NUCLEOTIDE SEQUENCES 1986/1987

VOLUME VIII

DATABASE DIRECTORY AND MASTER INDICES

A compilation from the
GenBank®
and
EMBL
data libraries

NUCLEOTIDE SEQUENCES 1986/1987

VOLUME VIII

DATABASE DIRECTORY AND MASTER INDICES

A Compilation from the
GenBank®
and
EMBL
data libraries

Compiled by

Edwin J. Atencio,* Howard S. Bilofsky,† June Bossinger,† Christian Burks,* Graham N. Cameron,‡ Michael J. Cinkosky,* Carol E. England,* Victor I. Esekogwu,* James W. Fickett,* Brian T. Foley,* Walter B. Goad,* Gregory H. Hamm,‡ David J. Hazledine,‡ Patricia Kahn,‡ Leslie Kay,* Frances I. Lewitter,† Natalie Lopez,* Kersti A. MacInnes,* Mia J. McLeod,* Deborah L. Melone,† Gerald Myers,* Debra Nelson,* Judith L. Nial,‡ Joanna K. Norman,* Eric D. Rasmussen,* Andrea A. Revels,* Wayne P. Rindone,† Carol R. Schermer,* Maura T. Smith,* Guenter Stoesser,‡ C. David Swindell,† Brian L. Trujillo,* and Chang-Shung Tung*

* *GenBank*
T-10 Mail Stop K710
Los Alamos National Laboratory (LANL)
Los Alamos, New Mexico 87545

† *GenBank*
BBN Laboratories Incorporated
10 Moulton Street
Cambridge, Massachusetts 02238

‡ *EMBL Nucleotide Sequence Data Library*
European Molecular Biology Laboratory
Postfach 10 22 09
D-6900 Heidelberg
Federal Republic of Germany

1987

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
Orlando San Diego New York Austin
Boston London Sydney Tokyo Toronto

ACADEMIC PRESS, INC.
Orlando, Florida 32887

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
24–28 Oval Road, London NW1 7DX

By purchasing or otherwise obtaining *Nucleotide Sequences 1986/1987*, recipient understands that the information contained in this compendium, which has been produced from the information contained in the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library and the GenBank® database ("the information"), has come from a variety of sources, published and perhaps unpublished. The information has been deposited in the GenBank® database and the EMBL Nucleotide Sequence Data Library, and it has been reproduced for inclusion in this compendium via a reliable and quality controlled procedure, but no such process is infallible. Therefore, Academic Press, Inc. (AP), Bolt Beranek and Newman Inc. (BBN), Los Alamos National Laboratory (LANL), The European Molecular Biology Laboratory (EMBL), and the U.S. Government make no representations or warranties regarding the content or accuracy of the information. By way of example, but not of limitation, AP, BBN, LANL, EMBL, and the U.S. Government make no representation or warranties of merchantability or fitness for a particular purpose, or that the use of the information will not infringe any patent, copyright, trade secret, or trademark of any third person. AP, BBN, LANL, EMBL, and the U.S. Government accept no responsibility for any expenses, losses, or action incurred or undertaken by the recipient as a result of the receipt or use of the information.

Note that GenBank® is a registered trademark for the Genetic Sequence Data Bank established by BBN and LANL under contract with the U.S. National Institutes of Health and should be used only in that context.

Information from this compendium may be duplicated, reproduced, or otherwise used by the recipient, but in no event may the GenBank® trademark be associated with such re-generated information, and in no event shall there be any remedy furnished by AP, BBN, LANL, EMBL, or the U.S. Government for such re-generated information, including but not limited to financial remuneration or technical interaction.

Please note that the proper attribution of *Nucleotide Sequences 1986/1987* as the source of your data and the public availability of this information in computer-readable form from BBN and EMBL will be appreciated.

Library of Congress Cataloging in Publication Data

Nucleotide sequences 1986/1987.

