The GDB™ Human Genome Data Base anno 1994

Kenneth H.Fasman*, A.Jamie Cuticchia and David T.Kingsbury

Johns Hopkins University School of Medicine, Baltimore, MD 21205-2100, USA

ABSTRACT

In 1991 the Genome Data Base at Johns Hopkins University School of Medicine was selected as the central repository for mapping data from the Human Genome Project, and was funded by NIH and DOE under a three year award. GDB has now finished 28 months of Federally funded operation. During this period a great deal of progress and many internal changes have taken place. In addition, many changes have also occurred in the external environment, and GDB has adapted its strategies to play an appropriate role in those changes as well. Recognizing the central role of mapping information in the genome project, it is important that GDB respond aggressively to the increasing demands of genomic researchers, as well as formulate a program of response to a number of long standing, but still unmet, needs of that community. It is even more important that GDB provide leadership in the genome informatics enterprise. Three themes described here are dominant in our future plans and represent the essence of the major changes made in the past year. They include: enhanced data acquisition, better map representation, and full integration into the collection of genomic databases.

INTRODUCTION

The goal of the Human Genome Project is to identify, map, and sequence all human genes. The GDB Human Genome Data Base (1,2,3) is, first and foremost, the central data resource for the human gene mapping effort. GDB's mission is to serve as the central public repository for human gene mapping data. The database stores information on the locations of genes in the human genome, the reagents used to identify and locate them, and supporting data.

However, simple information storage does not meet all of the needs of the genomic research community. For GDB to effectively support the mapping of the human genome, it must be an active tool of scientific discovery, not a static archive of past results. GDB must serve as more than just a repository for finished maps. Genome researchers must be able to combine their own raw mapping data with other map information in GDB to derive new results about gene location and genomic relationships.

With this goal in mind, the current aims of the GDB project are:

- Reorganize the GDB schema to build a more modular database, including independent units for the core mapping

data and literature citations. The new schema will be designed to support direct user curation of submitted data.

- Improve integration of GDB with other genomic databases, especially those containing nucleotide and amino acid sequences, protein physical data, and human genotypes.
- Improve map representation in GDB, particularly for physical mapping data. Build a graphical map editor, and interfaces to third-party map assembly tools.
- Build enhanced applications for data browsing, data editing and direct data submission.
- Mount a more aggressive data acquisition program focusing initially on literature screening, but rapidly moving to direct data submission.

In this paper we present a summary of activities undertaken over the last year towards these aims, and briefly discuss our future efforts.

FEATURES OF INTERIM RELEASES

There were four releases of the Genome Data Base in the past year: versions 5.1, 5.2, 5.3 and 5.4. Although none of these releases constituted a major redesign of the GDB schema, each version put into place a feature that would be essential to the large-scale changes envisioned for GDB version 6.0 (discussed below). These releases provided features that significantly improved the utility of the database for the human genome community in the interim.

Version 5.1

The main feature of GDB version 5.1 was the inclusion of support for accession numbers in the database schema and application software. The need for unique, unchanging, artificial external identifiers for entries in public databases is well recognized. They provide a shorthand for rapid retrieval of individual items of data, without the need to formulate queries so specific as to guarantee the retrieval of an individual record. They are particularly effective in situations where the use of more natural identifiers would cause problems, usually because the latter are not unique (e.g., Social Security number versus first and last name).

Accession numbers are essential to the proper functioning of the Genome Data Base for a number of reasons. For example, although genomic loci are assigned unique symbols by the HUGO Nomenclature Committee, many other important data items in GDB (probes, polymorphisms, cell lines, citations, etc.) have identifiers that are not guaranteed to be unique and unchanging,

*To whom correspondence should be addressed

and do not even have a well-defined syntax. Rapid retrieval of individual entries in these categories can be difficult. Even locus symbols are not stable. Many loci have a number of symbol aliases and some symbols actually are used to refer to entire gene families. Also, authors need a simple means of referring to a body of data submitted to the database in their manuscripts. Readers of those articles need a way of retrieving all of the related data as simply as possible. This is especially true for major mapping studies, for which very large data submissions are made to the database. Submission of data to the nucleotide sequence databases was greatly enhanced when journal editors encouraged authors to include database accession numbers in their manuscripts instead of publishing sequences. These numbers act as a receipt — a proof of submission. A similar mechanism for GDB would also help to insure that information is submitted to the database in a timely fashion.

