Release 110, December 2011

EMBL Outstation
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD
United Kingdom

Telephone: +44-1223-494499
Telefax : +44-1223-494468

Electronic mail: datasubs@ebi.ac.uk
URL: http://www.ebi.ac.uk/ena

Full version 

Text version 

This manual and the database it accompanies may be copied and redistributed freely, without advance permission, provided that this statement is reproduced with each copy.

CONTENTS

1 INTRODUCTION

This document describes the format and conventions used in EMBL-Bank. An attempt has been made to make the collected data as easily accessible as possible without restricting their usefulness to any particular type of computing environment. For this reason, the simplest possible organisation ("flat file") has been chosen.

The main body of this User Manual describes the features of the database which will remain stable, such as the flat file format and the use of line types to distinguish different kinds of information. Features of the database more likely to require change (such as journal abbreviations) are described in the appendices. Information which applies specifically to the current release of the database is presented in the Release Notes. The Release Notes also describe changes which are foreseen in future releases.

It is likely that the need to represent new kinds of information in the database will eventually necessitate changes or additions to the presentation of data.

Such changes will be made as far as possible in ways which have minimal impact on user programs and procedures. For example, a new type of data could be added to the database as a new line type (see Section 3) without affecting the processing of existing line types.

We would like to stress that both this manual and the database itself are free from any copyright restrictions (please see the statement on the title page). While we would appreciate acknowledgement if our efforts have been useful to you, we want to ensure that the data are freely available to anyone interested.

2 CONVENTIONS USED IN THE DATABASE

This section describes the general conventions which have been applied to The information in the database in order to achieve uniformity of presentation.

Specific abbreviations and symbol usage are summarized in the appendices.

2.1 Sequence Data

Nucleotide sequence data are generally presented in the database as they have been submitted or published, subject to certain conventions which have been established for the database as a whole. The sequences are always listed in the 5' to 3' direction, regardless of the published order. Bases are numbered sequentially beginning with 1 at the 5' end of the sequence.

The sequences are presented in the database in a form corresponding to the biological state of the information in vivo. Thus, cDNA sequences are stored in the database as RNA sequences, even though they usually appear in the literature as DNA. For genomic data, the coding strand is stored. Data containing coding sequences on both strands are stored according to the prevailing conventions in the literature. The stored data generally correspond to wild type sequences before mutation or genetic manipulation.

Sequences of tRNA molecules are stored as unmodified RNA sequences (equivalent to the mature transcript before any base modification occurs). This form (colinear with the genomic sequence) has been adopted to simplify both storage and analysis of the sequences. Thus, a modified base appears in the sequence as the corresponding unmodified base. However, each base modification is noted in the feature table, so that the mature tRNA sequence can be restored automatically by a simple computer program if this is desirable. The two-letter code used by Sprinzl and Gauss has been adopted for abbreviation of modified bases in the feature table.

2.2 Organism Identification and Classification

A unified taxonomy is used by the collaborating databases (DDBJ/EMBL-Bank/GenBank). Based on the NCBI's 'Taxon' project, this constitutes a taxonomy database which reflects current phylogenetic knowledge. It is a sequence-based taxonomy as far as possible, and is based upon published authorities wherever appropriate. Deciding criteria include a variety of physiological, ecological, morphological

characters, overall morphological similarity and common descent.
Evolutionary taxonomists tend to consider both overall similarity
and common descent when making and assigning a classification while
phylogeneticists attempt to reflect the branching pattern of the underlying
phylogenetic tree. There is of course no such thing as a single best method for
classifying organisms and the choice of one system over the other has to be made
with regard to the particular purpose of the classification. Because of the
inherent ambiguity of evolutionary classification and the specific needs of
database users (e.g. trying to track down the phylogenetic history of a group
of organisms or to elucidate the evolution of a molecule), the taxonomy strives
to reflect accurately current phylogenetic knowledge.
One of the major sources for classification are phylogenetic insights derived
from molecular evolution studies. New taxonomic information is included
as soon as it becomes available, but at the same time, efforts are made to
ensure that the arguments and evidence provided are reliable in order to avoid
frequent (and possibly unnecessary) changes to the classification system.
The OS/OC lines of all entries reflect the up to date taxonomic classification.
This classification is intended to be informative and helpful; no claim is
made that it is necessarily the best or most exact. This information is subject
to change in future editions.
According to the Feature Table Definition, an entry's sequence span has
to be covered by one source feature or a combination of several. 'Synthetic
constructs' are one type of sequence entry which typically contain several
source features. Here one of these source features spans the whole sequence
(/organism="synthetic construct"). The feature qualifier /focus is attached to
the preferred source feature and used to determine the taxonomic division. If
no translation table is specified, the organism with /focus will define the
translation table. Within an entry with several source features, only one will
exist with /focus on it.

