LBNL/PUB-919

A Catalog of Berkeley Lab Computing Sciences' Capabilities for Computational Biology and Data Management

Horst D. Simon, Editor
Computing Sciences Directorate
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720
HDSimon@lbl.gov

September 2004

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC 03-76SF00098; and by the National Institutes of Health, U.S. Department of Health and Human Services.

Table of Contents

Introduction

Computational biology represents one of the last remaining scientific frontiers for computational sciences. From helping researchers decode the vast amounts of complex data resulting from the Human Genome project to predicting protein folding to facilitating collaborations among scientists at sites around the world, computing and networking resources are playing an ever-increasing role in the biosciences.

Whether simulating complex systems on a supercomputer or analyzing and visualizing massive datasets, scientists today rely on advances in computer science, mathematics, and computational science, as well as large-scale computing and networking facilities, to increase our understanding of ourselves, our planet, and our universe. The Computing Sciences organization at the U.S. Department of Energy’s Lawrence Berkeley National Laboratory performs basic research and develops and deploys new tools and technologies to meet these needs and to advance research in many areas of science, including the life sciences.

Computing Sciences carries out its mission by operating national user facilities and programs, providing on-site services at Berkeley Lab, and conducting applied research and development in computer science, computational science, and applied mathematics — the three essential elements of computational modeling and simulation.

This paper describes current or recent biological research projects carried out in collaboration with the National Energy Research Scientific Computing (NERSC) Center Division and the Computational Research Division (CRD), as well as the computing infrastructure provided by the Information Technologies and Services Division (ITSD).

HPC Systems and Services (NERSC Center Division)

http://www.nersc.gov
Horst Simon, HDSimon@lbl.gov
Bill Kramer, WTKramer@lbl.gov

The NERSC Center provides reliable, state-of-the-art computing resources and user support to thousands of researchers nationwide while advancing the state of computational and computer sciences. NERSC operates one of the fastest computers in the United States available for general scientific research, a 6,656-processor IBM RS/6000 SP with a peak speed of 10 teraflop/s (~10 trillion calculations per second), 7.8 terabytes of aggregate memory, and a Global Parallel File System with 44 terabytes of storage. NERSC is a developer site of the R&D100 award-winning High Performance Storage System (HPSS), which provides 35 terabytes of disk cache and an archival storage capability of 8.8 petabytes. NERSC systems are robust and available 24 hours a day, 7 days a week. The systems are configured for high availability, tuned for good network bandwidth, and accessible through the DOE Science Grid. High-bandwidth access to NERSC from the U.S. and the world is available via DOE's Energy Sciences Network (ESnet) (http://www.es.net).

Software and Consulting Services

Francesca Verdier, FVerdier@lbl.gov

NERSC provides an array of computational biology tools optimized for our computational platforms, as well as staff scientists who can write or optimize software to fully exploit the high levels of parallelism in bioinformatics pipelines and map them to parallel supercomputers. One example of the benefits of parallel computing is the comparison of two whole genomes using the NCBI BLAST program. By breaking up both genomes into fragments, the whole problem can be decomposed into many smaller problems. By using many nodes of NERSC's IBM SP, each of these smaller BLAST problems is allocated to a single CPU, and the comparison proceeds much faster than on a serial workstation.

Data Management, Access and Storage

Nancy Meyer, NLMeyer@lbl.gov

NERSC is currently collaborating with the DOE's Joint Genome Institute (JGI) on a project to optimize genomic data storage for wide accessibility. This project will distribute and enhance access to data generated at JGI's Production Genome Facility, and will archive the data on NERSC's HPSS. The Production Genome Facility is one of the world's largest public DNA sequencing facilities, producing 2 million trace data files each month (25 KB each), 100 assembled projects per month (50–250 MB), and several very large assembled projects per year (~50 GB). NERSC storage will make this data more quickly and easily available to genome researchers. Storage will be configured to hold more data online and intelligently move the data to enable high-speed access.

NERSC and JGI staff are working closely to improve the organization and clustering of DNA trace data so that selected portions of an assembled dataset can be downloaded, instead of the entire project, which could contain millions of trace files. The Web interface being developed will for the first time correlate the DNA trace data with portions of the assembled sequence, simplifying the selection of trace data, reducing the amount of data that must be handled, and enhancing researchers' ability to generate partial comparisons. Cached Web servers and file systems will extend the archive to the wide area network, distributing data across multiple platforms based on data characteristics, then unify data access through one interface. When this project is completed, its storage and data movement techniques, together with its data interfaces, could serve as a model for other read-only data repositories.

Examples of Biological Research Using NERSC Capabilities

NERSC resources have been used for a variety of biological research projects funded by the Department of Energy's Office of Biological and Environmental Research, the National Institutes of Health, the National Science Foundation, and other agencies and foundations. A few representative examples are listed below.

Sequencing the *Fugu* Genome

The genome of the Japanese pufferfish *Fugu rubripes* was the first vertebrate genome to be draft sequenced after the human genome, and the first public assembly of an animal genome by the efficient but computationally demanding whole genome shotgun sequencing method. This achievement was made possible by a new computational algorithm, JAZZ, that was developed at DOE's Joint Genome Institute (JGI) to handle large genome assembly projects. JGI used NERSC's IBM SP to develop and test JAZZ, which reconstructs contiguous genome sequences by overlapping the short subsequences. [1]

Mapping the Universe of Protein Structures

Researchers at Berkeley Lab and the University of California at Berkeley, using NERSC's IBM SP, have created the first 3D global map of the protein structure universe (Figure 1). Based solely on empirical data and a mathematical formula, the map closely mirrors the widely used Structural Classification System of Proteins (SCOP), which is based on the visual observations of scientists who have been solving protein structures. This map provides important insight into the evolution and demographics of protein structures and may help scientists identify the functions of newly discovered proteins. [2]

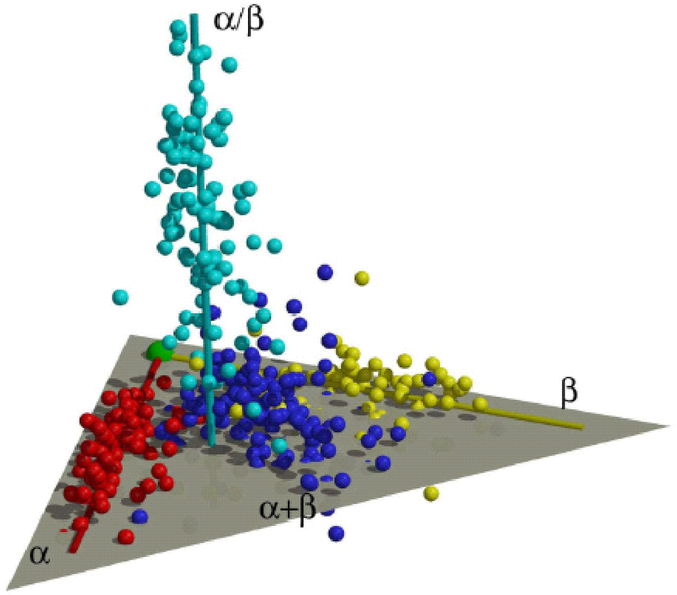


Figure 1. This 3D map of the protein structure universe shows the distribution in space of the 500 most common protein fold families as represented by spheres. The spheres, which are colored according to classification, reveal four distinct classes: α (red), β (yellow), $\alpha+\beta$ (blue), and α/β (cyan). It can be seen that the α, β, and $\alpha+\beta$ classes share a common area of origin, while the α/β class of protein structures is shown to be a late bloomer on the evolutionary flagpole.

