NOTICE

CERTAIN DATA CONTAINED IN THIS DOCUMENT MAY BE DIFFICULT TO READ IN MICROFICHE PRODUCTS.

CONF-930283--1

PNL-SA--21737

DE93 009482

GNOME VIEW: A TOOL FOR VISUAL REPRESENTATION OF HUMAN GENOME DATA

J. E. Pelkey
D. A. Thurman
R. J. Douthart
G. S. Thomas
V. B. Lortz

February 1993

Presented at the
1993 Symposium on
Applied Computing
February 14-16, 1993
Indianapolis, Indiana

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

PNL-SA-21737

GnomeView: A Tool for Visual Representation of Human Genome Data

Joanne E. Pelkey, Gregory S. Thomas,
David A. Thurman, Victor B. Lortz, Richard J. Douthart

Pacific Northwest Laboratory

Abstract

GnomeView is a tool for exploring data generated by the Human Genome Project. GnomeView provides both graphical and textual styles of data presentation; employs an intuitive window-based graphical query interface; and integrates its underlying genome databases in such a way that the user can navigate smoothly across databases and between different levels of data. This paper describes GnomeView and discusses how it addresses various genome informatics issues.

Key words: Human genome informatics, graphical database interfaces.

Introduction

The Human Genome Project is a worldwide research effort aimed at analyzing the structure of human DNA and determining the location of all human genes [12]. It is estimated that there are up to 100,000 genes [19] encoded by the 3.3 billion nucleotides [20] in the DNA sequence. With each nucleotide represented as a single letter in 10-point type, this sequence would be 5,000 miles long [4].

To provide effective ways to access such enormous quantities of data for biological researchers and students who are not intimately familiar with the details of data acquisition, storage, and retrieval employed by numerous and varied databases, GnomeView is being developed as a graphical user interface to information generated by the Human Genome Project. It targets one of the current informatics goals of the U.S. Human Genome Project, namely, "the creation of database tools that provide easy access to up-to-date [mapping] and sequencing information and allow ready comparison of the data in these several data sets" [12].

There are presently a variety of methods that can be used to access publicly available genome data. Most of these methods support only textual (i.e., c[illegible]-based) input and output. For example, electronic servers accept data and/or requests, perform [illegible] and/or analysis, and mail back results (e.g., [illegible]). Gopher or remote login services provide more interactive query facilities, but are still largely textual [14]. In contrast, visualization is a central theme of GnomeView. Graphical representation of data can reveal patterns that might otherwise be difficult to detect. GnomeView also provides the capability to view data textually; the user may select either or both style(s) of presentation.

The remainder of this paper is organized as follows. The second section provides pertinent information about the problem domain in which GnomeView operates, discussing genome mapping and genome databases. The key capabilities, system architecture, and implementation of GnomeView are detailed in the third section. The fourth section concludes by summarizing the current status of and future plans for GnomeView.

Problem Domain

Two aspects of GnomeView's problem domain—genome mapping and genome databases—are discussed in this section.

Genome Mapping

One of the primary goals of the Human Genome Project is the generation of genetic and physical maps of all human chromosomes. Each chromosome contains a long

molecule of DNA, the chemical of which genes are made. DNA is a double-stranded complex macromolecule in which each strand is a linear array of units called nucleotides. Each nucleotide consists of a sugar, a phosphate, and a base [5]. The four different bases are named A, T, C, and G. The order of the base pairs (each base is paired with its complement in the other strand) determines the information content of a particular gene or piece of DNA—the "genetic code." *Mapping* is the process of determining the position and spacing of genes, or other genetic landmarks, on the chromosomes relative to one another [12].

There are various types of biological mapping, each having an associated metric (or metrics). A metric describes the size of mapped objects and their position relative to one another. The different resolutions of the metrics typically imply a hierarchy of data. Consider an analogy to geographical mapping: a globe is a low resolution map expressed in terms of longitude and latitude, and a street map is a high resolution map expressed in terms of miles. Both maps describe the same thing, but at different levels of detail.

