

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed 56-character length restrictions, including spaces.</i>		LEAVE BLANK—FOR PHS USE ONLY.				
		Type	Activity	Number		
		Review Group		Formerly		
		Council/Board (Month, Year)		Date Received		
1. TITLE OF PROJECT Bioinformatics data and compute grids for bioscientists						
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES (If "Yes," state number and title) Number: PA-00-117 Title: Innovations in biomed inform sci & tech R21/R33						
3. PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR			New Investigator <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes			
3a. NAME (Last, first, middle) Gilbert, Donald, George		3b. DEGREE(S) Ph. D.				
3c. POSITION TITLE Associate Scientist		3d. MAILING ADDRESS (Street, city, state, zip code) Indiana University Jordan Hall 153 1001 E. 3rd St. Bloomington, IN 47405				
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Center for Genomics and Bioinformatics						
3f. MAJOR SUBDIVISION Office of Research and University Graduate School						
3g. TELEPHONE AND FAX (Area code, number and extension)		E-MAIL ADDRESS: gilbertd@bio.indiana.edu				
TEL: 812.855.0587		FAX: 812-856-9340				
4. HUMAN SUBJECTS RESEARCH <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		4a. Research Exempt <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes If "Yes," Exemption No. _____		5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		
		4b. Human Subjects Assurance No. M1167-02	4c. NIH-defined Phase III Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	5a. If "Yes," IACUC approval Date	5b. Animal welfare assurance no A4094-01	
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY)		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT		
From 12/01/02	Through 11/30/06	7a. Direct Costs (\$) \$ 100,000	7b. Total Costs (\$) \$ 149,000	8a. Direct Costs (\$) \$ 549,007	8b. Total Costs (\$) \$ 788,620	
9. APPLICANT ORGANIZATION Name Indiana University Address P.O. Box 1847 Bloomington, IN 47402-1847			10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input checked="" type="checkbox"/> State <input type="checkbox"/> Local Private: → <input type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged			
Institutional Profile File Number (if known) 20			11. ENTITY IDENTIFICATION NUMBER 1356001673A1 DUNS NO. (if available) 00-604-6700 Congressional District 8			
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Steven A. Martin Title Assistant VP for Research Address Indiana University P.O. Box 1847 Bloomington, IN 47402-1847 Tel 812.855.3963 FAX 812.855.9943 E-Mail spon2@iupui.edu			13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name George Walker Title VP for Research Address Indiana University P.O. Box 1847 Bloomington, IN 47402-1847 Tel 812.855.0516 FAX 812.855.9943 E-Mail rugs@indiana.edu			
14. PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR ASSURANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. I agree to accept responsibility for the scientific conduct of the project and to provide the required progress reports if a grant is awarded as a result of this application.			SIGNATURE OF PI/PPD NAMED IN 3a. (In ink. "Per" signature not acceptable.)		DATE	
15. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.			SIGNATURE OF OFFICIAL NAMED IN 13. (In ink. "Per" signature not acceptable.)		DATE	

Principal Investigator/Program Director (Last, first, middle): [Gilbert, Donald, George](#)

DESCRIPTION: State the application's broad, long-term objectives and specific aims, making reference to the health relatedness of the project. Describe concisely the research design and methods for achieving these goals. Avoid summaries of past accomplishments and the use of the first person. This abstract is meant to serve as a succinct and accurate description of the proposed work when separated from the application. If the application is funded, this description, as is, will become public information. Therefore, do not include proprietary/confidential information. **DO NOT EXCEED THE SPACE PROVIDED.**

The challenge addressed here is to enable teams of bioscience and biomedical researchers to effectively collaborate and share very large amounts of widely distributed data and computational resources. A system known as a "data grid" is an active area of research for scientific computing, and substantial investments are being made in the United States, Europe, and Asia to deploy international data grid infrastructures. Bioinformatics data grid methods are of economic importance to biomedical, pharmaceutical, agribusiness, and related industries. To empower bioscientists and medical researchers to answer questions using large, heterogenous bio-data sets, automation of data and compute resource discovery and distribution are essential.

Development, testing and deployment of a community-based, internationally integrated genomics and bioinformatics data grid is the main objective of this project. An essential component of this is development of grid-aware bioinformatics applications for use by bioscientists. Components and features of a bioinformatics grid include: resources discovery using computable directories of bioinformatics data and software; efficient network transport for high-volume data distribution; security and authentication in use of resources; and peer to peer methods for community, project oriented data sharing.

The US and Asia-Pacific bioinformatics data distribution project (Bio-Mirror), European Union's molecular biology network, national and regional bioinformatics centers are currently investigating grid technologies for this. As exemplified by the Globus package, grid methods are a good match to bioinformatics needs in reliance on standard protocols for data distribution, secure and authenticated resource use, and replica management. IUBio and Bio-Mirror projects at Indiana University have a decade of experience in bio-data distribution, and can help lead internationally in deploying grid methods.

PERFORMANCE SITE(S) (organization, city, state)

[Indiana University, Bloomington, Indiana](#)

KEY PERSONNEL. See instructions. Use continuation pages as needed to provide the required information in the format shown below. Start with Principal Investigator. List all other key personnel in alphabetical order, last name first.

Name	Organization	Role on Project
Donald George Gilbert	Indiana University	Principle Investigator
to be appointed	Indiana University	Research Associate

Disclosure Permission Statement. Applicable to SBIR/STTR Only. See instructions. Yes No

The name of the principal investigator/program director must be provided at the top of each printed page and each continuation page.

**RESEARCH GRANT
TABLE OF CONTENTS**

	<i>Page Numbers</i>
Face Page	1
Description, Performance Sites, and Personnel	2-
Table of Contents	3
Detailed Budget for Initial Budget Period (or Modular Budget)	4-5
Budget for Entire Proposed Period of Support (not applicable with Modular Budget)	6
Budgets Pertaining to Consortium/Contractual Arrangements (not applicable with Modular Budget)	
Biographical Sketch —Principal Investigator/Program Director <i>(Not to exceed four pages)</i>	7-9
Other Biographical Sketches <i>(Not to exceed four pages for each - See instructions)</i>	
Resources	10
 Research Plan	
Introduction to Revised Application <i>(Not to exceed 3 pages)</i>	
Introduction to Supplemental Application <i>(Not to exceed one page)</i>	
A. Specific Aims	11-12
B. Background and Significance.....	12-13
C. Preliminary Studies/Progress Report/ Phase I Progress Report (SBIR/STTR Phase II ONLY)	13-16
D. Research Design and Methods	16-20
d1. Milestones (R21)	21
a.-d. R33 phase	22-30
Inclusion of Women (Required if Item 4 on the Face Page is marked "Yes")	
Inclusion of Minorities (Required if Item 4 on the Face Page is marked "Yes")	
Inclusion of Children (Required if Item 4 on the Face Page is marked "Yes")	
Data and Safety Monitoring Plan (Required if Item 4 on the Face Page is marked "Yes" and a Phase I, II, or III clinical trial is proposed)	
F. Vertebrate Animals	
G. Literature Cited	31-32
H. Consortium/Contractual Arrangements	
I. Consultants.....	
J. Product Development Plan (SBIR/STTR Phase II and Fast-Track ONLY)	
Checklist	33
Appendix <i>(Five collated sets. No page numbering necessary for Appendix.)</i>	
<i>Appendices NOT PERMITTED for Phase I SBIR/STTR unless specifically solicited.</i>	<input checked="" type="checkbox"/>
Number of publications and manuscripts accepted for publication <i>(not to exceed 10)</i>	
Other items (list): 1. BioGridRunner preliminary application manual and CD 2. Notes on Globus grid services installation and testing	
 Research Plan, R33 phase	
a. Specific Aims	22
b. Background and Significance	23-25
c. Preliminary Studies/Progress Report	25
d. Research Design and Methods	25-30

A. Specific aims (R21)

The preliminary project will evaluate, design and test components of a bioinformatics grid to enable bioscientists to better use distributed and high-end computing resources. Components for easy use by bioscientists of grid resources, for locating and distributing bio-data among grid computers, and collaboratively and securely sharing computing resources in integration with bioinformatics applications and data will be part of this work. In the following R33 phase, these components will be fully developed, documented and published for public use.