Genome Data Base accession numbers (GDBID's) were implemented with a number of characteristics in order to meet these goals. Accession numbers are available for all major data objects in GDB: loci, probes, maps, polymorphisms, mutations, citations, cell lines, clone libraries, and populations. The accession number design allows for the addition of new classes of objects. A given entry's accession number is unique and unchanging in the database — GDBID's of deleted or merged objects are never recycled. Accession numbers have no inherent information content, other than a display format that indicates that they are GDB identifiers (as opposed to GenBank accession numbers or postal codes). They would not be able to adapt to an evolving schema if they contained information such as the data type of the assigned object.

When two (or more) entries are merged, the old accession numbers of the individual entries are mapped to the accession number of the new merged entry. A user is always able to determine the current status of an entry associated with any previously assigned GDBID. A new accession number is assigned as the primary identifier for the merged entry. The original identifiers are returned as alternative pointers to the data, but are not normally displayed when the merged entry is retrieved. When an entry is split, the two new entries are each assigned a new GDBID, with the old number becoming an alias for each.

Version 5.2

The most important feature of version 5.2 of the Genome Data Base was the separation of the literature component of GDB into a separate database. This was the first step towards the creation of a single citation repository for all public genomic databases (please see 'Citation Database' below). In addition, GDB 5.2 provided support in the database schema and application software for the maintenance of a single consensus locus order for each chromosome by the HUGO committees. It also included an improved representation of the location of genes on the mitochondrial chromosome. Version 5.2's Output Manager is capable of generating subscription reports with incremental updates. That is, a user can define a query to be run periodically, and the server will e-mail only the data satisfying the request that have been added or modified in GDB since the last time the report was run.

Version 5.3

Version 5.3 of the Genome Data Base was designed to eliminate a long-standing problem with maintaining the GDB remote nodes around the world. Previously, the database was updated weekly

at these sites by shipping database dumps on tape to each node via air courier. These tapes contained a complete, updated copy of the database each week. The greatest advantage to replicating the database in this way is its simplicity. No software mechanism, commercial or custom, is required to transfer the data and keep the databases synchronized. However, there are a number of disadvantages to this approach, chief among them is the lack of timeliness of data at the remote sites, the number of manual operations required to reload the database, and the expense of preparing and shipping tapes.

Fortunately, an elegant, inexpensive solution to the problem was available. The Genome Sequence Data Base (GSDB) staff at Los Alamos National Laboratory had already developed software to maintain satellite servers for their Sybase database. The GSDB satellite replication system relies on the fact that these databases have relatively low update rates when compared to airline reservation systems and the like. GDB's transaction rates are modest, and split second accuracy of replicated data is not required at other sites. Therefore, the system tracks each database row that has been modified, rather than every single update to each row. The changed rows are then replicated to remote sites as a 'snapshot' upon request. Unlike Sybase's Replication Server, if the same row is modified 1000 times between replications to a particular site, only the latest instance of the row is actually propagated to the site when it requests an update. In this way, satellites can replicate the database as frequently as necessary, while incurring minimal load on the master server.

The satellite replication scheme greatly simplifies the distribution of data to GDB's remote nodes. It is also an essential part of the plan to develop the CitDB literature database and other components of the genomic database federation. The Genome Data Base staff are working with the Genome Sequence DataBase to extend the replication system to a true peer-to-peer model, creating a rudimentary distributed database system. This can be accomplished by partitioning the database so that each site is responsible for rows in a table within a particular range of identifiers. The most robust system for accomplishing this gives every row a two column identifier, for example a 'database_id' and a 'row_id.'

Version 5.4

Modifications to the Genome Data Base for version 5.4 focused on additional support for direct electronic data submission. Currently, most new information is interactively entered into GDB by the data management staff and HUGO editors using the standard GDB client software. Additional data are submitted directly to GDB via paper and electronic forms. A mechanism was needed to identify easily all the data that are part of a single submission. Methods are also needed to associate all the data entered interactively from a single set of paper forms or from one or more related manuscripts.

Data submissions have certain basic attributes, such as the submitter, submission date, method of submission, and so on. They are relatively static entities — the record of a submission usually will not be deleted or modified in any way. All of the accession numbers assigned to the various objects in a single data submission should be tied together in some way. This provides a number of important capabilities:

- A full data set can be retrieved with a single handle. In that way, any manuscript could refer to a single GDBID regardless of the size of its submitted data set. This is not the same as using the accession number assigned to the paper itself, since

a submitter may choose to associate one data set with multiple references (particularly if some of them are personal communications and others are journal articles).

