Submission of Nucleotide Sequence Data to GenBank®

By PAUL GILNA,* LAURIE J. TOMLINSON AND CHRISTIAN BURKS

Theoretical Biology and Biophysics Group, T-10, Mail Stop K710, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Introduction

GenBank, the *genetic sequence data bank*, is chartered to provide a computer database of all published (and, increasingly, unpublished) DNA and RNA sequences and related bibliographic and biological information. Under the terms of this charter GenBank is intended to provide:

- an archive for long-term storage and preservation of sequence data and retrieval of nucleotide sequences, with delineation by sequence-specific, bibliographic, physical or functional criteria;
- a research platform, for examination and analysis of sequences grouped by sequence-specific or annotation-specific criteria.

The project is funded through a US National Institutes of General Medical Sciences contract with IntelliGenetics, Inc., which in turn contracts with the US Department of Energy acting on behalf of Los Alamos National Laboratory. Data collection and distribution are carried out in collaboration with the European Molecular Biology Laboratory (EMBL) Data Library and the DNA Data Bank of Japan (DDBJ). The database is updated and released quarterly. Information on how to access the data in GenBank is provided in the Appendix.

The Editors of the *Journal of General Microbiology (JGM)* have adopted the policy that papers reporting or analysing original nucleotide or amino acid sequence data will not be published unless the sequences reported therein have been deposited with GenBank. *JGM* and GenBank have established procedures to ensure the smooth and timely execution of this policy, imposing minimal delay on the publication process while considerably shortening the time it takes for such information to become publicly available upon publication. This article presents the history, need for, benefits, and mechanics of this policy.

Why a new scheme?

Several years ago, the data bank staffs and the user community began to be quite concerned about the timeliness and completeness of new nucleotide sequence data appearing in the data banks. The scientific community at large came to realize that a limited data bank staff cannot keep up with the ever-increasing flood of sequence data, and that the most natural and efficient way to address this problem was to shift a large part of the responsibility for data entry to the scientists generating the new sequence data.

More recently, an additional impetus has become evident: journals are rapidly becoming less willing to provide the principal forum for presenting primary sequence data (Burks, 1989). Many journals now ask authors to limit sequence figures to very limited spans that pertain most directly to the discussion in the manuscript and in some cases are publishing articles that discuss the results of sequencing without including any of the sequence data. In these cases, direct submission of the sequence data to GenBank is the only way that these data become generally accessible to the scientific community.

For the future, the efforts on the part of journals to enhance direct author-initiated sequence data submission to the data banks provide distinct advantages for each of the groups involved: the journals will tie into an alternative forum for presenting sequence data that give them more flexibility with sequence presentation in articles; GenBank will be better able to keep up with

* Corresponding author: telephone 505-665-3293; network, pgil%histone@lanl.gov.

the flow of sequence data; and the scientific community will gain more complete and timely access to the data they desire.

The data bank staffs have devoted considerable effort to finding ways of encouraging researchers to submit data directly to the data banks. In 1986 the nucleotide sequence data banks collaborated to construct a sequence submission form for distribution to authors. Later, the form was amended to reflect the information required by the amino acid sequence databases. The form allows authors to present information about the sequence in a systematic way, and enables the data banks to collect this information without having to extract and possibly misinterpret it from the published paper.

Why require rather than request submissions? Most journals began by requesting rather than requiring and enforcing data submission to GenBank. Although this has resulted in a marked improvement in the number of authors submitting data directly to GenBank, fewer than 70% of the authors comply with the journals' requests. This of course means additional effort for the data bank staff in tracking down and entering the withheld data, but perhaps – from the journals' point of view – more significantly, it creates confusion and resulting frustration for the journals' readership, who quickly develop the inexact impression (regardless of the distinction between 'request' and 'require') that, when accession numbers are printed, they will be able to find 100% of the data in GenBank at the time the article is published.

