

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

| PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 03-041 | | | | | FOR NSF USE ONLY | |
|---|------------------|---|---|---|----------------------------|--|
| NSF 02-058 | | | 07/14/03 | | NSF PROPOSAL NUMBER | |
| FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.) | | | | | 0345285 | |
| DBI - Database Activities | | | | | | |
| DATE RECEIVED | NUMBER OF COPIES | DIVISION ASSIGNED | FUND CODE | DUNS# (Data Universal Numbering System) | FILE LOCATION | |
| | | | | 006046700 | | |
| EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN) | | SHOW PREVIOUS AWARD NO. IF THIS IS <input checked="" type="checkbox"/> A RENEWAL <input type="checkbox"/> AN ACCOMPLISHMENT-BASED RENEWAL | | IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES <input type="checkbox"/> NO <input checked="" type="checkbox"/> IF YES, LIST ACRONYM(S) | | |
| 356001673 | | 0090782 | | | | |
| NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE | | | ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE | | | |
| Indiana University | | | Indiana University | | | |
| AWARDEE ORGANIZATION CODE (IF KNOWN) | | | P O Box 1847 | | | |
| 0018093000 | | | Bloomington, IN. 474021847 | | | |
| NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE | | | ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE | | | |
| PERFORMING ORGANIZATION CODE (IF KNOWN) | | | | | | |
| IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions) | | | | | | |
| | | <input type="checkbox"/> SMALL BUSINESS | | <input type="checkbox"/> MINORITY BUSINESS | | <input type="checkbox"/> IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE |
| | | <input type="checkbox"/> FOR-PROFIT ORGANIZATION | | <input type="checkbox"/> WOMAN-OWNED BUSINESS | | |
| TITLE OF PROPOSED PROJECT Data access systems for high-volume genomic and molecular bio-information | | | | | | |
| REQUESTED AMOUNT \$ 1,484,520 | | PROPOSED DURATION (1-60 MONTHS) 60 months | | REQUESTED STARTING DATE 01/01/04 | | SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE |
| CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW | | | | | | |
| <input type="checkbox"/> BEGINNING INVESTIGATOR (GPG I.A) | | | <input type="checkbox"/> HUMAN SUBJECTS (GPG II.C.11) | | | |
| <input type="checkbox"/> DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C) | | | Exemption Subsection _____ or IRB App. Date _____ | | | |
| <input type="checkbox"/> PROPRIETARY & PRIVILEGED INFORMATION (GPG I.B, II.C.6) | | | <input type="checkbox"/> INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.9) | | | |
| <input type="checkbox"/> HISTORIC PLACES (GPG II.C.9) | | | | | | |
| <input type="checkbox"/> SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.C.11) | | | <input type="checkbox"/> HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.E.1) | | | |
| <input type="checkbox"/> VERTEBRATE ANIMALS (GPG II.C.11) IACUC App. Date _____ | | | | | | |
| PI/PD DEPARTMENT Department of Biology | | | PI/PD POSTAL ADDRESS 1001 E. 3rd Street | | | |
| PI/PD FAX NUMBER 812-855-6705 | | | Bloomington, IN 47405 | | | |
| United States | | | | | | |
| NAMES (TYPED) | High Degree | Yr of Degree | Telephone Number | Electronic Mail Address | | |
| Donald G Gilbert | phd | 1981 | 812-855-0587 | gilbertd@bio.indiana.edu | | |
| CO-PI/PD | | | | | | |
| CO-PI/PD | | | | | | |
| CO-PI/PD | | | | | | |
| CO-PI/PD | | | | | | |

PROJECT SUMMARY

Intellectual merit: Biologists have discovered many millions of genes, proteins and other genome and molecular bio-information, now part of the biology "library" distributed on computers around the world. These data are complex in structure, large in volume, widely distributed, and rapidly changing. It is increasingly difficult for the scientist to find and computationally use those objects needed to answer research questions. IUBio Archive has provided a public distribution center for biology software and data over the last 15 years, extending the range of biology information that is easily usable by bioscientists, students and the public. This project will build on that record, with improvements and extensions to current collections of genome and molecular biology information. It will investigate emerging data access technology for retrieval and redistribution among national and international bioinformatics centers. Investigations of high-volume distributed search and retrieval methods, at databank and bio-object levels are planned. Use of summary data and metadata to build searchable directories of genomes and molecular objects will aid in the integration and federation of existing primary biology databases. Standards-compliant technologies for distributed data access will include Web Services, the Lightweight Directory Access protocol and Grid Services. Client software that is easy for the emerging group of data-mining biologists to employ in use of distributed databanks and objects will be developed in this proposal.

Specific aims of this project are:

A. Bio-information warehousing, access and distribution. Extend and investigate methods for biosciences data-warehousing, improve data distribution and public access and search services, in collaboration with the Bio-Mirror project.

B. Bioinformatics tools and services. Extend a public molecular biology software archive and web-based services at IUBio Archive, with applications to Bio-Mirror and other bioinformatics centers.

C. Genome information directory. Investigate improved methods for genome data federation, access and search services, in context of the euGenes eukaryote genome information project.

Broader impacts: This project is oriented to improving access to life sciences information for all citizens. This project and the PI help to educate a broad spectrum of citizens and students, consistent with national and industry goals of teaching and research in informatics. This includes continuing work with Indiana University programs in bioinformatics education, and interaction with museums and science education groups such as the American Museum of Natural History, the National Academy of Sciences Museum, and the National Science Teachers Association. This project will enhance the infrastructure of bioinformatics in the US and worldwide, and improve bioscience partnerships with Asian-Pacific, European and other nations. Improving security for biology data access and distribution is a component of this project that has applications for national security and bio-threat response.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.C.

| | Total No. of Pages | Page No.* (Optional)* |
|---|-------------------------------|----------------------------------|
| Cover Sheet for Proposal to the National Science Foundation | | |
| Project Summary (not to exceed 1 page) | 1 | _____ |
| Table of Contents | 1 | _____ |
| Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) (Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee) | 15 | _____ |
| References Cited | 4 | _____ |
| Biographical Sketches (Not to exceed 2 pages each) | 2 | _____ |
| Budget (Plus up to 3 pages of budget justification) | 7 | _____ |
| Current and Pending Support | 1 | _____ |
| Facilities, Equipment and Other Resources | 1 | _____ |
| Special Information/Supplementary Documentation | 5 | _____ |
| Appendix (List below.) (Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee) | _____ | _____ |
| Appendix Items: | | |

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

PROJECT DESCRIPTION

I. Introduction

How does one find, read, analyze and integrate biology information that is available at repositories throughout the world, including genes, proteins, and literature reports? There are at least 400 million such objects at public repositories such as the National Center for Biotechnology Information (NCBI) in USA, European Bioinformatics Institute (EBI) and National Institute of Genetics (NIG) in Japan, and other institutions. Tables 1 and 2 indicate the wealth of these, drawn from the EBI data repository and FlyBase Drosophila genome database in September 2002. Bio-data is collected, curated and computed by many research groups, and is continually being improved and expanded. New kinds and forms of bio-data are produced as new experimental methods replace older ones in life sciences. These data are widely distributed around the world, not just a few central repositories, and are exchanged and published in many ways via Internet access to databases, spreadsheets, structured and unstructured documents and files. These objects are categorized in many ways: by name, database identifiers, classes (DNA, proteins, genome features, literature, gene expression), organism sources, dates discovered, authors, key terms of biological research, and other attributes. Such attributes and metadata provide means for finding objects relevant to answering many biological questions.