- Submitters can be given responsibility for the curation of their own submissions. The submission record associated with each data object can be used to determine who has permission to modify or annotate that entry.

All electronic data submissions are assigned a GDBID for each data object in the submission. In addition, with the release of GDB 5.4, an accession number is assigned to the submission as a whole. This GDBID can be thought of as a 'submission number.' The database links all the accession numbers from a single submission to the submission number. In this way, a single number that refers to all related data items can be included in a publication. Users of the database are now able to retrieve individual records or all the records of a given type associated with a particular submission from any GDB data manager. They are also able to retrieve a list of the complete data set across all data types if so desired.

DATA DISTRIBUTION

GDB flat file format

Flat file databases do not allow one to perform very specific structured searches. They are more amenable to simple text searches than questions such as 'show me all cloned loci on the short arm of the X chromosome with heterozygosities greater than 0.8.' However, they are extremely easy for the novice to use for simple searching, particularly when combined with data distribution technologies such as CD-ROM, Gopher, WAIS, and the World Wide Web. Many of the public genomic databases (e.g., GenBank, EMBL Data Library) were first made available in a flat file format and later evolved into relational databases, but GDB began as a relational system from which a flat file had to be generated.

Flat files work best when the underlying data have a central data type (such as sequence for GenBank). However, there is no natural hierarchy of data in the Genome Data Base, and therefore no obvious way to organize the data into flat files. Locus, map, probe, polymorphism, and citation information forms a network of linked data in GDB, not a hierarchy. The only way to capture all of these interrelationships in flat file form is to create a number of files, each organized hierarchically from one of these data types. Thus there are locus, map, probe, polymorphism, and citation files in the GDB implementation.

The basic structure of a GDB flat file entry consists of the details of the 'base' object (locus, probe, etc.), followed by summary lists of the other objects in the database that are linked to it. Thus a locus record contains detailed information about the gene or DNA segment, followed by tables of associated polymorphisms, citations, maps, mutations, and probes. These are simply presented as one line summaries of each linked record. The addition of accession numbers to the Genome Data Base in version 5.1 was essential to the development of the flat file implementation, since the GDBID's provide a convenient way of accessing these linked entries.

World Wide Web server

The Genome Data Base has developed a World Wide Web server to provide access to the GDB and OMIM flat files. The World Wide Web is a system for managing and retrieving hyperlinked data across the Internet (4). The Web (also called WWW or W3)

was developed at CERN in 1990 for sharing manuscripts and other technical documentation among high energy physics laboratories around the world.

Client software was developed to display WWW documents and follow the embedded hypertext links from server to server. Mosaic is a multimedia WWW client that is also capable of accessing data using a variety of other Internet protocols, including FTP, Gopher, WAIS, and USENET news. It was developed at the National Center for Supercomputer Applications (NCSA) in Champaign, Illinois (5). It is capable of displaying embedded images in documents, and can invoke a wide variety of external viewer programs for the myriad of standard multimedia data formats that are currently available. Mosaic can display data entry forms dynamically as specified by a server document. Currently, when a new data type is added to GDB, the application software must be modified and distributed before users can query this information. With this new capability, we will be able to rapidly change forms-based query interfaces on our Web server and have those enhancements immediately available to anyone who accesses the database.

The World Wide Web allows GDB to combine the best aspects of relational and flat file databases using a 'two phase' approach. The data entry forms described above allow users to create very specific, structured queries. These queries are submitted to GDB's relational database server by our Web server as standard SQL queries. The relational database returns a series of GDB accession numbers that are then used to retrieve the appropriate entries from our WAIS flat file server. As these full text entries are returned to the Mosaic client, they are converted into HTML format on the fly. Not only does this allow the entries to be displayed in a very readable format, but it also allows the GDBID's in each entry to be marked as hypertext links. This enables the user to jump from a given data object to any of the objects that are linked to it in the database, including entries from OMIM. Furthermore, wherever GDB contains links to other databases (e.g., DNA sequence accession numbers), these links can be converted into active connections to other World Wide Web servers. The Web thus becomes a very simple way for biologists to obtain an integrated view of genomic data.

