Nucleic Acids Research, 1999, Vol. 27, No. 1 35–38

The Genome Sequence DataBase: towards an integrated functional genomics resource

M. P. Skupski, M. Booker, A. Farmer, M. Harpold, W. Huang, J. Inman, D. Kiphart, C. Kodira, S. Root, F. Schilkey, J. Schwertfeger, A. Siepel, D. Stamper, N. Thayer, R. Thompson, J. Wortman, J. J. Zhuang and C. Harger*

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

Received November 9, 1998; Revised and Accepted November 11, 1998

ABSTRACT

During 1998 the primary focus of the Genome Sequence DataBase (GSDB; http://www.ncgr.org/gsdb) located at the National Center for Genome Resources (NCGR) has been to improve data quality, improve data collections, and provide new methods and tools to access and analyze data. Data quality has been improved by extensive curation of certain data fields necessary for maintaining data collections and for using certain tools. Data quality has also been increased by improvements to the suite of programs that import data from the International Nucleotide Sequence Database Collaboration (IC). The Sequence Tag Alignment and Consensus Knowledgebase (STACK), a database of human expressed gene sequences developed by the South African National Bioinformatics Institute (SANBI), became available within the last year, allowing public access to this valuable resource of expressed sequences. Data access was improved by the addition of the Sequence Viewer, a platform-independent graphical viewer for GSDB sequence data. This tool has also been integrated with other searching and data retrieval tools. A BLAST homology search service was also made available, allowing researchers to search all of the data, including the unique data, that are available from GSDB. These improvements are designed to make GSDB more accessible to users, extend the rich searching capability already present in GSDB, and to facilitate the transition to an integrated system containing many different types of biological data.

INTRODUCTION

The genomic research community has embraced the idea of functional genomics, involving the integration and analysis of many different kinds of biological data at the level of complete genomes. In the past two years there have already been a number of studies examining functional genomics both by comparing complete genomes (1–7) and by examining the levels of expression of large numbers of coding regions in both yeast and human studies (8–15).

This emphasis on functional genomics has made it clear that the current paradigm of sequence databases is inadequate for the task of storing and linking all of the data that are necessary for functional genomic studies. Additional data types that are necessary include expression data, population polymorphism data and biochemical pathway data. NCGR believes that a system that incorporates all of these kinds of data will be necessary for biological discovery in the next five years, and that the most important requirement for such a system is that there be accurate links between the different types of data, so that a researcher can navigate from sequence data to polymorphism data to expression data to biochemical pathway data in any order and all combinations.

One of the most important components of such an integrated system is a high-quality sequence database. To establish links among the different data types, the quality of the sequence and annotation must be high. The concept of a gene has undergone radical change in the last 20 years, and the concept probably has limited utility in the current era of genomics (16). However, gene and product names can be associated with stretches of sequence that are known to produce a specific product, whether that stretch of sequence actually codes for an amino acid product, or has some regulatory role. NCGR believes that accurate annotation of gene and product names is essential in order to establish links with the other types of data necessary for an integrated system. Other data that must be accurate are taxonomy, chromosome, and map information, because they allow the establishment of meaningful links between different types of data.

Some of the organism specific databases have begun to integrate other kinds of data (17, http://genome-www.stanford.edu/Saccharomyces ; 18, http://WWW.informatics.jax.org). One of the major criticisms of these specialized databases is that they are not easily accessible to non-experts because they have been developed by and for experts in the speciality they serve (16). To avoid this problem and make an integrated system usable by researchers at all levels, the system must have visualization and analysis tools that are intuitive and easy to learn.

The Genome Sequence DataBase (GSDB) located at the National Center for Genome Resources (NCGR;

*To whom correspondence should be addressed. Tel: +1 505 982 7840; Fax: +1 505 995 4432; Email: cah@ncgr.org

http://www.ncgr.org) is a relational database that includes all of the data available from the databases that belong to the International Nucleotide Sequence Database Collaboration (IC). GSDB also provides support for data types that are not supported by the IC databases, including sequence analysis data, representations of pairwise alignments, and discontiguous sequences, which are groupings of individual sequences that have a known spatial relationship (e.g., all of the exons of a gene). This allows researchers to store meaningful relationships among sequences and make them available to others. NCGR has concentrated on using the strengths of the relational database system and the unique data types to provide better data access and analysis tools, interesting data collections, and more accurate data in certain fields.

In an effort to move toward an integrated system, GSDB has concentrated on data quality, new data collections, and access and analysis tools during the last year.

DATA QUALITY IMPROVEMENTS

Data curation

The GSDB staff has extensively curated the taxonomy of all organisms to ensure that the taxonomy followed within the database is consistent within each taxonomic group. The focus on taxonomy is driven first by a recognition that taxonomy will be very important when establishing links among different kinds of data that are incorporated into an integrated system. Second, new access and analysis tools that have been developed in the last year require that taxonomy be represented accurately within the database. Sequences that belong to new, unmatched species have their taxonomy updated within a day of entry into GSDB.

Additional efforts on curation are focused on the chromosome field for human sequences. This allows researchers working on human sequences to narrow the search if they know in which chromosome their sequence resides. It can also help place sequences if they can find significant homology to a sequence with a known location.