Table 1. Biology Databanks at EBI

| Databank | Contents | Entries |
|-----------|-------------------|------------|
| EMBL | DNA Sequences | 18,800,000 |
| SWALL | Protein sequences | 900,000 |
| InterPro+ | Protein motifs | 1,000,000 |
| HGBASE | SNP database | 1,500,000 |
| | Metabolic Paths | 250,000 |
| MEDLINE | Literature | 11,350,000 |
| Total | | 33,800,000 |

Table 2. Drosophila genome objects

| Contents | Entries |
|------------------------|---------|
| Literature References | 140,000 |
| Gene variants | 112,000 |
| Genome features | 50,000 |
| Genes | 40,000 |
| Transgene constructs | 37,000 |
| Chromosome aberrations | 16,000 |
| Fly Stocks | 15,000 |
| Drosophila researchers | 6,600 |
| Total | 416,600 |

Bioinformatics as a discipline, especially its information engineering side, has a strong focus on the rich and growing datasets of life sciences. With genome information in particular, a transition is underway from use of data files to an essential need for integrated database software (Stein 2003). Today biologists seek primary genome information from among many web services. Genomics web portals often lack methods for effectively mining large subsets of genomes, or are limited in the questions one can pose to the underlying complex data. The Ensembl project (Ensembl 2002, 2003) is an example of integrated software and data that bridges the gap in biology data access between bulk files and web portals.

Many data exchange formats are used in bioinformatics, with the majority using a key-value record object structure, such as EMBL and GenBank (Letovsky 1999). Several new exchange formats have been proposed in recent years, e.g. there are at least eight XML formats for biosequences (Gordon 2002), but community consensus on common formats is still a ways away. Collections of biology vocabulary and their relationships are now aiding integration (Covitz 2003) using semantic associations among gene,

protein, literature and related data sets, such as Gene Ontology and Global Open Biological Ontologies (GO 2000).

Developing and updating software to read databank structures and provide integration is a large and ongoing area of activity in bioinformatics. A wide range of bioinformatics tools provide such access: SRS - Sequence Retrieval System (SRS, Etzold 1993), PubMed (National Library of Medicine) and Entrez (NCBI), Ensembl (Ensembl 2002, 2003) and DiscoveryLink (IBM, Haas *et al.* 2001) database middleware layer, among others. An important approach to integration of large-scale heterogeneous bio-data is through knowledge maps (Parker *et al.* 2003) that are based on summary data and metadata describing relationships among core data. Catalogs or directories that integrate bio-data in basic but useful ways can be implemented as interfaces to existing access tools. Projects at DDBJ and EBI bioinformatics centers have pioneered such with Web Services and XML exchange (XEMBL; XDDBJ) by adding an interface layer on top of the centers' current database access tools.

Biology databases are among the most cross-referenced of science data, where many databases actively manage linkages to other databases. The Life Sciences Identifier (LSID) is a developing standard to offer consistent syntax for biology identifiers, and has reference implementations using Web Services and Lightweight Directory Access Protocol (LDAP). The OpenURL standard (NISO 2003) provides a related mechanism for network object access, an example is LinkOut used for biosequences and literature at NCBI bioinformatics center. These and other protocol standards in sciences and informatics are relevant to providing best access to collections of biology information.

Large bio-data sets need to be copied frequently among computers for bioscientists to perform experimental analyses. The task of keeping data sets integrated with others and in synchrony with their sources is a growing problem. The Bio-Mirror project distributes a core of 150 gigabytes of data among several global bioinformatics centers. Several of members of this project lack sufficient network bandwidth, as well as engineering personnel, to update data sets regularly. The problem is compounded for regional and laboratory data collections where bioinformatics engineers and infrastructure are more limiting resources.

Needs for automating the efficient selection, distribution and integration of bio-data are increasingly important in biological and biomedical research. This project will provide solutions to this through a combination of warehousing, redistribution and integrated access to a wide range of genome and biology data and software, using existing and newly developing high-performance search, retrieval and distribution methods; it will investigate basic, effective ways to federate or integrate bio-objects from local and distributed databases, developing directories or knowledge maps for improved access through summary and metadata about objects; and it will focus on eukaryote genome data as a set of widely interesting and complex data objects where improvements in integrated access are important.

Summary of IUBio projects. IUBio Archive (Gilbert 1989) has been providing public dissemination of biology information for nearly 15 years. IUBio Archive includes hundreds of biology software titles, many added during the current grant period, as well as a cataloged, searchable database of software. Along with software titles, 100+ GB of biology databanks is available in integrated searchable form via the SRS system. Web usage is about 75,000 pages/month (no robots), and 10,000 distinct hosts/month.

Bio-Mirror project (bio-mirror.net) has a goal of efficient world distribution of rapidly changing biology data. It is a production distribution system offering FTP, web and rsync access to daily updated, high-volume bioinformatics databanks and exchanges data with centers in Asia-Pacific and throughout the world. Current databank volume is 140 GB, with a 10-20% daily turnover, and transfers at this site are approaching one TB/week.

euGenes project (eugen.es) is a public access system that provides a common summary of gene and genome information from eukaryotic organisms. Summary information on these covers 900,000 genome features with maps and sequences, 200,000 named genes and their functions, with links to extended

information. Usage is about 20,000 pages/month (no robots), and 5,000 distinct hosts/month. A related project not in this proposal is FlyBase (flybase.net), a comprehensive database of *Drosophila* genetics and molecular biology (Table 2), with web usage of about 700,000 pages/month, and 30,000 distinct hosts/month. Recent developments with euGenes include redesign of the infrastructure for easy, rapid replication of the entire information system, in the context of the Generic Model Organism Database project (GMOD; www.gmod.org/argos/) and extensions to provide gene object catalogs as an aid toward federated access to source genome databases.

The aims of this proposal are:

A. Bio-information warehouse, access and distribution. Extend the current biology data and software archive, and investigate improved methods for biosciences data-warehousing or archiving, improve data distribution and public access and search services, in collaboration with the Bio-Mirror project. Extensions include development of object level access through directories or catalogs of bio-information, with emphasis on high-volume, high-availability distributed data access systems. Improvements in reliability, use of replicated sources, security options and efficiency can solve growing needs for biology data distribution.

B. Bioinformatics tools and services. Extend and improve the public molecular biology software archive and web-based services at IUBio Archive, with applications to Bio-Mirror and other bioinformatics centers. Integrate use of SRS, IBM DiscoveryLink and other data selection tools with bioinformatics analysis tools.

C. Genome information directory. Extend and investigate improved methods for genome data federation, access and search services, in context of the euGenes eukaryote organism genome project.

The research proposed focuses on information engineering for effective management of proliferating bio-information. One trend emerging among users of genome databases such as FlyBase is a move among bioscientists and proto-bioinformaticians to data-mining, that is search, retrieval and integration of large subsets of data, often drawn from many database sources, often focused on summary information for a range of common gene attributes. These results are used in spreadsheets and simple databases or analyses. This class of computer-savvy bioscientist is somewhere between a computer-naïve biologist needing strong computational support from data providers, and an experienced bioinformatics group who can 'roll their own' data-mining tools. The proposed aims will aid this growing class of bioscientist to more fully find and use the wealth of available bio-data, with a combined approach of warehousing, advanced data access tools, and database integration through summary information.

In overview the project goals are ambitious, but they are practical and can be achieved in the requested funding period to provide advanced bio-information management. This work will be done in conjunction with needs and efforts of other bioinformatics groups around the world that are investigating improved data access technologies. It will be carried out in collaboration with partners at US, European and Asian Pacific bioinformatics centers, including Rodrigo Lopez and Peter Stoehr (European Bioinformatics Inst., EBI), Ugawa Yoshihiro (Miyagi University, Japan), Markus Buchhorn (Australian National Univ.), Tan Tin Wee (Natl. Univ. Singapore), Rick Westerman (Purdue Univ.) and others. Letters of support are provided in a supplement to this proposal.

II. Results from Prior NSF Support

Current NSF support to the PI is 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, LDAP and other Internet protocols, at Internet addresses iubio.bio.indiana.edu, bio-mirror.net and eugen.es.org. Papers published from this ongoing project include (Gilbert 2002a, 2002b, 2003a, 2003b), with additional ones in preparation. This funding has been instrumental in training two bioinformatics graduate students, Danfeng Yao (M.S. in computer sciences, 2002) and Paul Poole (M.S. in bioinformatics, 2003), and many other students at Indiana University benefit from lectures and seminars with the PI on subjects supported by NSF through this project. This project has formed an important component for bioinformatics training and additional graduate student support is sought in this proposal.

