The genome data base (GDB)—a human gene mapping repository

P.L.Pearson
Department of Medicine, Welch Medical Library, Johns Hopkins Medical School, Baltimore, MD, USA

ABSTRACT

The types of gene mapping data and its organization in the Genome Data Base (GDB) recently established at Johns Hopkins Medical School are described. The database provides a continuous online environment for data perusal and editing and is used as the informatics core for running the human gene mapping workshops. Current development is primarily concentrated on extending the types of map object, means of defining map location, map storage and representation. Experimental data structures have been created that permit storage of any type of map information, physical or genetic.

INTRODUCTION

The need for a well organized informatics component to the Human Genome Initiative has been emphasized consistently in many reports and recommendations attempting to outline the general directions the Genome Initiative will take over the coming years (5,6). The current status of databases used for genome activities has recently been reviewed (1) and the distinction made between mapping, primary sequence and literature databases.

Consistent with the view that gene mapping will be one of the main activities during the first phase of the Genome Initiative (2,5), in June 1989 the Howard Hughes Medical Institute (HHMI) and Johns Hopkins Medical School initiated a new data base program under the direction of the author to serve the needs of the human gene mapping community. This followed a period which had witnessed the start of the Human Genome Initiative and an enormous explosion in the quantity of human gene mapping information being generated. One of the main functional requirements for the database was that the data structures needed to store, maintain and retrieve new types of gene mapping information could be added in a modular fashion without the necessity for redesigning the database every time data from a new mapping strategy was introduced. Further, the database had to support continuous on line access and data maintenance from remote sites (4). Other requirements were that the human disease information stored in the Online Mendelian Catalogue in Man (OMIM), maintained by Victor McKusick and colleagues at Johns Hopkins with support from HHMI should be integrated with the gene mapping information. OMIM is used as the basis for generating the hard copy volume of the McKusick Catalogue biannually and now in its 9th edition (3). The informatics and infrastructural support for the program was provided by the Laboratory for Applied Research in Academic Information of the Welch Medical Library under the leadership of Richard Lucier. The name of the database is the Genome Data Base or GDB for short.

The following 12 months were spent in designing, constructing and testing the database in time for its unveiling and first public use at the human gene mapping workshop held at St. Johns College, Oxford, in September 1990. Approximately 1 month prior to the Oxford meeting, data from five different sources was loaded into the database, the majority originating from the Human Gene Mapping Library (HGML) in Newhaven. This database had been sponsored by the HHMI for the previous five years up to and including the time of data transfer into GDB.

The following is a brief account of the type of information currently stored in GDB, of the ways in which data is entered, maintained, accessed and distributed, and new developments to be introduced during the coming year.

Types and storage of data

The database revolves around four important categories of data, namely map objects such as genes, DNA markers etc.; map location; genetic disease and locus description; and bibliographic references. Other types of information appended to these four major categories include DNA probe information, descriptions of polymorphisms and of the systems required to investigate polymorphic variation, a list of registered GDB users and contact persons to obtain probes, listings of mutations at individual loci giving rise genetic disease and known mouse homologies to mapped human genes. Currently, the main types of map object represented in the database include genes, DNA segments and fragile sites. Table 1 gives a breakdown of the information stored in GDB as of March 1, 1991. Some of the more interesting data include information on nearly four and a half thousand polymorphic systems involving two and a half thousand different loci. The number of probe submissions based on PCR primers and corresponding to the definition of site targeted sequences (1) is the modest total of just over 500. However, the submission rate of PCR probes is increasing rapidly and now represents approximately 50% of recent entries used for polymorphism studies. There are currently over four and a half thousand registered users representing an increase of three hundred percent since Sept 15, 1990 when GDB officially opened its doors to online access. The logon frequency is approximately two thousand per month.

The overall relationship of the various data types is depicted in a simplified fashion in figure 1. Other types of map object that will be included on the coming year include chromosome breakpoints, restriction sites, meiotic cross overs and partial

maps. This latter is necessary to permit maps to be included as map objects within higher order maps. At present map position is defined on the basis of cytogenetic band location. Other types of map information being built into the system include linkage maps, in situ hybridization information, radiation hybrid maps, contig maps etc. Data structures have been created that permit storage of map information irrespective of the type of map information. No distinction is made between physical or genetic map information in the ways in which the in formation is stored and presented. Essentially all map information can be defined as three global parameters applicable to all mapping procedures, namely order, distance, and where appropriate, likely hood estimates on the order and distance values. The mapping data are stored primarily as order information and the distance estimates app-ended to the order information. However, the distance information between any two map objects can always be retrieved if necessary.