A grid-aware, bioscientist-friendly application called “BioGridRunner” will be designed to use distributed computing infrastructure, as the primary component of work. This will use the CoG Toolkit of the Globus project [CF, FC], Java graphic user interface (GUI) and network directory (JNDI) technology, along with bioinformatics software toolkits. It will include directories of data and applications that are easy to search and manipulate, program dialogs for use of a wide range of bioinformatics applications, and methods for manipulating and processing biosequence data, and program results including graphics.

A second necessary component in this plan is the test construction of directories of bioinformatics resources at bioinformatics centers, including gene and genomic data, and application software. Such bio-information directories will be developed for use with BioGridRunner and other grid-aware software, using LDAP and other existing data directory technology. The Sequence Retrieval System (SRS) [ET] and/or other bio-data search and retrieval software will be a test case for data directory construction.

A third component is the test deployment of Grid infrastructure at bioinformatics centers and laboratories, for use with distributed computing of bioinformatics applications, including sequence databank searching, sequence analysis tools, and phylogenetic analysis tools.

A long-term goal of this work is the development of methods to search massive amounts of up-to-date bio-data, distributed among many servers, in an efficient way. Next, retrieve the data of interest to computers that the scientist has authority to use, whether personal desktop or institutional servers, with emphasis on efficient and secure data transport. Bio-applications are likewise located and installed as needed on these computers, analyzing the data, and returning results to the scientist. This can and should be done without the computer-naive biologist having to learn the informatics underpinnings. The BioGridRunner application proposed here exemplifies some of these features. It employs simple to use, established technology and emerging grid methods.

Potential collaborations with other bioinformatics and sciences groups on emerging Grid technologies will be identified for further development of interoperable systems. Potential collaborators, several of whom the PI has been in contact with informally about this work, include: BioMirror and BioGrid efforts of the global Asian Pacific Advanced Network (APAN) centers; the European Union Data Grid project with its bioinformatics working group; the international Virtual-Data Grid Laboratory (iVDGL) which has

interests in bioinformatics and partners at IU; Indiana-regional groups including bioinformatics centers at Purdue University and Indiana University Medical Center, and biotechnology companies; Indiana University e-Science informatics research groups in Computer Sciences, Physics, Informatics and Information Technology.

B. Background and Significance (R21)

One of the pressing challenges in bioinformatics is to enable geographically distributed teams of people to collaborate and share very large amounts of data and computational resources. The total volume of data that is on-line for biosciences research is many terabytes and increasing rapidly. Human Genome data, whole genome analyses, biochemical and protein structure predictions and proteomics data all challenge current computing resources. Improvements in distributed computing with Grid infrastructure are of economic importance to biomedical, pharmaceutical, agribusiness, and related industries. Bio-data is widely distributed, being collected and curated by many research groups, and published in many ways, from database and spreadsheet to many text data structures, to XML, HTML and other document structures. To empower bioscientists and medical researchers to answer questions using these large, heterogeneous data sets, automation of data and compute resource discovery and distribution are essential.

Combining the information and computational resources and abilities of many organizations and individuals may solve numerous biosciences research problems. However building a distributed information infrastructure that allows such scientific collaborations to share data and computational resources in secure manner over a wide area is a very difficult task. Such a system, known as a “data grid”, is currently a very active area of research in Computer Science [FK], and substantial investments are being made in the United States, Europe, and Asia to deploy international data grid infrastructures.

Collaborative initiatives in the U.S. have formed to build large-scale data grids for physics and astronomy [PG]. The Data Grid Project [EG] is a collaboration of particle physicists, earth scientists, and biologists building a grid infrastructure to enable next generation scientific exploration in these disciplines. This data grid’s biomedical applications working group [EB] has several goals in common with this proposal. Several national data grid initiatives in EU member states have been launched in the past year.

The objective of this proposal is to take initial steps in the development, testing and deployment of a community-based, internationally integrated genomics and bioinformatics grid infrastructure. An essential component of this is development of grid-aware bioinformatics applications for use by bioscientists of these grid resources.

Components and features of such a bioinformatics grid include: (a) Resources discovery using computable directories of bioinformatics resources, including public reference data and shared project data, documents, and software. (b) Efficient network transport methods for high-volume data distribution. (c) Security and authentication in distribution and use of resources. (d) Peer to peer methods for community, project oriented data sharing.

A goal of this project is improving data exchange among science centers within Indiana University and with global partners. The US and Asia-Pacific bioinformatics data distribution project Bio-Mirror [BM], European Union's molecular biology network, national and regional bioinformatics centers are currently investigating grid technologies for this. As exemplified by the Globus package [CF, FC], grid methods are a good match to bioinformatics needs in reliance on standard protocols for data distribution, secure and authenticated resource use, and replica management. Bioinformatics software requires improvements to take advantage of these. IUBio and Bio-Mirror projects at Indiana University have a decade of experience in bio-data distribution, and can help lead internationally in deploying grid methods.

Advances will require integration of heterogeneous computing, human, and data storage resources distributed locally, regionally and internationally. At Indiana University, we participate in several projects advancing grid technology for these fields, including science grid research and development (<http://www.cs.indiana.edu/Research/distributed/>) distribution of bioinformatics databases, the Grid Physics Network for particle physics (<http://lexus.physics.indiana.edu/griphyn/>), information technology (<http://www.indiana.edu/~ovpit/ipcres/>) and other areas. This project exploits investments made by Indiana University in several key areas of information technology, including: advanced networking (Internet2/Abilene, TransPAC, Global Network Operations Center), high performance computing, and massive data storage. This science informatics environment is synergistic for the development of bio-grid methods.