Includes indexes.
Contents: v. 1. Primates – v. 2. Rodents – v. 3. Other vertebrates and invertebrates – [etc.]
1. Nucleotide sequence–Tables–Collected works. I. Atencio, Edwin J. II. GenBank. III. Europaeiske molekylaerbiologiske laboratorium. IV. BBN Laboratories. V. Los Alamos National Laboratory.
QP 625.N89N85 1987 547.7'9 87-1782
ISBN 0–12–512518–6 (v. 8 : alk. paper)

PRINTED IN THE UNITED STATES OF AMERICA

87 88 89 90 9 8 7 6 5 4 3 2 1

NUCLEOTIDE SEQUENCES 1986/1987

VOLUME VIII

DATABASE DIRECTORY AND MASTER INDICES

Academic Press Rapid Manuscript Reproduction

Contents

VOLUME VIII
DATABASE DIRECTORY AND MASTER INDICES

SEQUENCE ENTRIES FOUND IN VOLUMES I, II, III, IV, V, VI, and VII

Preface

This eight-volume compendium of nucleotide sequences found in the GenBank and EMBL databases is the third edition resulting from the combined efforts of all of the technical and administrative staff at Los Alamos National Laboratory, the European Molecular Biology Laboratory, and BBN Laboratories Incorporated listed on the title page. Both the EMBL and GenBank databases have continued to grow at a remarkable rate, with each database doubling in size nearly once each year. We have organized this compendium in eight self-contained volumes, each of which is available separately. The first seven volumes each contain the same introductory and explanatory material, one or more sections of sequence entries, and several indices to the entries in that volume. Volume VIII contains a database directory and master indices to all of the volumes.

As a result of comments and suggestions we received in response to the previous edition, we have made several improvements in this edition. We hope that some slight adjustments in the layout and presentation of the sequence entries, including increasing use of mixed-case text and improvements in punctuation, will result in making them more easily readable than in the past.

Both databases are available in a variety of computer-readable forms. Additional information about obtaining the GenBank database can be obtained by writing to

Genbank
BBN Laboratories Incorporated
10 Moulton Street
Cambridge, Massachusetts 02238
USA

Further information about the EMBL Nucleotide Sequence Data Library can be obtained by writing to

EMBL Nucleotide Sequence Data Library
European Molecular Biology Laboratory
Postfach 10 22 09
D-6900 Heidelberg
Federal Republic of Germany

Wayne P. Rindone
Cambridge, Massachusetts
November 17, 1986

Introduction

Outline

1. Introduction

Nucleotide Sequences 1986/1987 is the third database compendium published as one result of a unique international collaboration between two leading nucleotide sequence data libraries, one based in the United States and one in Europe. The two databases are the EMBL Nucleotide Sequence Data Library, established by the European Molecular Biology Laboratory (EMBL), and the GenBank(R) Genetic Sequence Data Bank, which is a U.S. Government-sponsored nucleic acid sequence repository. Both databases serve molecular biologists and other investigators worldwide by collecting the large number of reported DNA and RNA sequences and making them available in computer-readable form. The primary distribution medium for both databases is magnetic tape.

The data in the compendium reflect the information found in GenBank Release 44.0 of August 1986. This information has been combined with the data included in EMBL Release 8.0, which was made available in May 1986. Regularly updated distribution tapes containing the EMBL Sequence Data Library are available four times annually. A new set of distribution tapes containing the entire GenBank database is also made available four times annually, and update tapes containing only entries that have been added or changed are available midway between each full GenBank release.

The sequences in this compendium are also available from GenBank on floppy diskettes. Because of limited storage capacity, only the sequences, some basic identifying information, and some of the biological annotations are included on this distribution medium. The remaining annotated information can be found in the compendium.

The GenBank database is available online on the DRR/NIH/PROPHET computer system, which can be accessed over Telenet, an international telecommunications network. The online database is updated every six weeks on the same schedule as the magnetic tape releases. This online service also provides users with access to the GenBank Software Clearinghouse, which contains information about commercially available software packages for analyzing and manipulating sequences.

For more information on the services provided by the GenBank and EMBL sequence libraries, please write:

GenBank
BBN Laboratories Inc.
10 Moulton St.
Cambridge, MA 02238
USA

European Molecular Biology Laboratory
Nucleotide Sequence Data Library
Postfach 10.2209
D-6900 Heidelberg
West Germany

1.1 Description of the compendium

The printed compendium makes the entire collection of information in both databases available to every member of the scientific community who wishes to use it, including investigators without access to computers. This compendium, drawn from the American and European databases, is the third printed compilation of substantially all nucleic acid sequences reported since 1967. These sequences and their associated annotations have been compiled from the published literature and from direct submissions from the authors by the GenBank staff at Los Alamos National Laboratory and by the EMBL data library staff at EMBL.