GnomeView currently supports two types of maps: chromosome maps and DNA sequence maps. The standard representation of a chromosome map is a stylized depiction of how the chromosome appears under a light microscope after chemical staining. The staining produces a pattern of bands in contrasting colors. Figure 1 shows a GnomeView map of chromosome 9. GnomeView computes band dimensions based upon data derived from the digitization of pictures in [6].

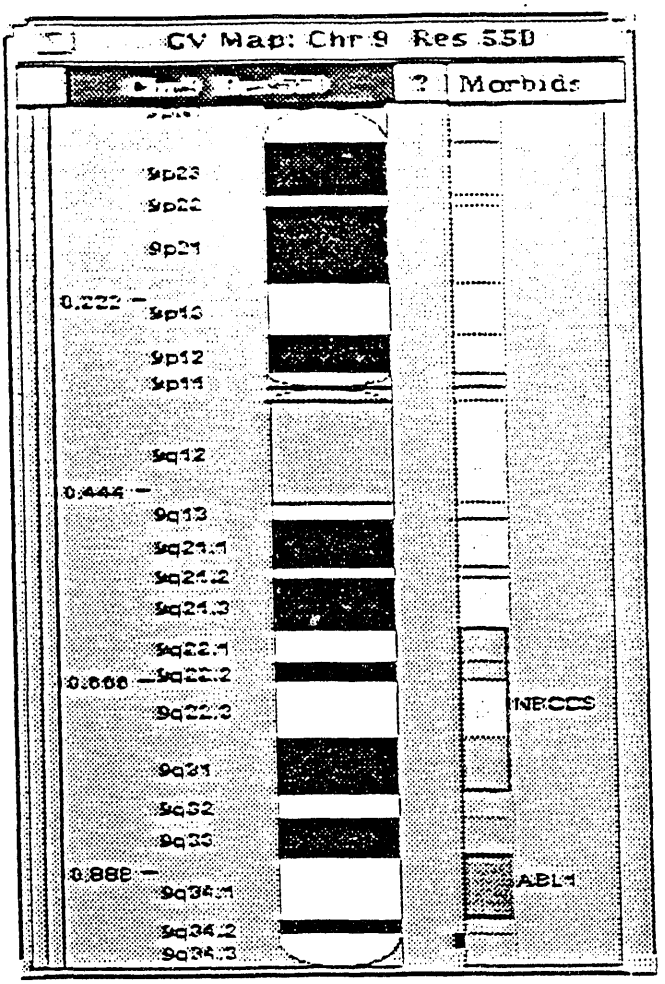

Figure 1: GnomeView Chromosome 9 Map

The metric at the chromosome level is chromosome bands. Each chromosome has a *centromere* separating the short (p) and long (q) arms of the chromosome. Bands are numbered outward in both directions from the centromere, as can be seen along the left side of the chromosome in Figure 1. Genes or other biological landmarks are located to the bands. For example, ABL1 (a locus related to leukemia) is located to a specific band, 9q34.1. The locus NBCCS (related to basal cell carcinoma syndrome) is located to a range of bands, 9q22.1–9q31.

DNA sequence maps provide a level of resolution which is orders of magnitude higher than that provided by chromosome maps. GnomeView represents the sequence of base pairs as a number line, drawing each base pair as a color-coded tick mark or letter as space permits.

The metric at the sequence level is base pair coordinates. Higher level objects, called features, are located to the sequence according to their base pair start and stop coordinates. In Figure 2, for example, the sequence HUMSODG1 is shown at three different magnification factors. At the highest magnification factor, a "repeat region" feature is seen to extend from base pair 196 (G) to base pair 208 (C).

Genome Databases

Another primary goal of the Human Genome Project is the development of capabilities for managing the data it produces. Computer databases will, of course, play a crucial role in any data management endeavor. There are presently a number of publicly accessible genome databases (e.g., [14, 8, 7, 3]), each of which tends to store a different type of data. Thus, GnomeView relies on one source (the Genome Data Base (GDB) [14]) for chromosome information and a different source (GenBank [8]) for sequence information.

