

Bioinformatics: Interdisciplinary Education and Research

Project Leaders

- Bruce Aronow, PhD, Associate Professor, Director, Bioinformatics, Division of Pediatric Informatics, Department of Pediatrics, College of Medicine, and Children's Hospital Research Foundation
- John J. Hutton, MD, Christian R. Holmes Professor and Dean, College of Medicine (as of July 1, 2002, Professor, Division of Pediatric Informatics, Department of Pediatrics, College of Medicine)
- Michael A. Lieberman, PhD, Distinguished Teaching Professor, Molecular Genetics, College of Medicine

Project Team

- Ralph F. Brueggemann, MBA, Director, Systems Development and Maintenance, Medical Center Academic Information Technology and Libraries (AIT&L)
- Ranajit Chakraborty, PhD, Kehoe Professor, Department of Environmental Health Sciences, College of Medicine, and Director, Center for Genome Information
- Ranjan Deka, PhD, Professor, Department of Environmental Health Sciences, College of Medicine
- Molly Gordon, MS, Director, Educational Services, UC Information Technology
- William Gordon, PhD, Applications Analyst, Systems Development and Maintenance, AIT&L
- Li Huang, MD, MS, Applications Analyst, Web Development, AIT&L
- Robert Kraft, BS, Applications Analyst, Systems Development and Maintenance, AIT&L
- Bruce Merz, BA, Applications Analyst, Systems Development and Maintenance, AIT&L
- Delores Mincarelli, BS, Research Associate, Systems Development and Maintenance, AIT&L
- John Pestian, PhD, MBA, Associate Professor, Director, Division of Pediatric Informatics, Department of Pediatrics, College of Medicine, and Children's Hospital Research Foundation
- Bhuvaneshwari Sakthivel, MS, Research Associate, Division of Pediatric Informatics, Department of Pediatrics, College of Medicine, and Children's Hospital Research Foundation
- Leslie Schick, MSLS, Director, Library Services, AIT&L
- Anshul Sharma, BTech, Applications Analyst, Web Development, AIT&
- Carman Wakefield, EdD, Instructional Designer, Information Services and Education, AIT&L

Oversight Committee

- John J. Hutton, MD (*Chair*)
- Marshall W. Anderson, PhD, Director, Center for Environmental Genetics, Professor and Chair, Department of Environmental Health, College of Medicine
- William S. Ball, Jr. MD, Interim Chair, Department of Biomedical Engineering, Colleges of Medicine and Engineering
- Michael A. Lieberman, PhD
- David E. Millhorn, PhD, Director, Genome Research Institute, Professor and Chair, Department of Genome Science
- John Pestian, PhD, MBA

Existing Programs/Organizations Involved

- Division of Pediatric Informatics, Department of Pediatrics, College of Medicine and Children's Hospital Research Foundation
- Department of Biomedical Engineering, Colleges of Medicine and Engineering
- Center for Genome Information, Department of Environmental Health, College of Medicine
- Center for Environmental Genetics, Department of Environmental Health, College of Medicine
- Genomics Research Institute, Medical Center
- University of Cincinnati Information Technology (UCit)
- Academic Information Technology and Libraries (AIT&L), Medical Center

Objectives/Needs

The UC Medical Center has invested millions of dollars in genomic research and in the information systems and biostatistical services necessary to support these efforts. Bioinformatics initiatives have emerged in units throughout the Medical Center. Our needs in bioinformatics are practical. As discussed later in this project description, most of the infrastructure is in place to support these needs. Bioinformatics tools and resources abound. However, we lack the mechanisms for researchers and students to use the existing resources optimally. Genomic research is growing rapidly and on many fronts. The amount of data and information produced is huge, creating an equally huge problem of managing the data and information and enabling people to transform it into knowledge. Typical of cutting edge endeavors, researchers are creating new resources and tools to meet their needs. There are few standards. Those who want to benefit from cutting edge endeavors must be able to filter, organize, understand, and apply the advances made at that edge. A major need in our institution is the translation of idiosyncratic data and applications into a common language that facilitates filtering and organizing.

Our proposed project focuses on cataloging available resources in a standardized way, indexing them so that investigators and students can readily choose the tools and databases they need, educating investigators and students about resources, and distributing to investigators and students the appropriate resources to support their research and education. Our objectives are:

1. To coordinate our bioinformatics programs, making them as efficient and effective as possible, and to develop a strategic vision for bioinformatics that all units agree to and pursue collaboratively.
2. To support our bioinformatics programs by developing tools to organize, catalog and deliver existing resources including software, hardware and faculty expertise.
3. To educate our students, faculty, and research assistants about fundamental principles of bioinformatics and the resources necessary to advance their research

Consistent with our Millennium Planning efforts, the Medical Center will form a Bioinformatics Oversight Committee in the summer of 2002 to assess our needs in bioinformatics and to devise institutional strategies to manage resource demands in this rapidly changing field. The Committee will report to the Senior Vice President and Provost for Health Affairs and will have a budget to facilitate

its work. IAIMS will provide a significant boost to this development.