Although the format chosen for entries in the printed compendium differs somewhat from that in either database, every entry contains information contributed both by EMBL and by GenBank. The final preparation of the data in the compendium was carried out by the GenBank staff at BBN Laboratories Incorporated (BBN); therefore, the format and conventions used in the compendium are somewhat closer to those used in the GenBank database than to those used in the EMBL data library. Technical Appendix E illustrates how the compendium format relates to the formats used in the two databases from which it was constructed. One of the goals of the collaboration between GenBank and EMBL is continued movement toward common standards and conventions for the two databases.

1.2 The two databases

The EMBL Nucleotide Sequence Data Library was established in 1980 by the European Molecular Biology Laboratory, an international center of fundamental research with its main emphasis in the fields of cell biology, molecular structures, differentiation, and instrumentation. EMBL, whose headquarters is in Heidelberg, Germany, is currently funded by the following member states: Austria, Denmark, France, Federal Republic of Germany, Finland, Greece, Israel, Italy, the Netherlands, Norway, Spain, Sweden, Switzerland, and the United Kingdom. The first release of the EMBL data library was in April 1982.

The GenBank database was created in 1982 by the National Institute of General Medical Sciences (NIGMS) of the U.S. National Institutes of Health (NIH). Los Alamos National Laboratory (LANL), which is operated by the University of California for the Department of Energy, is located in Los Alamos, New Mexico. LANL gathers, annotates, and organizes the database and transmits it to BBN Laboratories Incorporated, a research and consulting firm in Cambridge, Massachusetts. The collected information is prepared for release by BBN and distributed to subscribing institutions and scientists in regular updates. Cosponsors of the GenBank project include the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, the National Library of Medicine, the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, and the Division of Research Resources (DRR) of NIH, as well as the National Science Foundation, the U.S. Department of Energy, and the U.S. Department of Defense. GenBank's first release was in October 1982.

1.3 New features of this edition

- The Citation Index has been added to assist readers in finding bibliographical citations for journal articles. This new index lists journal title, volume number, page numbers, and year of publication for each article cited.

- As a result of limited resources and an ever-increasing rate of sequence publication, it has not been possible to collect and present all sequences in the fully annotated form that we would like. It is nevertheless vitally important

that at least as much raw sequence data as possible be presented. Therefore, we have a new section entitled Unannotated Sequences, which contains unannotated and unclassified sequences and citations. We hope that in the future we will have the resources to move this information rapidly into its proper position in the main database.

- A separate volume is now available that contains master indices for the entire database as well as a master directory for all of the entries in the database.

2. Contents of the compendium

As combined in this compendium, the two databases contain a total of nearly 8.5 million bases from 6700 articles. The following indices are provided to assist users in finding the information they need: the Keyword Phrase Index, the Taxonomic Classification Index, the Author Index, the Citation Index, the Accession Number Index, the EMBL Entry Index, and the GenBank Entry Index. Most of the entries are annotated to indicate the locations within the reported sequences of coding regions and other experimentally determined sites of biological significance. Full bibliographic information is included in every entry, and many of the entries also include comments abstracted from the original papers. Technical appendices located after the main data sections in each volume contain detailed explanations of information in the entries.

2.1 General organization of the compendium

The entries in the compendium are presented in thirteen sections; within each section the entries are grouped according to the source organism. These sections are arranged in eight volumes, as follows:

Volume I. Primates

Section 1. Primate Sequences

Volume II. Rodents

Section 2. Rodent Sequences

Volume III. Other Vertebrates and Invertebrates

Section 3. Other Mammalian Sequences
Section 4. Other Vertebrate Sequences
Section 5. Invertebrate Sequences

Volume IV. Plants and Organelles

Section 6. Plant Sequences
Section 7. Organelle Sequences

Volume V. Bacteria and Bacteriophage

Section 8. Bacterial Sequences
Section 9. Bacteriophage Sequences

Volume VI. Viruses

Section 10. Viral Sequences

Volume VII. Structural RNA, Synthetic, and Unannotated Sequences

Section 11. Structural RNA Sequences
Section 12. Synthetic Sequences
Section 13. Unannotated Sequences

Volume VIII. Database Directory and Master Indices

Each volume of the compendium contains this introduction, one or more sections of data, technical appendices, and indices to that volume. The Author Index, the Citation Index, the Taxonomic Classification *Index, the Keyword Phrase Index, the Accession Number* Index, the EMBL Entry Index, and the GenBank Entry Index in Volume VIII are master indices to all of the volumes in this edition.

