B. Project Summary

Shared genome database informatics tools and cyberinfrastructure

Don Gilbert

Intellectual merit: Bioscientists are assembling the genomes of thousands of organisms, comprising millions of genes, proteins and related bio-information, now part of biology's digital library distributed on computers around the world. Coping with this "glut of genomes" requires cost-effective, shared database tools, such as those built under the umbrella of the collaborative Generic Model Organism Database (GMOD) project. Software tools to fully assembly, analyze and compare these genomes are available, but the ability to employ them is limited to those with extensive computational resources and engineering talent. In this project, methods are being developed for use by existing and emerging model organism databases that will address genome database access needs and middleware for comparative analyses. Effective use of shared cyberinfrastructure, such as NSF-sponsored TeraGrid and other Grid systems, is a problem today for bioinformatics. This proposal addresses this with data grid methods that partition large genome database sets for effective use of Grid systems.

BIOSCI/Bionet, a network of biology news and discussion, is approaching its third decade of service to the life sciences community. This project will maintain and improve Bionet news for several model organism communities, and areas of molecular methods, bio-software, microbiology, proteins, plant biology and others. The Bio-mirror project for distribution of large genomic and biology databanks is administered under this project, and provides infrastructure for effective genome database analyses.

Broader impacts: This project improves access to life sciences knowledge for all citizens. It aids the education of a broad spectrum of scientists, students and the general public, consistent with national and industry goals of teaching and research in biology and informatics. The project is a component of the infrastructure of bioinformatics in the US and worldwide, and improves partnerships with Asian-Pacific, African, European and western hemisphere nations. BIOSCI/Bionet news, IUBio Archive and Bio-Mirror projects will continue a multi-decade record of serving the world community with public access to biology information.

genome informatics for 2006-2007. IU Genome Informatics Lab resources include two Terabyte storage dedicated to public genomics and bioinformatics datasets, including the world-collaborative Bio-mirror.net, and six dual-CPU servers with Grid tools for development and testing. Equipment additions are requested in this proposal as data grid sources, Bio-mirror and genome database hosting.

3.6 Bio-Mirror distribution of Biology data.

The Bio-Mirror project (Gilbert *et al.* 2004) is a worldwide bioinformatics public service for high-speed distribution and access to up to date DNA and protein biological sequence databanks. In genome research, public data sets are growing rapidly. Timely distribution of these is limited by current Internet transport routes. The Bio-Mirror project is devoted to facilitate timely access to important large data sets for this research, and has been operating since 1998. This project of the PI and colleagues has received support from NSF (award 0090782, see below) and other country agencies.

Data in the Bio-Mirrors project currently exceeds a Terabyte (December 2005, compressed for transport) and is updated daily from primary sources, where roughly 33% of data updates each month. Contents include DNA and protein biosequence and related databanks, metabolic paths, genome and model organism data. These are mirrored from originating sites in the US, Europe, Asia-Pacific and elsewhere. Nodes of the project include bioinformatics centers in Japan, Australia, Austria, Singapore, China, Korea, the USA and others. Indiana University high performance network infrastructure and collaborative help has been essential to this project, including Internet2, Trans-Pacific network, and Asia-Pacific Advanced Network (APAN) connections. This project also includes collaborations with EBI EMBL database services, and NCBI GenBank and other databases for production distribution.