2.3 Literature References

The references cited for an entry should be considered a pointer to the
literature and not a scientific credit for the elucidation of the
sequence. Although every effort is made to give complete reference
information, occasionally only a secondary source has been cited. This
has happened most frequently in cases where a secondary reference has
presented the data in a form easily entered. The speed and accuracy with
which data can be abstracted is very dependent on the form of presentation.
In such cases, we prefer to cite also the primary reference, and request
users who note such omissions to inform us so that the appropriate additions
may be made.

3 FORMAT OF THE DATABASE

EMBL-Bank is composed of sequence entries. Each entry corresponds to a single
contiguous sequence as contributed to the database or reported in the
literature. In some cases, entries have been assembled from several papers
reporting overlapping sequence regions. Conversely a single paper often
provides data for several entries, as when homologous sequences from different
organisms are compared.

3.1 Data Class

The data class of each entry, representing a methodological approach to the
generation of the data or a type of data, is indicated on the first (ID) line
of the entry. Each entry belongs to exactly one data class.

Class	Definition
CON	Entry constructed from segment entry sequences; if unannotated, annotation may be drawn from segment entries
PAT	Patent
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTC	High Thoughput CDNA sequencing
HTG	High Thoughput Genome sequencing
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TSA	Transcriptome Shotgun Assembly
STS	Sequence Tagged Site
STD	Standard (all entries not classified as above)

3.2 Taxonomic Division

The entries which constitute the database are grouped into taxonomic divisions, the object being to create subsets of the database which reflect areas of interest for many users.
In addition to the division, each entry contains a full taxonomic classification of the organism that was the source of the stored sequence, from kingdom down to genus and species (see below).
Each entry belongs to exactly one taxonomic division. The ID line of each entry indicates its taxonomic division, using the three letter codes shown below:

Division	Code
Bacteriophage	PHG
Environmental Sample	ENV
Fungal	FUN
Human	HUM
Invertebrate	INV
Other Mammal	MAM
Other Vertebrate	VRT
Mus musculus	MUS
Plant	PLN
Prokaryote	PRO
Other Rodent	ROD
Synthetic	SYN
Transgenic	TGN
Unclassified	UNC
Viral	VRL

3.3 Structure of an Entry
The entries in the database are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English, and the symbols and formatting employed for the base sequences themselves have been chosen for readability. Wherever possible, symbols familiar to molecular biologists have been used. At the same time, the structure is systematic enough to allow computer programs easily to read, identify, and manipulate the various types of data included.
Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data which make up the entry. In general, fixed format items have been kept to a minimum, and a more syntax-oriented structure adopted for the lines.
The two exceptions to this are the sequence data lines and the feature table lines, for which a fixed format was felt to offer significant advantages to the user. Users who write programs to process the database entries should not make any assumptions about the column placement of items on lines other than these two: all other line types are free-format.
A sample entry is shown in Figure 1.
Note that each line begins with a two-character line code, which indicates the type of information contained in the line. The currently used line types, along with their respective line codes, are listed below:

ID	identification	(begins each entry; 1 per entry)
AC	accession number	(>=1 per entry)
PR	project identifier	(0 or 1 per entry)
DT	date	(2 per entry)
DE	description	(>=1 per entry)
KW	keyword	(>=1 per entry)
OS	organism species	(>=1 per entry)
OC	organism classification	(>=1 per entry)
OG	organelle	(0 or 1 per entry)
RN	reference number	(>=1 per entry)
RC	reference comment	(>=0 per entry)
RP	reference positions	(>=1 per entry)
RX	reference cross-reference	(>=0 per entry)
RG	reference group	(>=0 per entry)
RA	reference author(s)	(>=0 per entry)
RT	reference title	(>=1 per entry)
RL	reference location	(>=1 per entry)
DR	database cross-reference	(>=0 per entry)
CC	comments or notes	(>=0 per entry)
AH	assembly header	(0 or 1 per entry)
AS	assembly information	(0 or >=1 per entry)

```
     FH - feature table header       (2 per entry)
     FT - feature table data         (>=2 per entry)
     XX - spacer line                (many per entry)
     SQ - sequence header            (1 per entry)
     CO - contig/construct line      (0 or >=1 per entry)
     bb - (blanks) sequence data     (>=1 per entry)
     // - termination line           (ends each entry; 1 per entry)
Note that some entries will not contain all of the line types, and some line
types occur many times in a single entry. As indicated, each entry begins with
an identification line (ID) and ends with a terminator line (//). The various
line types appear in entries in the order in which they are listed above
(except for XX lines which may appear anywhere between the ID and SQ lines). A
detailed description of each line type is given in the following sections.

ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="HSSP:P26205"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
```

```
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa       420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct       480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg       540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt       600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg       660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac       720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta       780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg       840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga       900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg       960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg      1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc      1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa      1140
     acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat      1200
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct      1260
     gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga      1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg      1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg      1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga      1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa      1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt      1620
     tggtgcacct aattaaataa aagaggttac tcttaacata tttttaggcc attcgttgtg      1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc      1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac      1800
     tttgaattaa aagtcttttt ttatttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa        1859
//
```

Figure 1 - A sample entry from the database

3.4 Line Structure

This section describes in detail the format of each type of line used in the database. Each line begins with a two-character line type code. This code is always followed by three blanks, so that the actual information in each line begins in character position 6.

