GenBank

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

ABSTRACT

The GenBank nucleotide sequence database now contains sequence data and associated annotation corresponding to 56,000,000 nucleotides in 45,000 entries. The input stream of data coming into the database has largely been shifted to direct submissions from the scientific community on electronic media. The data have been installed in a relational database management system and are made available in this form through on-line access, and through various network and off-line computer-readable media. 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Data **Bank**, provides the scientific community with a computer database of all published (and, increasingly often, unpublished) DNA and RNA sequences as well as related bibliographic and biological information that establish the physical, functional, and administrative context of the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 media 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).

SYSTEMS

Inquiries about acquiring access to the data should be addressed to IG;[2] inquiries about submiting 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 4.0C (similar to SunOS UNIX 4.0.3); at LANL, the database work is done on a network of Sun Microsystem (Mountain View, CA) 4/490 server and assorted workstation clients running under SunOS UNIX version 4.1. The GenBank database is maintained at both sites under the Sybase (Emeryville, CA) relational database management system (RDBMS). Software was developed in 'C' language.

All statistics presented here for the database correspond to the on-line version of the RDBMS form of GenBank as of February 11, 1991.

DATABASE ACCESS

The data in GenBank are now available on several media and through several different electonically-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 retrieve@genbank.bio.net for details). FASTA similarity searches are also available by e-mail (send 'HELP' to search@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 (5).

Off-line distribution

GenBank data can be obtained from IG on CDROM, 9 track tape, TK-50 and Sun 1/4' cartridges, 1.2 MB PC-AT floppy diskettes, and 800 kbyte Macintosh floppy diskettes.

* To whom correspondence should be addressed

Hard-copy distribution

The last printed distribution of the database was in 1987 (6); 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).

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'.

Highlights of the new features table are:

Complex features. Features which span non-contiguous sequences, and that were previously described in separate, consecutive entries, can now be represented within the confines of one features table. The new syntax allows specification of sequence spans in other entries, as well as flexibility in how their spans are related to one another.

Controlled descriptors. Previously much of the information of a given feature was collected loosely in a description attached to the feature. To promote the consistency and parsability of this information, it has been structured into 'qualifiers' (e.g., '/EC__number', '/function', and '/gene').

Protein coding regions. We have adopted the feature key 'CDS' to denote protein-coding regions. Every CDS feature will begin with a valid initiation codon and end with a valid termination codon, or provide frame information when the initiation codon is not known. This will continue (and improve) the support for software intended to automatically extract protein-coding sequences from GenBank.

Sequence Tagged Sites

We have designed GenBank, wherever possible, to anticipate, facilitate, and implement new and changing conventions adopted by the scientific community in describing nucleotide sequences and associated functionality (e.g., gene symbols and map locations (7,8)).

One recent example of such changes is the emergence of the Sequence Tagged Site (STS) approach to physical mapping of genomes (9) and the community interest in accessing these data as they evolve (10). An STS typically corresponds to several hundred base pairs of sequence unique to the genome it applies to, and can be defined by the sequences of a pair of PCR primers and a specific set of reaction conditions. These defining data can be expressed as features on nucleotide sequence data, and we have introduced corresponding fields to the database that are included in the flat file distribution format as keys in the features table. STS data will begin appearing in the features table in Release 68.0 of GenBank, as shown in the following (tentative) example (11):

```
FEATURES
  pcr__primer        join(1010..1029,1315..1331)
                     standard__name='hIL-7.1/8q12-13'
  STS                1010..1331
                     /map='8q12-13' 5note='A single HindIII fragment
                     of about 5 kbp was detected'
                     /label='hIL-7.1/8q12-13'/citation=[2]
```

Relational implementation of the database

GenBank is now maintained internally as a RDBMS (12) with a number of software tools specially designed for our needs. A general overview of the software architecture has been discussed elsewhere (4,13). Here, we present two particular aspects in greater detail.

Annotator's Workbench. The Annotator's Workbench (AWB) is the primary tool for browsing and editing the database. It is used by the GenBank staff to enter and annotate data submitted by researchers. Through the GOS (described above), researchers can also use it to browse the database.

The AWB displays entities from the database (e.g., sequences, papers, features, etc.) in windows or 'forms'. Typically, there are two kinds of data associated with an entity: its attributes and its links to other entities in the database. The AWB displays both types of data in the same window: attributes are directly editable (provided the user has permission), and links are displayed as descriptions of the linked entity. The user can ask to expand the link, which causes the AWB to open a new window containing the linked entity. As each link is expanded, the parent window is overlayed by the new window. When the user is finished with the expanded window, it can be closed to return to the previous level. In this way, one can explore an entity and its connections to related entities in the database.