Bulk file distribution from IUBio is rising, from 100 Gigabytes/month in year 2001, to three Terabytes/month in early 2003. Hardware capability includes 600 Gigabytes of storage and a 4 CPU Sun Enterprise server, funded by NSF Grant DBI-9982851 and the Indiana University High Performance Network Project. Currently 500 GB of biology data and software are stored at IUBio, available as part of Bio-Mirror (150 GB), IUBio services (200 GB), euGenes eukaryote genes database (100 GB), and Biology software and Bionet news archives (50 GB). The storage array is at capacity; an equipment fund supplement to the current award has been requested, and this proposal seeks additional server equipment to extend its services.

IUBio Archive

IUBio Archive for biology data and software has been an important Internet information resource for biologists since 1989 (Baxeavanis and Ouellete 1998). The PI's early work includes pioneering efforts to provide interactive searches of biosequence databanks over the Internet, including a graphical Internet search client (Gilbert 1990) and a WAIS server for searching biology data (Gilbert 1992, 1993). Public software available at IUBio is focused on molecular biology, with sections for alignment, sequence consensus, phylogenetics, pattern matching, primer selection, restriction enzymes, RNA structure, searching, and platform-specific programs (Gilbert 1999). This collection of publicly available software is a popular source of analysis tools for bioscientists.

A recent addition to IUBio Archive is the EPrints system (Steele 2002) for self-archiving of publications. Software authors can submit and maintain software packages with this system, lessening the burden on the archive administrator. Modifications were added to permit its use for software cataloging and distribution, and software file access through FTP and Rsync. EPrints includes the standard Open Archives Initiative (Lagoze and Sompel 2001; OAI 2001) interface for harvesting metadata about publications, used in distributed publication directories. These include OpenArchives.org, which catalogs the IUBio software archive on a regular basis. Additional software archiving developments include tools to automatically locate and retrieve new and updated public bioinformatics software packages. This work will help to ensure that IUBio Archive maintains current software with minimal intervention by an administrator.

IUBio Archive has for several years provided a US public site (Gilbert 1995) for the Sequence Retrieval System (SRS, Etzold and Argos, 1993; Zdobnov et al. 2002). SRS is an important and widely used tool for biosequence and related genomic data searches, as it provides a tested and working way of linking, or federating in a basic sense, many related databanks, as well as offering automated up to date access to these data. The SRS service at IUBio has been upgraded to new versions, and extended to add more biosequence data sets during the NSF funding period.

Continuing the service of archiving important public bioscience data and software tools, and expanding its role in bioinformatics data warehousing, is a goal for IUBio Archive. New tools of importance to

molecular biosciences are being added to IUBio user services. These include EMBOSS (Rice 2000), a developing, comprehensive, open-source package of sequence analysis tools; Pasteur Institute Software Environment (PISE; Letondal 2000), a framework for integrating bioscience analysis tools into a web server; and phylogenetics tools. An archive of the Bionet public news articles comprises a large and popular section of this archive. Besides serving the biosciences community, this news archive is a source of science information widely used by the general public. Bionet news contains much useful information on molecular biology materials and methods, software, organism and techniques oriented science news and discussion.

Bio-Mirror project

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 Asian-Pacific bioinformatics centers (APBionet) and IUBio Archive at Indiana University. It distributes databanks using Internet2 connections between continents. Many APBionet members are part of 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.

The Bio-Mirror project is provided as a public service by member centers with support of several organizations. Participants include APBionet and bioinformatics centers in Japan, Australia, Singapore, China, Korea, Malaysia, Taiwan, Thailand and the USA (Table 3). High performance network infrastructure and collaborative help have been essential to this project, including Trans-Pacific network (TransPAC) and Asia-Pacific Advanced Network (APAN).

Table 3. 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/ |
| Malaysia | Universiti Putra Malaysia | http://ingene2.upm.edu.my/ | ftp://ingene2.upm.edu.my/ |
| 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/ |

Contents of the Bio-Mirror databank set include core databanks from the three collaborating DNA databanks GenBank, EMBL and DDBJ. SwissProt, TrEMBL, PIR, PDB, Pfam and InterPro protein databanks are included, as well as BLAST and RefSeq databanks. Ensembl, NCBI Genomes, LocusLink, euGenes, Unigene, Gene Ontology, and related genome and other data are included. The current collection exceeds 140 Gigabytes (compressed, or about 500 GB uncompressed). Approximately 10% of these change daily, and most are updated in the course of 3 months. Complete databank file sets are mirrored from sources in the US, Europe, and Asia-Pacific. The European Bioinformatics Institute (EBI) and National Center for Biotechnology Information (NCBI) centers now include Bio-Mirror sites in their access information.

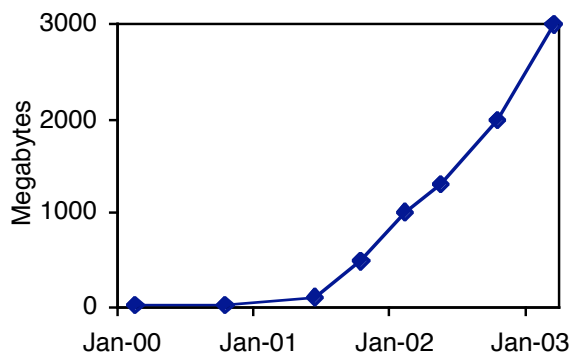


Figure 1. Monthly data transfer at US Bio-mirror.

Project sites currently serve many Terabytes per month to thousands of bioinformatics centers and labs. Bulk distribution has risen from 100 Gigabytes/month in year 2001 to 3 Terabytes/month in 2003 at the US node (Figure 1). File transfer (FTP) provides the best access, as FTP servers have been tuned for large file transfer. Bio-Mirror sites also offer search and analyses services of these data with SRS and other programs. The Perl FTP *mirror* package (McLoughlin 1998) is used at Bio-Mirror sites to maintain daily updates, with additional Perl tools for updating local databanks available in the project collection.

The Bio-Mirror project has been developed with colleagues Y. Ugawa, A. Mizushima (Japan), Tan Tin Wee (Singapore) and others, with the support of the Asian Pacific Bioinformatics group (APBionet). Nodes now include bioinformatics centers in Japan, Australia, Singapore, China, Korea, Thailand, Malaysia, and the USA. Indiana University high performance network infrastructure and collaborative help has been essential to this project, including TransPAC (Trans-Pacific network), and Asia-Pacific Advanced Network (APAN) connections.

***euGenes* genome data access system**

A bioinformatics project started in 1999 by the PI is a genome summary service for eukaryote organisms, *euGenes* (Gilbert 2002a). The genome data summarized in *euGenes* for eight eukaryote organisms, including humans, mouse, *Drosophila*, *C. elegans*, forms a set of attributes on genes that will be useful for developing directory systems to locate gene objects. Primary information on all known and inferred protein coding genes of these genomes is included, with whole chromosome DNA and feature annotations. There are usable whole-genome map displays with feature annotations, chromosome locations and molecular data integration. Consistent gene symbols, identifiers, and synonyms are used for the known, predicted and orphan coding genes from these organisms. Gene homologies are calculated using BLAST among these genomes. Standard vocabulary of molecular function, biological process and cell location is integrated into data searching and reporting.

euGenes is built with source data contained in the IUBio Archive / Bio-Mirror data distribution set. Its software derives from work by the PI in designing and developing the FlyBase project public interface as a portable, distributed web service for access to model organism data (FlyBase 1999). A recently funded NIH project that is related to this work is development of reusable, generic tools for genome data access for model organism genome databases. This focuses on developing practical tools for deploying directories of genome data in model organism databases, including FlyBase and *euGenes*. This is an

essential component for automated data access and grid computing. The NIH project complements the work proposed here, which focuses on data warehousing and distribution mechanisms.

