Nucleic Acids Research, 1996, Vol. 24, No. 1 **13–16**

The Genome Sequence DataBase (GSDB): meeting the challenge of genomic sequencing

Gifford Keen*, Jillian Burton, David Crowley, Emily Dickinson, Ada Espinosa-Lujan, Ed Franks, Carol Harger, Mo Manning, Shelley March, Mia McLeod, John O'Neill, Alicia Power+, Maria Pumilia, Rhonda Reinert, David Rider, John Rohrlich, Jolene Schwertfeger, Linda Smyth, Nina Thayer, Charles Troup and Chris Fields

National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505, USA

Received October 2, 1995; Accepted October 4, 1995

ABSTRACT

The genome sequence database (GSDB) is a complete, publicly available relational database of DNA sequences and annotation maintained by the National Center for Genome Resources (NCGR) under a Cooperative Agreement with the US Department of Energy (DOE). GSDB provides direct, client-server access to the database for data contributions, community annotation and SQL queries. The GSDB Annotator, a multi-platform graphic user interface, is freely available. Automatically updated relational replicates of GSDB are also freely available.

INTRODUCTION

The initiation of genome sequencing projects and the development of automated DNA sequencers have changed the style and pace of DNA sequencing. A laboratory with a single automated DNA sequencer can generate upwards of 50 000 bases of raw sequence data per day in steady state, and this rate is expected to increase by an order of magnitude within the next few years. Two complete microbial genomes have now been completed (1–2), several more are well underway, and dozens are expected to be finished by the end of the decade. The complete genome of *Saccharomyces cerevisiae* (3–5) is expected to be finished in 1996, and that of *Caenorhabditis elegans* (6–7) is expected to be complete by 1998. Over 80 million base pairs of human expressed sequence tags (8) have been sequenced and analyzed (9). The era of efficient, inexpensive, high throughput DNA sequencing has started. The genome sequence database (GSDB) is designed to meet the community wide challenges of managing, interpreting, and using DNA sequence data at an ever increasing rate.

High throughput, genome scale sequencing presents both technical and social challenges. Sequence production laboratories employ a variety of strategies, from high redundancy shotgunning producing highly accurate, completed sequences to directed strategies such as Ordered Shotgun Sequencing (10), sequence mapped gaps (11) or transposon based walking (12) to sampling strategies that produce ordered sequence fragments separated by gaps of known length (13). A laboratory may choose to release data to the public at any stage in the production process with any of these strategies; public sequence databases must, therefore, be prepared to manage sequence data in any form from complete, multi megabase contigs to collections of sub kilobase fragments related by order, orientation, and distance information. Moreover, multiple laboratories will often be involved in the completion of a sequenced region: one laboratory may generate mapped sequence samples, while many others produce highly accurate sequences of subregions or of corresponding cDNAs from expressed genes. As a region is further characterized, additional data showing diversity across individuals, alternative expression products, or even gross rearrangements may be added to the database. Finally, functional annotation of the sequence data may be the work of tens to hundreds of individual investigators. Capturing the latter data in a structured database is critical; too often, important functional results are presented only in the printed literature, and are therefore effectively lost.

GSDB is designed to meet the requirements for a community sequence database outlined by Waterman *et al.* (14). GSDB supports complex, *ad hoc* queries in a standard language, SQL (15). GSDB represents sequence data produced by any strategy, and supports the contribution of additional sequence data or structural or functional annotation by multiple researchers. GSDB extends the Electronic Data Publishing paradigm (16) from a model in which the database is viewed as a primary publication for data not appearing in the traditional literature to a model in which the database serves as a multi user laboratory database for the entire molecular biology community (17). Multiple sequences from a given region and structural and functional annotations on sequences are viewed, in this model, as independent observations, with authors and unique identifiers. GSDB provides multi user editing capabilities, with the necessary authorship, data security, integrity checking and versioning mechanisms needed to ensure that multiple authors do not overwrite each other's work. A mechanism is also provided for individuals or groups to define their own curated views of the data, which include whatever sequences and features they select.

