Electronic Data Publishing and GenBank

MICHAEL J. CINKOSKY, JAMES W. FICKETT, PAUL GILNA, CHRISTIAN BURKS

GenBank, the national repository for nucleotide sequence data, has implemented a new model of scientific data management, which we term electronic data publishing. In traditional publishing, both scientific conclusions and supporting data are communicated via the printed page, and in electronic journal publishing, both types of information are communicated via electronic media. In electronic data publishing, by contrast, conclusions are published in a journal while data are published via a network-accessible, electronic database.

SEVERAL YEARS AGO, IN AN EFFORT TO KEEP UP WITH AN exponentially increasing flow of nucleotide sequence data, we began looking at ways of combining two evolving technologies (namely, database management software and computer networks) into a more effective system for the communication of scientific data. We have designed and implemented a specialized form of electronic publishing, which we term electronic data publishing, where data (in our case DNA sequences and related annotation) are gathered, processed, and distributed electronically. Rather than compete with traditional scientific publications, electronic data publishing is designed both to complement and to support printed publications.

In this article we describe the changes that we have made to the GenBank project over the past several years in order to implement electronic data publishing. We also discuss the impact that this has had on the performance of the project as a whole. Finally, we explore some of the implications of this new form of scientific communication and some of the questions it raises.

The Traditional Model for Scientific Databases

In 1982 the Los Alamos Sequence Library (*1*) became GenBank (*2*), the national repository for nucleic acid sequence data. The standard model for public scientific databases at that time, and the one on which the original design of GenBank was based, is represented by the open arrows in Fig. 1. In this model, researchers gather data in an effort to answer a question or to test a hypothesis, and when sufficiently interesting scientific results have been obtained, a paper describing those results is published in a peer-reviewed journal. The authors of such an article would typically present some of the most significant data leading to their conclusions. The role of a database such as GenBank was to extract those data (in our case, nucleotide sequences) from the published article and to make them available in electronic form to the entire scientific community.

In 1982, nucleotide sequence data (represented as long strings of a's, c's, g's, and t's, along with associated biological annotation) were being generated at a rate of much less than one million nucleotides per year, so researchers expected to be able to analyze visually all of the data related to their area of interest. Essentially, people wanted to be able to read the database. For this reason, we designed the original database format as an ASCII text file, tailored for easy readability, which we distributed to users on a periodic basis, in both hard copy (*3*) and electronic form. The text file format also reduced the start-up costs by permitting the use of many common text-based tools for database maintenance (*4*) but led to considerable tension between the desire for an intuitively structured text document on the one hand and the requirements for machine usability on the other. An additional, hidden cost to this approach was that our annotation staff had to learn the details of our file format and, as the database developed, the formatting conventions became increasingly complex.

This model worked reasonably well for the first few years of the project, but it was not long before dramatic increases in the rate of generation of nucleotide sequence data began to tax the design severely. When GenBank was started, it was reasonable to expect that the volume of data would increase roughly linearly. In fact, however, the growth turned out to be exponential (*5*). Data are currently entering the database at the rate of more than 20 million nucleotides per year; this increase is due in part to dramatic improvements in sequencing technology. The effects of this dramatically higher rate of data creation were felt in several ways: (i) the amount of labor required to keep up with the flow increased in direct proportion to the volume of data; (ii) the lag time inherent in a quarterly distribution cycle became an increasing inconvenience to the people who depend on the database for their research; and (iii) as the volume of data increased, journals began to place tighter limitations on the amount of sequence data that they would publish. Authors were increasingly forced to present excerpts from their sequence data (for example, exons only). Although this is an understandable position for the journals to take (particularly in light of the limited usefulness of printed sequence data), large amounts of high-quality, interesting nucleotide sequence data would never reach GenBank if we were to continue relying on the printed page as our primary source of data.

Further, there was little evidence that the rate of increase in the creation of new data would diminish at any time in the near future. In 1985 and 1986, when we began considering the ideas presented in this article, the Human Genome Project was first being proposed. The ultimate goal of this project is to determine the complete sequence of human DNA, which is approximately 3×10^9 nucleotides in length. For this purpose, significant resources have been allocated to the development of technology that could easily result in a rate of data creation of one million nucleotides per day. It was clear

The authors are members of the Theoretical Biology and Biophysics Group of Los Alamos National Laboratory, F10, MS K710, Los Alamos, NM 87545.

that a radical redesign of the GenBank operation was needed before we could handle data at that rate.

