



# 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 00-2					<b>FOR NSF USE ONLY</b>	
<b>NSF 99-91</b>		<b>07/12/00</b>			<b>NSF PROPOSAL NUMBER</b>	
FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)					<b>0090782</b>	
<b>DBI - Database Activities</b>						
DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION	
				<b>006046700</b>		
EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)		SHOW PREVIOUS AWARD NO. IF THIS IS <input 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 ACRONYMS(S)		
<b>356001673</b>						
NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE			ADDRESS OF Awardee ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE			
<b>Indiana University Bloomington</b>			<b>Indiana University Bloomington</b>			
AWARDEE ORGANIZATION CODE (IF KNOWN)			<b>P.O. Box 1847</b>			
<b>0018093000</b>			<b>Bloomington, IN. 474021847</b>			
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.D.1 For Definitions) <input type="checkbox"/> FOR-PROFIT ORGANIZATION <input type="checkbox"/> SMALL BUSINESS <input type="checkbox"/> MINORITY BUSINESS <input type="checkbox"/> WOMAN-OWNED BUSINESS						
TITLE OF PROPOSED PROJECT <b>Integration, access and distribution of genomic and molecular bio-information</b>						
REQUESTED AMOUNT \$ <b>350,248</b>		PROPOSED DURATION (1-60 MONTHS) <b>36</b> months		REQUESTED STARTING DATE <b>01/01/01</b>		SHOW RELATED PREPROPOSAL NO., IF APPLICABLE
CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW						
<input type="checkbox"/> BEGINNING INVESTIGATOR (GPG 1.A.3) <input type="checkbox"/> VERTEBRATE ANIMALS (GPG II.D.12) IACUC App. Date _____						
<input type="checkbox"/> DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.D.1) <input type="checkbox"/> HUMAN SUBJECTS (GPG II.D.12)						
Exemption Subsection _____ or IRB App. Date _____						
<input type="checkbox"/> PROPRIETARY & PRIVILEGED INFORMATION (GPG II.D.10) <input type="checkbox"/> INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES _____						
<input type="checkbox"/> NATIONAL ENVIRONMENTAL POLICY ACT (GPG II.D.10) <input type="checkbox"/> FACILITATION FOR SCIENTISTS/ENGINEERS WITH DISABILITIES (GPG V.G.)						
<input type="checkbox"/> HISTORIC PLACES (GPG II.D.10) <input type="checkbox"/> RESEARCH OPPORTUNITY AWARD (GPG V.H)						
PI/PD DEPARTMENT <b>Department of Biology</b>			PI/PD POSTAL ADDRESS <b>1001 E. 3rd Street</b>			
PI/PD FAX NUMBER <b>812-855-6705</b>			<b>Bloomington, IN 47405</b>			
			<b>United States</b>			
NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address		
PI/PD NAME <b>Donald G Gilbert</b>	<b>phd</b>	<b>1981</b>	<b>812-855-0587</b>	<b>gilbertd@bio.indiana.edu</b>		
CO-PI/PD						
CO-PI/PD						
CO-PI/PD						
CO-PI/PD						

## **PROJECT SUMMARY**

Molecular biology and genomic sciences are dependent on bioinformatics for access to new data, analyses and knowledge bases. Hundreds of gigabytes of bio-information are published through Internet services. It is difficult for bioscientists to locate such sources, which often change, with answers relevant to their questions.

For example, where does one find current detailed information on a gene's function? A full, relevant answer can include links to research lab web pages, sequence and other databank reports, literature abstracts and journal articles, organism databases, metabolic pathway and gene expression data.

The need is growing for summarized, categorized bioscience knowledge, with up-to-date links to sources, produced in an automated fashion. Existing research in Internet information systems and knowledge integration suggests ways to produce this. The proposed research will employ these principles and tools to extend integration of wide ranging public bio-information.

IUBio Archive has a ten-year record of providing public access to molecular biology data and software, serving a worldwide community. The Bio-Mirror project is an international collaboration for rapid public access to such data. This project will extend IUBio Archive and Bio-Mirror roles in bioinformatics knowledge access and distribution.