* To whom correspondence should be addressed

+Present address: Huntsman Cancer Institute, Salt Lake City, UT, USA

Database users may choose to access the entire database, a particular view maintained by an editorial group that imposes particular standards, or a view based on data relevant to a particular set of interests. GSDB functions, therefore, both as a community laboratory database and as a collection of multiple, virtual, specialty databases.

GSDB 2.2: NEW DATA TYPES

GSDB version 2.2 replaces the notion of an 'entry' employed by the archival sequence databases with separate structures representing sequences and annotation. Both sequences and annotations are assigned their own individual, unique accession numbers that serve as permanent identifiers. This architectural change reflects the importance of structural and functional annotation as biological data, not merely notes about sequences. The addition of individual identifiers to annotations simplifies retrieval. It also allows multiple users to contribute annotation to a given sequence without updating an entire entry record with each change.

Aligned sets of sequences are represented as a fundamental data type in GSDB 2.2. The alignment structure allows the representation of multiple fragments from one clonal source (e.g., a single cosmid or bacterial artificial chromosome), multiple isolates of a single region from different sources (e.g., multiple isolates of a virus or disease gene), and multiple cDNAs from a genomic region (e.g., alternatively spliced mRNAs). Complete sequences are maintained in the alignment, with both coordinate based and sequence difference representations available for queries. The collection of covering sequences from each aligned set provides, moreover, a truly non-redundant sample of the sequences in the database.

Discontiguous sequences obtained by genome sampling strategies, EST projects or exon sequencing are maintained as sets of ordered, oriented fragments linked by distance and uncertainty values. This structure can also be used to represent sequence tagged site (STS) maps as discontiguous sequences spanning entire chromosomes. Discontiguous sequences can be included in alignments; hence regional or chromosomal sequences can be assembled by aligning multiple, independently derived sequences to a discontiguous sequence scaffold.

GSDB 2.2 represents sequence confidence with a span oriented real confidence value. This confidence value is qualified by flags indicating whether a sequence is single or multiple pass and whether one or both strands have been sequenced. Confidence values can, in the limit, be assigned to single bases.

The collection structure in GSDB 2.2 allows any set of sequences and annotations to be grouped into a queryable view. Any individual or group can create collections for their own or public use. Collections provide a mechanism for individuals or groups to curate or referee both sequences and annotations; they also provide a mechanism for assembling various non redundant or special interest sets of data in structured form.

GSDB 2.2: STRUCTURE AND ACCESS TOOLS

Architecture

GSDB employs a three tiered open architecture, as shown in Figure 1. The database is implemented with SYBASE, a commercial relational database management system (RDBMS) (Sybase Inc., Emeryville, CA). An object-oriented layer of

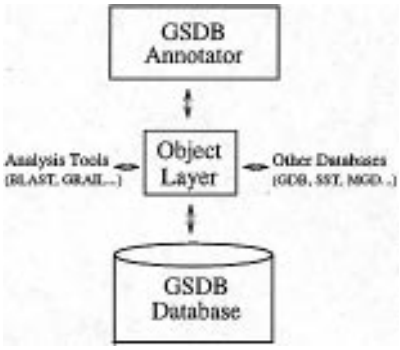

Figure 1. GSDB's open architecture supports links to analysis tools and other databases.

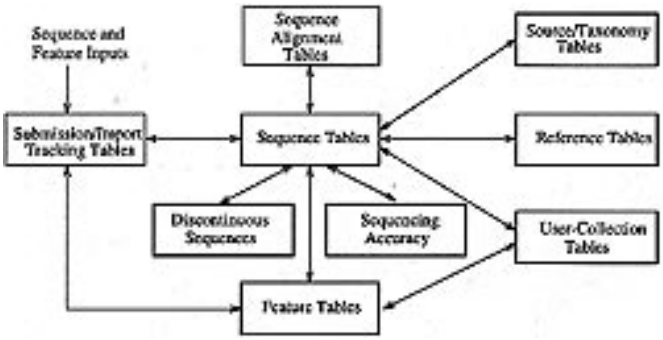

Figure 2. A simplified entity relationship-diagram of the GSDB schema structure.

software written in C++ sits on top of the database. This layer handles all database access and communication for a variety of applications. The GSDB Annotator, a multi platform graphic user interface, is the major system application provided as part of GSDB 2.2.