Two particular areas of rapidly advancing computer technology seemed especially applicable to our problem: the improvements in database management technology (*6*) and the advances in network design and availability (7). Over the past decade, database management system (DBMS) technology improved enough to make it possible to manage large amounts of data (gigabytes or more) efficiently on relatively inexpensive hardware. Although in the past it was commonly believed that large, production databases needed to be custom-built if they were to yield adequate performance, a number of commercial vendors have in recent years released powerful yet affordable DBMS products.

Over the same time period, the Defense Advanced Research Projects Agency Internet (*8*) has made it possible for most computers at major research facilities to be connected to one another by high-speed communication links, allowing scientists to exchange information with great ease. In recent years, many people in the scientific community have become accustomed to participating in global "conversations" as they unfold on various electronic bulletin boards around the Internet. Perhaps more than any other advance in computer science, computer networks have the potential to radically alter the way in which people access information.

Electronic Data Publishing

Confronted with the problem of keeping up with the flow of data in a timely fashion, we saw the need to create a new model for scientific databases in which the community was directly and naturally involved in the maintenance of the database and in which they would be given more direct access to the data. Our primary goals were the following: (i) to make data available electronically at the same time that a paper presenting conclusions based on those data appears in print; (ii) to develop a system capable of handling a sustained input rate of at least one million nucleotides (and associated annotation) per day; (iii) to structure the data so that it would be readily machine-parsable; and (iv) to have mechanisms in place to ensure the quality of the data.

As a first step, we [along with the staff of the European Molecular Biology Laboratory (EMBL) Data Library (*9*)] worked closely with journal editors in the development of policies requiring the submission of data to one of the databases before a paper could actually appear in print [see, for example, (*10*)]. This step alone, however, was not enough to enable us to reach all of our goals. If our fixed-size staff were to keep up with an exponentially increasing data flow, we would also need to receive the data in an automatically parsable, electronic form. We needed a new protocol that would allow for the development of software to assist scientists in the creation of direct submissions. And given the distribution delays inherent in a quarterly release cycle, we needed to design the system in such a way that it would naturally support automatic distribution of the data in essentially "real time."

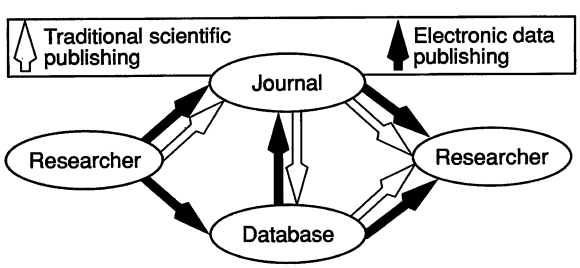


Fig. 1. In the traditional model for scientific databases, data were extracted from published articles by specially trained database staff. In the electronic data publishing model, data flows primarily from the original researcher to the appropriate database, bypassing the time-consuming step of manual transcription. In this model, the flow of data between journal and database is essentially reversed, with only a small subset of the data actually appearing in a journal article.

We developed the idea of electronic data publishing in order to reach these goals. Electronic data publishing uses a highly structured, network-based communication channel through which scientists can present their experimental results to others with a minimum of effort. This new communication method exists alongside the standard journal publication process without being dependent on the journals as a source of data. In fact, as shown in Fig. 1, with electronic data publishing the direction of the flow of data between journals and the database is reversed from that of the old model, with authors of articles now being able to excerpt or cite data from the database (for example: L. M. Corcoran *et al.*, GenBank 66, M12345). The role of the database staff in this model is less to gather data than it is to provide the required technology and quality assurance through data review and validation. [Early discussions of portions of the electronic data publishing paradigm, including the basic idea of community involvement in maintaining the database, can be found in (*11*). Other molecular biology database efforts that have implemented elements of the electronic data publishing concept are described in (*9, 12*)].

