

# **World-Wide High Performance Network Mirroring of Public Genome Data**

## **Proposal to High Performance Network Applications Program (HPNAP) of Indiana University**

June 1999

Principle investigator:  
Don Gilbert

[gilbertd@bio.indiana.edu](mailto:gilbertd@bio.indiana.edu)  
Associate Scientist, Bioinformatics  
Biology Department, IU Bloomington

### **Executive summary**

A central focus of new biology science and technology is the large and rapidly growing body of knowledge of the gene code fundamental to all life. The ongoing Human Genome Project and related international efforts are on track with a full unraveling of this code in the coming decade, and this is resulting in a flood of important data to biosciences researchers. The Internet is the primary means for distributing this public data. As primary genome data is growing to tens of Gigabytes, and rapid increase in use of the Internet has introduced delays, the ability to disseminate this data in timely manner has suffered. A need exists for use of high performance network methods to speed the distribution of this data.

This proposal to support a Bio-Mirror project, centered at Indiana University Biology Department, for high speed, world-wide distribution of gene code data will help address this need. Indiana University is well positioned with its High Performance Network initiatives to address needs for rapid distribution of these data. The principle investigator has established a ten year record of providing public access to bioinformatics data, through the well known IUBio Archive. Based on these strengths of Indiana University for bioinformatics network services, a Bio-Mirror project has been initiated to address rapid exchange of biosequence data.

This proposal requests equipment support of \$19,900 and collaboration with Indiana University High Performance Network initiative members (including TransPAC and Internet2/Abilene).

## **Application description and expected outcomes**

The Bio-Mirror project is a world-wide bioinformatics public service for high-speed access to up-to-date DNA and protein biological sequence databanks. In genome research, these databanks have been growing tremendously, so much that distribution of them is hampered by existing Internet speeds. The Bio-Mirror project is devoted to facilitate timely access to important large data sets for this research. High speed access is provided by Internet2 infrastructure of the Very High Speed Backbone Service (vBNS), Abilene, TransPAC, and the Asia-Pacific Advanced Network (APAN).

The original Bio-Mirrors proposal was developed by Japanese partner Y. Ugawa [1], following discussions with D. Gilbert about ways to address the needs of timely international distribution of biosequence data. Subsequently A. Mizushima (Japan) and Tan Tin Wee (Singapore) joined this group, bringing the support and approval of the Asian Pacific Bioinformatics working group (APBionet).

Indiana University's developing high performance network infrastructure and collaborative help from Rick McMullen, Doug Pearson (TransPAC), and others at IU enabled this project to start initial service in December 1998. Project documents, data sets and services are available at the Indiana University Biology department server, <http://www.bio-mirror.net/>, and at international partner sites (see below).

## **Importance to biological and related sciences**

The end of the 20th century has seen an explosion of information discovered from living organisms, especially in areas of molecular biology and genetics. The focus of bioinformatics is dealing with this flood of information, coming from academic, industry and government labs, and turning it into useful knowledge. At the center of this information flood is the rapidly unraveling genetic code of life, and new research directions in understanding this code, how life develops from it, and how it can be engineered to advance health and well-being of humans in many aspects.

The flow of information from the Human Genome Project and related world-wide efforts has revolutionized many fields of biosciences. This knowledge of the human genome code will revolutionize medical practice and biological research in the coming century, including understanding of most inherited diseases. The genome code coupled with new understanding of its organization, regulation and function in cells and in development of organisms will form the basis for designing new treatments for many diseases, and for understanding and modulating health problems with aging. Genome information is quickly becoming the basis for designing most new pharmacology and drugs. It is also central to the improvement of genomes of economically important crops and animals.

Detailed analysis and investigation of genome data has not been done, as the data are accumulating too rapidly to fully implement such analyses without the most powerful of computing methods and equipment. In fact this goal is the focus of several of the largest companies in the area of pharmaceuticals, medical, agricultural bioengineering and biotechnology. Celera, a new company focusing on the extraction and sale of knowledge from genome data is building a bioinformatics center for this that includes massively parallel computing rated at 1.3 teraflops, rivaled only by classified US government centers [2].

Current Internet speeds and delays are a drawback to timely sharing of genome data. This genome data is published electronically from world bioinformatics centers, including the U.S. National Center for Biotechnology Information (NCBI), home to the GenBank [3] DNA sequence databank. From primary collection centers, this data is distributed world-wide to researchers via the Internet. It is accessible to

researchers in many ways, including Web searches for small portions of the data, Internet file transfers for bulk exchanges, and even slower CD-ROM distribution. Timely up-to-date access to this daily changing data is of importance to the continued rapid advance of biosciences research and discoveries, as it forms the basis for integrating the common genome codes used by all life, so new discoveries pertaining to one section of the gene code carry over to discoveries in many other sections of the code, understanding of biological pathways and gene functions, and of different organisms.

