Nucleic Acids Research, Vol. 20, Supplement **2065–2069**

GenBank

Christian Burks, Michael J.Cinkosky, William M.Fischer, Paul Gilna*, Jamie E.-D.Hayden, Gifford M.Keen, Michael Kelly[1], David Kristofferson[1] and Julie Lawrence[1]
Theoretical Biology and Biophysics Group, T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545 and [1]IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, CA 94040, USA

ABSTRACT

The GenBank nucleotide sequence database now contains sequence data and associated annotation corresponding to 85,000,000 nucleotides in 67,000 entries from a total of 3,000 organisms. The input stream of data coming into the database is primarily as direct submissions from the scientific community on electronic media, with little or no data being keyboarded from the printed page by the databank staff. The data are maintained in a relational database management system and are made available in flatfile form through on-line access, and through various network and off-line computer-readable media. The data are also distributed in relational form through satellite copies at a number of institutions in the U.S. and elsewhere. In addition, GenBank provides the U.S. distribution center for the BIOSCI electronic bulletin board service.

INTRODUCTION

GenBank,[1] the **Gen**etic Sequence Data **Bank**, provides the scientific community with a computer database of all published (and, increasingly often, unpublished) DNA and RNA sequences as well as related bibliographic and biological information that establish the physical, functional, and administrative context of the sequence data.

The project is funded through an NIGMS contract with IntelliGenetics, Inc. (IG) which, in turn, contracts with the DOE acting on behalf of Los Alamos National Laboratory (LANL). The project is funded with co-sponsorship from other Institutes of the NIH, NLM, DRR, USDA, NSF, DOE, and DOD. Data collection and distribution are carried out in collaboration with the EMBL Data Library (1) and the DNA Data Bank of Japan (2).

We present a summarized overview of the GenBank database (and project), including submission of data to the database, mechanisms for maintaining the data, and the several means for distribution of the data to the scientific community. The historical origins of the project have been discussed previously (3). We focus here especially on an update of developments since recent, detailed descriptions of the database (4,5).

SYSTEMS

Inquiries about acquiring access to the data should be addressed to IG;[2] inquiries about submitting new data or revising data already in the database should be addressed to LANL.[3] Electronic mail addresses specific to the kind of query are provided in the sections below.

The GenBank On-line Service at IG utilizes a four processor (4×22 MIPS) Solbourne 5/800 running OS/MP 4.1A.1 (similar to SunOS UNIX 4.1.1); at LANL, the database work is done on a network of Sun Microsystem (Mountain View, CA) 4/690 server and assorted workstation clients running under SunOS UNIX version 4.1.2. The GenBank database is maintained at both sites under the Sybase (Emeryville, CA) relational database management system (RDBMS). Software was developed in the 'C' language.

All statistics presented here for the database correspond to the on-line version of the RDBMS form of GenBank as of March 13, 1992.

DATABASE ACCESS

The data in GenBank are available on several media and through several different electronically-based mechanisms, described below.

On-line system

The GenBank On-line Service (GOS) offers access to the latest GenBank, EMBL, and GenPept (protein translation of GenBank) data. The data are updated daily and are available by direct login to the system and by e-mail server. E-mail services include sequence retrieval from GenBank, EMBL, GenPept, and Swiss-Prot (send a message containing the word, 'HELP' to

* To whom correspondence should be addressed

[1]'GenBank' is a registered trademark of the U.S. Dept. of Health and Human Services.

[2]Genbank; IntelliGenetics, Inc.; 700 East El Camino Real; Mountain View, CA 94040; U.S.A. (e-mail, genbank@genbank.bio.net; fax 415–962–7302; telephone 415–962–7364).

[3]GenBank; Theoretical Biology and Biophysics Group; T-10, MS K710; Los Alamos National Laboratory; Los Alamos, NM 87545; U.S.A. (e-mail, genbank@life.lanl.gov; fax 505–665–3493; telephone 505–665–2177).

retrieve@genbank.bio.net for details). FASTA and BLAST similarity searches are also available by e-mail (send 'HELP' to search@genbank.bio.net or to blast@genbank.bio.net). GOS accounts offer access to a wide variety of sequence analysis programs and electronic communications facilities described further below. GOS has been described in greater detail elsewhere (6).