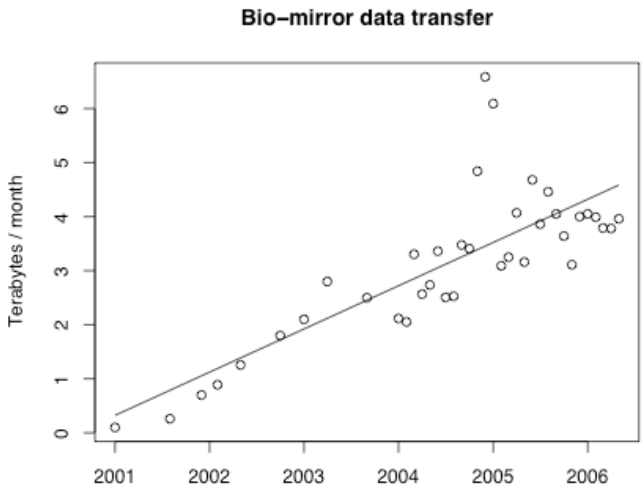

Figure 4. US Bio-mirror data transfers since year 2001, as Terabytes/month

This US Bio-mirror regularly serves bioinformatics centers in North, South, Central America, Asia, Pacific, Africa and Europe. Due to its extensive science network connectivity, the Indiana University node is 3x faster delivering data to TeraGrid centers at PSC, Pennsylvania, SDSC, California and NCSA, Illinois than fetching the same data from NCBI, Maryland. This makes it a good choice for bio-data distribution. This proposal requests 10% of its budget for administration efforts (new databanks, resolving problems), and equipment enhancements (disk storage, improved network connectivity hardware). Development plans discussed above for genome data grids will draw on Bio-mirror databanks.

3.7. BIOSCI/Bionet administration.

BIOSCI/Bionet is a set of electronic communication forums, the Bionet Usenet newsgroups and parallel e-mail lists, begun in the mid 1980s and used by biological scientists worldwide (Gilbert 2004). No fees are charged for the service. BIOSCI/Bionet reflects and facilitates the science community's

passion for public discourse to arrive at truths in an open manner that all can see and review. The Bionet archives provide important references to biology discussions and knowledge to many scientists and lay people. Bionet provides news for several model organism communities (*Annelida*, *Arabidopsis*, *C. elegans*, *Chlamydomonas*, *Maize*, yeast, fruitfly, zebrafish, and others) as well as general areas (molecular methods, bio-software, microbiology, cell biology, proteins, neuro-sciences, toxicology, *etc*.). Several scientists volunteer their time to moderate the messages for their communities. Figure 5 shows usage of Bionet since 1990, including average postings and spam per group, e-mail subscriptions, and Google searches.

Overall, these usage figures show a high rate of postings in the 1990s, with a sharp decline in 2000s coinciding with a high rate of spam. Spam is not the only cause of the use decline, but it is a significant one. Since 2005, there has been a large increase in usage noted by subscriptions, and Google searches, as well as on-topic postings and a cut-off of spam. Bionet moved to IUBio Archive, Indiana University in 2005, with the closing of MRC Rosalind Franklin Centre at Hinxton, UK, after its start at Stanford as part of the GenBank contract in the 1980s. IUBio Archive website serves 1.7 million requests/month excluding robots, with 75% of this as Bionet activity, ranking it among the most active science websites at Indiana University. Basic management needs include spam removal, news list facilitation, mail /subscription management, news archiving, and Usenet management, to be met with 20% of the funds requested in this proposal. This effort includes daily administration activities, handling requests from the public and moderators, weekly and monthly corrections for Usenet, mail and archive management.