## **Importance to Indiana University**

This project will build on University expertise and standing in the growing, important field of bioinformatics. Over the last few years, US biomedical research and funding has started to focus much more on biocomputing and bioinformatics as a nationally important field [4]. IU Biology department has a ten year presence as a center for distribution of bioinformatics data and software, at IUBio Archive (<http://iubio.bio.indiana.edu>).

Funding such a project will provide "seed" infrastructure, melded with IU's excellent high performance networking, to better establish IU as an important center for bioinformatics data distribution. With such infrastructure, national agencies (NHGRI[5], NSF, USDA, DOE) will see this university as worthy of additional funding and project support. Such institutional commitments to science centers can be critical in current national funding agency decisions for support. Additional funding for this project will be sought from the NSF Database and Informatics program, though any award will be more than a year away, necessitating curtailment of ongoing services.

This will be a research area associated with the developing School of Informatics. It may well be a focus point to attract students and further national and industry funding for this School. Indiana companies Eli Lilly, Dow Agrochemicals and others have growing needs for bioinformatics knowledge and reliance on discoveries from genome data.

## **Project background**

These servers are publicly available sites for high-speed access to up-to-date DNA/protein biological sequence databanks. High speed access between the sites is provided by the network infrastructure developed by Very High Speed Backbone Service (vBNS), TransPAC (Trans-Pacific network), and Asia-Pacific Advanced Network (APAN), and these sites are well connected to national research and education networks within each country.

DNA and protein biosequence databanks are essential for advanced studies in genome research. These sequence data have been collaboratively collected and prepared at bioinformatics centers in the US, Japan and Europe since 1984. Software search methods such as FASTA and BLAST are used commonly to find sequence homologies, and SRS for content searching. These require large local disk space to store the databases and analyze data.

These databanks are growing rapidly [3], doubling every 14 months since 1982, with the rate of increase accelerating as new genome sequencing projects are undertaken and new technology for this improves the output of high throughput genome sequencing centers for Human genome and other organism sequencing efforts.

Some mirror servers have been developed to provide most recently updated data for DNA/protein biological sequence database. In present, however, these servers often fail in timely updates due to the lack of existing network band width. Bi-directional and multi-directional mirror services of this genome data, taking advantage of new Internet2 infrastructure and new methods, will help to distribute this data

collected including in many countries and deliver them to researchers. This mirror service could not be established without high-speed and reliable Internet2 connections.

## **Current data sets**

Data in the Bio-Mirror project currently totals about 10 Gigabytes in compressed format, and are updated from the primary sources nightly. DNA biosequence data include GenBank, EMBL, DDBJ. Protein biosequence data include SWISS-PROT, TrEMBL, PIR. Other data include BLOCKS, ENZYME, PROSITE, REBASE. These are mirrored from originating sites for these data banks, in the US, UK, Switzerland, and Japan.

Additional large data sets of importance to this project include the protein structure databank (PDB), and several whole genome and genome sequence integration data sets. These data sets will be added when disk space becomes available.

The Indiana University Bio-Mirror offshoot of IUBio Archive, as well as several of the other Bio-Mirror project sites, provides search and analysis services on uncompressed sets of these data, requiring some 20 Gigabytes of disk space for data, indices and updates, as well as CPU processing power for the analyses.

## **Institutions currently participating in the project**

Agriculture, Forestry and Fisheries Research Council (AFFRC), Japan

File transfer: <ftp://bio-mirror.jp.apan.net/pub/biomirror/> Web: <http://bio-mirror.jp.apan.net/>

Contact: Akira Mizushima [goddila@maffin.ad.jp](mailto:goddila@maffin.ad.jp)

Indiana University, Department of Biology, USA

File transfer: <ftp://iubio:iubio@bio-mirror.net/> Web: <http://www.bio-mirror.net/>

Contact: Don Gilbert [gilbertd@bio.indiana.edu](mailto:gilbertd@bio.indiana.edu)

Advanced Computational Systems (ACSys) CRC, Australia

File transfer: <ftp://bio-mirror.au.apan.net/biomirrors/> Web: <http://bio-mirror.au.apan.net/>

Contact: Markus Buchhorn [markus@acsys.anu.edu.au](mailto:markus@acsys.anu.edu.au)