Have other journals adopted these policies and protocols? About three years ago, the *Journal of Biological Chemistry* began requesting authors to submit their nucleotide sequence data directly to GenBank; this request was accompanied by a copy of the GenBank data submission form. Since then, several other journals have begun distributing submission forms and urging authors to submit their data directly to the data banks. The following journals are doing so on behalf of GenBank: *Biochemistry*, *Cell*, *Genomics*, *Immunogenetics*, *Journal of Bacteriology*, *Journal of Biological Chemistry*, *Journal of Bone and Mineral Research*, *Journal of Virology*, *Molecular and Biochemical Parasitology*, *Molecular Biology and Evolution*, *Molecular and Cellular Biology*, *Molecular Plant-Microbe Interactions*, *Plasmid*, *Science*, and *Virology*. The *Proceedings of the National Academy of Sciences of the United States of America* (Burks & Tomlinson, 1989) and *Journal of Immunology* have recently adopted a policy similar to JGM, and it is anticipated that many other journals will follow suit.

[Note that the EMBL Data Library and the DNA Data Bank of Japan (DDBJ) have similarly developed ties with the journals that they cover and that GenBank does not. For example, *Nucleic Acids Research* established a policy in 1988 whereby authors are required to submit their data directly to the EMBL Data Library. *Plant Molecular Biology* is currently implementing a similar policy.]

What are the benefits of this policy? Among the numerous benefits to be obtained from this policy, two stand out.

First, the earlier a sequence is submitted to GenBank, the more beneficial are the results of the internal review process placed on the integrity of the submitted sequence; errors can be spotted and the author alerted in sufficient time to effect the necessary changes in the manuscript prior to publication. Thus GenBank provides an additional valuable step in the manuscript review process, helping to ensure greater integrity of data appearing in the printed article.

Second, that journals *require* submission is the only way to ensure 100% compliance; once this is achieved, every accession number published in the journal will be available in the database. Journals can at this point cease to print sequence information if they so choose, leaving more space for other scientific information and discussion.

How is the confidentiality of data preserved? The GenBank data submission form asks the author to indicate whether or not the data are to be made available immediately or at some later date up to and including the date of publication in the journal. *If the author asks that the data be held until a specific date, the data will not be released until that time.*

GenBank encourages authors to give permission for early release, as this allows more lead time for enabling data to get into a public version of the database by date of publication.

Note. When a sequence is submitted to GenBank prior to being accepted for publication, it is usually entered in the database as 'unpublished' and reconciled at the time of publication. Thus, if a journal declines to publish a manuscript, the sequence remains in the database as 'unpublished' or stays out until published elsewhere, according to the author's wishes.

How the new system works

Overview. As of 1 August 1989, *JGM* will require that all original nucleic acid or amino acid sequence data reported in submitted manuscripts have a corresponding GenBank accession number either at the time of initial submission of the manuscript or, at the very latest, by the time the manuscript is accepted for publication. An accession number, which is a unique identifier associated with each sequence submitted to the database, constitutes evidence that the data have been submitted to, and accepted by, GenBank. The GenBank staff will assign an accession number to data which appear complete and accurate, and will do so within seven days of receipt.

Thus, authors contemplating submission of a manuscript containing sequence data for publication in *JGM* should first send those data to GenBank and incorporate into the manuscript, in a footnote on the title page, the accession number sent to them by GenBank.

Authors who have not obtained such a number are required to do so by the time the manuscript is accepted for publication; they will be informed of this at the time the manuscript is submitted.

Submitting the data to GenBank. This section details the various methods an author can use to submit sequence data to GenBank and obtain an accession number.

The key aid to authors who wish to submit sequence data is the **GenBank sequence submission form**. This form solicits all the information needed to construct a database entry and has been structured to facilitate transcription not only of the sequence but of the data items that provide bibliographic, physical and functional contexts for the sequence. The form has been developed in collaboration with the EMBL and DDBJ nucleotide sequence data banks as well as the protein sequence data banks and is held in common among all these groups. The data submission form exists in both a paper and computer-readable version; the latter can be completed using a text editor.