With IAIMS, we will address the needs stated above by enhancing our distributive learning (see Appendix 8) and knowledge management (see Section B of the proposal) models to allow us to provide education and training resources that support research in our diverse and distributed environment. We will use these existing models to provide to students and researchers course and training materials that meet their individual bioinformatics needs, and the resources and tools they need for their research (including literature reference and full-text databases). These resources will all be organized for effective and efficient use by the user. These models have been designed to provide courses, training materials, and research tools on an “anywhere, anytime” basis via the Web, and will allow individuals to add materials that they have found useful.

Background/Local Context

The human genome is the overarching theme of the Medical Center’s primary strategic research areas of cancer, cardiovascular-pulmonary, neurobehavioral, and perinatal/early development, as well as the secondary areas of diabetes, environmental stress, and infectious diseases. In support of these strategic directions, the Medical Center continues to make large investments in bioinformatics and biostatistical support cores. In the following paragraphs we will discuss several existing cores that currently support bioinformatics initiatives.

The University of Cincinnati has recently created a new Department of Biomedical Engineering. Faculty of the Department are based in the colleges of Medicine and Engineering. Startup of the department was funded by the Whittaker Foundation with ongoing support from the University. The department will offer graduate degrees in bioinformatics. The Department of Computer Science is based in the College of Engineering. The departments of molecular genetics, cell biology, genome sciences, environmental health sciences, medicine, surgery, and pediatrics are in the College of Medicine, as are divisions of epidemiology, biostatistics, human genetics, and informatics. Research is a primary mission of the colleges of engineering and medicine and of the Children’s Hospital Research Foundation, where the College of Medicine’s pediatric faculty is based. Externally funded research awards to faculty of the Medical Center will total approximately \$180 million in 2002. Continuing growth in research is a fundamental part of the vision for the Medical Center, as articulated in our Millennium Plan. The Millennium Plan identifies bioinformatics as an essential core – a centrally supported activity shared by all research units – for this growth.

At the University of Cincinnati and Children’s Hospital Research Foundation we have extensive and expensive core facilities for research informatics, with investments exceeding \$10 million to date. As a matter of institutional strategy, the faculty and administration of the College of Medicine and Children’s Hospital chose to locate the major bioinformatics core within the Division of Pediatric Informatics of the Children’s Hospital Research Foundation. Three hundred faculty of the College are based at the Research Foundation. Initial funding for the bioinformatics core was provided by a Howard Hughes Medical Institute Research Infrastructure Improvement Grant, which was awarded to the College of Medicine. HHMI also provided funds for cores in proteomics (located in the College’s Medical Sciences Building) and genomics (located in the College’s Environmental Health Sciences Building). All three cores, bioinformatics, genomics, and proteomics, are now functioning and serving all faculty of the College.

The Division of Pediatric Informatics received start-up funds from the HHMI grant, as well as from the Children's Hospital Research Foundation and the College of Medicine. Recently the Division, with John Pestian as Principal Investigator, was awarded a \$2.3 million grant from the State of Ohio and Sun Microsystems to establish the nation's first supercomputer research center dedicated to pediatric informatics research. Computer hardware includes: two Sun Microsystems Star Fire 6800 with a total of 48 Sparc III processors and 96GB RAM, one Sun Microsystems Technical Computer Farm with 20 Sparc II processors and four TimeLogic Decipher BioAccelerator Field Programmable Gate Arrays, 36 Sun Microsystems Sparc II processors ((a Sun Microsystems E6500 (16 450Mhz Sparc II processors, 16GB RAM), a Sun Microsystems E4500 (8 450Mhz Sparc II processors, 8GB RAM), a Sun Microsystems E3500 (4 450Mhz Sparc II processors, 4GB RAM), 2 Sun Microsystems E450 (each with 4 450Mhz Sparc II processors, 4GB RAM)) that are available to the research base for advanced bioinformatics computational needs. All of these systems reside on a computational grid (Sun Gridware) that enables the sharing and scheduling of computational resources among specific systems. Additionally, there are a total of 32 Intel based processors housed in the Citrix Metaserver farm that provide centralized computational infrastructure for Windows 2000 bioinformatics software. Currently, a total of 3 terabytes of storage space is available to the research base. Storage capacity can be scaled up to 256 terabytes.

Web-based software tools include Oracle Application Server, BEA Weblogic, Apache web-server, and an Oracle 8i. PL/SQL, PHP and Java are used to develop end-user interfaces while Oracle and MySql are used for the storage and management of data. Numerous application software packages are available for use in research. These include SeqWeb, GeneSpring, GCG, HMMPPro, SAS, SPSS, Oracle WebDB and access to the proprietary Celera data and analytic tools.