2.2 Finding an entry

Users approaching the database for the first time must determine which section contains the sequence they are looking for. Most of the sections are self-explanatory, but it is helpful to point out the following conventions:

- Yeast and fungal sequences are in the Plant Sequences section.
- Plasmids and transposons isolated from bacteria are listed in the Bacterial Sequences section.
- The Structural RNA section includes the sequences of mature transfer RNA, ribosomal RNA, small nuclear RNA, and other structural RNA molecules. All structural RNA genes and most structural RNA precursor sequences are listed with their organisms in their particular sections.
- The Synthetic Sequences section includes any nucleic acid sequence that is created in a laboratory and does not occur naturally, including synthetic plasmids that are not included with the other bacterial sequences. The major exceptions to this rule are cDNA sequences, since they are regarded as a means of sequencing naturally occurring RNA sequences.

The individual entries within each section are arranged alphabetically by entry name. Summary tables and section directories are included at the beginning of each section to provide some guidance for locating the entries. Table 1 is an overall summary table of the entire database. This table shows the names of the sections, as well as the numbers of reported sequences, distinct entries, and nucleotide bases in each section. There are typically more reported sequences than entries because overlapping sequences are frequently merged into a single, combined entry.

A table that summarizes the entries appears at the beginning of each section. This table is called the Section Summary. The Section Summary for the Primate Sequences section, for example, lists, by organism (e.g., Ape), the corresponding organism code (e.g., APE), the number of reported sequences for that organism, the number of entries, the number of bases, and the page number on which this group of entries begins.

Note that the page numbers throughout are arranged separately for each section. The numbers are printed on each page with a short section prefix. For example, the first three pages of Section 1: Primate Sequences are numbered PRIMATE-1, PRIMATE-2, and PRIMATE-3. Table 1 shows the page number prefix for each section.

A detailed alphabetized directory for the section appears immediately after the Section Summary. The section directory contains one line of information for each entry in the section and serves as a complete table of contents for that section, listing the full entry name, the description and length of each entry (i.e., number of base pairs), and the page on which each entry appears.

3. How to Read an Entry

The entries for each section begin after the section directory. Each entry is separated from the next by a dashed line running the width of the page. There are two types of entries in the compendium: (1) self-contained, and (2) segmented. Segmented entries are used when noncontiguous pieces of the same nucleic acid molecule have been sequenced and the ordering of the pieces is known.

Table 1: Summary of Sequences Presented in Each Section

Section Number	Section Code	Section Description	Number of Sequences	Number of Entries	Number of Bases
1	PRIMATE	Primate Sequences	1492	1028	1240779
2	RODENT	Rodent Sequences	1638	1272	1111622
3	MAMMAL	Other Mammalian Sequences	293	245	244554
4	VERT	Other Vertebrate Sequences	557	474	400509
5	INVERT	Invertebrate Sequences	696	605	435280
6	PLANT	Plant Sequences	717	594	643365
7	ORGANELLE	Organelle Sequences	434	368	485666
8	BACT	Bacterial Sequences	1130	749	1031546
9	PHAGE	Bacteriophage Sequences	338	160	271817
10	VIRAL	Viral Sequences	1748	1093	1517025
11	RNA	Structural RNA Sequences	734	637	69232
12	SYNTHETIC	Synthetic Sequences	259	224	72029
13	UNANNOTATED	Unannotated Sequences	1377	1374	918933
		Overall Summary:	11413	8823	8442357

3.1 Summary of the entry fields

Each entry is composed of several kinds of information, referred to here as fields. Not every field appears in every entry, but the full list of possible fields, in the order in which they appear, is as follows:

- Entry Name – a short, unique name providing the label for the entry.
- Definition – a brief description of the sequence, beginning with the name of the source organism.
- Segment – indicates which segment this entry is in a series of separated sequences from the same molecule.
- EMBL ID – entry name(s) in the EMBL database that correspond to the entry names in this work.
- Accession Numbers – short codes that provide unique, unchanging identifiers for the data in each entry; the first number in the list is known as the primary accession number of the entry.
- Date – the year, month, and day when this form of the entry appeared in the GenBank version of the database, plus information on whether the entry is preliminary or complete.
- References – citations for all references used to construct each entry.
- Keywords – short phrases describing gene products and other information pertinent to looking up an entry.
- Source – most commonly used name of the source organism, followed by a formal scientific name.
- Comment – information that does not readily fall into the other fields, including information abstracted from the references and cross-references to other entries.
- Features and Sites Tables – tables designed to describe locations and regions of biological significance within the sequence.
- Origin – describes the start of a sequence in relation to an experimentally determined site.
- Sequence – statistics on the numbers and kinds of bases in the sequence, followed by the sequence itself.