The AWB provides several specific commands to facilitate data entry and annotation. For example, there are commands for importing sequence data and for automatically sending acknowledgements to authors upon receipt of sequence data and assignment of accession numbers. As well, there are several commands that perform correctness checks on annotated sequences and features.

Data integrity checks. One of the more important advantages gained by moving to a relational format for the GenBank database is the relative ease of performing data integrity checks. Under the current design, each data field has a strict definition and a series of checks which enforce it. Each field is assigned an abstract data type which defines the restrictions that fields of that type must meet in order to be legal; members of each of these groups may have further restrictions that apply to a subgroup or to that field alone. This means that there is a hierarchical system of checks pertinent to each field in the database. At the level of a row in a relational table, there are checks for consistency between the values of the fields. At the level of the tables, there are also checks to determine if the structure of the entire table is within bounds. Finally, at the level of the database as a whole there are statistical checks meant to determine if the current state of the database is consistent with past values. These checks are implemented in software, the Data Integrity Library (DIL), which contains the hierarchical checking functions and a master checking utility. This library is used to perform checks in three places: (i) in AWB, to enforce correct input; (ii) within the database, by batch processing, with errors saved for later examination and correction; and (iii) the flat-file-generating program.

The first method is most appropriate for field-by-field checks (e.g., that alphabetic names should not contain digits) and is implemented by calling the DIL functions from the AWB. The second method is most useful for the more complex and time-consuming checks (e.g., those that involve data from many tables). These are being implemented by means of a C library (the Data Integrity Library) containing the hierarchical checking functions and a master checking utility, which periodically scans the database for these problems and electronically mails the resulting reports to interested parties.

The third method has been implemented in the flat-file generator and an accompanying checking program. The correctness of the flat file output remains a serious matter for us, as our quarterly releases of the database continue to be made in this format. (Note, however, that we anticipate that much of the use of flat-files will shift to on-line database browsing.)

The increased data checking made possible by our having the data in relational form should enable us to provide our user community with an increasingly consistent, robust, and useful database.

Direct submission of data

The importance of the direct data submission paradigm has grown significantly over the past year. An increasing number of journals 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 import from 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' (13).

As described in earlier reports (4, 14–15), 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 80% of the data it collects as direct submissions from the community. Over 90% of our submissions are currently coming to us in electronic form. An increasing proportion (currently about 55%) of electronic submissions come in through electronic mail. (Direct submissions, and inquiries regarding them, should be directed to gb-sub@life.lanl.gov).

Receiving this amount of 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

IG released versions of the Authorin program (16) on both the IBM PC and the Macintosh during 1990 (a second release of the Macintosh version which will include additional functionality is planned for April 1991), and direct submissions generated by Authorin have begun to appear in the database. The Authorin program helps scientists annotate their data and outputs it in the proper format for rapid inclusion into the Genbank RDBMS. It helps ensure the 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).

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

Curator program

The GenBank Curator program, outlined previously (4), is now formally under way; the following scientists have begun work, and other proposals are now being brought into operation: Dr. R. Jones has begun working with us on a number of software modules, in part drawing on the power of highly-parallel hardware architectures for sequence database searches (17), that will 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 will allow us both to purge the entire existing database of such contamination and to routinely, automatically screen incoming data for similar problems. In addition, we are exploring the extension of this strategy to other computationally-intensive sequence-scanning checks that would benefit from the use of the Connection Machine at LANL that will be used for the vector scanning.

Dr. M. Berlyn is working with us to develop and implement automatic links between GenBank and the *Escherichia coli* Genetic Stock Center database (18) so that the Stock Center database will, in effect, become the master for GenBank *E. coli* nomenclature.

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 (9) to Expressed Sequence Tags (19), or 'EST'. Dr. A. Kerlavage has begun 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.

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. Today, with the community generating more than 20 million base pairs annually, the average lag time has been reduced to two weeks.

Perhaps the greatest impact of the data submissions policies lies in the area of data quality and integrity. For most scientific journals, the appearence 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 (20). This is clearly not the case. We estimate, from past scans (Gilna, unpublished results), that as many as 30% of sequences presented in journal articles contain one or more nucleotide errors. The most common source of errors resulted from the transcription steps used to create the sequence figure.

However, the fact that GenBank receives data prior to publication allows the database 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 above. Errors uncovered in the sequence data can be passed back to the author in time to have the data corrected for publication.

Some problems remain. Clearly, much of our data is available to the public before publication. However, a considerable portion of the data we receive is submitted with the proviso that the data be withheld from release until they appear in a publication. The existence of such a mechanism has been deemed important by the journals and scientific community alike. Ordinarily, we 'spot' publication 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. 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.

