DE-FG02-96ER62327

DATABASE MANAGEMENT RESEARCH FOR THE
HUMAN GENOME PROJECT

Final Progress Report
for period: 02/01/99 – 06/14/00

Carol J. Bult, Ph.D., Principal Investigator

The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609

Prepared for
THE U.S. DEPARTMENT OF ENERGY
AWARD NO. DE-FG02-96ER62327

DOE Patent Clearance Granted
MP Dvorscak 3.12.01
Date
Mark P. Dvorscak
(630) 252-2393
E-mail: mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office

Approved for the Institution:

Carol J. Bult, Ph.D.
Principal Investigator

Cookie Willems, CRA
Director, Office of Sponsored
Programs

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or 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 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

12 September 2000

Final Report for DE-96ER62327 Database Research Management for the Human Genome Project

Carol J. Bult, Ph.D.
Mouse Genome Informatics Group
The Jackson Laboratory
Bar Harbor, Maine 04609
cjb@informatics.jax.org

Project Goals and Accomplishments

This DOE grant was originally awarded to Dr. Nathan Goodman to investigate database technologies relevant to the Human Genome Project. In February of 1999, I was named as the replacement PI for the award. The goals of the revised project plan included the analysis of workflow and database management systems for large scale mutagenesis studies and the establishment of software and database products to enhance the integration of genome sequence data with existing biological knowledge.

Workflow Management for Mutagenesis

The identification of mutant phenotypes is a critical component of using animal models to help interpret the biology of the human genome. While much attention has been paid to the development of information systems to support genome sequencing, less attention has been paid to workflow management systems for large-scale phenotyping efforts. The PI completed a System Requirements and Workflow Analysis to support large-scale physiogenomics research. This analysis included interviewing researchers and animal care technicians to develop an understanding of various facets of animal management, phenotyping procedures, methods for detection of putative deviants, and procedures for the confirmation of deviants. It also included a review of information systems used by existing mutagenesis facilities in Germany and the UK. Although the final information management plan and database design was geared to the research environment at The Jackson Laboratory there are many components of the system that are applicable to phenotyping and mutagenesis at any institution.

The final product of this component of the revised project plan was a functioning, web-based prototype laboratory information management system to track mutagenesis and phenotyping activities in the Physiogenomics Program at The Jackson Laboratory. The system is being used internally by the Physiogenomics Program and currently tracks information on mutagenesis, breeding, and phenotyping for over 4,000 mice. Additional grant funds have been secured to move this system from a prototype to a production system and to make components of the database available to the general scientific community.

Data Integration

The genome sequence produced as the result of high-throughput sequencing efforts of the human genome and those of model organisms must be integrated with existing and emerging biological knowledge in order to fulfill the promise of the Human Genome Project. The development of software and database management strategies for the integration of sequence and biology requires attention to a very different set of criteria than does the development of information systems designed to track and manage the raw sequence data. Indeed, data integration requires that systems be designed to accommodate both automated computational methods of data analysis as well as analyses that require intensive human involvement.

A second goal of the revised project plan was to develop software and database tools to support the integration of the emerging mouse genome sequence with biological knowledge about mapping data, homology, gene function, and phenotype available for the laboratory mouse. The product produced for this component of the revised project plan was the MouseBLAST sequence similarity server. MouseBLAST is available to the community via the Mouse Genome Informatics Web Site (<http://www.informatics.jax.org>). This resource allows researchers to query the MGI database using protein or nucleotide sequences. If a sequence matches a gene that is already known and described in MGI, a link is provided to an html document that contains extensive hyperlinks to a wealth of knowledge about the mouse gene, related mapping data, polymorphisms, alleles, phenotypes, gene expression patterns, and mammalian homology.

Additional grant funds have been secured by the PI to continue the development and maintenance of the MouseBLAST software product and its related searchable sequence databases.