There are presently five full time faculty in Pediatric Informatics, with additional recruitments underway. These faculty both conduct their own research and oversee support services for others. Examples of research within the Division include Bruce Aronow's program in molecular biology with emphasis on use of microarrays to analyze expression of genes in heart, brain and gastrointestinal tract. He is developing an efficient algorithm called TraFac which identifies potential regulatory elements in genomic sequences and compares them across genes and species. John Pestian is developing an integrated clinical-genomic data repository which he calls the Discovery System. The System stores in a single Oracle repository the discharge summaries, pathology and radiology reports, clinic records and other clinical information for all patients seen at Cincinnati Children's Hospital, the most clinically active children's hospital in the United States. An integral part of the Discovery System is the Biological Sample Tracking System (BSTS), which was developed in Cincinnati. The BSTS is a web-based system for tracking and annotating biological samples of Children's Hospital Research Foundation and the University of Cincinnati core laboratories. Investigators can request services from the core labs and track the status of their request via the web. Using Palm Pilots, lab managers can scan bar coded samples, annotate any activities performed and synchronize with a centralized Oracle database. Software for The Discovery System and the Biological Sample Tracking System were developed in the Division of Pediatric Informatics. Faculty and staff of the Division are working with faculty in other units to develop software for mining data in the System as, for example, to analyze clinical outcomes, effectiveness of treatment, and utility of gene chips to predict outcomes in juvenile arthritis, diabetes, asthma, and other diseases. The important point of this discussion is to show that our core faculty have broad based skills in

informatics. They utilize existing software and databases as well as develop novel new approaches to data acquisition, storage and analysis.

Services of the Division of Pediatric Informatics that are available to faculty are focused on basic genomics research because the tools are well developed, although fragmented among multiple software packages, and are most highly in demand. Use of the Discovery System for clinical research is less than optimal at the moment because user friendly web based software is under development and is not ready for general release. The Division:

1. Aids investigators in the design and analysis of complex experiments that will shed light into the molecular basis of biological processes both through user training and focused project support,
2. Aids the microarray, cDNA library, and proteomics cores in the organization and management of tracking and reference data for gene clones, PCR products, protein samples and gel images,
3. Generates a web accessible database that allows public access to gene expression data, and analyses, from all components, and
4. Promotes the design and execution of experiments that allow multiple research groups to synergize in their efforts.

While the highest concentration of hardware for computational biology is located at the Children's Hospital Research Foundation, faculty with well funded research programs who utilize or develop tools of bioinformatics are located in several other departments and at other physical sites. The NIH funded Center for Environmental Genetics headed by Marshall Anderson and the Center for Genome Information headed by Ranajit Chakraborty are located in the Department of Environmental Health. The Genomics Research Institute and the Department of Genome Science headed by David Millhorn are both new ventures and represent a major expansion of College research programs. They are located in a newly acquired research facility of approximately 500,000 square feet which is approximately 10 miles from the medical college and Children's Hospital. Approximately 100 faculty from various departments will be located at that site when it is fully developed. Parts of the facility will be occupied by researchers from industry, such as Procter and Gamble Pharmaceuticals. These faculty and programs will heavily utilize resources in bioinformatics.

There are currently several major challenges for us as we continue to develop our bioinformatics program. If we are to effectively manage the expensive and complex resources that we have available today and the even more expensive and complex resources that are expected in the near future, we must develop a dynamic system that is capable of adjusting to developments in a research area that is rapidly and continuously changing.

The first challenge is to develop a mechanism by which we can identify, organize and catalog the many disparate bioinformatics resources. These resources currently include databases, analytical tools, literature, training opportunities, reference materials, and faculty expertise. A common vocabulary must be developed for cataloging and searching purposes.

The second challenge is to strengthen our bioinformatics research program and to support new

research efforts through the development of an on-line catalog of bioinformatics software tools and techniques. The catalog will use the common vocabulary referred to above. It will also capture the experience and expertise of users as enrichments and annotations to existing resources. Using our existing media repository, expanded individual profiles, and newly developed information portals, faculty and students will be guided to appropriate resources on an individualized basis. This knowledge base will be designed to allow faculty and students to identify and apply tools and techniques that will support their research programs on an ongoing basis. Our information management model will allow researchers to locate and evaluate these resources and their application to their research program. This model will also capture the experience and expertise of users as enrichments and annotations to existing resources.

Finally, using our model for distributive learning (see Appendix 8), we will develop an on-going system of bioinformatics education and training designed and customized to meet the needs of faculty, staff and students in the Medical Center. The strategy for the educational component of this project is to provide short, online education/training/information modules that complement formal education programs, fill the knowledge gaps of experienced bioinformatics users, and assist users in applying their existing knowledge to new tools and resources.