DATA ACQUISITION

Prior to September 1993, the data acquisition strategy of GDB could best be defined as passive. GDB has processed data submissions in a timely fashion (generally within 48 hours of receipt) and worked with research groups to gather their data for bulk submission to GDB. It was our expectation that researchers would contribute their data to GDB as part of the journal article submission process.

However, it has become clear that many laboratories do not place a high priority on the data submission process. Although most genome centers and large laboratories do have efficient methods for producing GDB submissions, most smaller groups have not yet felt compelled to devote the resources necessary for submission. GDB staff are not in a position to exert pressure on the community to contribute data. That is a task that belongs to the funding agencies and the community at large. Additionally, many members of the scientific community feel that journal editors should also insist that mapping data be submitted to GDB prior to the publication of an article, as is currently done with sequence data.

Direct data submissions come to GDB in two ways. Most small

research groups submit their data to the database using paper forms. These forms are available on the GDB FTP server as PostScript files, and are also mailed or faxed upon request. GDB has been in the process of simplifying these forms, which are currently too complex to be easily completed. In cases where the contributor has already collated the data in their own fixed format (such as a spreadsheet table), GDB staff will arrange to process the submission provided that the minimum amount of information has been included either within the table or in accompanying documents or files.

Currently, the majority of contributors prepare paper submissions. However, since February 1992 electronic submissions account for 78% of new loci and 77% of new probes. Presently, most U.S. genome centers and other large laboratories such as Génethon submit their data to GDB using the designated ASCII file format. Templates for these files are also available on our FTP server. These groups generate the appropriate files directly from their own laboratory databases, allowing for efficient transfer of the data. Additionally, Manfred Zorn of the Lawrence Berkeley Laboratory is collaborating with GDB on a program for preparing electronic submission files, much as individuals now use AuthorIn to prepare submissions for GenBank and GSDB. In fact, in its second iteration, this submission tool will allow for simultaneous submission to both GDB and the nucleotide sequence databases.

As large mapping groups have been pressured by the community to release data more quickly, there has been a corresponding growth in FTP servers as a means of providing this information. It is GDB's responsibility to work with these groups to develop a common format for those data that will both serve the members of their community and provide easy loading into the database. As the trend in data acquisition will be toward larger data sets loaded electronically, a consensus format for data exchange developed between the research community and GDB will promote efficient transfer of data.

Although increasing numbers of research groups contribute data directly to GDB, there are still cases where data are presented in the scientific literature without being relayed to the database. Beginning in September 1993, as part of our more active data acquisition strategy, GDB devoted additional resources toward scanning the literature for information that had not been previously submitted. Approximately 85% of all data added to GDB in the last two years have been linked to citations in one of 25 journals. Although the literature scanning procedure is still being refined, it is estimated that the data are available in GDB within 2–4 weeks after publication in the scanned journals. Medline searches are also performed to pick out promising articles from other journals.

Literature scanning is only an interim measure to make the database more useful to the scientific community. As the data continue to grow dramatically (Table 1 and Figure 1), it is not feasible to expect the resources available for literature scanning to grow at the same rate. Therefore, it is hoped that the community will continue its trend of contributing data to GDB prior to publication as GDB continues to develop tools to aid the submission process.

Scientific curation continues to be a vital component of the data acquisition strategy of the Genome Data Base. The Human Genome Organisation (HUGO) appoints a group of editors annually that are responsible for defining consensus views of GDB data, such as the cytogenetic locations of genes and consensus maps of each human chromosome. In light of the

Table 1. GDB statistics. Totals as of June 15 of each year.

	1991	1992	1993	1994
LOCUS:				
Total Genes	2326	3121	3942	4670
Total D-segments	6359	10224	15586	46932
Mapped Genes	2145	2665	3083	3308
Mapped D-segments	6359	8273	13633	32291
Mapped Fragile Sites	113	113	116	116
Mapped Breakpoints	0	69	257	537
Map Sets	0	243	601	894
Total Mapped Loci	8617	11363	17690	37146
Total Loci	8798	13770	20502	53160
PROBE:				
PCR	606	2016	7221	12961
ASO	424	440	450	464
Clones	15269	21156	27454	92301
Total Probes	16299	23612	35125	105726
POLYMORPHISMS:				
Polymorphic Genes	764	844	891	899
Polymorphic D-segments	2605	3271	4629	7420
Total Polymorphisms	4781	6175	7866	11313
CITATIONS:				
Journal Articles	16706	20511	26459	33994
Abstracts	0	0	194	888
Personal Communications	5577	5921	6362	6760
Books	12	25	40	42
Theses	1	1	2	3
Citations Linked to Data	13808	17621	20363	23009
Total Citations	22296	26458	33057	41687