Internal structure: A relational database. In order for the data to be most useful to the community, they must be delivered in such a way that software can make efficient use of the biological annotation connected with each sequence. If this condition were not required, a simple, unstructured electronic bulletin board would suffice as the communication channel. However, if people are to be able to write software, for example, to automatically translate protein coding regions of DNA, the information identifying such regions must be stored and presented in a highly structured form.

The design of this internal structure is important because the database appearing at the center of an electronic data publishing system defines not only the types of objects that can be discussed but also the kinds of statements that can be made about those objects. For example, one central difficulty for nucleotide sequence databases is in adequately representing different, and seemingly conflicting, views of a sequence. For the molecular biology community, a sequence is, in one sense, a publication. The task of the database then is to give proper credit, to represent the literature, and to provide a stable, unchanging archive for the data as originally presented. However, in the context of biological reality, a nucleotide sequence as represented in a database is an imperfect and incomplete picture of an idealized molecule: imperfect, because experimental data always have errors; incomplete, because there is always more to learn; and idealized, because we speak of the sequence of a gene in a species, ordinarily ignoring individual variation. In this context the charter of the database must be to present a continually changing picture that represents our most accurate, complete, and up-to-date understanding. Our solution is to store unchanging reports of the sequences as originally presented but also to provide a syntax by which a complete, correct, and up-to-date picture of the biological reality can be built up out of a composite of these reports.

More generally, rather than simply support a particular distribution format, the database design should model the underlying reality as accurately as possible. Aside from raw sequence data, GenBank includes information about the physical context in which the sequences are found (such as taxonomy and laboratory host information); the functional context (what the sequence does in nature); and the bibliographical context (who determined the sequence) (*13*).

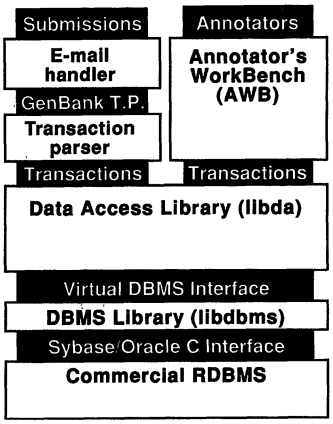


Fig. 2. The software architecture of the GenBank project.

The internal design of the database is based on the relational model, in which data can be thought of as being represented in rows and columns of tables. Although we considered using the object-oriented database model (*14*) because of its ability to handle complex data types, we chose to use the relational model because of its far greater maturity in the marketplace (*15*). This will ensure that we will have adequate vendor support and a stable foundation for our software development.

Internal structure: Support software. Building an electronic data publishing system was in part a matter of assembling components of existing technology [for example, the Internet and commercial relational DBMS (RDBMS) products] and in part applying well-known ideas in a new area (for example, a relational schema for GenBank is conceptually quite different, and more complex, than one for accounting data). In addition, significant custom development was necessary, including the creation of the Data Access Library and the Annotator's WorkBench, in order to achieve the desired configuration (Fig. 2).

The Data Access Library is a collection of C language functions that implement an entity-based interface to the relational database. Entities are similar in many respects to the objects of object-oriented programming; in particular, entities are complex data types that typically combine information from several relational tables. This library has the ability to execute transactions on the various entities in the database (for example, "add a sequence" or "update an author"). The functions in this library are called by programs to retrieve data from the database, add data to the database, or modify data already in the database. The application programs written to use the Data Access Library do not depend on the schema of the underlying database; they only depend on the entity definitions. Further, this library implements many of our quality checks, thereby ensuring that the data have been validated before entering the database.

Because the Data Access Library is highly portable, with well under 500 lines of DBMS-specific code, and because all database functionality is implemented at this level (including, for example, relational integrity), we are able to run all of our software on several different RDBMS platforms with little or no change. In addition, because the Data Access Library has its own DBMS-independent transaction mechanism, by simply recording all successful transactions, this library fulfills the major requirements for supporting satellite copies of the database (see below).