## TABLE OF CONTENTS

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

Section	Total No. of Pages in Section	Page No.* (Optional)*
Cover Sheet (NSF Form 1207) (Submit Page 2 with original proposal only)		
A Project Summary (not to exceed 1 page)	1	_____
B Table of Contents (NSF Form 1359)	1	_____
C Project Description (plus Results from Prior NSF Support) (not to exceed 15 pages) <b>(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	10	_____
D References Cited	3	_____
E Biographical Sketches (Not to exceed 2 pages each)	2	_____
F Budget (NSF Form 1030, plus up to 3 pages of budget justification)	8	_____
G Current and Pending Support (NSF Form 1239)	1	_____
H Facilities, Equipment and Other Resources (NSF Form 1363)	1	_____
I Special Information/Supplementary Documentation	4	_____
J Appendix (List below. ) <b>(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	_____	_____
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

Internet information services are now a primary means for bioscientists to solve problems, learn new concepts and details, and analyze their data in many areas of biology, especially in the genomic sciences whose central data is the rapidly unraveling gene code. Integration of a wide range of large and diverse public Internet information sources for molecular biosciences can be improved, with a goal of providing simpler, more comprehensive access to up-to-date knowledge for bioscientists.

The range of primary public data in molecular biosciences has expanded well beyond the sequence databanks, to include genomic, polymorphism, gene expression, metabolic, phylogenetic, literature and other classifications. Increasingly, a large body of relevant information originates at bioinformatics centers and investigator labs, found on Internet servers at educational, industry, government and other locations. Much of this latter is as yet poorly integrated with bulk databanks.

What bio-information should be integrated for scientific uses? Consider the heterogeneity of public data available from NCBI, EBI, genome sequencing and bioinformatics centers, genome databases, personal or laboratory web pages, and publications. This project frames the answer in terms of what a scientist wants or expects as answers to biological questions. For example, where does one find information on particular genes, their function, and expression? A full, relevant answer will include research laboratory and summary web pages, sequence and related data bank entries <ref Baxevanis, A., 1998>, PubMed references, on-line journal articles, organism databases (e.g. FlyBase, WormBase, Mouse Genome Database), metabolic pathways (e.g., KEGG <ref Kanehisa 1999>, WIT <ref Overbeek *et al.* 1999>), gene expression data, and more.

Human curated knowledge summaries, commoner in biosciences than automated summaries, are prone decline in usefulness, from lack of personnel time. Internet search engines (e.g., AltaVista <ref Digital Equipment Corp 1995>, Googlebot, Infoseek) and classification systems (e.g., Yahoo <ref Filo and Yang, 1994>) now form a basis for general knowledge discovery, for bioscientists along with the general public. But these waste a scientist's time sifting through much information that is irrelevant to their questions. A helpful advancement for the biology community is the use of search robots such as MedHunt <ref HON Foundation, 1999> that are targeted to biosciences content.

General strategies for focused web-crawling <ref Chakrabarti *et al.*, 1998, 1999 > can produce relevant, category-oriented collections in an automated way. Such topical collections provide a way of finding answers more relevant to the specialized questions of bioscientists. Focused web collecting should be an important component of building a bioscience-oriented knowledge base.

Data structures of this bio-information universe are heterogeneous and often fuzzy or weakly structured, so that text and concept analysis methods from information systems and digital library research can be more useful than relational database methods for well-structured data. Text analysis methods used in information retrieval (IR) <ref Kowolski, 1997> and digital library research, provide a basic level of integration of these semi-structured, heterogeneous databanks and Internet documents. These methods can be used to produce dictionaries of bioscience vocabularies, which are then used as building blocks for a more integrated knowledge discovery system.

These text analysis, including data parsing to produce dictionaries of terms and phrases, can handle the very large data sets involved (over 30 Gigabytes for current sequence databanks), and offer rapid searches based in inverted indices of the dictionaries. These methods offer commonly used search methods including data field specific queries, Boolean operator refinements, phrase and fuzzy search methods, and word stemming. Such allows fairly sophisticated queries of a wide range of data, but is dependent on the specific vocabulary sets of source data. The Sequence Retrieval System (SRS, <ref Etzold and Argos 1993, Carter *et al.*, 1999>) is one such text analysis system, and it extends the above methods to include databank linking, analogous to relational database table join selections, which permits one to formulate queries across multiple databanks. The SRS system also has the advantages of wide usage at bioinformatics centers and a large number of protocols for parsing hundreds of bioinformatics data sets.

Text analyses however are not enough to provide categorizations based on ontologies of biology concepts and their relationships. Knowledge integration of available bioscience data is necessary to provide this useful categorization. The Kleisli project <ref Davidson *et al.*, 1999> and TAMBIS project <ref Baker *et al.* 1998, 1999> offer much promise for integration at a higher level.

These analyses will produce *metadata* describing the contents of source data and collections of data in an integrated fashion. Answers to questions put to this metadata are *where* full information is on the Internet, with a summary, title, keywords and placement in concept categories. This metadata will be provided to the public in categorized and searchable fashion, and in bulk and subsets that may prove useful to other bioinformatics services. Biologists using it will be able to answer questions such as "What Internet resources dealing with protein function have appeared or changed since I last looked, ranked by importance to my research community?" It may also help with answers to more detailed questions such as "Which resources can tell me about genes in selected organisms that encode a beta-catenin protein expressed in the nervous system?"