C. Preliminary Studies (R21)

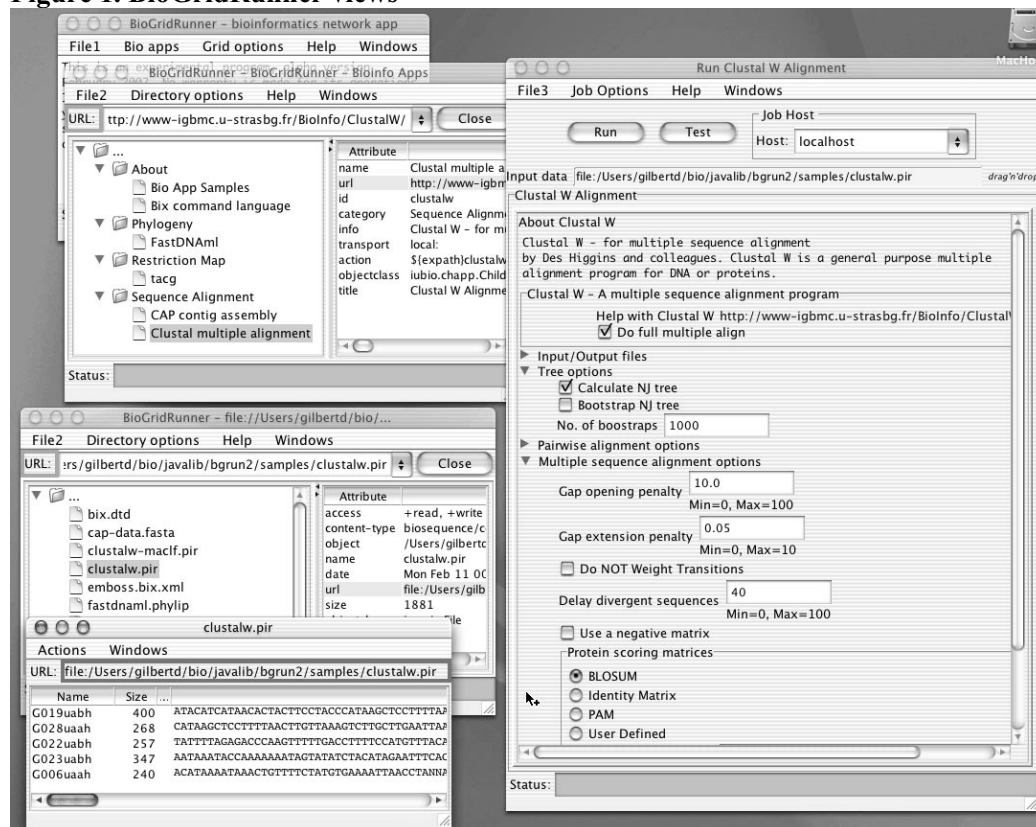
There are a number of aspects involved in grid computing; many deal with software, data and information resources that are installed or need to be added to server / resource computers. The example tested is the Globus grid tool set (<http://www.globus.org/>), for servers and clients. Appendix 2 provides preliminary notes on installation and testing of the Globus packages for use with bioinformatics applications. The Globus Grid server software, versions 1 and 2 (beta), have been deployed for testing at Indiana University's Center for Genomics and Bioinformatics (CGB) servers.

As well, perhaps more importantly, for bioinformatics information, are directories of where to find, how to get, data and software. Computable directories unlike web pages, are built with such as LDAP directory software, and can be searched, federated among widely disparate sources of information, to list the "names" of any networked object, along with its properties (if it is software, what operating systems, etc.; if it is data, what kinds). The Lightweight Directory Access Protocol (LDAP, www.openldap.org) provides directory services, separately from Grid LDAP tools, and have been installed and populated with sample bioinformatics data, including preliminary data directories for the Bio-Mirror project [BM], directories of genome map and sequence information from the euGenes project [G19], and sample directories using SRS bio-data search and retrieval as a back end. Find the main test server at [ldap://iubio.bio.indiana.edu/](http://iubio.bio.indiana.edu/) which links to other services including a start at Bio-Mirror data packages, and euGenes genome data.

euGenes genome data sets provides a test case for trying LDAP directories of bio-data, to see if it will handle search and retrieval in a automated way. This evaluation is not finished; it seems like it may be an efficient, automatable method (e.g., ldap://eugen.es.org:3891/o=euGenes). It looks like one will be able to add gluttonous amounts of data to an LDAP lookup interface. SRS search and retrieval of GenBank, SwissProt and others may be a fairly simple addition. One can also add RDBMS or SQL databases as backends to LDAP directory systems.

Of notable interest, LDAP services can be federated in such a way that your search can transparently traverse any number of servers. Given cooperation with this among bioinformatics centers, one could search all the public bio-data in the world with one query.

BioGridRunner is an experimental distributed computing application for bioinformatics, incorporation directory services (data and software), grid computing methods (security, authentication, data transport and remote jobs), and gene sequence and genomic data processing methods. Figure 1 shows a set of views from this preliminary bio-grid client, including a directory of bioinformatics applications (top left), a directory of sequence data on local computer (middle left), an open sequence document (bottom left), and a dialog to align this sequence data using the ClustalW application (right). This application and the data can be run on the local computer or on a grid server, depending on the job host selected.

Figure 1. BioGridRunner views

This preliminary Java application for using bioinformatics software and data in a Grid environment will potentially aid bioscientists to better use high-performance distributed computing resources. Appendix 1 contains the manual for use of BioGridRunner. This program is built using the Globus Commodity Grid (CoG) toolkit [CG], along with parts of the SeqPup program, Readseq and related bioinformatics software. Public access to BioGridRunner is available for testing at <http://iubio.bio.indiana.edu/grid/runner/>.

Related work

SeqPup is a biosequence editor and analysis framework for molecular biology [GI6]. *Readseq* [GI2, GI8] is a widely used biosequence conversion tool and a component of SeqPup. *gnomap* is a whole-genome map display and discovery tool, used in the euGenes genome information project [GI9]. *Phylodendron* is a phylogenetic tree-drawing program [GI5]. These form part of the Java software framework in which the BioGridRunner application is developed, and are components used in its design.

Data and document translation methods in the current source class set include the Readseq selection of some 20 biosequence formats, including standard SAX and DOM interfaces to XML data structures. They include file and printing classes for standard

graphics formats of Postscript, GIF, and PDF, including hyperlinked versions of these for web-CGI use.

IUBio Archive [GI1] and the Bio-Mirror project [BM] form a backbone for bio-grid software and data distribution. Bio-Mirror, co-developed by the PI with Asian-Pacific partners, has been distributing public databanks around the globe since 1999, using high-speed network and daily mirroring of sources. It is currently serving on the order of a Terabyte of data each month to bioinformatics centers and labs, and is an official alternate source for EBI-EMBL and NCBI-GenBank databanks. IUBio Archive provides public bioinformatics software, as well as data search and retrieval with SRS [ET, ZL]. Table 1 lists the current contents and data sizes of Bio-Mirror. Many of these databanks change daily or weekly, so that timely distribution and use of current data in analyses is an important facet of bio grid methods.

Table 1 Bio-Mirror USA at IUBio, Databank status

70 Gigabytes (compressed) Sunday, 17 March 2002

Section	Mbytes	Updated	Databank source
blast	8053	06-Mar-2002	Biosequence databases for BLAST searches
blocks/data-blocks	31	12-Aug-2001	Highly conserved regions of proteins
blocks/data-prints	16	05-Jan-2002	PRINTS from NCBI
ddbj/daily_updates	3618	15-Mar-2002	DDBJ daily from AFFRC/bio-mirror
ddbj/regular_release	8563	18-Jan-2002	DNA Data Bank of Japan
embl/new	1316	15-Mar-2002	EMBL daily from EBI
embl/release	9824	12-Mar-2002	The EMBL Nucleotide Sequence Database
enzyme	2	23-Oct-2001	Enzyme nomenclature database
eugenes	184	16-Mar-2002	Eukaryote Genes Summary Databank
eugenes/arabidopsis	115	11-Mar-2002	Arabidopsis locus data from TAIR
eugenes/celegans	10	27-Dec-2001	C. elegans WormBase data from Sanger
eugenes/fish	1	10-Mar-2002	Zebrafish Genome data from ZFIN
eugenes/fly	40	15-Mar-2002	Drosophila Genes from FlyBase
eugenes/human	16	16-Mar-2002	LocusLink data from NCBI
eugenes/mouse	4	15-Mar-2002	Mouse Genome data from MGD
eugenes/yeast	1	15-Mar-2002	Saccharomyces genome data from SGD
genbank	22541	17-Mar-2002	GenBank Sequence Database
genbank/daily-nc	2648	17-Mar-2002	GenBank daily nc updates from NCBI
geneontology	91	15-Mar-2002	Vocabularies of gene functions and roles
interpro	56	15-Nov-2001	InterPro Protein databank
ncbigenomes	4875	16-Mar-2002	Whole genome sequence section of GenBank
pdb	8252	15-Mar-2002	Protein Data Bank of 3-D macromolecules
pfam	312	15-Mar-2002	Pfam database of protein domains and HMMs
pir/release	188	02-Mar-2002	NRL 3D Protein Sequence--Structure
prosite	7	07-Mar-2002	Database of protein families and domains
rebase	1	04-Feb-2002	The Restriction Enzyme Database
refseq/cumulative	671	16-Mar-2002	NCBI Reference Sequences
swissprot	67	07-Mar-2002	Annotated protein sequence database
swissprot/updates	3	07-Mar-2002	SwissProt new data from ExpASy
taxonomy/ebi	5	11-Mar-2002	Taxonomy data
taxonomy/ncbi	47	17-Mar-2002	Species names
trembl	190	15-Mar-2002	A supplement to SWISS-PROT
unigene	1078	15-Mar-2002	Unique Gene Sequence Collection

D. Research Design and Methods (R21)

This plan includes design of grid client software, of data directories, and grid server implementations for bioinformatics. This includes methods for information search and retrieval [KO] of data from public bioinformatics centers, integration of standard bioinformatics programs for analysis of sequence, genome, and gene expression information for use with large, genome-sized, and heterogeneous bio-data.