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

One of the projects the GenBank software team has been working on is the installation and software support of satellite databases. A satellite database is a copy of the master genbank database available to a remote site for on-line use. Changes to the database will only be made on the master version maintained at LANL, but each of the satellite databases will be updated, daily and automatically.

The local software that is used to modifiy the data base is written so that there is only one entry point — the Data Access Library — into the database. The Data Access Library records all commands that result in changes to the database using the GenBank transaction protocol.

We have developed a program that runs at both the master and satellite sites to handle the communication of these transactions. At the master site, it periodically breaks up the transaction log into numbered packets and distributes them by electronic mail to the satellite databases. At the satellite end, this daemon checks its mail for incoming packets. When a packet is received, the packet number is checked to verify that it is the next packet expected. (If a packet is received out of order, for example packet # 257 is received before packet # 256, the later packet is held aside and a request for retransmission of any missing packets is sent back to the program at the master site. Once the correct packets have been received, they are executed on the satellite database.)

By this method a satellite database will remain synchronized with the master database and data integrity at the remote site will be assured. As a result the satellite databases will never be more than three or four hours out of date with the master database and within that amount of time the satellites will be an exact reflection of the master database.

Presently, there is one satellite installed at IG. Several other experimental satellite sites are planned for the immediate future; interested institutions can contact LANL by e-mail at satellite@life.lanl.gov.

BIOSCI Bulletin board system

GOS is the distribution center in the Americas for the BIOSCI electronic newsgroup service. Currently BIOSCI consists of 20 newsgroups on a variety of topics of interest to biological scientists, including one used for discussing issues related to GenBank. Recently, newsgroups have been started for the mapping effort on human chromosome 22 and for the *Arabidopsis* genome community. 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 (21) 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 (22). 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. These factors are moving us rapidly towards a continuum of computer-based platforms for harvesting, organizing, interpreting, and distributing nucleotide sequence data (23).

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) Meth. Enz. 183, 23–31 (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. Horne, D.A., Houghton, F.K., Kelkar, P.M., Kelley, T.A., Kelly, 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. Benton, D. (1990) Recent changes in the GenBank On-line Service. *Nucl. Acids Res.*, **18**, 1517–1520
6. 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).
7. Stephens, J.C., Gilna, P., Maglott, D.R., Cavanaugh, M.L., Doute, R.C., Hutchings, G.A., Hayden, J., and Burks, C. (1989) Enhancement and expansion of the links between the GenBank and ATCC databases. *Cytogenet. Cell Genet.*, **51**, 1085–1086.
8. Stephens, J.C., Cavanaugh, M.L., Gradie, M.I., Mador, M.L., and Kidd, K.K. (1990) Mappng the human genome: current status. *Science*, **250**, 237–244.
9. Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) A common language for physical mapping of the human genome. *Science*, **245**, 1434–1435.
10. Roberts, R.J. and Olson, M.V. (1990) Editorial. *Nucl. Acids Res.*, **18**, issue #5, editorial pages.
11. Brunton,L.L. and Lupton,S.D. (1990) An STS in the human IL7 gene located at 8q12–13. *Nucl. Acids Res.* **18**, 1315–1315.
12. Date, C. (1990) *An Introduction to Database Systems, Fifth Ed.*, Addison-Wesley, Reading, MA.
13. Cinkosky, M.J., Fickett, J.W., Gilna, P., and Burks, C. (1991) Electronic data publishing and GenBank. *Science*, in press.
14. Burks, C. and Tomlinson, L.J. (1989) Submission of data to GenBank. *Proc. Natl. Acad. Sci., USA*, **86**, 408–408.
15. Gilna, P., Tomlinson, L.J., and Burks, C. (1989) Submission of nucleotide sequence data to GenBank. *J. Gen. Microbiol.*, **135**, 1779–1786.
16. Moore, J.F., Benton, D. and Burks, C. (1989) The GenBank nucleic acid data bank. *BRL Focus*, **11(4)**, 69–72.
17. 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.
18. Berlyn, M. and S. Letovsky (1990) The E. coli Genetic Stock Center database: a relational representation of genomic information. In *Biomatrix Meeting Abstracts, July 1990*, George Mason University, Vienna, VA.
19. Venter, C. and Kerlavage, A.R. (1991) Manuscript in preparation.
20. 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.
21. 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 \(rg database via Usenet. *Comp. Applic. Biosci.*, **7**, 111–112.
22. 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.
23. Burks, C. (1989) The flow of nucleotide sequence data into data banks: role and impact of large-scale sequencing projects. In *Computers and DNA*, G.I. Bell and T. Marr, Ed., Addison-Wesley, Reading, MA, pp. 35–45.

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

[2] GenBank, IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, CA 94040, USA (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, USA (e-mail, genbank@life.lanl.gov, fax 505-665-3493, telephone, 505-665-2177).