18–23 *Nucleic Acids Research, 1997, Vol. 25, No. 1*

The Genome Sequence DataBase version 1.0 (GSDB): from low pass sequences to complete genomes

Carol Harger*, Marian Skupski, Ethan Allen, Christopher Clark+, David Crowley+, Emily Dickinson+, David Easley, Ada Espinosa-Lujan, Andrew Farmer, Chris Fields+, Leandrita Flores, Lynn Harris+, Gifford Keen+, Maurice Manning, Mia McLeod, John O'Neill, Maria Pumilia+, Rhonda Reinert, David Rider+, John Rohrlich+, Yolanda Romero, Jolene Schwertfeger, Gustavo Seluja, Adam Siepel, Gautam Singh, Linda Smyth+, David Stamper, Judy Stein+, Randy Suggs, Rajini Takkallapalli, Nina Thayer, Gary Thompson+, Colleen Walsh+, Frederick Wedgeworth III+ and Peter A. Schad

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

Received October 21, 1996; Accepted October 22, 1996

ABSTRACT

The Genome Sequence DataBase (GSDB) has completed its conversion to an improved relational database. The new database, GSDB 1.0, is fully operational and publicly available. Data contributions, including both original sequence submissions and community annotation, are being accomplished through the use of a graphical client-server interface tool, the GSDB Annotator, and via GIO (GSDB Input/Output) files. Data retrieval services are being provided through a new Web Query Tool and direct SQL. All methods of data contribution and data retrieval fully support the new data types that have been incorporated into GSDB, including discontiguous sequences, multiple sequence alignments, and community annotation.

INTRODUCTION

The age of efficient, inexpensive, high throughput DNA sequencing has arrived. To date two bacterial genomes (1,2), one archeon genome (3), one blue-green algae genome (4) and one yeast genome (5–11) have been completely sequenced and are available to the public. It is anticipated that ~100 or more additional genomes will be completed in the next few years. In addition, the effort to elucidate the complete human genome sequence is in high gear and has already produced over 25 000 sequence tagged sites in addition to many genomic sequences (12). 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 a variety of technical and social challenges. A variety of strategies, from high redundancy shotgunning to directed strategies such as Ordered Shotgun Sequencing (13), sequence mapped gaps (14) or transposon based walking (15) to sampling strategies that produce ordered sequence fragments separated by gaps of known length (16) are used by sequencing labs, and data may be released to the public at any state in the production process. Public sequence databases must be able 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. Multiple laboratories are often involved in the completion of a sequenced region, with one laboratory generating mapped sequence samples, while others produce highly accurate sequences of subregions or cDNAs from expressed genes. Further annotation of the sequence showing diversity among individuals, alternative expression products, or the functions of portions of the sequence may be the work of many other investigators. It is particularly important to include these latter data in a structured database. Often important functional results are documented only in the printed literature and are effectively lost to the community.

GSDB is designed to meet the requirements for a community sequence database outlined by Waterman *et al.* (17). Complex, ad hoc queries in a standard language, SQL (18) are supported. GSDB extends the Electronic Data Publishing paradigm (19) 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 (20). In this model, multiple sequences of one region and functional and structural annotations on sequences are considered to be independent observations and each observation has its own authors and unique identifiers. Critical to making structural, functional and diversity data useable to the community, GSDB supports multi-user editing capabilities, with the necessary authorship, data security, integrity checking and versioning mechanisms

* To whom correspondence should be addressed. Tel: +1 505 982 7840; Fax: +1 505 982 7690; Email: cah@ncgr.org