Off-line distribution

GenBank data can be obtained from IG on CDROM, 9 track tape, TK-50 and Sun 1/4″ cartridges. As of March 1992, copies of the release on 1.2 MB PC-AT floppy diskettes and 800 kbyte Macintosh floppy diskettes are no longer being distributed.

Hard-copy distribution

The last printed distribution of the database was in 1987 (7); this eight-volume publication was undertaken in collaboration with the EMBL Data Library (1) and corresponded to GenBank Rel. 44.0 (August 1986) and EMBL Rel. 8.0 (May 1986). The distribution included 8823 entries representing 8,442,357 bases, or approximately 10% of the current database.

UPDATE OVERVIEW

We describe recent developments in the mechanisms supporting the flow of data into the database and out to the community.

New features table

Rel. 65.0 (September 1990) was distributed in the new features table, a computer-parsable syntax (developed jointly by DDBJ, EMBL, and GenBank) which expresses the complex biological features associated with nucleotide sequences. The different features table formats and annotation standards adopted by the EMBL Data Library and GenBank at their outsets created significant difficulties for both the data banks' data sharing efforts and the user community. One of the goals of the new features table was to set out a common format and facilitate common annotation standards. Documentation for the new features table format and content can be retrieved either through anonymous FTP from genbank.bio.net [134.172.1.160] in pub/doc, or by sending an e-mail message to bioserve@genome.lanl.gov containing the word 'gb-feature'.

New features

One of the more significant additions to the feature table during the course of the last year was the inclusion of a feature qualifier to the CDS feature key, which reports the translated amino acid sequence. The CDS feature key presents an *instruction set* that allows the determination of the conceptual amino acid translation by software designed to parse the feature table. The CDS key (along with other keys) is capable of joining spans of sequences from multiple entries (e.g., seperate exon sequences) to assemble the final translated product.

However, not many software tools currently take advantage of the richness of these data, and users must frequently resort to manual interpretation of the CDS feature key to generate the corresponding translation. To accommodate this deficiency, the conceptual translation derived from the instructions in the CDS feature key are now being presented in the flatfile distribution of the database in a **/translation=** qualifier, where the sequence data are presented as a string of one letter amino acids.

Direct submission of data

The importance of the direct data submission paradigm has grown significantly over the past years. The majority of journals presenting reports of sequenced entities have changed their editorial policies with respect to the amount of sequence data that will appear in an article reporting those data. Editorial policies frequently decree that only portions of the sequence data directly relevant to the paper may appear in a figure. This trend suggests that the presence of sequence data in a paper will, over time, be restricted to the use of such figures for illustrative purposes only, and not to report the data. Perhaps the most significant implication of this trend is that the conventional scientific journal will no longer be the primary forum for reporting of sequence data; rather, the community will turn to the databases both as the alternative source for the data and as a new public forum for data dissemination. We have termed this new paradigm 'Electronic Data Publishing' (9).

As described in earlier reports (4,5,10,11), GenBank, in collaboration with EMBL (1) and DDBJ (2), has been working with the editorial staff of the scientific journals and with the scientific community to encourage direct submission of data to the databanks. This collaboration has met with great success. At this time, the GenBank project receives approximately 95% of the data it collects as direct submissions from the community. Over 98% of our submissions are currently coming to us in electronic form. An increasing proportion (currently about 75%) of electronic submissions come in through electronic mail. Direct submissions, and inquiries regarding them, should be directed to: **gb-sub@genome.lanl.gov**

Receiving 98% of our data ahead of publication has had a significant positive impact on the completeness and timeliness of the database, as well as on the quality of the data which ultimately reaches the community.

Authorin

Versions of the Authorin program (12) on both the IBM PC and the Macintosh are currently available from IG. The Authorin program helps scientists annotate their data and outputs them in the proper format for rapid inclusion into the GenBank RDBMS. It helps ensure the correctness of the data, controls the vocabulary used in annotating the data, and reduces the time needed to release the data to the scientific community. The program is available free of charge from IG (requests may be sent to authorin@genbank.bio.net, and should specify whether the PC or the MAC version is desired).