Figure 1. Number of known human genes. Figures prior to 1991 reflect the number of human genes as reported at the Human Gene Mapping Workshops (HGM's). Figures from 1991 to the present were taken from the Genome Data Base as of June 15th of each year.

increased rate at which data are being generated, it is becoming ever more difficult for editors to perform their curatorial duties. For that reason, several countries (France, Germany, the United Kingdom, and the United States) have allocated funds for the support of editorial assistants to aid in the curation efforts. The assistants are provided with guidance and technical support from GDB (although they are not GDB employees). They work with the HUGO editors to insure that consensus views of the data are maintained in the database and to insure that the data in GDB reflect the current state of knowledge. Editorial assistants have also played an increasing role in assisting the scientific community in data submission, as well as entering data from the literature scanning process discussed earlier.

A FEDERATION OF BIOLOGICAL DATABASES

The next release of the Genome Data Base and its application software is being designed to be one component in an interacting collection of public database systems, rather than functioning as an independent entity with its own idiosyncratic characteristics.

The need for a federation of genomic databases was discussed at a meeting of bioinformatics experts, sponsored by the Department of Energy, that was held in Baltimore in April 1993. The conclusions of that group have implications for all of the community databases, especially GDB and the nucleotide sequence databases. The report from the April meeting states:

Successful HGP data management requires the development of a federated information infrastructure, with data flowing electronically over networks from producers to databases to users. Achieving this must be a top priority.

[...] Achieving coordination and interoperability among genome databases and other informatics systems must be of the highest priority. We must begin to think of the computational infrastructure of genome research (indeed, of biological research) as a federated information infrastructure of interlocking pieces. The distributed nature of biological research will require the development of multiple software and database projects. [...] Users must be able to retrieve related data from multiple databases such as GDB, PIR, Medline, PDB, and GenBank without having to make separate queries to the databases, then integrate the results themselves. (7)

GDB recognizes the desire for transparent linking of the sequence and mapping databases. We have been working actively to establish links within the current data set, and to find a suitable means for insuring future linking at the time of data submission. Because of the lack of a consistent nomenclature for the designation of human genes, and the clear identification of most submitted human sequences, we have proposed a mechanism for linking based on clones and not loci (8). We expect this procedure to be implemented in a future version of sequence and mapping data submission software.

The responsibilities of the individual databases participating in the federation were also laid out:

Community databases are shared resources, open to the entire research community. Although they have traditionally emphasized refined information over raw data, the trend now is for community databases to play an increasing role in making raw data (i.e., data that can be used for further analyses) available as a research resource. Community databases must be robust, since the majority of their users will be located off site. Formal user support, in the form of manuals, help lines, regularly staffed help desks, and training are required. Because these systems must meet the consensus needs of the entire community, care must be taken in their design to ensure that the databases are sufficiently flexible and robust to address the needs of different user communities. Since users need to integrate findings from several different community databases, **each community database should be designed as a component of a larger information infrastructure for computational biology** (emphasis added). Specifically, community databases should recognize the biological interdependence of information in multiple databases and should provide support for integrated queries involving multiple databases. To this end, they should be built from standard components and must be well documented. (7)

The Genome Data Base already possesses a number of these properties. However, the next release of GDB will incorporate even more of the aspects described above. In the following sections, we will describe how GDB's design will reflect the federation in microcosm, by creating a core database of basic mapping data along with a number of supporting databases.

PLANS FOR GDB VERSION 6.0

The Genome Data Base began as a reimplementations of Yale's Human Gene Mapping Library (9) using the relational data model. Fundamentally, the database schema has changed very

little since that time. GDB Version 6.0 is intended to be the most comprehensive overhaul of the Genome Data Base and its associated application software since its inception. The redesign of the database is focusing on a number of areas, including a revised model for the entry and maintenance of data, improved representation, display, and editing of genomic maps (particularly physical maps), and features that will better enable GDB to participate in the proposed genomic database federation.