See Example 1 for an example of a typical pair of entries.

3.2 The fields in detail

ENTRY NAME
EMBL "ID" Names and GenBank "Locus" Names

The entry name is a short, unique name that provides the label for an entry. In order to organize this compendium in a coherent fashion, it was necessary to choose a uniform method for naming all of the entries, regardless of which database the information was extracted from. By mutual agreement, we have presented the entries under the names assigned to them in the GenBank database. The conventions for choosing these names, which include abbreviations for the organisms from which the nucleic acids were isolated, are described in detail in Technical Appendix A: Entry Name and Molecule Type Conventions.

The GenBank entry names have been called "locus names" throughout this book, and there are many occasions where one entry refers to another "locus" or another group of "loci"; this terminology is simply a way of referring to other entries. The entry names used for the corresponding information in the EMBL Sequence Library are given after the label "EMBL ID:" in the second line of each entry. Not all entries have been assigned EMBL ID names at this stage of our collaboration, but eventually all entries will be assigned names in both databases, and we are actively moving toward a common naming system for corresponding entries in the two databases.

The GenBank Entry Index lists all of the GenBank entry names alphabetically, together with the section name and page number on which the entry begins. The other indices refer to GenBank entry names, not page numbers, since these are the names used in organizing the book. The page numbers must be looked up in the GenBank Entry Index.

