A new system for direct submission of data to the nucleotide sequence data banks

The EMBL Data Library and GenBank(R) staff

The Executive Editors of Nucleic Acids Research (NAR) have decided to establish a policy that manuscripts reporting or discussing primary nucleotide sequence data will be considered for publication only when accompanied by evidence that these data have been deposited with the EMBL/GenBank nucleotide sequence data banks. Here we discuss how NAR and the data banks will implement this policy and the implications for authors who submit manuscripts to NAR.

1 WHY A NEW SCHEME?

The rate at which nucleotide sequence data are being generated worldwide is increasing dramatically, a trend which will undoubtedly continue. EMBL and GenBank are being swamped by this enormous volume of data, to the point where current processing systems are inadequate. Furthermore, it is becoming increasingly difficult for the data bank staff to maintain the necessary biological expertise to abstract the relevant information from all published articles. The volume of sequence data is also problematic for scientific journals, which are becoming increasingly unable and/or unwilling to print these data.

At the same time, it is clear that access to a complete, up-to-date collection of primary sequence data is essential to virtually all areas of molecular biology, including research in basic biological processes such as growth control, differentiation and oncogenesis, as well as in biotechnology and microbial and genetic diagnostics. It is therefore essential that all sequences which have been determined in laboratories around the world are rapidly made available to the research community. Since both EMBL and GenBank currently collect much of their data by scanning journals for papers containing primary sequences, the trend against publication of these data must be accompanied by the establishment of alternative mechanisms for collecting and annotating them.

Solutions to these problems will require, among other things, a radical revision of the mechanisms by which sequence data and related information enter the databases. The present data capture mechanisms need to be replaced by ones in which data capture is uncoupled from publication and researchers take a more active role in preparing their data for inclusion in the databases.

© IRL Press Limited, Oxford, England.
Nucleic Acids Research

2 PRESENT SCHEME FOR DIRECT SUBMISSION TO THE DATA BANKS

Both EMBL and GenBank have devoted significant efforts to finding ways of encouraging researchers to submit data directly to the data banks. The most recent step in this direction was the establishment of a system in which journals distribute EMBL/GenBank data submission forms to authors of manuscripts containing nucleotide sequences. The form requests that authors send their sequence along with relevant descriptive information (annotation) to EMBL or GenBank, and gives instructions about how to do so. This system has been functioning since August 1986 and has been supported with enthusiasm by many journals and researchers.

This scheme has greatly increased the number of researchers who submit data to the data banks, but it is clearly insufficient in the long run. Compliance by authors is voluntary, and therefore many do not respond; EMBL presently gets responses from approximately 35% of the authors who presumably received a data submission form from a journal. Only about 30% of the responses include the sequence data in computer-readable form; the remaining 70% of the sequences are sent as computer printouts and must therefore be typed into the computer by data bank staff, despite the fact that they clearly already reside in the researcher’s computer.

However, this system was not conceived as an end in itself but as a first step in a broad revision of data capture mechanisms. The next goal is to ensure that all primary sequence data relevant to a publication reach the database, regardless of whether a journal actually prints them. The new policy of NAR and the data banks is intended as a step towards achieving this objective.

3 GENERAL DESCRIPTION OF THE NEW SYSTEM

This description summarizes the main points of the ultimate system as we presently envisage it.

(a) NAR will consider manuscripts containing or discussing nucleotide sequence data for publication only when they are accompanied by evidence that the data have been deposited in the EMBL/GenBank databases. The data bank staff will assign an accession number to submitted data which are seen to be complete and accurate.

(b) EMBL and GenBank will make these sequence and annotation data available (via an on-line service) by the time the article referring to them is published.

(c) Reviewers of NAR manuscripts will have on-line access to the primary sequence data, enabling them to evaluate critically the analyses and interpretations of these data.

4 IMPLEMENTATION

The system will be implemented gradually, beginning with a relatively simple scheme that provides the functionality described in point (a) above and gradually incorporating more sophisticated features which address the remaining points.
4.1 Launching the new system

As of 1 January 1988, manuscripts submitted to NAR and containing or referring to primary sequence data should be accompanied by a database accession number. This number, which permanently identifies a sequence in the databases, constitutes evidence that the data have been submitted to and accepted by one of the data banks.

4.2 STAGE 1: Ensure that sequence data published in NAR have been deposited with the EMBL/GenBank databases

This is the period beginning with the launching of the scheme and lasting until the time when EMBL starts offering an on-line service (see section 7).

Initially, authors will be required to submit their data to EMBL; at some point in the future, they will be given the option of submitting data to either EMBL or GenBank (see section 4.5).

The criteria for granting an accession number will be that the author has submitted the following:

(a) the sequence itself, in computer-readable form (via computer network mail, on magnetic tape or on a floppy diskette). Printouts will be accepted ONLY in the unlikely event that the authors have no access to any sort of computer.

(b) complete annotation; that is, a data submission form properly filled out. The data submission form is printed at the end of this article, and can be copied and used for this purpose. It is also available on various computer-readable media (see section 5.1).

(c) a computer network address, a telex number or a telefax number (advisable but not required) (see section 5.2).

The data will be checked for completeness and for certain types of errors. Submissions containing errors will not be given an accession number until the mistakes are corrected. Complete and annotated data will be assigned an accession number as described in section 5.