3.4.1 The ID Line

The ID (IDentification) line is always the first line of an entry. The format of the ID line is:

ID <1>; SV <2>; <3>; <4>; <5>; <6>; <7> BP.

The tokens represent:

1. Primary accession number
2. Sequence version number
3. Topology: 'circular' or 'linear'
4. Molecule type (see note 1 below)
5. Data class (see section 3.1)
6. Taxonomic division (see section 3.2)
7. Sequence length (see note 2 below)

Note 1 - Molecule type: this represents the type of molecule as stored and can be any value from the list of current values for the mandatory mol_type source qualifier. This item should be the same as the value in the mol_type qualifier(s) in a given entry.

Note 2 - Sequence length: The last item on the ID line is the length of the sequence (the total number of bases in the sequence). This number includes base positions reported as present but undetermined (coded as "N").

An example of a complete identification line is shown below:

ID CD789012; SV 4; linear; genomic DNA; HTG; MAM; 500 BP.

3.4.2 The AC Line

The AC (ACcession number) line lists the accession numbers associated with the entry.

Examples of accession number lines are shown below:

```
 AC   X56734; S46826;
 AC   Y00001; X00001-X00005; X00008; Z00001-Z00005;
```

Each accession number, or range of accession numbers, is terminated by a semicolon. Where necessary, more than one AC line is used. Consecutive secondary accession numbers in EMBL-Bank flatfiles are shown in the form of inclusive accession number ranges.

Accession numbers are the primary means of identifying sequences providing a stable way of identifying entries from release to release. An accession number, however, always remains in the accession number list of the latest version of the entry in which it first appeared. Accession numbers allow unambiguous citation of database entries. Researchers who wish to cite entries in their publications should always cite the first accession number in the list (the "primary" accession number) to ensure that readers can find the relevant data in a subsequent release. Readers wishing to find the data thus cited must look at all the accession numbers in each entry's list.

Secondary accession numbers: One reason for allowing the existence of several accession numbers is to allow tracking of data when entries are merged or split. For example, when two entries are merged into one, a "primary" accession number goes at the start of the list, and those from the merged entries are added after this one as "secondary" numbers.

Example: AC X56734; S46826;

Similarly, if an existing entry is split into two or more entries (a rare occurrence), the original accession number list is retained in all the derived entries.

An accession number is dropped from the database only when the data to which it was assigned have been completely removed from the database.

3.4.3 The PR Line

The PR (PRoject) line shows the International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry. Full details of INSDC Project are available at http://www.ebi.ac.uk/ena/about/page.php?page=project_guidelines.

Example: PR Project:17285;

3.4.4 The DT Line

The DT (DaTe) line shows when an entry first appeared in the database and when it was last updated. Each entry contains two DT lines, formatted as follows:

```
DT   DD-MON-YYYY (Rel. #, Created)
DT   DD-MON-YYYY (Rel. #, Last updated, Version #)
```

The DT lines from the above example are:

```
DT   12-SEP-1991 (Rel. 29, Created)
DT   13-SEP-1993 (Rel. 37, Last updated, Version 8)
```

The date supplied on each DT line indicates when the entry was created or Last updated; that will usually also be the date when the new or modified Entry became publicly visible via the EBI network servers. The release number indicates the first quarterly release made *after* the entry was created or last updated. The version number appears only on the "Last updated" DT line.

The absolute value of the version number is of no particular significance; its purpose is to allow users to determine easily if the version of an entry which they already have is still the most up to date version. Version numbers are incremented by one every time an entry is updated; since an entry may be updated several times before its first appearance in a quarterly release, the version number at the time of its first release appearance may be greater than one. Note that because an entry may also be updated several times between two quarterly releases, there may be gaps in the sequence of version numbers which appear in consecutive releases.

If an entry has not been updated since it was created, it will still have two DT lines and the "Last updated" line will have the same date (and release number) as the "Created" line.

3.4.5 The DE Line

The DE (Description) lines contain general descriptive information about the sequence stored. This may include the designations of genes for which the sequence codes, the region of the genome from which it is derived, or other information which helps to identify the sequence. The format for a DE line is:

```
DE   description
```

The description is given in ordinary English and is free-format. Often, more than one DE line is required; when this is the case, the text is divided only between words. The description line from the example above is

```
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
```

The first DE line generally contains a brief description, which can stand alone for cataloguing purposes.