DEFINITION

The definition of an entry provides a brief description of the sequence. This definition is used to construct the listing for the entry in the section directory. Typically it includes the name of the organism and other important information describing the entry. Information about the type of molecule and whether the sequence presented is circular or a complete tandem repeat is included in brackets at the end of the definition for most entries. The conventions used in specifying the molecule type are described in detail in Technical Appendix A: Entry Name and Molecule Type Conventions.

```
ANIMTCYB1:  a.nidulans mt apocytochrome b (coba) gene; exon1. [DNA]     SEGMENT: 1 of 2
 EMBL ID: MIAN02                          ACCESSION NUMBERS:  J01388 V00651                DATE:  updated   83-11-01
 REFERENCES:     [1] (bases 1 to 838) Waring,R.B., Davies,R.W., Lee,S., Grisi,E., Berks,M.M.  and Scazzocchio,C.; "the
          mosaic organization of the apocytochrome b gene of aspergillus nidulans revealed by dna sequencing"; Cell
          27, 4-11 (1981)
 KEYWORDS:  cytochrome; apocytochrome.
 SOURCE:    aspergillus nidulans.     Mitochondrion Aspergillus nidulans
 COMMENT:   Single intron of about 1050 bp occupies same position as I3 in "long" S. cerevisiae gene.  Open reading
          frame of exon 1 continues at least 200bp into ivs.  TGA codes for trp.  See ‹hummt› and ‹ystmtcyb›.  See
          other loci beginning ‹animtcyb›.
 SITES:                                                    FEATURES:
 key        site span   description                       key        from     to   description
 refnumbr     1    1    numbered -125 in [1]; zero not used. pept       126 +   631   apocytochrome b (exon 1)
 -›pept     126    1    coba coding sequence start        FEATURES:
 pept/IVS   632    0    coba ivsl start (exonl end)       key        from     to       description
                                                          CDS         126    631       apocytochrome b part 1 (631 is
                                                                                       2nd base in codon)
                                                          IVS         632   ›838       intron I
 ORIGIN:    near hind iii site in bgl ii fragment 4.
 SEQUENCE:    838 bp   320 a   112 c   132 g   274 t
       1 atataaaaca gtaattaata aataaaataa ttactttaat ctttagattt  ttaaatctga taaataaaaa aaaaaaaata aataaataaa ttaagtgaag
     101 aaaaaaaaaa ataaaaaaat aaaaaatgag aattttaaaa agtcatcctt  tactaaaaat agtaaattcg tatataatag attcacctca accagctaat
     201 ttaagttatt tatgaaattt cggatcatta ttagctttat gtttaggtat  acaaatagta acaggtgtta cattagctat gcattataca cctagtgtat
     301 cagaagcatt taattctgta gagcatatta tgagagatgt aaataatcca  tgattagtac gttacttaca ctctaataca gcttcagctt tcttcttttt
     401 agtatactta cacataggaa gaggtttata ttatggatct tacaaaacac  ctagaacttt aacatgagct attggaacag taatactaat agttatgatg
     501 gccacagcct tcttaggtta tgttttacct tatggtcaaa tgagtttatg  aggtgctaca gttattacta acctaatgag tgctatacct tgaataggtc
     601 aagatattgt tgagtttatt tgaggaggtt tatacacaga tgaaccacaa  tgcggtgacg tattgttaaa aatcctgctt aatgctggaa aatccccaat
     701 cttaggattt gcatacgact tattctttat aatagtatta ttaataggcg  tgaaaattgc aatgacacgg ggaaaatcag caggggtgag aagtttacat
     801 acttcagaag cctctcagag actacatgca ggagatct
-------------------------------------------------------------------------------------------------------------------
ANIMTCYB2:  a.nidulans mt apocytochrome b (coba) gene; exon2. [DNA]     SEGMENT: 2 of 2
 EMBL ID: MIAN03                          ACCESSION NUMBERS:  J01389 V00652                DATE:  updated   83-11-01
 REFERENCES:     [1] (bases 1 to 1082) Waring,R.B., Davies,R.W., Lee,S., Grisi,E., Berks,M.M.  and Scazzocchio,C.; "the
          mosaic organization of the apocytochrome b gene of aspergillus nidulans revealed by dna sequencing"; Cell
          27, 4-11 (1981)
 KEYWORDS:  cytochrome; apocytochrome.
 SOURCE:    aspergillus nidulans.     Mitochondrion Aspergillus nidulans
 COMMENT:   Single intron of about 1050 bp occupies same position as I3 in "long" S. cerevisiae gene.  Open reading
          frame of exon 1 continues at least 200bp into ivs.  TGA codes for trp.  See ‹hummt› and ‹ystmtcyb›.  See
          other loci beginning ‹animtcyb›.
 SITES:                                                    FEATURES:
 key        site span   description                       key        from     to   description
 IVS/pept     77    0    coba exon2 start (ivsl end)       pept     +    77    734   apocytochrome b (exon 2)
 pept‹-     734    1    coba coding sequence end          FEATURES:
                                                          key        from     to       description
                                                          CDS          77    731       apocytochrome b part 2 (77 is
                                                                                       3rd base in codon)
                                                          IVS          ‹1     76       intron I
 ORIGIN:    about 750 bp after animtcybl
 SEQUENCE:    1082 bp   373 a   123 c   140 g   446 t
       1 gatcaataaa gaaatttatt gcgtatagta agaggattta atatttatat  taaatctgta actatcaaca taaatgctct gtaaataatg caactttaaa
     101 cagattcttt gcattacatt tcttattacc ttttgtatta gctgctttag  cattaatgca tttaatagct atgcatgata cagtaggatc aggtaatcct
     201 ttaggtattt ctgctaatta cgatagatta ccttttgctc cttattttat  atttaaagat ttaataacta tatttatatt ctttattgta ttatcaatat
     301 ttgttttctt tatgcctaat gctttaggtg atagtgaaaa ttatgttatg  gctaatccta tgcaaactcc acctgctata gttccagaat gatatctttt
     401 acctttctat gctattttaa gatctatacc taataaatta ttaggtgtta  tagctatgtt tgctgctata ttagcattaa tggttatgcc tataactgat
     501 ttatctaaat taagaggagt acaatttaga cctttaagta aagtagtatt  ctatattttt gtagctaact tcttaatatt aatgcaaata ggtgcaaaac
     601 acgttgaaac tccatttatt gaatttggac aaatttctac tattatttat  tttgcatatt tctttgtaat agttcctgtt gttagtttaa ttgaaaatac
     701 tttagtagaa ttaggaacta aaaaaaactt ttaattctta gtcctcttag  gaaaaaaaaa caaatttatt aaaacagtcg aaatttaatt tatgaaaatg
     801 atattagaca aaaaatttta aaaagaatta gatagctaca tttgattata  atcaatttat taatattttg gttttcatct atactttgta gttaatcata
     901 agtatgatgt aataaatagt aatatctttt taaagtagac ttgaccttta  aaatttttaa tataattatt attatcttgt tagagtataa ttaaatacaa
    1001 tataatattg tatattagga gtttgagcaa atggttttgc gttttgattg  caaattgaaa tatagggatt cgatttcccc gg
-------------------------------------------------------------------------------------------------------------------
```

Example 1. Two segmented entries from the Organelle Sequences section.