+Present address: Molecular Informatics Incorporated, Santa Fe, NM 87505, USA

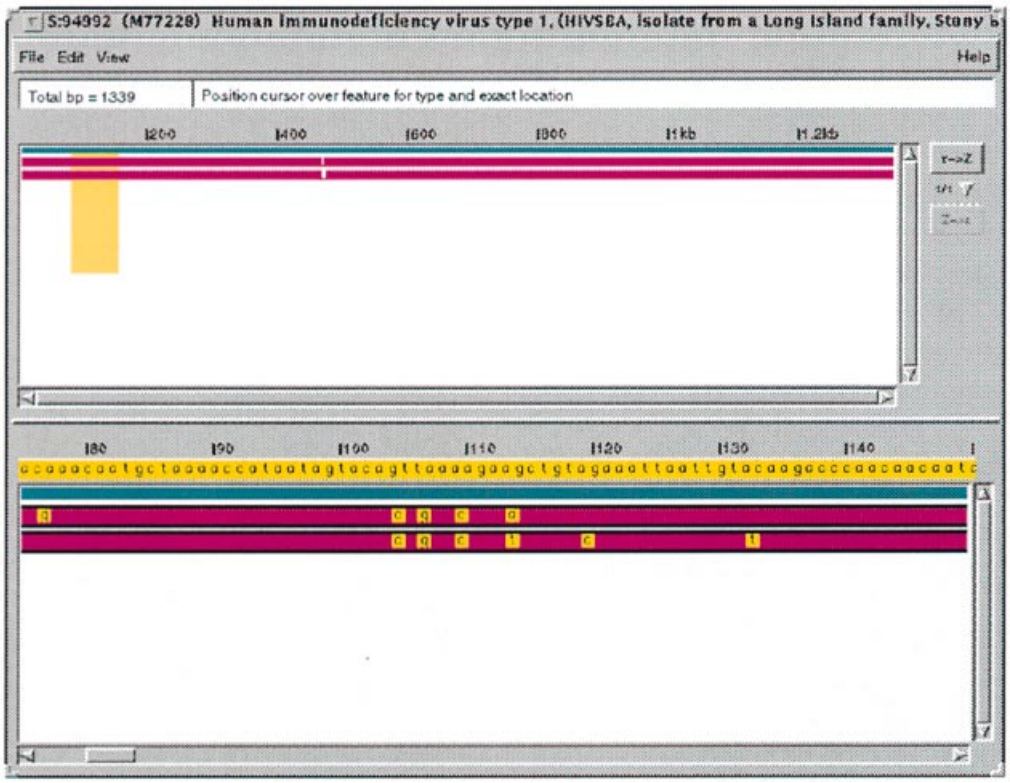

Figure 1. The Annotator can align multiple sequences to the primary sequence. Alignments are displayed in a similar way as features: the Annotator shows differences with the yellow-highlighted letters in the magenta display bar.

needed to ensure that multiple authors do not overwrite each other's work.

GSDB 1.0: A NEW LOOK, FEEL AND FUNCTIONALITY

The conversion of GSDB to a new architecture and schema was completed August 1996. This schema was previously referred to as GSDB 2.2 (21). Since this will be the first version of the database which is a community curated and owned database, a decision was made to rename it GSDB 1.0. The new GSDB replaces the notion of an 'entry' employed by the archival sequence databases with discrete structures to represent sequences and biological annotation. Both sequences and biological features receive accession numbers that serve as permanent identifiers which can be referenced in journals. By providing a stable identifier for each feature as well as each sequence, GSDB 1.0 supports community (third party) annotation of sequences by multiple research groups and simplifies the retrieval of biological annotation independent of sequences.

New data types

The ability to provide the research community with direct access to GSDB to deposit and curate data was one of the reasons for converting to the new schema. Another, equally important, reason was the incorporation of new data types that are useful for high throughput DNA sequencing data. The new data representations in GSDB include: sequence alignments, discontiguous sequences, data ownership, analysis data, and sequencing confidence.

Sequence alignments

Sequence alignments are used to represent a variety of sequence relationships. Currently alignments are used in: the building of large contiguous sequences from smaller fragments, the comparison of allelic variations within a specific gene or variations within a gene family, phylogenetic comparisons, and the representation of genomic sequence-to-product relationships. GSDB is capable of representing all of these kinds of sequence alignments. Multiple sequences can be aligned to your primary sequence of interest. What is actually stored in the database is the relationship between each aligned sequence and the primary sequence of interest. These stored relationships denote the exact base pair spans that are corresponding between the primary sequence and an aligned sequence. In addition, differences which occur between the primary sequence and any aligned sequence are easy to identify when they are displayed in the GSDB Annotator (Fig. 1) and they are stored in the database for ease of querying.