Data acquisition

Data from the IC databases are imported into GSDB each night, using a suite of special import programs to place each data element into its proper location. Over 99% of all data from the IC databases has been correctly imported into GSDB within 24–48 hours of its initial release by the database of origin. Improvements have been made to the import suite to ensure that the data from IC flatfiles is placed into the proper fields, and reports are generated nightly of fields frequently in need of curation so that they can be reviewed and corrected by the biology staff as necessary.

Biology staff members routinely survey the Internet to find sequence data not included in any of the other sequence databases (‘data mining’), and format these data for entry into GSDB using GIO (the GSDB Input/Output file format). Genome center web sites supported by the Department of Energy and the National Institutes of Health are a special focus. Appropriate credit is given to the center from which the data was mined, and links back to their website are included. About 4.5% of the data within GSDB are unique to GSDB, providing a rich source of information to researchers worldwide.

DATA COLLECTIONS

STACK

The Sequence Tag Alignment and Consensus Knowledgebase (STACK), a public database of human expressed gene sequences, was analyzed and developed by the South African National Bioinformatics Institute (SANBI; 19,20). These sequences consist of consensus sequences derived from aligned expressed sequence tags (ESTs) and clustered sequences (discontiguous sequences) that are part of a single gene. SANBI chose to deposit these data in GSDB for public access because GSDB is the only public sequence database with the ability to represent both alignment and discontiguous sequence data.

STACK data are extremely valuable, because the consensus sequences that were built from the analysis that SANBI performed are longer than the consensus sequences that have been made by other groups and are longer than the individual ESTs that were used to construct the consensi. The increase in length makes identification of individual genes much more likely, and improves the ability of a researcher with a sequence that matches one of the ESTs to find a match.

STACK data can be accessed and retrieved using a special version of Maestro, a database query tool, developed specifically for retrieving these collaborative data, SANBI Maestro (http://www.ncgr.org/cgi-bin/sanbi/maestro/front.pl).

Human sequence-based maps

The human discontiguous sequences constructed using the sequence tagged sites (STS) markers from the Whitehead Institute for Biomedical Research and Stanford Human Genome Center (21) have been expanded to include genomic sequences. Genomic sequences are placed on these maps by determining homology with sequences already in place. On average, 60 genomic sequences have been placed on each chromosome, and newly acquired data are checked regularly to locate new matches and to place them on each chromosome.

Accession numbers of these maps can be obtained from the ‘What’s New’ section of the GSDB web site (http://www.ncgr.org/gsdb), and the discontiguous sequences can be retrieved in flatfile format using Maestro (http://www.ncgr.org/gsdb/maestro/Index.html). The flatfile format of discontiguous sequences includes all sequences that are part of the discontig and any information about order and distance that is known among the sequences.

DATA ACCESS AND ANALYSIS IMPROVEMENTS

Sequence Viewer

Sequence Viewer is a new platform-independent graphical viewer for sequence data, implemented in Java as an Internet-based applet that can be launched from a web browser or used as a stand-alone application with increased functionality. The applet version of Sequence Viewer is integrated with Maestro, the web-based search tool for GSDB, and sequences can be viewed directly from Maestro results. The application version of Sequence Viewer can be downloaded from the Sequence Viewer web site (http://www.ncgr.org/gsdb/sv).

Specific advantages of the Sequence Viewer include: easy visualization of the features of a sequence including placement of features, length, type and name; easy access to feature-associated

sequence due to integration with Excerpt, GSDB's sub-sequence retrieval script; simple identification of STS markers through display as vertical bars; simple text display of additional information about features and sequences, such as reference data or gene and product information; clear indication of the strand on which a feature occurs; easy color customization of feature types; clear viewing of large sequences that contain extensive annotation on any platform.

Sequence Viewer is intended as the first module of a tool that will eventually allow viewing and editing of sequences, multiple alignments and discontiguous sequences.

Table 1. Summary of the organization of the searchable sets into finer subsets to allow more precision in searching

Data set	Subsets
Archaea	
Eubacteria	
Virus	RNA viruses
	DNA viruses
	Unclassified viruses
Chordates	*Fugu* (puffer fish)
	All other chordates (excludes mammals)
Mammals	Rodent
	Mouse
	Rat
	Other rodents
	Non-human primates
	All other mammals
Human	Individual chromosomes
	ESTs
	STSs
	unmapped sequences
Non-chordate animals	*Caenorhabditis elegans*
	Drosophila melanogaster
	All other non-chordates
Plants	Eudicots
	Arabidopsis thaliana
	Other dicots
	Monocots
	Non-flowering vascular plants
	Other plants
Fungi	*Saccharomyces cerevisiae*
	All other fungi
Miscellaneous eukaryotes	All eukaryotes that are not in the kingdoms Animalia, Plantae, or Fungi
Organelles	Human mitochondrion
	Plant mitochondrion
	Chordate mitochondrion
	Chloroplast
	Non-vertebrate, non-plant mitochondrion
	Other organelles
Non-human ESTs	
Non-human STSs	
Synthetic sequences	
Unidentified sequences	Sequences of uncertain origin
Complete genomes	Each microbial complete genome

Sequence similarity searching

A TimeLogic DeCypher II was recently acquired to provide a server for homology searching using several algorithms. Currently BLAST is available to use in searching nucleotide sequences from GSDB and protein sequences from several protein databases (http://seqsim.ncgr.org). GSDB contains sequence data that is not available elsewhere. This unique data can be searched using the BLAST similarity search.