We view a coherent vision of the future in bioinformatics, coupled with an integrated information management strategy, as essential elements necessary to support our overarching missions in research and education. The institutional vision and strategic plan will be provided by the Bioinformatics Oversight Committee, which will bring together our experts in bioinformatics and biomedical research to look to the future. This vision will guide institutional strategies for managing our bioinformatics enterprise in a way that assures that we have available the resources to support our extensive programs in research, education, and patient care.

National Context

The National Center for Biotechnology Information (NCBI), a unit of the National Library of Medicine, defines bioinformatics as the application of computer technology to the management of biological information. It is clear that genomic technologies and computational advances are leading to an information revolution in medicine and biology which will generate a need for new tools to study, track, control, prevent, and treat complex diseases. Bioinformatics provides essential tools for mining large databases such as those provided by the Human Genome Project and by microarray analyses of gene expression. Its sophisticated tools are absolutely essential to identify molecular lesions in common multifactoral human diseases of high prevalence such as diabetes, obesity, arthritis, hypertension, and cancer.

Bioinformatics has emerged as the current great challenge for most research intensive medical centers. The challenges are anchored in the explosive growth of biomedical information contained in large databases, the complexity of organizing and mining the databases, the cost of the informatics infrastructure, and the relative paucity of people who are trained in both the information and biomedical sciences. Many research intensive medical centers have established programs in bioinformatics, and are working to maximize their purchasing power in order to establish central facilities for the institution. Separate silos of activity have tended to develop to address various components of bioinformatics and have proved difficult to integrate. The problem of integrating

databases and systems for handling information is international. For example, in the United States the National Institutes of Health and the National Library of Medicine, and internationally the European Molecular Biology Organization, have developed databases and searching facilities to allow bioinformatics investigators to place and retrieve data. Each of these databases and programs has idiosyncrasies and is not always easily utilized in a complementary way. Institutions planning or sustaining initiatives in bioinformatics always work in a national or international context.

Description

Provide to students, faculty, and staff the bioinformatics resources and tools that they need for their research and education (including databases, software, and literature references), all of which are cataloged and organized for effective use when and where they are needed.

Our core services in bioinformatics are intended to support our entire research enterprise. Because of the diversity of needs among investigators, we provide access to a very large number of databases and tools. Individual investigators, however, have focused research programs and use a subset of tools and databases. Given the very large number of research tools and materials available, and differing levels of expertise, many researchers expend large amounts of time finding appropriate research materials and tools. Much of the problem associated with locating and accessing bioinformatics information, applications, and other tools is related to shortcomings in current mechanisms for cataloging and searching for these resources. The area of bioinformatics is relatively new and very complex. Resources in this area have grown up around small, mostly proprietary research enterprises. This has made sharing information and tools very difficult. Additionally, a new lexicon of terms has emerged that is both unique to bioinformatics and typically not included in existing vocabulary standards such as MeSH and UMLS. This project will develop a sophisticated common vocabulary for bioinformatics resources as a foundation of a cataloging system. We will base vocabulary choices on MeSH and UMLS whenever possible, and will work collaboratively with MeSH and UMLS developers to guarantee our decisions are consistent with standards as they evolve.

The four digital services at the heart of our knowledge management model, enabled by individual profiles built into the integrated database and middleware application, will take advantage of our new, developing vocabulary to more effectively filter bioinformatics information for the researcher and student. For example, a researcher will be able to effectively search the media repository (pull) to find software and resources related to their needs. To reduce the amount of effort researchers spend finding tools and materials, this project will allow individual researchers to store profiles of their research projects and self-assessed expertise levels. The learning model will push to each researcher's desktop the appropriate resources, possibly pre-selected for level of reliability and ease of use. We will also provide the names of other investigators in our institution who have had experience with the relevant databases and software. Experienced colleagues could provide helpful advice, particularly to the novice user. Some of this expertise will be captured in the resource repository in the form of comments and ratings. This information, in turn, will be available to further assist users in assessing new and existing resources. We envision providing customized, web based bioinformatics resources to our faculty and students who are located at multiple geographic sites. Identified researchers will be notified of new knowledge, tools or publications in their field of expertise using our built-in alert system.

As an example of the complexity of the current research environment, some members of our faculty have a large NIH grant to sequence the genome of *Pneumocystis carinii*, a pathogenic fungus that commonly infects immunocompromised patients. The organism has a large genome and contains approximately 8 megabase pairs of DNA distributed over 15 or 16 chromosomes. Sequencing the DNA requires high throughput technologies that generate data very quickly. Software such as Phred converts the traces from the automated sequencer into sequences and indicates the probable accuracy of the base calls. Software such as Phrap assembler combined with consed or a variety of commercial proprietary packages is used to assemble long segments of the genome. The genomic sequences must then be analyzed to find the introns, exons and regulatory elements of genes (software such as Genescan and MatInspector). The tools that investigators must choose and understand in order to sequence the *Pneumocystis* genome are quite different from those required by another research group which has been funded by NIH to discover genes in the mouse which differentiate the nervous system from other tissues. Here the basic tool is microarray analysis of expression of genes in different tissues. In this project, GeneSpring software was used to identify 653 cDNAs that were highly expressed in the nervous system, out of 8734 on the array. The cDNAs were used to identify genes, some of known function, others unknown. Genomic sequences were then identified using Celera databases and associated Celera software. Some of the sequences were complete genes while others were only partial sequences of the relevant genes. If a partial gene was identified, efforts were made to identify contigs that could be used to assemble the full sequence of a gene. Putative transcription factor binding sites could be identified in the genomic sequences using software such as MatInspector. Other software programs, such as our locally developed algorithm, TraFac, were then used to identify clusters of phylogenetically conserved cis-elements among genes highly expressed in the nervous system. These were putative regulatory elements.