## II. Research plan

This project seeks to extend research in information retrieval systems for biosciences, and extend public services for use of important biology data and software, through IUBio Archive and Bio-Mirrors. Integration of summary knowledge from a wide range of available molecular biosciences public data is a focus of this research. It will build on existing bioinformatics and information technology research and methods, seeking a further integration and summarization

of knowledge in the areas. Methods for improved user access and user analyses of such integrated data will be developed. Methods for improved data sharing among world bioinformatics centers will be tested.

This project has three related sections:

- **Bioinformatics tools and services.** Extend and improve a public molecular biology software tool archive and web-based services at IUBio Archive, with applications to Bio-Mirror and other bioinformatics centers.
- **Bio-information warehousing and distribution.** Extend and investigate improved methods for biosciences data-warehousing, re-distribution and user search services, in collaboration with the Bio-Mirror project.
- **Bioinformatics knowledge integration.** Research areas of knowledge integration for a broad spectrum of this molecular bioscience information, including automated Internet collecting and categorization.

## IUBio Archive

IUBio Archive for biology data and software <url <http://iubio.bio.indiana.edu/>> has been an important Internet information resource for biologists since 1989 <ref Baxeavanis and Ouellete 1998>. The PI's early work includes pioneering efforts to provide interactive searches of biosequence databanks over the Internet, including a graphic-user interface Internet search client <ref Gilbert 1990> and a WAIS-based search server <ref Gilbert 1992, 1993> at IUBio. Public software available at IUBio is focused on molecular biology, with sections for alignment, codon, browsing, consensus, phylogenetics, pattern matching, primer selection, restriction enzymes, RNA structure, searching, and platform-specific programs <ref Gilbert 1999>. This large collection of publicly available software is a popular source of analysis tools for bioscientists.

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. As part of this project, new tools of importance to molecular biosciences will be added to IUBio user services. These include EMBOSS <ref Rice 2000>, a developing, comprehensive, open-source package of sequence analysis tools; Pasteur Institute Software Environment (PISE <ref Letondal, 2000>), a framework for integrating bioscience analysis tools into a web server; phylogenetic tools, based on popularity of Phylodendron tree drawing. Other such tools will be added pending review and feasibility for available hardware and databases at IUBio.

IUBio Archive has for several years provided a US public site <ref Gilbert 1995> for the Sequence Retrieval System. 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 simple sense, many related databanks, as well as offering automated up to date access to these data.

An archive of the Bionet public news articles comprises a large and widely used section of this archive. Besides serving the biosciences community, this news archive is a source of science information well 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.

Recently expanded hardware capability includes 300 Gigabyte of RAID storage and a 4 CPU Sun Enterprise server, funded by NSF Grant DBI-9982851 and the Indiana University High Performance Network Project. Currently about 100 GB of biology data and software are stored, available as part of Bio-Mirror (30 GB), IUBio services (60 GB), euGenes eukaryote genes database (5 GB), and software and Bionet news archives (3 GB). Further storage and processor expansion, as well as student and data processing workstations for use with this research will be sought through equipment grants from Sun Microsystems and other vendors interested in public bioinformatics projects.

The archive is currently managed through time volunteered by the PI and those who contribute software and data. The popularity of IUBio archive has at times strained this volunteer effort. This proposal seeks salary support in part to provide time for needed improvements of IUBio archive.

## **Bio-Mirror project**

In 1998, IUBio became a central data-warehouse for the Bio-Mirror project for high-speed worldwide mirroring of public bioinformatics data using Internet2 infrastructure <url <http://www.bio-mirror.net/>>. The Bio-Mirror project is a worldwide bioinformatics public service for high-speed access to up to date DNA and protein biological sequence databanks. In genome research, public data sets have been growing tremendously; so much so that distribution is hampered by existing Internet speeds. The Bio-Mirror project is devoted to facilitate timely access to important large data sets for this research.

To facilitate timely distribution and access to these data, this project will investigate improved methods for data exchange. The Internet2 Distributed Storage Infrastructure (I2-DSI, <url <http://dsi.internet2.edu/>>) project may offer such methods. DSI methods can improve network throughput using bi-directional and multidirectional methods, intelligent data subset selection and compression, with tools such as rsync+ and multicast distribution <ref Dempsey and Weiss, 1999>.

The Bio-Mirror project has been developed with colleagues Y. Ugawa <ref Ugawa 1997>, 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, and the USA, with a Malaysia node expected soon. Other centers are being encouraged to join this project. Indiana University's high performance network infrastructure and collaborative help been essential to this project, including TransPAC (Trans-Pacific network), and Asia-Pacific Advanced Network (APAN) connections.

