BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3238–3240
doi:10.1093/bioinformatics/bth219

Bio-Mirror project for public bio-data distribution

Don Gilbert[1,*], Yoshihiro Ugawa[2], Markus Buchhorn[3], Tan Tin Wee[4], Akira Mizushima[5], Hyunchul Kim[6], Kilnam Chon[6], Seyeon Weon[7], Juncai Ma[8], Yoshihiro Ichiyanagi[8], Der-Ming Liou[9], Somnuk Keretho[10] and Suhaimi Napis[11]

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA, [2]Miyagi University of Education, Japan, [3]ANU Internet Futures Project, Australian National University, Australia, [4]BioInformatics Centre, National University of Singapore, Singapore, [5]National Agricultural Research Organization, Japan, [6]Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [7]Bioinformatics Research Laboratories, Co., Taejon, South Korea, [8]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China, [9]Bioinformatics Research Center, National Yang-Ming University, Taiwan, [10]Kasetsart University, Thailand and [11]Universiti Putra Malaysia, Malaysia

Received on May 14, 2003; revised on December 14, 2003; accepted on January 29, 2004
Advance Access publication April 1, 2004

ABSTRACT
Summary: Timely worldwide distribution of biosequence and bioinformatics data depends on high performance networking and advances in Internet transport methods. The Bio-Mirror project focuses on providing up-to-date distribution of this rapidly growing and changing data. It offers FTP, Web and Rsync access to many high-volume databanks from several sites around the world. Experiments with data grids and other methods offer future improvements in biology data distribution.
Availability: http://www.bio-mirror.net/
Contact: bio-mirror@apan.net

INTRODUCTION

The Bio-Mirror project for distribution of public biosequence and bioinformatics data is a collaborative service of several worldwide bioinformatics centers. As multi-gigabyte public bioinformatics databanks grow and change daily, access to them is hampered by limits on Internet bandwidth. The Bio-Mirror project addresses this problem with rapid redistribution from several sites. Such mirrors reduce the burden on source providers, and mitigate Internet outages and slow distant connections.

The Bio-Mirror project was formed in 1998 as a collaboration of Asia-Pacific bioinformatics centers (APBionet) and IUBio Archive at Indiana University. It distributes databanks using Internet2 connections between continents. Many APBionet members are part of the new and growing bioinformatics centers, where Bio-Mirror sites provide a short path to current data from US and European sources. Project sites serve data to a range of educational, government and industry bioinformatics groups. We also investigate new technologies in science data grid informatics to improve data distribution.

CONTENTS AND METHODS

Contents of the Bio-Mirror databank set include core databanks from the three collaborating DNA databanks: GenBank, EMBL and The DNA Databank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/). SwissProt, TrEMBL, PIR, PDB, Pfam and InterPro protein databanks are included, as well as BLAST and RefSeq mixed databanks. Ensembl, NCBI Genomes, LocusLink, euGenes, Unigene, Gene Ontology, and related genome and other data are included. The current collection exceeds 140 GB (compressed, or about 500 GB uncompressed). Approximately 10% of these change daily, and most are updated in the course of three months. Complete databank file sets are mirrored from sources in the US, Europe and Asia-Pacific. The European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/) and National Center for Biotechnology Information (NCBI, http://www.ncbi.nih.gov/) centers now include Bio-Mirror sites in their access information.

The Bio-Mirror project is provided as a public service by member centers with the support of several organizations. Participants include APBionet and bioinformatics centers in Japan, Australia, Singapore, China, Korea, Malaysia, Taiwan, Thailand and the USA (Table 1). High performance network

*To whom correspondence should be addressed.

Table 1. Bio-Mirror sites

Country	Host	Web site	Bulk data access
Australia	Australian National University	–	rsync: and ftp://bio-mirror.au.apan.net/biomirror/
China	Institute of Microbiology, Chinese Academy of Sciences	http://bio-mirror.cn.apan.net/	ftp://bio-mirror.cn.apan.net/pub/biomirror/
Japan	Computer Center for AFFRC	http://bio-mirror.jp.apan.net/	ftp://bio-mirror.jp.apan.net/pub/biomirror/
Korea	Korea Advanced Institute of Science and Technology	http://bio-mirror.kr.apan.net/	ftp://bio-mirror.kr.apan.net/pub/biomirror/
New Zealand	University of Auckland	http://www.biomirror.org.nz	ftp://biomirror.auckland.nz
Singapore	National University of Singapore	http://bio-mirror.sg.apan.net/	ftp://bio-mirror.sg.apan.net/biomirrors/
Taiwan	National Yang-Ming University	http://bio-mirror.ym.edu.tw/	ftp://bio-mirror.ym.edu.tw/biomirror/
Thailand	Kasetsart University	http://bio-mirror.ku.ac.th/	http://bio-mirror.ku.ac.th/biomirror/
USA	IUBio Archive, Indiana University	http://www.bio-mirror.net/	rsync: and ftp://bio-mirror.net/biomirror/

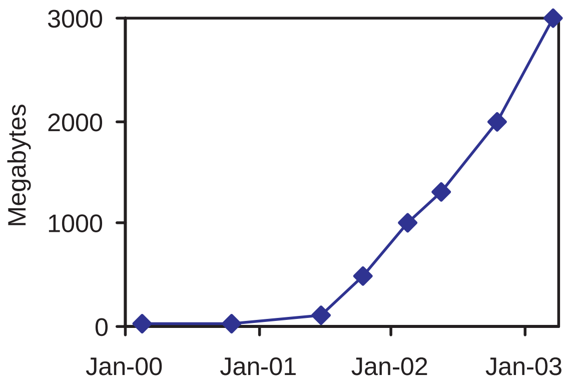

Fig. 1. Monthly data transfer at US Bio-mirror.

infrastructure and collaborative help have been essential to this project, including Trans-Pacific network (TransPAC) and Asia-Pacific Advanced Network (APAN).