Web functions in the euGenes service provide for search, retrieval, map display and summaries of genome data are provided for easy use, including ability to combine queries with gene ontology, homologies, features and other attributes. Public usage has risen from less than 1% of the FlyBase system in late 2000 to 5% of it in 2002. Many people use this system on a daily basis. These include bioscientists from all parts of the globe, in academia, government and biotechnology and pharmaceutical industries. The genome maps are particularly popular, for their ease of use and functionality. Reference database links are included for source data. Cross-links to euGenes from WormBase (Stein *et al.* 2001) and GeneCards (Rebhan *et al.* 1998) are currently available. Web logs show frequent and returning use by those referred from these services, an indication that this enhances WormBase and GeneCards services. Much of the open source software developed for euGenes is in Perl, with use of bioinformatics tools such as SRS, BLAST, and Unix tools including MySQL and Berkeley DB. Java programs (*gnomap* and *Readseq*, D. Gilbert, unpublished 2000) that produce genome maps and extract annotated sequence from genomes are component tools that are also used in other contexts.

III. Research plan

IUBio Archive data and software collections, including its component Bio-Mirror and euGenes data distribution and integration projects, will be maintained and updated in the proposal period. A five-year project span is requested to maintain continuity for current and new public services. Approximately half of this project effort will be directed to improving current information infrastructure, and one half toward research in new data access means. New databank additions will focus on the emerging genome sets along with other widely used bioinformatics databanks. These include searchable Medline literature data, and searchable Ensembl genome data. Software packages will be added to the bio-software collection through author-archiving and automated collection means.

Improvements planned for the euGenes system include automated replicability for distribution to other computers. Addition of new eukaryote organisms is planned, drawing on more source databases including Ensembl (vertebrates) and Gramene (rice). Web Services search and retrieval access genome summary data will be added, drawn from any common practice examples in genome database informatics, and additional human-usable web interface features will be added, drawn from available open-source genome informatics projects such as GMOD.

In addition, experimental work will be undertaken to extend access to these collections. Experiments with data grid and alternate technology to improve distribution of biology data, and especially genome data as found in euGenes and FlyBase projects are in progress, and offer future improvements in this area. Extensions planned for IUBio Archive include development of object level access through directories or catalogs of bio-information (Gilbert 2002c, 2002d), in context of grid computing and web services, with emphasis on high-volume, high-availability distributed data access systems. Improvements for the SRS data access system, including work with SQL and XQuery integration using IBM DiscoveryLink (Hass *et al.* 2001; renamed Information Integrator), and addition of bio-application services including EMBOSS (Rice, 2000) are planned.

A. Bio-information warehousing, access and distribution.

Improved bio-data distribution. The Rsync protocol (Tridgell, 2002), recently added at Indiana and Australian Bio-Mirror servers, has proven a useful alternative to FTP for file distribution, as it includes file system synchronization that updates only changed files. Rsync can also synchronize only changed sections within files, though for binary-compressed large databanks this may reduce efficiency. The SDSC Storage Resource Broker (Baru *et al.* 1998) provides access heterogeneous data sets based on attributes rather than physical locations, and will be tested in the Bio-mirror context for improvements to bio-data distribution. 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. Additional work to identify and deploy improvements for distribution will be undertaken.

Object versus Bulk distribution. Efficient search and retrieval for subsets of databanks, as object or record-oriented data exchanges, has several advantages over the current common practice of bulk databank distribution. A data record in bioinformatics can be any ID/Accession record from the source data model. Record-oriented exchanges could greatly improve distribution efficiency: only changed records need to be updated, rather than entire databanks. For efficient grid computing, one wants to be able to distribute selected subsets to many compute nodes, rather than whole databanks. Components of biology data grids will need to include directories or catalogs of data at file and object levels, to locate objects and groupings for replication, as well as efficient bulk data transport and automatic replication

management. The directory itself should be replicated and contain information on replication servers for each data set such as Bio-Mirror sites.

B. Bioinformatics tools and services.

In addition to improving the data and software collections, new interactive web services for use of the housed data will be developed. Addition of the EMBOSS suite of molecular biology analysis functions, and other bio-data analysis tools for web-based analyses, with links to SRS for data selection, will provide public access to integrated data analysis functions. SRS has several features recommending it for producing data directory information. It provides a tested and practical way of federating many related databanks, as well as offering automated, up-to-date access to these data. It has parsing methods for hundreds of biology databanks that are actively maintained by Lion Bioscience as data formats change. SRS provides summary metadata for each databank and data object, which can be adjusted to project needs. It is one of the fastest search and retrieval systems suited for the basic biological queries that are commonly text-oriented, and is efficient for large biology databanks, using inverted text indexing and regular expression parsing. Vocabulary databases such as Gene Ontology can be used for classifying DNA and protein objects through federated indexing (following cross-links between databases).

C. Genome information directory.

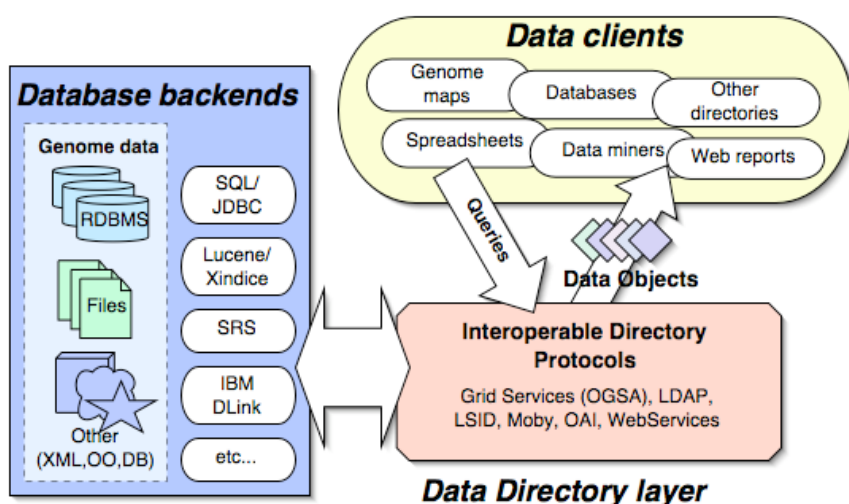
Work with euGenes genome collections will focus on providing summary catalogs or directories of primary information on eukaryote genes and genome features drawn from source databases. These summary data include gene identifiers, names and synonyms, database cross references, and summary information on function (Gene Ontology), feature location, and homologies, plus metadata relating to database sources and access links, update dates and versions. This will provide a *knuckle* service, a broad survey of gene information in source databases, in the *knuckles-and-nodes* approach suggested by Lincoln Stein (2003). The current genome collection will be extended to newly available eukaryote genomes. Search and retrieval of this genome summary data will be made available through the common data directory interfaces discussed. This service will be made highly replicable, drawing on the related project work for GMOD, so that genome catalogs can be readily copied and accessed from many Internet sites and within government and industry projects that require secure, local access.

A goal common to the above three sections of this project is the examination and use of available technology for an efficient bio-data access system that works at the object level. This can include existing and new genome and bio-database systems such as Ensembl, DiscoveryLink and SRS, with additions of LDAP for high efficiency, distributable query/retrieval of large volume object directories, and Web Services plus emerging Grid Services for wider access to object directories. Given no current common practice standard for distributed data access in bioinformatics, each of these viable contenders will be evaluated and employed where they prove most feasible and useful to improving community access.

New technology for providing computable and efficient high-volume access to large databanks is being developed in the context of data grids (Avaki Corp., 2002; OSGA-DAI, 2002; iVDGL, 2002) in several sciences. The developing grid computing infrastructure includes methods for efficient transport of high volume data, project-oriented sharing of data, security and resource authentication. Core technologies include Web Services, Grid Services and LDAP. Web Services provide computable web access to data, with examples for bio-databanks (XEMBL, XDDBJ). Lightweight Directory Access Protocol (LDAP) is a computer industry standard for distributed searching and updating large collections of metadata and their underlying objects. An example LDAP directory of bio-data is available at IUBio (Gilbert, 2002). The developing Open Grid Services Architecture (OGSA) standard has growing support (Foster and Kesselman, 1999), including bioinformatics projects (EGD-WP10, 2000; MyGrid, 2002). OGSA Database Access and Integration (OSGA-DAI, 2002) is particularly relevant to developing collection and

directory level access to science data. Additional technologies such as network performance monitoring (NMI, 2002) can usefully improve reliable, rapid distribution of databanks.