SEGMENT

In those cases where an entry is segmented, a segment field is used to indicate which segment this entry is in a series of noncontiguous sequences from the same molecule. Segmented entries contain a label after the molecule type, at the end of the first line of the entry, which indicates the position of this entry in the group of segmented entries. The number at the end of the entry name also indicates which segment of a complete sequence is contained in this entry.

ACCESSION NUMBERS

Accession numbers have been assigned to all entries in the EMBL Sequence Library and the GenBank database, using a system that was worked out jointly by the two database teams. These arbitrary labels for the data included in an entry consist of a single letter followed by five digits; unlike the entry names, they carry no information about the type or nature of the entry. Accession numbers never change, but rather follow along with the data they point to no matter how the entry or entries in question might be reorganized in future versions of either database. For example, if two different entries with different accession numbers are merged into a single entry, both accession numbers are included in the new entry.

The first number in the list of accession numbers is the primary accession number for the entry. If you cite information contained in an entry in either database, you are encouraged to include the primary accession number in your citation, since this number will enable your readers to find the data in question in future releases of either database. The Accession Number Index lists all accession numbers that have been assigned to date and the entries to which they have been assigned.

DATE

The date field contains a date in the form year-month-day, preceded by the word "entered", "updated", "pre-entry", or "unannotated". The date given is the date of the most recent GenBank release in which this entry underwent any major revisions. If the word "entered" appears before the date, this means that this entry was first entered in the database on the date given, and that it has not undergone any substantial revisions since that release. If there has been some substantial revision, such as the addition of another reference, the date of the most recent revision is given preceded by the word "updated." If the word "pre-entry" appears, it indicates that although the complete sequence and some portion of the annotations appear in this entry, there are additional annotations that will be included once review of the article in question is completed. Pre-entries sometimes undergo several rounds of updates and revisions before they are upgraded to full entries. The date field in all of the entries in the Unannotated Sequences section begins with the word "unannotated".

REFERENCES

This field includes the number (in brackets) assigned to each cited paper; a brief statement of which information in the entry comes from each reference (shown in parentheses); and the actual citation of the article. Sequence data submitted directly to the EMBL or GenBank databases and not published elsewhere are listed as unpublished references. These references generally have no title; they simply list the contributor as the author and include the word "Unpublished" followed by the year in parentheses and the address of the contributor. See Technical Appendix B: Reference Citation Conventions for further information on reference format.

KEYWORDS

The keywords field contains short words or phrases that identify gene products and other useful identifying characteristics of the entries. The Keyword Phrase Index provides a means for looking up all entries that have common keyword phrases; if a particular entry has no keyword phrases, this indicates that none of the phrases in the keyword index apply to it. A small number of pre-entries show the word "unassigned" in place of any keyword phrases; this word indicates that the entry has not yet been reviewed to determine which, if any, of the keyword phrases apply to it.

SOURCE

The source field is an English language statement about the organism and tissue from which the sequence was isolated. If the entry contains a viral sequence, the host organism is usually also named. This statement is followed by a formal designation of the source organism, typically consisting of the full scientific name. The Taxonomic Classification Index lists all entries according to their formal taxonomic classifications.

COMMENT

The comment includes information that does not readily fall in other fields, such as statements abstracted from the references and cross-references to other entries.

FEATURES AND SITES TABLES

Three different tables can appear in each compendium entry: the EMBL features table, the GenBank features table, and the sites table. All three are designed to describe regions and locations of biological significance within the sequence. The two features tables show regions, with starting and ending points for each feature of interest. The sites table, on the other hand, shows individual locations of interest within the sequence, together with a number that indicates whether the location is a single point or encompasses multiple bases. The features tables presented come from both databases; those with lowercase keywords in the key column are in GenBank format, while those with uppercase keywords are in EMBL format. The EMBL features tables are included in those cases where the numbering system used in the EMBL entry corresponds to the numbering used in the GenBank entry and the EMBL table provides some information that augments the information found in the other two tables. There is often considerable redundancy in the information contained in these three tables. Full information about the conventions used in constructing these tables appears in Technical Appendix C: Sites and Features Tables.

ORIGIN

The origin field describes the start of the sequence (the 5' terminus) in relation to some experimentally determined site, such as a restriction enzyme cutting site.

SEQUENCE

The first line in the sequence field gives the total number of base pairs, and the number of adenines, cytosines, guanines, thymines (or uracils), and other bases reported in the sequence. After this information, the actual sequence is listed. The bases in the sequence are numbered line by line, and they are presented one hundred bases per line, in groups of ten.