3.4.6 The KW Line

The KW (KeyWord) lines provide information which can be used to generate cross-reference indexes of the sequence entries based on functional, structural, or other categories deemed important.

The format for a KW line is:

```
     KW   keyword[; keyword ...].
```

More than one keyword may be listed on each KW line; the keywords are separated by semicolons, and the last keyword is followed by a full stop. Keywords may consist of more than one word, and they may contain embedded blanks and stops. A keyword is never split between lines.

An example of a keyword line is:

```
     KW   beta-glucosidase.
```

The keywords are ordered alphabetically; the ordering implies no hierarchy of importance or function. If an entry has no keywords assigned to it, it will contain a single KW line like this:

```
     KW   .
```

3.4.7 The OS Line

The OS (Organism Species) line specifies the preferred scientific name of the organism which was the source of the stored sequence. In most cases this is done by giving the Latin genus and species designations, followed (in parentheses) by the preferred common name in English where known. The format is:

```
     OS   Genus species (name)
```

In some cases, particularly for viruses and genetic elements, the only accepted designation is a simple name such as "Canine adenovirus type 2". In these cases only this designation is given. The species line from the example is:

```
     OS   Trifolium repens (white clover)
```

Hybrid organisms are classified in their own right. A rat/mouse hybrid, for example, would appear as follows:

```
     OS   Mus musculus x Rattus norvegicus
     OC   (OC for mouse)
```

If the source organism is unknown but has been/will be cultured, the OS line will contain a unique name derived from the what is known of the classification. The unique name serves to identify the database entry, which will be updated once the full classification is known. In the case of an unknown bacterium, for example:

```
     OS   unidentified bacterium B8
     OC   Bacteria.
```

For environmental samples where there is no intention to culture the organism and complete taxonomy cannot be determined, collective names are used in the OS line and the classification given extends down to the most resolved taxonomic node possible, for example:

```
     OS   uncultured proteobacterium
     OC   Bacteria; Proteobacteria; environmental samples.
```

For naturally occurring plasmids the OS/OC lines will contain the source organism and the plasmid name will appear on the OG line. For example:

```
     OS   Escherichia coli
     OC   Prokaryota; ... Enterobacteriaceae.
     XX
     OG   Plasmid colE1
```

For artificial plasmids the OS line will be "OS Cloning vector" and the sequence will be classified as an artificial sequence. For example:

```
     OS   Cloning vector M13plex17
```

```
     OC   Artificial sequences; vectors.

Where only a naturally occurring part of a plasmid is reported, the plasmid
name will appear on the OG line and the OS/OC lines will describe the natural
source.
For example:
     OS   Escherichia coli
     OC   Prokaryota; ... Enterobacteriaceae.
     XX
     OG   Plasmid pUC8
```

3.4.8 The OC Line

```
The OC (Organism Classification) lines contain the taxonomic classification
Of the source organism as described in Section 2.2 above.
The classification is listed top-down as nodes in a taxonomic tree in which
the most general grouping is given first.  The classification may be
distributed over several OC lines, but nodes are not split or hyphenated
between lines. The individual items are separated by semicolons and the
list is terminated by a full stop. The format for the OC line is:
     OC   Node[; Node...].

Example classification lines:
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   euphyllophytes; Spermatophyta; Magnoliophyta; eudicotyledons; Rosidae;
OC   Fabales; Fabaceae; Papilionoideae; Trifolium.
```

3.4.9 The OG Line

```
The OG (OrGanelle) linetype indicates the sub-cellular location of non-nuclear
sequences.  It is only present in entries containing non-nuclear sequences
and appears after the last OC line in such entries.
The OG line contains
a) one data item (title cased) from the controlled list detailed under the
/organelle qualifier definition in the Feature Table Definition document
that accompanies this release or
b) a plasmid name.
Examples include "Mitochondrion", "Plastid:Chloroplast" and "Plasmid pBR322".

For example, a chloroplast sequence from Euglena gracilis would appear as:
     OS   Euglena gracilis (green algae)
     OC   Eukaryota; Planta; Phycophyta; Euglenophyceae.
     OG   Plastid:Chloroplast
```

