Nucleic Acids Research, 1998, Vol. 26, No. 1 21–26

The Genome Sequence DataBase (GSDB): improving data quality and data access

C. Harger*, M. Skupski, J. Bingham, A. Farmer, S. Hoisie, P. Hraber, D. Kiphart, L. Krakowski, M. McLeod, J. Schwertfeger, G. Seluja+, A. Siepel, G. Singh, D. Stamper, P. Steadman, N. Thayer, R. Thompson, P. Wargo, M. Waugh, J. J. Zhuang and P. A. Schad

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

Received November 3, 1997; Accepted November 4, 1997

ABSTRACT

In 1997 the primary focus of the Genome Sequence DataBase (GSDB; www.ncgr.org/gsdb) located at the National Center for Genome Resources was to improve data quality and accessibility. Efforts to increase the quality of data within the database included two major projects; one to identify and remove all vector contamination from sequences in the database and one to create premier sequence sets (including both alignments and discontiguous sequences). Data accessibility was improved during the course of the last year in several ways. First, a graphical database sequence viewer was made available to researchers. Second, an update process was implemented for the web-based query tool, Maestro. Third, a web-based tool, Excerpt, was developed to retrieve selected regions of any sequence in the database. And lastly, a GSDB flatfile that contains annotation unique to GSDB (e.g., sequence analysis and alignment data) was developed. Additionally, the GSDB web site provides a tool for the detection of matrix attachment regions (MARs), which can be used to identify regions of high coding potential. The ultimate goal of this work is to make GSDB a more useful resource for genomic comparison studies and gene level studies by improving data quality and by providing data access capabilities that are consistent with the needs of both types of studies.

INTRODUCTION

A new era in biological research involving the comparison and functional analysis of complete genomes has begun. This revolution is the direct result of several technological advances. First, the advent of efficient, inexpensive, high throughput DNA sequencing strategies (1–5) was necessary in order to produce large volumes of accurate sequences. The rate at which progress has been made in the Microbial Genome Initiative is evidence of the effect that high throughput DNA sequencing has had on biological research. In the last year the number of complete microbial genomes, that are available in the public nucleotide sequence databases, has more than doubled and another 100 are expected in the next few years. Singly, each of these sequences and its associated biological annotation contribute to the advancement of the understanding of gene function and microbial biochemistry/physiology (1). Collectively the availability of these complete genomes provides researchers with raw data to perform comparative genomic studies. For example, last year a minimal gene complement for cellular life was determined based upon the comparative analysis of the *Mycoplasma genitalium* genome with that of *Haemophilus influenzae* (6). In addition, a comparative analysis (at the protein level) of eubacterial, archaeal, and unicellular eukaryotic genomes was conducted to investigate the origins of archaea and novel protein functions (7). These examples mark the beginning of the era of comparative and functional genomics.

The availability of complete genome sequences has not only had a significant impact on the study of microbes and the Human Genome Project, in terms of sequencing and mapping efforts, but also upon other areas of biology such as agriculture (8) and bioremediation (9).

The second technological advancement that was necessary for the realization of comparative and functional genomics was the development and improvement of sequence analysis algorithms (10–12). These advancements are evident by the increased numbers of biological feature prediction software and homology identification software that are available at web-sites like Pedro's BioMolecular Research Tools (www.public.iastate.edu/~pedro/research_tools.html) and BCM Search Launcher (13, http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html). Improvements to tools like Grail (12) and GeneFinder (http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html) to predict introns, exons and protein coding regions in a variety of organisms and homology searching tools like Beauty (10) and BLAST (11) have also been important to the revolution. The GSDB web site also contains a tool to detect matrix attachment

*To whom correspondence should be addressed. Tel: +1 505 982 7840; Fax: +1 505 995 4439; Email: cah@ncgr.org

+Present address: SmithKline Beecham Pharmaceuticals, Bioinformatics, UW 2230, 709 Swedeland Road, King of Prussia, PA 19460, USA

regions, which are regions that often mark concentrated areas of transcription factor binding (14, www.ncgr.org/MarFinder).