4. Two Sample Entries

Example 1 is a reproduction of a typical pair of segmented entries found in the Organelle Sequences section. In this example, the first entry consists of the 5' portion of a particular sequence and the second consists of the 3' portion of the same sequence. The two entries in this excerpt have only been partly converted to full mixed-case representation. When there are free text portions of an entry that are not converted to mixed case, they are shown in lower-case characters. The common prefix for the two entry names is ANIMTCYB; therefore, the first entry is called ANIMTCYB1 and the second ANIMTCYB2. The segment field at the end of the first line of each of these entries states explicitly that the first entry is the first of two segments and the next is the second of two segments.

The entry names in this example consist of several parts. The first three characters in both names are "ANI", an abbreviation for the fungus Aspergillus nidulans. The letters "MT" indicate that these are mitochondrial sequence entries, and "CYB" is an abbreviation for apocytochrome b. The last character in each entry name is the segment number. See Technical Appendix A for a description of the conventions and abbreviations used in assigning entry names.

The definition following the entry name gives the name of the source organism and other information describing the sequence. For example, "a. nidulans mt apocytochrome b (coba) gene; exon 1" indicates that the entry called ANIMTCYB1 contains exon 1 of the mitochondrial coba gene coding for apocytochrome b in Aspergillus nidulans. The notation "[DNA]" immediately following the definition indicates that the sequence represents a double-stranded DNA molecule.

On the second line of each entry, the corresponding EMBL ID names are given. The entry called ANIMTCYB1 corresponds to the EMBL entry called MIAN02, and ANIMTCYB2 corresponds to MIAN03. The accession numbers appear next. Each of these entries has two accession numbers, since these particular entries were originally entered independently in each database. The date field at the end of the second line indicates that these entries were most recently revised in the GenBank release dated 1 November 1983.

The list of references begins on the third line of these entries. The notation "[1] (bases 1 to 838)" indicates that the reference in the first entry is referred to as reference [1] and that it is the source of the bases numbered from 1 to 838 in the entry. The remainder of the reference listing is a fairly conventional citation for the particular reference.

The keywords field appears next. These two entries can be looked up using the two keywords "cytochrome" and "apocytochrome" in the Keyword Phrase Index. The next field, the source, lists first the commonly used name of the source organism, Aspergillus nidulans, followed by the scientific name used to classify the organism in the taxonomic classification index, Mitochondrion Aspergillus nidulans. In many cases, such as those in this example, the two parts of the source field are somewhat redundant.

The comment, beginning on the next line of the entry, gives brief information about the entry (taken from the reference) and refers the reader to other loci (entries) that have names beginning "ANIMTCYB".

Three tables appear next in each of these entries: a sites table and two features tables. All three of these tables indicate the portions of these sequences that code for apocytochrome b; this includes the bases numbered from 126 to 131 in ANIMTCYB1 and bases 77 to 734 in ANIMTCYB2. The EMBL features tables were included in these two entries since they explicitly state the reading frame for the codons in these regions. These two entries also illustrate the use of some of the special symbols that can appear in both styles of features tables. The "+" signs in the GenBank features tables indicate that the coding region resumes in another segment. The "<" and ">" characters in the EMBL features tables indicate that the intervening sequence extends beyond the ends of the reported sequences. The apparent differences in the numbers reported in the three tables in these entries are due to systematic differences in the conventions that have been used in the EMBL and GenBank databases. For example, the GenBank features table reports that the coding sequence in ANIMTCYB2 terminates at base 734, while the corresponding EMBL features table reports the end of the coding region as base 731. This apparent difference merely reflects the GenBank convention of including the termination codon in the reported coding region and the EMBL convention of excluding the termination codon.

The origin lines, which follow the features and sites tables in these entries, indicate that the sequence in ANIMTCYB1 starts near a particular restriction enzyme cut site and that the sequence in ANIMTCYB2 is separated from that in ANIMTCYB1 by approximately 750 bases.

The sequence field is the last field in the entry. The first line of the first entry shows that there are 838 base pairs in this sequence, which includes 320 adenines, 112 cytosines, 132 guanines, and 274 thymines. Finally, the actual sequence is listed and clearly numbered.

Please refer to the technical appendices after the data sections in each volume for additional details about the conventions used in presenting the entries in this collection. The indices at the end of each volume include brief explanations and instructions for their use.

Section 1: Primate Sequences