3.4.10 The Reference (RN, RC, RP, RX, RG, RA, RT, RL) Lines

```
These lines comprise the literature citations within the database.
The citations provide access to the papers from which the data has been
abstracted. The reference lines for a given citation occur in a block, and
are always in the order RN, RC, RP, RX, RG, RA, RT, RL. Within each such
reference block the RN line occurs once, the RC, RP and RX lines occur zero
or more times, and the RA, RT, RL lines each occur one or more times.
If several references are given, there will be a reference block for each.
Example of references :

RN   [5]
RP   1-1859
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.).";
RL   Plant Mol. Biol. 17:209-219(1991).

The formats of the individual lines are explained in the following
paragraphs.
RN   [2]
RP   1-1657990
RG   Prochlorococcus genome consortium
RA   Larimer F.;
RT   ;
RL   Submitted (03-JUL-2003) to the EMBL/GenBank/DDBJ databases.
RL   Larimer F., DOE Joint Genome Institute, Production Genomics Facility,
RL   2800 Mitchell Drive, Walnut Creek, CA 94598, USA, and the Genome
```

```
RL   Analysis Group, Oak Ridge National Laboratory, 1060 Commerce Park Drive,
RL   Oak Ridge, TN 37831, USA;
```

3.4.10.1 The RN Line

The RN (Reference Number) line gives a unique number to each reference Citation within an entry. This number is used to designate the reference in comments and in the feature table. The format of the RN line is:

```
RN   [n]
```

The reference number is always enclosed in square brackets. Note that the set of reference numbers which appear in an entry does not necessarily form a continuous sequence from 1 to n, where the entry contains "n" references. As references are added to and removed from an entry, gaps may be introduced into the sequence of numbers. The important point is that once an RN number has been assigned to a reference within an entry it never changes. The reference number line in the example above is:

```
RN   [5]
```

3.4.10.2 The RC Line

The RC (Reference Comment) linetype is an optional linetype which appears if The reference has a comment. The comment is in English and as many RC lines as are required to display the comment will appear. They are formatted thus:

```
RC   comment
```

3.4.10.3 The RP Line

The RP (Reference Position) linetype is an optional linetype which appears if one or more contiguous base spans of the presented sequence can be attributed to the reference in question. As many RP lines as are required to display the base span(s) will appear.

The base span(s) indicate which part(s) of the sequence are covered by the reference. Note that the numbering scheme is for the sequence as presented in the database entry (i.e. from 5' to 3' starting at 1), not the scheme used by the authors in the reference should the two differ. The RP line is formatted thus:

```
RP   i-j[, k-l...]
```

The RP line in the example above is:

```
RP   1-1859
```

3.4.10.4 The RX Line

The RX (reference cross-reference) linetype is an optional linetype which contains a cross-reference to an external citation or abstract resource. For example, if a journal citation exists in the PUBMED database, there will be an RX line pointing to the relevant PUBMED identifier.

The format of the RX line is as follows:

```
RX  resource_identifier; identifier.
```

The first item on the RX line, the resource identifier, is the abbreviated name of the data collection to which reference is made. The current set of cross-referenced resources is:

Resource ID	Fullname
PUBMED	PUBMED bibliographic database (NLM)
DOI	Digital Object Identifier (International DOI Foundation)
AGRICOLA	US National Agriculture Library (NAL) of the US Department of Agriculture (USDA)

The second item on the RX line, the identifier, is a pointer to the entry in the external resource to which reference is being made. The data item used as the primary identifier depends on the resource being referenced.

For example:

```
RX   DOI; 10.1016/0024-3205(83)90010-3.
RX   PUBMED; 264242.
```

Note that further details of DOI are available at http://www.doi.org/. URLs formulated in the following way are resolved to the correct full text URLs:

```
http://dx.doi.org/
eg. http:/dx.doi.org/10.1016/0024-3205(83)90010-3
```

3.4.10.5 The RG Line

The RG (Reference Group) lines list the working groups/consortia that produced the record. RG line is mainly used in submission reference blocks, but could also be used in paper reference if the working group is cited as an author in the paper.

3.4.10.6 The RA Line

The RA (Reference Author) lines list the authors of the paper (or other work) cited. All of the authors are included, and are listed in the order given in the paper. The names are listed surname first followed by a blank followed by initial(s) with stops. Occasionally the initials may not be known, in which case the surname alone will be listed. The author names are separated by commas and terminated by a semicolon; they are not split between lines. The RA line in the example is:

```
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
```

As many RA lines as necessary are included for each reference.

3.4.10.7 The RT Line

The RT (Reference Title) lines give the title of the paper (or other work) as exactly as is possible given the limitations of computer character sets. Note that the form used is that which would be used in a citation rather than that displayed at the top of the published paper. For instance, where journals capitalise major title words this is not preserved. The title is enclosed in double quotes, and may be continued over several lines as necessary. The title lines are terminated by a semicolon. The title lines from the example are:

```
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
```

Greek letters in titles are spelled out; for example, a title in an entry would contain "kappa-immunoglobulin" even though the letter itself may be present in the original title. Similar simplifications have been made in other cases (e.g. subscripts and superscripts). Note that the RT line of a citation which has no title (such as a submission to the database) contains only a semicolon.

