Cytogenet Cell Genet 58:1833-1838 (1991)

Report of the informatics committee

C.J. Rawlings, C. Brunn, S. Bryant, R.J. Robbins, and R.E. Lucier

Introduction

HGM 10.5 marked the introduction of the Genome Data Base (GDB) as the official system for maintaining and accessing mapping data in support of the international Human Genome Project. While the prototype system was successful in compiling and making available the HGM consensus map, critical work remained in four areas prior to HGM 11: (1) enhancing editorial modules based upon the expressed needs of the HGM 10.5 attendees; (2) extending the content domain covered by the database to include additional information, especially physical mapping data; (3) increasing the functionality of the General User Interface and related query tools used by the wider scientific community; and (4) improving the system's robustness, as part of the general maturation process in moving from an operational prototype to a full production system.

GDB †

At HGM11 GDB once again provided the data management tools to support the works of the Chromosome, Nomenclature and DNA committees. For GDB to be used at HGM11, GDB in Baltimore was "frozen" and a copy taken to London and installed on computers at the meeting. This copy was updated throughout the meeting, then taken back to Baltimore, where the updated version was installed and made available once more. By the Tuesday following HGM11, all of the data entered and approved during the meeting were available to the scientific community.

The main features of GDB, its implementation in Sybase using the APT screen design tools as well as the Sybase server client architecture were described in the HGM10.5 Informatics Committee Report (Rawlings and Lucier, 1990). Since HGM10.5 there have been 3 major upgrades to the software - the 2.0 version released in February, the 3.0 version release in April, and the 4.0 released in July. The number of upgrades in rapid succession reflects the evolution of the software, the need to tailor the modules to meet the users' requirements, and the need to create standard interfaces that are intuitive to users. The major enhancements for the software releases between the HGM10.5 and HGM11 workshop are outlined in Table 1.

Table 1. The major enhancements to GDB between HGM10.5 and HGM11.

Version	Date	Enhancements
1.0	Aug 1990	Used at HGM10.5 in Oxford
2.0	Feb 1991	- Various enhancements to the individual data managers - Additional output reports available - New locus retrieve screen
3.0	Apr 1991	- Map manager introduced with read access capabilities - Consistency in 'look and feel' among data managers added. - Revised source manager implemented.
4.0	Jul 1991	- Map manager editing available - New proposal system implemented - Inclusion of mouse homology and MIM disorder data - Messaging system revised - Breakpoints can be entered
4.1	Aug 1991	Used at HGM11 in London

The most significant enhancements to the version of GDB software used at the HGM11 workshop were the addition of the map manager and the revised proposal system for entering and updating locus information.

The *Map Manager*, as used at HGM11, was the initial prototype for a module to allow the entering of an array of map information derived from various methods including linkage maps, radiation hybrid maps, and contig maps.

