

Genome Sequence DataBase
1800 Old Pecos Trail, Suite A
Santa Fe, NM 87505

Peter Schad, Vice-President
Bioinformatics and Biotechnology
505/995-4447, Fax: -4432
cnc@ncgr.org

Carol Harger
GSDB Manager
505/982-7840, Fax: -7690
cah@ncgr.org

http://www.ncgr.org

The National Center for Genome Resources (NCGR) is a not-for-profit organization created to design, develop, support, and deliver resources in support of public and private genome and genetic research. To accomplish these goals, NCGR is developing and publishing the Genome Sequence DataBase (GSDB) and the Genetics and Public Issues (GPI) program.

NCGR is a center to facilitate the flow of information and resources from genome projects into both public and private sectors. A broadly based board of governors provides direction and strategy for the center's development.

NCGR opened in Santa Fe in July 1994, with its initial bioinformatics work being developed through a cooperative 5-year agreement with the Department of Energy funded in July 1995. Committed to serving as a resource for all genomic research, the center works collaboratively with researchers and seeks input from users to ensure that tools and projects under development meet their needs.

Genome Sequence DataBase

GSDB is a relational database that contains nucleotide sequence data and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data and to furnish useful data-retrieval services.

GSDB adheres to the philosophy that the database is a "community-owned" resource that should be simple to update to reflect new discoveries about sequences. A corollary to this is GSDB's conviction that researchers know their areas of expertise much better than a database curator and, therefore, they should be given ownership and control over the data they submit to the database. The true role of the GSDB staff is to help researchers submit data to and retrieve data from the database.

GSDB Enhancements

During 1996, GSDB underwent a major renovation to support new data types and concepts that are important to genomic research. Tables within the database were restruc-

ured, and new tables and data fields were added. Some key additions to GSDB include the support of data ownership, sequence alignments, and discontinuous sequences.

The concept of data ownership is a cornerstone to the functioning of the new GSDB. Every piece of data (e.g., sequence or feature) within the database is owned by the submitting researcher, and changes can be made only by the data owner or GSDB staff. This implementation of data ownership provides GSDB with the ability to support community (third-party) annotation—the addition of annotation to a sequence by other community researchers.

A second enhancement of GSDB is the ability to store and represent sequence alignments. GSDB staff has been constructing alignments to several key sequences including the *env* and *pol* (reverse transcriptase) genes of the HIV genome, the complete chromosome VIII of *Saccharomyces cerevisiae*, and the complete genome of *Haemophilus influenzae*. These alignments are useful as possible sites of biological interest and for rapidly identifying differences between sequences.

A third key GSDB enhancement is the ability to represent known relationships of order and distance between separate individual pieces of sequence. These sets of sequences and their relative positions are grouped together as a single discontinuous sequence. Such a sequence may be as simple as two primers that define the ends of a sequence tagged site (STS), it may comprise all exons that are part of a single gene, or it may be as complex as the STS map for an entire chromosome.

GSDB staff has constructed discontinuous sequences for human chromosomes 1 through 22 and X that include markers from Massachusetts Institute of Technology–Whitehead Institute STS maps and from the Stanford Human Genome Center. The set of 2000 STS markers for chromosome X, which were mapped recently by Washington University at St. Louis, also have been added to chromosome X. About 50 genomic sequences have been added to the chromosome 22 map by determining their overlap with STS markers. Genomic sequences are being added to all the chromosomes as their overlap with the STS markers is determined. These discontinuous sequences can be retrieved easily and viewed via their sequence names using