Lastly, the development of effective computer and software systems to manipulate and manage large volumes of data was necessary for the realization of comparative and functional genomics (15,16). The ability to query the public nucleotide sequence databases in an efficient manner, when they contain more than one billion bases (15–17) and associated biological annotation is the direct result of improvements in computer, database and network systems.

The ability to produce, analyze, and store large volumes of nucleotide sequence and associated data are the technological cornerstones of comparative and functional genomics.

The Genome Sequence DataBase (GSDB) located at the National Center for Genome Resources (www.ncgr.org) is poised to support many types of data that are essential to comparative and functional genomics studies. In August 1996 GSDB completed a database conversion (15,18) that included upgrading of the Sybase system and improvement of the database schema. As a result of this conversion, GSDB can now store a complete genome (of any number of base pairs) as a single sequence, sequencing confidence data, sequence analysis results/scores, sequence alignments, discontiguous sequences (discontigs), data ownership and researcher defined features. While all of these changes are useful to the researcher community as a whole, these changes are crucial to the support of genomic level sequence comparisons and functional analyses. In most public nucleotide sequence databases it is difficult to adequately evaluate the significance of a biological feature. For example, comparing a protein coding region that has been hypothesized in one sequence based upon significant homology to another sequence which contains a protein coding region is not possible given the data stored in most nucleotide databases. Therefore comparison of the two sequences can only be accomplished by repeating the homology search on the same data set using identical search parameters. Storage of search parameters and scores with hypothetic features would greaty diminish the need to recreate homology searches. GSDB is the only public nucleotide sequence database designed to support the representation of such data.

DATA QUALITY IMPROVEMENTS

One of the primary driving forces behind both GSDB data related and programming projects are the needs of the research community. In order for a public database to be useful to researchers, the quality of data within it must be relatively high. In a relational database like GSDB improvements in data quality can occur from either the removal of erroneous data, or from the addition and creation of more meaningful data annotation and data relationships. Discussed below are two ongoing projects; one of which involves the identification and removal of erroneous sequence data (vector contamination) and one of which involves organizing multiple related sequences into meaningful groupings through the use of the database's ability to represent sequence alignments and discontiguous sequences.

Vector contamination

In this age of high volume sequence production, many researchers are relying more heavily on homology with annotated sequences in the public nucleotide sequence databases to determine the gene content of their sequence(s). Consequently, it is critical that the sequences in the public nucleotide sequence databases not contain vector contamination. The incorporation of vector sequences into the public nucleotide sequence databases has been a problem for a long time. Various researchers have examined this problem since 1992 (19–25). Recently, the GSDB staff reassessed this data quality issue and analyzed the sequences in the database to determine the extent and progression of vector contamination.

The initial phase of the analysis involved the creation of a database that contained ~200 unique multiple cloning sites that are commonly used in cloning vectors. A homology search was then performed between each sequence in GSDB and the multiple cloning site sequence database. This initial screen was used to identify potentially contaminated sequences. The initial screening was limited to only the multiple cloning sites because these sequences are rather rare in natural sequences and they are frequently adjacent to the cloned sequence of interest. Other regions of cloning vectors are frequently derived from naturally occurring sequences and they are infrequently adjacent to the cloned sequence of interest. By only searching against the multiple cloning sites the number of sequences that were incorrectly identified as containing vector contamination was minimized. This method of vector identification was tested on the data set that was available to Lamperti *et al*. in 1992 (25). The percentage of identified contamination that we obtained was slightly higher (0.3% versus 0.23%) but in concurrence with their published results (25).

Following initial identification, a potentially contaminated sequence was compared to a database of complete vector sequences in order to identify the entire span of vector contamination. In addition, these sequences were examined for the existence of a restriction enzyme site at the junction between the potential vector sequence and the cloned sequence of interest. The finding of a restriction site at the proposed vector-cloned sequence junction further supported the argument that all of the vector contamination in the sequence had been identified correctly.