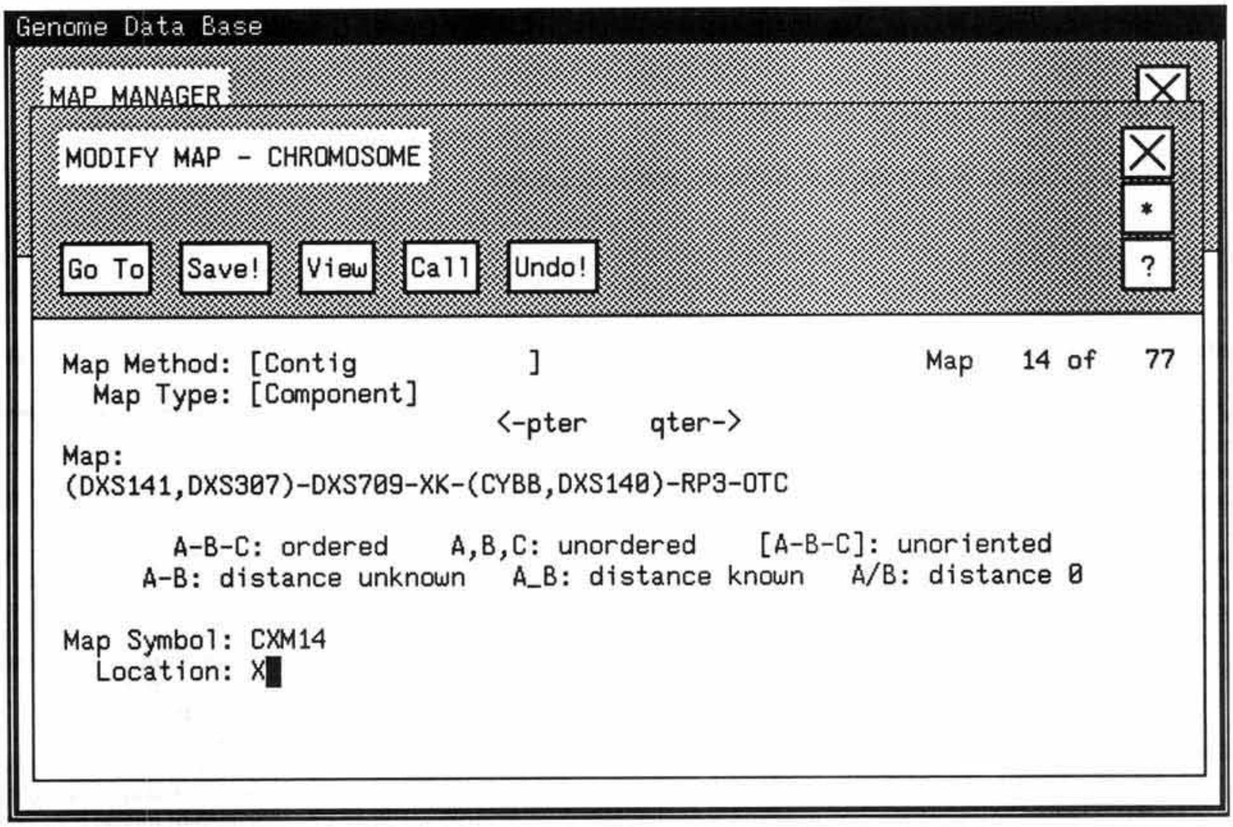
Data structures in GDB have been designed to capture the hierarchical nature of map information within the relational model and to accommodate the various forms these data may take with advances in this area of research. With these data structures, the strengths of a relational database can be exploited to allow the user to query the data in flexible and creative ways. The character-based interface to this information provides a simple means of entering data using a standard concise grammar which is flexible, but unambiguous in semantics. (Figure 1)

Graphical tools to display, analyze and compare this information will be developed in the future.

The proposal system model used to support scientific discussion and validation of the data was unchanged from HGM10.5, but the underlying software was modified to make it easier to add new data and to change dynamically which individual committee members are interested in these data elements. The screen displays were enhanced to facilitate the editorial process and make it easier for the user to compare validated and proposed data. (Figure 2).

Independently from GDB, the Nomenclature, Comparative mapping (mouse) and Genetic Linkage committees used their own

† GDB was originally developed as a Howard Hughes Medical Institute project in collaboration with the Johns Hopkins University and the William H Welch Medical Library. On 1 September 1991, the National Institutes of Health and the Department of Energy began providing support for GDB.



```
Map Method: [Contig          ]                          Map   14 of   77
  Map Type: [Component]
                              <-pter    qter->
Map:
(DXS141,DXS307)-DXS709-XK-(CYBB,DXS140)-RP3-OTC

       A-B-C: ordered    A,B,C: unordered    [A-B-C]: unoriented
     A-B: distance unknown   A_B: distance known   A/B: distance 0

Map Symbol: CXM14
  Location: X
```

Figure 1. Screen taken from the GDB Map Manager showing how maps are entered using a standard grammar.

individual database systems at HGM11 to manage data maintained separately by them.