These examples illustrate the amount of knowledge of bioinformatics and the variety of tools required to perform genomic research, and the extent to which pre-selection of research materials and tools could help investigators. Our two examples are not intended to present cutting edge research in bioinformatics per se. We could have done this, but we believe that the major need in our institution at this time is to help the large number faculty and students, both current users and potential users of bioinformatics resources, to understand and apply those resources optimally.

Provide courses and training materials on an “anywhere, anytime” basis via the web to teach students, faculty, and staff to choose and utilize knowledgably our resources in bioinformatics.

There are several levels of educational needs for faculty, students, and staff in bioinformatics. The baseline educational needs in this area are met by formal educational programs. We have well developed graduate courses that cover the basic theory and science necessary to conduct research in genomics. Bioinformatics, however, is a very rapidly growing field that requires researchers to constantly learn new analytic techniques, new information management applications, and new programs for transferring and organizing data. In the context of an active research enterprise, it is rare that new applications or information are so uniquely different that users must take significant time away from their projects to learn large amounts of new material. On the contrary, existing knowledge of programs, applications, and databases is easily transferred to learning new information and skills.

Our educational strategy is to identify knowledge gaps and needed instructional bridges to support

and facilitate ongoing research. We are basing our strategy on an adult learning model in which the learner identifies his/her information/knowledge gaps as problems that are retarding progress or optimal performance. These gaps are typically both frequent and small. Individuals will have the opportunity to access short educational modules as needed. We will be developing a growing library of these online modules that will be cataloged and accessible from our resource repository. We will also use the same profile-driven "push" model to alert individuals to potentially relevant new training information. Often the training information will be paired with "pushed" information on the availability of new resources. People will receive information on "this is what's new" as well as "...and this is the instruction you will need to use it."

We recognize that faculty, staff, and students have a wide range of educational needs. These needs run the gamut from "How do I transfer data from this program to my database?" to "What are the statistical implications of running 8,000 analyses that result in 200 clustered "significant" results?" Our goal is to work closely with researchers to monitor the needs of users and to quickly develop appropriate educational modules that are available online. To accomplish this, we will develop mechanisms to monitor the information/education needs of researchers and staff. Additionally, we need to have organized resources that can translate the information/education gap into short, functionally appropriate modules.

Currently, there are a number of traditional ways that researchers and their staff use to try to meet their educational needs. Regular internal seminars are centered on topics identified by key researchers. While this strategy has identified some needs that are common to everyone, it does not fully address all of the educational needs of researchers and staff. The resulting research seminars provide a good opportunity to fully address specific educational topics, but they are not helpful to those who cannot attend. They generally assume a reasonably sophisticated prior knowledge of the field and the content of the seminars is only available the single time it is offered. Our knowledge management model will "push" recorded seminars to individuals profiled to view such content. It will also "alert" individuals on an ad hoc basis to critical training and professional development events. We will use the current informal methods of talking with key researchers for some needs assessments. However, we will also monitor resource use in the electronic repository, requests for assistance among researchers, and federal and state compliance mandates to develop a broader institutional perspective and to plan a more coherent program of education.

The Medical Center has gained considerable experience in the development of online training modules. Our Distributive Learning Collaboratory has developed several training programs for required education by OSHA and for human subjects researchers. Over the next year the Collaboratory is scheduled to complete at least four additional training modules in these areas. The modules use off-the-shelf technology in the development of small competency-based segments that allow the trainee to verify his/her knowledge of a particular learning objective. This model allows individuals to exit the program at any time and then to return to the point they exited the program at a later time. In areas where the educational content is readily available, a 10-15 minute module requires approximately 10 weeks to design and program. In cases where an educational module is urgently needed, the development time could be cut in half if resources are made available.

We will develop series of Web-based training modules for the following three broad topics. Genomics researchers and bioinformaticians at UC identified these topics as most important:

- The computational infrastructure in bioinformatics
- Sequence analysis using proprietary and public web-based software
- Expression analysis using proprietary and public web-based software

The online educational modules are designed to complement, rather than to supplant, formal courses on bioinformatics topics. New researchers and students will continue to take advantage of formal courses that provide state-of-the-art knowledge of:

- a) basic principles of design, selection, and application of algorithms for the creation and mining of biological databases;
- b) local and internet based resources for genomic, proteomic, and genetic analyses;
- c) contemporary problems in biology and medicine which can be addressed using these resources.