In the nearly 1 000 000 sequences that were examined, approximately 0.36% were found to contain vector sequences. Most of these sequences only contained vector contamination at either the 5′ or 3′ end, while a few contained vector contamination at both ends. The vector contamination has been denoted with annotation in all of these sequences and they are in the process of being removed from the sequence and stored in a comment attached to the sequence. Over 75% of the sequences that were identified as containing vector contamination were also found to contain a restriction enzyme site at the junction (26).

A much smaller number of sequences appeared to contain vector contamination not at the ends of the sequence, but at internal positions of the sequence. This potential vector contamination has been annotated as such, but will not be removed from the sequence until the data submitter has been consulted.

In the data submitted during the 5 year period that was examined, the overall level of vector contamination in the database appeared to remain relatively constant at a value of <1%. However, the study also revealed that >50% of the contamination that was incorporated into the database in the last 2 years was contained in EST and STS sequences. Since these sequences, especially the ESTs are routinely used in homology searches, it is imperative that the vector contamination be identified and removed from the database, or else research may be influenced by erroneous homology matches.

The immediate goal of identifying sequences within the database that contain vector contamination has been successfully completed and the procedure/program to remove this contamination is in development. It is anticipated that the subsequent removal of these vector sequences will be completed by May 1998. In addition, the automation procedure to identify vector contamination prior to its incorporation into the database is being designed.

Unique sequence and annotation representation

One of the primary ways in which GSDB differs from the IC (International Collaboration—DDBJ, EMBL and GenBank) databases is in the types of data which can be stored in the database and in data representation. First, GSDB stores a contiguous string of nucleotides as a single sequence in the database regardless of size. Consequently a complete bacterial genome or fungal chromosome can be retrieved as a single file. In addition GSDB assigns ownership of data within the database to the contributing researcher. Finally, GSDB supports the incorporation of many other types of data that are not supported in the IC databases. Included in this set of unique data types are: sequence alignments, sequence analysis data, sequencing confidence values, and discontiguous sequences. During the last year the GSDB staff has been utilizing these unique capabilities to represent sequence alignments and discontiguous sequences (Table 1) to augment sequence annotation and representation in the database.

Sequence alignments

In the past 2 years several complete microbial genomes, viral genomes, fungal chromosomes, and naturally occurring plasmid sequences have been incorporated into GSDB. During 1997 the GSDB staff made significant progress in aligning small pieces of genomic sequences to the complete genomic sequence of the same organism. Base pair differences, as well as the base pair spans over which two sequences align are retained in the database. Approximately 2100 HIV 1 sequences were compared to the complete genomic sequence of HIV 1 strain HXB 2 (GSDB:S:135829) the HIV research community reference standard sequence. Alignments were created between this sequence and any other HIV 1 sequence that showed >80% homology over 95% of the length of the smaller sequence as determined by BLAST 2.0 and FASTA (11,28).

A similar comparison was performed with the complete genome of *Escherichia coli* (27, GSDB:S:1649882) and other smaller *E.coli* sequences in the database. BLAST 1.4.9 and BLAST 2.0 were used to identify sequences that had >95% identity over the total length of the smaller sequence. In total, >2000 of the *E.coli* sequences within the database were incorporated into the complete genome record as sequence alignments. In this case, the base pair differences that occur between the complete genome and other aligned sequences are more likely the result of sequence variations or sequencing errors.

The GSDB staff has not defined the meaning of these differences, as it is not our role. It is our role to present the data to the research community in an unbiased format, so that individual researcher can decide the value of the differences for him/herself.

The HIV 1 and *E.coli* complete genomes are only two examples of sequences for which the GSDB staff has performed intraspecies homology searches. A listing that includes all genomes that were analyzed prior to October 1997 is provided in Table 1. A listing of sequences that have been analyzed since October 1997 is provided in the 'What's New' section of the GSDB Web site (Fig. 1, www.ncgr.org/gsdb). The GSDB capability of storing sequence alignment information is flexible in the kinds of sequence alignments that can be represented. In addition to representing intraspecies homology alignments, this capability can also be used to represent the relationship between a completed clone sequence and the subclone sequences were assembled into the complete clone. It can also be used to represent the relationship between a transcribed molecule (mRNA, rRNA, tRNA, etc.) and the corresponding genomic sequence. Lastly, it can be used to represent interspecies homology comparisons, especially those that are utilized to predict that a specific gene is in a sequence because of its similarity to a sequence with that gene from another organism.