Many international APBionet project members are part of new, growing bioinformatics centers. Providing user services of software and server tools for analyses of these data sets is part of the

need at these centers. The proposed work for IUBio server tool extensions will be done in a portable manner, available to other bioinformatics centers.

Data in the Bio-Mirrors project currently encompasses 30 Gigabytes (compressed) updated from the primary sources nightly. Contents include DNA and protein biosequence and related databanks, genome and model organism data (Table 1). These are mirrored from originating sites in the US, UK, Switzerland, and Japan. Planned additions include the protein structure databank (PDB), Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways, and others.

**Table 1. Bio-Mirrors data sets**

Databank	Description	Home site
BlastDB	Biosequence databases for BLAST searches	NCBI
Blocks	Highly conserved regions of proteins	NCBI
DDBJ	DNA Data Bank of Japan	NIG
EMBL	The EMBL Nucleotide Sequence Database	EBI
Enzyme	Enzyme nomenclature database	ExPASy
GenBank	GenBank Sequence Database	NCBI
Genomes	Whole genome sequence section of GenBank	NCBI
InterPro	InterPro Protein databank	EBI
PIR	Protein Information Resource	NBRF
Pfam	The Pfam database of protein domains and HMMs	WUSTL
Prosite	Database of protein families and domains	ExPASy
Rebase	The Restriction Enzyme Database	NEB
RefSeq	NCBI Reference Sequences	NCBI
SRS Databanks	List of active SRS databases around world	EBI
SWISS-PROT	Annotated protein sequence database	ExPASy
Taxonomy	Species names	NCBI
TrEMBL	A supplement to SWISS-PROT	EBI
UniGene	Unique gene sequence collection	NCBI
euGenes	Eukaryote Genes Summary Databank	IUBio

## Knowledge integration for biosciences

This project has a general goal of examining the possibility of collecting and integrating all Internet published bioscience information in a knowledge base, for use of this community. That this possibility is feasible is based on current research outlined in the Introduction, and the practical success of "all-the-Internet" search services.

Methods for automated information collection are available, and have been demonstrated to work for the large universe of all published Internet documents. The kinds of integration from all Internet sources is necessarily very simple compared to what can be done with a focused subset such biosciences documents. This project will expand its knowledge integration by



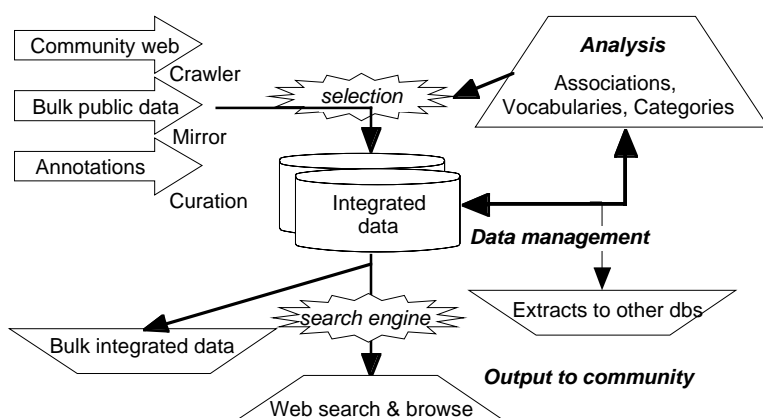
basing it on existing biosciences knowledge domains and ontologies, and providing an extensive analysis of collected documents to assign relevance and categories to these domains.

Bioscience documents will be gathered automatically using Internet robots. The integrated data and analyses will be packaged for output to public search and use in several ways. Automated gathering of new documents and URL links, as well as and public searches and output will rely on existing methods and tools. The analysis phase will impact on gathering as well as output. An overview of the bio-information integration process is shown in Figure 1.

An initially restricted scope is planned as test beds for this research. Two subsets of Internet publications will be examined:

- bioinformatics software and tools information
- eukaryotic organism genome information (gene function and structure, expression data, metabolic relations, map information, literature, web documents)

These subsets are chosen by virtue of wide interest to biologists, and as they are areas of expertise and related research by the PI. Extension of this research to a wider set of bioscience information will be possible given success with these subsets.



**Figure 1. Knowledge integration overview**

Data sources for the project include:

- Curated or man-made annotations relevant to biosciences, including dictionaries, vocabularies, thesauri, ontologies, concept documents, and URL lists. These will be used for primary weighting and selection of automatically collected documents and document servers.
- A dictionary of non-biological terms generated

- or drawn from other sources, for negative weightings, URL screening and stoplists.
- Bulk databanks with relevant documentation (see Table 1, plus PubMed, PDB and others).
- Internet-collected documents, with URL lists for crawling to be primed from lists of bioscience servers and potential servers from a broad list, including most .edu domains, and selected .gov, .intl, and .com host domains.
- URL lists for crawling will be supplemented through queries, formed from dictionaries from the integrated data, to "all-the-web" search services.