A major problem with the current set of databases is lack of integration. There has been no convenient way to follow references across database boundaries; the researcher must make a manual transition from one database to the other and know how to use each of the

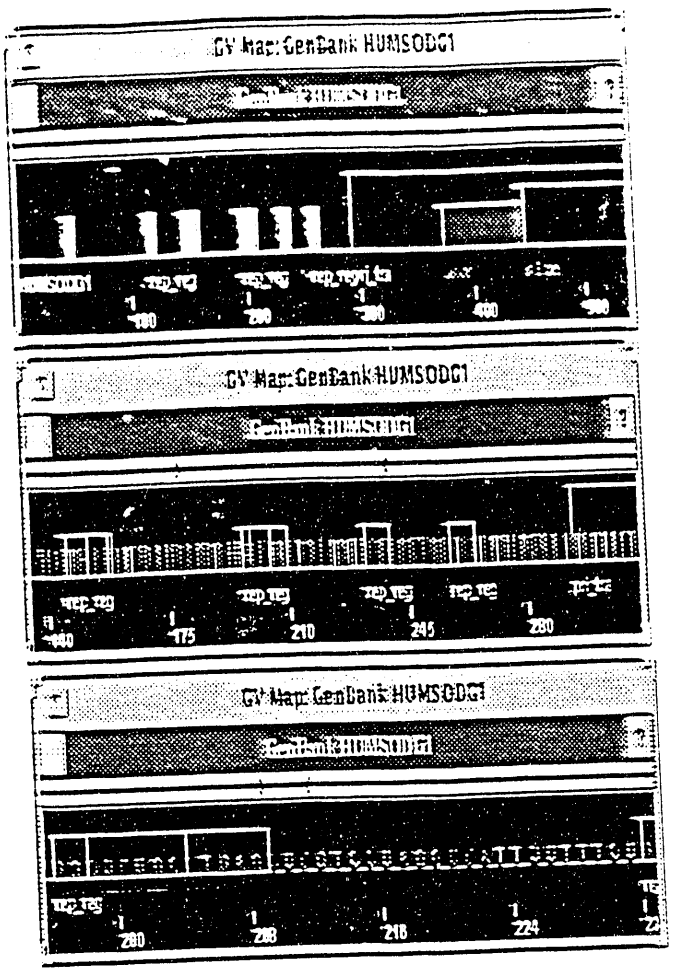

Figure 2: GnomeView Sequence HUMSODG1 Map

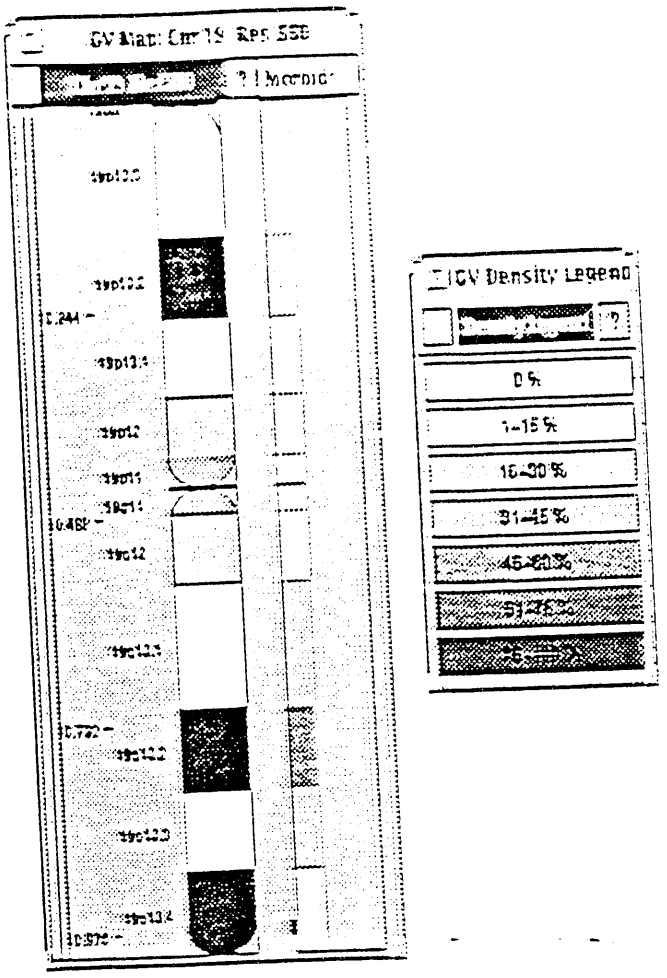

Figure 3: GnomeView Chromosome 19 Map

relevant databases. This makes it difficult to track an object through different levels of the mapping hierarchy, because most databases contain data from only one particular level.

A library that insulates the application developer from the mundane issues of database manipulation is described in [21], but this work does not address the issue of simultaneous access to multiple databases, nor does it relieve the developer from the burden of having to understand the individual database schemas.

GnomeView

GnomeView allows a user to access and graphically view data at different mapping levels. It strives to make navigation through mapping levels and hierarchies of data as easy and consistent as possible, obviating the need for the user to even know which underlying data sources are being accessed. The key capabilities, system architecture, and implementation of GnomeView are discussed in this section.

Capabilities

This section describes three of the salient capabilities of GnomeView: visualization, a graphical user interface, and integrated data presentation.

Visualization

GnomeView allows the user to select a graphical and/or textual style of data presentation. The heart of the graphical representation is the genomic map. As discussed earlier, GnomeView currently supports chromosome maps and DNA sequence maps.

In addition to these maps, GnomeView provides *density* maps—a type of color-coded histogram. A density map indicates the distribution of objects over an associated genomic map, lending an important sense of topology to the data. Figure 3 shows a map of chromosome 19 and an associated density map of all the morbid loci on chromosome 19. From this, one can readily see that morbid loci on chromosome 19 are concentrated on

band 19q13.2.