That Authorin has now become an established component of the data submission mechanism is best illustrated by the fact at at least 65% of submissions now use this tool, compared to only 15% one year ago.

A modification of the PC version of Authorin (called 'PatentIn') for submitting patent applications containing sequence data was produced by IG and released to the U.S. Patent Office in November 1990. Copies of PatentIn are available from the U.S. Patent Office.

Curator program

The GenBank Curator program, outlined previously (4,5), is now formally under way; the following scientists have begun work, and other proposals are now being brought into operation:

Dr. R. Jones began working with us last year on a number of software modules, in part drawing on the power of highly-

parallel hardware architectures for sequence database searches (13), that would routinely check incoming sequences for vector sequence contamination. Although we have verified and, where appropriate, corrected vector contamination in the past on a case-by-case basis as it was brought to our attention, this work was designed to allow us both to purge the entire existing database of such contamination and to routinely, automatically screen incoming data for similar problems. Today, that purge has largely taken place and such routine checks are now in place, and authors are being informed of the presence of such possible contaminant sequences when discovered by this software. While they are being resolved, their presence in the database is flagged by a */misc feature* in the features table.

There has been renewed interest quite recently in developing extensive (and even comprehensive) gene expression 'maps' of genomes by directed cDNA-based sequencing, and extending the notion of the STS strategy (8) to Expressed Sequence Tags (14), or 'EST's'. Last year Dr. A. Kerlavage began working with us to to examine the special constraints placed on either direct data entry protocols or annotation descriptors by this (and similar) large-scale sequencing projects. As a result of this work EST data from this project are now submitted in transaction form to the project where they are read directly into and out from the database in a matter of minutes; recently 2303 such EST sequences were made available in the public on-line systems on the same day as their publication (15). Data from other cDNA projects have also been submitted to the project by similar mechanisms.

Several curators have been provided with online access and special permissions to conduct their work on the database by remote links over the internet computer network. Their work consists of annotation correction and addition and is focused on specific areas of the database related to their expertise. Amongst the topics being covered are:

Correction and refinement of *S. pombe* and *S. cerevisae* annotation and nomenclature; Dr. John Hill, NYU Medical Center.

Added value annotation and correction of members of the transmembrane helix superfamily of genes, including addition of previously missing sequence data; Dr. Kevin Lynch, University of Virginia.

Added value annotation and correction of herpesvirus sequence entries, with particular emphasis on strain identification and sequence overlap annotation; Dr. Hal B. Jenson, U. Texas, San Antonio.

Annotation and correction of heat shock sequence entries, including resolution of redundancy; Dr Lawrence Moran, University of Toronto.

Turn-around time

The primary goal of the direct data submissions program is to enable availability of sequence data in retrievable, electronic form at the point of conventional journal publication of the data.

Five years ago, when the rate of data production in the scientific community was approximately 2 million nucleotides per year, the average lag time between the appearance of sequence data in a published article and the subsequent appearance of those same data in a public release of the database was about thirteen months. In 1991, the databank staff handled more than 20 million base pairs, yet the average lag time had been reduced to two weeks.

The report on GenBank progress presented in the 1991 version of this supplement (5) reported the presence of 45,000,000 nucleotides in the database as of February, 1991, whereas in this report it is twice that; the doubling time of the database is about one year. Thus, in 1992, GenBank will have processed at least twice the amount of data as in the previous year.

Today, our data processing procedures are designed to issue an accession number or set of numbers to submitting authors within 24–48 hours of receipt of the data, regardless of the medium. Once received, the data are processed to completion and released to the online servers consistently within a further ten days. To ensure the continuation of that consistency of service to the community, we have established a number of internal production controls that are designed to ensure that we process data at a rate that is continually matched to the volume of incoming data; since our resources do not keep pace with the increases in data, we must instead double our efficiency each year to enable us to deal with a doubling of the data.