LDAP is interesting for high-volume data access: it uses common object schema, allows federated queries over distributed collections, is efficient for high volumes of objects by using binary object transport, and is flexible for accessing a range of data structures including flat files, object-oriented data and relational databases. Web Services offer much flexibility and garner much of the current interest in bioinformatics object exchange, but are not as fully developed a standard for distributed query as LDAP, and WS are not as efficient for high-performance scientific computing compared to binary transport methods (Chiu *et al.* 2002). Though these are different protocols, they share many similarities, and there are available translation tools to make Web Services and LDAP interoperate (e.g. DSML).



Data access middleware will be developed to link existing genomics and bioinformatics databanks to grid and related data retrieval protocols. Preliminary research of Dr. Gilbert has produced initial versions of these for Web Services and LDAP (Gilbert 2002c, 2002d). A directory service can support several interface standards, as they refer to the same underlying attributes and metadata, as diagrammed

Figure 2. Bio-object access components

in Figure 2. The major components of this diagram are the existing bioinformatics databases, comprising various relational databases, flat file databases, and others, with the software components that interact with these. Several of these are currently employed in IUBio Archive and euGenes projects. A data directory middleware layer to interact with these existing backends will provide distributed query and retrieval interfaces that are compliant with existing and developing standards for Internet distributed directory systems. Methods of abstracting collection interfaces will be employed in constructing a software library for this, usable by client programs developed for this project as well as for other projects. The construction of this biology collection service will include schema of suitable consensus attributes and metadata, and techniques for linking collections to existing databases.

Directories provide a “broad and shallow knuckle” of integration that can add significant value to existing databases. Such directories provide a means for researchers who develop laboratory databases that use links to reference data, to access and retrieve up-to-date gene or other object attributes in an efficient manner. They provide the means for analysis and visualization tools, such as genome map viewers, to likewise access and combine current data from multiple sources, useful for instance in comparative genome studies.

Directories provide a “broad and shallow knuckle” of integration that can add significant value to existing databases. Such directories provide a means for researchers who develop laboratory databases that use links to reference data, to access and retrieve up-to-date gene or other object attributes in an efficient manner. They provide the means for analysis and visualization tools, such as genome map viewers, to likewise access and combine current data from multiple sources, useful for instance in comparative genome studies.

An object entry in this directory will have primary attributes that are common to bio-data records: stable identifiers, name(s), functional categories, location of a gene on a genome, along with meta-data such as when it was updated, where among data centers this record can be located, what formats it comes in. Summaries of genome data available in GeneCards, LocusLink and euGenes provide working cases to define these primary attributes: identifiers, names and cross-database links are important for

directories, detailed attributes, experimental and computed evidence are less so. Categorizations such as Gene Ontology, taxonomic, developmental and anatomical attributes are important for semantic and high-level searches.

This directory layer should use common community-defined data definitions, schema and attributes. These will include object (gene, protein, reference) names, synonyms and Ids, cross-references to other databases, dates of updating, available formats, source and locations for data, categories defined by Gene Ontology, MIAME, taxonomy and other community standards. Directory linking via cross-references can join several projects, so that a number of related sites can be searched in one pass (distributed searches are a feature of LDAP that is not yet standard for Web Services). Community access to add and update directory entries is also a useful feature for enabling shared projects. Given appropriate security and access controls, as available with LDAP, this is straightforward to implement. For client access to directories, software will be built on existing libraries for Web Services and LDAP in Java and Perl languages. Computer-savvy bioscientists and bioinformaticians will be able to use client software with minimal programming.

Preliminary tests using data transport protocols LDAP, SOAP, and SRS-Web for search and retrieval of volumes of biosequence objects have been done (Gilbert, 2002d). Tests using the same database backend system differ in the object directory interface, transport protocol, and data reformatting required. Preliminary results indicate that LDAP is 10 times faster than Web Services methods (SOAP and SRS-Web), due to a more compact binary-encoded transport and a more direct route from server file system to client applications.

This project will test and develop public access via a variety of standard protocols: Web and Grid Services, LDAP and Open Archives Initiative. We intend to build information retrieval services that can be used via simple Java and Perl clients to enable researchers to access genome and molecular biology data, use filters to select data subsets of interest, and run programs that use the data without needing detailed knowledge of the underlying computer and operating system infrastructure. Open source software tools and libraries now exist for sciences information technology realms to provide the Data Directory Layer components (Figure 2) and associated client software to interact with it. Existing, open-source toolkits and libraries in Perl and Java languages for access to OGSA, OAI and LDAP compliant collections will provide the access support for information retrieval tools. LDAP client access is available with standard Java Directory Services and Perl libraries. Tomcat, Axis and SOAP projects of the Apache.org open-source community effort provide widely used web services frameworks that share access to common databases. OGSA-DAI and Community Grid toolkits (von Laszenwski *et al.* 2000) provide libraries for developing data grid client programs.

IV. Relation to present state of field

The need for efficient distributed computing methods in bioinformatics is increasingly evident, as seen from several projects that address this in different ways. The current common means for bioinformatics centers and laboratories to offer network services are web-based, which has the value of universal accessibility, but has limitations in computable access, and lack of standard user interfaces and distributed use of resources. Technology for finding bio-data currently relies primarily on use of FTP servers, Web page servers and general Web robot indexing such as Google and Alta-Vista. There are a few examples of services for finding and accessing bio-data using CORBA or related (RPC, Java RMI) technology.

EMBNet (www.embnet.org) in European and elsewhere, and APBionet (www.apbionet.org, collaborating in Bio-Mirror) in Asia-Pacific, are networks of bioinformatics centers that share data and tools in a fashion similar to the Bio-Mirror project. These networks, along with regional and institution centers currently rely primarily on standard FTP for bulk databank exchange. They have needs for improved data distribution and access that this project addresses. Many of these centers provide data access using SRS; future versions of this bio-data access system may include a network federation mechanism based on LDAP or Web Services.

GeneCards (Rebhan *et al.* 1998) is an integrated summary of human gene information available on the Internet and via bioinformatics databases. This project combines curation and automated information retrieval and analysis to produce a very easy to use summary of knowledge, along lines similar to the euGenes genome service, but more specialized in its area of genome information. The Pise project (Letondal 2000) provides a good example of Web-centric bioinformatics services. Pise has a command description language (XML based), which encapsulates the input, output and options needed for running bio-applications. The Embreo (Tribble 2001) and Jembooss (Carver and Mullan 2002) projects are developing Web Services methods for distributed computing of bioinformatics applications, particularly the EMBOSS package.

Current directories of software and data for biosciences produce either web-based HTML documents, and/or metadata in various formats. Many of these are human curated, and often suffer from going out-of-date after a time. Perhaps the best example of automatic or semi-automatically maintained biology resource directories is the BioNetBook project (Letondal *et al.* 1999). A public distributed genome annotation system (BioDAS, biodas.org) provides common access to genome information from WormBase, Ensembl, UCSC Human Genome Project, TIGR and FlyBase. BioDAS is a specific solution to a specific need, that of providing network access to reference genome annotations, with requirements for development and maintenance of client and server software focused on this one function. The myGrid project (www.mygrid.org.uk) is a developing bio-grid effort that will include research on data distribution methods.

Collaborative initiatives in the US, Europe and elsewhere have formed to build large-scale data grids for physics, astronomy, climate sciences and biology (iVDGL, PPDG, EGD-WP10). Data grid research approaches in Globus and iVDGL projects include studies on high volume distribution of data files as well as object level exchange (Allcock *et al.* 2002; Stockinger *et al.* 2002). Another high-performance data access technology is the SDSC Storage Resource Broker (Baru *et al.* 1998). This client-server middleware provides a uniform interface for distributing heterogeneous data over a network and for managing replicate data sets. It provides a way to access data sets and resources based on their attributes rather than names or physical locations.