One of the most powerful features GnomeView provides is the ability to interactively magnify and scroll any map. As the user selects smaller regions of a map to view, more detail becomes visible. At any particular magnification, a map may be scrolled in either direction. A complete DNA sequence map, for example, may contain too many base pairs to display them individually. At this level, only the high level features of the sequence are displayed. Base pairs appear as the user "zooms in," initially as tick marks and eventually as letters. The tick marks and the letters are color-coded to improve the ease with which patterns (e.g., AT-rich regions) can be discerned. Figure 2 shows a sequence at three different magnification levels.

Graphical User Interface

The window-based query interface provided by GnomeView allows the user to specify (potentially complex) queries simply through the use of buttons, menus, and minimal text entry. We believe that a graphical user interface (GUI) is more intuitive than one based on traditional database forms packages or SQL (Standard Query Language) interfaces. GnomeView empowers users to explore the data and its interrelationships because it relieves them from having to learn query languages and database details.

A GnomeView window used to query for chromosome loci is shown in Figure 4. The current settings in this window will retrieve all loci related to Alzheimer's disease. The user merely pushes [Disorder], types `Alzheimer*`, and pushes [Search]. Performing this query without using GnomeView would require knowledge of both SQL and the schema of the underlying database (GDB).

For example, using the following relevant GDB schema definitions,

```
locus = (....
        locus_annot :  text,
        locus_cloned : tinyint,
        locus_id : int,
        locus_name : varchar(255),
        locus_symbol : char(12))

locus_mim_eref = (
        locus_id : int,
        locus_mim_annot : text,
        mim_ref : char(12))

mim_data = (
        mim_annot : text,
        mim_data_status_key : tinyint,
        mim_id : int,
        mim_ref : char(12))

mim_disorders = (
        mim_disorder_annot : text,
        mim_disorder_map_meth_key : tinyint,
        mim_disorder_name : varchar(255),
        mim_disorder_num : tinyint,
        mim_id : int)
```

