BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 8(5)

Distribution of biological databases over low-bandwidth networks

S Sikander Azam[1, 2], Shamshad Zarina[1*]

[1]National Center for Proteomics, University of Karachi, Karachi-75270, Pakistan; [2]National Center for Bioinformatics, Quaid-e-Azam University, Islamabad; Shamshad Zarina – Email: szarina@uok.edu.pk; Phone: 0092-21-34656511; Fax: 0092-21-34650726; *Corresponding author

Received February 07, 2012; Accepted March 03, 2012; Published March 17, 2012

Abstract:
Databases are integral part of bioinformatics and need to be accessed most frequently, thus downloading and updating them on a regular basis is very critical. The establishment of bioinformatics research facility is a challenge for developing countries as they suffer from inherent low-bandwidth and unreliable internet connections. Therefore, the identification of techniques supporting download and automatic synchronization of large biological database at low bandwidth is of utmost importance. In current study, two protocols (FTP and Bit Torrent) were evaluated and the utility of a BitTorren based peer-to-peer (btP2P) file distribution model for automatic synchronization and distribution of large dataset at our facility in Pakistan have been discussed.

Background:
During last couple of years, scientific community among developing countries including Pakistan has developed interest in bioinformatics research **[1, 3].** Most of the bioinformatics applications are database dependant and to access, search and retrieve data, reliable internet connections are required. To facilitate maximum utilization, bioinformatics resources and facilities around the globe prefer to download these databases on their local servers. There has been an exponential growth in database records as a consequence of major advances in genomics and proteomics technologies, stressing need of frequent updates with the latest releases. Many developing countries face a major problem in regular update of databases due to lack of infrastructure, slow/unreliable internet connectivity and low bandwidth. It is expected that in future, databases size would outgrow existing rate of transfer at current bandwidth, thus it is imperative to develop efficient tools for obtaining automatic updates on a regular basis. To address such issues, a Bio-Mirror project was also launched which uses FTP mode for data transfer **[4].**

Updates are usually managed by client server approach (FTP or WWW) or P2P (Peer-to-Peer) file sharing applications. FTP has been a traditional method for file sharing and downloading from remote server and is very popular for downloading large files. However, it requires large network bandwidths and suffers from scalability bottleneck. As an alternative, P2P applications have become immensely popular for fast and efficient distribution of files in recent years. P2P architecture operates in a distributed autonomous system mode that does not rely on a specific server system. Torrent protocol working environment is based on peer-to-peer (P2P) technique in which every user is connected to each other with mesh technology. On the other hand, FTP protocol working environment is completely dependent on a single server which means it may create single point of failure. The performance of traditional FTP file sharing applications deteriorates rapidly as the number of clients increase while in P2P module, more peers means better performance. There are many P2P file sharing applications such as Kazza, Gnutella, Napster, BitTorrent to name a few. Among these applications, BitTorrent P2P file sharing system has been analyzed in many studies **[5, 6].**

Considering the existing scenario and future difficulties, techniques supporting automatic synchronization of databases at low bandwidth are of utmost importance. In current study, efficiency of FTP and BitTorrent applications are compared in order to download large sized database (Gigabyte) and using

 © 2012 Biomedical Informatics

them through local servers without delay and response time out message. With the help of btP2P protocol, the problems of updating the enormous data of biological databases and at the same time, avoiding the network connection issues have been addressed.

Methodology:

Computational Resources

Two Ultra20M2 of Sun Microsystems based nodes with dual core processor were utilized for current study. These servers were selected as they are stable, reliable and provided maximum uptime **[7].**

Database Selection

NCBI database website **[8]** was used to download databases. NCBI website supports FTP protocol and all the databases such as PubMed, Nucleotide, EST, Protein, Structure, SNPs, conserved Protein Databases, etc are available in FTP servers.

Selection of Application for database downloads

Multiple applications for FTP and torrent protocols are available. Filezilla **[9]** and Bitcomet **[10]** were selected as representatives of FTP and BitTorrent procedures, respectively. These programs are among the best clients, having the ability to download data at the same interval of time. Filezilla is a single server based solution that does not support torrent file, becomes slower with increasing number of users and lacks resume facility after internet link failure. Bit Torrent on the other hand is a peer based solution that uses mesh technology and supports resume facility as well as both torrent and FTP files.

Performance Evaluation

Most of the bioinformatics databases are usually uploaded on FTP server. Downloading of database was performed using both applications and was monitored for the span of fifteen hours. In order to make sure that the load on the network should be same for both of the methods during the test period; the whole procedure was carried out on different machines with similar specifications, and same network. Both the btP2P and FTP performances were evaluated.

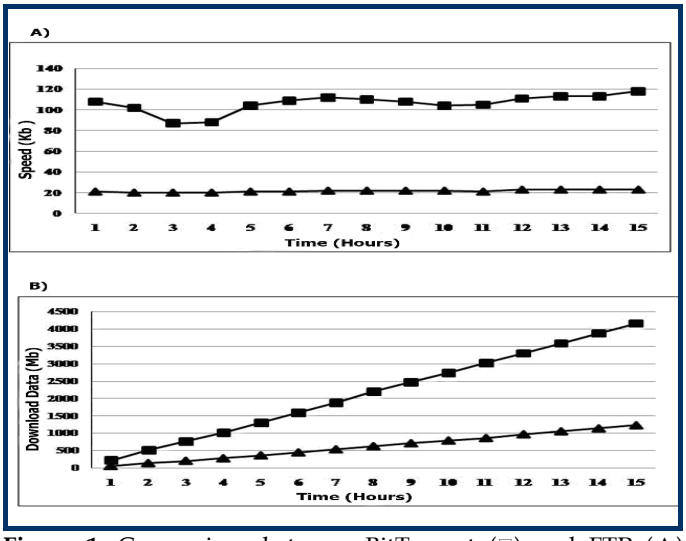

Figure 1: Comparison between BitTorrent (□) and FTP (▲) protocol indicating speed **(A)** and downloaded data **(B)** in 15 hours

