The Genome Channel and Genome Annotation Consortium*

Ed Uberbacher, Richard Mural, Manesh Shah, Ying Xu, Sheryl Martin
Sergey Petrov, Jay Snoddy, and Morey Parang
Computational Biosciences Section
Life Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6480

Manfred Zorn, Sylvia Spengler, Donn Davey
Lawrence Berkeley National Laboratory

Terry Gaasterland
Argonne National Laboratory

Peter Schaad
The National Center for Genome Resources

Stan Letovsky and Bob Cottingham
The Genome Database

David Haussler
University of California Santa Cruz

Pavel Pavzner
University of Southern California

Chris Overton
University of Pennsylvania

"The submitted manuscript has been authorized by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

To be presented at the The Sixth Department of Energy Contractor and Grantee Workshop, Santa Fe, NM, November 9-13, 1997.

*Research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.
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, make 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.
Summary

Human and model organism sequencing projects will soon be producing data at a rate which will require new methods and infrastructure for users to be able to effectively view and understand the data. A multi-institutional project was recently funded to provide large-scale analytical processing capabilities and we will present the results of several pilot efforts related to this project. The goals of the project are as follows:

- Provide an environment where annotation can be constructed based on multiple interoperable analysis tools and significant available computing power.

- Provide an environment where characterization of long genomic sequence regions can be facilitated and analysis can maintained and updated over time.

- Provide an interactive graphical environment where predictions, features and evidence from many tools can be combined by users into high-quality annotation and visualized by the community.

- Provide high-throughput automated analysis methods which can be configured by genome centers for their use in constructing annotation and facilitating data submission.

- Provide high-quality annotation to large genomic sequence regions which would otherwise go unannotated.

- Provide the community with the best sequence level view of genomes possible.
The components of this system are a number of services, a broker that oversees the distribution and management of tasks and a data warehouse, with services implemented using distributed object technology. Multiple gene prediction is accomplished using several gene finding tools including the GRAIL-EST system and gene annotation from databases such as Genbank is also captured. The data warehouse supporting the Genome Channel view is updated daily by automated Internet agents and event triggers which facilitate analysis procedures. Real time operation of the Genome Channel browser will be demonstrated. A more detailed description of the basic components follows:

**The Genome Channel**

The Genome Channel provides a graphical user interface to comprehensively browse and query assembled sequence placed in the public domain by the Human Genome Project and sequencing of model organisms. It is a JAVA interface tool which relies on a number of underlying data resources, analysis tools and data retrieval agents to provide up-to-date view of genomic sequences, as well as computational and experimental annotation. Navigation from a whole chromosome view to contigs provided by sequencing centers allows one to zoom in on regions of interest to see information about clones, markers, ESTs, computationally and experimentally determined genes, the sequence and sequence source information, related homology and functional information, and hyperlinks to numerous underlying primary data resources.

**Analysis Methods**

Current analysis methods which combine pattern recognition with EST information and protein similarities are capable of accurate and automated analysis of large genomic regions containing complex multiple gene structures. Analysis methods will include the GRAIL EST/Protein homolog system, Procrustes (Pavel Pevsner et al.), and Genie (David Haussler et al.) as well as other tools. The results of multiple tools can be viewed in a common environment, combined mathematically in user specified ways, and used as the basis for the automated or interactive construction of annotation.

**Data Mining Agents**

Maintenance of an up to date description of genomic regions will be based on the use of data mining agents formulated for particular information resources. These agents will make use of several different technologies, such as OPM (Victor Markowitz), Kleisli (Chris Overton), and database indices to facilitate meaningful information retrieval from remote sources. We expect new information related to each gene or genome region to continue to be discovered and actively linked in a long-term ongoing process. The current state of these links will be maintained in the data warehouse.

**The Data Warehouse**
The data warehouse provides and maintains a snapshot of the current view or status of the genomes or genomic regions analyzed in the project. The warehouse is designed to facilitate rapid access by users, visualization tools, and analysis systems. The views contained in the warehouse will be constructed and maintained by processes within this project (such as sequence analysis and information retrieval agents), with additional help from central databases like GDB. It will contain and make available a synthesized best view of genomes from multiple underlying sources.

**Internet Object Request Broker**

Services for data input, analysis, visualization and submission will be facilitated with a distributed underlying Internet architecture using CORBA with an object request broker to manage processes. Compute platforms, analysis servers, databases, etc. will be at a variety of locations and in some cases duplicated depending on need. Specialized computing hardware will be used to facilitate some tasks.

This project is jointly sponsored by the Computational Grand Challenge program of the Office of Computational and Technology Research and the Human Genome Program of the Office of Biological and Environmental Research of the Department of Energy.