Users and developers may gain access to GSDB and its applications either through the object layer or by direct database access using SQL. Data analysis applications such as Blast (18) or Grail (19) are integrated into the GSDB Annotator at the object level. A public applications programming interface (API) to support the integration of analysis tools into the object layer is under development.

Schema structure

We have implemented a new schema structure to respond to the changing needs of the genome community. Figure 2 provides a high level view of database element interrelationships, incorporating the new data types discussed. A complete description of the schema is available by anonymous ftp from ftp.ncgr.org.

Database inputs and outputs

GSDB accepts input data interactively via the GSDB Annotator, via direct database–database transactions or via parsable files, and delivers output via the GSDB Annotator, SQL queries and traditional flatfiles.

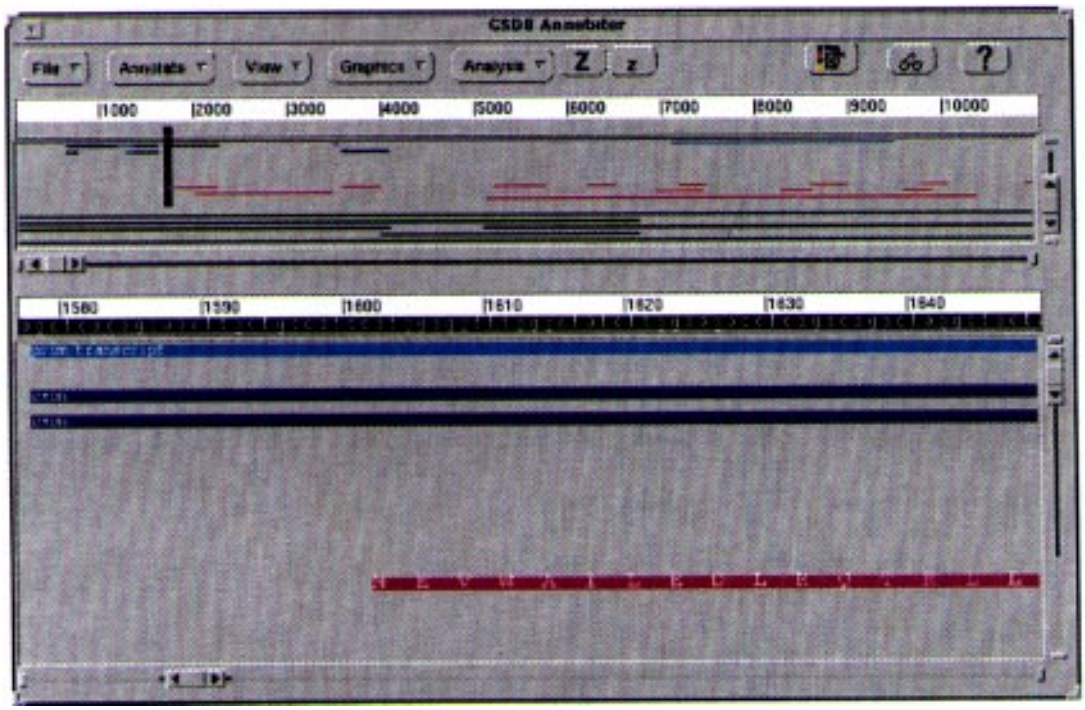

Figure 3. The GSDB Annotator: a database browser and editor. Sequences and annotation are displayed as editable graphics at both low and high resolution. Mouse-driven fill-in forms allow easy addition of annotation to sequences.