The HGM11 computer network

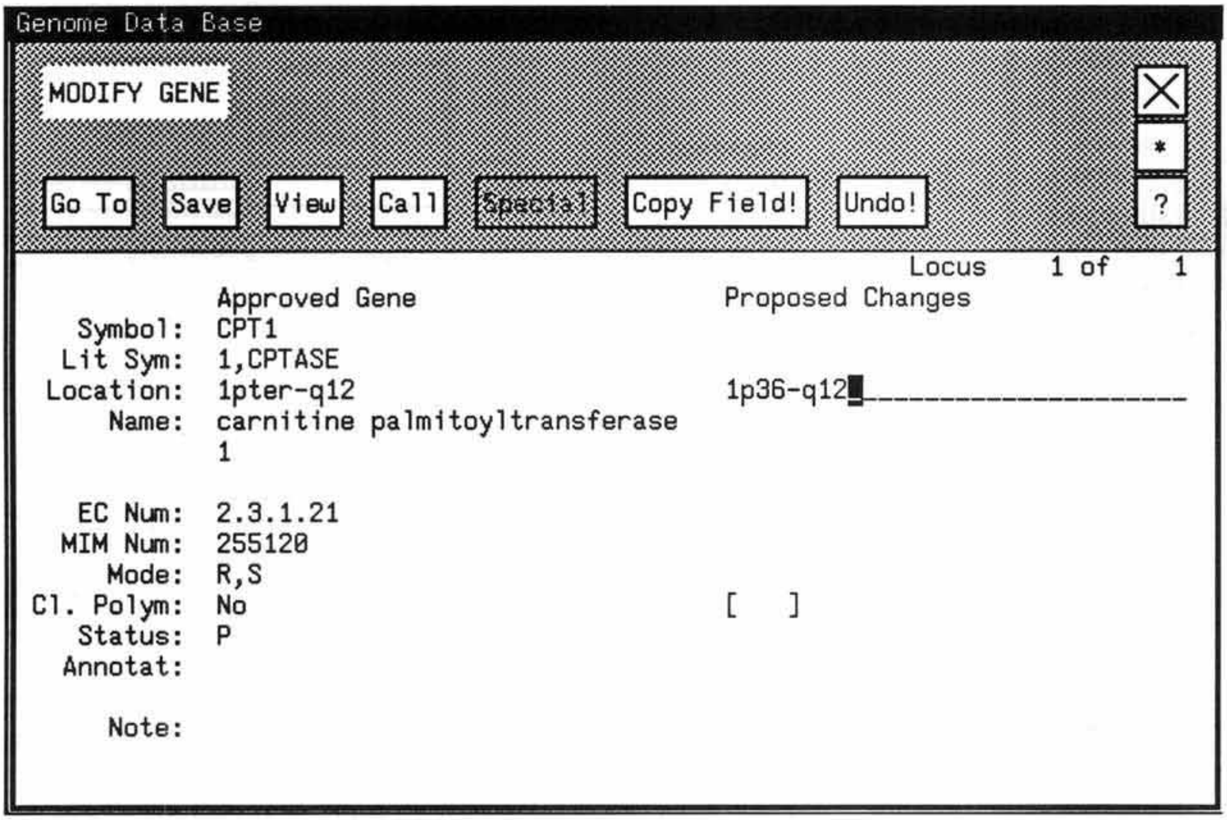
The HGM11 computer system comprised 57 SUN SparcStation workstations for use by the workshop committees, the GDB and HGM11 computing teams and the demonstrations of genetics databases. The committees were provided with diskless SparcStation SLCs configured with 16 Megabytes of memory. The network (Figure 3) was separated into 5 separate ethernet segments. in order to keep to minimum the amount of competing network traffic reaching the backbone segment. Apart from the one reserved for the HGM11 computer support group, the diskless SUN workstations were supported by two file servers per segment; one being a Sun SparcServer 4/330 and the other a SparcStation with external disk. Each segment was supported with at least 1 Gigabyte of disk - used mainly as virtual memory for the diskless SparcStation SLCs.

Each committee was also provided with either an IBM PS/2 model 30s or an Apple Macintosh personal computer. These were used for word-processing during the preparation of committee reports. They were also used with terminal emulation software to provide additional devices able to run the GDB user interfaces. The terminal emulation was achieved using the Kermit program providing the equivalent to Digital Equipment Corporation (DEC) VT320 video terminals (Kermit Version 3.0; Columbia University) with direct RS232 serial connections to the local SLC workstation. A total of 35 IBM PS/2s and 25 Apple Macintoshes were used. In order to simplify the computer installation, the personal computers were not connected to the ethernet. PCs and Macintoshes dedicated to printing were placed strategically in 10 locations around the HGM11 site. Laser printers for the Unix (SUN) network were connected to 8 of the Unix file server systems.