"Hands on" instruction in the application of computer technology to the management of biological information provides:

- a) knowledge of the databases and software tools available on commonly used public-access web sites such as NCBI and EMBL;
- b) knowledge of locally maintained databases and software for mining these databases, including the advantages and limitations of each;
- c) practical experience selecting, accessing and using software and databases to solve typical problems encountered in biological research.

Seminar series, mini-symposia, and discussion groups will continue to be held in order to update investigators about new developments in the rapidly evolving field of bioinformatics. These will be planned, organized, and funded by the Bioinformatics Oversight Committee. The content of these live educational activities will be captured in a variety of electronic formats in order to make the content available for review via the Internet.

The University of Cincinnati has approved a new graduate program in bioinformatics, which will require us to recruit additional faculty and to strengthen education and research in bioinformatics. There is a critical need for well trained investigators in this field. The sequences of the genomes of numerous organisms are now available in public databases and are rapidly being expanded. Mining these data pose enormous technical challenges. To these data are added data on gene expression. It has become technically easy to generate large amounts of data about gene expression using microarrays; the task of analyzing the data becomes more complex as the size of microarrays increase. Similarly, as the ability to genotype huge numbers of people is a reality in our institution, the complexity of database management and analysis to identify genes involved in multi-genic diseases is apparent. Professionals who understand the mathematics and biology involved in these analyses will be required to properly interpret the experiments. New algorithms for efficient and meaningful data analyses are needed, as is basic research on such issues as determining statistical and

biological significance of results from complex data mining. These individuals need to call upon knowledge of biology, medicine, mathematics, and computer science in order to knowledgeably analyze the data and to effectively communicate with those who created the data. Such individuals will be crucial for bioinformatics to fulfill the role expected of it in the future. We need to utilize these learning accelerators to expand our ability to train an increasing number of individual researchers and students. The University has responded to this challenge by creating two major new programs. First, our new Department of Biomedical Engineering has bioinformatics as one of its three areas of focus and has been given the authority by the Ohio Board of Regents to grant BS, MS and PhD degrees. It has received large grants from the Whittaker Foundation to assist in the start up of its programs. Second, within the College of Medicine, a Center for Genomics Information, headed by Dr. Ranajit Chakraborty, has been created with graduate student and postdoctoral education as its major aims.

Program Uniqueness/Benefits

Bioinformatics is well-positioned for institution-wide planning in part because the need for a cost effective program of high quality is easily recognized by faculty, students, staff, and senior administrators. Wide spread recognition of this new discipline as key to success of multiple missions – research, education, and patient care – places bioinformatics in a unique position for broad support. Because of high costs and the scarcity of well-trained specialists to staff bioinformatics programs, the benefits of a coordinated institution-wide initiative are more obvious than in most other fields.

Management

The Bioinformatics Oversight Committee, chaired by Dr. Hutton, will meet regularly to determine the strategic direction, to ensure coordination of bioinformatics efforts at the UC Medical Center, and to monitor the progress of and provide advice to the development of the education and training initiatives undertaken as part of the IAIMS program. The Committee consists of the directors and chairs of the major programs and departments that currently have bioinformatics programs. Membership will be modified appropriately as new programs come into existence and/or the focus of the program changes. The committee will report to the Senior Vice President and Provost for Health Affairs, who chairs the IAIMS Steering Committee, and will have its own budget.

The bioinformatics project team, led by Drs. Aronow, Hutton and Lieberman, will be responsible for the development and deployment of the proposed education and training initiatives. The project team will meet regularly to assess progress, solve problems, and redirect activities as required. They will receive direction from the Oversight Committee and will make adjustments throughout the course of the grant funded program based on data provided by the formal process of evaluation.

For the past several years, we have been using the Unified Method of software development, and especially Use Cases, to determine and refine user needs and develop software that addresses these needs. We have found that this method encourages user participation in the development phase and gives users a sense of “ownership” of the resulting products, while providing ample opportunities for “tuning” the software during the development stages to ensure that user needs are fully addressed.

Technical Description

The integrated database described in the main body of the IAIMS application and the Information Technology and Resources Self-Study (Appendix 3) will be expanded to include the new definitions and data types associated with bioinformatics resources and software. User profiles will be modified to allow linkages between user profiles and the information stored in the media repository. Information portals will be designed to allow researchers to search the media repository and to automatically receive information from the repository.

The Evolutionary Toolset

The enhanced distributive learning model combined with the knowledge management model developed for IAIMS combine to form a toolset that can morph and evolve to meet the needs of rapidly changing fields like bioinformatics. We will expand the integrated database to include definitions and categorizations related to the evolving field of bioinformatics research. Also, the media repository will be expanded to include resource types (for example, descriptions of gene databases, protein databases, and software tools) associated with bioinformatics, and to support a new research service that is, in effect, an annotated bibliography of research tools and resources in bioinformatics.