Bioinformatics Centre (BIC), National University of Singapore, Singapore

File transfer: <ftp://bio-mirror.sg.apan.net/biomirrors/> Web: <http://bio-mirror.sg.apan.net/>

Contact: Mark De Silva [mark@bic.nus.edu.sg](mailto:mark@bic.nus.edu.sg)

The Institute of Microbiology, Chinese Academy of Sciences (IMCAS), China

File transfer: <ftp://bio-mirror.cn.apan.net/> Web: <http://bio-mirror.cn.apan.net/>

Contact: Juncai MA [ma@sun.im.ac.cn](mailto:ma@sun.im.ac.cn)

The addition of a Bio-Mirror site in Korea is expected soon. Bioinformatics centers in other countries, including Canada and Europe, have expressed interest in the project. The support of Asia Pacific Bioinformatics Network APBioNet (<http://www.apbionet.org/>) through the APBioNet-APAN advanced networking project, with the Agriculture Working Group and Bioinformatics Working Group of APAN, has been instrumental in aiding this project.

## Internet2 infrastructure organizations

The following organizations for Internet2 infrastructure are employed by the Bio-Mirror project. Expertise and information from these organizations will be important to the most efficient and rapid exchange of Bio-Mirror data.

Abilene - <http://www.internet2.edu/abilene/>

Asia-Pacific Advanced Network (APAN) - <http://www.apan.net/>

Trans-Pacific network, TransPAC - <http://www.transpac.org/>

Very High Speed Backbone Service (vBNS) - <http://www.vbns.net/>

Singapore Internet Next Generation Advanced Research and Education Network (SINGAREN) - <http://www.singaren.net.sg/>

Agriculture, Forestry and Fisheries Research Council (AFFRC), Japan and Advanced Computational Systems (ACSys) CRC, Australia, are well connected to APAN, sharing data through APAN between the sites, and through TransPAC with Indiana University, and through SINGAREN with National University of Singapore. SINGAREN is connected to STARTAP vBNS, to CA\*Net2 and to Tokyo APAN.

Indiana University is well connected to high-speed initiatives of Internet2, as a member in vBNS, and center of Abilene network operations and the US TransPAC connections to APAN. Abilene is an advanced research and education network in the United States. See <http://www.iuinfo.indiana.edu/ocm/-releases/Abilene2.htm>

## Current IU Biology Bio-Mirror support

The IUBio Archive and Bio-Mirror service relies now on equipment purchased for other purposes. These include:

- Sun Microsystems E3000 server and 24 Gigabyte RAID array, purchased for and used by the FlyBase genome informatics project (<http://flybase.bio.indiana.edu>). Currently the Bio-Mirror and other public bioinformatics services piggy-back on this project equipment, but are now interfering with FlyBase project needs.
- one 18 Gigabyte disk purchased by IU Biology department for this project. This has been filled to capacity, necessitating removal of data and delay in plans to add new bioinformatic data sets.
- a 100baseT connection through an ethernet switch to the IU campus backbone network.

As well, Indiana University's developing high performance network infrastructure and collaborative help from Rick McMullen, Doug Pearson (TransPAC), and others, have been essential to this project. The P.I. has and will devote needed time to maintaining this hardware, network connectivity, and biology data and services for the Bio-Mirror project.

## Future directions

Potential additions for the Bio-Mirror project may include new biosequence data such as the annotated rice genome sequence, estimated at 450 megabases, for which several international partners have planned to accomplish under the Rice Genome Project (RGP) [6]. This project biosequence data would find a useful partner in the high-speed data mirroring of the Bio-Mirror project, as many of the countries participating in rice sequencing also have partners in the Bio-Mirror project through APAN/APBionet centers for bioinformatics. Japan is the leading country for RGP, with Korea, China, Singapore, India, USA, Canada, and European Union involvement expected.

Publishing this data electronically in a timely manner, in a way which all participating countries can have up-to-date access to the data is important. Indiana University could become an important partner in this project by offering a US hub for high-speed daily transfer of this data to other Bio-Mirror sites, supporting the international partners in this Rice Genome project. It can be expected that the full public data generated and released by this project will be several gigabytes, with significant daily or weekly changes as new sequence is finished and released in three stages set by international data release agreement. The US effort for rice sequencing is funded by USDA, NSF and DOE agencies, in partnership with international groups, and is expected to start the fall of 1999.

A new bioinformatic project in development at IU Biology department is the Model Eukaryotic Organism Genome Information Database, which has a preliminary service at <http://iubio.bio.indiana.edu/meow/>. This project may also provide a highly important resource to the US and world biosciences community [7]. It is designed to be portable and mirrored to other sites around the world, and would make a good addition to the Bio-Mirror high-speed distribution framework.

