The Genome Sequence DataBase


National Center for Genome Resources, 1800 Old Pecos Trail, Suite A, Santa Fe, NM 87505, USA and
Celera Genomics, 45 West Guide Drive, Rockville, MD 20650, USA

Received October 7, 1999; Revised and Accepted October 13, 1999

ABSTRACT

The Genome Sequence DataBase (GSDB) is a database of publicly available nucleotide sequences and their associated biological and bibliographic information. Several notable changes have occurred in the past year: GSDB stopped accepting data submissions from researchers; ownership of data submitted to GSDB was transferred to GenBank; sequence analysis capabilities were expanded to include Smith–Waterman and Frame Search; and Sequence Viewer became available to Mac users. The content of GSDB remains up-to-date because publicly available data is acquired from the International Nucleotide Sequence Database Collaboration databases (IC) on a nightly basis. This allows GSDB to continue providing researchers with the ability to analyze, query and retrieve nucleotide sequences in the database. GSDB and its related tools are freely accessible from the URL: http://www.ncgr.org

INTRODUCTION

The Genome Sequence DataBase (GSDB), operated by the National Center for Genome Resources (NCGR), is a relational database of publicly available nucleotide sequences and associated biological and bibliographic annotation. The database contains both genomic and expressed nucleotide sequences from essentially all organisms for which some sequence data has been determined.

DATA ACQUISITION

The amount of nucleotide sequence data that is currently accessible in the public databases is approximately 5 million sequences consisting of approximately 4 billion nucleotides and it is doubling in volume about every 9–12 months. The amount of resources needed to adequately process, review and update this quantity of data is large. For the past several years, researchers have been provided with the ability to directly submit their sequence data to GSDB. However, with the termination of the Department of Energy grant in the fall of 1998, NCGR decided that the resources and staff efforts that had been dedicated to handling data submissions would be more beneficial if applied to other projects. Consequently, GSDB no longer accepts direct data submissions from researchers not associated with a specific collaborative project.

The content of GSDB remains up-to-date and reflects the content of the IC databases [which include DDBJ (1), EMBL (2) and GenBank (3)]. Publicly available sequence data are automatically imported nightly from the IC databases. During the past year improvements in exception handling have been made to the parsers which are responsible for incorporating IC data into GSDB. As a result significantly fewer parse errors now occur so that much less manual review of the data is needed. Upon review of the errors, data exceptions are manually incorporated into GSDB, while actual data errors are returned to the IC database of origin so that they can be corrected.

It is NCGR’s intention to continue updating GSDB in this manner so that data analysis and retrieval capabilities can be provided to researchers and so that GSDB can be utilized for internal and collaborative research projects at NCGR.

TRANSFER OF DATA OWNERSHIP

Since GSDB no longer handles data submissions and/or updates to existing sequence records, ownership of all of the data and associated revision histories were transferred to GenBank in the fall of 1999. This transfer provides researchers with continuity of service and allows them to continue to update database records as needed. In October 1999, attempts were made to contact all data submitters via postal and electronic mail to inform them of the data transfer. Special efforts were made to contact data submitters whose data was still being held in confidence (awaiting publication of the associated journal article) prior to the transfer of data. GSDB was unable to notify some data submitters whose contact information was outdated. Any questions concerning the transfer of data and their associated histories should be directed to gsdb@ncgr.org

CONTINUING DATA ACCESS

For the past several years, NCGR has provided a variety of tools to aid researchers in their efforts to identify and retrieve sequences of interest from GSDB. NCGR will continue to provide researchers with the ability to retrieve sequences from GSDB via web-based tools. These tools include a simple query tool, Maestro (http://www.ncgr.org/gsdb/maestro) (4), which allows users to search the database on terms such as organism name, gene name and keywords. Searches can be performed on single terms or a combination of terms by using the boolean