The Life Sciences ID (LSID) initiative is a proposal for developing a standard naming system for biology data objects that would improve consistency in locating these objects, and could form a common, stable naming system for use with LDAP and other directory systems. The LDAP distributed search system seems particularly suited for summary metadata describing biology data objects, and recently is receiving attention for uses in development of biology data directory systems. The LSID proposal

includes reference implementations in both Web Services (LSID-WS 2003) and LDAP (LSID-LD 2003). There is bioinformatics community support for both, with arguments in favor of the LDAP mechanism (Phillips 2003). Robert Kincaid has been investigating LDAP for constructing bio-object directories, using NCBI LocusLink and related source data for gene lookups with gene expression and other database interactions (Kincaid 2002). LDAP is widely used in current computer operating systems, and is employed for directories of people and organization information, computer authentication and resource discovery. It has a well-developed security model, including access controls at any level of object hierarchy and fields. It is used in Grid computing (Czajkowski et al 2001), and forms part of the NSF Middleware initiative (NMI 2002), for use in locating large video objects and other directory functions.

A thoughtful review of current and new approaches to integrating biology databases is provided by Lincoln Stein (2003). In this he discusses several practices, from warehousing through distributed federation, and proposes a functional compromise to answer current needs with the difficult problem of integration. This “knuckles and nodes” approach builds onto source node databases (such as WormBase, FlyBase, GenBank) an additional group of knuckle or joining data services, which link and consolidate specific reference information among the source nodes. Examples of knuckle services include euGenes and LocusLink for joining gene databases by common identifiers, names and orthologies among organism genes. It should be clear that there is no easy nor single solution to integration of a large, heterogenous collection of bio-data, but that several approaches need to be tried, and best values of each used or added to other methods where possible. The BioMoby project (Wilkinson and Links 2002) is working to provide a centralized repository system for access to a broad range of biology data, via object names, using Web Services technology. The BioMoby project, MyGrid, implementations of LSID (LSID-WS 2003), and other bio-information integration projects have similarities with the research outlined in federated access to bio-objects. This project complements these others, emphasizing integration via directories of simple summary data and metadata that many genome and bio-data objects have, analogous to the Dublin Core for literature, and by focusing on aspects of efficient high-volume selection and transport used for finding and replicating bio-objects among computers.

The Open Grid Services Architecture (OGSA) standard for distributed computing has growing support in software industry and academic research (Foster and Kesselman 1999), including the NSF Middleware project (NMI 2002) and bioinformatics projects (EGD-WP10 2000, MyGrid 2002). Grid computing infrastructure includes standards and methods for efficient transport of high volume data, project-oriented sharing of data, security and resource authentication. The OGSA Grid Service interface defines Web Services operations for querying and retrieving data, with support for query languages such as XQuery. OGSA Database Access and Integration is particularly relevant to developing collection and directory level access to science data. Web Services, including the developing OGSA Grid Services, are very suitable for small message-sized interactions. But they are not efficient for high-performance scientific computing with large volume data transport, compared to available binary transport methods (Chiu *et al.* 2002).

The OpenURL standard (Van de Sompel and Beit-Arie 2001; NISO 2003) from Digital Library research is another developing standard that is important for many web-based scholarly information environments, including bioinformatics. OpenURL provides a mechanism for delivery of objects that originate in various domains including PubMed and Open Archives Initiative. LinkOut is an implementation of this for biosequence and literature data in NCBI Entrez and PubMed systems. These and other protocol standards in sciences and informatics are relevant to providing best access to collections of biology information.

V. Outcomes, publication and distribution

Outcomes include new methods for biology data access, distribution and federation and improved use of this knowledge for bioscientists. Extended bioinformatics software tools at IUBio Archive will be made available to public use and redistribution. Integration of these tools in servers at Bio-Mirror and other bioinformatics centers is underway, including directories of bio-information distributed from IUBio. Software, data and documentation produced have been and will continue to be made publicly available as an open-source effort, in electronic and traditional publications read by bioscientists and bioinformaticians.

VI. Schedule and Milestones

Year 1. IUBio server software archive and tools enhancements: upgrade SRS to version 7; add EMBOSS and related biodata analysis tools for web-based analyses; add Medline/PubMed literature database to SRS web search services; investigate use of IBM DiscoveryLink/Information Integrator with biology databases for web and distributed database searches. Add Ensembl genome organism databases for integration with euGenes summary data. Investigate, with Bio-Mirror partners, data replication using directories of data objects.

Year 2. Focus on functional developments and evaluation of best standards in data access of LDAP, Web and Grid Services methods. Identify use cases for laboratory and projects wanting distributed data query and retrieval for large subsets of genome and molecular biology data. Continue year 1 extensions for new databases and data access tools. Extended metadata production for integration of bio-information, with preliminary public access and distribution.

Year 3. Focus on ease-of-use improvements in euGenes genome summary web interfaces; addition of robust GMOD tools to euGenes. Addition of emerging best practices for data grid and bio-data cataloging systems, depending on functionality, including Web Services and Grid Services for data access, BioMoby, LSID implementations. Work on distribution, documentation and implementation help for bioinformatics server tools to Bio-Mirror and other bioinformatics centers.

Year 4. Work on practical problems for developed data access systems, and on ease of replication and installation at partner and other bioinformatics centers. Focus on solving problems with use cases of client laboratories and projects. Improvements to documentation and distribution mechanisms for advanced data access tools, in conjunction with Bio-Mirror partners and other bioinformatics centers. Revisiting of best bioinformatics community practices and tools for data access, with modification and addition of access methods and data sets as required.

Year 5. Focus on solving production access problems with integrated bio-data directories, and on problems arising in partnership centers and laboratory use cases. Extended publication and dissemination of available access systems.

VII. Broader impacts

Contributions to Bioinformatics: This project enhances access to genomic and molecular biology data and software in several ways: it collects from primary sources and redistributes over global network; it organizes new data and software into consistent frameworks for easy use (SRS, euGenes, bioinformatics web); it develops improved methods for public use of bioinformatics software and data

including emerging data/compute grid methods; euGenes summarization of genome data and skeleton for new genome information system; and e-Prints database with its Open-Archive Initiative access for software publication and organization. The research proposed focuses on information engineering for effective management of proliferating bio-information, rather than fundamental new experiments in computational biology. This project will enhance the infrastructure of bioinformatics in the US and worldwide, as a data warehouse connected with high-speed Internet network infrastructure. It will enhance partnerships with Asian-Pacific, European and western hemisphere nations in areas of bioinformatics and high-speed networking. Benefits to society include improved world scientific partnerships and technology transfer to areas that are now building bioinformatics and Internet centers for sciences.

Education: Teaching bioinformatics methods to graduate students, including those working on this project and two dozen others enrolled in the Indiana University School of Informatics Masters in Bioinformatics program, is a component of the PI's efforts. Talks and lectures by the PI on this area of bioinformatics to students and a range of academic and industry scientists are ongoing. Informal teaching and advising to biology graduate students and faculty on bioinformatics methods also is an educational contribution. The international bioinformatics organizations APBioNet (Asia-Pacific Bioinformatics network; www.apbionet.org) and EMBNet (European Bioinformatics network; www.embnet.org) are collaborative partners with Indiana University bioinformatics efforts of Dr. Gilbert. These organizations include educational objectives, workshops, courses and efforts to promote bioinformatics education in numerous countries.

The euGenes genome service is useful for and used by general public, and provides a reference source for museums and other science education organizations for information about the Human and related genomes. The National Science Teachers Association (NSTA.org) has selected it as a teaching resource (www.scilinks.org). A genomics exhibit at American Museum of Natural History, created by Dr. Robert DeSalle (<http://www.amnh.org/exhibitions/genomics/>), was aided by collaboration with the PI. The National Academy of Sciences' Marian Koshland Science Museum is developing a genomics exhibit (Dr. Erika Shugart, managing editor), which also may well contain information derived from the euGenes.

Contributions to other Sciences and beyond Science and Engineering: This project is oriented to improving access to life sciences information for all citizens. This PI and Indiana University are enthusiastically committed to enabling participation of all citizens, especially those often underrepresented in technological and engineering disciplines, including minorities, women and persons with disabilities. We will seek to involve and educate a broad spectrum of citizens with this project. The archive of Bionet news at IUBio is widely used by general public for learning about active research in many bioscience areas. The PI is involved in discussions in the State of Indiana on ways to improve bioinformatics practices and education for commerce.