Perhaps the greatest impact of the data submissions policies lies in the area of data quality and integrity. For most scientific journals, the appearance of a manuscript implies that the content of the article has been reviewed by a panel of the authors' peers. There is a common misconception, however, that the scientific data (in this case, sequence data) have traditionally undergone a similar degree of review, and that the allowance for unpublished data would therefore invite sequence data of significantly lower quality into the database (16). This is clearly not the case. Sequence data *per se* are not reviewed by the journal editorial staff; it is not the task of the editorial review process to review the data, but rather the experimental methodology and the scientific conclusions obtained as a result.

However, the fact that GenBank receives data prior to publication allows the databank to take on the task of verifying data integrity. All sequences entering the database are subjected to a growing number of automated checks as they proceed through the annotation process. Examples of checks which are currently applied include verification of coding regions, verification of intron/exon splice junctions, and examination for the presence of common vector sequences (see the discussion of the curator program above). These (and similar) checks are incorporated into a Sequence Validation Suite, which is a subset of the DIL described in (5). Errors uncovered in the sequence data can be passed back to the author in time to have the data corrected for publication.

In addition, all submissions are routinely passed back to the author for review at the time of annotation and release, and many respond with updates and corrections, both at that point, and frequently at points beyond the publication of their data.

Confidential data

Though much of our data are available to the public before publication, a considerable portion of the data we receive is submitted with the proviso that they be withheld from release until they appear in a publication. Ordinarily, we make the link from publication to submitted data through our normal journal scanning process, and quickly release the data so that they will have appeared in the database within two weeks of publication. However, we are still likely to miss published data by this mechanism. It is no longer the case that 80–90% of sequences appear in a few major journals, hence the problem of having to manually scan journals to determine publication of submitted, confidential data is compounded by there being a greater diversity of journals in which the data may ultimately appear.

We have begun addressing this problem in a number of ways, including examination of advance releases of journals' tables of contents. Presently, if an accession number which appears in a journal article does not return a sequence from the on-line servers (thereby inferring that we are still holding these data confidential), users are asked to inform GenBank through e-mail at update@life.lanl.gov. However, these *after the fact* measures will only present the databanks with an increased amount of labour with an increasingly diminished return. Further, it is unlikely that we will continue to scan journals or perform any data entry in the years to come. Instead, we are moving to place the responsibility for data release directly with the submitting author by encouraging owners of confidential data to release their data in advance of publication. Authors are reminded of this responsibility at the time of submission and periodically during the course of the holding of these data.

Anecdotal evidence exists to suggest that this shift in responsibility is already taking effect as the number of authors who inform us of the impending publication of their data (or who request their release regardless of publication status) has increased significantly over the course of the last year.

The direct submission of data (and the tools developed to assist authors in this endeavour) have enabled the database to impart significant increases to both the turnaround and the quality of the data in the database. As procedures continue to improve to the point where we are meeting our goal with all sequences, the sequence databases will become an even more integral component of the scientific publication process.

Satellite installations of GenBank

The relational version of GenBank as maintained at Los Alamos, and described previously (5, 9) is available for installation at remote sites in satellite form. These satellite copies are automatically updated nightly using software developed at Los Alamos. This allows sites around the world to access the most current copy of all data.

The following sites currently maintain satellite copies of the database:

- IntelliGenetics (Mountain View, CA)
- Lawrence Berkeley Laboratory (Berkeley, CA)
- The Australian National Genome Information Service (ANGIS, Sydney, Australia)
- Walter and Eliza Hall Institute (Melbourne, Australia)
- Applied Math Group, DSIR (Wellington, New Zealand)
- University of Texas High Peforrnance Computing Center (Austin, TX)
- Cold Spring Harbor Laboratory (Cold Spring Harbor, NY)
- DNA Data Bank of Japan (DDBJ, Mishima, Japan)
- Biomedical Centre, Uppsala University (Uppsala, Sweden)
- Johns Hopkins University (Genome DataBase, Baltimore, MD)

Currently, the satellite software is available for Sun Sparc and compatible computer systems. The software requires the use of the Sybase relational database management system (RDBMS). The source code is available for porting to other computer architectures and operating systems. It has been ported to, and is running on Digital Corporation VAX systems running Ultrix.