The main Sybase server for GDB was a SUN 4/490 SparcServer configured with 64Mbytes of memory and 2 Gbytes of disk. The server was located on the HGM11 site. In order to reduce the likelihood of a prolonged loss of service in the event of a hardware failure on the main server, a reserve server in the ICRF Central Computer Unit machine room was available at all times via a 2 megabit/second network connection. The reserve server was a Sun 4/280 with 64Mbytes of memory and 2 892Mbyte disks. The reserve database was kept synchronous by frequent incremental updates from the main one.

Wide area networking

The HGM11 network was linked to three wide area networks : the UK Joint Academic Network (JANET), the US Internet via the University of London Computer Centre and the private TCP/IP ICRF network ICNET. The most important networking during HGM11 and in the period immediately preceding the workshop was the Internet connection which provided a high bandwidth link through to the GDB group for transfer of data, the use of GDB by UK committee chairs and co-chairs and exchange of electronic mail.



```
                                                      Locus   1 of   1
             Approved Gene                    Proposed Changes
   Symbol:   CPT1
  Lit Sym:   1,CPTASE
 Location:   1pter-q12                        1p36-q12█_________________
     Name:   carnitine palmitoyltransferase
             1

   EC Num:   2.3.1.21
  MIM Num:   255120
     Mode:   R,S
Cl. Polym:   No                               [   ]
   Status:   P
  Annotat:

     Note:
```

Figure 2. The screen from the GDB Locus Manager used to propose changes to locus information.

GDB Online

GDB is a communications system, as well as a database with applications modules for maintaining and accessing data. In order to fulfill this function, GDB depends upon the national and international networking infrastructure. The original thinking of the technical and scientific designers was that GDB would transform the manner in which the community maintains the consensus map, from a mode of collecting and inputting data primarily at annual workshops to the continuous review and refinement of the data. The limitations of current networks and the tendency of groups to alter their behavior slowly have limited progress towards this goal. The Johns Hopkins University and the U.K. Medical Research Council collaborated on the development and implementation of strategies to make it possible for chromosome editors in the U.S., U.K., and Europe to edit and access GDB on an ongoing basis.

Continuing enhancements are being made to GDB to improve the ease with which users and editors may interact with the system. First among these will be the establishment of "official" remote GDB nodes at several locations around the world. To ensure that scientists do not have to compare different official nodes to see which is the best, or has the most up to date data, an official node agreement has been developed that defines the mutual guarantees that must be made to establish an official remote site. All of these guarantees are designed to ensure consistency across multiple official sites. For example, GDB will provide each remote site with an update of the database every week and the remote site must agree to install all updates promptly. Also, GDB will provide each remote site with copies of all user documentation and the remote sites must agree to distribute full documentation to all of its users. Other agreements of a technical nature (to ensure hardware and software compatibility) are also required. Based on discussions currently underway, we anticipate that several remote sites will be online in early 1992, including (but not limited to) Japan, Australia, Germany, and Sweden.

Distributed GDB software

To facilitate user access to the GDB system, GDB will distribute copies of the front-end software to any who request it. For technical reasons, the software is of use only to those with Sun workstations that are connected to the Internet. Because the GDB front-end software cannot run without licensed products from Sybase, the software can only be distributed to sites that have an appropriate Sybase license. Proof of an appropriate license is required before the software can be distributed. For the convenience of those who do not have a Sybase license, but who wish to obtain one, GDB is making arrangements so that it can act as a reseller and provide the minimum necessary license at a special rate. The exact price that must be charged for each such license is still under negotiation with Sybase, but it is expected to be available at a reasonable rate. GDB will not charge for the codes which it developed, but to offset actual expenses incurred a minimal annual service fee will be required for those wishing regular updates and maintenance. For further information, contact the GDB office in Baltimore.