Bio-grid client design

Although this project includes methods for grid technology deployment and data distribution at bioinformatics server centers, the focus is on design and development of a client program useable by bioscientists for directing their data analysis needs. This client needs to include the following functions:

- computer and data resource allocation, including authentication for use of distributed computers, data transfer and job execution;
- user data management and transport among computer resources;
- search and retrieval of both data and analysis applications multiple bioinformatics services and archives;
- control of program commands, options for analyses, and management of input and output results, including automated joining or pipelining of multiple analyses.

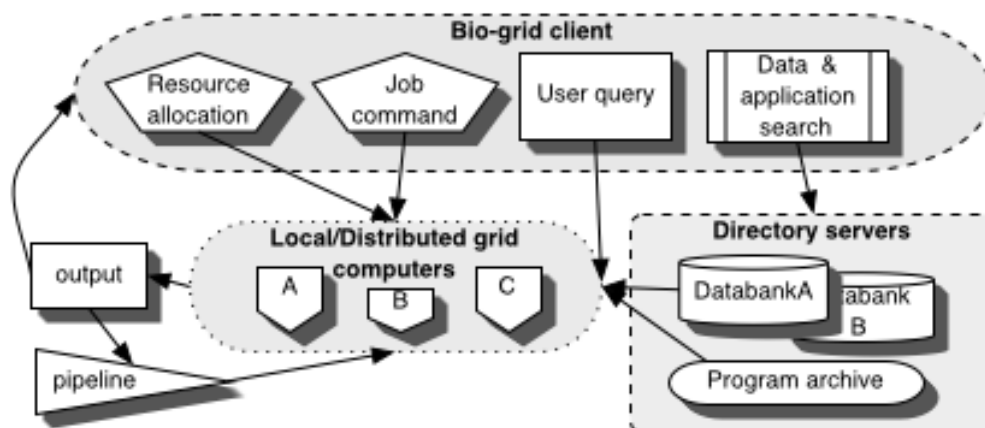
These functions are based on commodity grid toolkit methods (CoG of the Globus project), with extensions for specific needs of bioinformatics. The resource allocation methods of the CoG toolkit are used for digital authentication certificates, selection of grid computers, and for transport of data and messages.

Data and software location, search and retrieval methods are encapsulated in a directory interface design, where directories of resources are hierarchical views of items and their attributes. This is an interface familiar to most computer users for file system access, and can be generalized to networked resources, software and data resources, each with attributes indicating content, categories of information, dates and sizes. These attributes can be searched, in either restricted or global ways, to locate resources without the scientist knowing where they are located or their specific names.

Control of program commands is achieved with controlled descriptions of program functions, inputs, outputs and options. These descriptions are used to provide a program dialog to the bioscientist, who then selects in it the parameters and computing resources for use with the analysis. It collects input data from directory views, allows selection of options suitable to those inputs, and selection of options for computers to run the program on, and how to handle the program results.

A schematic of the proposed bio-grid client software and its interactions with computing and data resources is shown in Figure 2.

Figure 2. Bio-grid client-server architecture



Bio-data grid methods

Development of data-grid methods for bioinformatics will need to answer questions including

1. Do any of the current or proposed grid methods improve upon existing distribution among bioinformatics centers such as Bio-mirror project sites? The Globus GRID project uses FTP as its primary data transport, with improvements for parallel transfers, and replica (mirror) management.
2. Is there any current or proposed grid method to reduce need for redistribution of complete databanks, in favor of record-oriented data updates? Rather than transfer gigabyte-sized databanks, can we transfer only updated records, which may still be a major subset of the full databank? This assumes some kind of data record indexing and updating, with database software. The Globus Replica management methods fall in this category.
3. Are there agreeable standards for exchanging data records? Candidates include current flat-file bio-data formats, XML, ASN.1, CORBA and perhaps others.

ASN.1 – Developed for bioinformatics use at NCBI [NC] in 1990s, though not used much for bioinformatics outside of NCBI, it is used in several software industry Internet messaging protocols including LDAP and SSL. It has much the same flexibility as XML, and includes the precision of primitive data types and compression for efficiency.

CORBA - a software-to-software exchange that is efficient but has high software development costs, no storage format, and is useful for database exchanges.

XML – has wide software industry support and increasingly is of interest to bioinformatics. This is much less efficient in handling high-volume data exchange than ASN.1 or CORBA, but offers more human readability.

Each of these has positive and negative aspects; the main point is that bioinformatics groups haven't one common data exchange standard, and anything adopted for this

project should be flexible and multi-format capable. The idea behind (2) and (3) is that data grid methods for researcher-initiated analyses will need efficient data transport to compute servers; having these as part of bioinformatics center distribution such as for the Bio-Mirror project would be synergistic. Replication among centers and distribution to end-point applications both require efficient transport and can use the same component methods.

Data and resource directories

Another aspect of grid computing is automated resource discovery; how does a researcher specify where to pick up data for an analysis? It should be more automated than researcher specifying a URL -- he/she will, if he knows the data source at all, pick the source and ignore the mirror sites which can give more efficient transport.

A distributed directory of data resources is part of what we need. There are various standards and software now for such. LDAP (lightweight directory access protocol) has become one widely used (included in newer Unix systems and with open source packages; www.openldap.org), and is used in the Globus Grid package. Information searching thus becomes a search of distributed directories.

Methods for providing computable directories of bioinformation are a core component of a bioinformatics grid. LDAP provides a mature technology for this. It is capable of large-scale directory management, is flexible for implementing the range of data structures and sources from flat file thru relational database, object oriented data, and bioinformatics special search and retrieval tools such as SRS. In conjunction with Grid technologies for data location replication, which also use LDAP, this can provide the needed resource locating method for bio-grid deployment.

An alternative directory technology that will be investigated is the SDSC Storage Resource Broker [BA]. 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.

Data transport and resource allocation

This project will test the Globus methods for directory services (GIS) and data transport (GASS) on a set of bioinformatics data from the Bio-mirror project. The Globus data grid methods are a good starting point for Bio-mirror data, including:

- GridFTP -- A high-performance, secure, robust data transfer mechanism

- Globus Replica Catalog -- A mechanism for maintaining a catalog of dataset replicas.

- Globus Replica Management -- A mechanism that ties together the Replica Catalog and GridFTP technologies, allowing applications to create and manage replicas of large datasets.