Installation of new satellite sites is handled semi-automatically. The installation program, the program for automatic updates of the satellite and the latest dump of the relational database can be retrieved from Los Alamos by anonymous ftp to genome.lanl.gov, [128.165.24.151]. The file named 'requirements' includes full instructions for configuring Sybase. Installation, including transfer and loading of a copy of the current database, is handled automatically by the installation program.

BIOSCI Bulletin board system

GOS is the distribution center in the Americas for the BIOSCI electronic newsgroup service. Currently BIOSCI consists of 26 newsgroups on a variety of topics of interest to biological scientists, including one used for discussing issues related to GenBank. BIOSCI bulletins are distributed around the world, and scientists with e-mail access can receive the newsgroups free of charge by sending a request to biosci@genbank.bio.net.

Those with access to USENET news do not need e-mail subscriptions and can participate by reading the 'bionet.*' newsgroups on USENET (consult your local computer systems manager for details). The USENET newsgroup bionet.molbio.genbank.updates distributes the latest GenBank data, and software (17) for extracting the data automatically from the newsgroup on VAX and UNIX systems (for more information, contact smith@mcclb0.med.nyu.edu or roy@alanine.phri.nyu.edu). This software allows new GenBank entries to be distributed as soon as the sequences are available over existing networks, using existing Usenet software and infrastructure.

SUMMARY

It is clear that nucleotide sequence data will continue growing exponentially, as was predicted several years ago (18). Computer technology continues to advance, and there is an increasing availability of computer hardware in molecular biology laboratories as well as network links between them. Sequencing technology is advancing to the point where the task of sequencing DNA has become routine. These factors are placing significant burdens on the traditional methods of dissemination of sequence data to the community, the scientific journals, to the point where alternative pathways must be found if these data are to continue to appear in the public domain. The systems and procedures designed by GenBank, collectively termed 'Electronic Data Publishing'(9), provide an alternative pathway for the publication of biological sequence data.

The developments we described above place GenBank in an excellent position to handle the growth and further interpretation of nucleotide sequence data into the near future and beyond; but, as in the past, we will rely heavily on input from the scientific community allowing us both to anticipate new kinds of data and to provide the data in the most useful forms.

ACKNOWLEDGEMENTS

We are grateful to past and present GenBank staff, our many collaborators, and advisors (formal and informal) for their ongoing contributions to the GenBank Project. This work was funded by a contract from the NIH (NO1-GM-7−2110), and the work at LANL was also supported under the auspices of the U.S. Dept of Energy.