Determining the Structure of a Carcinogen

The mutagen PhIP is the most abundant of the heterocyclic amines (HA), which are formed in meat and fish during cooking and have been shown to form DNA adducts, which are implicated in the etiology of colon, lung, and breast cancers. Because of difficulties in synthesizing the large quantities of highly purified samples needed, no structural studies of HA-DNA adducts had been reported until the work of Brown et al., who optimized the synthesis of the C8-dG-PhIP adduct in an 11-mer DNA sequence and determined its structure through a combination of nuclear magnetic resonance imaging and molecular mechanics computation at NERSC. Determination of the structure of this adduct is an important first step in elucidating the mechanisms by which DNA damage by PhIP can lead to cancer. This study received a Certificate of Merit Award in February 2002 at the Sixth International Symposium on Predictive Oncology and Intervention Strategies. [3]

Computational Modeling of Protein Switches

The Src-family kinases and their close relatives, such as Abl, function as molecular switches and are important targets for therapeutic intervention; for example, the c-Abl inhibitor Gleevec is used to treat chronic myolegenous leukemia. The development of specific inhibitors is complicated by the fact that human cells contain ~500 different protein kinases, all of which catalyze the same chemical reaction and contain highly conserved active sites. Because crystal structures provide only static views of certain states of each kinase, Nagar et al. used molecular dynamics simulations to help understand specific switching mechanisms. The simulations confirmed that the regulatory mechanism of the proto-oncogenic Abl protein shares an important feature with the regulatory mechanism of the related Src-family tyrosine kinases (Figure 2). This result was verified by *in vitro* experiments. [4]

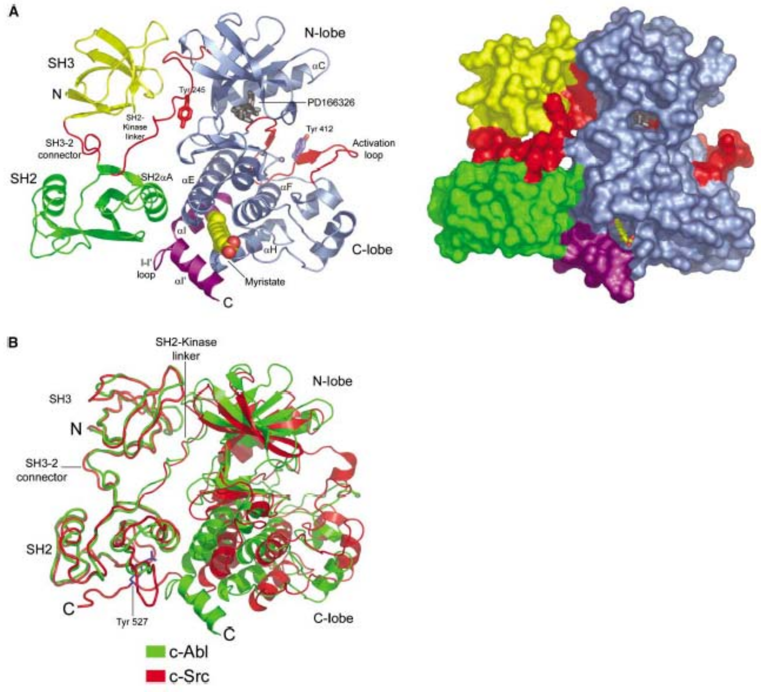


Figure 2. (A) Ribbon and surface representations of c-Abl. (B) Superposition of the structure of c-Src.

The Impact of Carcinogens on DNA Structure

Understanding the 3D structure of DNA that has been modified by chemical carcinogens is important in understanding the molecular basis of cancer and chemical toxicology. Peterson et al. studied [POB]dG, which is produced from metabolically activated tobacco-specific nitrosomines. This compound modifies the base guanine, posing a significant cancer risk. The results of NMR spectra for [POB]dG were used as inputs to the DUPLEX code, which produced a minimum energy structure that agreed with the NMR data. [5]

Tools and Technologies (Computational Research Division)

http://crd.lbl.gov
Horst Simon, HDSimon@lbl.gov

The Computational Research Division (CRD) creates computational tools and technologies that enable scientific breakthroughs, by conducting applied research and development in computer science, computational science and applied mathematics. CRD consists of three departments: the Biological Data Management and Technology Center, the Distributed Systems Department, and the High Performance Computing Research Department.

Biological Data Management and Technology Center

http://crd.lbl.gov/html/BDMTC/
Victor Markowitz, VMMarkowitz@lbl.gov

Translating the massive increases of complex data related to biomedical research into improved health care will require improved methods of data management and software tool development, according to two National Institutes of Health reports: the "Biomedical Information Science and Technology Initiative," prepared by the Working Group on Biomedical Computing Advisory Committee to the NIH Director, and the "NIH Roadmap for Accelerating Medical Discovery to Improve Health: Bioinformatics and Computational Biology." Both documents recommend employing advanced data management technologies for developing interoperable biomedical databases and software engineering principles for delivering robust and reliable tools for biomedical research.

Data management and bioinformatics software development challenges are also discussed in DOE's Genomes to Life (GTL) program report "User Facilities for 21st Century Systems Biology: Providing Critical Technologies for the Research Community." GTL envisions different types of facilities generating data that would be organized in a variety of databases, including expression, proteomic, protein-function, chemistry, and pathway databases. Data will be collected, archived and passed through a number of processing stages, including data annotation and integration, whereby a "seamless and effectively centralized capability to deal with data" in the form of data centers collecting and integrating effectively large-scale biological data is seen as key to GTL's success.

To help meet these needs, Berkeley Lab has established the Biological Data Management and Technology Center (BDMTC). BDMTC serves as a source of expertise in and provides support for data management and bioinformatics tool development projects at the Joint Genome Institute (JGI), the Life Sciences and Physical Biosciences Divisions at Berkeley Lab, Biomedical Centers at the University of California San Francisco, and other similar organizations in the Bay Area. BDMTC enables collaborating organizations to share experience, expertise, technology and results across projects. BDMTC employs industry practices in developing data management systems and bioinformatics tools, while maintaining academic high standards for the underlying data generation, interpretation and analysis methods and algorithms.

BDMTC provides support in addressing key data management challenges, including:

- the massive increase in the amount and range of biological data,
- the difficulty of quantifying the quality of data generated using inherently imprecise tools and techniques, and
- the high complexity of integrating data residing in diverse and sometimes poorly correlated repositories.

BDMTC's strategy involves adapting existing technology and methods to a specific application in order to address immediate data management and bioinformatics requirements. Cost-effectiveness and the ability to take advantage of rapid technological advances without loss of quality, time, or cost are built into solutions that are inherently evolving. Critical data management problems that cannot be resolved using existing technology are pursued as part of longer-term R&D activities.