Physics research projects are developing virtual data grid methods, in the GriPhyN project and the International Virtual-Data Grid Laboratory (iVDGL) [VD], for practical deployment and application of data grids. The methods being developed here will likely have application to bioinformatics data management, and the discovery phase of this project. The Grid Data Mirroring Package is a component of the Virtual Data Toolkit [VD] and Data Grid projects, with applications for bioinformatics data replication. The

Virtual Data Toolkit, currently in development, will form part of this project as it becomes practical.

This project will rely on commodity grid toolkit methods for data transport, along with authentication of resource use. The grid methods for resource authentication are based on electronically signed certificates produced by certificate authorities already widely used for HTTPS and other secure Internet information exchange. This method is preferable to passwords and other authentication mechanisms, in that certificates can be exchanged securely; they can be used to provide time-limited proxy certificates for short-term access to a resource, and are an available, tested technology.

Public release of project information

Publication of the current work is at <ftp://iubio.bio.indiana.edu/biogrid/>. This includes source code, documentation for bio-grid server, client and some standard biology applications for use with this, including NCBI BLAST, EMBOSS and others. Binary executables for various computer systems including MacOSX, Linux, Solaris will be made available for standard bioinformatics applications. To encourage collaborations with other bio-grid groups develop, other means for collaborative work will be added, such as shared CVS source repositories, possibly at <http://www.sourceforge.net/>

Milestones for R21 phase

1. Design the essential components and their integration for a bioinformatics grid client that will be usable by bioscientists.
2. Determine a usable technique, LDAP or other, for automated, globally searchable directories of bio-data. Incorporate needed directory components in the bio-grid client.
3. Determine grid protocols and tools that are needed for the bio-grid client and servers for use with bioinformatics resources. This determination will be based on the current and forthcoming protocols and interfaces supported by public data services.
4. Determine which test set of bioinformatics applications are suited for grid-based analyses without major modifications, to be feasible to support in this project.
5. Identify collaborative groups and projects to test and co-develop bio-grid methods planned for this project. Global, regional and institution collaborations are of interest to match expected uses bio-grid methods.
6. Publicly release a working prototype bio-grid client that works with supporting grid infrastructure. This public release is deemed essential for gaining feedback on its functionality from the bioscience community.

The preliminary design of a bio-grid client (1) to assess feasibility of this project has begun, but several aspects remain to finish before the R33 phase: The common directory access to distributed resources lacks functional searching and retrieval options. Integration of the components is limited and the preliminary design suffers from over-complexity. The command language and its processing in command dialogs is primitive and lacks bio-data context sensitivity (e.g., options differ for input of protein versus nucleic acid). Tests on the utility of LDAP and other techniques for bio-data directories (2) need to be made before the project can proceed to the next phase. Development of data directories is a new area for bioinformatics, and technology must be able to handle existing and forthcoming heterogeneous, high-volume data sets. Grid protocols and implementations are evolving rapidly in the academic and, more recently, commercial sectors of sciences and information technology. Successful completion of (3) requires a flexibility to work with current implementations of grid technology while keeping options open to incorporate forthcoming ones that may be better suited to bioinformatics uses. This project hopes to produce a bio-grid client that is functional for many uses, but will need to select a few test applications (4), based on wide utility plus feasibility of use in a grid-computing environment. Identifying collaborators (5) who will aid this project in a cooperative way by sharing infrastructure development and its bioinformatics applications is important for progress to the next phase, in order to produce work that will have widest possible utility in the biosciences. A necessary concomitant of this is public release (6) of working prototypes.

A. Specific aims (R33)

The extended project phase will build and deploy functional bioinformatics grid infrastructure and application software as determined from the preliminary phase. A data and computational grid will be established for public use at CGB bioinformatics public servers. Grid-aware application software (BioGridRunner) will be fully developed for effective use of this and other bio-grid infrastructure. Directories of bioinformatics resources with gene and genomic data, bioinformatics software, will be established and deployed for public use. Partnerships with other bioinformatics centers and science groups will be established for testing and deploying bio-grid technologies.

The grid-aware, bioscientist-friendly application designed in preliminary phase will be fully developed as the primary component of work. Ease of use by bioscientists is a fundamental criterion. While data grid methods offer much promise for enabling effective use of bioinformatics resources, they are at present complex and difficult to use. User interface methods that simplify this, such as consistent and familiar directory displays of distributed resources, and drag'n'drop methods for moving or linking these resources among directories, will be used. Flexibility to allow bioscientists to configure and add new analysis methods and resources will be included within this paradigm of directories. Standard formats for biosequences, phylogenetic and map data, for gene expression will be supported. The grid client will use standard network protocols that are extended for Grid computing (FTP, HTTP, LDAP, with security enhancements). It will provide for use of bioinformatics component tools, making it interoperable with other bioinformatics applications.

Directories of up-to-date data sources and mirrors will be an important addition to distributed use of bio-data, including the Bio-Mirror project [BM] for global data distribution. Such directories include source and replicate URLs of data, update times and contents, providing a method for automatically finding and computing the most efficient network transport for data to the user's computational resources. Directories of genomic databases are to be included in this, so that bioscientists can search and retrieve the most current reference genome data to use in genome analyses.

Specific test cases for grid data analyses, as determined to be most useful and feasible in the preliminary phase, will be deployed at the project site and with collaborators for public use. These likely include BLAST [AG] sequence databank matching, EMBOSS [RI] sequence analysis toolkit, and PHYLIP [FE] or related phylogenetic analyses. Methods for distributing data and these applications in a grid environment will be developed to be under control of the bioscientist through grid-aware client software.

Along with deployment of Grid infrastructure at the project site resources, we will examine the feasibility of deploying Grid infrastructure among computing resources shared by bioscientists, such as departmental or laboratory computers, to enable their use in bioinformatics analyses.

B. Background and Significance (R33)

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 some limitations in user interface and distributed use of resources.

In the field of physics, the individual scientist/lab needs to analyze terabyte to petabyte-sized data sets extracted from data centers [PG]. The emerging framework of distributed data and compute grids being developed for physics and other sciences applies well to bioinformatics. Basic data analysis services are available from bioinformatics centers, but there are limits to what they can provide to answer individual research questions. Large bio-data sets need to be copied frequently among computers to allow needed analyses. The task of keeping these data sets in synchrony with their sources such as NCBI and EBI is pushing network limits. The Bio-Mirror project distributes a core set of 150 gigabytes of data to several global bioinformatics centers, on the order of a terabyte per month. Several of these groups lack sufficient network bandwidth to update data sets regularly. The problem is compounded for regional and laboratory data collections where bioinformatics engineers are in shorter supply.

The European Union Data Grid project [EG] has a working group focused on biomedical applications [EB]. This group of 10 bioinformatics laboratories in Europe is working along lines similar to those proposed here, is a good prospect for collaboration. This group has several focal projects, involving integration of one or more grid technologies with particular bioscience research tasks. One of these is in the optimization of bioinformatics algorithms with grid-aware methods, not part of but applicable to this project. A second focus of this working group meshes well with the proposed work on data distribution in a grid environment. The Network Protein Sequence Analysis (NPS, <http://npsa-pbil.ibcp.fr/>) bioinformatics center is developing grid methods for BLAST, ClustalW and other data-intensive sequence analyses to use within their web-services environment. Their initial plan is to provide replicate databanks on each grid node of their service center for BLAST and similar data processing methods. The current project differs from this by moving away from a data and analysis center paradigm, to allow researcher-initiated selection and distribution of databanks. Where large databanks are needed for analyses, one option to be examined is the partitioning of databanks into smaller sets for distribution to compute nodes. This requires a merging of partitioned analyses, which is feasible for some applications such as sequence similarity searches and phylogenetic bootstrap analyses. A third focus of this working group is phylogenetic analyses. Their example case is with the HOVERGEN and HOBACGEN databases of vertebrate and bacterial genes [DP1, DP2]. These currently total 16 gigabytes of sequence data, requiring 4 weeks of CPU time for phylogenetic analyses, which a grid-based system can substantially improve on, to the point of enabling important analyses that are currently not feasible. This group is developing a client-server system called PhyloJava, with aspects similar to the planned BioGridRunner, and benefits for co-operative development between these projects are evident.