Table 1. Sequence alignments

GSDB accession no.	Sequence description
GSDB:S:1649882	*E.coli* complete genome and sequence alignments
GSDB:S:43226	*H.influenzae* complete genome and sequence alignments
GSDB:S:135829	HIV complete genome and sequence alignments
GSDB:S:941927	*M.janaschii* complete genome and sequence alignments
GSDB:S:43082	*M.genitalium* complete genome and sequence alignments
GSDB:S:1492975	*Rhizobium* sp. pNGR234a complete sequence and alignments
GSDB:S:1025912	*S.cerevisiae* chrom. III complete sequence and alignments
GSDB:S:1650518	*S.cerevisiae* chrom. V complete sequence and alignments
GSDB:S:848927	*S.cerevisiae* chrom. VI complete sequence and alignments
GSDB:S:1327451	*S.cerevisiae* chrom. VIII complete sequence and alignments
GSDB:S:1637738	*S.cerevisiae* chrom. IX complete sequence and alignments
GSDB:S:1347770	*Synechocystis* sp. complete genome and sequence alignments

Discontiguous sequences

A discontiguous sequence is a set of sequence fragments that are part of a larger contiguous sequence. (This set of sequences is provided a single GSDB sequence accession number, so that the set is easy to reference and retrieve.) Consequently there is a known physical relationship between these fragments, but they do not abut to one another, nor do they overlap. The known physical relationship between these fragments may be as simple as the knowledge that the fragments are all from the same cosmid (or any sequence of known size) and therefore cannot be more than X Kb apart. Alternatively the relationship may be as specific as knowing both the order in which the fragments occur within the