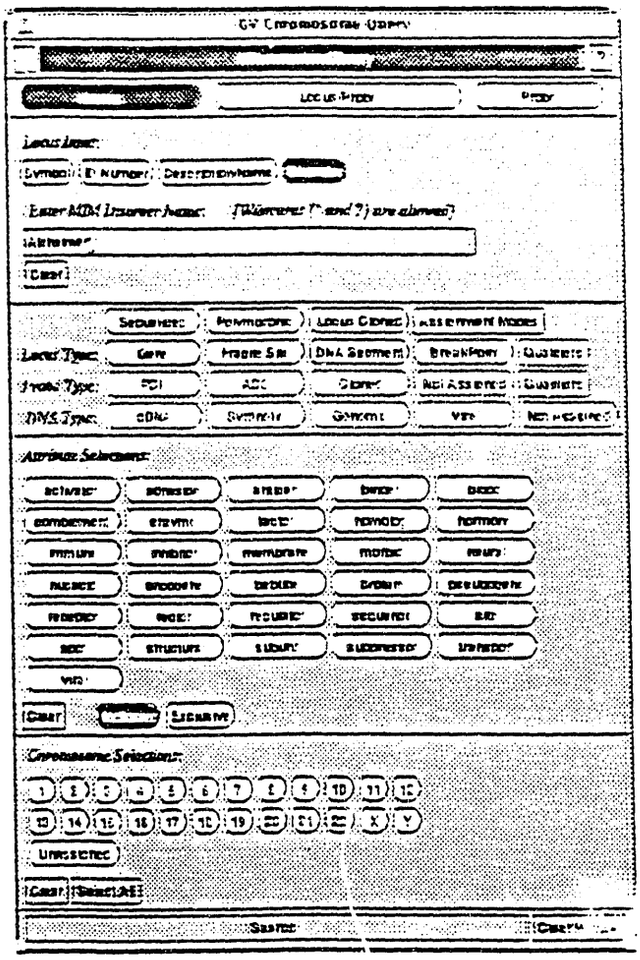

Figure 4: GnomeView Chromosome Query Window

an appropriate SQL statement based on this schema might look like:

```
select *
from locus
where
   mim_disorders.mim_disorder_name = "Alzheimer*"
   and mim_disorders.mim_id = mim_data.mim_id
   and mim_data.mim_ref = locus_mim_eref.mim_ref
   and locus_mim_eref.locus_id = locus.locus_id
```

The *locus* relation, with primary key *locus_id*, contains

the highest level of information about genomic loci in GDB. The *mim_data* and *mim_disorders* relations, with primary keys *mim_ref* and *mim_id*, respectively, contain detailed information about disorders contained in [11]. The *locus_mim_ref* relation is essentially used to join the locus and mim relations to find the specific disorder information associated with a particular locus.

As queries become more sophisticated, a GUI that shields the user from this level of detail becomes ever more imperative.

Integrated Data Presentation

As discussed earlier, there is a decided lack of integration among the current public genome databases. GnomeView addresses this problem by presenting to the user a facade of seamless database integration—the ability to transition smoothly across databases and between different levels of data. GnomeView can achieve this because among the data elements stored in its local database are the cross references to other databases provided by each of its constituent databases. Thus, given a locus in GDB, for example, GnomeView can easily retrieve all the associated sequences in GenBank, and vice versa, without requiring any manual intervention on the part of the user.

The degree of seamlessness in the database integration is influenced by the cross-referential integrity of the various databases. As researchers peruse data relationships more often, cross-referential integrity becomes more important. Even at this early stage, GnomeView's multiple database capability has already been useful in exposing inconsistencies in references across databases.

System Architecture

Figure 5 depicts the system architecture of GnomeView. ASCII flat files are downloaded from the public databases to a local disk via anonymous FTP (File Transfer Protocol) over the Internet. The local GnomeView database is then loaded from the flat files. The user interacts with the local GnomeView database via the GUI.