The available 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 [LB]. The current project will not address the issue of long-term maintenance of such directories, but seeks to test methodology for improved automation of biology resource directories.

The Pise project [LC] 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, in a way very similar to the command descriptions under development for this project. The W2H project [SE] relies on HTTPS for secure authenticated use of bioinformatics applications on central server computers, including both GCG and EMBOSS package configurations. It however is not easily extended to other bio-applications and is practically limited to single central servers.

The Embreo (<http://www.hgmp.mrc.ac.uk/~ptribble/Embreo/>) and Jembooss (<http://www.uk.embnet.org/Software/EMBOSS/Jembooss/index.html>) projects are developing alternate methods for distributed computing of bioinformatics applications, particularly for the EMBOSS package. These rely on the developing XML protocols including SOAP (Simple Object Access Protocol) and UDDI (Universal Description, Discovery and Integration). A public distributed genome annotation system (DAS, <http://biodas.org>) is under development among project members from WormBase, Ensembl, UCSC Human Genome Project, TIGR and FlyBase. DAS will offer reference genome annotation, with Internet client and server software, to provide bioscientists ready access to current genome maps, sequence and annotations. DAS, like Embreo, is a bioinformatics specific solution that applies to a specific data structure and content. A general technology like LDAP can provide similar and more extensive bio-data distribution methods.

These web server solutions do not address efficient data transport for high performance distributed computing with large volumes of data. Other differences from data grid infrastructure include server-centric rather than distributed, peer-to-peer computing model; the need to install and maintain specific bioinformatics-domain infrastructure; and a more limited security model for resource authentication and secure transport. New work to merge grid and web technologies such as XSOAP [GA] will provide interoperability between these technologies.

Data grid technology is specifically designed for high volume and high performance distributed computing. It is approaching the stage of being a commodity technology, with computer industry as well as academic support. It will be deployed widely in scientific and technology computing centers, departments and laboratories. It will not be necessary for bioinformatics centers to install and manage specific solutions, nor for bioscientists to rely on service centers to manage the computing resources. This is a version of peer-to-peer computing for sciences [GA], where research groups will be able to deploy grid

technology to lab or departmental resources, and/or use university or industry resources where grid technology is shared by multiple disciplines. The research group will be able to offer secure and authenticated use of data and compute resources to group members or public as needed.

Bioscientists are eclectic in their use of computer operating systems. MS Windows, Macintosh and Unix systems are all part of the mixture. Based on usage of biology-specific web servers (iubio.bio.indiana.edu, flybase.net, eugenes.org) during years 2001-2002, the logs of browser type show about 70% of these bioscientists use MS Windows systems, about 25% use Macintoshes, and a small set (2-5%) use a Unix system for web browsing, though many more use shared Unix systems for bioinformatics analyses. The Java language is important in allowing one to build biologist-friendly applications that work well on all of these systems.

C. Preliminary Studies (R33)

The R21 phase discusses and forms the basis for R33 phase preliminary studies.

D. Research Design and Methods (R33)

Data directories and transport

The basics of a data grid make sense for bio-data: (1) a catalog of data entries is used to locate records and logical groups of data for replication. The catalog itself can be replicated and contain information on replica-hosting servers. (2) FTP is used for data transport with improvements to use security, parallel data streams, third-party initiation, partial records and other efficiency measures. (3) Management software is used to make replication automatic and ensure that replica collections are up-to-date (replica collections are subsets from the source data for local needs).

Bio-data distribution projects such as Bio-Mirror should consider whether it is time use record-oriented data exchanges, and find commonly agreed on formats. A data record in bioinformatics is any ID/Accession record from the source data model (GenBank, EMBL, Swissprot, FlyBase, WormBase, etc.). The record-oriented exchanges have a second goal besides improving distribution efficiency: for client uses, one wants a data grid to be able to send just selected subsets, not the whole databank. The end-point is that analysis software has to get what it needs in data formats, but intermediate replication methods can be tuned to use efficient formats.

A general design for record-oriented data exchanges is

- (1) Create directory entries for each data record including fields for update time and ID/accession, databank division. At the initial stage, we want to keep it 'light-weight', and not put all relevant fields in such entries. In the long term this could be done, where the backend database for any directory system could be Entrez, SRS, or RDMBS software.
- (2) Directories or replica catalogs should be used for determining data to be mirrored or fetched, and the catalogs themselves can be replicated and exchanged.

(3) Local storage format is at the choice of the data center as long as the replication mechanism works, and the content needed by given analysis software can be generated at the endpoint.

(4) Provide a mechanism for locally reconstituting data in appropriate format for end-user software. E.g. for an SRS server, make flat file data sections, for BLAST, etc. make Fasta data records. This might be left to end-point software; that is a bio-mirror can store and exchange data in its transfer format.

For initial tests with the Bio-Mirror project, it isn't necessary to restructure source files, but we want to see if updating large databanks like GenBank via grid methods is more efficient than bulk flat file FTP. A next step would be seeing if this all improves efficiency of daily updates to end-use software such as SRS or BLAST. Whatever is used should be efficient enough to query and transport a large chunk of Gigabyte data sets in a timely manner. One should be able to ask for 2/3 the data records and transport them in no more than 2/3 the time it takes to fetch a full databank. There are over 15 million sequence records in GenBank release 128, totaling 60 Gigabytes, so directory service and record-oriented transfers need to be good.

Globus documents suggest it is capable of such efficient data transport. The methods from the Globus version 2 grid package rely on LDAP for data catalog search and retrieval, with replica catalog management software in development, and on an enhanced FTP server/client modified for parallel transfers, large TCP buffers, other efficiency and security methods [GD]. Work to develop virtual data grid methods is underway [VD] for physics and other sciences. These developments, including its proposed Virtual Data Toolkit, will likely form a basis for planned bio-data grid work. An eventual goal of this is to produce software meeting these design criteria that would be made available in a package suitable for use with other bio-applications.

Data grid methods

Methods for data exchange will be needed for use with grid data distribution, including data exchange for biosequences, map related, phylogenetic, and gene expression information. Standard bioinformatics formats including GenBank, EMBL, GFF, ASN.1, XML, tabular data and relational database SQL exchanges will be supported with existing components.

In the short term, the end-user analysis software will define what data model and format is needed. But it would better for long-term data integrity to have a common replica of a given data record, with reformatting or filtering done as needed at the end-point analysis. It may be good to move away from GenBank, EMBL and related flat formats for sequence data to something capable of mingling that data with other content, which XML and ASN.1 seem to be best at. Newer data analyses want richer content: parts of sequence, literature, structure, expression and phylogenetic combined.

ASN.1 is a potential common transport format because of its combination of flexibility, compressibility and specificity, as well as NCBI standard bio-data descriptions [NC]. Both XML and ASN.1 offer the flexibility of encapsulating and merging most bio-data

models. XML is a verbose format, which is a serious concern for efficient transport of high-volume data (an ASN.1 databank expands 10 times in XML format). XML as a choice is best for messages of small sizes where human readability aids its development, but it adds significant overhead to transport and processing of large volume data. ASN.1 supports several primitive data types such as integer and real numbers, Booleans, binary data and others, compared to XML's character data primitive. ASN.1 is a good fit with the LDAP, SSL and other network protocols that also use binary encoded ASN.1 as a transport structure. Jim Ostell suggests (personal communication) that NCBI may look at record-oriented bio-data access, to support for data caching at regional centers. They may find it easier to do this with ASN.1 format. The primary weakness of ASN.1 is limited access to software tools for its parsing and processing. A C/C++ language parser based on the NCBI-toolkit can serve for bioinformatics groups, but Java and Perl, language parsers for ASN.1 currently cannot fully handle the complexity of bio-data descriptions.