3.4.10.8 The RL Line

The RL (Reference Location) line contains the conventional citation information for the reference. In general, the RL lines alone are sufficient to find the paper in question. They include the journal, volume number, page range and year for each paper.

Journal names are abbreviated according to existing ISO standards (International Standard Serial Number)

The format for the location lines is:

```
     RL   journal vol:pp-pp(year).
```

Thus, the reference location line in the example is:

```
     RL   Plant Mol. Biol. 17:209-219(1991).
```

Very occasionally a journal is encountered which does not consecutively number pages within a volume, but rather starts the numbering anew for each issue number. In this case the issue number must be included, and the format becomes:

```
     RL   journal vol(no):pp-pp(year).
```

If a paper is in press, the RL line will appear with such information as we have available, the missing items appearing as zeros. For example:

```
     RL   Nucleic Acids Res. 0:0-0(2004).
```

This indicates a paper which will be published in Nucleic Acids Research at some point in 2004, for which we have no volume or page information. Such references are updated to include the missing information when it becomes available.

Another variation of the RL line is used for papers found in books or other similar publications, which are cited as shown below:

```
     RA   Birnstiel M., Portmann R., Busslinger M., Schaffner W.,
     RA   Probst E., Kressmeann A.;
     RT   "Functional organization of the histone genes in the
     RT   sea urchin Psammechinus:  A progress report";
     RL   (in) Engberg J., Klenow H., Leick V. (Eds.);
     RL   SPECIFIC EUKARYOTIC GENES:117-132;
     RL   Munksgaard, Copenhagen (1979).
```

Note specifically that the line where one would normally encounter the journal location is replaced with lines giving the bibliographic citation of the book. The first RL line in this case contains the designation "(in)", which indicates that this is a book reference.

The following examples illustrate RL line formats that are used for data submissions:

```
     RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
     RL   M.A. Hughes, UNIVERSITY OF NEWCASTLE UPON TYNE, MEDICAL SCHOOL, NEW
     RL   CASTLE UPON TYNE, NE2  4HH, UK
```

Submitter address is always included in new entries, but some older

submissions do not have this information.
RL lines take another form for thesis references.
For example:
RL Thesis (1999), Department of Genetics,
RL University of Cambridge, Cambridge, U.K.
For an unpublished reference, the RL line takes the following form:
RL Unpublished.
Patent references have the following form:
RL Patent number EP0238993-A/3, 30-SEP-1987.
RL BAYER AG.
The words "Patent number" are followed by the patent application number, the patent type (separated by a hyphen), the sequence's serial number within the patent (separated by a slash) and the patent application date. The subsequent RL lines list the patent applicants, normally company names.
Finally, for journal publications where no ISSN number is available for the journal (proceedings and abstracts, for example), the RL line contains the designation "(misc)" as in the following example.
RL (misc) Proc. Vth Int. Symp. Biol. Terr. Isopods 2:365-380(2003).

3.4.11 The DR Line
The DR (Database Cross-reference) line cross-references other databases which contain information related to the entry in which the DR line appears. For example, if an EMBL-Bank sequence is cited in the IMGT/LIGM database there will be a DR line pointing to the relevant IMGT/LIGM entry.
The format of the DR line is as follows:
DR database_identifier; primary_identifier; secondary_identifier.
The first item on the DR line, the database identifier, is the abbreviated name of the data collection to which reference is made.
The second item on the DR line, the primary identifier, is a pointer to the entry in the external database to which reference is being made.
The third item on the DR line is the secondary identifier, if available, from the referenced database.
An example of a DR line is shown below:
DR MGI; 98599; Tcrb-V4.

3.4.12 The AH Line (in TPA and TSA records only)
Third Party Annotation (TPA) and Transcriptome Shotgun Assembly (TSA) records may include information on the composition of their sequences to show which spans originated from which contributing primary sequences. The AH (Assembly Header) line provides column headings for the assembly information. The lines contain no data and may be ignored by computer programs.
The AH line format is:
AH LOCAL_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP

3.4.13 The AS Line (in TPA and TSA records)
The AS (ASsembly Information) lines provide information on the composition of a TPA or TSA sequence. These lines include information on local sequence spans (those spans seen in the sequence of the entry showing the AS lines) plus identifiers and base spans of contributing primary sequences (for EMBL-Bank primary entries only).

a) LOCAL_SPAN base span on local sequence shown in entry
b) PRIMARY_IDENTIFIER acc.version of contributing EMBL-Bank sequence(s) or trace identifier for Trace Archive sequence(s)
c) PRIMARY_SPAN base span on contributing EMBL-Bank primary sequence or not_available for Trace Archive sequence(s)
d) COMP 'c' is used to indicate that contributing sequence originates from complementary strand in primary entry