Project sites currently serve many Terabytes per month to thousands of bioinformatics centers and laboratories. Bulk distribution has risen from 100 GB/month in year 2001 to 3 TB/month in 2003 at the US node (Fig. 1). File transfer protocol (FTP) provides the best access, as FTP servers have been tuned for large file transfer. The rsync protocol (Tridgell, 2002, http://rsync.samba.org/) is supported at some Bio-Mirror servers as an efficient alternative. Bio-Mirror sites also offer search and analyses services of these data with SRS (Zdobnov *et al*., 2002) and other programs. The Perl FTP mirror package (McLoughlin, 1998, http://sunsite.org.uk/packages/mirror/) is used at Bio-Mirror sites to maintain daily updates, with additional Perl tools for updating local databanks available in the project collection.

DISCUSSION

The quantity of popular bio-data served through this project has grown from under 10 GB in 1998 to 150 GB in 2003. Much of this has come with growth in core databanks. As Human Genome and related projects matured, genome databases such as Ensembl, euGenes and NCBI Genomes have added significant quantities of new data. European bioinformatics centers in the The European Molecular Biology network (EMBnet, http://www.embnet.org/) group offer similar bio-data distribution, and many bioinformatics groups see the need for improving data distribution. Others are encouraged to join the Bio-Mirror project.

Improving bio-data distribution. Rsync is a useful alternative to FTP, as it includes file system synchronization similar to the Perl mirror package that updates only changed files. Rsync also attempts to synchronize only changed sections within files, though for binary-compressed large databanks this may reduce efficiency. GridFTP (Allcock *et al*., 2002), another possible improvement, supports parallel transfers and other efficiency methods. Although it can double transfer rates in a local network, our tests indicate problems such as lack of anonymous transfers, limited support for 64-bit systems, limited mirroring options and bandwidth costs associated with parallel transfer.

Data grids. New technology for computable access to large databanks is being developed in the context of science data grids (Avaki Corporation, 2002, http://www.avaki.com/news/release20021209.html; OSGA-DAI, 2002, www.ogsadai.org.uk/; iVDGL, 2002, http://www.ivdgl.org/). Web services examples in bioinformatics include bio-databank access XML Central of DDBJ (XDDBJ, http://xml.nig.ac.jp/) and EMBL Nucleotide Sequence data in XML (XEMBL, http://www.ebi.ac.uk/xembl/). Lightweight Directory Access Protocol (LDAP) directories of bio-databanks are also available (Gilbert, 2002, http://iubio.bio.indiana.edu/biogrid/directories/). Open Grid Services Architecture (OGSA, Foster and Kesselman, 1999) has growing support in bioinformatics (EGD-WP10, 2000, http://edg-wp10.healthgrid.org/, http://marianne.in2p3.fr/datagrid/wp10/; MyGrid, 2002, http://www.mygrid.org.uk/), with a database access and integration component (OSGA-DAI, 2002) that is relevant.

Object versus Bulk distribution. Bioinformatics groups should consider investing in object or record-oriented exchange. When only changed records are updated, transport time and costs are minimized. Record-oriented transport has many uses, where selection of data subsets tied to search systems is often of most interest. For grid computing, one wants to rapidly distribute task and data subsets to many computer nodes. One practical goal for the Bio-Mirror project is to develop more efficient object-level distribution using available technology. Tests are in progress with FTP, LDAP, SOAP and SRS to search and retrieve gigabytes of biosequence records (Gilbert, 2002) and suggest that LDAP matches or surpasses FTP when data selection is included, and is about 10 times faster than Web services, due to its compact binary-encoded transport and more direct route from server storage to client applications.

ACKNOWLEDGEMENTS

Rick McMullen, Doug Pearson and Hisashi Eguchi (TransPac and APAN), Kazuya Tamura (Forestry and Forest Products Research Institute, Japan), Peter Stoehr and Rodrigo Lopez (EBI), Mark Cavanaugh and Scott McGinnis (NCBI) have provided valuable assistance. Bio-Mirror project has received support from Australian National University, Australia; Agriculture, Forestry and Fisheries Research Council, Japan; APBioNet of Asia-Pacific Advanced Network; Chinese Academy of Sciences, China; Korea Advanced Institute of Science and Technology, Korea; National Yang-Ming University, Taiwan; National Science Foundation and Indiana University, USA; and National University of Singapore, Singapore.

REFERENCES

Allcock,B., Bester,J., Bresnahan,J, Chervenak,A.L., Foster,I., Kesselman,C., Meder,S., Nefedova,V., Quesnal,D. and Tuecke,S. (2002) Data management and transfer in high performance computational grid environments. *Parallel Comput. J.*, **28**, 749–771.

Avaki Corporation (2002) Avaki Data Grid 3.0, Java-based data grid software.

EGD-WP10 (2000) Grid-aware biomedical applications for datagrid testbed assessment, WP10 of European Union Data Grid Project.

Foster,I. and Kesselman,C. (eds) (1999) *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers, San Francisco, CA.

Gilbert,D.G. (2002) Directories of Bio-data.

iVDGL (2002) The International Virtual-Data Grid Laboratory.

McLoughlin,L. (1998) Mirror perl package.

MyGrid (2002) The MyGrid project, University of Manchester, UK.

OSGA-DAI (2002) Open grid services architecture: data access and integration project.

Tridgell,A. (2002) Rsync, remote file synchronization system.

Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—recent developments. *Bioinformatics*, **18**, 368–373.