The Annotator's WorkBench (Fig. 3) is, in many ways, the most visible component of our restructuring. This is a window-based, interactive interface to GenBank in relational form. Built on the Data Access Library, the Annotator's WorkBench extends the entity-based view of the relational database up to the user level, allowing one to browse freely through the database and significantly improving the efficiency of both annotation and retrieval.

In the Annotator's WorkBench, users work with worksheets, which are arbitrary collections of entities from within the database. With a single keystroke, entities contained within a worksheet can be opened, or expanded, to display their contents. Each entity in turn contains other, nested entities that can also be expanded to show increasing levels of detail. In this way, the Annotator's Workbench, which can be run either locally or remotely via either network or modem, provides intuitive, consistent access to all of the data in the database. The Annotator's WorkBench (along with a satellite copy of the database) has been installed at IntelliGenetics and is available as part of the GenBank On-Line Service (*16*).

Data collection methods. Our work with journal editors led to a dramatic change in the way in which we received data, with the majority of sequence data destined for printed publications coming to us before publication. However, as long as data entering GenBank had to be processed manually by database staff, data entry would remain a labor-intensive operation, with the staff size required to grow proportionally to the size of the data input stream. Because most scientists doing DNA sequencing store their data on computers at their laboratory, we defined a machine-parsable format that can be used to transmit these data to us for automatic processing. We call this format the GenBank Transaction Protocol.

To help users create these transactions, the GenBank team at IntelliGenetics has developed a program called Authorin (*17*) (running on both the IBM PC and the Apple Macintosh) which gathers all of the appropriate information from the user and packages it as a transaction that can then be sent to Los Alamos for automatic processing. Currently, transactions received from Authorin (either on floppy or by e-mail) are checked by hand before processing, but the software is in place to allow these submissions to be processed without any human intervention at all. Once this happens, the time required to process a submission and issue an accession number (which both identifies the sequence and verifies submission) will be reduced to minutes as opposed to the days required in the manual system.

In addition, the Annotator's WorkBench enables us to significantly improve the method by which we gather expert advice from our curators (scientists who have been selected to oversee particular subject areas of the database). In the past, curators needed to travel to our site and to work closely with a member of our annotation staff to edit the database. This was required because many complicated formatting conventions had to be learned before one was able to perform updates on the database itself. The Annotator's WorkBench reduces that overhead considerably by eliminating the need for significant amounts of training. Thus, with the Annotator's Work-

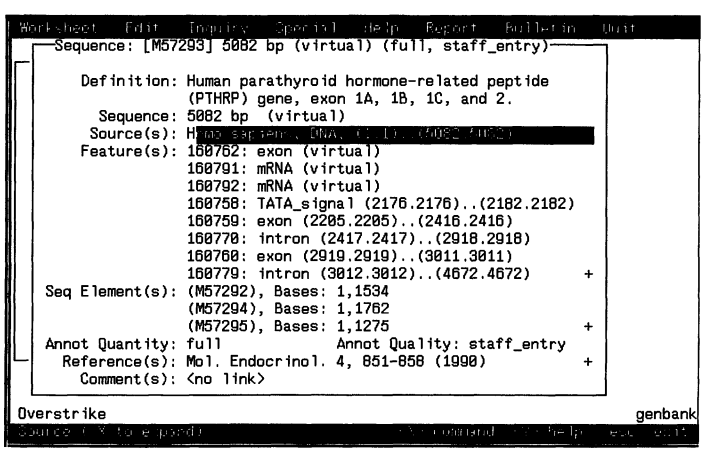


Fig. 3. The GenBank Annotator's WorkBench.

Bench, curators can access the official copy of GenBank remotely, using an easy-to-learn interface, and can perform updates to the database without having to rely on assistance from our staff. This greatly enhances the value of the curators to the database and will in the long run provide one of the key mechanisms by which the quality of direct submissions can be maintained.

Data distribution methods. In 1989, we began updating the GenBank On-Line Service (*16*) nightly, thus making newly entered data immediately available. This service, coupled with the fact that we receive the majority of our data before publication, makes it possible for users to retrieve the data from published articles at the same time that they receive the articles themselves. Although this increased functionality meets many of the needs of our users, there are still good reasons for a user to want a complete copy of the database on a local computer system in relational form. To meet these needs, we designed a mechanism by which many sites that were receiving magnetic tape releases could instead be updated automatically every day. This strategy was possible because all database updates go through the Data Access Library, which archives all successful transactions. These transactions can then be forwarded to satellites in much the same way that direct submissions are sent to us from researchers. Each satellite database runs the GenBank software, including the e-mail transaction-handling software required to process the forwarded transactions. In this way, we are able to bring each copy of the database up to date with the master each day.