For the main part GDB operates under a commercially available relational database management system (RDBMS) called Sybase. This permits the data elements to be retrieved and compared in an extremely flexible fashion. A main advantage of the relational system is that each data element is stored in only one location within the database despite the fact that it may be used many different times in different applications. This makes maintenance of data integrity much simpler when new data is added or existing data modified than with traditional hierarchical database management systems. However, the information on genetic disease loci and mendelian phenotypes assembled by Victor McKusick and colleagues is stored in another type of informational system called Irx, which is not strictly a database management system but a means of rapidly scanning text strings for particular words or groups of words. We have connected the Sybase and Irx systems so that users can pass transparently from the genetic disease to the map information or visa versa easily without first having to logoff from the one system and then logon to the other to access the information. Users have the choice of either accessing the map or the genetic disease information when they first come into GDB dependent upon their primary interests. There after, they can traverse freely between the two types of information with the proviso that when they exit one system to go into the other and then back again, their point of re-entry will be at the original exit point.

Data entry

The GDB is a public database and tries to provide a consensus view of the sum total of human gene mapping information. Most of the information is derived from the literature and assembled by an editorial staff directly supported by GDB. The major types of data entered at this level include gene names and symbols, DNA probes, DNA polymorphisms, genetic disease information and information on disease mutations. The consensus map information for each chromosome is maintained by the chair and co-chair person for chromosome committees nominated by the human gene mapping workshops. GDB provides an integrated environment for editors to enter new data and pass the information from one editor to another during an approval cycle. After a newly mapped marker has been fully approved the information then becomes accessible to the general public, prior to that it only being visible to the editors concerned. A typical example would be that a chair of particular chromosome notes the mapping of a new gene on their chromosome from the literature. They enter this information into the database and this initiates an approval cycle involving nomenclature to check that the gene has been appropriately named, or the DNA committee if clones and/or polymorphisms are involved in the mapping. A characteristic of the GDB is that it does not permit a entry to become approved unless a source reference is linked to the entry. This means that users can always trace the origin of information. The system permits personal communications and currently about twenty five percent of references are in the form of personal communications. We may expect this proportion to increase rapidly over the coming years as journals refuse to publish all map information and many researchers go over to direct submission as the means of data entry. In this respect, the increase in proportion of direct entries will directly parallel that already observed for the DNA sequence databases.

GDB provides an environment in which map information can be integrated and permits the development of consensus or derivative maps by permitting interactive integration of map information already present in the database. This function is still in its infancy and perhaps more so than any other aspect of the data base will undergo changes during the coming months as experience suggests better ways of integrating and presenting map information. Indeed there are no universally accepted standards for storing and representing map information and we must regard the present implementation in GDB as experimental. However it will serve to initiate a discussion within the gene mapping community of what map data representations and manipulations best serve the communities needs.

Links to other public databases

Accession numbers to DNA sequences stored in Genbank are available in GDB for those genes and/or probes known to be sequenced. In future we hope to make access to the sequence information more or less transparent to the user so that users can move from the map to the sequence information and visa versa. Links to other databases are currently being established for the probe information stored in the ATCC catalogue, the CEPH family data for linkage studies and the information stored in the Gbase, the mouse gene mapping database maintained at the Jackson Laboratory.

Data distribution and International Access

Besides permitting remote online access GDB also serves the needs of the gene mapping community by providing an interactive environment for running the human gene mapping conferences or chromosome specific workshops. For example, at the height of the last conference held in Oxford in September, 1990 there were over one hundred people logged on to the system simultaneously. The database was mounted on its own dedicated computer system at the workshop and all data changes were entered into the one memory bank via a local area network of Sun workstations and terminals. Following the conference the data tapes were transferred to the mother system in Baltimore and since then the Baltimore system has operated as the sole node for all editorial changes to the data set. However, in collaboration with the British Medical Research Council a read only version of GDB has been established in the UK. This provides easier and cheaper accessibility to the information for users located in the UK and Europe. The copy in the UK is updated once a week from the mother system in Baltimore. There are also well advanced plans to establish other read only copies of GDB at the German Cancer Research Center in Heidelberg under the auspices of the EEC and at the EMBO laboratories in Heidelberg.

Interest has also been expressed by the French, Japanese, Australians and Swedes to mount read only copies of GDB. We envisage that eventually there will be a net of GDB 'clones' being updated on a regular basis from the mother system and providing a ready online access to GDB users in their own countries.