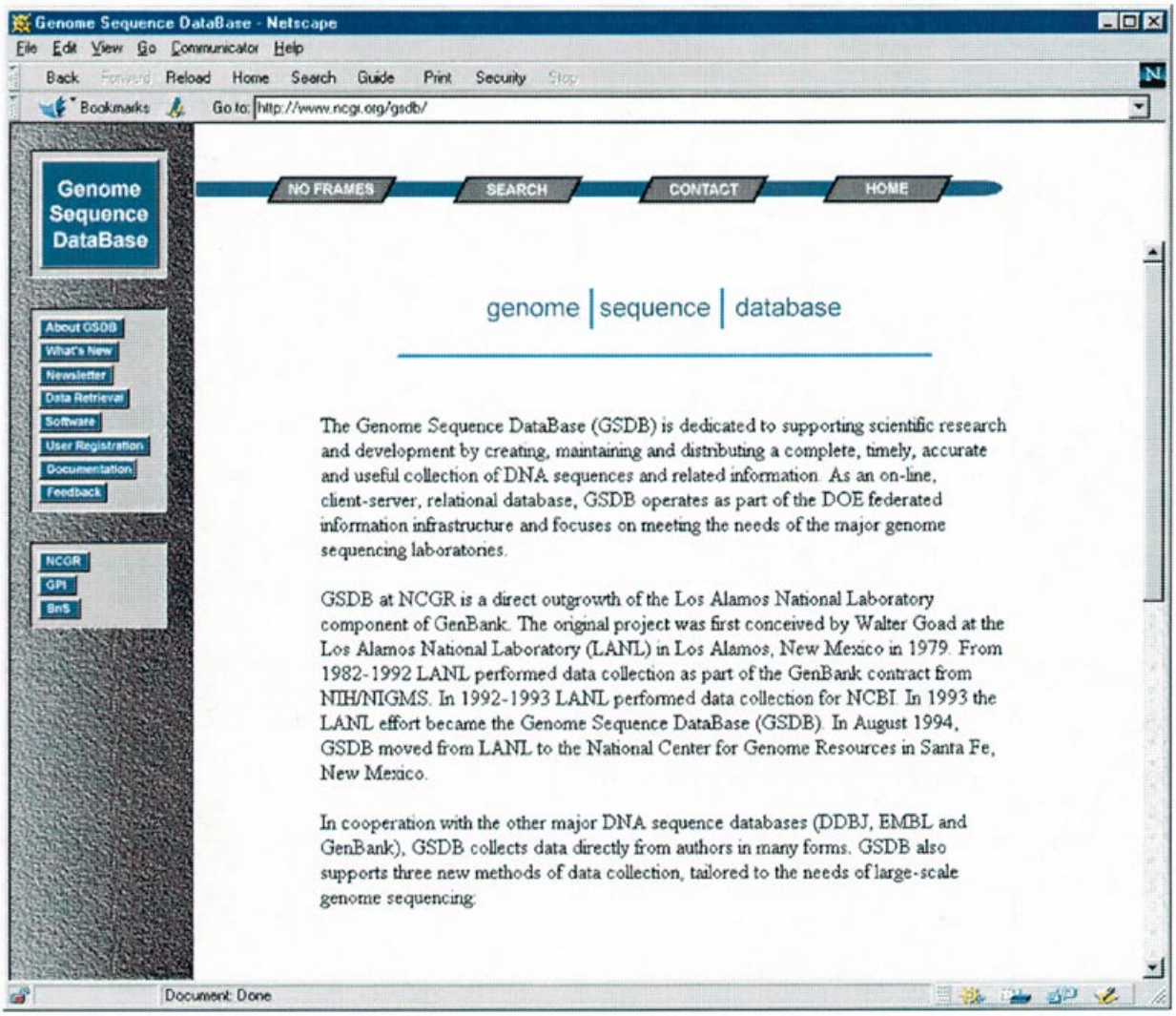

Figure 1. The GSDB web site.

larger contiguous sequence and the physical distance between them.

By constructing discontiguous sequences for sets of sequences in the database where physical relationships between the sequences are known, the GSDB staff has been organizing the data into more meaningful sets. One of the goals of the staff is to continuously review these sets and to add new fragments to the sets when they become available in the database. This effort will continue until all the gaps between the fragments are filled in and a single contiguous sequence has been 'built'.

In the past year the GSDB staff constructed two sets of discontiguous sequences, corresponding to the genomes of two agriculturally important crops, corn and rice. A single discontiguous sequence was constructed for each chromosome in each of these genomes. In addition, the discontiguous sequences that correspond to the human chromosomes 1–22 and chromosome X (15) were updated by the addition of a significant number of sequence fragments. Accession numbers and descriptions of all of these discontiguous sequences can be found in the 'What's New' section of the GSDB web site (www.ncgr.org/gsdb). As more discontiguous sequences become available, accession numbers will be posted in 'What's New'.

The physical position of each sequence fragment is stored in the database in terms of kilobases from the left end of the discontiguous sequence and an uncertainty value. By convention the left end of the discontiguous sequence is always the terminal region of the short arm (p arm) of the chromosome. In the case of the discontiguous sequences corresponding to the human chromosomes, most of the newly added fragments and all of the fragments that were used initially to construct these discontiguous sequences are STSs. The staff has also been working to place other, larger genomic sequences into the discontiguous sequences. This is being accomplished by performing BLAST homology searches between human genomic sequences and the sequence fragments that comprise a discontiguous sequence. When a genomic sequence is found to overlap/contain a specific sequence fragment, the genomic sequence is incorporated into the appropriate position of the discontiguous sequence.

In the upcoming year, efforts to improve data quality will continue both through the identification and correction of

inaccurate data and data relationships, and through the incorporation of appropriate sequence relationships and other meaningful annotation.

DATA ACCESSIBILITY IMPROVEMENTS

During 1997 the GSDB staff also focused on improving the ease with which researchers can access sequences and annotation. Four major advances in data access were accomplished in this year. These include the release of the Sun and Macintosh versions of the graphical database interface tool, Annotator, the implementation of a timely update process for the web-based query tool, Maestro, the development of a web-based tool to retrieve selected portions of sequences, Excerpt, and the development and distribution of GSDB flatfiles that include the annotations which are unique to GSDB.