The Editors of *JGM* would strongly urge that the sequence and annotation data be provided to GenBank in *computer-readable form*; typed versions of the sequence or annotation data should only be used where no access to a computer is available.

A **computer-readable form** may be obtained as follows.

- With all releases of the GenBank, EMBL and DDBJ databases since 1987.
- From GenBank, EMBL or DDBJ by electronic mail. The data bank staffs will send you an on-line version of the form in ASCII format; this can be filled out with any text editor and returned to GenBank by electronic mail or on floppy disc. Instructions are given in the Appendix.
- From GenBank, EMBL or DDBJ on Macintosh or MS-DOS 5¼-inch and 3½-inch diskettes. Refer to the Appendix.
- Through BIONET, the National Computer Resource for Molecular Biology. Subscribers may obtain information about submission by typing the command HELP SUBMIT-SEQ at the system prompt. This 'help topic' describes BIONET's XGENPUB program, which assists users in filling out an on-line version of the sequence submission form and then automatically mails the completed form and sequence data to GenBank. For more information on BIONET contact IntelliGenetics Inc., 700 East El Camino Real, Mountain View, CA 94040, USA (telephone 415-962-7337; e-mail bionet@net.bio.net).

A **printed GenBank sequence submission form** is available from the *JGM* Editorial Office, or from GenBank, EMBL or DDBJ. Details on how to contact the data banks are given in the Appendix.

Formats for submitted data. A data submission should include the following.

- The sequence itself, in computer-readable form (electronic mail, MS-DOS or Macintosh diskettes, or magnetic tape). Typed copy will only be accepted if the authors have no access to a computer.
- A completed computer-readable data submission form for each submitted sequence: if you have two or more sequences, use a separate form for each sequence. A printed form will only be accepted if the authors have no access to a computer.
- A computer network address, or a telefax number. Although not essential, their presence will speed up the return of the accession number.

In addition, the information supplied on the form should conform as closely as possible to the following standards.

- Each sequence submission should include the names of the authors as they appear on the corresponding manuscript (if any).
- Each distinct sequence should be listed separately using the same number of bases/residues per line and clearly indicating its length in bases/residues.
- Enumeration of distinct sequences should begin with a '1' and ascend in the direction 5' to 3' (amino- to carboxy-terminus).
- Amino acid sequences should be listed using one-letter code. The code for representing the sequence should conform to the IUPAC-IUB standards, which are described in the following references: *Nucleic Acids Research* **13**, 3021–3030 (1985) for nucleotides, and *Journal of Biological Chemistry* **243**, 3557–3559 (1968) for amino acids.

How to send your submission to GenBank. GenBank can sequence and annotation data submitted in any of the following ways.

- **Electronic file transfer.** Files can be sent via computer network to **gb-sub%life@lanl.gov**. This address can be reached via various gateways from BITNET, INTERNET, USENET, JANET, JUNET, etc. Ask your local network expert how to send it, or telephone us for help at (505) 665-2177. GenBank maintains a list of personnel at university and other computer sites who can assist locally with electronic submission.
- **Floppy- or micro-diskette.** Macintosh or DOS systems (all sizes and densities). If using word-processing software, the file should be sent as an ASCII text file rather than as a software-specific file.
- **Magnetic tapes.** GenBank can read ½-inch magnetic tapes in any format; and ¼-inch magnetic data cartridges written on UNIX workstations.
- **Printed copy.** *As a last resort only!* Please do not reduce the size of the letters in the sequence.

Refer to the Appendix for more information on how to contact and send material to GenBank.

How quickly will GenBank acknowledge a scientist's submission? Submissions will be acknowledged *within seven days of receipt*; electronic mail submissions will be acknowledged *within 48 hours*. Acknowledgment consists either of an assigned accession number or, if the submission is incomplete or in error, a request for additional information.