Discontiguous sequences

In addition to actual sequence data, many of the sequencing strategies that are being employed today provide the researcher with relative position and orientation information as it applies to a set of disconnected pieces of sequence. This information is often valuable to the research community as a whole. In the past, this information has been difficult to incorporate into a database record in a standard way. As a remedy to this problem, GSDB assigns each set of disconnected sequence pieces (for which relative order, distance, and orientation are known) a sequence

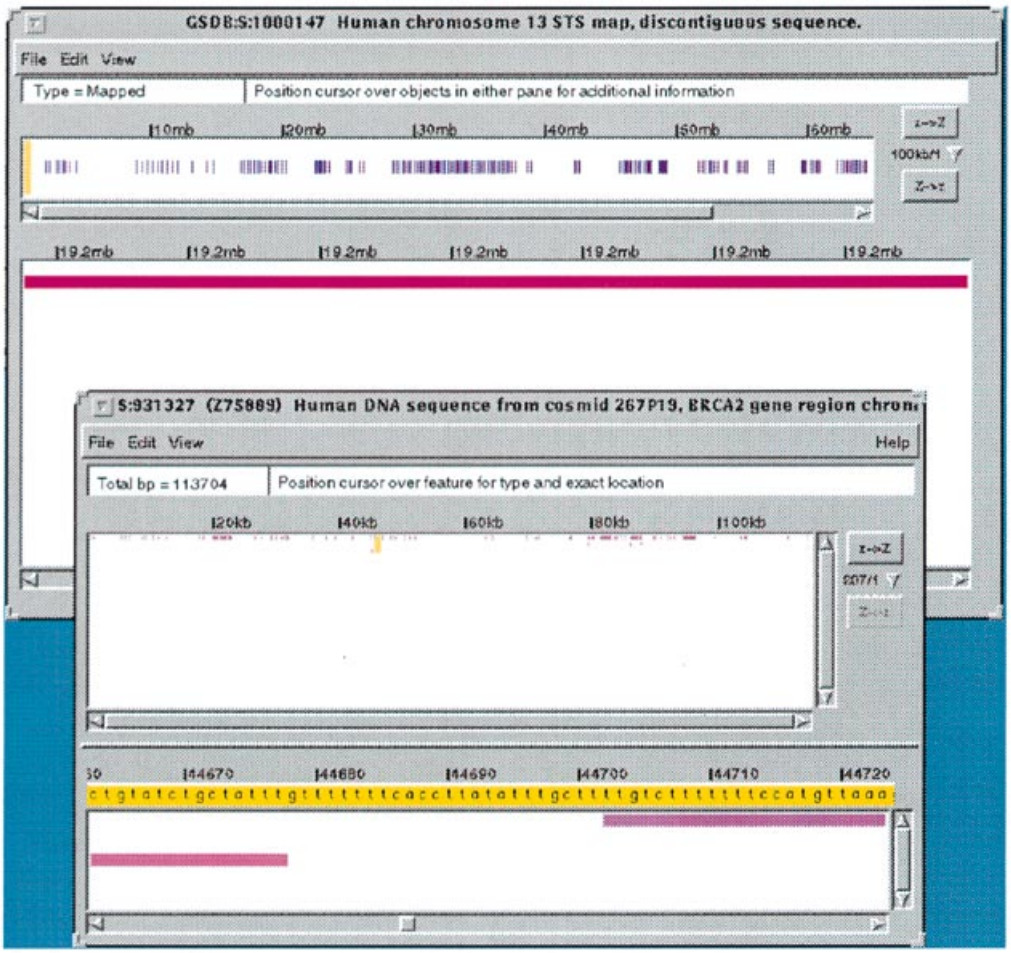

Figure 2. GSDB Annotator display of a human chromosome STS map. The Annotator's discontiguous sequence display allows multiple resolution displays of a multi-Mb sequence. The Annotator display will zoom in several steps from >100 Mb to individual sequence pieces. The discontiguous sequence piece opens below to display its annotation at a higher resolution.

accession number and the set is referred to as a discontiguous sequence. A discontiguous sequence can be used to represent actual physical maps (e.g., human chromosome STS maps, Fig. 2), ordered sequences (the genomic exons of a gene where the introns were not sequenced), or binned sequences (several pieces of sequence are all from the same clone but for which information about relative order is not available). All of these types of discontiguous sequences are easily displayed in the GSDB Annotator.