GSDB Annotator. The GSDB Annotator is an interactive browser and editor client for GSDB. It runs on multiple platforms including SUN, MacIntosh and PCs with Windows systems (available mid 1996). The GSDB Annotator displays database sequences and analysis results in a graphic format that reflects the way researchers think about sequences. Figure 3 shows the main interface window of the Annotator.

Web. Our World Wide Web Server (http://www.ncgr.org) provides several different access mechanisms to GSDB. Applications are available for adding new sequences, reporting updates, and retrieving responses to internal as well as cross database queries.

Accounts. GSDB user accounts are available on request. All external users may also access the public readonly account. All accounts allow interactive SQL access from any machine on the Internet.

Data exchange with archival databases. Each day GSDB exchanges data in the form of flatfiles with the European Molecular Biology Laboratory, the DNA DataBank of Japan, and GenBank at the National Center for Biotechnology Information. This exchange process enables all sequence databases to maintain a complete and up to date collection of nucleotide sequences.

Bulk data contributions. Large sequencing centers make up the majority of bulk data contributors. To facilitate the entry process of bulk sequence data, GSDB supports both file based and direct SQL approaches. With the file based approach, users generate custom files that are processed through the GSDB object layer into the database. Direct SQL, however, allows users to download data from their local database directly into GSDB via custom software.

Versioning

GSDB's versioning of changes to sequences and annotation provides citability analogous to that offered by scientific journals. In addition to citability, versioning enables researchers access to any given dated data set in order to perform repeated or more extensive analysis with a fixed set of data.

Versioning in GSDB is accomplished through a dual database design in which the production database contains only the most recent versions of all data, and a versioning database contains all historical data. All versioning is done at the row level and each row that is inserted or updated in the production database results in a new row being added to the versioning database. All rows in the versioning database are time-stamped.

FUNCTIONALITY PLANNED FOR GSDB 3.0

Design of the next version of GSDB, 3.0, is well underway. The major structural change between GSDB 2.2 and GSDB 3.0 will be the capability of representing features as shared across multiple sequences. For example, an exon feature may be shared, in GSDB 3.0, across multiple genomic samples that show sequence variants, and across multiple cDNA sequences that show alternative splicing. Shared features will greatly enhance the queryability of GSDB.

The representation of aligned sets of sequences will be enhanced in GSDB 3.0 to explicitly include relationship types, including substrate–product (e.g. genomic to mRNA) and instance–type (e.g. individual isolate to selected representative) relations between sequences. Adding these relations will improve the representation of both expression pathways and collections of isolates that provide mutation or other diversity data. GSDB 3.0

will also include an improved representation of binding relations between sequences.

As the richness of biological annotation grows, the ability to link to external databases that provide additional detail or other types of data increases. GSDB 3.0 will provide capabilities for linking a variety of data to multiple external databases in a way that will support complex inter database joins.

Development plans and draft specifications for GSDB 3.0 will be posted on http://www.ncgr.org/gsdb as they become available. Comments on these documents and requests for additional functionality are appreciated and will be acknowledged.

THE FUTURE: FEDERATED BIOLOGICAL DATABASES

DNA sequence data may have application in any area of biology, biomedicine, agriculture or biotechnology, from basic research in molecular mechanisms, developmental genetics or evolutionary systematics to epidemiology, vaccine design, plant breeding, bioremediation or industrial process development. Conversely, results obtained in any area of biological or biochemical research may be relevant to the structural or functional characterization of a sequenced region. No single database or collection of curators can capture all of biology, or foresee the requirements of all potential applications of biological knowledge. GSDB is, therefore, designed to function as one node in a federation of autonomous, independently curated biological databases spanning many domains and interest areas. As commercial quality tools for distributing complex queries across the Internet become available, the sequence and annotation data maintained in GSDB will become progressively more tightly linked with relevant information from other parts of biology and biotechnology.

CONTACTING GSDB AND NCGR

Information about GSDB as well as data contribution and access tools are available on our World Wide Web site, http://www.ncgr.org, or contact us at one of the addresses below.