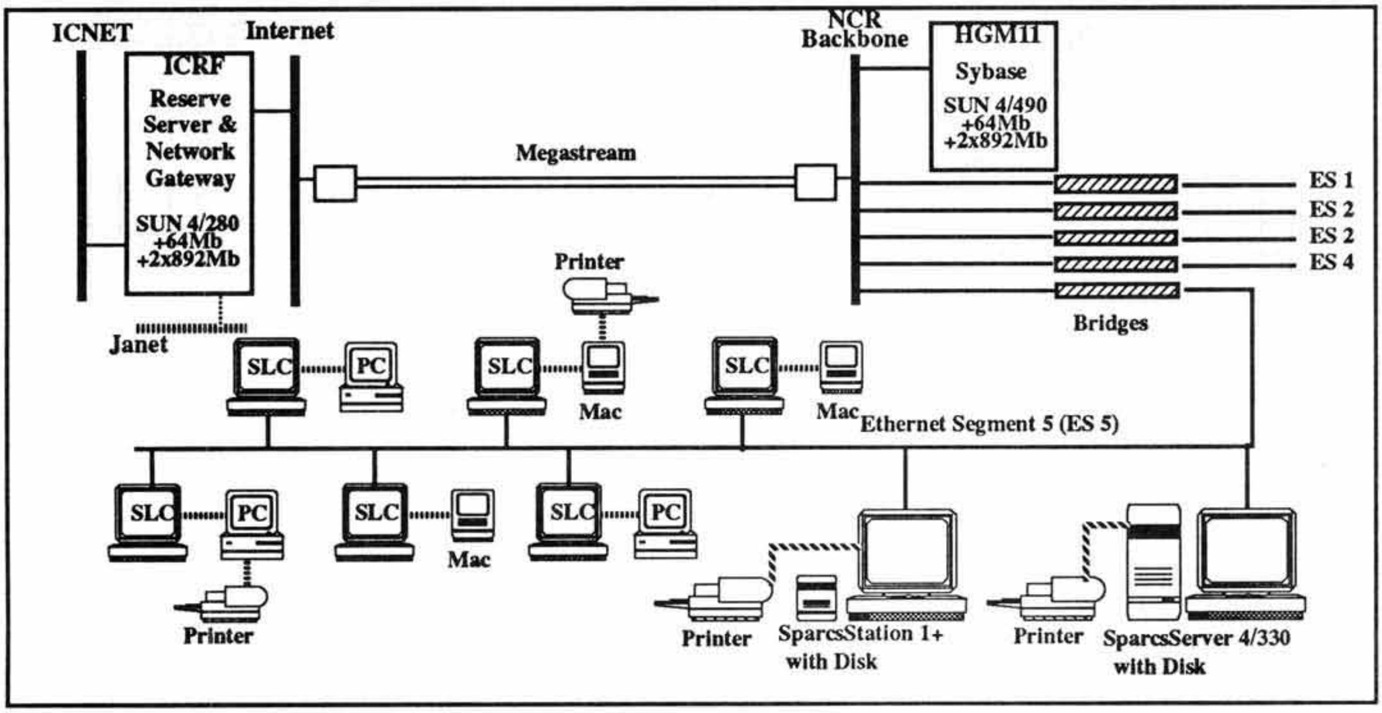


Figure 3. The HGM11 workshop network in the New Connaught Rooms. Solid lines between devices depict ethernet; discontinuous lines, serial RS232 connections.

GDB at single chromosome workshops

Until now, it has only been possible to update GDB data in one database at a time. Although the relocation of the database works well for large HGM meetings, it simply will not work for individual chromosome workshops, several of which may well be scheduled at the same time in different locations. Some individual workshops may be held in locations where high-quality network connections will allow the participants to use the main copy of GDB in Baltimore for recording their new entries. To accommodate those meetings that may occur at sites with inadequate network facilities, GDB staff are working to develop a system that would allow a copy of GDB to travel to the site of the workshop for updating, without requiring that the main database in Baltimore be frozen for the duration of the meeting. This capability is expected to be available with the release of the next version of GDB software.