GenBank currently maintains a satellite copy of the database at IntelliGenetics in Mountain View, California, where it is made available to users through the GenBank On-Line Service. In addition, we are in the process of setting up satellites at several other sites in the United States and abroad. The Transaction Protocol will also be used to greatly improve the exchange of data with other related databases, such as the Genome Data Bank (GDB) at the Welch Medical Library of Johns Hopkins University. In cases such as this, we only need to exchange a very limited subset of our data (for example, official gene symbols and GenBank accession numbers).

Discussion

Now that there is a working electronic data publishing system in operation, it is appropriate to step back and evaluate the idea as it has been implemented at GenBank. At the practical level, electronic data publishing at GenBank has been quite successful. In 1984, it took on average over 1 year to get nucleotide sequences from journals to the users. This year, even though we processed ten times as much data (14.1 million nucleotides in 1990, as opposed to 1.38 million in 1984), the average delay between the time that an article appears and the time that the data are available in the database is 2 weeks. (And this was accomplished at the same time that the cost of the operation, on a per nucleotide basis, dropped by two orders of magnitude from approximately $10 to roughly $0.10 per base pair.)

At a more conceptual level, however, some larger issues bear discussion. In particular, we must consider questions such as the peer review (and thus the quality) of submitted data and the academic credit that will be associated with database entries (*18*). Although it is natural to question the quality of data that has not gone through the standard peer-review process along with an associated scientific paper, on closer examination these concerns turn out to be unfounded. Our experience indicates that electronic data publishing has actually resulted in a higher level of quality for data in GenBank and therefore also in the journals. Because data are only handled electronically (once they have been entered into a computer by the original researchers), several sources of transcription errors have been eliminated. (Our experience has consistently indicated that the largest source of errors in nucleotide sequence data is transcription errors in the creation of figures for printed publication.) Further, most journals and their reviewers tend to concentrate on the quality of the scientific reasoning embodied in the paper, typically leaving the data essentially unreviewed. We, on the other hand, specialize in the evaluation and analysis of nucleotide sequence data; the (increasingly automatic) checks that we perform on sequence data [see, for example, (*19*)] are far more extensive than most reviewers would have the time to perform. When anomalies in the data are discovered, the person who submitted the data is contacted for clarification. In short, the concept of electronic data publishing has actually increased the level of review that data receive before being made public.

We are similarly encouraged on the question of academic credit. Several years ago, it was not clear whether the scientific community would actively participate in a system dependent on direct submission without assurance of some form of academic recognition for their contribution. However, having journal editors encourage or require submission of data before publication of a paper has been met with far greater acceptance on the part of authors than might have been anticipated. We are currently receiving approximately 80% of our data as a direct electronic submission before publication of any related paper. The research community will always decide de facto the relative importance of the generation of data; what we (and others who have acted as proponents of direct submission) have accomplished is the establishment of a system that researchers are coming to view as the natural mechanism for the communication of large amounts of scientific data.

Our success with electronic data publishing has created a few problems as well. Today, for example, the greatest single delay in the process of getting data out to the user community results from our preserving the confidentiality—when requested—of data submitted before publication. As long as the sequencing of DNA is carried out as an integral component of scientific research (as opposed to being done, say, on a contractual basis), the raw data will remain proprietary until results based on it are published. To accommodate this situation, we offer to hold submitted data in confidence until a related paper appears in print, but this policy puts considerable burden on the database staff to match confidential submissions with publications. To help alleviate this problem, we have implemented software that attempts to match each manually scanned citation with potentially related submissions (a person still makes the final judgment about the correctness of a match), but significant additional effort is still required to minimize the impact of confidential data on our productivity.