Implementation

GnomeView is written in C [9]. The user interface is based on the X Window System, Version 11 [17], the X Toolkit [1, 10], and the Athena widget set [15]. GnomeView employs a network model database, db_VISTA [16], as its internal database. The target

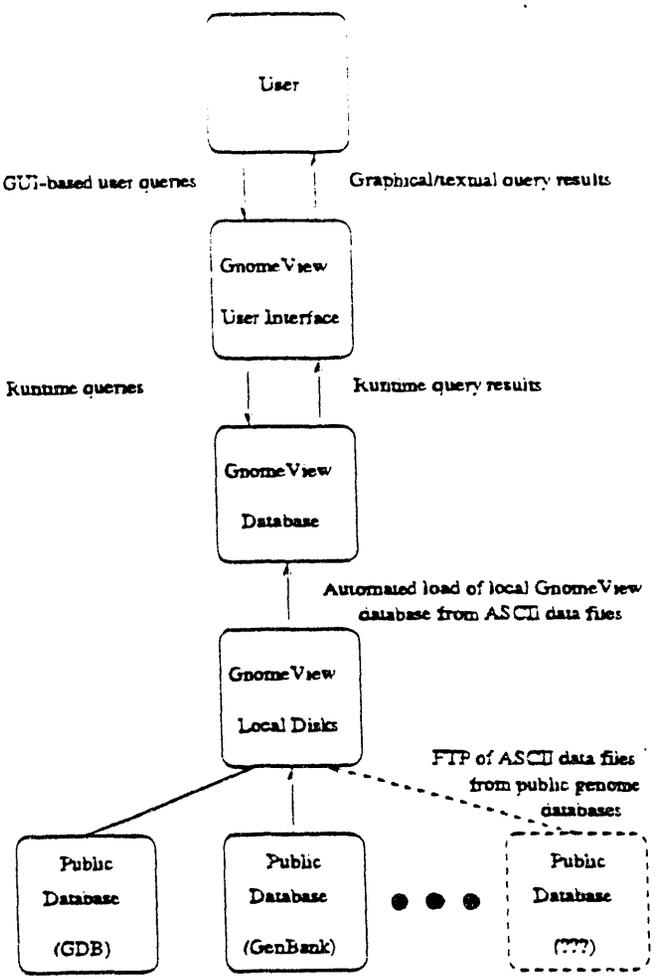

Figure 5: GnomeView System Architecture

platform is a workstation running UNIX and X, with an attached high resolution monitor. The development environment consists primarily of Sun SPARCstations running SunOS.

Current Status and Future Plans

GnomeView entered a beta test phase in October, 1992. It will be made available to the general community in 1993, after feedback from the beta test sites is received and incorporated into the system.

Future work on GnomeView will address several areas:

- **Expansion of Data Coverage**

  In addition to more data from the current sources being made available (GnomeView does not presently use all of the data available from GDB and GenBank), the number of data sources accessed by GnomeView will be increased. Some of the new sources will provide data for mapping levels other than the two GnomeView now supports

(chromosome and DNA sequence). It is likely that GnomeView's first new source will be contig mapping [12] data from the human genome center at Lawrence Livermore National Laboratory.

- **Direct Runtime Access to Public Databases**

  As explained earlier, GnomeView loads a local database after downloading data from the public databases. This gives maximum flexibility in how the data is represented locally, but has the disadvantage that all data used by GnomeView must be replicated and stored locally. It is estimated that only one tenth of one percent of the data that the Human Genome Project will eventually generate is currently available; while the current local storage requirement of 225 megabytes is acceptable, a thousand times this figure is not. Direct runtime access to the public databases over the Internet would solve the problem of lack of sufficient local storage capacity.

  Databases such as GDB already provide some services along these lines, but the interfaces are still schema-dependent. The development of high level, schema-independent application programmer interfaces to the public databases is necessary for direct runtime access to be feasible.

- **Analysis Capability and Map Creation**

  Eventually GnomeView will incorporate data analysis capability similar to that of various sequence analysis packages, e.g., CAGE/GEM [2]. GnomeView will also provide a mechanism for users to create and modify their own local maps, thus allowing them to dynamically investigate such things as chromosome cross-overs and mutations.