Other means of data access under development include providing direct electronic entry from and to GDB with other remote databases. In this context it is worth noting that the Sybase system provides a ready means for one Sybase system to communicate and exchange information with another using the client−server model normally used for communicating between a single database and workstations connected to it. This opens up the possibility of using GDB as a window into the more specialized data residing in databases at centers concentrating on mapping particular parts of the human genome and storing only summary information in GDB.

There are also plans to distribute information as flat files, either on CD ROMs or magnetic disks for those who wish to mount the information on their own system. The advantages of this include compatibility with other local programs and hardware. The disadvantages include a lower level of timeliness of the data, the loss of flexibility in retrieving and linking data elements provided by the Sybase system and the need to provide their own software to manipulate the information.

User Information

Potential users can obtain information for accessing the database from :

The Genome Data Base,
William H. Welch Medical Library,
1830 E. Monument Street,
Baltimore, MD., 21205 , USA.,
TEL (301)-955-9705 (General Information)
TEL (301)-955-7058 (User Support)
FAX (301)-955-0054
E-MAIL help@welch.jhu.edu

or alternatively in the UK. from:

The Human Genome Program Resource Center (HGMP-RC)
Clinical Research Center,
Watford Rd.,
Harrow, Middlesex,
HA13UJ, UK.,
TEL (081)-869-3446
FAX (081)-869-3807

REFERENCES

1. Pearson, M. L, and Zoll, D., (1991) FASEB. J. 5, 35−39
2. McKusick, V., (1991) FASEB. J. 5, 12−20
3. McKusick, V., (1990) Mendelian Inheritance in man, 9th ed., Johns Hopkins Univ. Press., Baltimore
4. Pearson, P. L., Lucier, R., and Brunn, C. (1991) In Etiology of Human Disease at the DNA level, ed. Lindsten and Petterson, Raven Press. Ltd., 23−33
5. National Research Council. Committee on Mapping and Sequencing the Human Genome. (1988) National Academy Press. Washington, DC.
6. U.S. Department of Health and Human Services and U.S. Department of Energy. (1990) National Technical Information Service, U.S. Department of Commerce, DOE/ER-0452P
7. Rawlings, C. J., and Lucier, R. E., (1991) Report of the Informatics Committee, HGM 10.5, Cytogenet. Cell Genet. 55, 779−782

Table 1.

Marker Loci	
Total genes	2217
Mapped genes	1883
Mapped DNA segments	5369
Mapped fragile sites	113
Total mapped loci	7365
Disease loci and mendelian phenotypes (OMIM)	
Total	5248
Probes	
PCR	519
ASO	432
Clones	14032
Total probes	14983
Polymorphisms	
Polymorphic genes	521
Polymorphic DNA segments	2145
Total Polymorphisms	4435
References	
Journal articles	15467
Personal communications	5508
Theses	12
Books	1
Total references	20988
Users and contacts	
Users only	3315
Users/contacts	1432
Contacts only	1922
Total People	6669

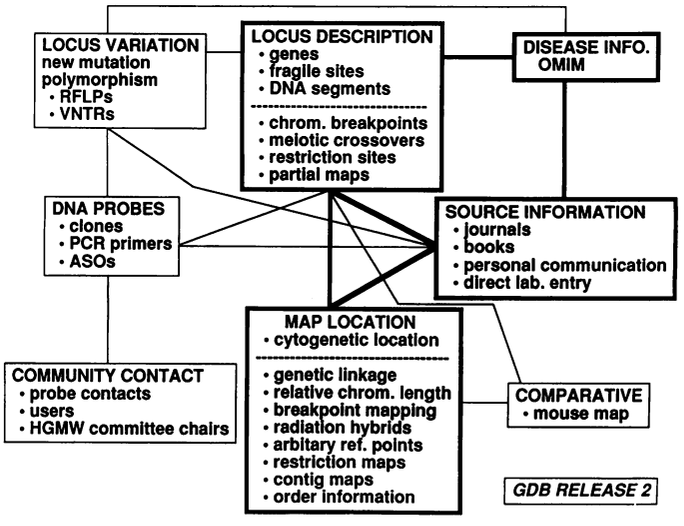

Fig 1. The overall data relationships are depicted in a simplified fashion. The data elements below the dotted line in the boxes for map location and locus description will be implemented by the time of Release 2 of GDB at the 11th Human Gene Mapping Workshop in London, Aug., 1991. The thicker lines indicate the most important data elements and their relationships. In practice the relational database management system permits many more data interconnections to be made than shown.