The analysis will be designed to produce an integrated dictionary and concept hierarchy of biology information. Bulk data parsing and indexing to dictionaries will use available SRS parsers and methods, as this package includes parsing patterns for a large set of the important bioinformatics data sets of interest.

Document relevance to categories of bioscience knowledge will be determined from:

- a weighted sum of scores for terms and phrases matching existing dictionary entries;
- a score for biological concepts, based on word phrase analysis in relation to other documents scored for concepts;
- a score in a URL linkage matrix with other documents <ref Brin and Page, 1998, Kowalski, 1997>.

Dictionary entries have a confidence level depending on their source. Curated entries have high confidence, bulk bioscience data medium, and automatically collected data have lower confidence. Methods outlined in <ref Brin and Page, 1998> for document weighting based on a linkage matrix and other criteria will be employed.

Analysis needs to promote classifications and terms up and down linkage (web server) document hierarchy; e.g. a general question "where can I find *C. elegans* genes?" should rank the WormBase main page highly over single page with these terms.

Standard text analysis methods <ref Kowalski 1997, Klein 1999>, including misspelling and stemming dictionaries will be employed. Concept indexing and categorization <ref Kowalski 1997> will be used to map from terms to concepts, starting with concepts from curated data, then to categorize documents in concept hierarchies. Factor analysis of the term-document matrix will be used to provide such concept indexing <ref Chakrabarti *et al.* 1998>. The document processing will develop matrices of document term relevance, concept relevance, Internet linkage. These will be subject to correlation analysis and data mining methods <ref Chakrabarti *et al.* 1998; Witten and Frank, 1999>.

Existing software components will be used as much as possible, and design flexibility for changing component instances in the protocol. Components that have been used or reviewed for possible use include

- Web crawlers: httdig <url <http://www.httdig.org/>>, libwww robot framework <url <http://www.w3.org/Robot/>>, a focused crawler <ref IBM, 1999>, Harvest <url <http://www.tardis.ed.ac.uk/harvest/>>, Googlebot <ref Brin and Page, 1998> and others.
- Indexers: glimpse <ref Manber and Wu, 1994>, essence, Isearch <ref Isearch developers, 1996>, FreeWais, SRS
- Analyzers: R statistics <ref R developers, 2000>, text analysis tools <ref Klein 1999>, data mining tools <ref Witten and Frank, 1999; kdnuggets 2000>, data management tools <ref MySQL developers, 1999, PostgreSQL developers, 1999>.

Results of the analysis will be used to feed back to the web-gathering phase, to screen and weight new documents. The results will provide relevance feedback for public searching and browsing of documents, and will be used in categorizing documents and links for user selection.

The primary outcome of this research will be metadata describing the contents of source data and collections of data in an integrated fashion. This metadata will be provided to the public in categorized and searchable fashion, using an XML architecture suitable for query processing <ref Abiteboul *et al.*, 2000>, and in bulk and subsets that may prove useful to other bioinformatics services.

This work will be designed to be self-maintaining and updating, over a period of years with minimal effort. Integrated data and software will be open source, freely available, for customizing, maintenance and longevity as an adaptable community resource. Should it prove to be of wide interest, commercial licenses for use can be used to fund its continued maintenance.

### III. Relation to present state of field

EMBNNet <url <http://www.embnnet.org/>> is a network of primarily European bioinformatics centers which share data and tools in a fashion similar to the developing Bio-Mirror project. The work planned in this proposal for is more focused on improved data and tool distribution methods than the general bioinformatics collaboration exemplified by EMBNNet. The Canadian Bioinformatics Resource (CBR, <url <http://www.cbr.nrc.ca/>>) is an extensive bioinformatics service that is a model for some of the proposed expansion of services for IUBio Archive. The BioCatalog <ref Rodriguez-Tom 1998> is a summary of publicly available software tools for molecular biosciences, similar in concept to the tool cataloging efforts at IUBio archive, but produced by curation rather than automated collection.

Amos Bairoch's links <url <http://www.expasy.ch/alinks.html>> and BioWorld <url <http://search.ebi.ac.uk:8888/compass>> are curated collections of links to categories of Internet bio-information, similar in aspects to the proposed knowledge integration. These have special value by virtue of the curation and editing effort provided by the expert knowledge of their authors. They do not provide as comprehensive nor current information as is possible with automated tools.

GeneCards <ref 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 proposed work, but more specialized in its area of genome information.

