Submission of data to GenBank

CHRISTIAN BURKS AND LAURIE J. TOMLINSON

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

In response to both the ever-increasing rate of determining nucleotide sequences (1) and the growing trend among journals to allow articles to appear that describe the results of determining a sequence without explicitly presenting the sequence (1), GenBank*(2–5) and a number of the journals that publish nucleotide sequence data are working together to promote the direct, timely submission of nucleotide sequence data to GenBank. The policy being established by the PROCEEDINGS is described in the editorial on p. 407; here, we will provide a brief summary, in the context of this policy, of the mechanics of submitting data directly to GenBank.

Contacting GenBank. Requests for submission forms, as well as submissions of new data and suggested revisions of data, can be directed to

GenBank Submissions
T-10, MS K710
Los Alamos National Laboratory
Los Alamos, NM 87545 USA

505-665-2177 (telephone)
505-667-1754 (telefax)

gb-sub%life@lanl.gov (electronic network mail).

Submission Form. All direct submissions of data to GenBank should use the GenBank submission form. (This form has been developed in collaboration and is held in common with our collaborators in nucleotide sequence collection, the EMBL Data Library and the DNA Data Bank of Japan, as well as with those—the Protein Identification Resource and its collaborators—interested primarily in amino acid sequences.) This form has been structured to facilitate the transcription of not only the sequence but of all the data items that provide a bibliographic, physical, and biological context for the sequence.

Acquiring the Submission Form. This form is available on both computer-readable (e.g., over electronic mail network or on floppy disk) and printed media from the address above. It is also available on a number of electronic bulletin boards (e.g., BIONET), and a number of institutional computer centers have current versions available. Finally, it is available on Release 48.0 of GenBank and Release 11.0 of the EMBL data base and all releases forward.

Filling Out the Form. Particular attention should be paid to filling out the form completely so as to avoid any delays in processing the form once it is submitted to GenBank. All relevant questions on the submission form should be filled out. Sequence data should be checked for typographical integrity and consistency with the corresponding assigned functional features.

For a given sequence, the number of functional features that could be assigned is, of course, somewhat open-ended. We consider delineation of transcribed (where known) and translated regions to be a minimum requirement; guidelines for further annotation are given on the submission form.

Submitting Data. We prefer to receive the submission form and corresponding sequence data on computer-readable media (e.g., electronic mail, floppy disk, or magnetic tape—
in order of decreasing preference). If these media are not available to the individual correspondent, we can also process the printed version of the submission form and/or clean, printed listings of the sequence data.

Electronic file transfer. Files can be sent by computer network to the network GenBank submission address listed above. This address—in most cases with no modification—
can be reached from various networks, including BITNET, ARPA NET, USENET, JANET, JUNET, etc. Ask your local network or system expert how to send electronic mail, or contact us for help. Floppy disks. We can read Macintosh diskettes or 5½-inch diskettes written on MS-DOS systems. We prefer that the submitted data be written as flat text files rather than in a format specific to a given word processing system. Magnetic tapes. We can read ½-inch magnetic tapes in any format, and ½-inch magnetic data cartridges written on SUN workstations. Printed copy. Please do not reduce the size of the sequence, and submit a copy of the lowercase sequence free of additional graphical overlay.

Acknowledgment and Receipt of Accession Number. When GenBank receives a submission, it will normally acknowledge receipt within 1 week (usually 1–2 days for electronic mail submissions). For submissions that are determined to contain all the requisite data, the acknowledgment will include an accession number that will be permanently associated with the submitted sequence in the data base. For submissions that appear incomplete, acknowledgment will consist of a request for additional data. Incomplete data items on the submission form and inconsistencies between the assigned functional features and the corresponding sequence are the biggest source of delays in processing submission forms.

The author will provide the accession number to the PROCEEDINGS office to be included in a footnote to the published paper. This accession number should be used when referring to the corresponding data in the future.

How Quickly Will Submitted Data Appear in GenBank? It currently takes data that arrive at GenBank 3–5 months to make their way into a public release, depending on arrival time relative to the quarterly updating cycle. Thus, data submitted several months prior to publication are likely to appear in the data base at or just before the time of publication.

Note that we are currently restructuring our data maintenance system so that we will be able to implement more continuous updating of the data base; this will considerably reduce the time it takes for newly submitted data to get into a public release of the data base.

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


*Inquiries regarding accessing the data base should be directed to: GenBank, IntelliGenetics, Inc., 700 El Camino Real East, Mountain View, CA 94040 (telephone, 415-962-7364; e-mail, genbank@bionet-20.bio.net.).