Example:

AH	LOCAL_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
AS	1-426	AC004528.1	18665-19090	
AS	427-526	AC001234.2	1-100	c
AS	527-1000	TI55475028	not_available	

3.4.14 The CO Line (in CON records only)
Con(structed) sequences in the CON data classes represent complete

```
chromosomes, genomes and other long sequences constructed from segment entries.
CON data class entries do not contain sequence data per se, but rather the
assembly information on all accession.versions and sequence locations relevant
to building the constructed sequence. The assembly information is represented in
the CO lines.
Example:
CO   join(Z99104.1:1..213080,Z99105.1:18431..221160,Z99106.1:13061..209100,
CO   Z99107.1:11151..213190,Z99108.1:11071..208430,Z99109.1:11751..210440,
CO   Z99110.1:15551..216750,Z99111.1:16351..208230,Z99112.1:4601..208780,
CO   Z99113.1:26001..233780,Z99114.1:14811..207730,Z99115.1:12361..213680,
CO   Z99116.1:13961..218470,Z99117.1:14281..213420,Z99118.1:17741..218410,
CO   Z99119.1:15771..215640,Z99120.1:16411..217420,Z99121.1:14871..209510,
CO   Z99122.1:11971..212610,Z99123.1:11301..212150,Z99124.1:11271..215534)

Gaps of undefined length are represented using the expression 'gap(unk100)'.
These gaps contribute to the sequence length for the entry (as shown in the
ID line).
Example: CO   join(AL358912.1:1..39187,gap(unk100),AL137130.1:1..40815,...
Gaps of defined length are represented via 'gap(#)' where # is the
gap length. These gaps also contribute to the sequence length for the entry (as
shown in the ID line).
Example: CO   AE005330.1:61..14164,AE005331.1:61..3773,gap(4001),...
Below are the relevant sections of a Bacillus subtilis CON entry providing
construct information for the assembly of the Bacillus subtilis genome.

ID   AL009126; SV 2; circular; genomic DNA; CON; PRO; 4214630 BP.
XX
AC   AL009126;
XX
DT   18-JUL-2002 (Rel. 72, Created)
DT   07-JUL-2003 (Rel. 76, Last updated, Version 3)
XX
DE   Bacillus subtilis complete genome.
XX
KW   complete genome.
XX
OS   Bacillus subtilis subsp. subtilis str. 168
OC   Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
...
CITATION INFORMATION
...
FH   Key             Location/Qualifiers
FH
FT   source          1..4214630
FT                   /organism="Bacillus subtilis subsp. subtilis str. 168"
FT                   /strain="168"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:224308"
XX
CO   join(Z99104.2:1..213080,Z99105.2:51..202768,Z99106.2:31..195912,
CO   Z99107.2:51..202089,Z99108.2:51..197409,Z99109.2:41..198743,
CO   Z99110.2:41..201241,Z99111.2:41..191980,Z99112.2:41..204263,
CO   Z99113.2:41..207829,Z99114.2:41..192961,Z99115.2:51..201375,
CO   Z99116.2:31..204537,Z99117.2:31..199173,Z99118.2:31..200707,
CO   Z99119.2:51..199922,Z99120.2:51..201059,Z99121.2:51..194692,
CO   Z99122.2:51..200690,Z99123.2:31..201139,Z99124.2:51..203901)
//
```

3.4.15 The FH Line

The FH (Feature Header) lines are present only to improve readability of an entry when it is printed or displayed on a terminal screen. The lines contain no data and may be ignored by computer programs. The format of these lines is always the same:

```
     FH   Key             Location/Qualifiers
     FH
```

The first line provides column headings for the feature table, and the second line serves as a spacer. If an entry contains no feature table (i.e. no FT lines - see below), the FH lines will not appear.

3.4.16 The FT Line

The FT (Feature Table) lines provide a mechanism for the annotation of the sequence data. Regions or sites in the sequence which are of interest are listed in the table. In general, the features in the feature table represent signals or other characteristics reported in the cited references. In some cases, ambiguities or features noted in the course of data preparation have been included. The feature table is subject to expansion or change as more becomes known about a given sequence.
Feature Table Definition Document:
A complete and definitive description of the feature table is given in the document "The DDBJ/EMBL/GenBank Feature Table: Definition".
URL: ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/FT_current.txt
Much effort is expended in the design of the feature table to try to ensure that it will be self-explanatory to the human reader, and we therefore expect that the official definition document will be of interest mainly to software developers rather than to end-users of the database.
Annotation Guides:
To help submitters annotate their sequences annotation guides are available from the EMBL-EBI Web servers:
WebFeat: A complete list of feature table key and qualifier definitions, providing full explanations of their use.
URL: http://www.ebi.ac.uk/embl/WebFeat/index.html
EMBL-Bank Annotation Examples.
A selection of EMBL-Bank approved feature table annotations for some common biological sequences (i.e., ribosomal RNA, mitochondrial genome).
URL: http://www.ebi.ac.uk/embl/Standards/web/index.html