Annotator

The graphical database interface tool, Annotator, was released for both Sun and Macintosh platforms during 1997. This tool is designed to support the submission of individual sequences to GSDB and to allow researchers to review, edit and update the information contained within a single sequence. Annotator displays sequences, features and alignments from the database in a graphic format that is biologically relevant. Annotator is free software that is available from the 'Software' section of the GSDB Web site (www.ncgr.org/gsdb).

The scope of Annotator is somewhat limited because it is platform dependent. However, the utility of Annotator for viewing and editing data has reinforced the need for a graphical sequence viewer that can interact with the database. GSDB staff has begun development of a web based, graphical sequence viewer that will be both platform independent and more flexible than Annotator.

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

Update process for Maestro

Maestro allows users to retrieve sequences from the database by querying on 18 different fields, including accession number, author, gene symbol and product name. The data can be represented in three different formats: GSDB flatfile, fasta or GIO (GSDB Input/Output format). To allow quick retrieval of sequences that match a given criterion, Maestro is implemented using query tables that are snapshots of the database from a single timepoint. We have recently implemented a procedure that updates these query tables once each week, to allow querying on all data available in GSDB.

During the next year we will be adding additional fields to the querying capability of Maestro. Suggestions regarding other fields that should be queryable are welcome.

Excerpt

The number of large (>100 000 bp) sequences in the database is growing rapidly. In part because of the number of complete bacterial and archaeal genomes that are being sequenced, but also because researchers are building larger contigs from human and model organism sequences. Many researchers are interested only in a relatively small segment of these large sequences, and are limited in the size of sequences they can utilize because of software or hardware restriction. Excerpt allows users to select a subset of any sequence within GSDB based on base pair span, gene or product name, all genes, all intergenic regions, or the sequence broken into equal sized spans. This provides researchers with easy access to all sequences, regardless of their computing power. Excerpt is available from the 'Data Retrieval' section of the GSDB web site (www.ncgr.org/gsdb).

GSDB flatfile definition

The GSDB staff has defined a flatfile format that represents the unique data types from GSDB. While this representation is not graphical, it allows all users to access discontiguous sequence data and alignments from GSDB, regardless of the computer platform they use. The new flatfile will show both the sequences that are aligned to the sequence that is displayed and the sequences to which it is aligned. In addition, the sequence accession number of any discontiguous sequence of which it is a piece of will be included. Flatfiles of discontiguous sequences will include the sequence accession numbers and descriptions of all of the sequences that are part of the discontig, and any information in the database on the order or distance from left of the sequence pieces.

Additional data types that are supported by GSDB, but not the IC databases that are included in the GSDB flatfile representation include sequence confidence data, user defined features, analysis information, and owner information. When GSDB flatfiles are viewed using the GSDB web pages, the flatfile will contain a hyperlink to the contact information for the owner of the data.

NEW ANALYSIS TOOLS

Our web site now includes a new tool for locating matrix attachment regions (MARs). The MAR sequences are 100–1000 bp long and anchor chromatin loops to the nuclear matrix. Adjacent MARs often delineate areas where transcription factor binding sites are concentrated, and can thus be used to identify areas within a sequence to search for coding potential.

This tool was developed by Singh *et al.* (14) and is available through the 'Software' section of the GSDB web site (www.ncgr.org/gsdb), and relies on the combination of a 'database' of known MAR sequences and a set of decision rules to determine if a MAR sequence is present in a sequence. MARs are identified based on the probability that the MAR motifs occur at random in a given window of the sequence being analyzed. Where known, predictions of MARs using this tool closely corresponded to their experimentally determined locations (14).

FUTURE DIRECTIONS

During the next year the GSDB staff will continue to make changes that will improve data quality and data access. We will continue to improve the unique data sets that we have created, and will create new unique data sets that will be useful to researchers throughout the biological community. To further improve data quality, we will implement biologically based data checks that will prevent bad data from being entered into GSDB.