BDMTC is led by Victor M. Markowitz, who until 2003 was CIO and Senior VP, Data Management Systems at Gene Logic, where he was responsible for the development and deployment of the data management and analysis platform for the company's gene expression data. Prior to joining Gene Logic in 1997, he was a staff scientist at Berkeley Lab, where he led the development of data management tools applied to biological databases.

BDMTC is currently working with JGI on developing microbial genome and metagenomics data management systems, and is involved in the UC Berkeley proposal for an NIH National Center for Biomedical Computing, where it provides the infrastructure, data management, and software development cores. BDMTC aims to become a QB3 affiliated center.

Distributed Systems Department

http://www.dsd.lbl.gov/
Deb Agarwal, DAAgarwal@lbl.gov

The Distributed Systems Department (DSD) researches and develops software components that allow scientists to address complex and large-scale computing and data analysis problems in a distributed environment, such as the DOE Science Grid. Just as the World Wide Web and browser software make millions of information sources easily available to a desktop computer, a distributed computing environment or Grid integrates computing, data storage, and instrumentation systems that are managed by various organizations in dispersed locations so that they function and appear to the user as one system.

DSD has several areas of expertise which could be applied to Grid-based biomedical informatics, as described below.

Production Grid Services

http://www.doesciencegrid.org/
Keith Jackson, KRJackson@lbl.gov

Berkeley Lab is leading several important Grid development efforts. The DOE Science Grid project is developing the components necessary to use Grids in production. The next-generation Grid will be based on Web services, and DSD is a core developer of these Grid Web service capabilities. Members of DSD play a role in defining Grid-capable Web services standards (WS-RF), developing concrete implementations of Grid services (pyGridware), participating in interoperability testing between different Grid service implementations (WS-RF Interop), and implementing production Grids (DOE Production Science Grid). We can provide real-world experience in the design, development, and management of production Grids.

Grid Services Architecture

http://dsd.lbl.gov/firefish/
Wolfgang Hoschek, whoschek@lbl.gov

http://dsd.lbl.gov/NetLogger/
Brian Tierney, BLTierney@lbl.gov

DSD has the necessary experience to assist in implementing the architecture underlying a new production Grid. One major challenge to Grid implementations is to develop an architecture which supports dynamic, programmatic access to local and remote data sources and metadata repositories without sacrificing the high-performance requirement. The DSD Firefish project proposes to address this problem through a peer-to-peer service infrastructure which supports queries across autonomous data repositories, including dynamic and heterogeneous information sources. Another challenge is diagnosing failures in complex distributed computing environments. The DSD Netlogger project provides a high performance network logging toolkit which has been useful in diagnosing failures in Grid computing environments

Security/Authentication/Access Control

http://www.dsd.lbl.gov/SGT/
Mary Thompson, MRThompson@lbl.gov

The DSD Distributed Secure Grid Infrastructure group has considerable experience in developing and deploying authorization services for heterogeneous, widely distributed resources that require a combination of local and centralized access control. In particular, the Akenti authorization service which they have developed builds on certificate-based authentication of users and signed, verifiable user attributes to determine resource access based on both centrally and locally determined access policies. This experience is directly applicable to providing secure access control of confidential data, such as patient medical records, which are geographically and administratively distributed. The SciShare project, a secure filesharing system developed at LBNL, proposes to apply these principles of policy-based access control both to group query protocols and to individual peer-to-peer data sharing responses.

Workflow

Charles McParland, CPMcParland@lbl.gov

Workflows are managed, automated processes with a predefined control structure and data flow. Although workflow is typically used to describe standardized business processes, many scientific projects have developed workflows which are regularly used in their research. Practically every online biological database is built using a custom-designed workflow, implemented in a hand-crafted workflow management system. These workflow management systems are often brittle and must be manually monitored and altered to adapt to unanticipated failures. A new generation of workflow languages and implementations, based on the orchestration of data and control flow between independent Web services, promises to make design and management of workflows simpler and more reliable.

DSD is developing a graphical user interface for composition and monitoring of Grid-based workflows. This GUI will allow scientists to design workflow networks by “dragging and dropping” existing Grid service components and visually monitoring the resulting workflow execution. The underlying workflow is based on the Business Process Execution Language (BPEL). Through careful design of the workflow file formats and programmatic interfaces between Grid services, DSD’s tools enable end users to submit complex queries through a friendly Web interface, while also allowing the production Grid managers to execute and monitor regularly scheduled complex, compute- and data-intensive production workflows.

Collaboration Technologies

http://www.dsd.lbl.gov/Collaboratories/
Deb Agarwal, DAAgarwal@lbl.gov

Distributed scientific collaborations need tools to support development of topical communities of researchers working together regardless of actual physical location. The Pervasive Collaborative Computing Environment (PCCE) is targeted at providing connectivity between collaborating researchers and promoting a sense of community. At its core, the PCCE provides presence, instant messaging and file sharing capabilities. Support for videoconferencing and application sharing capabilities is also provided by leveraging the available commercial and research solutions.

Semantic Grids

Deb Agarwal, DAAgarwal@lbl.gov

One exciting area of research that DSD is beginning to explore is the Semantic Grid. The Semantic Grid draws from the Semantic Web, a common framework which allows data to be shared and reused across administrative boundaries. The Semantic Grid adapts the Semantic Web to the service-oriented architecture being used to develop next-generation Grid services. We believe this approach has the potential to form the fundamental architecture for Grid-based biomedical informatics similar to the caBIG project described below.

caBIG

http://cabig.nci.nih.gov
David Konerding, DEKonerding@lbl.gov

The National Cancer Institute's (NCI) support for Grid development is currently focused on the Cancer Biomedical Informatics Grid (caBIG), a cancer-oriented biomedical informatics network which connects cancer-related elements of data, tools, individuals and organizations that leverage their strengths and expertise globally. The Grid's primary goal, which directs its infrastructure and software development, is to eliminate suffering of cancer through the integration of disparate cancer research data.

The caBIG project has made progress in implementing a controlled vocabulary, a standard for metadata based on the vocabulary, and an object library based on the metadata standard. Their object library implementation uses the Web services paradigm for exposing operations on the metadata. These facts suggest that caBIG is well poised to take advantage of standard Grid middleware built around WS-RF, the primary collection of standards for implementing Grid services on top of Web services. However, there is a large gap between developing software that uses Web services technology and implementing a successful high performance distributed computing Grid.

High Performance Computing Research Department

http://hpcrd.lbl.gov/
Juan Meza, JCMeza@lbl.gov

The High Performance Computing Research Department (HPCRD) conducts research and development in mathematical modeling, algorithmic design, software implementation, and system architectures, and evaluates new and promising technologies. They collaborate directly with scientists, in fields ranging from materials sciences to climate modeling to astrophysics, to solve computational and data management problems. They also create visualizations to help scientists gain new physical insights and make the data more comprehensible. Some current projects are described below.

ProteinShop: Solving Protein Structures from Scratch

http://www-vis.lbl.gov/Research/ProteinShop/
Sylvia Crivelli, SNCrivelli@lbl.gov

One of the grand challenges in biology is the protein structure prediction problem: how to determine the three-dimensional structure of a protein, or its “native structure,” given only its sequence of amino acids. Many protein structure prediction methods are based on the Nobel Prize-winning theory developed by Chris Anfisen, which specifies that the native structure of a protein is that in which the free energy is at the minimum value. However, finding the global minimum is a complex task, as the number of 3D permutations increases exponentially with the number of amino acids in the sequence. There is no direct analytic solution. Only an exhaustive search of the entire solution space is guaranteed to produce an optimal answer, and there are theoretically an infinite number of combinations to be tested.