- **Implementation Enhancements**

  Migration from X11R4 to X11R5 is imminent. Longer range plans include porting to a different widget set (e.g., Motif), reimplementation in C++ [18], and the replacement of GnomeView's network model database with an object oriented database.

Acknowledgements

This work is supported by the U.S. Department of Energy under contract number DE-AC06-76RLO 1830.

GDB data used in preparing this paper was derived from version 4.2 of the GDB(TM) Human Genome Data Base on the general access computer at Johns Hopkins University in Baltimore, updated on September 22, 1992.

GenBank data used in preparing this paper was derived from GenBank Release 73 on genbank.bio.net, updated on September 22, 1992.

References

[1] P. J. Asente and R. R. Swick. *X Window System Toolkit*. Digital Press, 1990.

[2] R. J. Douthart, J. J. Thomas, S. D. Rosier, J. E. Schmaltz, and J. W. West. Cloning simulation in the CAGE environment. *Nucleic Acids Research*, 14(1):285–297, January 1986.

[3] European Molecular Biology Laboratory. *EMBL Nucleotide Sequence Data Library Release Notes*, January 1988. Release 14.

[4] K. A. Frenkel. The human genome project and informatics. *Communications of the ACM*, 34(11):41–51, November 1991.

[5] L. Gonick and M. Wheelis. *The Cartoon Guide to Genetics*. HarperPerennial, New York, NY, updated edition, 1991.

[6] D. G. Harnden and H. P. Klinger, editors. *ISCN 1985: An International System for Human Cytogenic Nomenclature (1985)*. Karger, Basel, Switzerland, 1985. Published in collaboration with Cytogenetics and Cell Genetics.

[7] Human Gene Mapping Library. *Regional Localization of Genes and DNA Segments on Human Chromosomes*. Howard Hughes Medical Institute, New Haven, CT, February 1988.

[8] IntelliGenetics, Inc. *GenBank(R) Release 73.0*, September 1992.

[9] B. W. Kernighan and D. M. Ritchie. *The C Programming Language*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[10] J. McCormack, P. J. Asente, and R. R. Swick. *X Toolkit Intrinsics—C Language Interface, X Window System, X Version 11, Release 4*. Digital Equipment Corporation.

[11] Victor A. McKusick, M.D. *Mendelian Inheritance in Man (MIM)*. The Johns Hopkins University Press, Baltimore, MD, 1992.

[12] National Institutes of Health and U.S. Department of Energy. Understanding our genetic inheritance, the U.S. human genome project: The first five years, FY 1991–1995. Report DOE/ER-0452P. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health; U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, April 1990.

[13] J. Ostell. GenInfo backbone database overview. Technical report, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, June 1990.

[14] P. L. Pearson. The Genome Data Base (GDB)—a human gene mapping repository. *Nucleic Acids Research*, 19 (Supplement):2237–2239, 1991.

[15] C. D. Peterson. *Athena Widget Set—C Language Interface. X Window System. X Version 11. Release 4*. MIT X Consortium.

[16] Raima Corporation, Bellevue, WA. *dl_VISTA User's Guide*, sixth edition, 1990.

[17] R. Scheifler and J. Gettys. *X Window System*. Digital Press, third edition, 1992.

[18] B. Stroustrup. *The C++ Programming Language*. Addison-Wesley, Reading, MA, 1986.

[19] U.S. Department of Energy. Human genome 1989–90 program report. Report DOE/ER-0446P, U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, March 1990.

[20] U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research. *DOE Human Genome Program, Report of the Second Contractor-Grantee Workshop*, Santa Fe, NM, February 1991. Available as report Conf-9102129.

[21] M. Wagner, T. M. Yeh, T. Slezak, E. Branscomb, and A. Carrano. *Multiple Database Interface Library*. Lawrence Livermore National Laboratory, October 1990.

END

DATE
FILMED
5 28 93