What effect will electronic data publishing have on traditional publishing? We expect that the effect will be positive for a number of reasons. First, by insisting that the data supporting a submitted manuscript be accepted, and thus validated, by a database, the reviewers can have increased confidence in the quality of the presented work. (We frequently delay issuance of a database accession number briefly until questions about submitted data can be resolved, and we are actively expanding the scope of checks performed before a sequence will be accepted.) Second, having the data on-line when a paper is submitted to a journal makes it easier for reviewers to perform sophisticated analysis of those data as a component of the review process. (Important security issues must be dealt with before we can provide the community with this service, but we hope that they can be resolved.)

Another advantage of electronic data publishing over printed journals for the communication of scientific data is in the handling of errata. Sequences are often revised and corrected after publication, yet the corrections are rarely made public, and, when they are, there is no "forward reference" in the original article, informing the

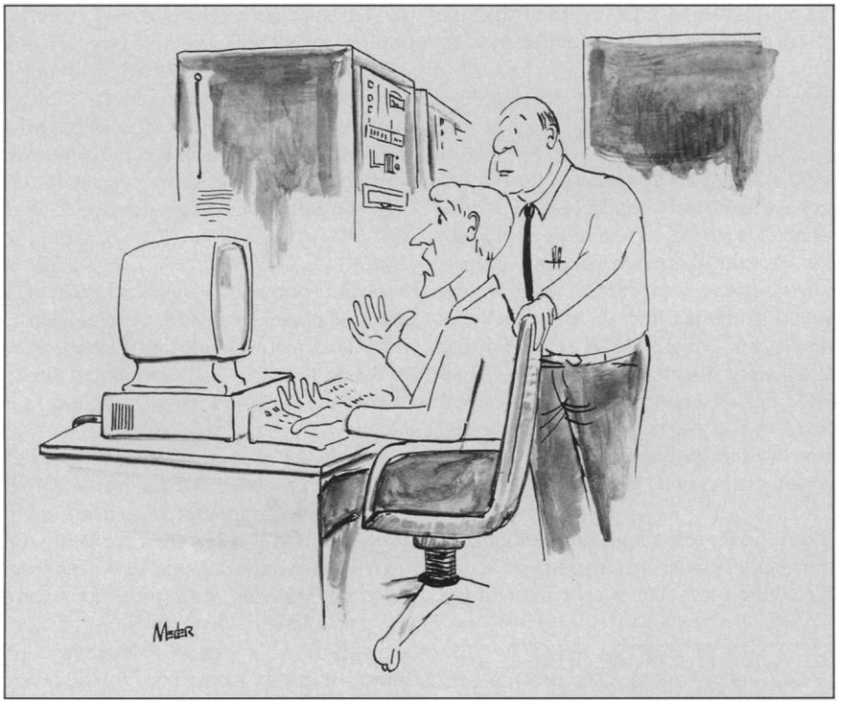
reader of errata to come. On the other hand, when a separate and easily accessible communication route is available for data, and when the data originators "own" the data publication through direct submission, we have found that researchers will more consistently maintain the correctness of the public view of the data. The end result is that users gain access to the most correct and up-to-date information available.

We believe that this model of electronic data publishing has implications for the scientific community as a whole. GenBank's example has made it clear that electronic data publishing, implemented with safeguards on quality and in coordination with traditional publishing, can be an effective method for making scientific data available to researchers at the same time that they receive the conclusions based on those data.