3.4.17 The SQ Line
The SQ (SeQuence header) line marks the beginning of the sequence data and Gives a summary of its content. An example is:

```
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
```

As shown, the line contains the length of the sequence in base pairs followed by its base composition. Bases other than A, C, G and T are grouped together as "other". (Note that "BP" is also used for single stranded RNA sequences, which is not strictly accurate, but has been used for consistency of format.) This information can be used as a check on accuracy or for statistical purposes. The word "Sequence" is present solely as a marker for readability.

3.4.18 The Sequence Data Line
The sequence data line has a line code consisting of two blanks. The sequence is written 60 bases per line, in groups of 10 bases separated by a blank character, beginning at position 6 of the line. The direction listed is always 5' to 3', and wherever possible the non-coding strand (homologous to the message) has been stored. Columns 73-80 of each sequence line contain base numbers for easier reading and quick location of regions of interest. The numbers are right justified and indicate the number of the last base on each line.
An example of a data line is:

```
aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
```

The characters used for the bases correspond to the IUPAC-IUB Commission recommendations (see appendices).

3.4.19 The CC Line
CC lines are free text comments about the entry, and may be used to convey any sort of information thought to be useful that is unsuitable for inclusion in other line types.

3.4.20 The XX Line
The XX (spacer) line contains no data or comments. Its purpose is to make an entry easier to read on a page or terminal screen by setting off the various types of information in appropriate groupings. XX is used instead of blank lines to avoid confusion with the sequence data lines. The XX lines can always be ignored by computer programs.

3.4.21 The // Line
The // (terminator) line also contains no data or comments. It designates the end of an entry.

APPENDIX A
STANDARD BASE CODES

These are the official IUPAC-IUB single-letter base codes (reference 1 below).

Code	Base Description	
G	Guanine	
A	Adenine	
T	Thymine	
C	Cytosine	
R	Purine	(A or G)
Y	Pyrimidine	(C or T or U)
M	Amino	(A or C)
K	Ketone	(G or T)
S	Strong interaction	(C or G)
W	Weak interaction	(A or T)
H	Not-G	(A or C or T) H follows G in the alphabet
B	Not-A	(C or G or T) B follows A
V	Not-T (not-U)	(A or C or G) V follows U
D	Not-C	(A or G or T) D follows C
N	Any	(A or C or G or T)

APPENDIX B
MODIFIED BASE CODES

The following table is taken from Sprinzl M. and Gauss D.H.
(reference 2 below). The codes appear in database entries as values for the
/mod_base qualifier in the feature table.

Code	Modified Base
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxymethyl)uridine
cm	2'-O-methylcytidine
cm5u	5-carbamoylmethyluridine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
d	dihydrouridine
fm	2'-O-methylpseudouridine
gal q	beta,D-galactosylqueuosine
gm	2'-O-methylguanosine
i	inosine
i6a	N6-isopentenyladenosine
m1a	1-methyladenosine
m1am	2'-O-methyl-1-methyladenosine
m1f	1-methylpseudouridine
m1g	1-methylguanosine
m1i	1-methylinosine
m22g	2,2-dimethylguanosine
m22gm	N2,N2,3'-trimethylguanosine
m2a	2-methyladenosine
m2g	2-methylguanosine
m3c	3-methylcytidine
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine
mam5s2u	5-methylaminomethyl-2-thiouridine
mam5u	5-methylaminomethyluridine
man q	beta,D-mannosylqueuosine
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine
mcm5u	5-methoxycarbonylmethyluridine
mo5u	5-methoxyuridine
ms2i6a	2-methylthio-N6-isopentenyladenosine
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurin-6-yl)carbamoyl)

	threonine
mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl) threonine
mv	uridine-5-oxoacetic acid methylester
o5u	uridine-5-oxyacetic acid(v)
osyw	wybutoxosine
p	pseudouridine
q	queuosine
s2c	2-thiocytidine
s2t	5-methyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
t	5-methyluridine
t6a	N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine
tm	2'-O-methyl-5-methyluridine
um	2'-O-methyluridine
x	3-(3-amino-3-carboxypropyl)uridine,(acp3)U
yw	wybutosine

APPENDIX C
REFERENCES FOR ABBREVIATIONS AND SYMBOLS

1. Cornish-Bowden A., Nucl. Acids Res. 13:3021-3030(1985).
2. Sprinzl M., and Gauss D.H., "Compilation of tRNA Sequences", Nucl. Acids Res. 10:r1-r55(1982).

Revised: 9-MAR-2011