With this in mind, researchers at Berkeley Lab, Lawrence Livermore National Laboratory, and the University of California, Davis, have developed ProteinShop, a software application that facilitates a solution to the protein prediction problem through a combination of unique features and capabilities. These include:

- Helping researchers automatically generate tertiary protein structures “from scratch” by using the sequence of amino acids and secondary structure predictions obtained from public prediction servers accessible via the Internet.
- Enabling users to apply their accumulated biochemical knowledge and intuition during the interactive manipulation of structures.
- Facilitating interactive comparison and analysis of alternative structures through visualization of free energy computed during modeling.
- Accelerating discovery of low-energy configurations by applying local optimizations to user-selected protein structures.
- Accelerating time-to-solution by enabling a user to interact with and “steer” a remotely running global optimization code.

One important advantage of ProteinShop over all other modeling tools is that it supports interactive, real-time motion of protein parts with respect to each other without breaking the protein’s chemical structure (Figure 3). To maintain biochemical integrity during manipulation, ProteinShop uses an inverse kinematics (IK) algorithm to constrain molecule motion. IK is typically used in robotics and character animation applications for modeling articulated motion of jointed limbs. In the case of ProteinShop, the joints being articulated are the dozens to hundreds of dihedral angles in a typical protein.

Using an IK-based constraint system during molecular transformation means that all the dihedral angles along the molecular backbone move “in tandem” in a physically meaningful way, like the joints in our limbs. This translates into preservation of the chemical validity of the molecule during manipulation. Our implementation runs at interactive rates even for large molecules (hundreds of atoms) that may contain many hundreds of joints.

ProteinShop cannot solve the entire protein structure prediction problem on its own. However, it helps transform an unsolvable problem into one that can be approached through computational means by providing researchers with specific features that are

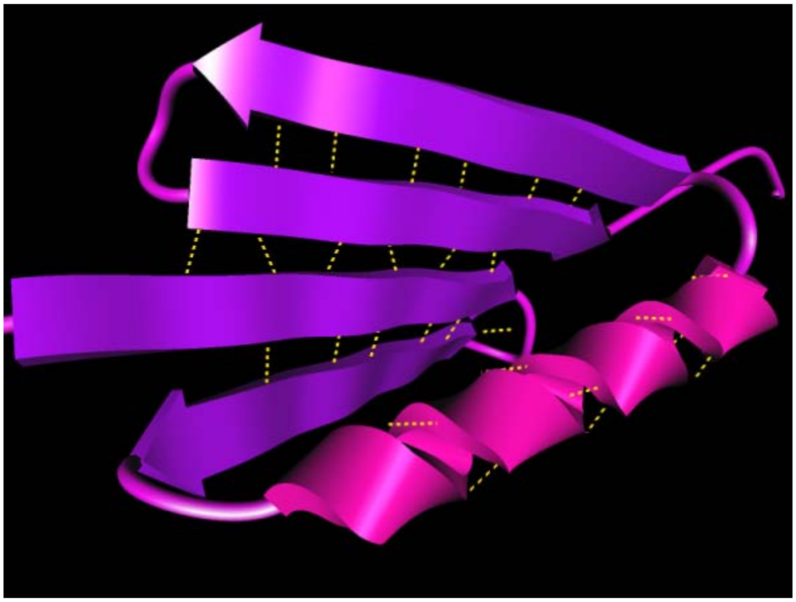


Figure 3. Final 3D protein structure developed with ProteinShop, ready for optimization.

not present in any other software program. ProteinShop provides the tools that, combined with human knowledge, enable researchers to narrow an infinite search space into the one best structure far more rapidly than ever before. While crystallography can provide accurate information on protein structure, it is an extremely time-consuming process. The promise of computational methods is that they will evolve to achieve the accuracy of crystallography while taking only a fraction of the time. The potential impact on humanity from a better understanding of a protein's shape and function is profound, including the development of new cures for diseases that are presently not curable.

While ProteinShop has proven to be a powerful tool in advancing solutions to the protein structure prediction problem, its fundamental capabilities can contribute to progress in a wide range of biological applications, including the design of new drugs and biomaterials. As additional capabilities are added to the program, ProteinShop will help research in the following areas:

- **Drug docking.** With the addition of a capability for visualizing multiple proteins and/or drug compounds simultaneously, it will be possible to see how proteins bind with existing drugs.

- **Drug design.** Pharmaceutical researchers will be able to use ProteinShop for direction in developing new drugs before undertaking the expensive process of synthesizing and then testing each new compound. By using ProteinShop, scientists will be able to model new compounds and then conduct virtual experiments to see how well they bind with the targeted protein. ProteinShop's ability to manipulate protein structures and potential binding sites will enable researchers to shape these compounds until they perform the desired functions.

Phylo-VISTA: Visual Analysis of Sequence Data

http://www-gsd.lbl.gov/phylovista

Wes Bethel, EWBethel@lbl.gov

Large-scale genome sequencing efforts are producing an abundance of sequence data for an increasing number of organisms. Comparative analysis of DNA sequences from multiple species is a powerful strategy for identifying functional elements such as genes and their regulatory sequences. This approach is based on the assumption that functionally important elements evolve more slowly than nonfunctional genomic regions due to selective constraints. Several efforts are ongoing to sequence and analyze targeted genomic regions for conservation across many evolutionarily diverse species.

This work is motivated by the need to perform visual data analysis of sequence data across an arbitrary number of species, along with the need to perform such analysis at varying levels of resolution to accommodate the explosive growth in the size of sequence data.[6] The system, called "Phylo-VISTA" (short for Phylogenetic Visualization Tools for Alignments), supports visualization of multi-species sequence comparison using phylogenetic trees as a framework to guide the display and analysis of conservation levels across tree nodes. Phylo-VISTA supports interactive visual analysis of pre-aligned multi-species sequences by performing the following functions:

- display of a multiple alignment sequences with the associated phylogenetic tree
- computation of a similarity measure over a user-specified window for any node of the tree
- visualization of the degree of sequence conservation by a line plot (Figure 4)
- presentation of comparative data together with annotations.

Berkeley Lab's successful VISTA comparative genomics software was used as a basis for the visualization of multiple alignments along with an associated phylogenetic tree. In order to visualize multiple alignment data, several extensions to VISTA were developed. For pairwise comparison, VISTA requires a user to select one of the sequences as the base sequence. To create a VISTA plot, the user moves a window over an alignment, and VISTA calculates the percent-identity between the base sequence and the aligned sequence over a window surrounding each base pair. The x-axis represents the base sequence, and the y-axis represents percent-identity. The alignment data is projected on the base sequence, and annotations are also presented in the plots. VISTA displays the size and location of gaps in the aligned sequence. VISTA can show annotations only for the base sequence, and gap information can be displayed only for the sequences other than the base. Thus, using one sequence as base results in loss of information about the gaps in the base sequence and corresponding data of other sequences.