4.3 STAGE 2: Make data available on-line by time of publication

This stage will begin when EMBL launches its on-line service and can make the sequence data pertaining to published articles available on-line to readers of NAR by the time the journal is printed.

Efforts are underway to work out efficient methods of transmitting data to GenBank outside the normal release schedule. Data relevant to NAR publications will then also be available via the GenBank on-line service (see section 7).

4.4 STAGE 3: Make data available on-line to reviewers

The data banks will establish a system whereby reviewers of a manuscript will (in confidence, where required) be able to access the data on-line and to
check the validity of various claims made by the authors (for example, that a consensus sequence or an unidentified open reading frame is present).

4.5 Other features to be implemented

Once the system is functional, authors will be given the option to submit data to GenBank as well as to EMBL. This will be announced in NAR.

5 COMMUNICATION AND INFORMATION FLOW

5.1 How will authors know what annotation to provide?

EMBL and GenBank entries include not only primary sequence data but also descriptive information such as the source of the sequenced segment (e.g., organism, strain, tissue) and the location of interesting features/regions within the sequence (e.g., signal peptide, coding regions, regulatory signals). If the sequence encodes a protein, information such as the location of sites subject to post-translational modification and of functional domains within the protein is also collected. Data relevant to proteins is transmitted to the protein sequence data banks.

To ensure that data submissions are sufficiently well annotated, authors should fill in an EMBL/GenBank data submission form for each sequence they submit. (A properly completed form will be one of the requirements for receiving an accession number.) At present, the data submission form exists in both a paper and a machine-readable version; the latter can be completed using a text editor. These two versions are presently available from a number of sources, and there are many possibilities for establishing additional ones:

(a) Hard copy form: printed at the end of this article and available from EMBL and GenBank.

(b) Machine-readable form: available by electronic mail (computer network) or on floppy disk from EMBL or GenBank, on all off-line (e.g., magnetic tapes) distributions of EMBL or GenBank, and through the SEQNET (Cambridge, U.K.) and BIONET (Mountain View, California, U.S.A.) molecular biology bulletin boards. It will be available for download via EMBL and GenBank on-line services, and could also be sent to various local on-line services or computer centres. (Local computer centres can contact EMBL or GenBank if they wish to maintain a copy.)

In the longer term, the data banks intend to replace the machine-readable form with a data entry program that can be used interactively and that will perform data validation.

5.2 How long will it take for authors to receive an accession number?

Authors can expect a response from EMBL within 7 days of receipt of their data. This response will consist either of an accession number (assuming that all information has been provided and no errors have been discovered) or, if data are missing or errors are detected, an explanation of what additional information is needed. The 7 day estimate is based on the assumption that the
authors have provided a telex/telefax number or a computer network address; if not, notification will be by regular post and will therefore take longer. It will not be possible for the data bank staff to contact authors by telephone. Authors can speed up the process by sending data either electronically or via telefax.

6 DATA SECURITY

The present EMBL/GenBank data submission form asks authors whether they agree that their submitted data can be made available to the public immediately or whether it should be withheld until publication. Roughly 80% choose the former option and 20% the latter. Withholding data until publication poses no problems to our current data management procedures; this group of entries is rendered inaccessible to users outside the Data Library staff until we find the published article during our weekly journal scan.

It will be more complicated to withhold data from the general user community and yet make them available on-line to reviewers, and a new system will therefore be required. The difficulty is compounded by the facts that the reviewers differ for every manuscript (i.e., they are not simply a small group of people who can be given "privileged" access to data), and that only the journal editors know who the reviewers for any given manuscript are.

It is worthwhile to point out that the need for such a mechanism will probably decrease over time. Once journals stop printing sequences, submission to the relevant data bank is likely to play an increasingly important role as a means by which scientific credit/priority for sequence data is assigned. As a result, it may become far less common for authors to request that data are withheld until a related publication appears. But these times are not yet upon us, and at present some molecular biologists would probably be unhappy with a direct submission system that made no provisions for data to remain confidential.

7 ON-LINE SERVICES RELEVANT TO THE PROPOSED SCHEME

EMBL plans to introduce an on-line service by the beginning of 1988. The system will start out with modest capabilities and will gradually be expanded to include more sophisticated features.

The initial system will allow users connected to the BITNET/EARN network to request that small datasets be sent over the network to their local machine. The data of interest will be identified by citation, accession number, keyword, author name or taxonomic classification. This service will enable a researcher who reads about a sequence in a recent journal to obtain it in computer-readable form without having to wait for the next release of the database.

The system will be expanded to support dial-up access for remote users, who can then log onto the EMBL machine and download data. This will enable personal computer users not attached to any network to access the on-line service. Other improvements to the system will be aimed at providing more extensive retrieval capabilities so that users can identify data of interest by a variety of criteria such as organism, keyword, or date of entry into the database.
GenBank presently has an on-line service which provides access to data from the most recent release. This service is accessible via dial-up to PC users with modems.

* Address correspondence to:

Dr. P. Kahn  
The EMBL Data Library  
EMBL  
Postfach 10.2209  
6900 Heidelberg  
West Germany  

Telephone: (06221) 387409  
Telefax: (06221) 387306  
Telex: 461613 (embl d)  

Computer network: datalib@embl.earn