REFERENCES AND NOTES

1. J. Fickett, W. Goad, M. Kanehisa, *Los Alamos Sequence Library* (Los Alamos Unclassified Rep. LA-9274-MS, 1982).
2. C. Burks *et al., Comput. Appl. Biosci.* **1**, 225 (1985); C. Burks *et al.*, *Methods Enzymol.* **183**, 3 (1990).
3. J. S. Anderson *et al., Nucleotide Sequences 1984: A Compilation from the GenBank and EMBL Data Libraries*, parts I and II (IRL Press, Oxford, United Kingdom, 1984); J. Armstrong *et al., Nucleotide Sequences 1985: A Compilation from the GenBank and EMBL Data Libraries*, parts I to IV (IRL Press, Oxford, United Kingdom, 1985); E. Atencio *et al., Nucleotide Sequences 1987: A Compilation from the GenBank and EMBL Data Libraries*, parts I to VIII (Academic Press, Orlando, FL, 1987).
4. J. W. Fickett and C. Burks, in *Mathematical Models for DNA Sequences*, M. S. Waterman, Ed. (CRC Press, Boca Raton, FL, 1987), pp. 1–34.
5. C. Burks, in (*20*), pp. 17–26.
6. C. Date, *An Introduction to Database Systems* (Addison-Wesley, Reading, MA, ed. 5, 1990).
7. J. S. Quarterman and J. C. Hoskins, *Comm. ACM* **29**, 932 (1986).
8. D. Comer, *Internetworking with TCP/IP: Principles, Protocols, and Architecture* (Prentice-Hall, Englewood Cliffs, NJ, 1988).
9. P. Kahn and G. Cameron, *Methods Enzymol.* **183**, 23 (1990).
10. P. Kahn and D. Hazledine, *Nucleic Acids Res.* **16**, i (1988); C. Burks and L. J. Tomlinson, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 408 (1989); I. Dawid, *ibid.*, p. 407.
11. J. W. Fickett, in (*20*), pp. 295–302; M. J. Cinkosky, D. Nelson, J. W. Fickett, T. Marr, *The Restructuring of GenBank* (Los Alamos Unclassified Rep. 88-1255, 1988).
12. F. C. Bernstein *et al., J. Mol. Biol.* **112**, 535 (1977); J. C. Stevens, I. H. Cohen, K. K. Kidd, in (*21*), pp. 69–81; S. Miyazawa, *ibid.*, pp. 47–61.
13. The complete schema describing the internal design of GenBank may be obtained by sending an e-mail message containing only the text "gb-schema" to bioserve@genome.lanl.gov.
14. W. Kim and F. H. Lochovsky, Eds., *Object-Oriented Concepts, Databases, and Applications* (ACM Press, New York, 1989).
15. Descriptions of our earlier work exploring the conversion of GenBank to relational form can be found in M. Kanehisa, J. W. Fickett, W. B. Goad, *Nucleic Acids Res.* **12**, 149 (1984); J. W. Fickett and C. Burks, in *Mathematical Methods for DNA Sequences*, M. S. Waterman, Ed. (CRC Press, Boca Raton, FL, 1989), pp. 1–34.
16. D. Benton, *Nucleic Acids Res.* **18**, 1517 (1990). For more information on this service, contact IntelliGenetics at genbank@genbank.bio.net.
17. J. F. Moore, D. Benton, C. Burks, *Focus (Rochester, N.Y.)* **11**, 69 (1989).
18. A symposium several years ago attracted papers on this topic: C. Burks, in (*20*), pp. 327–333; R. D. Wells, *ibid.*, pp. 335–337; P. Kahn and L. Philipson, *ibid.*, pp. 339–343.
19. J. W. Fickett, *Trends Biochem. Sci.* **11**, 190 (1986).
20. R. R. Colwell, Ed., *Biomolecular Data: A Resource in Transition* (Oxford Univ. Press, New York, 1989).
21. G. I. Bell and T. G. Marr, Eds., *Computers and DNA* (Addison-Wesley, Redwood City, CA, 1990).

22. We thank M. Cassidy, K. Cumella, G. Keen, T. Kelley, D. Sorensen, R. Sutherland, and all former members of the GenBank software development team for their work in implementing the software described here; L. Tomlinson, J. Hayden, D. Nelson, T. Marr, G. Bell, W. Goad, all other past and present members of the GenBank project, the GenBank advisory boards, and the staffs of the EMBL and DNA Data Bank of Japan databases for their contributions to the development and implementation of the ideas presented here; and all of the journal editors whose enthusiastic cooperation has contributed to the success of this effort. GenBank is a registered trademark of the NIH. This work was funded by a contract from the NIH (GM-7-2110) and a field task from the Department of Energy, Office of Health and Environmental Research (ERW-F116), and was done under the auspices of the U.S. Department of Energy.



"No matter how I play with the parameters, the model keeps coming back with the same thing: 'Rain! . . . 40 days and 40 nights'."