Data ownership

The research community is the expert with respect to the data in GSDB and as such the community should have the capability to curate and editorialize the data. GSDB has been designed so that data is owned by the contributing researcher. Ownership of data is assigned based upon the database account (login and password) that was used to access the database while the data was being deposited. GSDB enforces the rule that only the owner of a piece of data has the permission to edit or change that data. Since each piece of data is owned by the contributing researcher, GSDB can now allow the research community to add biological annotation to any sequence in the database regardless of who owns the sequence. This type of data contribution is referred to as community annotation or third party annotation and may occur over an extended period of time. The e-mail address, phone number, and postal address of each owner are accessible through the GSDB Annotator and the Web Query interface.

Analysis data

Determining the biological 'content' of a sequence can be quite a lengthy and arduous task. In addition, the current high rate at which sequences are being produced makes it necessary to streamline the process of determining the biological 'content' of each sequence. As a consequence, many researchers are turning to analysis software to assist them in identifying potential biological features within their sequences. This information is often very valuable to the research community. GSDB now provides a standard way to record the program name, the run date, the parameters used, and the results in the database. All of these data can be viewed in the GSDB Annotator.

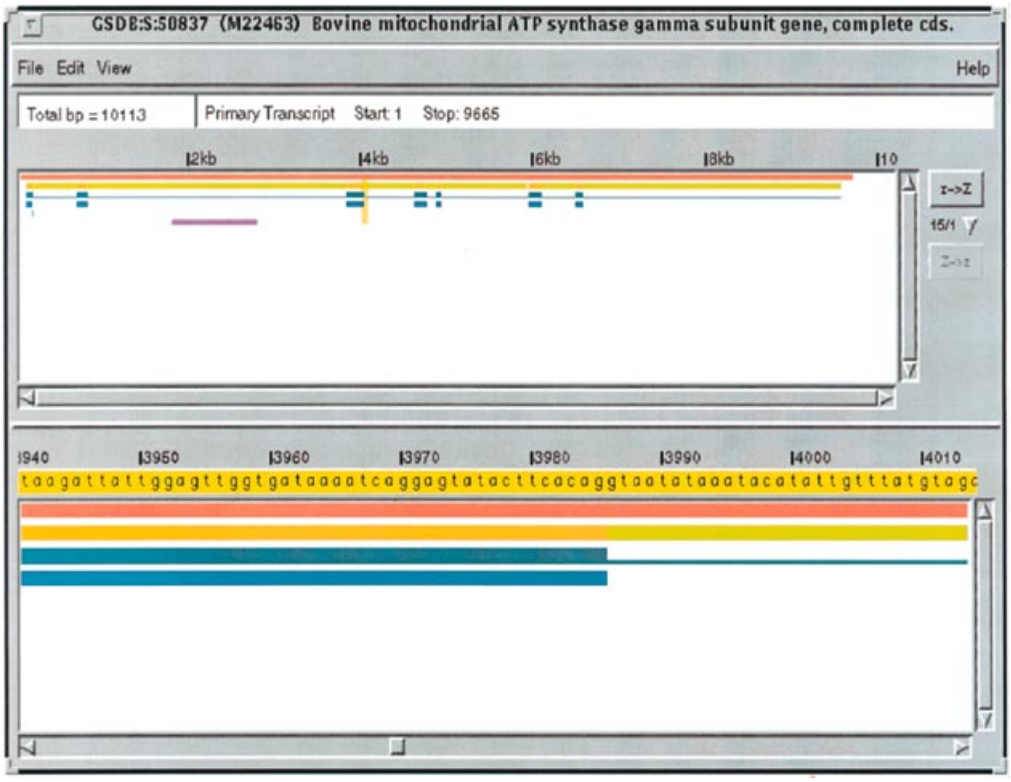

Figure 3. The GSDB Annotator: a database browser, sequence editor, and submission tool. Sequences and annotation are displayed as editable graphics at both low and high resolution. Mouse-driven fill-in forms allow easy addition of annotation to sequences.