We will develop a web application – an enhanced personal profile maintainer – that researchers can use to specify their areas of research activity, their specific interests in bioinformatics, and their level of expertise in area and tools knowledge. With this web application, researchers will be able to contribute to the research service by rating tools and resources already included and by adding new tools and resources to the service.

Researchers' personal portals will provide direct access to resources, tools, and courses appropriate to their areas of research activity, bioinformatics interest, and technical expertise.

Estimated Resources

Stage	Person Days	Roles
Project Management	12	Project planning, scheduling, and monitoring
Requirements	53	Customer needs collection and define workflow
Analysis	53	System analysis and preparation of specifications
Design	173	Database and web interface design
Implementation	262	Database development / middleware development / interface development
Quality Assurance	31	Interface (browser testing) / Usability / Error / Load / DB Verification.
Evaluation	1	Evaluate project
Maintenance	2	Ongoing maintenance

In accordance with our distributive learning model, we will convert portions of the introductory bioinformatics courses to web-based training modules. The modules may include any combination of text, illustrations, animations, video, audio, and other electronic media that can be delivered via a web browser. The ideal will be to use state-of-the-art tools to build each module so that it can be delivered

to a student or researcher across a variety of platforms. One option is to use an instructional shell like Blackboard for course delivery; another option is to use the Research Training system currently in use at the University of Cincinnati to deliver web courses. The overall goal of the learning model is to facilitate learning across a variety of delivery methods for access anywhere, at any time. Based on feedback, modular content will be developed and delivered using appropriate technologies in the following topics:

The computational infrastructure in bioinformatics

Estimated Resources:

Stage	Person Days	Roles
Project Management	12	Project planning, scheduling, and monitoring
Requirements	48	Determining Audience needs
Analysis	18	Preparing Content
Design	33	Design
Implementation	86	Module Development
Quality Assurance	18	Testing.
Evaluation	1	Evaluate project
Maintenance	9	Ongoing maintenance – 5 hours per week.

Sequence analysis using proprietary and public web-based software

Estimated Resources

Stage	Person Days	Roles
Project Management	12	Project planning, scheduling, and monitoring
Requirements	44	Determining Audience needs
Analysis	17	Preparing Content
Design	32	Design
Implementation	85	Module Development
Quality Assurance	17	Testing.
Evaluation	1	Evaluate project
Maintenance	9	Ongoing maintenance – 5 hours per week.

Expression analysis using proprietary and public web-based software

Estimated Resources

Stage	Person Days	Roles
Project Management	12	Project planning, scheduling, and monitoring
Requirements	44	Determining Audience needs
Analysis	17	Preparing Content

Design	25	Design
Implementation	92	Module Development
Quality Assurance	17	Testing.
Evaluation	1	Evaluate project
Maintenance	16	Ongoing maintenance – 5 hours per week.

Resources Required

This project will require 2.25 full-time equivalent staff for each of four years to complete. These personnel will fall into the following categories:

Faculty/content experts	.75 FTE
Technology specialists	1.40 FTE
Administrators	.10 FTE

Evaluation

The evaluation of this project will focus on the increased capacity of students and researchers to develop and conduct research in bioinformatics and genomics. These objectives are predicated on the assumption that improving the organization and access to genomics data as well as the improving the bioinformatics skills of faculty and students will result in a more efficiently functioning research enterprise. The summative and formative evaluations will employ both quantitative and qualitative methodologies for gathering data. The evaluation reports will be integrated into the ongoing management of this project so as to provide feedback and guidance to project administrators, particularly the Bioinformatics Oversight Committee.

The first step in this evaluation will be to identify specific project milestones and to develop appropriate success indicators that can be measured. The dynamic nature of this project is likely to change or add milestones throughout the project. However, the table below illustrates several planned milestones and their associated evaluation strategies.

Objective 1: Provide to students, faculty, and staff the bioinformatics resources and tools that they need for their research and education (including databases, software, and literature references), all of which are cataloged and organized for effective use when and where they are needed.

Milestone	Evaluation Measures	Feedback
Development of a cataloging scheme for databases, tools, and other resources	<ul style="list-style-type: none"> ❑ User reviews of vocabulary terms ❑ Match between needs taxonomy and cataloging scheme (level of lexicon overlap) ❑ User assessment (survey and focus groups) of utility of cataloging scheme 	<ul style="list-style-type: none"> ❑ Monthly presentation of evaluation data to project leaders for first 12 months. ❑ Quarterly reports thereafter.
Individualized "need" profile	<ul style="list-style-type: none"> ❑ Time needed for users to complete profile. ❑ User assessments (survey and focus 	<ul style="list-style-type: none"> ❑ Monthly presentation of data to project leaders for 1st six months.