There are several things authors can do to minimize the time it takes to get an accession number:

- Be sure that submissions include all the necessary materials and that all the relevant questions on the data submission form have been answered.
- Check the data to be sure that they do not contain inconsistencies or errors and that the sequence is the correct length. Check also that the sequence you are sending is the correct one (different sequence or incorrect length are common errors found in direct submissions).
- If you have one, be sure to include your computer network address or a telefax number: in the absence of these GenBank will use regular mail to send notification of accession numbers (airmail to Europe can take up to ten days).

GenBank will not issue an accession number over the telephone; there are too many possibilities for error. Similarly, authors should not use the telephone to transfer accession numbers to Editors or the Editorial Office; rather they should send a copy of the notification received from GenBank by mail or telefax, or insert the number at proof stage.

Authors are strongly encouraged to send their sequence data to GenBank as early in the manuscript preparation stage as possible. Early submission averts the possibility of delays in publication while waiting for notification of an accession number. In addition, the checks applied to the sequence at the time of submission may alert the author to possible errors in the data, allowing corrections to be made in the manuscript.

Where does the accession number go? The accession number should appear in a footnote on the title page of the article when it is printed. For the *JGM* the footnote will read: 'The nucleotide [or amino acid] sequence data reported in this paper have been submitted to GenBank and have been assigned the accession number M00000'. Other journals may have adopted different formats or locations for printing the accession number.

How quickly will the submitted data appear in GenBank? At present, it ordinarily takes data that arrive at Los Alamos three to five months to make their way into a public GenBank release. This lag is due in part to the quarterly update schedule GenBank now adheres to and in part to the time it takes to process and check new data.

However, if GenBank receives a direct, computer-readable sequence at least five weeks before we close for update, the sequence will be in the next release of the database.

Note that GenBank is currently restructuring its data maintenance system so that more continuous updating of the database will be possible; the time it takes for newly submitted data to get into a public release of the database will be greatly reduced when this is implemented (Burks *et al.*, 1989).

Will data submitted to GenBank appear in other databases? GenBank is currently exchanging data with the DNA Data Bank of Japan (DDBJ) and the EMBL Data Library; thus, data submitted to and incorporated into GenBank are subsequently incorporated into these other nucleotide sequence data banks (and vice versa). (Note that DDBJ, EMBL and GenBank have divided the journals up so that any given journal only needs to interact with one of the data banks.) Similarly, the amino acid sequence data banks (PIR, MIPS and JIPID) also receive GenBank releases. Additionally, in the case of PIR, GenBank forwards all electronic mail submissions to PIR at the time they are received.

APPENDIX

Access to GenBank data

The following are the current options for accessing the database; new releases of floppies are prepared semi-annually, and tapes and on-line data are updated quarterly. For further details on obtaining access to GenBank data, enquiries should be addressed to IntelliGenetics, Inc., 700

East El Camino Real, Mountain View, CA 94940, USA [telephone, (415) 962-7364; e-mail, genbank@bionet-20.bio.net].

Magnetic tape. The data are currently available as:

- Industry-standard 9-track tape, written as unlabelled ASCII files in fixed-length records of 80 characters.
- VAX/VMS Backup saveset (on 9-track tapes or TK-50 cartridges).
- Compressed UNIX tar archives (on 9-track tapes or SUN $\frac{1}{4}$ -inch QIC 24-format tape cartridges).

Floppy or microdiskette. The data are available on:

- $5\frac{1}{4}$ -inch, 360 kbyte, double-sided, double-density disks, which can be read by IBM or IBM-compatible computers running PC-DOS or MS-DOS (versions 2.0 or greater).
- 1.2 Mbyte, double-sided, high-density disks, which can be read on IBM or IBM compatible systems as above.
- $3\frac{1}{2}$ -inch, 800 kbyte, Macintosh-format microdiskettes, which can be read by Apple Macintosh computers.

Dial-up on-line service. Ten-minute sessions of access to an on-line edition of the most recent release of the database are currently offered free of charge. To gain access to GenBank on-line, a researcher needs only a modem and communications software. Access is accomplished via the Telenet public data network. Entries can be identified and extracted by accession number, author name, entry name, keyword phrase, or organism name. Extracted entries can then be downloaded to your local computer. For a fee, users may request individual on-line accounts, which offer access to additional searching and alignment programs, bulletin boards, and electronic mail. Contact IntelliGenetics for more information.