GDB Access

For assistance with access to GDB contact: Product Services Genome Database, The Johns Hopkins University, 1830 E. Monument Street, 3rd Floor, Baltimore, MD 21205, USA; Phone: +1-410-955-7058, Fax: +1- 410-955-0054. Electronic Mail: help@welch.jhu.edu

In order to facilitate those who cannot easily reach Baltimore access is also available at the UK Medical Research Council, Human Gene Mapping Resource Centre (HGMP-RC). For assistance with access to GDB through the UK HGMP-RC contact either Christine Bates, Administrator, HGMP Resource Centre, Clinical Research Centre, Watford Road, Harrow, Middlesex, HA1 3UJ, UK; Phone: +44-81-869-3446, Fax: +44-81- 869-3807.

HGM Chairs and Co-Chairs outside the United States and the United Kingdom should contact either Baltimore or London depending upon their location and networking capabilities.

Processing of HGM11 abstracts

The abstracts from papers submitted to HGM11 were collected electronically using software (ABS11) written by the HGM11 computing group in the FoxBase database management system.

ABS11 was developed to gather the data necessary to transfer the abstracts into GDB so that they were available to Chairs and Co-chairs as GDB sources for online access prior to HGM11. The abstracts were also processed into the Maker Markup Language (MML) for preparation in the FrameMaker desktop publishing software ready for printing and distribution as the book of abstracts for HGM11.

ABS11 was developed for both IBM-PC and Apple Macintosh personal computer systems. It was distributed together with installation and de-installation software and documentation to approximately 580 scientists (380 IBM-PC and 200 Macintosh users). Disks with completed abstracts were returned from 342 groups (207 IBM-PC and 135 Macintosh) containing a total of 787 abstracts.

Idiogram Production

As is evident in the reports from the Chromosome and Neoplasia committees in this volume, the preparation of schematic diagrams of G-banded chromosomes (idiograms) labelled at the appropriate cytogenetic location with the symbols of mapped genes or the abbreviations for neoplastic diseases associated with chromosome abnormalities, is an important part of the preparation of HGMW committee reports.

In anticipation of requests for many idiograms to be prepared at HGM11 and for the data to be drawn directly from GDB, the HGM11 computing group developed a program (IDIO) which behaves as a separate client (written in the C language using the Sybase DB-library) to GDB and the FrameMaker Open Interface library of remote calls to a FrameMaker server to prepare publication-quality idiograms directly as a FrameMaker document on a workstation screen.

In addition to directly providing publication quality idiograms, this approach has the additional advantages of providing positional accuracy and labelling with up to date gene symbols through the direct coupling to GDB as well as allowing the sophisticated FrameMaker editing and publishing tools to be used to make detailed changes to the idiogram after the genetic mapping data has been drawn.

Genetic Database Exhibition

For the first time at an HGM workshop an opportunity was provided for developers of databases relevant to the Human Genome project to demonstrate their software and systems. The exhibition proved very popular and was well attended throughout the meeting. The following reports summarise each of the systems demonstrated.

GENOGRAPHICS

Ross Overbeek, Ray Hagstrom and Dave Zawada, Argonne National Laboratory USA

GENOGRAPHICS offers a graphical representation of the *E. coli* genome. It is based upon the ISO Abstract Syntax Notation (ASN.1) interface, developed in collaboration with the National Centre for Biotechnology Information (NCBI). It includes information on Kohara clones and can be interrogated using both a graphical and a natural language interface. The data includes sequence information and pattern matching data.

GENOGRAPHICS uses the SUN MicroSystems Network Extensible Windowing System (NeWS) to display graphical data. NeWS can run within the Sun OpenWindows standard window management system. As well as the *E. coli* system, there is also a version based on human data from Argonne. It is particularly good at the juxtaposition and and merger of maps based on different kinds of scale e.g. cytogenetic and restriction maps. Scales can be normalized for display purposes.

The linguistic interface, based on Prolog, grows by accretion as new questions are posed. The interface has been matured by considerable interaction with biologists. Questions are posed in a language approximating English and results are returned as text.

The front-end interfaces are available for distribution. The data sets may be available by arrangement.