Federation features of GDB 6

The next release of GDB will reflect the proposed genomic database federation in microcosm, by creating a core database of basic mapping data along with a number of supporting databases. Rather than using internal database identifiers such as artificial simple keys to link these components, this collection of databases will be linked by the same stable external identifiers (i.e., accession numbers as described previously) that will link databases across the whole federation. There are several important reasons for doing this. First, the management of identifiers in the revised GDB will serve as a model for what needs to be done across the whole federation. Second, it will allow any of the supporting databases to be separated from the GDB core and managed by another group, perhaps one more appropriate to the task. For example, the proposed citation database could be turned over to the National Library of Medicine.

GDB will be split into a mapping core, consisting of map objects, mapping reagents, and maps (see below), and a loosely coupled group of supporting databases. For example, literature references will be separated from GDB and combined with those of other genomic databases in the CitDB database. Information about GDB users and clone contacts will be placed in a separate, expanded genome people database. The general principle is to separate large, heterogeneous databases into a collection of smaller, more homogeneous ones, so that each collection becomes a 'mini-federation.' This basic organizational scheme can then be applied to the larger federation as well.

Whenever possible, the Genome Data Base must avoid storing data that are already being maintained in other scientific databases. Just as the principle of normalization is important in the design of an individual relational database, it is also essential in creating a well-organized federated database. Wherever practical, a particular class of information should be kept in a single, homogeneous database. This greatly reduces the need for synchronizing content across databases, just as a properly normalized relational database allows updates to occur in only one place. Rather than storing DNA sequence or genotype information in GDB, better links will be established to other databases such as GSDB, GenBank, EMBL, DDBJ, CEPH and CHLC. Only in cases where the existing resource databases are truly inadequate to the task will the data be stored also in GDB.

New editorial model

The Genome Data Base's original editorial model grew out of the Human Gene Mapping workshops. Newly acquired data were entered into the database with a provisional status. These data were then vetted by the HUGO editors, and only made available to the general public when they were fully approved by all relevant committees. There are a number of problems with this model. It does not scale well with increasing data volume, since all contributions must be vetted by the editors. Likewise, the mapping data in GDB does not consist solely of consensus information. Some data objects in the Genome Data Base are

simple experimental observations, not really subject to editorial approval. For example, when a submitter claims to have constructed a clone with certain properties, it requires actual experimentation to evaluate the claim. Other data, such as the assignment of a gene to a specific cytogenetic band, are consensus in nature, being built upon a number of supporting and conflicting observations. Many users of GDB may not be in a position to resolve the scientific debate for themselves. They will simply want the current consensus position as defined by the 'experts.'

The database must accommodate both editorially approved consensus data and experimental observations representing the results of a single experiment or experimental series. Consensus data are derived from the analysis of many user-supplied results. Perhaps, at some future point, there will be a universally accepted view of the structure of the human genome. Until that time, GDB will need to store the myriad glimpses of the underlying genomic organization that are being generated by individual mappers, as well as the current consensus. Representatives of the community will continue to play the same critical role in the maintenance of the consensus view of the data in GDB. That role will become more important when raw, potentially conflicting information is also in the database.

We expect to allow users to directly curate their data submissions as well as annotate submissions by others. Direct submission is the only solution to the data input problem that scales well as the amount of submitted information and the number of submitters grows. Direct curation makes the submitters primarily responsible for the annotation and management of their own data. Direct curation is important not only because of the explosive growth of the available data, but also because it encourages timely updates of the information by those best qualified to make them.

The negative aspect of direct data submission and curation is its potential for diminishing the quality of the database. For example, the nucleotide sequence databases are often criticized for the large number of poorly (or even erroneously) annotated sequences they contain. Direct user submission and curation must be accompanied by additional expert curation of the database, both by full time GDB professional staff and researchers who are selected by the genome community for their expertise in relevant subject areas. Regular review of submissions by database editors and curators will help to maintain data quality. Thus we end up with an editorial model where all data can be annotated by any user, and a consensus view of the genome is provided by representatives of the community.

Improved map representation

The Genome Data Base is currently capable of representing information about many of the basic genomic landmarks (genes, anonymous DNA segments, fragile sites, etc.), as well as cataloguing the numerous cloned and uncloned reagents used to locate these markers. A scheme also exists for representing basic map information, such as marker order and intermarker distance, as well as probabilities on these values. We are well aware of the limitations of the current database design in this regard. We are currently working to improve GDB's ability to capture today's map information, as well as that which will be collected in the coming years.

We propose to introduce the following improvements to the representation of map data in the Genome Data Base:

- Increase the variety of biological and experimental map

objects, and provide support for all of the common mapping techniques.