With the tools and resources that we currently have available, GSDB is positioned to be a useful resource for computational and functional genomics. We will continue to create tools and data sets that will be useful for both computational questions and for functional questions.

CONTACT INFORMATION

GSDB can be contacted at: National Center for Genome Resources, 1800 Old Pecos Trail, Suite A, Santa Fe, NM 87505, USA. Tel: +1 505 982 7840 or +1 800 450 4854; Email: ncgr@ncgr.org or gsdb@ncgr.org; URL: http://www.ncgr.org

ACKNOWLEDGEMENTS

We would like to acknowledge S.A.Krawetz and J.A.Kramer of Wayne State University Medical School for allowing us to host the MAR-Finder software tool. We would also like to acknowledge the assistance of J.H.Horton in building the maize discontiguous sequences.

REFERENCES

1 Koonin, E.V. *et al.* (1997) *Curr. Biol.*, **6**, 404–416.
2 Burland, V. *et al.* (1995) *Nucleic Acids Res.*, **23**, 2105–2119.
3 Kunst, F. *et al.* (1995) *Microbiology*, **141**, 249–255.
4 Ogasawara, N. *et al.* (1995) *Microbiology*, **141**, 257–259.
5 Devine, K.M. *et al.* (1995) *Trends Biochem. Sci.*, **11**, 429–431.
6 Fraser, C.M. *et al.* (1995) *Science*, **270**, 397–403.
7 Koonin, E.V. *et al.* (1997) *Mol. Microbiol.*, **25**, 619–637.
8 Tanksley, S.D. and McCouch, S.R. (1997) *Science*, **277**, 1063–1066.
9 Maymo-Gatell, X. *et al.* (1997) *Science*, **276**, 1568–1571.
10 Worley, K.C. *et al.* (1995) *Genome Res.*, **5**, 13–184.
11 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
12 Uberbacher, E.C. *et al.* (1994) In Adams,M., Fields,C. and Venter,J. (eds) *Automated DNA Sequencing and Analysis*. Academic Press, London, pp. 307–312.
13 Smith, R.F. *et al.* (1996) *Genome Res.*, **6**, 454–462.
14 Singh, G.B. *et al.* (1997) *Nucleic Acids Res.*, **25**, 1419–1425.
15 Harger, C.A. *et al.* (1997) *Nucleic Acids Res.*, **25**, 18–23.
16 Bairoch, A. (1997) In Wilkins,M.R., Williams,K.L., Appel,R.D. and Hochstrasser,D.F. (eds) *Proteome Research: New Frontiers in Functional Genomics.* Springer-Verlag, New York, pp. 93–147.
17 Benson D.A. *et al.* (1996) *Nucleic Acids Res.*, **24**, 1–5. [See also this issue *Nucleic Acids Res.* (1998) **26**, 1–7].
18 Keen, G. *et al.* (1996) *Nucleic Acids Res.*, **24**, 13–16.
19 Kristensen, T. *et al.* (1992) *J. DNA Sequencing Mapping*, **2**, 343–346.
20 Bork, P. and Bairoch, A. (1996) *Trends Genet.*, **12**, 425–427.
21 Lopez, R. *et al.* (1992) *Nature*, **355**, 211.
22 Reynolds, T.L. (1994) *BioTechniques*, **16**, 1124–1125.
23 Rothberg, J.M. (1992) *Nature*, **356**, 738.
24 Savakis, C. and Doelz, R. (1993) *Science*, **259**, 1677–1678.
25 Lamperti, E.D. *et al.* (1992) *Nucleic Acids Res.*, **20**, 2741–2747.
26 Seluja, G.A. *et al.* in prep.
27 Blattner, F.R. *et al.* (1997) *Science*, **277**, 1453–1462.
28 Pearson, W.R. (1994) In Griffin,A.M. and Griffin,H.G. (eds) *Methods in Molecular Biology, vol. 24: Computer analysis of sequence data, part 1*. Humana Press, Totowa, NJ, pp. 307–331.