Developing collaborations with other sciences in grid computing are expected to yield multi-disciplinary computing infrastructure useful to many disciplines. Development of directory systems in which security and authenticated access to biology data are fundamental components has applications for national security and bio-threat response research. Improvements in access to public genome and bioscience data will aid in the general improvement of science/technology understanding and integration into commercial and personal activities.

REFERENCES CITED

- Allcock, B, J. Bester, J. Bresnahan, A. L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnal, S. Tuecke, 2002. Data Management and Transfer in High Performance Computational Grid Environments. *Parallel Computing Journal*, Vol. 28 (5), pp. 749-771.
- Avaki Corporation, 2002. Avaki Data Grid 3.0, Java-based data grid software. URL: <http://www.avaki.com/news/release20021209.html>
- Baru, C., R. Moore, A. Rajasekar, M. Wan 1998. The SDSC Storage Resource Broker, Proc. CASCON'98 Conference, Toronto; also <http://www.npaci.edu/DICE/SRB/>.
- Baxevanis, A. and B.F.F. Ouellete, eds, 1998. *Bioinformatics: A practical guide to the analysis of genes and proteins*. J. Wiley and Sons, Inc., NY.
- Bio-Mirror project, 1997. A public service for distribution and access to biosequence and bioinformatics data. URL: <http://www.bio-mirror.net/>
- Carver, TJ, Lisa J Mullan, 2002. A new graphical user interface to EMBOSS. *Comparative and Functional Genomics*, 3(1): 75-78. <http://www.uk.embnet.org/Software/EMBOSS/Jemboss/>
- Chiu, K., Govindaraju, M., Bramley, R. 2002. Investigating the Limits of SOAP Performance for Scientific Computing. Technical report TR559, IU Extreme Lab. URL: <http://www.extreme.indiana.edu/xgws/papers/soap-hpdc2002/soap-hpdc2002.pdf>
- Covitz, P.A., 2003. To Infinity, and Beyond: Uniting the Galaxy of Biological Data. *Omics A Journal of Integrative Biology*, 7: 21-22
- Czajkowski, K., S. Fitzgerald, I. Foster, C. Kesselman, 2001, "Grid Information Services for Distributed Resource Sharing", Proc. Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), IEEE Press, August 2001. Also <http://www.globus.org/datagrid/>
- DDBJ. The Dna Databank of Japan. URL: <http://www.ddbj.nig.ac.jp/>
- EBI. Brooksbank, C. et al. 2003. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, 31: 43-50 DOI: 10.1093/nar/gkg066 URL: <http://www.ebi.ac.uk/>
- EGD-WP10, 2000. WP10 of European Union Data Grid Project. Grid-aware biomedical applications for datagrid testbed assessment. URL: <http://marianne.in2p3.fr/datagrid/wp10/>
- EMBnet, 2000. The European Molecular Biology network. URL: <http://www.embnet.org/>
- Ensembl. Hubbard, T. et al., 2002. The Ensembl genome database project. *Nucleic Acids Research*, 30: 38-41; Clamp, M., et al., 2003. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Research*, 31: 38-42. doi: 10.1093/nar/gkg083 URL: <http://www.ensembl.org/>
- Etzold, T, and Argos, P., 1993. SRS – an indexing and retrieval tool for flat file data libraries. *Comp. Appl. Biosci.*, 9 : 49-58.
- FlyBase Consortium, 1999. The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res.*, 27: 85-88.
- Foster, I. and Kesselman, C., eds., 1999. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann.
- Foster, I., C. Kesselman, J. Nick, S. Tuecke, 2002. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002.

- Gene Ontology Consortium, 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.*, 25: 25-29. URL: <http://www.geneontology.org/>
- Gilbert, D., 1989. IUBio Archive for biology software and data. FTP and <http://iubio.bio.indiana.edu>
- Gilbert, D., 1990. Two Hypercard calculators for molecular biology. *Comput. Appl. Biosci.* 6: 113-116. GenBank Search, a Hypercard stack for network searching and fetching sequences from the GenBank e-mail server. October 1990.
- Gilbert, D., 1992. Booleans, partial words and other WAIS mods in biology data searches. *Bionet.Software*, Nov. 1992. URL: <ftp://iubio.bio.indiana.edu/util/wais/>
- Gilbert, D., 1993. The Global Library. *Trends in Biochem. Sci.* 18: 107-108.
- Gilbert, D., 1995. New SRS node at IUBio, Indiana. *Bionet.Software.SRS*, March 1995. URL: <http://iubio.bio.indiana.edu/srs/srsc>
- Gilbert, D., 1999. Free Software in Molecular Biology for Macintosh and MS Windows Computers. in *Bioinformatics Methods and Protocols* (S. Misener and S.A. Krawetz, eds.) Humana Press, NJ. Also <http://iubio.bio.indiana.edu/soft/molbio/Listings.html>
- Gilbert, D.G., 2002a. euGenes, a eukaryote organism genome information service. *Nucleic Acids Res.*, 30, 145-148 URL: <http://iubio.bio.indiana.edu/eugenest/>
- Gilbert, D.G., 2002b. Pise, software for building bioinformatics webs. *Briefings in Bioinformatics*, 3(4): 405-409.
- Gilbert, D.G., 2002c. BioGridRunner, a preliminary bioinformatics client application for data grid computing. <http://iubio.bio.indiana.edu/grid/>
- Gilbert, D.G., 2002d. Directories of Bio-data. <http://iubio.bio.indiana.edu/biogrid/directories/>
- Gilbert, D.G., 2003a. Protein family alignment annotation. *Briefings in Bioinformatics*, 4(2) (in press).
- Gilbert, D.G., 2003b. Shopping in the genome market with EnsMart. *Briefings in Bioinformatics*, 4(3) (in press).
- Gordon, P. 2002. Biosequence XML formats. URL: <http://www.visualgenomics.ca/gordonp/xml>
- Haas, L. M., P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope, 2001. DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, 40(2): 489-511 URL: <http://www.research.ibm.com/journal/sj/402/haas.html>
- iVDGL, 2002. The International Virtual-Data Grid Laboratory. URL: <http://www.ivdgl.org/>
- Kincaid, R., 2002. BNS: A DNS-Inspired Biomolecule Naming Service, Poster 157B(i), ISMB 2002, 8/5/2002. <http://openbns.sourceforge.net/>
- Lagoze, C., H. Van de Sompel, 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Roanoke VA, June 24-28, 2001, pp. 54-62
- Letondal, C. 2000. PISE, a tool to generate Web interfaces for Molecular Biology programs. Pasteur Institute. <http://www-alt.pasteur.fr/~letondal/Pise/>
- Letondal, C., S. Bortzmeyer, A. Thebault, I. Wang, 1999. Bio Netbook, <http://www.pasteur.fr/recherche/BNB/bnb-en.html>
- Letovsky, S., (ed), 1999. *Bioinformatics: Databases and Systems*, Kluwer Academic: Boston.

LSID, 2002. Life Sciences Identifier standard, www.i3c.org. URL: http://www.i3c.org/workgroups/technical_architecture/resources/lsid/docs/index.htm

LSID-LD, 2003. LSID reference implementation in LDAP. <http://sourceforge.net/projects/lsid>

LSID-WS, 2003 LSID reference implementation in WebServices. <http://www.i3c.org/wgr/ta/resources/lsid/docs/index.htm>

McLoughlin, L., 1998. Mirror perl package. URL: <http://sunsite.org.uk/packages/mirror/>

MyGrid, 2002. The MyGrid project, Univ. Manchester, UK. URL: <http://www.mygrid.org.uk/>

NCBI. National Center for Biotechnology Information NCBI, URL: <http://www.ncbi.nih.gov/>

NIG. National Institute of Genetics. URL: <http://www.nig.ac.jp/>

NISO, 2003. National Information Standards Organization. NISO Linking Solution -- the OpenURL. URL: <http://www.niso.org/standards/resources/OpenURL-release.html>

NMI, 2002. The NSF Middleware Initiative. URL: <http://www.nsf-middleware.org/>

OGSA-DAI, 2002. Open Grid Services Architecture: Data Access and Integration project. URL: <http://www.ogsadai.org.uk/>

Parker, D.S., Gorlick, M., Lee, C. 1993. Evolving from Bioinformatics in-the-Small to Bioinformatics in-the-Large. *Omics A Journal of Integrative Biology*, 7: 37-48

Phillips, J. 2003. Recommendation for the Support of a LSID Resolution Proposal. URL: ftp://ftp1.nci.nih.gov/pub/cacore/caBIO/lsid/lsid_memo.doc

PPDG. The Particle Physics Data Grid (<http://www.ppdg.net/>); The Grid Physics Network (<http://www.griphyn.org/>)

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D., 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14: 656-664.

