 1) it is easily accessible through the web, and 2) it notifies new updates to users through email. Although it is mainly for local users who want to download large databases frequently, BOD will be developed as a general omics database management system in the future.

General usage and webpage interface

Fig. 1A shows the main page of BOD, found at http://ftp.kobic.kr. The red box shows a list of databases that will be updated when the user accesses the webpage. The blue box shows a list of databases updated in the most recent 3 days. When the user clicks the data list (e.g., Interpro, NCBI_GENE, DBEST) in the blue box, the KOBIC FTP site will be linked. "Lastest" is connected to the current directory of the FTP site with the latest data. "Detail" shows a specific list of the newest data, and each "Detail" is linked to the file via FTP. When the user clicks "Download List" on the top left of Fig. 1A, as in Fig. 1B, all the data lists registered in BOD can be seen. Lists with complete downloads and the newest data are highlighted in blue and are linked to the FTP service. Lists that are downloading show when the download will be completed. The user can click "Daily Update Log" from the navigation panel in the side menu located on the left to view only the lists with completed downloads.



■ We checked the availability of these db's as of [20070817]

Bio_Mart	:	MSD UNIPROT MART
DBEST	:	dbest_daily_update
EnsEMBL_MART	:	EnsEMBL_MART
GBXXX	:	gbxxx_daily_update
GO	:	EBI GO
MINT	:	MINT- interaction proteome
NCBI_PATENT_nt	:	NCBI BLAST fasta file
NR_Recreate	:	nonredundant protein
Pfam_database_files	:	Pfam database files
PubChem	:	PubChem confirming...
genbank	:	Downloading GenBank
integr8	:	INTERGR8
pfam	:	pfam server & database files

■ We updated these DB's in the last 3 days

nterPro	:	20070816	Lastest	Detail
NCBI_GENE	:	20070816	Lastest	Detail
NCBI_PATENT_aa	:	20070816	Lastest	Detail
GBXXX	:	20070816	Lastest	Detail
DBEST	:	20070816	Lastest	Detail

■ Download List

BIND	:	Biomolecular Interaction Network Database	[remain 3 days]	more
BioGRID	:	BioGRID - General repository for interaction Datasets	[remain 26 days]	more
Bio_Mart	:	MSD UNIPROT MART	[today]	more
Blast	:	Blast of NCBI	[remain 23 days]	more
COG	:	Clusters of Orthologous Groups of proteins	[remain 5 days]	more
CYGD	:	The MIPS Comprehensive Yeast Genome Database	[remain 23 days]	more
ChEBI	:	Chemical Entities of Biological Interest	[remain 1 days]	more
DBEST	:	dbest_daily_update	[today]	more
DDD_FSSP	:	DDD_FSSP	[remain 19 days]	more
DSSP	:	retrieve secondary structure data	[remain 3 days]	more
EBI_Genome_Review	:	EBI_Genome_Review_SQL	[remain 3 days]	more
ENSEMBL	:	Ensembl	[remain 4 days]	more
EnsEMBL_MART	:	EnsEMBL_MART	[today]	more
EnsEMBL_MULTI	:	EnsEMBL_Multi_Species	[remain 1 days]	more

□ NRDB90	a nonredundant sequence database [more]
□ NR_Recreate	nonredundant protein [more]
□ NT_Recreate	Nucleotide Sequence [more]
□ OMIM	OMIM [more]
□ PIR	Protein Information Resource [more]
□ Pfam_database_files	Pfam database files [more]
□ PubChem	PubChem [more]
□ RefSeq	RefSeq [more]
□ SCOP	SCOP [more]
□ TAXONOMY	NCBI_Taxonomy [more]
□ TransFac	TransFac [more]
□ UNIGENE	NCBI_Unigene [more]
□ UniProt	UniProt [more]
□ UniRef100	UniRef100 [more]
□ dbSNP	dbSNP [more]
□ genbank	Downloading GenBank [more]
□ integr8	INTERGR8 [more]
□ pfam	pfam server & database files [more]
□ string	STRING - Search Tool for the Retrieval of Interacting Ge

E-mail :

Password : Confirm password :

daeui [latest login : 20070910]

GEO

Site Type	ftp
Site	ftp.ncbi.nih.gov
Target Path	/pub/geo
User	anonymous
Password	anonymous@anonymous.com
Description	GEO
Description Link	
Schedule	Monthly
Detail Schedule	10
Path Type	Directory
Download Path	Create folder
Create Date	20051104

Fig. 1. (A) BOD main page. The red box shows the list of data that will be updated when a user connects to the site. The blue box shows the list of recent updates. (B) Download list shows a list of all the data registered in BOD. The data registered on the list but not completely downloaded are shown in black. The lists with the newest data are shown in blue and are linked to the FTP sites. The list also shows the date of the latest update and the expected update date. (C) Email Registration. When the user checks a database and inputs an email address and password, the BOD system notifies the user whenever the checked lists are updated. (D) Download Manager. The user can register new target databases for downloading and can modify an existing list.

Table 1. Detailed description of Download Manager

Filed Name	Description
Site Type	The downloading protocol (http or FTP)
Site	The download target address
Target Path	The target path
Description	Description of the target
Description Link	The links for referencing outside sources
Schedule	The update period (Daily, Weekly, Every 10 days, Monthly)
Detail Schedule	The user should choose the number according to the period (Weekly, Every 10 days, Monthly)
Path Type	The user selects whether the target is the entire directory or just certain files
Download Path	Create Folder → creates a new folder for each update Same Folder → puts over the folder for each update

Receive an email about the newest update information

BOD provides the update status of the data lists for which the user has registered through email. To use this service, the user should click “Email Registration,” on the upper left side of Fig. 1B. New users should register their emails, and current users can log in via registered emails and modify their lists of database of interests or stop receiving emails. To modify lists, the user can log in via email and modify the checkboxes of the lists and press the “Alert me by email” button, as shown in Fig. 1C. To stop receiving email, uncheck any boxes, then BOD will assume the user has stopped using its service.