Sequencing confidence

The issue of whether to believe the base call at each position of a sequence is critical for both the researcher producing the sequence and for a researcher in the community who is interested in the sequence. It is most beneficial to both individuals for the producer of the sequence to be able to specify the degree of confidence that he has in a specific base call. GSDB now provides the ability to assign a sequencing confidence value to a single base or to a range of bases. We are currently encouraging researchers to assign values to those bases which fall below an acceptable minimum value. This allows researchers who view sequence data from GSDB to quickly and easily identify base calls which are suspect within the sequence.

Data submission and exchange

The onset of high throughput sequencing and the availability of large contiguous sequences (e.g., complete bacterial genomes or chromosomes) has had considerable impact on the type and quality of submission methods that GSDB provides to the research community. We currently provide two methods for the submission of sequences and biological annotation; the GSDB Input/Output (GIO) file format and the GSDB Annotator. Regardless of the method of data submission, all data that goes into GSDB is associated with a GSDB user account so that ownership of the data is properly assigned. These accounts are free to the public and can be obtained through our web site.

The GIO file submission method is designed to be an automated bulk submission method and is specifically designed for labs producing large amounts of sequence data. The file format is easy to produce via simple scripts and we are offering assistance in the development of these scripts. Moreover, the generation of GIO files can be integrated into the sequence processing regimen of most laboratories. In addition, the GIO file is parsable and can be automatically read into GSDB without human intervention. This streamlines the submission process at both the laboratory and database end. GIO files can also be used to perform updates to existing sequence data that were previously submitted.

The GSDB Annotator

The GSDB Annotator is designed to support the submission needs of individuals who occasionally need to submit sequences to the database and who need to review, edit or update the information of an individual sequence. The GSDB Annotator is a graphical, interactive editor client for GSDB. It runs on multiple platforms, including UNIX, Macintosh Power PC and Windows NT. The GSDB Annotator displays database sequences, features, analysis results, and alignments in a graphic format that is user friendly and biological relevant (Fig. 3). One way to ensure a higher quality of information associated with each sequence that is contributed to GSDB through the use of the Annotator is the incorporation of many software checks into the interface. For example, if you indicate that there is a coding region within your sequence and you declare that it is complete at both ends, the