Experiments on producing rich link databases <ref Achard and Dessen 1998, Achard *et al.* 1998> or meta-databases <ref Cheung *et al.* 1998> for biosciences are similar in concept to the proposed knowledge integration phase. Marvin and BioHunt <ref HON Foundation, 1999> provides an example of a web robot focused on biosciences topics. The TAMBIS project <ref Baker *et al.* 1998, 1999; Paton *et al.* 1999> is developing methods for categorizing diverse bio-information using ontologies. The Kleisli project <ref Davidson *et al.*, 1999> is a knowledge base project for bio-information. The approach here is to provide a uniform, comprehensive structure for integrating heterogeneous databases.

### IV. Relation to other work of PI

A bioinformatics project recently undertaken by the PI is a genome summary service for eukaryote organisms, *euGenes* <url <http://iubio.bio.indiana.edu/eugenesis/>>. This project's goal is

the next step beyond single organism databases <ref Model Eukaryotic Organism Workshop, 1998>, providing an integrated set of tools and data for bioscientists to learn of the essential knowledge found in genome databases. EuGenes is automated, collecting data from relevant organism data servers, analyzing and presenting it, in a fashion similar to that of proposed work. It provides one integrated view of eukaryote gene information, including summaries and analyses not available in other services, such as BLAST-measured gene associations, a genome map view locating genes on organism's sequence map. It is designed to be portable to other sites, and builds on successful FlyBase and other technologies. The proposed project can provide an important means for augmenting this model organism summary information.

As part of the FlyBase project <ref FlyBase Consortium, 1999>, the PI has designed and developed much of the current portable, distributed web service for public access to this model organism data. Developing software acceptable to, easily used by, and widely available to bioscientists, as well as employment of Internet client-server methods to bioinformatics, have been continuing research goals of this PI <ref Gilbert 1989, 1990 *et seq.*>. This project builds on these goals to produce an expanded information service for bioscientists.

## **V. Outcomes, publication and distribution**

Outcomes of new methods for biology data federation and improved use of this integrated knowledge for bioscientists are expected. Extended bioinformatics software archive and 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 will be an outcome. A metadata knowledge base of integrated bio-information will be published for use at IUBio and other bioinformatics projects. Future extensions include extracting subsets of metadata relevant to specific bioscience community databases. Software, data and documentation produced will be made publicly available, in electronic and traditional publications read by bioscientists and bioinformaticians.

## **VI. Broader impacts**

Students in bioinformatics will help in this project, as graduate research assistants and for class projects. This research is associated with the Indiana University School of Informatics <url <http://informatics.indiana.edu/>>, which has a goal to educate students in bioinformatics, in line with national and industry goals of teaching and research in informatics.

This project is a public-oriented information service for all bioscientists, though the specific focus is on genome-related sciences. It will improve the integrated understanding of, and access to, a range of bioscience information for average biologists, the general public, and for bioinformaticians working closely with large data sets. The results will be disseminated broadly to enhance scientific and technological understanding. All integrated data, methods and software developed to produce it, will be made available freely to other informaticians.

This project will enhance the infrastructure of bioinformatics in the US and worldwide, as a data warehouse connected with high-speed Internet2 network infrastructure. It will enhance partnerships with Asian-Pacific 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.

## **VII. Timeline**

Year 1. IUBio server software archive and tools enhancements. Investigations with Bio-Mirror partners in rsync distribution of data. Preliminary robotic collection and analyses of Internet bio-information.

Year 2. Extended detailed analyses of bio-information, with preliminary public access and distribution. Continued enhancements to IUBio server tools, and Bio-Mirror distribution methods. Distribution, documentation and implementation help for bioinformatics server tools to Bio-Mirror and other bioinformatics centers.

Year 3. Refinements to bio-information collection and public services, with wide testing by bioscientists. Extraction of integrated bio-information for use with euGenes, genome databases, and other projects. Continued enhancements to IUBio server tools and their re-distribution to other centers.