Data models differs based on perspective: a research question may be sequence-centric, or literature- or gene-expression-centric, and will need a different model, merging parts from source data models. Putting together the appropriate data model here is in the realm of end-point data analyses, but it impacts data-grid issues in that end-point software needs options for fine-grain selection from sources: just sequence ID fields from one set, abstracts or keywords from literature, etc. It remains to be determined where on a grid it is best to have fine-grain selection mechanisms - at source, end-point, and regional centers. This is one aspect of study for this project.

Bio-application command description

A command description language is an essential component for controlling a range of biology applications in both web and grid computing environments. The input, output and command options needed for running bio-applications must be described in a software-parsable way. This project uses an XML description suitable for controlling biology-specific applications, based on prior work of the PI with CORBA-based application command descriptions [GI3]. This command language works for a range of bio-software, including EMBOSS, NCBI tools, PHYLIP and other phylogenetic programs, Genetics Computer Group (GCG) software and others. Similarly, the Pise project uses an XML command description, EMBOSS project uses a command description called ADC, and GCG uses application descriptions for its SeqLab and SeqWeb user interface programs. One goal for this project is to work toward a common bio-application description language, which can be shared among projects.

Bio-grid client application

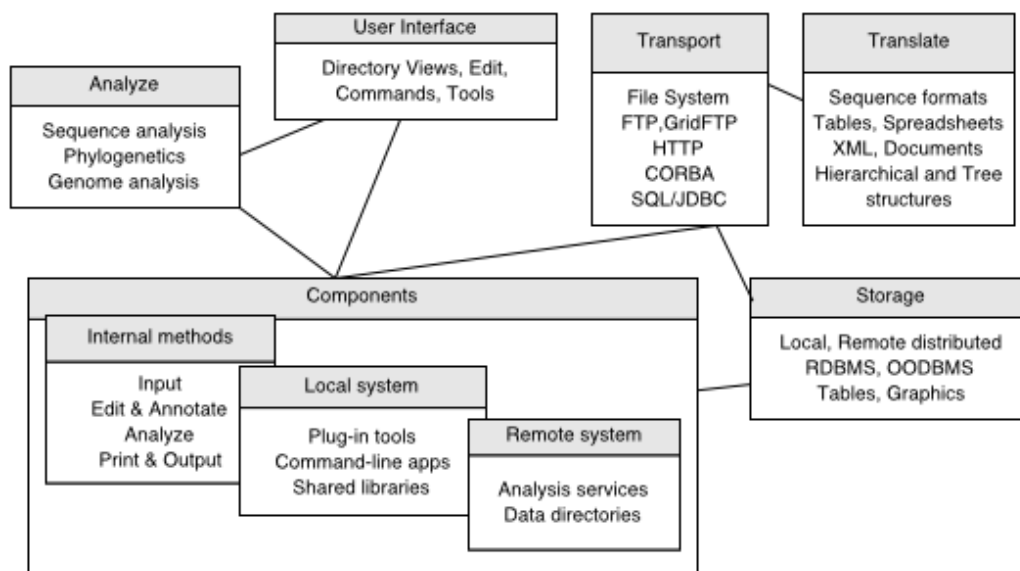
The bio-grid client software will be designed to be accessible to and usable by most bioscientists, through use of the Java language and standard library of user-interface, graphics, and network classes. Easy-to-use graphic user interfaces will be built with the Java Swing application framework. This grid client will employ data and analysis transports for interchangeable use of local workstation and networked computing resources. Component tools will be linked through a component plug-in architecture and network transport protocols for computing tasks. Configuration, selection and addition of

tools will be feasible for the bioscientist by using standard interfaces to local and networked compute and data services and applications. The components of a bio-grid client are outlined in Figure 3.

Underlying data and analysis methods is the transport system, for linking local and remote server information. Transports including file system, Internet protocols HTTP and FTP, LDAP for directory services, the database protocols SQL and JDBC, and structured data exchange of CORBA can be included at the transport level. These can be interchanged with a large degree of transparency to the components needing data, query and document transport. Between transport and component use is a format translation component for reading and writing data. For persistent storage, standard database, text and graphic data components will be used. Database interfaces for SQL, JDBC, XML, tabular and graphic types will be included where suitable.

The integration of frames or views is based in part on successful prior work with SeqPup and related software. As existing software engineering techniques and this prior work shows, an object-oriented class library of component methods for the range of application functions allows one to build applications that share many common features. This object-oriented structure also allows one to present and manipulate data using various distinct views of this data.

Figure 3. Bio-grid client framework



Recent work of the PI in the area of interfaces to bioinformatics tools shows that it is practical to support the range of user-interface (UI) methods useful to bioscientists in each tool. Thus, current versions of Readseq and Phylodendron include graphic UI, web (CGI) and command-line interfaces. A design lesson relevant to this multiple UI support is to construct abstract interfaces to the UI aspects of each tool, supporting methods such as linking, annotating and editing of data. Specific functions are implemented with

subclasses and helper classes. This makes it possible for the tool to support visual, interactive displays in the GUI instantiation, and to produce a range of web graphic maps in its command-line or CGI instantiation.

A Java class library of bioinformatics methods developed by the PI, which forms the code base of SeqPup [GI6], Phylodendron [GI5], Readseq [GI8], and gnomap will provide a starting point for development of the proposed framework. Table 2 lists main packages currently in this library and source size (7,100 Kb is about 3,500 printed pages). Publication of this source library is at <ftp://iubio.bio.indiana.edu/molbio/java/source/>

Table 2. Java class packages

Class package		Size (Kb)
iubio	biosequence application framework; genome maps; biosequence data objects (following packages)	4600
bioseq	Biosequence data model	364
chap	External child applications interface	512
directory	Directory services (LDAP, Files, JNDI)	560
drawseq	Biosequence drawing and genome maps	656
grid	Grid distributed computing methods	184
jni	Java Native application interfaces	40
readseq	Biosequence input/output	832
seqapp	Sequence manipulation and editing methods	868
swing	Java Swing GUI methods	376
views	Other GUI methods	168
dtree	phylogenetic trees	600
bopper	CORBA transport	600
flybase	utilities, graphics library	1300

Data and analysis transports for interchangeable use of local and networked computing resources will be built with an abstract transport component which can be designed from various structure object methods including Grid transport, CORBA transport [OH], Web (HTTP) and others.

Software engineering methods suited to large genome data sets will be used in this work. Methods for file system persistent storage [MY, OD] and efficient memory-to-disk integration for handling data strings of gigabyte sizes are available [WM]. Data summarization and indexing techniques for efficient display, searching and manipulation of large data sets are to be employed to allow efficient use of all appropriate data.

The planned component plug-in architecture uses a mixture of system process calls to command-line interface application, shared libraries built with a Java Native Interface (JNI) and remote method invocations. Use of Java's runtime class loader is a part of this, where plug-in tools are loaded as needed into the application framework, a method in common use for extending functions of commercial applications. Selection and addition of tools will be feasible for the bioscientist by using configuration dialogs.

Sequence of project (R33)

Year 1. Establish bioinformatics test grid

Develop bio-grid client to alpha stage. Develop directories of bioinformatics data and software for public sources. Test deploy grid infrastructure with bioinformatics applications. Initiate collaboration with partners identified in R21 phase.