The Internet2 Distributed Storage Infrastructure (I2-DSI) project may offer important methods for the Bio-Mirrors project. Bio-Mirror members have agreed to continue for now with tested methods of FTP mirroring for this production service, but will investigate potential of DSI for improving network throughput and distribution of these data. As Indiana University is a partner in the I2-DSI project, it may be possible to tap knowledge of this for Bio-Mirrors.

The IUBio biology data and software archive has been an ongoing project of this IU department for this decade. It's international repute and wide use as a public bioinformatic resource formed the starting point for the Bio-Mirror project.

## **Budget summary and justification**

The funding request is for hardware to support the large biosequence data sets, in disk storage and computer processing power to support public use of this data.

Sun Microsystems A1000 RAID array with 8 x 18GB disks (144GB total)	\$13,995
Upgrade existing Sun workstation to Sun Ultra 10 Model 360 Workstation	\$4,655
512-Mbyte memory upgrade (2 x 256MB DIMMs)	\$1,195
<hr/>	
Total request	\$19,845

Current equipment is Sun Microsystems based, and upgrading that will be more cost-effective than selections from other vendors.

A 144 Gigabyte RAID array is requested to hold the primary data of the Bio-Mirror project, providing room for its growth over the next few years, as well as room for public search and analysis services using these data. This equipment should also allow for adding important new data sets to the project, such as from the rice genome sequencing project, or the model organism database project.

A Sun Ultra 10 will be used to provide public biosequence search and analysis services in conjunction with the Bio-Mirror project. The currently used equipment for this is owned by another project, for which Bio-Mirror services are becoming a burden and have had to be reduced and in part eliminated to protect the needs of the owner project. A memory upgrade to this server is important for rapid indexing and searching of this multi-gigabyte data set.

## References

- [1] Y Ugawa, 1997. Development of Mirror Server by using High Speed Data Transfer in Genome Science. Proposal to Asian-Pacific Advanced Network (APAN) organization, <http://www.jp.apan.net/HPIIS-Applications/JP-AFFRC>
- [2] E. Marshall, 1999. A High-Stakes Gamble on Genome Sequencing. **Science**, 284, p. 1906-1908, 18 June 1999.
- [3] DA Benson, et al, 1999. GenBank. **Nucleic Acids Res** 1999 Jan 1;27(1):12-7 See also <http://www.ncbi.nlm.nih.gov/>
- [4] D. Malakoff, 1999. Biocomputing: NIH Urged to Fund Centers to Merge Computing and Biology. **Science**, 284, p. 1742, 11 June 1999.
- [5] National Human Genome Research Institute (NHGRI) <http://www.nhgri.nih.gov/>, and Genome Informatics, [http://www.nhgri.nih.gov/About\\_NHGRI/Der/ginform.htm](http://www.nhgri.nih.gov/About_NHGRI/Der/ginform.htm)
- [6] B. Burr, 1997. An International Collaboration To Sequence The Rice Genome, <ftp://genome1.bio.bnl.gov/pub/maize/rice.html>
- [7] Model Eukaryotic Organism Workshop, December 1998. Workshop report. <http://www.nhlbi.nih.gov/nhlbi/sciinf/modeldb/model.htm>

## **Appendix**

### **Bio-Mirror Project milestones**

- June 1997 -- initial discussions between Y Ugawa and D Gilbert about Internet 2 uses for biosequence data between USA and Japan.
- August 1997 -- "Development of Mirror Server by using High Speed Data Transfer in Genome Science" proposed by Y Ugawa to APAN organization. See <http://www.jp.apan.net/HPIIS-Applications/JP-AFFRC>
- February 1998 -- APAN Singapore link with APAN Japan established.
- April 1998 -- an expanded Bio-Mirrors project, by A Mizushima and Y Ugawa, approved by APBioNet.
- May 1998 -- APAN resource allocation secretariat approved the resource allocation which A Mizushima and Tan Tin Wee applied for through APBioNet.
- June 1998 -- Bio-Mirrors project approved by APAN.
- September 1998 -- Transpac link to APAN Japan established.
- December 1998 -- Initial mirroring between bio-mirror.jp.apan.net and bio-mirror.us.apan.net, including GenBank, EMBL, and protein data.
- January 1999 -- Australian site bio-mirror.au.apan.net joins; DDBJ databank added.
- March 1999 -- TransPAC, the international connection between the vBNS and APAN is operational at 70 Mbps.
- April 1999 -- Singapore site bio-mirror.sg.apan.net joins.
- May 1999 -- China site bio-mirror.cn.apan.net joins.