*To whom correspondence should be addressed. Tel: +1 505 982 7840; Fax +1 505 995 4432; Email: cah@ncgr.org
operators (‘and’, ‘or’ and ‘but not’) which are provided in the tool. Maestro is intuitive to use, however searching via this tool is limited to approximately 20 of the most commonly queried fields in the database. NCGR also provides a tool, Ad hoc query tool (http://www.ncgr.org/gsdb/adhoc ), which can be used to query any field in GSDB. This tool allows the user to use structured query language (SQL) to directly formulate and execute queries against GSDB. Documentation is provided to assist researchers in their use of SQL. Another tool, Excerpt (http://www.ncgr.org/gsdb/excerpt ), allows researchers to identify interesting regions of a sequence(s) in GSDB and retrieve only those regions, instead of the entire sequence(s). For example, the interesting regions could correspond to the protein coding or intergenic regions of a sequence.

In mid 1998 the set of tools associated with GSDB was expanded to include a java-based data visualization tool. Sequence Viewer (5) is a customizable tool that allows researchers to navigate and examine GSDB sequences and visually inspect their associated biological features. At the time it was originally made publicly available, the version of java used to develop the tool was not supported on the Macintosh platform. This has changed in the past year and Sequence Viewer (http://www.ncgr.org/gsdb.sv ) is now available to Macintosh users. Sequence Viewer and the other tools discussed above have been integrated with one another. As a result, researchers can easily move data from one tool to another with a simple ‘click of a button’. All of these tools can be accessed from NCGR’s home page at http://www.ncgr.org

SEQUENCE ANALYSIS

In 1998 Blast sequence similarity searching (http://seqsim.ncgr.org/newBlast.html ) was added to the repertoire of sequence analysis (5) and retrieval capabilities associated with GSDB. NCGR is in the process of making other similarity algorithms, including Smith–Waterman (6) and Frame Search, available to the public. Both algorithms are more sensitive than Blast (7,8) at identifying sequence similarities. These algorithms are implemented on a TimeLogic Decypher II system rated at 2.5 billion Smith–Waterman calculations per second (http://www.timelogic.com ).

Frame Search, which is based upon the Smith–Waterman algorithm, is a valuable tool to use with single-pass sequences, in that it can compensate for frame shifts in either the query sequence or the target database. This is accomplished by dynamically translating a nucleotide query sequence and switching between reading frames as needed to continue extending a match between the query sequence and a sequence in the target protein database.

In addition, Frame Search can be used to search a protein query sequence against a dynamically translated nucleotide database. As a consequence of the Smith–Waterman basis of this algorithm and the dynamic translation of the query sequence or target database, Frame Search is a computationally intensive algorithm. However, based upon internal testing and benchmarking of Frame Search on the Decypher system, we have found that it is comparable in speed to tblastx searches.

The amount of time necessary to complete a Frame Search search is primarily driven by the size of the database being searched and/or the amount of data being translated into protein sequence. SWISS-PROT, release 38.0 (9) contains ~2% the amount of data found in GSDB. Consequently searches against the entire nucleotide sequence database or against translations of the entire nucleotide sequence database are considerably slower than searches against protein databases. The web-based interfaces to these algorithms are currently undergoing beta-testing and they should be available to the general public soon.

NCGR believes that access to algorithms like Frame Search will become increasingly important to researchers as the amount of EST and other single-pass sequences in the public nucleotide databases continues to increase. NCGR is pleased to be able to provide this tool to the general research community and to collaborators as a way to increase the information content of sequence analysis projects.

In the coming year NCGR plans to continue using GSDB as its warehouse of nucleotide sequence data. Therefore importing data from the IC databases will continue, as well as sequence submissions from our direct collaborators. In addition, parsers will be improved as necessary to maintain the effortless import of data into GSDB. Public access to the data in GSDB and its associated tools will continue to be provided. This is to provide additional value to our collaborators, give maximal relevance to our internal projects, and as part of our public service mission to continue to provide high quality data and intuitive data mining tools to the public research community.

ACKNOWLEDGEMENT

This work was supported by the National Center for Genome Resources.

REFERENCES