Therefore, Phylo-VISTA uses the entire multiple alignment as a base. As a result, Phylo-VISTA is capable of displaying location and length of gaps in all sequences. In addition, to visualize all available data for each sequence, Phylo-VISTA provides annotations beyond those associated with a single base sequence. Multi-species plots allow a user to analyze desirable features in a single visualization (e.g., to view and analyze gaps and annotations of all sequences being compared). A sum of weighted pairwise similarity measures is used for comparing more than two sequences.

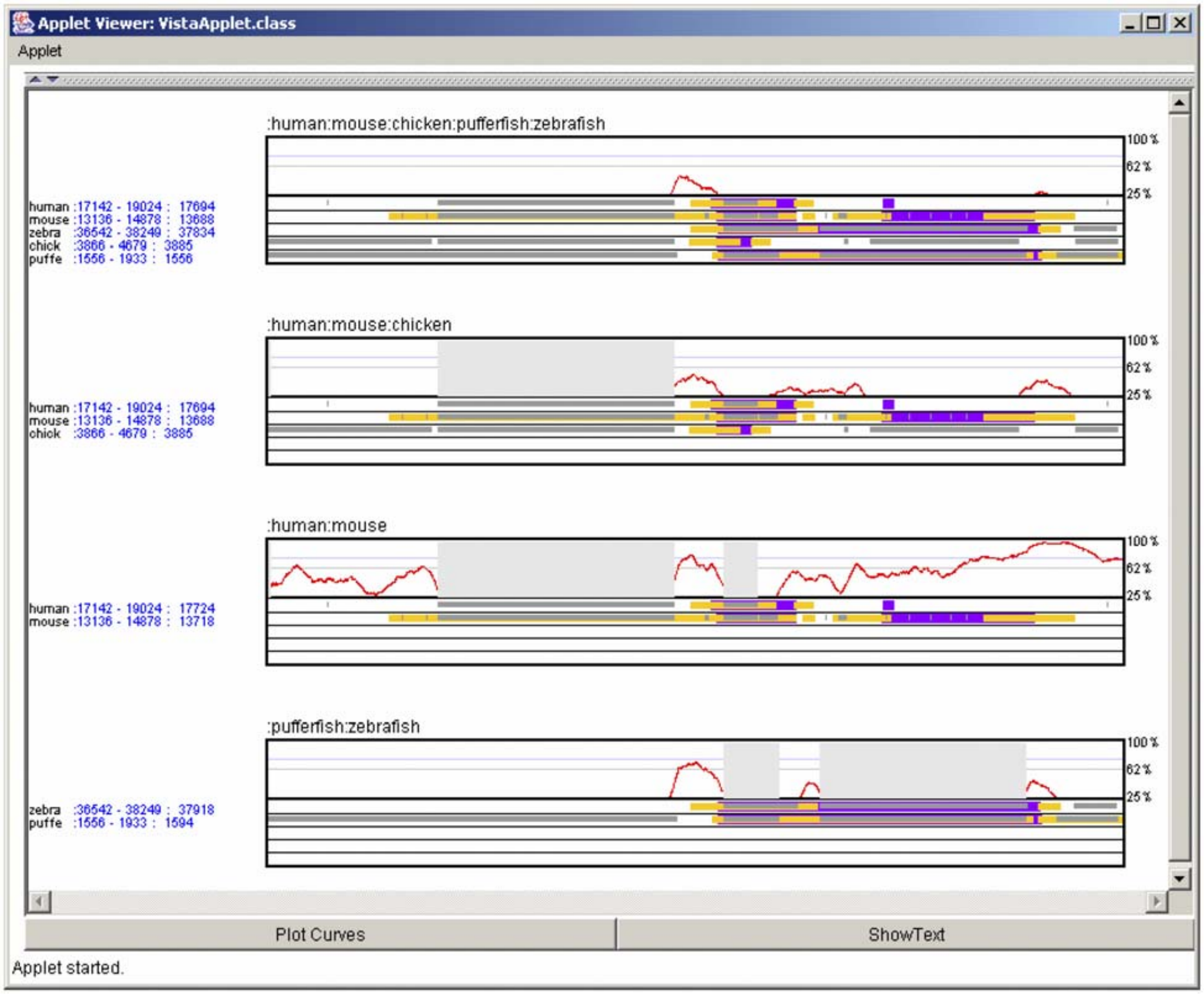

Applet Viewer: VistaApplet.class
Applet
:human:mouse:chicken:pufferfish:zebrafish
100 %
62 %
25 %
human :17142 - 19024 : 17694
mouse :13136 - 14878 : 13688
zebra :36542 - 38249 : 37834
chick :3866 - 4679 : 3885
puffe :1556 - 1933 : 1556
:human:mouse:chicken
100 %
62 %
25 %
human :17142 - 19024 : 17694
mouse :13136 - 14878 : 13688
chick :3866 - 4679 : 3885
:human:mouse
100 %
62 %
25 %
human :17142 - 19024 : 17724
mouse :13136 - 14878 : 13718
:pufferfish:zebrafish
100 %
62 %
25 %
zebra :36542 - 38249 : 37918
puffe :1556 - 1933 : 1594
Plot Curves
ShowText
Applet started.

Figure 4. Similarity plots for human, mouse chicken, pufferfish, and zebrafish. The height in the line plot corresponds to percent-similarity. Minimum conservation is set to 25%. Below the line plots the annotations for each sequence are given. Gray rectangles indicate gaps, blue rectangles represent exons, and yellow rectangles show user-supplied conserved features. The text on the left side of the annotations shows the name of the sequence, its selected start and end positions, and the current cursor position. The peak visible in all plots indicates a region conserved in all sequences.

Phylo-VISTA has been integrated into the global alignment program LAGAN, available at (http://lagan.stanford.edu). The developers plan to integrate Phylo-VISTA with a search engine for transcription factor binding sites. Limited display area and limited display resolution are physical restrictions which must be considered when developing interactive sequence data exploration methods for the comparison of several hundred sequences, each one consisting of several million base pairs. Additional innovative visualization techniques will be developed to meet this challenge.

DEB: A Data Entry and Browsing Tool for Biology Data

Arie Shoshani, AShoshani@lbl.gov

As part of DOE's Genomes to Life project, the Scientific Data Management Group at Berkeley Lab has developed a Web-based Data Entry and Browsing (DEB) tool that will facilitate capturing the metadata from experiments in a computer-searchable form. The system is built on top of the Object-Based Database Tools (the OPM database tools developed previously at Berkeley Lab), and the data is stored in the Oracle database system.

The design of the tool is based on input and insight from the biologists on the project as to the features that a biologist will find useful. The interface design mimics the familiar laboratory notebook format. The DEB system can generate pages that can be used to populate a physical notebook.

One of the strong features of the design is the ability to browse through previous experiments and describe the next experiments with minimal effort based on existing entries in the database. Another important feature is providing security to the object-instance level by users and groups, where each experiment or related objects can be assigned read, write, and delete permission to other users/groups.

The most powerful capability of the DEB system is that it is schema-driven, that is, the interfaces are generated automatically from the schema definition. Therefore, new database schemas can quickly be used to generate DEB interfaces as well as the underlying Oracle database for them. This feature makes this tool immediately applicable to new and/or changing databases.