Discussion:

Databases are usually downloaded using client-server architecture like FTP. If server becomes overloaded, response time might increase. P2P file sharing protocols have gained popularity as an alternative procedure to FTP. In current communication, both the applications Filezilla and Bitcomet were compared as representatives of FTP and Bit Torrent (P2P) protocols. Our results indicate that BitTorrent protocol is more efficient in downloading large data (GB) in less time period **Table 1** (**see supplementary material**). In first hour, downloading speed of Bitcomet was 87 KB with 234 Mb while downloading speed using Filezilla was 21 KB with 66 MB. In successive hours, torrent downloading speed kept on improving than that of FTP and by the end of fifteenth hour, torrent downloading speed was at least four times higher than the FTP downloading speed.

The speed comparison of Torrent and FTP protocols with respect to time is shown in **(Figure 1A)**. The results show the slow speed of FTP as compared to the torrent speed. **Figure 1B** represents the downloaded data comparison between the two protocols in specified time limit. This further demonstrates that the torrent is more reliable as compare to FTP protocol. In recent years, a significant part of internet bandwidth is being used by P2P traffic. BitTorrent is a popular P2P application that aims to avoid bottleneck of FTP servers while delivering large and popular files **[11].** An earlier communication has clearly shown the better performance of btP2P protocol than traditional FTP for automatically synchronizing large amounts of biological databases across the three countries of Asia-Pacific region **[12].** However, they have compared FTP and P2P file sharing applications using Azureus as a Bit Torrent representative. For current study, Bitcomet was selected, which is a client written in C++. Bitcomet can run in windows environment and offers a preview download mode so that users can preview download content although the file has not been completely finished. It allows users to create their own torrents and can be used for HTTP/FTP download, a format usually used for most of the bioinformatics database download. The results obtained from our study demonstrate that BtP2P techniques can be applied to scale database servers and can outperform client-server based applications. With two available nodes, it is concurred that the performance using btP2P is better than that of FTP. The results of our study showed significant improvement in download performance using btP2P than conventional File Transfer Protocol (FTP). Our study has exhibited the reliability of btP2P in the transmission of continuously growing multi-gigabyte biological databases without failure. Furthermore, the download performance for btP2P can be further intensified by including more nodes from various parts within the country. This study suggests that the btP2P technology is highly appropriate for file sharing applications as this is effective, viable and self scalable.

Conclusion:

Based on above mentioned observations, it can be concluded that the Torrent protocol is almost four times faster than FTP protocol. Hence torrent protocol is recommended as a better tool for updating and synchronization of the biological data sets using low bandwidth. Results obtained from this study support the findings of Sangket *et al.* **[12]** who compared the downloading performance between FTP and btP2P on different

 © 2012 Biomedical Informatics

subnets among developing countries. Most of the databases use FTP protocol and as Bitcomet client supports both FTP and torrent, it may offer a better choice. The download performance for btP2P can be improved further by including more nodes from other institutes and Research and Development (R&D) organizations. It is suggested that btP2P technology may be an appropriate application for file sharing, automatic synchronization and distribution of biological databases and software over low-bandwidth networks.

Acknowledgments:
Authors are grateful to the Higher Education Commission, Pakistan for the financial support for this work (grant no: 20-752).

References:

[1] Ilyas M *et al. PLoS Comput Biol*. 2011 **7**: e1001135 [PMID: 21750669]
[2] Ranganathan S *et al. Appl Bioinformatics*. 2002 **1**: 101 [PMID: 15130849]
[3] Ranganathan S *et al. BMC Bioinformatics*. 2008 **9**: SI [PMID: 18315840]
[4] Gilbert D *et al. Bioinformatics***.** 2004 **20:** 3238 [PMID: 15059839]
[5] Pouwelse J *et al. Peer-to-Peer Systems1V*. 2005 **3640**: 205
[6] Guo L *et al. IEEE J Selected Areas Commun*. 2005 **25**: 155
[7] Garud R & Kumaraswamy A, *Strategic Management Journal*. 2006 **14:** 351
[8] http://www.ncbi.nlm.nih.gov
[9] http://filezilla-project.org/
[10] www.bitcomet.com/
[11] Wei *et al. Future Generation Computer systems* 2007 **23**: 983
[12] Sangket U *et al. Bioinformatics*. 2008 **24**: 299 [PMID: 18037613]

Edited by P Kangueane
Citation: **Azam** & **Zarina,** Bioinformation 8(5): 239-242 (2012)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

BIOINFORMATION

open access

Supplementary material:

Table 1: Time and speed comparison between BitTorrent and FTP software

Bit Torrent Software				FTP Software			
S. No.	Torrent Download Speed (KB)	Torrent Downloaded Data (Mb)	Time in Hours	S. No.	FTP Download Speed (KB)	FTP Downloaded Data (Mb)	Time in Hours
1	87	234	1	1	21	66	1
2	82	522	2	2	20	150	2
3	67	774	3	3	20	216	3
4	68	1020	4	4	20	288	4
5	83	1302	5	5	21	372	5
6	88	1596	6	6	21	456	6
7	90	1884	7	7	22	546	7
8	88	2202	8	8	22	636	8
9	86	2466	9	9	22	720	9
10	82	2742	10	10	22	798	10
11	84	3030	11	11	21	870	11
12	88	3300	12	12	23	978	12
13	90	3588	13	13	23	1062	13
14	90	3876	14	14	23	1146	14
15	95	4164	15	15	23	1236	15

© 2012 Biomedical Informatics