Download Manager

This function targets a public biological database to be listed on the KOBIC FTP server through FTP or http protocol. Currently, new lists can be listed by KOBIC researchers only. When a user clicks the “LogIn” button on the upper right side of the webpage and inputs a user name and password, the Download Manager menu of BOD is created on the navigation panel on the left side of the webpage. When the user clicks on that menu, the interface of the Download Manager can be seen. In Fig. 1D, a registered user is eligible to modify specific details of input information. BOD not only classifies a redundant database by DB Name, but also checks whether the database has the Target Path with the same site.

Conclusion

BOD is a web-based, automatic biodata download management system. It is automated and effective for large-scale database handling. BOD at present is a web service for national bioinformation dissemination within South Korea. It provides omics bioinformation of such entities as the genome, transcriptome, proteome, interactome, variome (such as SNP data), and textome (literature). For better service, BOD focuses on the synergistic effects in operation, by sharing and dividing tasks among biological researchers. BOD’s unique characteristics are the following: 1) it is easily accessible through the web, and 2) it notifies users of new updates through email. Although it is mainly for local users who want to download large databases frequently, BOD will be developed as a general biological database management system in the future.

Acknowledgments

We thank our colleagues at KOBIC and Maryana Bhak for editing the manuscript. This project was supported by the KRIBB Research Initiative Program, 07-38 program by Korea Agency Digital Opportunity and Promotion, and M10508040002-07N0804-0021 grant by MOST.

References

Baxevanis ,A. D. (2003). The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Research*. 31, No. 1, 1-12.

Han, A.R., Kang, H.J., Cho, Y.B., Lee, S.H., Kim, Y.J., and Gong, S.S. (2006). SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. *Nucleic Acids Research*. 34(Web Server issue), W642-W644.

Han, A.R., Kim, W.Y., and Park, S.M. (2007). SNP2NMD: A database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. *Bioinformatics*. 23(3), 397-399.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. *Nucleic Acids Research*. 32(Database issue), D711-D715.

Gilbert, D., Ugawa, Y., Buchhorn, M., Wee,T.T., Mizushima, A., Kim, H., Chon, K., Weon, S., Ma, J., W. J., Ichiyanagi, Y., Liou, D.M., Keretho, S., and Napis, S. (2004). Bio-mirror project for public bio-data distribution. *Bioinformatics*. 20, 2338-2340.

Maglott, D.R., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*. 35(Database issue), D26-D31.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004). IntAct-an open source molecular interaction database. *Nucleic Acids Research*. 32(Database issue), D452-D455.

Park, J.H., Hwang, J.H., Lee, Y.S., Kim, S.C., and Lee, D.H. (2007). SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms. *Nucleic Acids Research*. 35(Database issue), D711-D715.

Kang, H.J., Hong, T.H., Chung, W.H., Kim, Y.U., Jung, J.H., Hwang, S.H., Han, A.R., and Kim, Y.J. (2006). D2GSNP: a web server for the selection of Single Nucleotide Polymorphisms within human disease genes. *Genomics and Informatics*. 4(1), 45-47.

Pruitt, K.D., and Maglott, D.R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*. 29, No. 1, 137-140.

Kim,S.B., Yang, S.J., Kim, S.K., Kim, S.C., Woo, H.G., Volsky, D.J., Kim, S.Y., and Chu, I.S. (2007). GAzer: Gene Set Analyzer. *Bioinformatics*.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. 29, No. 1, 308-311.

Gong, S.S., Kim, T.Y., Oh, J.S., Kwon, J.K., Cho, S.A., Bolser, D.M., and Bhak, J. (2006). BioCC: An openfree hypertext Bio Community Cluster for Biology. *Genomics and Informatics*. 4(3), 125-128.

Gong, S.S., Yoon, G.S., Jang, I.S., Bolser, D., Dafas, P., Schroeder, M., Choi, H.S., Cho, Y.B., Han, K.S., Lee, S.H., Choi, H.H., Lappe, M., Holm, L., Kim, S.S., Oh, D.H., and Bhak, J. H. (2005). PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*. 21(10), 2541-2543.

Gong, S.S., Park, D.B., Choi, H.S., Ko, J.S., Jang, I.S., Lee, J.S., Bolser, D.M., Oh, D.H., Kim, D.S., and Bhak, J. (2005). A protein domain interaction interface database: InterPare. *Bioinformatics*. 6, 207.

Bogusk, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST-database for “expressed sequence tags”. *Nature Genetics*. 4, 332-333.

Shapovalov, M.V., Canutescu, A.A., and Dunbrack, R.L.Jr. (2007). BioDownloader: bioinformatics downloads and updates in a few clicks. *Bioinformatics*. 23(11), 1437-1439.

Murzin , A.G., Brenner, S.E., Hubbard , T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*. 247(4), 536-40.

Woo, T., Shin, G., Kang, T., Kim, B., Seo, J., Kim, S.S., and Kim, C.B. (2005). PrimateDB: Development of Primate Genome DB and Web Service. *Genomics and Informatics*. 3(2), 73-76.