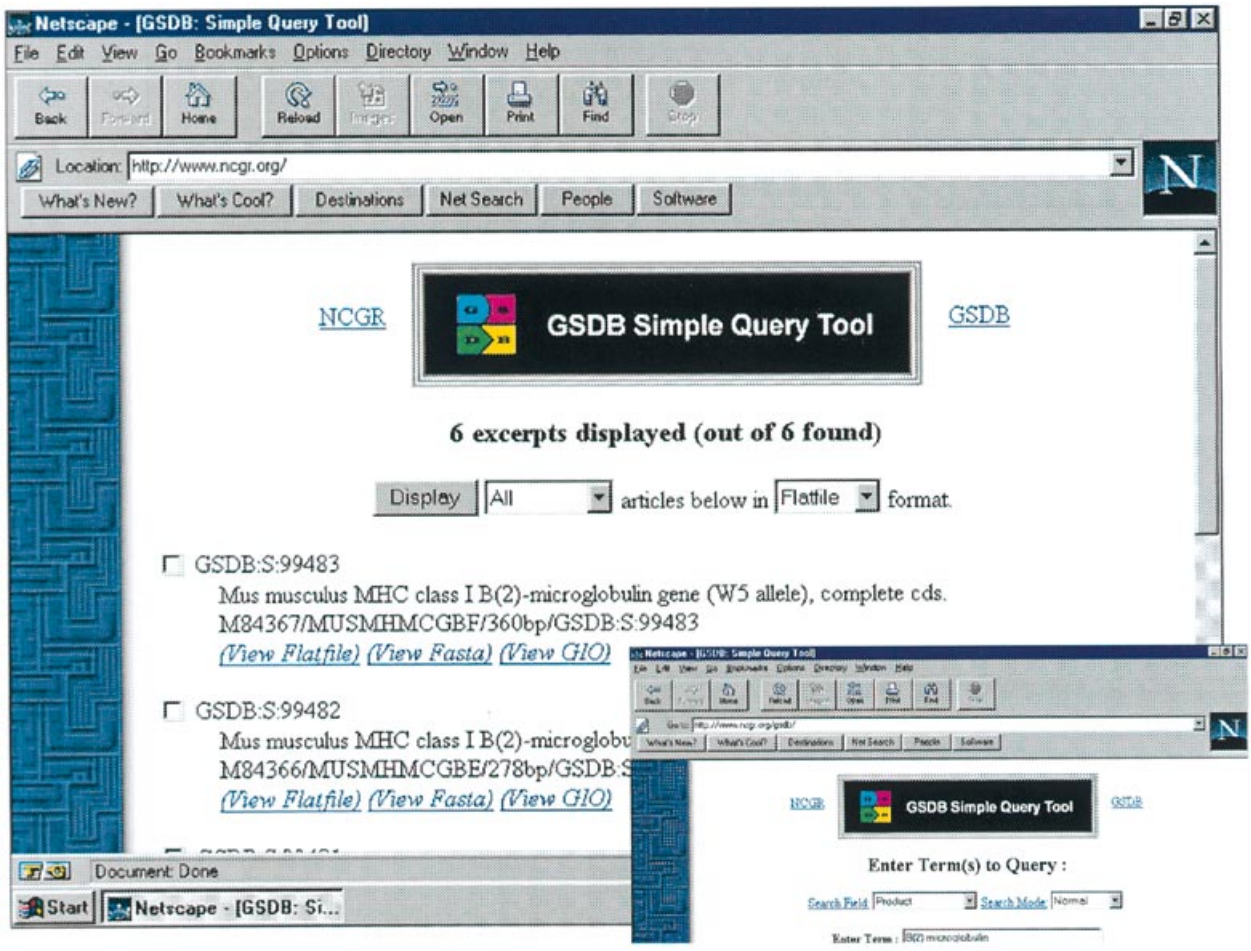

Figure 4. GSDB simple query tool. The user can search GSDB on several database fields, including gene, keyword, organism, and author, among others. The input window is shown above, the output below. By clicking on the hotlinks in the query output the user can view sequences of interest in several types of file formats.

software will return an error message if the defined coding region does not begin with an accepted start codon or end with an accepted stop codon. The GSDB Annotator is free software that can be obtained from our web site.

As part of GSDB's effort to provide the community with a useful data set, GSDB exchanges data with the archival databases. Data is exchanged nightly in the form of flatfiles with the DDBJ, EMBL and GenBank databases. Data from these databases is incorporated into GSDB and is available for the addition of annotation via our community annotation mechanism.

Improving data quality

To improve the quality of the data and its usefulness to the community, we have begun to assess the extent of vector contamination in GSDB. Sequences that are identified as vector contamination will be removed from the sequence and included in a comment section that can be viewed by any researcher. This will allow each researcher that uses the sequence to determine whether the contamination is truly vector contamination. We have begun to remove vector contamination from sequences in the rodent and primate divisions of GSDB, and will continue with other divisions.

Data retrieval

GSDB delivers output via direct SQL, the query tool available on the web, and the Annotator. Direct SQL can be done using the ad hoc SQL option available on the GSDB web site. The GSDB Simple Query Tool available on the GSDB web site allows users to retrieve data by searching on accession number, author, organism, gene name, product, and several other fields (Fig. 4). The data retrieved can be saved in any of three formats: flatfile, fasta or GIO (GSDB input/output format). The GIO output can be viewed with all annotation using the Annotator. Advanced queries based on the length of the sequence, alignments with the sequence, the confidence in the sequence, and others will be available soon.

FUTURE DIRECTIONS

One of the major goals for GSDB is to improve the richness of the annotation in the database. The greater the level of annotation, the more useful the sequence and the database will be to molecular biologists. To further this goal, we have begun to build ordered discontiguous sequences of the human chromosomes, including sequence tagged sites, expressed sequence tags, and genomic

sequences. We have also begun building alignments of related sequences that can be accessed using the annotator.

Another major goal for GSDB is to provide flexibility in retrieval and viewing of sequences from the database. It is important to provide access to redundant sequences that provide data on variation within and among populations, while still allowing access to individual sequences to researchers that wish only for one sequence.