BGDM: Biopathways Graph Data Manager

http://pueblo.lbl.gov/~olken/graphdm/graphdm.htm

Frank Olken, FOlken@lbl.gov

Berkeley Lab is constructing a general-purpose graph data management system, which will be adapted (by means of suitable schemas) to support biopathways and protein interaction network databases for microbial organisms**.** The ultimate vision is a common graph database management system and query language that would greatly ease the development and use of such biological databases. The type of graphs targeted include metabolic pathways, signaling pathways, gene regulatory networks, protein interaction networks, taxonomies, and contact graphs.

The project is building a query language and graphical user interface for expressing queries over such graphs, as well as a query processing engine. Graph query processing will take place largely in main memory. A relational DBMS will be used as the back end to provide persistent storage of the graphs and some (simple) portions of the query processing.

Many of these graph queries are motivated by interest in comparative biology over biopathways and other graph data. Comparative biology has proven to be a fruitful approach when the comparisons concern sequence data (DNA, RNA, or protein sequences) or shape comparisons (e.g., protein structures).

The developers anticipate that as metabolic pathways, signaling pathways, and genetic regulatory networks are determined for many microbes, there will be mounting interest in doing comparative biology computations on these graphs.

CryoEM Reconstruction Software

http://www.macro-em.org
Esmond Ng, EGNg@lbl.gov

High-resolution cryo-electron microscopy (CryoEM) has emerged as a powerful means of determining the molecular structural of complex biological assemblies. The estimation of three-dimensional density map of a large molecule or molecular assembly from a large number of two-dimensional electron microscopy images of isolated (single) particles with random and unknown orientations can now be obtained at a resolution of 10–12 angstrom or higher. Examples of successful applications are the human transcription complex [7], the *E. coli* ribosome [8], bovine NADH:ubiquinone [9], and spliceosomal U1 snRP [10]. With this kind of resolution, it is now possible to extract detailed features of a molecular assembly to allow structural biologists to elucidate the mapping of molecular subunits and active sites within a complex and to visualize the different assembly stages and conformational changes that accompany biological activities.

The LBNL team has been working with structural biologists from Berkeley Lab, University of California at Berkeley, Wadsworth Center in Albany, NY, Baylor School of Medicine, and the University of Texas Houston Medical School on extending the capability and efficiency of the current generation of single-particle CryoEM reconstruction software. The work is currently funded by an NIH program project. CRD’s involvement in this project aims at bridging the gap between CryoEM research and the current hardware and algorithmic innovation in scientific computing. A new framework for CryoEM reconstruction has been developed by posing the problem as a nonlinear optimization problem. The team is in the process of developing efficient and stable numerical algorithms for solving this type of optimization problem.

Because CryoEM computation involves an extremely large volume of data (e.g., 106 particle images, each with 256 × 256 pixels), parallelization is absolutely necessary in order to achieve high throughput. We have been actively engaged in the parallelization of the current generation of CryoEM reconstruction software for large-scale distributed memory systems to achieve scalable performance.

Imaging and Informatics

http://vision.lbl.gov/
Bahram Parvin, B_Parvin@lbl.gov

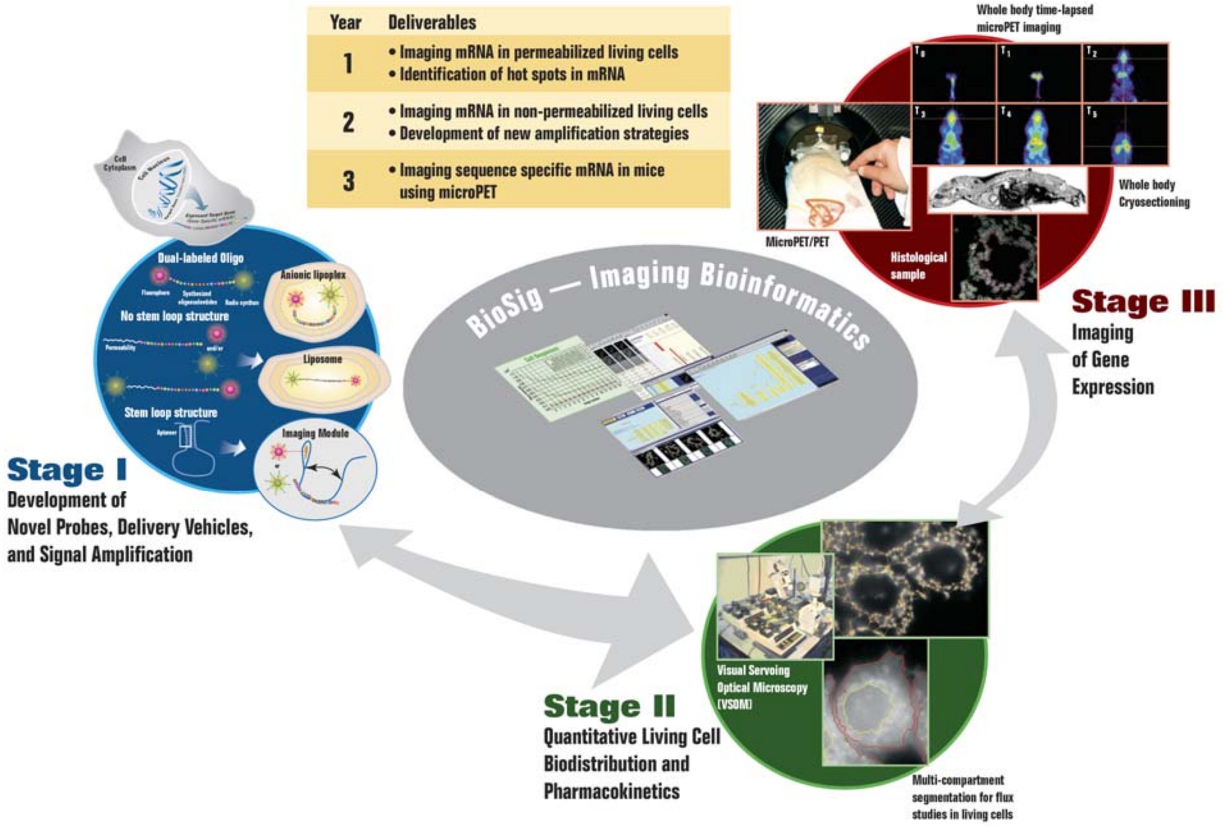
The Imaging and Informatics Group has three joint collaborative projects with Berkeley Lab’s Life Sciences Division: Nuclear Imaging of Gene Expression, Mechanism of Tissue Response to Low Dose Radiation, and Mechanism of Human Epithelial Cell Response to HZE Radiation. The common thread in these projects is quantitative representation of spatio-temporal data through the development of novel algorithms and the informatics tool annotation of images and information about images. The development of algorithms has been driven by model-based geometric representation of data or leaning by example. The development of the informatics framework is being driven through common data models (whenever possible), controlled vocabulary, and functional portals for data analysis and representation.