Contact: Dr. Ross Overbeek, Mathematics and Computer
Science Division Argonne National Laboratory, 9700
South Cass Avenue Argonne, IL 60439-4844, USA
[overbeek@mcs.anl.gov]

LDDB and LND

Robin Winter, Kennedy Galton Centre UK

Lisa Fine and Dr. Michael Baraitser demonstrated two commercial databases from OUP. The London Dysmorphology Database (LDDB) and a Neurology Database (NDB). Originally developed by Dr. Robin Winter and Dr. Michael Baraitser, the LDDB is a system designed to aid the clinician in the diagnosis of non-chromosomal, multiple congenital anomaly syndromes. The data contained within LDDB are drawn from over 1000 journals that are regularly scanned.

LDDB and LND both ran under MS-DOS with a colour graphics display. They were not mouse-based and did not run under a window manager such as MS-Windows but were very easy and quick to use.

A menu system offers the user a choice of information categories that can be browsed. Searches can be refined and the results viewed in detail or printed. LDDB is of particular value in diagnosing rare syndromes.

The NDB is a similar diagnostic tool and reference source for clinicians and contains data on over 2000 syndromes of the nervous system.

Both databases are regularly updated and are available by commercial subscription.

Contact: Janet Caldwell, Electronic Publishing, Oxford
University Press, Walton Street, Oxford OX2 6DP, UK

RLDB

Hans Lehrach, ICRF UK

Guenther Zehetner demonstrated the ICRF Reference Library Data Base (RLDB) which was written in the Oracle Database Management System and is running on the ICRF VAX 8700 cluster.

RLDB was developed in order to administer both the distribution of clones and the collation of hybridisation results from collaborating laboratories. It contains information on the libraries, clones picked into microtiter plates and their transfer to large scale grid filters, as well as probe data and hybridisation signals. It can also generate summary statistical information about the usage of clones and the results obtained.

Richard Mott showed an early version of software for ordering clones by constructing contigs from hybridisation data, partly obtained from RLDB. The software implemented a minimum spanning tree algorithm to perform the ordering and ran under the UNIX operating system with output presented in Postscript

Contact: Dr Hans Lehrach, Genome Analysis Laboratory,
Imperial Cancer Research Fund, 44 Lincoln's Inn
Fields, London WC2A 3PX, UK
[h_lehrach@uk.ac.icrf]

EMG

Thomas Snell, The Jackson Laboratory USA

The Encyclopedia of the Mouse Genome (EMG) provides standardised access to a range of data sets from different research groups, all concerned with the mouse genome. It is designed to be comprehensive, with the emphasis on simplicity.

The primary access to the data is via genetic maps, and thence to the literature on nomenclature or mapping methods. Both linkage and cytogenetic maps can be displayed. Using the (computer) mouse to select a gene symbol causes the information pertaining to that symbol to be displayed. It is possible to zoom in and out of chromosomal regions and to search for particular symbols in the map.

EMG currently includes data from the Genomic Database of the Mouse (GBASE) the mouse Homology Database and Programs (HMDP) and the Mouse Cytogenetic Database (MCD) Future releases will include committee reports, biological resource lists and physical mapping data.

EMG requires a Sparcstation running SunOS 4.0.1 or later, and uses the Sunview windowing system. Future releases will be written for the X windowing system, and will therefore be available on a larger number of hardware platforms.

Contact: Dr Thomas Snell, The Jackson Laboratory, 600 Main
Street, Bar Harbor, ME 04609, USA
[tcs@jax.org]

ACEDB

Richard Durbin, MRC-LMB, Cambridge UK

A *C. elegans* Data Base (ACEDB) was demonstrated by Dr. Richard Durbin. ACEDB runs under X-windows (although SunView is supported as well) on the UNIX operating system. It has been constructed using advanced object modelling techniques and provides both view and edit facilities to laboratories involved in analyzing the genome of *C. elegans*. ACEDB emphasises the graphical presentation of data and both genetic and physical maps can be displayed with information about their components displayed by clicking with the mouse. Objects appear in their own windows and can be further explored. Context-sensitive help is available and there is also a powerful underlying query language that can be exploited by the experienced user.

Contact: Dr. Richard Durbin, MRC Laboratory of Molecular
Biology Hills Road, Cambridge UK
[rdo@uk.ac.cam.mrc-lmb]

CIS

Manfred Zorn, Lawrence Berkeley Laboratory, USA

The Chromosome Information System (CIS) is based on data obtained from GDB. It is a semi-private research tool that embodies both browsing and editing facilities.