Year 2. Demonstrate value of bioinformatics grid

Develop and document grid client to beta stage. Populate and use bioinformatics directories for Bio-Mirror data replication. Test and improve integration with bio-applications for grid use.

Year 3. Incorporate partnership bioinformatics centers and public access

Improvements for practical use with partner groups and working bioscientists, with revisions suggested by them. Improvements following from development of related works in genome informatics and science grid informatics. Final documentation, user help and full public release will be completed in year 3.

Public release of project information

All software, design methods, data and analysis protocols developed in this project will be made publicly available for free use by bioscientists and bioinformaticians working on other projects, as the PI's past works have been. Publication of the current work is at <ftp://iubio.bio.indiana.edu/biogrid/>. To encourage collaborations with other bio-grid groups develop, other means for collaborative work will be added, such as shared CVS source repositories, possibly at <http://www.sourceforge.net/>.

G. Literature Cited

- [AB] Abiteboul, S., P. Buneman, J. Gray, 1999. Data on the Web : From Relations to Semistructured Data and Xml. Morgan Kaufmann Publishers
- [AG] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410. <http://www.ncbi.nlm.nih.gov/BLAST>
- [BA] Baru, C., R. Moore, A. Rajasekar, M. Wan 1998. The SDSC Storage Resource Broker, Proc. CASCON'98 Conference, Toront; also <http://www.npaci.edu/DICE/SRB/>.
- [BM] Bio-Mirror project, 1997. A public service for distribution and access to biosequence and bioinformatics data, <http://www.bio-mirror.net/>
- [CF] 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/>
- [CG] 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. Also <http://www.globus.org/cog/>
- [DP1] Duret L., Perriere G. Gouy M. (1999) "HOVERGEN: comparative analysis of homologous vertebrate genes". In Bioinformatics: databases and systems, (ed. Letovsky S.I.), pp. 21-35. The Netherland: Kluwer Academic Publishers.
- [DP2] Perriere G., Duret L. Gouy M. (2000) "HOBACGEN: database system for comparative genomics in bacteria". Genome. Res. 10:379-385
- [EB] WP10 of European Union Data Grid Project, 2000. Grid-aware biomedical applications for datagrid testbed assessment. <http://marianne.in2p3.fr/datagrid/wp10/>, http://marianne.in2p3.fr/datagrid/wp10/documents/DataGrid-10-D10_2-0109-2-0.pdf
- [EG] The European Union Data Grid Project, <http://www.eu-datagrid.org/>
- [ET] Etzold, T, and Argos, P, 1993. SRS – an indexing and retrieval tool for flat file data libraries. *Comp. Appl. Biosci.*, **9** : 49-58.
- [FC] Foster, I., C. Kesselman, S. Tuecke, 2001. The anatomy of the Grid: enabling scalable virtual organizations. *Int'l J. Supercomputer Applications*, 15(3).
- [FE] Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package). Distributed by the author. Department of Genetics, University of Washington, Seattle. <http://evolution.genetics.washington.edu/phylip.html>
- [FK] Foster, I. & C. Kesselman, Eds, 1999. "The Grid, Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, Inc.
- [GA] Gannon, D. R. Bramley, G. Fox, et al. 2001. Programming the Grid: Distributed software components, P2P, and grid web services for scientific applications. GRID 2001, 2nd Intn'l Workshop on Grid Computing.. <http://grids.ucs.indiana.edu/ptliupages/publications/sc01gridgannon.pdf>
- [GI1] Gilbert, D., 1989. IUBio Archive for biology software and data. FTP and <http://iubio.bio.indiana.edu>
- [GI2] Gilbert, D., 1990. Readseq, a biosequence conversion tool. Bionet.Software, February 1990. Also in <http://iubio.bio.indiana.edu/soft/molbio/readseq/classic/>
- [GI3] Gilbert, D., 1996b. Bopper, a new Internet biocomputing office protocol for client-server data analysis. Bionet.software.gcg, July 1996 Also <ftp://iubio.bio.indiana.edu/util/dclap/source/bopper.tar.gz>

- [GI5] Gilbert, D.G., 1997, 1999. PhyloDendron, Java software for phylogenetic tree drawing. Bionet.Software, January 1997. See also <http://iubio.bio.indiana.edu/treeapp/>
- [GI6] Gilbert, D., 1998. SeqPup, biosequence editor & analysis platform. Bionet.Software, August 1998. Also <http://iubio.bio.indiana.edu/soft/molbio/seqpup/>
- [GI7] Gilbert, D., 1999a. 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>
- [GI8] Gilbert, D., 1999b. Readseq version 2, an improved biosequence conversion tool, written in the Java language. Bionet.Software, August 1999. Also in <http://iubio.bio.indiana.edu/soft/molbio/readseq/java/>
- [GI9] Gilbert, D.G., 2002. euGenes, a eukaryote organism genome information service. *Nucleic Acids Res.*, 30, 145-148 Also <http://iubio.bio.indiana.edu/eugenest/>
- [GD] Globus data-grid documents, 2001-2002. <http://www.globus.org/datagrid/software.html>,
<http://www.globus.org/datagrid/replica-management.html>,
<http://www.globus.org/datagrid/deliverables/ReplicaManagementService.pdf>
- [KO] Kowalski, G., 1997. Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers
- [LB] Letondal, C., S. Bortzmeyer, A. Thebault, I. Wang, 1999. Bio Netbook, <http://www.pasteur.fr/recherche/BNB/bnb-en.html>
- [LC] Letondal, C. 2000. PISE, a tool to generate Web interfaces for Molecular Biology programs. Pasteur Institute. <http://www-alt.pasteur.fr/~letondal/Pise/>
- [LE] Letovsky, S., (ed), 1999. Bioinformatics: Databases and Systems, Kluwer Academic: Boston.
- [MY] MySQL developers, 1999. MySQL, a Structured Query Language database server. <http://www.mysql.com/>
- [NC] NCBI data models, ASN.1 and XML. <http://www.ncbi.nih.gov/IEB/>,
<http://www.ncbi.nih.gov/IEB/ToolBox/XML/ncbixml.txt>
- [OD] Object Design, 1998. ObjectStore PSE/PSE Pro for Java. <http://www.objectdesign.com/>
- [OH] Orfali, R. and D. Harkey. 1997. Client/server programming with JAVA and CORBA. John Wiley & Sons, New York, NY. 657 pp.
- [OL] Overbeek, R. N. Larsen, N. Maltsev, G. D. Pusch, E. Selkov 1999. WIT/WIT2: Metabolic reconstruction systems. *in* Levotsky, Bioinformatics: Databases and Systems.
- [PG] The Particle Physics Data Grid (<http://www.ppdg.net/>); The Grid Physics Network (<http://www.griphyn.org/>)
- [RI] Rice, P. 2000. EMBOSS - The European Molecular Biology Open Software Suite <http://www.sanger.ac.uk/Software/EMBOSS/>
- [SE] Senger, M. 1996. W2H, WWW interface to sequence analysis software packages. <http://www.w2h.dkfz-heidelberg.de/>
- [VD] Virtual Data Grid projects, including GriPhyN, for partial physics, <http://www.griphyn.org/>; the International Virtual-Data Grid Laboratory (iVDGL), <http://www.ivdgl.org/>; Grid Data Mirroring Package (GDMP) <http://cmsdoc.cern.ch/cms/grid>
- [WM] Witten, I., A. Moffat, T. Bell, 1999. Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann Publishers
- [ZL] Zdobnov, E.M., R. Lopez, R. Apweiler, T. Etzold, 2002. The EBI SRS server – recent developments. *Bioinformatics*, 18, 368-373.