The primary objective of the Nuclear Imaging of Gene Expression (NIGE) consortium — Ames National Laboratory at Iowa State University, Berkeley Lab, and Brookhaven National Laboratory — is to overcome the technical challenges of directly imaging the expression of any gene in living cells, animals, and humans. The

NIGE consortium combines the expertise of these three institutions in an integrated three-step approach to developing gene-specific mRNA imaging agents: probe chemistry and signal amplification, kinetics and uptakes and washout in living cell studies, and in vivo biodistribution and pharmacokinetics (Figure 5). Synthetic antisense compounds and novel aptamer-based probes that can be detected using both fluorescent and radioactive labels will be used in conjunction with fluorescence microscope, single-photon emission computed tomography (SPECT), and positron emission tomography (PET) imaging modalities.

As a result of technical challenges associated with direct gene imaging, the consortium is focusing on the same gene target (uterocalin/24p3) and employing the same nucleotide sequence, the same model cell systems, and the same animal models. Classic biochemical and molecular biology techniques are being used to obtain and verify the selection of a small (15–20 nucleotide) consortium antisense sequence that will target an accessible region of the uterocalin mRNA target.

This multi-institutional project is being supported through the development of a new bioinformatics system for molecular imaging with four distinct concepts: imageable agent, imageable target, amplification technique, and imaging instrument. Quantitative techniques are being integrated throughout the programmed project to characterize kinetics and amplification strategies associated with each imageable probe.



Year	Deliverables
1	• Imaging mRNA in permeabilized living cells • Identification of hot spots in mRNA
2	• Imaging mRNA in non-permeabilized living cells • Development of new amplification strategies
3	• Imaging sequence specific mRNA in mice using microPET

Cell Cytoplasm
Cell Nucleus
Dual-labeled Oligo
Fluorophore
Synthesized oligonucleotides
Radio synthon
No stem loop structure
Permeability
and/or
Stem loop structure
Aptamer
Anionic lipoplex
Liposome
Imaging Module
Stage I
Development of Novel Probes, Delivery Vehicles, and Signal Amplification
BioSig — Imaging Bioinformatics
Whole body time-lapsed microPET imaging
T0
T1
T2
T3
T4
T5
MicroPET/PET
Whole body Cryosectioning
Histological sample
Stage III
Imaging of Gene Expression
Visual Servoing Optical Microscopy (VSOM)
Multi-compartment segmentation for flux studies in living cells
Stage II
Quantitative Living Cell Biodistribution and Pharmacokinetics

Figure 5. The NIGE consortium has initiated an integrated three-step approach to developing gene-specific mRNA imaging agents.

Furthermore, plans are under way to develop a computational model of RNA folding in isolating hot spots for probe development.

Studies associated with radiation biology have the potential to generate large amounts of data targeting protein and gene expression under exogenous conditions. These data need to be cataloged and quantified for phenotypic characterization and predictive modeling. Toward this objective, the BioSig imaging bioinformatics system has been developed to manage images and information about images.[11]

The central organizing metaphor of BioSig is a networked graph representation that relates experimental parameters, images collected through different modes of microscopy, quantitative representation of images, and portals for data analysis. The networked graph represents the flow of data where each measurement can be tracked to a particular image, a sample and how it was conditioned and stained, and the species that sample came from.

Biological heterogeneity requires a substantial number of experiments (for each endpoint), which leads to a large amount of images and annotation data. The BioSig analytical and informatics framework enables annotation of experimental designs (metadata), conversion of images to quantitative features through novel algorithms, presentation of different views of pertinent information, and construction of probabilistic predictive models based on phenotypic analysis. The informatics framework is being revised to leverage data standards in microscopy, gene expression, and animal models. Plans are under way to integrate graphical-based learning models with the informatics component to facilitate predictive modeling.

Modeling Cellular Networks

Phil Colella, PColella@lbl.gov

Current research is creating enormous amounts of data on how the development and behavior of cells is governed by a complex network of reacting chemical species that are transported within a cell and among cellular communities. The Applied Numerical Algorithms Group is working with the Arkin Laboratory in Berkeley Lab's Physical Biosciences Division and UC Berkeley's Bioengineering and Chemistry departments to develop the software infrastructure to turn this data into quantitative and predictive models. These models will contribute to a deeper understanding of cellular regulation and development, more efficient discovery of cellular networks, and a better ability to engineer or control cells for medical, agricultural, or industrial purposes.

XMDR: Extended Metadata Registry Project

http://hpcrd.lbl.gov/SDM/XMDR/
Bruce Bargmeyer, BEBargmeyer@lbl.gov

Berkeley Lab is leading a research project that is a part of an inter-agency international collaboration on ecoinformatics, which fosters application of information science and information technology for the environment. The project involves EPA, DOD, USGS, DOE, HHS, USDA, European Environment Agency, United Nations Environmental Program, Mayo Clinic, and other organizations. The initial emphasis is on environmental and biomedical data/metadata; however, the project can extend to any domain of information. The XMDR project's goal is to advance state-of-the-art metadata registries to record difficult semantic structures and to provide semantic services.

A key element of the semantic web, semantic grids, and semantics-based computing is the expression of vocabularies, so that computers can take much of the drudgery out of accessing and utilizing information. The data is made “smarter” so that Web-based tools can do intelligent things with the data without programmers having to hardwire understanding about the data into computer programs. The work involves the ISO/IEC 11179 Metadata Registry family of standards, a variety of terminology standards, and W3C recommendations and best practices. The project focuses on data and metadata, which may be unstructured or may be structured as data elements, thesauri, taxonomies, graphs, entity-relations, axiomatized ontologies, and more. Structures of increasing complexity are necessary to represent the complexity inherent in health and environmental data and knowledge. Registering interrelationships between the structures — for example, between data elements and terminology structures — is important as data management techniques progress toward deeper semantics management.

The XDMR project builds on efforts that scientists, organizations, and agencies have made to develop data standards for databases, XML schemas, data interchange, and more. It goes to the next level of integrating and sharing health and environmental data, making it more useful to wide spectrum of internal and external users. It also builds on many large- and small-scale efforts to define vocabularies, glossaries, thesauri, and other terminologies.

The participants will test and demonstrate the extended capabilities with a reference implementation researched and developed at Berkeley Lab. The extended capabilities will be documented as proposals for a revision of ISO/IEC 11179 Parts 2 and 3 to become Version 3 of that family of standards. The revised standard will serve as the design for operational implementations of ISO/IEC 11179 metadata registries by the participating agencies, other organizations and vendors. It is expected that the reference implementation will also be useful in further research on semantic services for Grid computing and semantics-based computing.

Data Mining and Machine Learning Algorithms for Genomics

http://crd.lbl.gov/~cding/papers/
Chris Ding, CHQDing@lbl.gov

The Scientific Computing Group is applying data mining and machine learning algorithms to a variety of problems in genomics. Projects include:

- Using support vector machines (SVMs) to predict protein folds, protein binding, etc. Researchers used the SVM method to predict the fold of remote homologs (sequence similarity less than 25% identity) with 50% accuracy. [12]
- Using partially labeled learning methods. Very often, we have only positive examples, such as the known functional RNAs, among vast number of unlabeled samples. We developed a method to predict new functional RNAs using this positive-only learning mechanism.
- Prediction of new protein complexes and protein-protein interactions using spectral clustering methods. This is equivalent to finding densely connected regions in a protein interaction network. We have developed several very effective clustering algorithms using eigenvectors. These newer methods perform significantly better than standard hierarchical clustering methods.
- Multi-level graph contraction for RNA secondary structure. We are developing a new systematic graph representation for RNA that unifies the existing tree and dual graphs, and incorporates more information such as base stacking and coaxial stacking.