To allow more flexibility in homology searching, we provide searchable subsets, primarily broken down along taxonomic lines so that researchers can choose to search only those sequences that are relevant (Table 1). The domains of Archaea and Eubacteria are available as separate search sets. Within eukaryotes, each kingdom is available as a search set, with many of the kingdoms split into smaller subsets to allow maximum flexibility in choosing a search set. Additional search sets that are broken out include STS, EST and STACK data, as well as individual human chromosomes.

Users can choose a single data set, or can choose multiple data sets from the subsets provided. This allows researchers to search a smaller set of sequences if they are trying to determine if their sequence is already in the database for the species from which it was obtained. An advantage to researchers working on humans is that they can search on individual chromosomes to determine if their sequence occurs in a larger contig that has been sequenced or to help place the sequence along the chromosome.

To support the breakdown of the data into these search sets, the GSDB staff have been extensively reviewing and curating data fields used to identify subsets, including taxonomy and chromosome data. Ensuring accuracy of these fields means that the search subsets will contain all data relevant to the researcher's needs.

The DeCypher includes accelerated versions of Smith–Waterman, Frame Search, Symmetric Frame Independent and Profile Search that will become publicly available for searching data from GSDB during the first half of 1999.

Integration of tools

Another focus during the past year has been on integration among the tools that are used for accessing data from GSDB. For example, a search done using Maestro, the web-based query tool that searches GSDB based on certain fields, will return options to retrieve a sequence in flatfile format; using Excerpt, a script that retrieves only a portion of a sequence; and Sequence Viewer. Also, Sequence Viewer has options that allow the retrieval of the sequence flatfile, and of a portion of the sequence using Excerpt.

FUTURE DIRECTIONS

During the next year GSDB will become a component of an integrated biological knowledge system that is designed to house additional types of biological data, including expression data, biochemical pathway data, and population polymorphism data. As part of this transition, NCGR intends to offer continued public access to GSDB. NCGR will continue to create new tools and databases that are useful to the biological research community.

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

This work was funded by Cooperative Agreement DE-FC03-95ER62062 with the Office of Biological and Environmental Research of the Department of Energy.

REFERENCES

1 Koonin,E.V. and Galperin,M.Y. (1997) *Curr. Opin. Genet. Dev.*, **7**, 757–763.
2 Mushegian,A.R. and Koonin,E.V. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
3 Koonin,E.V., Mushegian,A.R. and Rudd,K.E. (1996) *Curr. Biol.*, **6**, 404–416.
4 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K. *et al.* (1997) *J. Bacteriol.*, **179**, 7135–7155.
5 Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) *Curr. Biol.*, **6**, 279–291.
6 de Rosa,R. and Labedan,B. (1998) *Mol. Biol. Evol.*, **15**, 17–27.
7 Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) *Nucleic Acids Res.*, **25**, 701–712.
8 DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) *Science*, **278**, 680–686.
9 Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 13057–13062.
10 DeRisi,J., Penland,L., Brown,P.O., Bittner,M.L., Meltzer,P.S., Ray,M., Chen,Y., Su,Y.A. and Trent,J.M. (1996) *Nature Genet.*, **14**, 457–460.
11 Schena.,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.
12 Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Vogelstein,B. and Kinzler,K.W. (1997) *Science*, **276**, 1268–1272.
13 Heller,R.A., Schena,M., Chai,A., Shalon,D., Bedilion,T., Gilmore,J., Woolley,D.E. and Davis,R.W. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 2150–2155.
14 Wodicka,L., Dong,H., Mittmann,M., Ho,M.H. and Lockhart,D.J. (1997) *Nature Biotechnol.*, **15**, 1359–1367.
15 Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) *Science*, **270**, 467–470.
16 Gelbart,W.M. (1998) *Science*, **282**, 659–661.
17 Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) *Nucleic Acids Res.*, **26**, 73–80.
18 Blake,J.A., Eppig,J.T., Richardson,J.E. and Davisson,M.T. (1998) *Nucleic Acids Res.*, **26**, 130–137.
19 Hide,W.,Burke,J., Christoffels,A. and Miller,R. (1997). In Miyano,S. and Takagi,T. (eds.) *Genome Informatics*. Universal Academy Press Inc. Tokyo, Japan, pp. 187–196.
20 Burke,J., Wang,H., Hide,W. and Davison,D. (1998) *Genome Res.*, **8**, 276–290.
21 Harger,C., Skupski,M., Bingham,J., Farmer,A., Hoisie,S., Hraber,P., Kiphart,D., Krakowski,L., McLeod,M., Schwertfeger,J. *et al.* (1998) *Nucleic Acids Res.*, **26**, 21–26.