Dendrome, A GENOME DATABASE FOR FOREST TREES

B.K. Sherman and D.B. Neale [1/]

Abstract.-- The Dendrome project is creating an archive of genome maps, analysis tools, and data visualization technologies of particular interest to forest molecular geneticists. Services are provided through the Internet worldwide computer network. These services afford access to genome databases, images, announcements, software and expertise. Connections to technically related archives and services are supported.

Keywords: Pinus taeda, database, bioinformatics

INTRODUCTION

The Dendrome project is an attempt to construct a framework for the acquisition, storage and retrieval of genome data of forest trees. Dendrome databases are designed to be research tools for forest geneticists and other forest biologists. The first genome database offered is of loblolly pine (Pinus taeda L.). This software is APtDB (A P. taeda Database) and it includes the maps and data resulting from mapping projects at the Institute of Forest Genetics. This database offers a sophisticated user interface and allows graphical display of both genetic and physical maps. It is intended that maps from other laboratories will be brought into congruence and the data merged into a consensus map. Certain other services are in place and can be accessed via computer networks from locations around the world. These include retrieval of images of autoradiograms, textual representations of data in APtDB, announcements, protocols and interconnection with related services.

The genome maps represent chromosomal locations of genes. The databases will also include information associated with genes such as nucleotide and amino acid sequences, patterns of gene expression, links to metabolic pathways, and sizes of gene families. Information of this type could be used by molecular biologists and physiologists. There will be information on the amount and type of allelic variability of mapped genes. This information could be used by population and evolutionary geneticists. There will be information on map position of quantitative trait loci for important traits. Quantitative

[1/]Computer Scientist and Molecular Geneticist, Institute of Forest Genetics, USDA Forest Service, Pacific Southwest Research Station, 800 Buchanan St., Albany, California, 94716. Internet: bks@s27w007.pswfs.gov dbn@s27w007.pswfs.gov

geneticists and breeders might use this information for experimental purposes or even practicing marker-assisted breeding.

The Dendrome project is funded by the United States Department of Agriculture Agricultural Research Service Office of Plant Genome. It is one component of a three to five year collaborative project to construct prototypes for conifers and other species: Arabidopsis thaliana, maize, soybeans, and wheat. Data collected by the various projects will be cross-referenced into a centralized Plant Genome Database to be administered by the National Agricultural Library.

HARDWARE AND NETWORKING

The primary computer for Dendrome is a Sun Microsystems SPARCstation 2. This computer has 64 megabytes of RAM, and five gigabytes of storage on hard disk drives in addition to tertiary storage on tape media and optical disks. This machine acts as the main server for resources described below.

A LAN (Local Area Network) provides connections to various smaller computers, printers and input devices. A router on the LAN provides a connection to BARRNet (Bay Area Regional Research Network) and the Internet. A Forest Service computer on the LAN gives a connection to the Forest Service wide area network. For a complete discussion of the relationship of the Internet to other computer and telecommunication networks, see Quarterman (1990).

Image acquisition is done with a Cohu charge-coupled device camera, Epix framegrabber, and an Intel-architecture computer. The image data is transferred to the Sun workstation for analysis, manipulation and export. A Umax scanner is available for high resolution color images.

SOFTWARE

The primary repository of genome data is APtDB which employs the ACEDB software (Durbin 1990). Genome data is characterized by deep knowledge of a few items and little about most. This model is not tractable by conventional relational database systems. ACEDB is an attempt to use an object-oriented approach to encode both genetic and physical data. It uses graphical displays, extensive cross-referencing and both a keyboard and mouse to allow easy navigation and visualization of the data and their relationships. Images of the autoradiograms that were used to create the genetic map of loblolly pine are indexed into APtDB.

The X Window System provides a platform-independent mechanism for giving users access to a common graphical interface. Using X, APtDB, which at present runs only on Unix platforms, can be running on the Sun workstation yet have the window displays and user input occur on other machines on the network, an Apple Macintosh, say, or an Intel 80486 computer running Microsoft Windows.

Remote retrieval of data from Dendrome does not require sophisticated computers. A terminal, a modem and a telephone line are sufficient to use two main services: gopher and WAIS. These protocols allow purely textual interaction with the Dendrome databases, but still offer a powerful and concise mechanism for information retrieval.

Gopher is an interactive menu-based tool that allows one to navigate the Internet searching for services. Moving from machine to machine is accomplished without passwords or knowledge of machine names or locations. Machine architectural differences are hidden by the protocol. Gopher can offer access to Wide Area Information Servers (WAIS). WAIS allows huge collections of textual data to be indexed by every word in the data. The index is very large, often larger that the data but once the index is built, novel searches can be accomplished extremely quickly. Both the index and the data are made available to WAIS clients by WAIS servers. Dendrome offers several WAIS indexed collections. Information located via gopher and wais can be stored in a file on the user's local computer or electronically mailed to a colleague. For a more complete description of these protocols, see Krol (1992).

Scientists at the Institute of Forest Genetics are experimenting with various genetic linkage analysis software including: GMendel, Mapmaker, Crimap, and Joinmap. Various image manipulation software is being used and evaluated including: xv, pbmplus and HIPS.

CONCLUSION

Contemporary genetics is characterized by exponentially increasing quantities of data. The quantity and structure of genome data imply the collaborative use of computers and high speed communication networks. Making biological inferences based on that data requires ways of visualizing the data at various levels of detail. Efforts of individual researchers can be amplified if data can be easily shared between laboratories. Many resources are already available to biologists only via the Internet (Smith 1993). The Dendrome project is an attempt to make such technology available for forest biology investigations. The ultimate measure of Dendrome will be whether or not researchers use it.

HOW TO CONTACT Dendrome

Internet electronic mail: Dendrome@s27w007.pswfs.gov

Gopher Server: s27w007.pswfs.gov port 70

The IP address of s27w007.pswfs.gov is 192.131.1.21

Surface mail: **Dendrome Project**
Institute of Forest Genetics
P.O. Box 245
Berkeley, CA 94701

LITERATURE CITED

Durbin, R. and Thierry-Mieg J. 1990 & 1991. ACEDB --A Caenorhabditis elegans Database. (Computer software and genome data in electronic media. Contact Dendrome, or the authors, for current acquisition information.).

Krol, E. 1992. The Whole Internet: Catalog & User's Guide. O'Reilley & Associates, Inc., Sebastopol, CA. 376 p.

Quarterman, J.S. 1990. The Matrix: Computer Networks and Conferencing Systems Worldwide. Digital Press, Burlington, MA. 719 p.

Smith, U.R. 1993. A Biologist's Guide to the Internet. (Available via anonymous ftp from rtfm.mit.edu in pub/usenet/news.answers/biology/guide.) Approx. 20 p.