Signaling Pathways Analysis for Cancer Treatment

Juan Meza, JCMeza@lbl.gov
Ali Pinar, APinar@lbl.gov

Many aspects of cancer result from the deregulation of interacting signaling pathways which control a cell. Studies of these deregulated pathways have revealed genomic, epigenomic, and biological differences between tumor and normal cells that can be targeted therapeutically. Over 2,000 potential targets have been identified, and about 800 therapeutic agents are currently being clinically evaluated.

Although this wealth of targets and therapeutic agents bodes well for future cancer treatment, it will take years of research and enormous financial commitment to capitalize on these developments. One important hurdle is the remarkable genomic and biological variability between clinically and pathologically similar tumors. Similar tumors respond differently to targeted therapeutics due to underlying genomic differences. This calls not only for developing therapeutics, but also identifying combinations of drugs and targets for effective cancer treatment.

A collaboration between the Scientific Computing Group and the Life Sciences Division at Berkeley Lab is developing a computational model to analyze signaling networks, to identify subsets of patients that will respond well to pathway-targeted therapeutics, and conversely to find combinations of therapeutics that will be effective for specified patient groups.

Signaling pathways can be considered as a graph, where signals correspond to edges and signaling biological entities correspond to vertices. Inhibiting cancer development then can be considered as “cutting” this graph, that is, removing a subset of edges of minimum cardinality to disconnect all paths reaching a specified target.

Although the problem of merely disconnecting a graph has been studied as the maximum-flow minimum-cut problem, additional biological constraints make this problem much harder. When we cut the signaling network, we not only need to stop the cancer, but also allow the cell to retain its normal activities. We are currently working on designing algorithms for this purpose.

Infrastructure/Support (Information Technologies and Services Division)

http://www.lbl.gov/ITSD
Sandy Merola, AXMerola@lbl.gov
Tammy Welcome, TSWelcome@lbl.gov

The Information Technologies and Services Division (ITSD) supports Berkeley Lab's scientific mission by developing and maintaining the Laboratory's computing, information and communications infrastructure. ITSD provides desktop computing support, centralized email and calendaring services, networking, telecommunications, cybersecurity, and business information systems to laboratory scientists, business professionals, and management.

ESnet: Energy Sciences Network

http://www.es.net/
William E. Johnston, WEJohnston@lbl.gov

ESnet is a high-speed network serving thousands of DOE scientists and collaborators worldwide. A pioneer in providing high-bandwidth, reliable connections, ESnet enables researchers at national laboratories, universities and other institutions to communicate with each other using the collaborative capabilities needed to address some of the world's most important scientific challenges.

Managed and operated by the ESnet staff in the ITSD division, ESnet provides direct connections to all major DOE sites with high performance speeds, as well as fast interconnections to more than 100 other networks. Funded principally by DOE's Office of Science, ESnet services allow scientists to make effective use of unique DOE research facilities and computing resources, independent of time and geographic location.

ESnet's collaboration services include video, audio, and data conferencing; security tools; and the DOEGrids Certificate Service, which issues identity certificates to individual subscribers within virtual organizations and service certificates for Grid services.

Scientific Cluster Support

http://scs.lbl.gov/
Tammy Welcome, TSWelcome@lbl.gov
Gary Jung, GMJung@lbl.gov

The Scientific Cluster Support program provides startup and ongoing system administration services to Linux clusters at Berkeley Lab. The program is designed to promote and facilitate the cost-effective use of Linux clusters in scientific research. It helps research groups with choosing and procuring the right cluster configuration, setting up and configuring the system, helping researchers get their applications onto the cluster, then providing ongoing systems administration and cybersecurity support.

The SCS program currently supports 13 small to medium-sized individual clusters for various scientific projects throughout Berkeley Lab, including two Life Sciences Division clusters: a genomics cluster for Mike Eisen's work to determine transcription regulation networks in model organisms, and another cluster for Priscilla Cooper and John Tainer to determine structures of DNA repair proteins and other large proteins from data collected from the novel Sybils beam line at the Advanced Light Source.

IT Consulting and Support

Tammy Welcome, TSWelcome@lbl.gov

In addition to the providing the general computing infrastructure for the lab, ITSD provides specialized consulting and support to various science projects to meet their specific scientific computing requirements. This includes areas such as the following:

- consulting on computing, infrastructure, and facility requirements for science and center proposals
- assessments of existing scientific division-provided IT infrastructure
- consulting and support services to meet special IT needs, including cybersecurity and networking.

References

[1] S. Aparicio et al., “Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes,” Science **297**, 1301 (2002).

[2] J. Hou, G. E. Sims, C. Zhang, and S.-H. Kim, “A global representation of the protein fold space,” PNAS **100**, 2386 (2003).

[3] K. Brown et al., “Solution structure of the 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine C8-deoxyguanosine adduct in duplex DNA,” PNAS **98**, 8507 (2001).

[4] B. Nagar et al., “Structural basis for the autoinhibition of c-Abl tyrosine kinase,” Cell **112**, 859 (2003).

[5] L. A. Peterson et al., “Solution structure of an O6-[4-oxo-4-(3-Pyridyl)butyl]guanine adduct in an 11mer DNA duplex: Evidence for formation of a base triplex,” Biochemistry ASAP, DOI: 10.1021/bi035217v (2003).

[6] N. Shah, et al., Olivier Couronne, Len A. Pennacchio, Michael Brudno, Serafim Batzoglou, Wes Bethel, Edward M. Rubin, Bernd Hamann and Inna Dubchak, “Phylo-VISTA: Interactive visualization of multiple DNA sequence alignments,” Bioinformatics **20,** 636–643 (2004).

[7] Frank Andel III, Andreas G. Ladurner, Carla Inouye, Robert Tjian, and Eva Nogales, “Three-dimensional structure of the human TFIID-IIA-IIB complex,” Science **286**, 2153–2156 (1999).

[8] Irene S. Gabashvili, Michelle Whirl-Carrillo, Michael Bada, D. Rey Banatao, and Russ B. Altman, “Ribosomal dynamics inferred from variations in experimental measurements,” RNA **9**, 1301–1307 (2003).

[9] N. Grigorieff, “Three-dimensional structure of bovine NADH:ubiquinone oxidoreductase (complex I) at 22 Å in ice,” J. Mol. Biol. **277**, 1033–1046 (1998).

[10] H. Stark, P. Dube, R. Luhrmann, and B. Kastner, “Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle,” Nature **409**, 539–42 (2001).

[11] B. Parvin, Q. Yang, G. Fontenay, and M. H. Barcellos-Hoff, “BioSig: An Imaging Bioinformatics System for Phenotypic Analysis,” IEEE Transactions on Systems, Man, and Cybernetics **33**, 814–824 (2003).

[12] C. Ding and I. Dubchak, “Multi-class protein fold recognition using support vector machines and neural networks,” Bioinformatics **17**, 349 (2001).

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.