- Represent map objects as genomic extents, rather than points on a line.
- Develop a generic representation of map objects that is useful as input to integrated map construction tools.
- Include lower level mapping data in GDB so that information from the database can be used to construct new integrated maps, not just report finished results. However, GDB will not store all the raw data required for 'primary' map building (e.g., clone fingerprints, haplotype data).
- Introduce global and local coordinate systems for representing genomic location.
- Improve support for uncertainty in map object size, order, and location.
- Remove human karyotype assumptions and cytogenetic nomenclature wherever possible from the schema, as a step towards an organism-independent database.

A detailed discussion of the proposed representation is beyond the scope of this article. We refer the reader to the GDB World Wide Web server, where a detailed description of the proposed schema may be found.

Map editor

It is currently too difficult to enter mapping data into GDB because of its awkward character interface. As map data are inherently graphical, a graphical user interface for creating and entering small numbers of moderately complex maps is required. This tool should allow GDB users to enter new maps and retrieve, display, and modify existing ones. Many investigators are now using MacDraw or other personal computer vector drawing programs to create maps for publication and presentation at meetings. These programs are easy to obtain and run, have elegant user interfaces, and produce high quality output. However, they have two drawbacks: they are not specialized in any way for map drawing, and they do not provide any assistance in submitting the completed map to the Genome Data Base. In fact, the published diagram must be 'reverse engineered' by GDB staff when they are discovered in the literature scanning process.

We have recently begun to design a map editor for creating genetic and physical maps to address these shortcomings. It is our intention to create a program with an interface similar to MacDraw that does know about genomic maps. When a user completes a diagram with this program, it will generate two products: a high quality hard copy output for publication and a file that can be transmitted to GDB via e-mail, FTP, or mailing a floppy disk. This file will contain all the information defined in the map diagram using an electronic data submission protocol that is under development. It should be noted that groups with large numbers of maps, or extremely complex maps, will create these submission files directly from their own laboratory databases, bypassing the interactive map editor completely.

Citation Database (CitDB)

One of the stumbling blocks for implementing the integration of a federation of genomic databases is the management of literature information. Clearly, a lot of effort is being duplicated managing literature references for each database. GDB has created a literature citation database to serve public biological databases. The initial collaborating databases in the CitDB project are GDB, OMIM, the Genome Sequence DataBase, The Institute for Genomic Research's EGAD and SST (Chris Fields, personal

communication), TBASE (6), and the ProLink integrated protein database being developed at Boston University (Kathleen Klose, personal communication). We hope to add additional public databases (e.g., GenBank, PDB, PIR, SWISS-PROT, MGD, FlyBASE, etc.) to the collaboration over time.

At first glance, it may not be obvious why CitDB is needed at all. The National Library of Medicine already maintains the extensive Medline database of biomedical literature. Many of the databases participating in CitDB obtain the bulk of their reference information directly from Medline. However, this database is not sufficient to serve the needs of the Human Genome Project because Medline:

- is not on-line in an industry standard searchable form,
- does not track all journals,
- does not track conference and workshop proceedings and abstracts,
- does not track books,
- does not accommodate personal communications associated with direct submissions to biological databases,
- and does not provide the 'submitter' (citation author) with the ability to edit/amend/annotate their entry.

The CitDB genomic citation database was created to address all of the above problems. It is intended to provide a single convergent literature reference source for all genomic databases. We hope that it will eliminate the redundant development and maintenance efforts for scientific literature information at each of the participating databases.

The next version of the CitDB schema is currently being designed with the other collaborating databases. It will include several minor improvements, and will contain a few additions to the schema required for compatibility with the initial collaborators. Most importantly, it will incorporate a modified peer-to-peer database replication scheme that GDB is developing in cooperation with the Genome Sequence DataBase. This will allow each participating database to enter citations into their local copy of CitDB and have those records replicated to all other participating sites. Each site will only have direct editorial access to the citations that they submitted to the database. Proposed changes to other references will be sent on to the originating database. Sites that want access to CitDB without actually contributing entries can be easily accommodated.