Rice, P. 2000. EMBOSS - The European Molecular Biology Open Software Suite <http://www.sanger.ac.uk/Software/EMBOSS/>

SRS. Sequence Retrieval System, Lion Bioscience. URL: <http://srs.ebi.ac.uk/>

Steele, C. 2002. E-prints: the future of scholarly communication? *inCite*, October 2002. <http://www.alia.org.au/incite/2002/10/eprints.html>

Stein, L. 2003. Integrating Biological Databases. *Nature Reviews Genetics*. 4: 337-345. doi:10.1038/nrg1065

Stein, L., Sternberg, S., Durbin, R., Thierry-Mieg, J., and Spieth, J., 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, 29: 82-86.

Stockinger, H., A. Samar, B. Allcock, I. Foster, K. Holtman, and B. Tierney, 2002. File and Object Replication in Data Grids.; *Journal of Cluster Computing*, 5(3)305-314.

OAI. The Open Archives Initiative. URL: <http://www.openarchives.org/>

Tribble, P. 2001. Embreo. <http://www.hgmp.mrc.ac.uk/~ptribble/Embreo/>

Tridgell, A. 2002. rsync, remote file synchronization system. URL: <http://rsync.samba.org/>

Valencia, A. 2002. Search and retrieve: Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Reports* 3(5): 396 - 400. doi: 10.1093/embo-reports/kvf104

- Van de Sompel, H. and O. Beit-Arie, 2001. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine*, 7(3). ISSN 1082-9873
- von Laszenwski, G. I. Foster, J. Gawor, W. Smith and S. Tuecke, 2000. CoG Kits: A bridge between commodity distributed computing and high-performance grids. *ACM 2000 Java Grande Conference*. URL: <http://www.globus.org/cog/>
- Wilkinson, MD and Links, M., 2002. BioMOBY: an open-source biological web services proposal. *Briefings In Bioinformatics* 3:4. 331-341.
- XDDBJ. XML Central of DDBJ, URL: <http://xml.nig.ac.jp/>
- XEMBL. EMBL Nucleotide Sequence data in XML. URL: <http://www.ebi.ac.uk/xembl/>
- Zdobnov, E.M., Lopez, R., Apweiler, R., Eitzold, T., 2002. The EBI SRS server – recent developments. *Bioinformatics*, 18, 368-373.

Biographical sketch for Donald G. Gilbert

Address

Biology Department, Indiana University, Bloomington, IN 47405 gilbertd@bio.indiana.edu

Education

Post-doctoral, Evolutionary and computational biology, University of Chicago, 1983-1984; Syracuse University, 1981-1983

Ph.D., Evolutionary Biology, Indiana University, 1981

B.S., Biology, 1977, B.S., Psychology, 1974, University of Illinois

Employment

Senior Scientist, Bioinformatics, Biology Dept. & Ctr. for Genomics & Bioinformatics, IU, 2000 –

Associated faculty, School of Informatics, Indiana University, 1999-

Senior bioinformatician with FlyBase genome informatics project, 1992-

Biocomputist and computer manager, Indiana University, 1988-2001

Software developer and consultant, Bloomington, Indiana, 1984-1988

Active bioinformatics projects

IUBio Archive of Biology software and data, ftp and <http://iubio.bio.indiana.edu/>

FlyBase, Database of the Drosophila Genome, <http://flybase.bio.indiana.edu/>

Bio-Mirror project for world distribution of large bioinformatics data sets, <http://www.bio-mirror.net/>

EuGenes, Genome Informatics for Model Eukaryotic Organisms, <http://eugen.es.org/>

Talks, Honors, et cetera

Invited talk, Globus World Life Sciences Workshop (San Diego), January 2003

Invited talk, I-Light Applications Workshop (Indianapolis), December 2002

Invited talk, Eli Lilly Bioinformatics group (Indianapolis), November 2002

Invited talk, Daphnia Genome Consortium Workshop (Indiana U), September 2002

Invited talk, Sequence Retrieval System Workshop, ISMB02 (Edmonton, AB.), August 2002

Invited talk, DOE conference *Computational Strategies for Whole Genome Analysis*, Argonne Natl. Lab., May 1996

Invited talk, National Center for Biotechnology Information (NCBI), February 1991

Editorial board, **Briefings in Bioinformatics**, 2002-

IUBio Archive chosen as software repository for **Bioinformatics** journal, 1996 - present

Member of NSF, USDA and DOE review panels for IT Research, Rice, Arabidopsis projects, 1996-1999

Editorial board, **Gene-Combis** and **Journal of Computational Biology**, 1995 - 1997

Internet biology community support for IUBio Archive, 1993

Reviewer for Bioinformatics, Evolution, Science, Genetics, Theoretical & Applied Genetics, US and UK grant agencies.

Publications

- Gilbert, D.G., 2003. Shopping in the genome market with EnsMart. **Briefings in Bioinformatics**, 4(3) (in press).
- Gilbert, D.G., 2003. Protein family alignment annotation. **Briefings in Bioinformatics**, 4(2) (in press).
- Gilbert, D.G., 2002. Pise, software for building bioinformatics webs. **Briefings in Bioinformatics**, 3(4): 405-409.
- Gilbert, D. G. 2002. BioGridRunner, a preliminary bioinformatics client application for data grid computing. <http://iubio.bio.indiana.edu/grid/>
- Gilbert, D.G., 2002. Sequence file format conversion with command-line Readseq. In **Current Protocols in Bioinformatics**, A. Baxevanis and D. Davison, eds. Wiley, (in press).
- Gilbert, D. G. 2002. euGenes, a eukaryote organism genome information service. **Nucleic Acids Res.** 30: 145-148. Also <http://iubio.bio.indiana.edu/eugenesis/>
- Gilbert, D.G., 1999. Free Software in Molecular Biology for Macintosh and MS Windows Computers. in **Bioinformatics Methods and Protocols**, S. Misener and S.A. Krawetz, eds. Humana Press, NJ. Also <http://iubio.bio.indiana.edu/soft/molbio/Listings.html>
- Gilbert, D.G., 1996, 1998. SeqPup, biosequence editor and analysis software for molecular biology Bionet.Software, July 1996. See also <http://iubio.bio.indiana.edu/soft/molbio/seqpup/>
- Gilbert, D.G., 1996. Portable FlyBase server available for Drosophila data. Bionet.Announce, Also <http://flybase.bio.indiana.edu/docs/Portable-server/>
- Gilbert, D.G., 1995. Genbank searches at IUBio archive. Bionet.Molbio.Genbank, <http://iubio.bio.indiana.edu/Genbank-Sequences/>
- Gilbert, D.G., 1995. New SRS node at IUBio, Indiana Bionet.Software.SRS, Also <http://iubio.bio.indiana.edu/srs/>
- Gilbert, D.G., 1993. The Global Library, **Trends in Biochem. Sci.**,18: 107-108.
- Gilbert, D.G., 1990. Two Hypercard calculators for molecular biology. **Comput. Appl. Biosci.** 6: 113-116.

Collaborators (past four years)

W. M. Gelbart (Harvard Univ.), A. de Grey (Univ. of Cambridge), M. Juncai (Chinese Academy of Sciences), T. Kaufman (Indiana Univ.), K. Matthews (Indiana Univ.), A. Mizushima (AFFRC, Japan), Y. Ugawa (AFFRC, Japan), T. Tan Wee (National Univ. of Singapore)

Advisor to students

Paul Poole, Nihar Sheth, Manish Anand, Danfeng Yao, Lily Han Xu,

Graduate and Postgraduate Advisors

Rollin Richmond (Indiana Univ.), William T. Starmer (Syracuse Univ.), Russel Lande (Univ. of Chicago)