REFERENCES

1. Kahn,P. and Cameron,G. (1990) EMBL Data Library. *Meth. Enz.* **183**, 23–31.
2. Miyazawa,S. (1990) DNA DataBank of Japan: Present status and future plans. In *Computers and DNA*, G.I.Bell and T.Marr, Ed., Addison-Wesley, Reading, MA, 47–61.
3. Burks,C., Fickett,J.W., Goad,W.B., Kanehisa,M., Lewitter,F.I., Rindone,W.P., Swindell,C.D. and Tung,C.-S. (1985) The GenBank nucleic acid sequence database. *Comp. Applic. Biosci.* **1**, 225–233.
4. Burks,C., Cinkosky,M.J., Gilna,P., Hayden,J.E.-D., Abe,Y., Atencio,E.J., Barnhouse,S., Benton,D., Buenafe,C.A., Cumella,K.E., Davison,D.B., Emmert,D.B., Faulkner,M.J., Fickett,J.W., Fischer,W.M., Good,M., Home,D.A., Houghton,F.K., Kelkar,P.M., Kelley T.A., Kelley M., King,M.A., Langan,B.J., Lauer,J.T., Lopez,N., Lynch,C., Lynch,J., Marchi,J.B., Marr,T.G., Martinez,F.A., McLeod,M.J., Medvick,P.A., Mishra,S.K., Moore,J., Munk,C.A., Mondragon,S.M., Nasseri,K.K., Nelson,D., Nelson,W., Nguyen,T., Reiss,G., Rice,J., Ryals,J., Salazar,M.D., Stelts,S.R., Trujillo,B.L., Tomlinson,L.J., Weiner,M.G., Welch,F.J., Wiig,S.E., Yudin,K. and Zins,L.B. (1990) GenBank: Current status and future directions. *Meth. Enzymol.* **183**, 3–22.
5. Burks,C., Cassidy,M., Cinkosky,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.-D., Keen,G.M.,Kelley,T.A., Kelly,M., Kristofferson,D. and Ryals,J. (1991) GenBank. *Nucleic Acids Res.* **19 Suppl**, 2221–2225
6. Benton,D. (1990) Recent changes in the GenBank On-line Service. *Nucleic Acids Res.* **18**, 1517–1520.
7. Atencio,E.J., Bilofsky,H.S., Bossinger,J., Burks,C., Cameron,G.N., Cinkosky,M.J., England,C.E., Esekogwu,V.I., Fickett,J.W., Foley,B.T., Goad,W.B., Hamm,G.H., Hazledine,D.J., Kahn,P., Kay,L., Lewitter,F.I., Lopez,N., MacInnes,K.A., McLeod,M.J., Melone,D.L., Myers,G., Nelson,D., Nial,J.L., Norman,J.K., Rasmussen,E.D., Revels,A.A., Rindone,W.P., Schermer,C.R., Smith,M.T., Stoesser,G., Swindell,C.D., Trujillo,B.L., and Tung,C.-S. (1987) *Nucleotide Sequences 1986/1987: A Compilation from the GenBank and EMBL Data Libraries*. Academic Press, Orlando, FL (published as Volumes I–VIII).
8. Olson,M., Hood,L., Cantor,C. and Botstein,D. (1989) A common language for physical mapping of the human genome. *Science* **245**, 1434–1435.
9. Cinkosky,M.J., Fickett,J.W., Gilna,P. and Burks,C. (1991) Electronic data publishing and GenBank. *Science* **252**, 1273–1277.
10. Burks,C. and Tomlinson,L.J. (1989) Submission of data to GenBank. *Proc. Natl. Acad. Sci.* USA **86**, 408.
11. Gilna,P., Tomlinson,L.J. and Burks,C. (1989) Submission of nucleotide sequence data to GenBank. *J. Gen. Microbiol.* **135**, 1779–1786.
12. Moore,J.F., Benton,D. and Burks,C. (1989) The GenBank nucleic acid data bank. *BRL Focus* **11(4)**, 69–72.
13. Jones,R., Taylor,W.R., Zhang,X., Mesirov,J.P., and Lander,E. (1989) Protein sequence comparison on the Connection Machine CM-2. In *Computers and DNA*. G.I. Bell and T. Marr, Eds., Addison-Wesley, Reading, MA, pp. 1–9.
14. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Olde,B., Moreno,R.F., Kerlavage,A.R., McCombie,W.R. and Venter,J.C. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651–1656.
15. Adams,M.D., Dubnick,M., Kerlavage,A.R., Moreno,R.F., Kelley,J.M., Utterback,T.R., Fields,C. and Venter,J.C. (1992) Sequence identification of 2375 human brain genes. *Nature* **355** 632–634.
16. Burks,C. (1989) Sources of data in the GenBank database. In *Biomolecular Data: A Resource in Transition*, R.R.Colwell, Ed., Oxford University Press, England, pp. 327–334.
17. Smith,R.H., Gottesman,S., Hobbs,B., Lear,E., Kristofferson,D., Benton,D. and Smith,P.R. (1991) A mechanism for maintaining an up-to-date GenBank database via Usenet. *Comp. Applic. Biosci.* **7**, 111–112.
18. Burks,C. (1989) How much sequence data will the data banks be processing in the near future? In *Biomolecular Data: A Resource in Transition*, R.R. Colwell, Ed., Oxford University Press, England, pp. 17–26.