It is written for the Apple Macintosh using Supercard (a Hypercard derivative) and the Sybase Hyper-DB library. Regions of chromosomes are selectable using a mouse and different types and resolutions of maps can be displayed. The current version supports cytogenetic, genetic, restriction and radiation hybrid maps. Map items can be bands, loci, fragments, sequences, STS or PCR primers. A hypertext-like system lets the user switch to entries about a particular locus or reference, and then examine probes defining a particular locus.

Contact: Dr. Manfred Zorn, Human Genome Computing,
Lawrence Berkeley Laboratory, 1 Cyclotron Road,
Berkeley, CA 94720 USA
[zorn@csr.lbl.gov]

ENTREZ:Sequences

David Lipman and Jim Ostell, NCBI, USA

ENTREZ:Sequences is a merger of GENBANK, PIR and about 70000 Medline references. About 50% of the citations contain sequence data that links directly to entries in the databanks. GENBANK sequences, translated into protein, similarly link to Medline UIDs. As well as these hard-coded links, emergent links within the databases are derived from the results of sequence comparisons that identify similar sequences.

ENTREZ provides an easy way into the data, being a hypertext-like navigation tool that runs on Apple Macintoshes and IBM PCs. After a short time using ENTREZ it is possible to swiftly move through the the information that represents the domain of enquiry.

ENTREZ is largely dependent on ASN.1, an ISO external data representation standard. NCBI have put substantial effort into developing a collection of software (ASNTOOLS) which serve to facilitate data exchange between electronic information sources. They have also produced a library of ANSI-compatible C routines (CORETOOLS), which are available to developers to encourage the spread of standard, maintainable software.

ENTREZ is updated every 8 weeks on to CD-ROM and NCBI should be contacted directly for information on distribution. The distribution requires either a Macintosh with at least 1MB of memory, and a hard disk or a PC-compatible computer with Microsoft Windows (version 3.0 or later), at least 2MB of RAM and a hard disk.

Contact: Dr. David Lipman, National Center for Biotechnology
Information, National Institutes of Health, National
Library of Medicine, Building 38A, Room 8N803,
Bethesda, MD 20894, USA
[lipman@ncbi.nlm.nih.gov]

BROWSER

Elbert Branscomb LLNL USA

BROWSER is an application using X-Windows (SUN OpenWindows) software that communicates with a Sybase server that can be physically located anywhere on a TCP/IP network. For the purposes of the demonstration, the server was located at the Lawrence Livermore Laboratory with access via the International Internet. The network connection was troublesome, but robust enough to permit successful demonstrations.

BROWSER is concerned with reconstructing cosmid contigs. Contigs are displayed and are queried using a point-and-click mechanism. Various kinds of information are kept about contigs, including likelihood ratios for overlapping contigs. Contig building is done largely automatically, and the contigs can be browsed for statistically weak areas, i.e., those with low likelihood ratios for overlap.

Contact: Dr. Elbert Branscomb, Biomedical Sciences Division,
Human Genome Center, Lawrence Livermore
National Laboratory, PO Box 5507 L-452, Livermore,
CA 94551 USA
[elbert@alu.llnl.gov]

JHGML

Japanese Human Gene Mapping Library, Keio University School of Medicine, Japan

The Japanese Human Gene Mapping Library system (JHGML) runs under X-Windows and models a similar domain of genome information to that stored in GDB. One of the most striking features of JHGML is the incorporation of Kanji and Katakana characters as an integral part of the database and browsing interface.

Users are given the ability to display stylised banded chromosomes from which a subregion can be selected using the mouse or directly from the keyboard. Buttons are available to display, in another window, information about genetic loci from the selected region. This can be the approved symbol or selected citations from the literature. Searches can be made using keywords and the results expanded and output to a file or printed.

Contact: Dr. Nobuyoshi Shimizu, Keio University School of
Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160,
JAPAN
[nshimizu@mt.cs.keio.ac.jp]

References

Rawlings CJ, Lucier RE. Report of the Informatics Committee. Cytogenet Cell Genet 55:779-782 (1990).