GSDB WWW Site	http://www.ncgr.org/gsdb/gsdb.html
GSDB WebSub Submission Tool	http://www.ncgr.org./gsdb/WebSub.html
GSDB WWW on line updates	http://www.ncgr.org/gsdb/update.html
Assistance with WebSub	WebSub@gsdb.ncgr.org
GSDB Submissions	datasubs@gsdb.ncgr.org
GSDB Updates	update@gsdb.ncgr.org
Anonymous FTP	ftp.ncgr.org
Relational Replication Startup	dnr@ncgr.org
GSDB User & Public Accounts	gsdb@ncgr.org

The National Center for Genome Resources (NCGR) is a private, not for profit [501c(3)] corporation established in 1994 to provide information and other resources generated by the genome projects and related research and development to the public and private sectors. NCGR's public projects include GSDB and a database of information on the ethical, legal and social (ELSI) issues raised by genome projects. Requests for further information about NCGR can be addressed to ncgr@ncgr.org or to:

Coordinator of External Affairs
National Center for Genome Resources
1800 Old Pecos Trail
Santa Fe, NM 87505, USA
[1] 505 982-7840 (voice) 7690 (fax)

ACKNOWLEDGEMENTS

The Genome Sequence DataBase is supported by Cooperative Agreement 95ER62062 between the National Center for Genome Resources and the US Department of Energy, Office of Health and Environmental Research. NCGR is also supported by the US Small Business Administration under Award SB OT 94-001.

REFERENCES

1 Fleischmann, R.D. *et al.* (1995) *Science*, **269**, 496–512.
2 Fraser, C.M. *et al*. *Science*, **270**, 397–403
3 Oliver, S.G. *et al.* (1992) *Nature*, **357**, 38–46.
4 Dujon, B. *et al.* (1994) *Nature*, **369**, 371–378.
5 Johnston, M. *et al.* (1994) *Science*, **265**, 2077–2082.
6 Sulston, J. *et al.* (1992) *Nature*, **356**, 37–41.
7 Wilson, R. *et al.* (1994) *Nature*, **368**, 32–38.
8 Adams, M. D. *et al*. (1991) *Science*, **252**, 1651–1656.
9 Adams, M. D. *et al.* (1995) *Nature*, **37** (Suppl.), 3–174.
10 Chen, E., Schlessinger, D. and Kere, J. (1993) *Genomics*, **17**, 651–656.
11 Richards, S., Muzny, D., Civitello, A., Lu, F. and Gibbs, R. (1994) In Adams, M., Fields, C., and Venter, J.C. (eds), *Automated DNA Sequencing and Analysis.* Academic Press, London, pp. 191–198.
12 Martin, C., Mayeda, C., Davis, C., Strathman, M. and Palazzolo, C. (1994) In Adams, M., Fields, C., and Venter, J.C. (eds), *Automated DNA Sequencing and Analysis.* Academic Press, London, pp. 60–64.
13 Smith, M.W., Holmsen, A.L., Wei, Y.H., Peterson, M. and Evans, G.A. (1994) *Nature Genetics*, **7**, 40–47.
14 Waterman, M. *et al*. (1994) *J. Computational Biology*, **1**, 173–190.
15 U.S. Department of Commerce (1993) *Federal Information Processing Standard Publication 127-2: Database Language SQL.* National Institute of Standards and Technology.
16 Cinkosky, M., Fickett, J., Gilna, P. and Burks, C. (1991) *Science*, **252**, 1273–1277.
17 Fields, C. In T. Beugelsdijk (ed.), *Automated Technologies for Genome Characterization.* Wiley, New York, in press.
18 Alschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Molecular Biology*, **215**, 403–410.
19 Uberbacher, E., Einstein, J., Guan, X. and Mural, R. (1993) In Lim, H., Fickett, J., Cantor, C. and Robbins, R. (eds), *Bioinformatics, Supercomputing, and Complex Genome Analysis.* World Scientific, Singapore, pp. 465–476.