Secondary sources. The GenBank project was not chartered to develop or distribute software for manipulation and analysis of the data by the user community; however, such software can be acquired from a number of independent distributors and vendors (Rawlings, 1987). In many cases, these software packages also include a redistributed (and often reformatted) version of GenBank data as part of the package.

How to contact the nucleotide sequence data banks for data submission

GenBank. JGM's primary association is with the GenBank nucleotide sequence data bank. EMBL and DDBJ will distribute forms and accept submissions (see below), but to avoid complications you are urged to communicate directly with GenBank.

- Electronic mail:
General enquiries, error reports: genbank%life@lanl.gov
Sequence submissions and forms: gb-sub%life@lanl.gov
- Telephone: USA (505) 665-2177
Outside office hours (8.00 a.m. to 5.00 p.m., Mountain Standard Time) a taped message on a telephone answering machine will request that you leave your name, number and any other information you wish.
- Telefax: USA (505) 665-3493
Note. Do not use this fax number to send submissions; the quality of the output is not as good as the original. If you have to send a printed submission, please use conventional mail.

- Postal address: GenBank Submissions
Group T-10, Mail Stop K710
Los Alamos National Laboratory
Los Alamos
NM 87545
USA.

EMBL. Authors in Europe who experience difficulties in connecting with computer networks in the USA may obtain an electronic or hardcopy GenBank submission form from EMBL: the form follows exactly the same format. In addition, EMBL will accept electronic mail and hardcopy sequence submissions connected with this and other GenBank journals on GenBank's behalf, but will pass them on to GenBank for accession number issuance following acknowledgment of receipt. The delays imposed by forwarding of electronic mail submissions are minimal but those by forwarding of postal submissions may be significant and such submissions should thus be mailed directly to GenBank. Please allow plenty of time if you need or plan to use any of these paths.

- Electronic mail:
 - General enquiries: datalib@embl.earn
 - Sequence submissions and forms: datasubs@embl.earn
- Telephone: +49-6221-387-258
- Telefax: +49-6221-387-306
- Telex: 461613 (embl d)
- Postal address: Data Submissions
EMBL Data Library
Postfach 10.2209
6900 Heidelberg
Federal Republic of Germany

DDBJ. Authors in Japan who have difficulty with mail or computer network connections with the USA may obtain a submission form from DDBJ. DDBJ staff will acknowledge submissions, but will forward submissions connected with this journal to GenBank.

- Electronic mail:
 - General enquiries: ddbj@niguts.nig.junet
 - Submission forms: ddbjsub@niguts.nig.junet
- Telephone: 559 75 0771
- Telefax: 559 75 6040
- Postal address: Data Submissions
Laboratory of Genetic Information Analysis
Center for Genetic Information Research
National Institute of Genetics
111 Yata
Mishima Shizuoka 411
Japan

This work was supported by a National Institutes of Health GenBank Contract, NO1-GM-7-2110 and also under the auspices of the US Department of Energy.

'GenBank' is a registered trademark of the US Department of Health and Human Services.

REFERENCES

- BURKS, C. (1989). How much sequence data will the data banks be processing in the near future? in *Biomolecular Data: a Resource in Transition*, pp. 17-26. Edited by R. R. Caldwell, D. G. Swartz & M. T. McDonald. Oxford: Oxford University Press.
- BURKS, C. & TOMLINSON, L. J. (1989). Submission of data to GenBank. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 408.
- BURKS, C., CINKOSKY, M. J., GILNA, P., HAYDEN, J. E.-D. AND OTHERS (1989). GenBank; current status and future directions. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences (Methods in Enzymology vol. 183)*. Edited by R. F. Doolittle. New York: Academic Press (in the Press).
- RAWLINGS, C. (1987). *Software Directory for Molecular Biologists*. London: Macmillan.