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 will provide capabilities for linking a variety of data to multiple external databases in a way that will support complex inter-database joins.

We anticipate making incremental improvements in the schema, submission methods, and retrieval capabilities. These improvements will be announced before implementation and are likely to occur on a quarterly basis.

CONTACTING GSDB AND NCGR

Information about GSDB as well as data access tools are available on our World Wide Web site given below or by contacting us at one of the addresses below.

gsdb@ncgr.org	GSDB Annotator & Data Support Info.
gsdbinfo@ncgr.org	GSDB General User Information
update@ncgr.org	GSDB Update Information
dbadmin@ncgr.org	Satellite copies of GSDB
webmaster@ncgr.org	Questions or comments about the GSDB Web page
www.ncgr.org/gsdb	GSDB Home Page
www.ncgr.org/gsdb/sql.html	GSDB Ad hoc SQL Queries
www.ncgr.org/gsdb/user-account.html	GSDB User Acct Registration
www.ncgr.org/cgi-bin/test.ff	GSDB Sequence & Feature Retrieval
www.ncgr.org/gsdb/query.html	GSDB Simple Query Tool

The National Center for Genome Resources (NCGR) is a private, not for profit 501(c)(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 the Genetics and Public Issues program, a resource for information on a wide variety of genetically inherited diseases. Requests for further information about NCGR can be addressed to ncgr@ncgr.org or to: Director of External Affairs, National Center for Genome Resources, 1800A Old Pecos Trail, Santa Fe, NM 87505, USA. Tel: +1 505 982 7840; Fax: +1 505 982 7690; 1 800 450 4854

ACKNOWLEDGMENTS

The Genome Sequence DataBase is supported by Cooperative Agreement DE-FC03-95ER62062 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 SBAHQ-95-1-0013.

REFERENCES

1 Fleischmann, R.D. *et al*. (1995) *Science*, **269**, 496–512.
2 Fraser, C.M. *et al*. (1995) *Science*, **270**, 397–403.
3 Bult, C.J. *et al*. (1996) *Science*, **273**, 1058–1073.
4 Kaneko, T. *et al*. (1996) *DNA Res.*, **3**, 109–136.
5 Bussey, H. *et al*. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 3809–3813.
6 Feldmann, H. *et al*. (1994) *EMBO J.*, **13**, 5795–5809.
7 Oliver, S.G. *et al*. (1992) *Nature*, **357**, 38–46.
8 Murakami, Y. *et al*. (1995) *Nature Genet.*, **10**, 261–268.
9 Johnston, M. *et al*. (1994) *Science*, **265**, 2077–2082.
10 Galibert, F. *et al*. (1996) *EMBO J.*, **15**, 2031–2049.
11 Dujon, B. *et al*. (1994) *Nature*, **369**, 371–378.
12 Adams, M. D. *et al*. *Nature* (in press).
13 Chen, E., Schlessinger, D. & Kere, J. (1993) *Genomics* **17**, 651–656.
14 Richards, S., Muzny, D., Civitello, A., Lu, F. & Gibbs, R. (1994) In *Automated DNA Sequencing and Analysis* (eds. Adams, M., Fields, C., & Venter, J. C.) Academic Press, London, pp. 191–198.
15 Martin, C., Mayeda, C., Davis, C., Strathman, M. & Palazzolo, C. (1994) In *Automated DNA Sequencing and Analysis* (eds. Adams, M., Fields, C., & Venter, J. C.) Academic Press, London, pp. 60–64.
16 Smith, M. *et al*. (1994) *Nature Genet.* **7**, 40–47.
17 Waterman, M. *et al*. (1994) *J. Computat. Biol.* **1**, 173–190.
18 Department of Commerce, Federal Information Processing Standard Publication 127–2: Database Language SQL. National Institute of Standards and Technology (1993).
19 Cinkosky, M., Fickett, J., Gilna, P. and Burks, C. (1991) *Science* **252**, 1273–1277.
20 Fields, C. In *Automated Technologies for Genome Characterization* (ed. T. Beugelsdijk) Wiley, New York (in press).
21 Keen, G. *et al*. (1996) *Nucleic Acids Res.*, **24**, 13–16.