## REFERENCES CITED

---

- Abiteboul, S., P. Buneman, J. Gray, 1999. *Data on the Web : From Relations to Semistructured Data and Xml*. Morgan Kaufmann Publishers
- Achard, F and P. Dessen, 1998. GenXref VI: automatic generation of links between two heterogeneous databases. *Bioinformatics*, 14: 20-24.
- Achard, F, C. Cussat-Blanc, E. Viara and E. Barillot, 1998. The new Virgil database: a service of rich links. *Bioinformatics*, 14: 342-348. Also <http://www.infobiogen.fr/services/virgil/>
- Baker, P.G., A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, 1998. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *in Proc. 6th Intl. Conf. on Intelligent Systems for Molecular Biology, ISMB98, Montreal*. <http://img.cs.man.ac.uk/-tambis/papers/TAMBIS-ISMB98.zip>
- Baker, P.G., C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, A. Brass, 1999. An ontology for bioinformatics applications. *Bioinformatics*, 15: 510-520.
- Baxeavanis, 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.
- Baxeavanis, A., 1998. Information retrieval from biological databases. *in Bioinformatics: A practical guide to the analysis of genes and proteins* (A. Baxeavanis and B.F.F. Ouellete, eds). pp.98-120. J. Wiley and Sons, Inc., NY.
- Brin, S. and L Page, 1998. The anatomy of a large-scale hypertextual Web search engine. *in Proc. of the 7th WWW Conf.*, pages 107—117. <http://www7.scu.edu.au/programme/-fullpapers/1921/com1921.htm>
- Chakrabarti, S, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, 1998. Automatic Resource by Analyzing Hyperlink Structure and Associated Text. *in Proc. of the 7th WWW Conf.*. <http://www7.scu.edu.au/programme/fullpapers/1898/com1898.html>
- Chakrabarti, S., B. Dom, R. Agrawal, P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB Journal*, 1998. [http://www.almaden.ibm.com/cs/k53/irpapers/VLDB54\\_3.PDF](http://www.almaden.ibm.com/cs/k53/irpapers/VLDB54_3.PDF)
- Chakrabarti, S., M. Van den Berg, B. Dom, 1999. Focused crawling: a new approach to topic specific resource discovery 8th WWW conf.. <http://www.cs.berkeley.edu/~soumen/-doc/www99focus/html/>
- Cheung, K.-H., P. Nadkarni and D.-G. Shin, 1998. A metadata approach to query interoperation between molecular biology databases. *Bioinformatics*, 14 (6): 486-497.
- Dempsey, B and D. Weiss, 1999. Towards An Efficient, Scalable Replication Mechanism for the I2-DSI Project, UNC SILS, Technical Report TR-1999-01 <http://www.ils.unc.edu/ils/-research/reports/TR-1999-01.pdf>.
- Digital Equipment Corp., 1995. AltaVista Internet search service. <http://www.altavista.com/>
- Etzold, T and P. Argos, 1993. SRS — an indexing and retrieval tool for flat file data libraries. *Comp. Appl. Biosci.*, 9: 49-58.

- Carter, P., T. Coupaye, D. Kreil, T. Etzold, 1999. SRS: Analyzing and using data from heterogenous textual databanks. *In* Bioinformatics: Databases and Systems by Letovsky, Stanley, Ed.; Kluwer Academic: Boston, 1999
- Filo, D. and J. Yang, 1994. Yahoo!, Yet Another Hierarchical Official Oracle.  
<http://www.yahoo.com/>
- Frishman, D., K. Heumann, A. Lesk and H.-W. Mewes, 1998. Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics*, 14(7): 551-561.
- FlyBase Consortium, 1999. The FlyBase database of the Drosophila genome projects and community literature. *Nuc. Acid Res.*, 27:85-88. Also <http://flybase.bio.indiana.edu/>
- Gilbert, D., 1989. IUBio Archive for biology software and data. ftp, <http://iubio.bio.indiana.edu>
- Gilbert, D., 1990. GenBank Search, a Hypercard stack for network searching and fetching sequences from the GenBank e-mail server. October 1990.  
<ftp://iubio.bio.indiana.edu/molbio/mac/gbsearch-ncbi.hqx>
- Gilbert, D., 1992. Booleans, partial words and other WAIS mods in biology data searches. *Bionet.Software*, Nov. 1992. Also <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. Also <http://iubio.bio.indiana.edu/srs/srsc>
- Gilbert, D., 1996. Portable FlyBase server available for Drosophila data. *Bionet.Announce*, April 1996 Also <http://flybase.bio.indiana.edu/docs/Portable-server/>
- 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>
- HON Foundation, 1999. MARVIN, a multi-agent retrieval vagabond on information networks.  
<http://www.hon.ch/MedHunt/Marvin.html>
- Hornik, K. 2000, The R FAQ: Frequently Asked Questions on R.  
<ftp://franz.stat.wisc.edu/pub/R/doc/FAQ/R-FAQ.html>
- IBM, 1999. CLEVER project and automated text analysis, <http://www.almaden.ibm.com/-cs/k53/clever.html>
- Isearch developers, 1996. The CNIDR ISEARCH Text Searching System.  
<http://www.cnidr.org/ir/isearch.html>
- Kanehisa, M. 1999. KEGG: From genes to biochemical pathways. *in* Levotsky, Bioinformatics: Databases and Systems.
- kdnuggets, 2000. Software for Data Mining and Knowledge Discovery.  
<http://www.kdnuggets.com/software/index.html>
- Klein, H. 1999. Links to text analysis software, <http://www.intext.de/TEXTANAE.HTM>
- Kowalski, G., 1997. Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers

- Letondal, C., S. Bortzmeyer, A. Thebault, I. Wang, 1999. Bio Netbook,  
<http://www.pasteur.fr/recherche/BNB/bnb-en.html>
- Letondal, C. 2000. PISE, a tool to generate Web interfaces for Molecular Biology programs.  
Pasteur Institute. <http://www-alt.pasteur.fr/~letondal/Pise/>
- Letovsky, S., (ed), 1999. Bioinformatics: Databases and Systems , Kluwer Academic: Boston.
- Manber, U. and S. Wu., 1994. GLIMPSE: A Tool to Search Through File Systems. Usenix  
Winter 1994 Technical Conf.. Also, Technical Report #TR 93-34, Dept Computer  
Science, U. Arizona, <ftp://ftp.cs.arizona.edu/reports/1993/TR93-34.ps>.
- Model Eukaryotic Organism Workshop, 1998. Workshop report, December 1998.  
<http://www.nhlbi.nih.gov/nhlbi/sciinf/modeldb/model.htm>
- MySQL developers, 1999. MySQL, a Structured Query Language database server.  
<http://www.mysql.com/>
- Overbeek, R. N. Larsen, N. Maltsev, G. D. Pusch, E. Selkov 1999. WIT/WIT2: Metabolic  
reconstruction systems. *in* Levitsky, Bioinformatics: Databases and Systems.
- Paton, N.W., Stevens, R., Baker, P.G., Goble, C.A., Bechhofer, S., and Brass, A, 1999. Query  
Processing in the TAMBIS Bioinformatics Source Integration System, Proc. 11th Int. Conf.  
on Scientific and Statistical Databases (SSDBM), IEEE Press, 138-147
- PostgreSQL developers, 1999. PostgreSQL, an open-source, object relational database  
management system. <http://www.postgresql.org/>
- R developers, 2000. The Comprehensive R Archive Network.  
<ftp://franz.stat.wisc.edu/pub/R/index.html>
- Rebhan, M., V. Chalifa-Caspi, J. Prilusky and D. Lancet, 1998. GeneCards: a novel functional  
genomics compendium with automated data mining and query reformulation support.  
Bioinformatics, 14 (8): 656-664. Also <http://bioinformatics.weizmann.ac.il/cards/>
- Rice, P. 2000. EMBOSS - The European Molecular Biology Open Software Suite  
<http://www.sanger.ac.uk/Software/EMBOSS/>
- Riethoven, J-J. R. Harper., 1999. BioWurld - Where all Bioinformatics related URLs live.  
<http://search.ebi.ac.uk:8888/compass>
- Rodriguez-Tom , P. 1998. The BioCatalog. Bioinformatics, 14(5): 469-470. Also  
<http://www.ebi.ac.uk/biocat/biocat.html>
- Ugawa Y, 1997. Development of Mirror Server by using High Speed Data Transfer in Genome  
Science. Proposal to Asian-Pacific Advanced Network (APAN) organization,  
<http://www.jp.apan.net/HPIIS-Applications/JP-AFFRC>
- Witten, I., A. Moffat, T. Bell, 1999. Managing Gigabytes: Compressing and Indexing  
Documents and Images , Morgan Kaufmann Publishers
- Witten, I, and E. Frank, 1999 Data Mining: Practical Machine Learning Tools and Techniques  
with Java Implementations , Morgan Kaufmann Publishers. also  
<http://www.cs.waikato.ac.nz/~ml/weka/index.html>



## Current and Pending Support

(See GPG Section II.D.8 for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: **Donald Gilbert**

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ \*Transfer of Support

Project/Proposal Title: **World-Wide High Performance Network Mirroring of Public Genome Data**

Source of Support: **Indiana University**

Total Award Amount: \$ **19,900** Total Award Period Covered: **09/01/99 - 08/01/00**

Location of Project: **Indiana University**

Person-Months Per Year Committed to the Project. Cal:**0.00** Acad:**0.00** Sumr: **0.00**

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ \*Transfer of Support

Project/Proposal Title: **IUBio Archive and World-Wide High Performance Network Distribution of Bioinformatics Data**

Source of Support: **NSF - BDI 99 Grant# DBI-9982851**

Total Award Amount: \$ **67,000** Total Award Period Covered: **01/01/00 - 12/30/00**

Location of Project: **Indiana University**

Person-Months Per Year Committed to the Project. Cal:**0.00** Acad:**0.00** Sumr: **0.00**

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ \*Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: \$ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ \*Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: \$ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ \*Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: \$ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

\*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.