CONTACTING THE GENOME DATA BASE

All questions about GDB, from obtaining a user account to inquiries about database content, should be directed to GDB User Services via one of the following methods:

Mail

GDB Human Genome Data Base
Johns Hopkins University School of Medicine
2024 East Monument Street, Suite 1-200
Baltimore MD 21205-2100, USA

Electronic mail

help@gdb.org
Phone: (410) 955-9705
Fax: (410) 614-0434

Data contribution

All inquiries regarding the contribution of data to the Genome Data Base can be sent in care of 'Data Acquisition and Curation'

at the above mailing address, telephone, or fax numbers. Preferably, inquiries should be sent via electronic mail to: data@gdb.org

World Wide Web server (WWW)

The Universal Resource Locator (URL) for the GDB/OMIM server is <http://gdbwww.gdb.org/>. GDB and the Welch Laboratory for Applied Bioinformatics also maintain a more general bioinformatics server at <http://www.gdb.org/Hopkins.html>.

FTP file server

GDB maintains a collection of data files, documentation, standard reports, software, etc. which are available via anonymous FTP from <ftp.gdb.org>.

Electronic mail server

GDB and OMIM data can be obtained from an electronic mail query system based on simple keyword searches. For more information on this service, send an e-mail message containing the single word 'help' in the body of the message (no subject line is necessary) to: mailserv@gdb.org.

Wide Area Information Server (WAIS)

The Genome Data Base maintains a number of WAIS information sources, one for each major category of data, as organized in flat file form. They are:

gdb-cell-line.src
gdb-citation.src
gdb-contact.src
gdb-library.src
gdb-locus.src
gdb-map.src
gdb-mutation.src
gdb-polym.src
gdb-probe.src
omim.src

Information on accessing these WAIS sources can be obtained from the WAIS Directory of Servers at think.com.

Gopher

The Welch Laboratory for Applied Bioinformatics and GDB maintain a computational biology Gopher server at the Internet address gopher.gdb.org.

GDB SITES

The Genome Data Base is provided at the following international sites to enable easy worldwide access to the database:

- ANGIS, University of Sydney, Sydney, Australia
- CAOS/CAMM Center, University of Nijmegen, Nijmegen, Netherlands
- DKFZ (German Cancer Research Center), Heidelberg, Germany
- CNRS-INSERM, Villejuif, France
- JICST, Tsukuba Science City, Japan
- UK Human Gene Mapping Project Resource Center, Hinxton, United Kingdom
- Uppsala Biomedical Center, Uppsala, Sweden
- Weizmann Institute of Science, Rehovot, Israel

ACKNOWLEDGEMENTS

The GDB Human Genome Data Base is funded by a grant from the U.S. Department of Energy (DE-FC02-91ER61230) and the U.S. National Institutes of Health (NIH HG00586-01). GDB receives additional support from the Science and Technology Agency of Japan, the Medical Research Council of the United Kingdom, and the INSERM of France.

REFERENCES

1. Pearson, P.L. (1991) *Nucleic Acids Res.*, 19-Supplement, 2237–9.
2. Pearson, P.L., Matheson, N.W., Flescher D.C., and Robbins R.J. (1992) *Nucleic Acids Res.*, 20-Supplement, 2201–2206.
3. Cuticchia, A.J., Fasman, K.H., Kingsbury, D.T., Robbins, R.J., and Pearson, P.L. (1993) *Nucleic Acids Res.*, 21–13, 3003–6.
4. Berners-Lee, T.J., Cailliau, R., Groff, J-F., Pollermann, B. (1992) World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*. 2:52–58. Meckler Publishing, Westport, CT.
5. Andreessen, M. (1993) NCSA Mosaic technical summary, revision 2.1. National Center for Supercomputing Applications, Champaign, Illinois.
6. Woychik, R.P., Wassom, J.S., Jacobson, D.A., and Kingsbury, D.T. (1993) TBASE: a computerized database for transgenic animals and targeted mutations. *Nature* 363: 375–376.
7. Branscomb, E., Cartwright, P., Cinkosky, M.J., Davison, D., Fasman, K.H., Fields, C., Gilna, P., Kingsbury, D.T., Marr, T., Slezak, T., Smith, F.R., Spengler, S., Uberbacher, E., and Waterman, M. (1993) DOE informatics summit meeting report. 26–27 April 1993. Baltimore, Maryland.
8. Fasman, K.H. (1994) Restructuring the Genome Data Base: a model for a federation of biological databases. *J. Computational Biology*. In press.
9. Stephens, J.C., Cohen, I.H., and Kidd, K.K. (1990) The Human Gene Mapping Library: present status and future directions. In: *Computers and DNA, Santa Fe Institute Studies in the Sciences of Complexity*. Vol. VII. G. Bell and T. Marr, eds. Addison-Wesley, Redwood City.