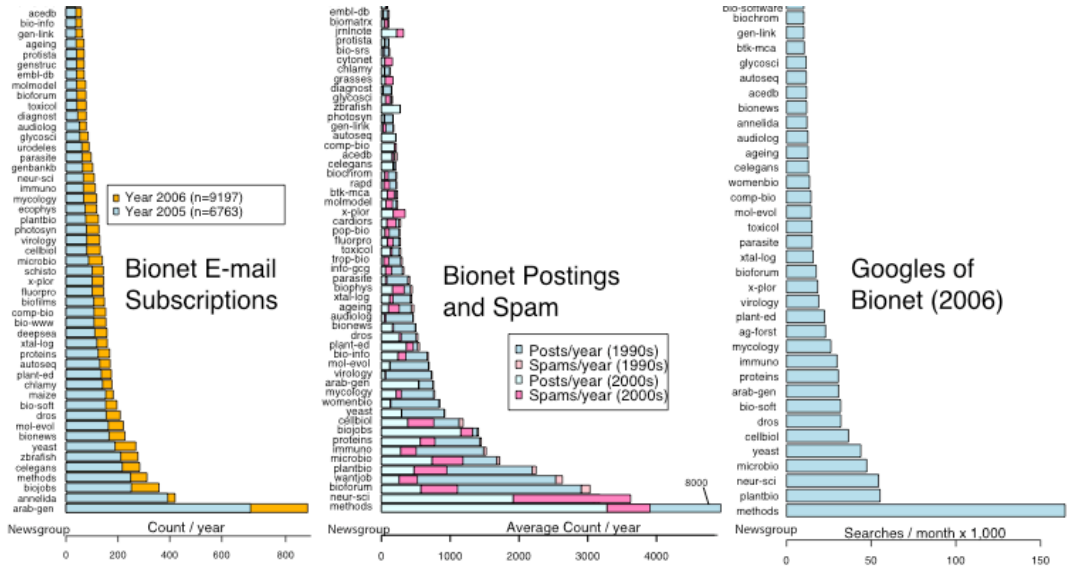

Figure 5. Bionet activity since 1990 for e-mail subscriptions (including 2005-2006 increase), message counts, spam counts, and Google searches (2006 only).

Of the roughly 50 active Bionet groups, Arabidopsis news has most e-mail subscriptions (1,000 of 10,000 total), while Molecular-methods is the most active with 4,000 posts/year, mostly via Usenet. This recent posting from the moderator of protein-crystallography news encapsulates the state of groups and some of the improvements in Bionet.

> From: Cathy Lawson <cathy.lawson@rutgers.edu>
> Newsgroups: bionet.xtallography
> Subject: xtal-log activity
> Date: Sat, 01 Jul 2006 15:10:01 -0400
> Archive: http://www.bio.net/bionet/mm/xtal-log/2006-July/004984.html

> BTW, it seems that the ccp4bb is much more active on discussing protein
>crystallography. I wonder how much this xtal-log forum is active !

Dear Jorge,

The xtal-log bulletin board has not so active in the past few years. It is easy to see this by looking at the storage size for each month's correspondence on the xtal-log archive page: http://www.bio.net/bionet/mm/xtal-log/. When I took on the job of moderating in 2001 there had already been a significant reduction in use, in part because of spam problems, but also because other crystallography bulletin boards (e.g., for ccp4, cns, o) welcome "off topic" posts.

However, there is good news:

The xtal-log bb continues to actively used by a number of scientists to post job opportunities, conference and software announcements, currently reaching more than 150 e-mail subscribers. I would therefore say that it is still a very appropriate place to ask questions about protein crystallography. To reach a larger set of experts, I would suggest cross-posting with one of the other bbs.

Moderation has become super easy. The bionet bb's were changed over from MRC to the group at Indiana in June 2005, and better moderation software was implemented. They must have a terrific spam filter because nearly all the messages that come to my attention are list-appropriate. Also, I always promote good posters to "automatically accept" status.

So ... bring on the crystallography questions!

Spam reduction and more rapid of distribution of news, along with improved news/mail management methods, RSS news feeds, enabling of Google searching with a high page rank that makes Bionet posts readily visible, are result of unfunded efforts of the PI during 2005-2006 year. Recently Indiana U. retired the news.indiana.edu Usenet server, and colleagues at Purdue stepped in to provide news.purdue.edu for Bionet management. There are numerous options for bioscience discussion lists today, including institution/project mail lists, Google Groups, wiki and blog methods. Most of these media lack the broad public distribution and “findability” of Bionet, and have centralized points of failure. Usenet, as a distributed news system, will continue Bionet if administration at net.bio.net fails.

4. Results from Prior NSF Support

Prior NSF support to the PI has been through NSF-DBI award 0090782 of $250,000 for 2/1/2001 to 01/31/2004. This project, “IUBio Archive: Access and Distribution of Genomic and Molecular Bio-information”, serves public biology software and data via FTP, Web/HTTP, Rsync and other Internet protocols, at Internet addresses iubio.bio.indiana.edu, bio-mirror.net and eugenes.org. Papers published from this ongoing project include Gilbert (2002, 2003), Gilbert *et al.* (2004). This funding has been instrumental in training two bioinformatics graduate students, and many other students at Indiana University benefit from lectures and discussion by the PI on subjects supported by NSF through this project. NSF-sponsored use of TeraGrid resources has been provided to the PI in 2005 as a development allocation BIR050001 for “Development of GMOD Genome Database Community Grid Resources”.