Milestone	Evaluation Measures	Feedback
	<ul style="list-style-type: none"> group) of accuracy and utility of profiles □ Time delay between changes in needs and changes in user profiles 	<ul style="list-style-type: none"> □ Quarterly presentations thereafter.
Coding and cataloging resources	<ul style="list-style-type: none"> □ Number of miscoded resources (reported by users) □ Number of "null" searches □ Time to code and catalog new resources 	<ul style="list-style-type: none"> □ Weekly reports to project leaders for six months after coding begins. □ Monthly reports thereafter
"Pushing" resources to users	<ul style="list-style-type: none"> □ Match between user needs profiles and resources pushed to users (survey of users) □ Utilization of pushed resources by users (survey and use statistics) □ Time to change "pushed" resources when user needs change 	<ul style="list-style-type: none"> □ Profile/resources match reports will be reported quarterly to project leaders. □ Utilization and Update time statistics will be reported monthly for 12 months after resources are available. Reported quarterly thereafter.
Resource sharing among users	<ul style="list-style-type: none"> □ Number of first time access to resources by users (survey) □ Number of new collaborations (focus group and survey) □ Number of projects sharing resources/tools 	<ul style="list-style-type: none"> □ Statistics will be reported to project leaders biennially.

Objective 2: Provide courses and training materials on an "anywhere, anytime" basis via the web to teach students, faculty, and staff to choose and utilize knowledgeably our resources in bioinformatics.

Milestone	Evaluation Measures	Feedback
Identify educational and instructional needs of users	<ul style="list-style-type: none"> □ Comparison of questions asked among researchers and staff to content of online training modules □ User ratings of proposed instructional material (survey) □ User self-assessments (survey) □ Enrollment/registration statistics for existing instruction □ Use rates of resources 	<ul style="list-style-type: none"> □ User needs survey data will be presented to project team biennially. □ Enrollment and use rate statistics will be presented quarterly.
Develop a system for timely creation of instruction	<ul style="list-style-type: none"> □ Time from need identification to availability of instruction □ User comparison of our instruction vs. other sources of instruction 	<ul style="list-style-type: none"> □ Regular meetings will be held between the instruction development team and

Milestone	Evaluation Measures	Feedback
	<ul style="list-style-type: none"> ❑ Quality of instruction (user feedback) ❑ Amount and type of resources used to develop instruction 	the project leaders for review of these data.
Provide easy access to instructional resources anytime and anyplace	<ul style="list-style-type: none"> ❑ Log of technical problems/issues related to online instruction. ❑ Online user feedback focused on IT access/use issues. ❑ Online instruction use statistics 	<ul style="list-style-type: none"> ❑ Data will be reviewed on an ongoing basis by the technical team and instruction development team. Quarterly reports will be given to the project leadership team.
Impact of online instruction on user needs and behavior	<ul style="list-style-type: none"> ❑ Changes in content and number of questions directed to researchers by other researchers and staff ❑ Content evaluations of online materials by users. ❑ Changes in use rates of data resources after completing instruction. ❑ Yearly "impact of instruction" surveys to users. 	<ul style="list-style-type: none"> ❑ Content evaluation data reported to instruction development team and project leadership quarterly. ❑ Data resource use statistics and annual surveys reported annually.

Bibliography

1. Gross, LJ. Education for a biocomplex future. *Science* **288**:807 (2000).
2. Krawetz, SA and Womble, DD. Design and implementation of an introductory course for computer applications in molecular genetics. A case study. *Molecular Biotechnology*, **17**:27-41 (2001).
3. Luscombe, NM, Greenbaum, D and Gerstein M. What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, **40**:346-358 (2001).
4. Zauhar, RJ. University bioinformatics programs on the rise. *Nature Biotechnology*, **19**:285-286 (2001).

Timeline

Our focus on interdisciplinary education and research in bioinformatics will span the duration of the operations grant. It will require four years to develop and establish a full range of web based educational modules for the use by scientists interested in utilizing bioinformatics. Each educational module will be completed by first assessing what is available, learning the content, taking the content and developing a template for web based training, using a variety of tools, testing the materials on a small group, making modifications, testing with another small group, and finally going "live". "Going live" will not mark the end of development for each module.

We recognize that the field of bioinformatics is progressing rapidly. There will need to be continuous

reassessment of algorithms, databases, and theories of analysis and interpretation to keep our knowledge management and learning applications relevant and current. The Bioinformatics Oversight Committee, through its ongoing assessment and planning of our total institutional effort in bioinformatics, will have the responsibility for making sure we adjust to the rapid and sometimes unpredictable advances.

The detailed implementation schedule that follows shows a breakdown of tasks needed to complete the project, the time/effort required for each task, and the order in which tasks will be completed. Because of the rapid evolution of this field and a relatively high degree of uncertainty about future developments, this timeline is less detailed than the timelines for the other two projects in our IAIMS proposal.